Departamento de Informática y Análisis Numérico

Programa de Doctorado en Computación Avanzada, Energía y Plasma

MODELO DE EVALUACIÓN ENTRE PARES CON EL ENFOQUE DE ANÁLISIS DE SENTIMIENTO

PEER ASSESSMENT MODEL WITH THE SENTIMENT ANALYSIS APPROACH

Tesis Doctoral presentada por:

Maricela Pinargote Ortega

Directores:

Dr. Sebastián Ventura Soto
Dr. Jaime Meza Hormaza

Córdoba Mayo de 2023

TITULO: *MODELO DE EVALUACIÓN ENTRE PARES CON EL ENFOQUE DE ANÁLISIS DE SENTIMIENTO*

AUTOR: *Jemmer Maricela Pinargote Ortega*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

TÍTULO DE LA TESIS: Modelo de evaluación entre pares con el enfoque de análisis de sentimiento

DOCTORANDO/A: Jenmer Maricela Pinargote Ortega

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

El trabajo de investigación realizado por la doctoranda Maricela Pinargote Ortega dirigido por los doctores Sebastián Ventura Soto y Jaime Meza Hormaza, cumple los requisitos formales de originalidad y calidad, y mantiene el rigor científico y académico exigibles como para que sea presentado a defensa pública, y evaluado en Comisión Académica, en orden a la posible adquisición del grado de Doctor.

El trabajo ofrece a la comunidad científica información válida de evaluación entre pares basada en análisis de sentimiento. Concretamente, en la tesis doctoral se ha realizado una revisión de la literatura científica sobre evaluación entre pares, minería de texto y técnicas de computación blanda. Se aplicó la metodología de investigación-acción. Durante el desarrollo del modelo predictivo se obtuvo dos conjuntos de datos en español, uno de evaluación de tarea y otro de evaluación de calidad de la evaluación. Se aplicó el enfoque de aprendizaje automático supervisado para obtener una puntuación de sentimiento correspondiente a una retroalimentación textual específica. Seguidamente, se obtuvo un modelo de cálculo que contribuyó a mejorar la confiabilidad del proceso de evaluación entre pares, donde la puntuación de cada criterio de evaluación de tarea y evaluación de calidad de evaluación se generó con la técnica computacional de lógica difusa correlacionando puntuación numérica y sentimiento. Se probó la validez del modelo propuesto en tres escenarios de educación superior: virtual asincrónico, virtual sincrónico y presencial. Finalmente, se obtuvo un modelo de calibración que contribuyó a mejorar la fiabilidad en el proceso de evaluación entre pares, ya que, mediante el ajuste de la puntuación individual de cada tarea en función del rendimiento y índice (rating) de confianza del evaluador, se logró que la relación entre la puntación del colectivo y puntuación que proporciona el docente tendiera a subir.

Como principales frutos del trabajo realizado se han derivado artículos publicados/ aceptados en conferencias y revistas científicas (JCR/SJR):

1. Pinargote-Ortega M., Bowen-Mendoza L., Meza J., and Ventura S., “Peer assessment using soft computing techniques,” J. Comput. High. Educ., vol. 33, no. 3, pp. 684–726, 2021, doi: 10.1007/s12528-021-09296-w. Impact Factor: 4.045 (Q1).
2. Pinargote Ortega M., Mendoza Bowen L., Hormaza J., and Ventura Soto S., “Accuracy’ measures of sentiment analysis algorithms for spanish corpus generated in peer assessment,” ACM Int. Conf. Proceeding Ser., 2020, doi: 10.1145/3410352.3410838.
3. Pinargote-Ortega M., Bowen-Mendoza L., Meza J. and Ventura S., "Sentiment Analysis Techniques for Peer Feedback: A Review," 2023 Ninth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 2023, pp. 1-8, doi: 10.1109/ICEDEG58167.2023.10122085.
4. Pinargote-Ortega M., Bowen-Mendoza L., Meza J., and Ventura S., “Sentiment Analysis Techniques for Peer Feedback: A Review”. Aceptado para publicación en AIP (American Institute of Physics) Conference Proceedings. 2023.
5. Pinargote-Ortega M., Bowen-Mendoza L., Meza J., and Ventura S., “Peer Feedback Sentiment Analysis Prototype”. Aceptado para publicación en RISTI (Revista Ibérica de Sistemas y Tecnologías de la Información). 2023.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 15 de mayo de 2023

Firma de los directores

Firmado digitalmente por VENTURA SOTO SEBASTIAN EMILIO - 30510000V
Fecha: 2023.05.26 19:51:26 +02'00'

Fdo: Sebastián Ventura Soto

Firmado electrónicamente por: JAIME ALCIDES MEZA HORMAZA

Fdo: Jaime Meza Hormaza

La tesis titulada “Modelo de evaluación entre pares con el enfoque de análisis de sentimiento”, que presenta Maricela Pinargote Ortega para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado "Computación Avanzada, Energía y Plasmas", del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, bajo la dirección de los doctores Sebastián Ventura Soto y Jaime Meza Hormaza cumpliendo, en su opinión, los requisitos exigidos a este tipo de trabajos.

Córdoba, mayo de 2023

La Doctoranda

Firmado electrónicamente por:
JENMER MARICELA
PINARGOTE ORTEGA

Fdo: Maricela Pinargote Ortega

El Director

Firmado
digitalmente por
VENTURA SOTO
SEBASTIAN
EMILIO -
30510000V
Fecha: 2023.05.26
19:50:51 +02'00'

Fdo: Sebastián Ventura Soto

El Director

Firmado electrónicamente por:
JAIME ALCIDES MEZA
HORMAZA

Fdo: Meza Hormaza

Dedico este trabajo a mi familia.

En especial a mis angelit@s:

Luis, Andrés, Jean Carlos, Mily, Jolieth, Mía Sabella y Luis José.

Como ejemplo que con perseverancia se obtiene grandes metas.

Augurando que lograrán mucho más.

AGRADECIMIENTOS

A mis directores de tesis Dr. Sebastián Ventura Soto y Dr. Jaime Meza Hormaza por todo el apoyo que me han brindado, por su disposición y por compartir conmigo sus conocimientos y experiencias.

A mi emblemática Universidad Técnica de Manabí, al señor Rector Dr. Santiago Quiroz Fernández, ex Rector Dr. Vicente Véliz Briones, y a las autoridades de la Facultad de Ciencias Informáticas por brindarme el apoyo y las condiciones para realizar mis estudios de doctorado.

A la Universidad de Córdoba, la cual ha sido un importante pilar en mi formación como investigadora.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ACRÓNIMOS

Acrónimos	Significado
SA	Sentiment Analysis
NLP	Natural Language Processing
TF-IDF	Term Frequency-Inverse Document Frequency
SW	Stop-Words
BoW	Bag of Words
NB	Naïve Bayes
K-NN	K-Nearest Neighbor
MNB	Multinomial Naive Bayes
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
DT	Decision Trees
VE	Vote Ensemble
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
SOM	Smallest of Maximum
MOM	Middle of Maximum
LOM	Largest of Maximum

RESUMEN

La evaluación de los trabajos de respuesta abierta es una tarea que debe ser realizada por un experto; sin embargo, suponen una importante carga de trabajo de corrección para el docente. En este contexto, la evaluación entre pares se ha considerado como un enfoque alternativo para abordar el problema. Este tipo de evaluación no solo proporciona la reducción en la carga de trabajo de corrección, sino que también aporta beneficios adicionales, como la posibilidad de que el estudiante verifique diferentes soluciones para un mismo problema y la provisión de retroalimentaciones útiles.

Los cambios en los paradigmas educativos han promovido la integración de métodos de evaluación que pretenden ir más allá de la evaluación de conocimientos (sumativa), que estén más integrados en el proceso de formación y aprendizaje (formativa). La evaluación formativa contribuye significativamente en la calidad de aprendizaje que los estudiantes obtienen al dar y recibir retroalimentación, y en el acceso inmediato que los docentes pueden tener sobre el progreso de la clase. Las instituciones educativas actualmente buscan obtener el conocimiento inmerso de estos textos no estructurado. Por lo tanto, el objetivo general de esta tesis ha sido diseñar un modelo de evaluación entre pares, que coadyuve a los docentes a mejorar sus procesos de enseñanza-aprendizaje mediante métodos de análisis de sentimiento.

Se aplicó la metodología de diseño investigación-acción, en primer lugar, se realizó el estado del arte sobre evaluación entre pares, minería de texto y técnicas de computación blanda. Subsecuentemente, se diseñó un modelo que combina la evaluación entre pares con el aprendizaje colaborativo y el método calibrado en varias fases: a) se formó grupos de estudiantes que participan en realizar el trabajo de manera colaborativa, con la finalidad de tener grupos similares, pero tener diferencias individuales en el proceso de evaluación entre pares para beneficiarse de la colaboración entre estudiantes; b) se diseñó una rúbrica para la recolección de datos, donde los evaluadores evaluaron aspectos específicos del trabajo, proporcionando por cada criterio una puntuación numérica y retroalimentación textual; c) los evaluados evaluaron la calidad de evaluación de la tarea (evaluación inversa) para obtener el rating de confianza del evaluador; d) los grupos corrigieron el trabajo basándose en las retroalimentaciones dadas por los evaluadores en la primera ronda (evaluación en dos rondas); e) la puntuación de evaluación de tarea se calibró en función del rendimiento e índice (rating) de confianza del evaluador.

Durante el desarrollo del modelo se obtuvo dos conjuntos de datos en español, uno de evaluación de tarea y otro de evaluación de calidad de la evaluación. Se aplicó el enfoque de

aprendizaje automático supervisado para obtener una puntuación de sentimiento correspondiente a una retroalimentación textual específica. Se analizó distintas técnicas de minería de texto y procesamiento de lenguaje natural sobre la tarea de clasificación de sentimiento como Bag of Words, combinaciones de (N-Grams + Term Frequency-Inverse Document Frequency + Stop-Words), y Word2Vec/Glove pre-entrenados para formar los distintos vocabularios. Se evaluó algoritmos de aprendizaje automático clásico (Naïve Bayes, Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, Decision Trees), de aprendizaje automático moderno (Vote Ensemble), y de aprendizaje profundo (Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM)). Se obtuvo dos modelos predictivos con mejor rendimiento. Un modelo con Bi-LSTM utilizando representación de Glove, para predecir la puntuación de sentimiento de la retroalimentación textual de evaluación de tarea; y un modelo con LSTM utilizando representación de Glove, para predecir la puntuación de sentimiento de la retroalimentación textual de evaluación de calidad de la evaluación.

Seguidamente, se obtuvo un modelo de cálculo que contribuyó a mejorar la confiabilidad del proceso de evaluación entre pares. La puntuación de cada criterio de evaluación de tarea y evaluación de calidad de evaluación se generó con la técnica computacional de lógica difusa correlacionando puntuación numérica y sentimiento, determinando que los métodos de defuzzificación (máximo más chico, media de máximo y máximo más grande) fueron los más apropiados para este estudio. La puntuación individual de cada evaluador se obtuvo con cálculos de media de todos los criterios. La puntuación del colectivo de evaluación de tarea y rating de confianza del evaluador se obtuvo con cálculos de media/mediana del conjunto de puntuaciones individuales, determinando que la mediana tiene el mejor ajuste para generar una puntuación del colectivo confiable.

Se probó la validez del modelo propuesto en 3 escenarios de educación superior: virtual asincrónico, virtual sincrónico y presencial. Se correlacionó mediante Pearson la puntación que recibe el estudiante del colectivo con la puntuación que proporciona el docente, obteniendo similaridad fuerte en el 8% de las actividades en virtual asincrónico (r=0.718-0.790), en el 25% de las actividades en virtual sincrónico (r=0.741 a 0.971) y en el 40% de las actividades en presencial (r=0.780 a 0.951), determinando que el modelo se puede aplicar en todos los escenarios de educación evaluados, y con mayor efectividad en el presencial.

Finalmente, se obtuvo un modelo de calibración que contribuyó a mejorar la fiabilidad en el proceso de evaluación entre pares, ya que, mediante el ajuste de la puntuación individual de

cada tarea en función del rendimiento y índice (rating) de confianza del evaluador, se logró que la relación entre la puntación del colectivo y puntuación que proporciona el docente tendiera a subir el 46% de las actividades en escenario virtual asincrónico, 69% en virtual sincrónico y 60% en presencial.

Además, se evaluó si existe mejora del rendimiento estudiantil en la segunda ronda aplicando el modelo en el proceso de evaluación entre pares, mediante la prueba t de Student, se determinó que el 100% de las actividades evaluadas obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, el incremento en la segunda ronda del rendimiento del estudiante en virtual asincrónico fue de 3%-12%, en virtual sincrónico de 7%-22%, y en presencial de 15%-34%. En tal sentido, el modelo de evaluación entre pares basado en análisis de sentimiento podría implementarse como una herramienta pedagógica para apoyar al docente en enriquecer el proceso de enseñanza-aprendizaje, ya que los estudiantes dieron y recibieron retroalimentaciones detalladas sobre lo correcto o incorrecto de un trabajo específico, y pudieron refutar sobre las retroalimentaciones dadas; lo que a su vez indujo que mejoraran el trabajo y el rendimiento en la segunda ronda.

ABSTRACT

The assessment of open response work is a task that must be carried out by an expert; however, they represent a significant correction workload for the teacher. In this context, peer assessment has been considered as an alternative approach to address the problem. This type of assessment not only reduces the correction workload but also brings additional benefits, such as the possibility for the student to verify different solutions for the same problem and the provision of useful feedback.

Changes in educational paradigms have promoted the integration of assessment methods that aim to go beyond (summative) knowledge assessment, which is more integrated into the training and learning process (formative). Formative assessment contributes significantly to the quality of learning students gain from giving and receiving feedback, and the immediate access teachers can have to class progress. Educational institutions currently seek to gain immersed knowledge from these unstructured texts. Therefore, the general objective of this thesis has been to design a peer assessment model that helped teachers improve their teaching-learning processes through sentiment analysis methods.

The research-action design methodology was applied, firstly, the state of the art on peer assessment, text mining, and computational techniques was carried out. Subsequently, a model was designed that combines peer assessment with collaborative learning and the calibrated method in several phases: a) groups of students were formed to participate in carrying out the work collaboratively, to have similar groups, but have individual differences in the peer assessment process to benefit from collaboration among students; b) a rubric was designed for data collection, where the evaluators evaluated specific aspects of the work, providing a numerical score and textual feedback for each criterion; c) the evaluators evaluated the quality of the task assessment (inverse assessment) to obtain the evaluator's confidence rating; d) the groups corrected the work based on the feedback given by the evaluators in the first round (evaluation in two rounds); e) the task assessment score was calibrated based on the performance and confidence rating of the evaluator.

During the development of the model, two sets of data were obtained in Spanish, one for task assessment and the other for assessment of the quality of the assessment. The supervised machine learning approach was applied to obtain a sentiment score corresponding to specific textual feedback. Different text mining and natural language processing techniques were analyzed on the sentiment classification task, such as Bag of Words, combinations of (N-Grams+Term

Frequency-Inverse Document Frequency+Stop-Words), and Word2Vec/Glove pre-trained to form the different vocabularies. Algorithms were evaluated of classic machine learning (Naïve Bayes, Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, Decision Trees), modern machine learning (Vote Ensemble), and deep learning (Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM)). Two predictive models with better performance were obtained. A model with Bi-LSTM using Glove's representation, to predict the sentiment score of task assessment textual feedback; and a model with LSTM using Glove's representation, to predict the sentiment score of the assessment quality assessment textual feedback.

Thereafter, a calculation model was obtained that contributed to improving the reliability of the peer assessment process. The score for each task assessment criterion and quality assessment was generated with the fuzzy logic computational technique correlating numerical score and sentiment, determining that the (Smallest of Maximum, Middle of Maximum, and Largest of Maximum) defuzzification methods were the most appropriate for this study. The individual score of each evaluator was obtained with mean calculations of all the criteria. The task assessment collective score and the evaluator confidence rating were obtained with mean/median calculations of the set of individual scores, determining that the median has the best fit to generate a reliable collective score.

The validity of the proposed model was tested in 3 higher education scenarios: virtual asynchronous, virtual synchronous, and face-to-face. Using Pearson, the score received by the student from the group was correlated with the score provided by the teacher, obtaining strong similarity in 8% of the activities in virtual asynchronous (r=0.718-0.790), 25% of the activities in virtual synchronous (r=0.741 to 0.971) and 40% of the activities in face-to-face (r=0.780 to 0.951), determining that the model can be applied in all the education scenarios evaluated, and with greater effectiveness in face-to-face.

Finally, a calibration model was obtained that contributed to improving the reliability of the peer assessment process, since, by adjusting the individual score of each task based on the performance and confidence rating of the evaluator, it was achieved that the relationship between the score of the group and the score provided by the teacher would tend to increase in 46% of the activities in asynchronous virtual settings, 69% in synchronous virtual settings and 60% in face-to-face settings.

In addition, it was evaluated if there is an improvement in student performance in the second round by applying the model in the peer assessment process, using the student's t-test, it was determined that 100% of the activities evaluated obtained the average score in the second round greater than the first round with a significance value of less than 0.05, the increase in the second round of student performance in asynchronous virtual was 3%-12%, in synchronous virtual 7%-22%, and in face-to-face 15 %-3. 4%. In this sense, the peer assessment model based on sentiment analysis could be implemented as a pedagogical tool to support the teacher in enriching the teaching-learning process, since the students gave and received detailed feedback on the correct or incorrect of a specific work, and they were able to refute the feedback given; which in turn induced improved work and performance in the second round.

1. INTRODUCCIÓN

Este capítulo proporciona una visión general acerca de la investigación realizada en esta tesis doctoral. Se describe el planteamiento del problema a resolver, posteriormente se expone la motivación de este trabajo, así como sus objetivos. También se expone la metodología utilizada a lo largo de la investigación realizada. Por último, se detalla la estructura del documento de tesis.

1.1. Descripción del problema

En la educación superior, la calificación de actividades es un proceso arduo para el docente, ya que el número de estudiantes es una limitante [1]. Por lo general, el tipo de evaluación que se aplica es cuantitativa (sumativa). En esta situación el docente es quien califica al estudiante, según los criterios establecidos en la cátedra, juicio que pudiera considerarse muy genérico, ya que no se considera las competencias, habilidades para asignar una calificación más equilibrada [2][3].

Motivo por el cual, se busca incluir en el ámbito universitario tipos de evaluaciones con los que se logre que la evaluación deje de ser una actividad llevada a cabo al final del proceso, que no solo se evalúen las competencias cognitivas sino también las reflexivas, criticas, colaborativas, y procedimentales siendo el estudiante el protagonista de los hechos [4]–[6]. Donde el estudiante participe a través de la autoevaluación, o la evaluación entre iguales o la coevaluación entre estudiantes y docentes [5][7].

Tomando en consideración, que el docente puede aplicar distintos procedimientos evaluativos como estrategia integrada en el proceso de enseñanza-aprendizaje, se hace necesario sistemas de evaluación caracterizados por su fiabilidad y validez que permitan recolectar los resultados del aprendizaje tanto con evaluación cuantitativa y cualitativa (retroalimentación formativa) para que el estudiante tenga conocimiento de sus avances, niveles de logros, limitaciones o fallas, y carencias, propiciando la autorregulación de los aprendizajes [8]–[11], la cual se basa en la revisión y mejora de las actividades de formación o aprendizajes.

La evaluación entre pares basada en retroalimentación es poco aplicada, debido a que el docente debe leer e interpretar una ingente cantidad de información textual no estructurada recolectadas mediante rúbricas de evaluación, y establecer una puntuación de cada retroalimentación textual, además buscar métodos para relacionarla con la puntuación numérica;

y todo esto, con un escaso tiempo, que conlleva a que el estudiante obtenga los resultados de manera tardía.

Las instituciones de educación superior se enfrentan actualmente a varios retos frente a los sistemas de evaluación informatizados para conseguir llegar al conocimiento inmerso de textos no estructurado. Por lo tanto, esta investigación proyecta responder las siguientes preguntas:

- ¿Como aplicar evaluación entre pares cuantitativa, cualitativa, inversa, en dos rondas y calibrada en escenarios de educación superior?
- ¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?
- ¿Cómo correlacionar puntuación numérica (evaluación cuantitativa) con retroalimentación textual (evaluación cualitativa), y generar una puntuación equilibrada entre estas dos evaluaciones?
- ¿Cómo determinar un índice (rating) de confidencia de los pares evaluadores (evaluación inversa)?
- ¿Qué método de tendencia central (media/mediana) tiene el mejor ajuste para generar una puntuación del colectivo confiable en procesos de evaluación entre pares?
- ¿Cómo calibrar la puntuación de evaluación de tarea, que establezca fiabilidad en el proceso de evaluación entre pares?
- ¿Cuál es la incidencia de la modalidad de educación en procesos de evaluación entre pares?
- ¿De qué manera el modelo de evaluación entre pares basado en análisis de sentimiento podría contribuir en el proceso de enseñanza-aprendizaje?

1.2. Motivación de la tesis

Las instituciones de educación superior ante los avances tecnológicos y la pandemia de Covid-19, se han esforzado por encontrar medios para garantizar que los estudiantes puedan continuar sus estudios a pesar de la crisis y/o el distanciamiento social. El aprendizaje presencial o en línea puede adoptar muchas formas diferentes, incluidas aquellas que son más innovadoras y atractivas desde el punto de vista pedagógico [12]. Algunos estudiantes sienten que es necesario realizar evaluaciones periódicas para mantener el proceso de enseñanza-aprendizaje en el camino correcto. Los docentes pueden utilizar herramientas y técnicas innovadoras para el

mismo [13]. Además, las nuevas teorías pedagógicas sugieren que el estudiante sea protagonista, así como responsable de su propio aprendizaje; el protagonismo del estudiante, en su proceso de formación, ha evolucionado que incluye actividades como la evaluación entre pares en el aprendizaje permanente, por lo tanto, una solución potencialmente efectiva es la evaluación de tareas por parte de sus compañeros de aula [7].

La evaluación entre iguales o por pares es considerada como una forma específica de aprendizaje colaborativo en el que los aprendices realizan una valoración sobre el proceso o producto de aprendizaje de todos o de algún estudiante o grupo [14], y pueden realizarse como evaluaciones sumativas y/o formativas. El propósito de la evaluación sumativa es la calificación y evaluación del aprendizaje de los estudiantes. La evaluación formativa se enfoca en el desarrollo de los procesos de aprendizaje de los estudiantes. La retroalimentación se considera un componente principal de la evaluación formativa y uno de los factores que más influyen en el aprendizaje [3]. Para comprender la retroalimentación formativa se plantea: 1) el papel del procesamiento de información humana y las características individuales de los estudiantes para la eficiencia de la retroalimentación; 2) cómo brindar retroalimentación significativa a los estudiantes en los dominios de estudio donde el trabajo de los estudiantes es difícil de evaluar; y 3) cómo las fuentes de retroalimentación humana (estudiantes pares) pueden ser apoyadas por interfaces de usuario y análisis de la retroalimentación por la tecnología [15].

Varios escenarios exitosos han sido evidenciados en su aplicación, sin embargo, suelen producirse problemas si no se aplica un procedimiento fiable. En este sentido, algunos de los problemas y cuestionamientos destacados con los que concuerdan múltiples autores se presentan: carga de trabajo en revisar, y puntuar cada evaluación cualitativa (retroalimentación), resultados tardíos de la evaluación, aplicar métodos para relacionar la evaluación cualitativa con la cuantitativa y obtener una calificación equilibrada, revisar la calidad de las revisiones realizada por los pares, la efectividad de la retroalimentación, ratio de correlación entre pares, bonificación, penalización, entre otros [16]–[23].

Con el propósito de disminuir estos inconvenientes, se han utilizado técnicas computacionales que permitan generar una puntuación de la evaluación cualitativa, destacándose el análisis de sentimiento. Este abre un abanico de opciones en los procesos de minado sobre datos no estructurados, además el uso de este enfoque ha sido escasamente explorado en procesos de evaluación entre pares con enfoque cualitativo [24][25]. También se han utilizado técnicas para relacionar la puntuación de la evaluación cualitativa con la cuantitativa,

destacándose la lógica difusa que permite modelar la inexactitud y la incertidumbre inherentes a la evaluación [22], [23], [26]–[31].

Por tanto, el diseño de un modelo de evaluación entre pares con enfoque de análisis de sentimiento y lógica difusa utilizando modelos computacionales, se podría considerar una contribución a la comunidad científica para coadyuvar en los procesos de gestión académica y en la práctica evaluativa entre pares de trabajos colaborativos en cursos universitarios tradicionales de lengua española, ya que la mayoría de las investigaciones se han centrado en los MOOC [21], [23], [27]; en la evaluación de respuestas abiertas [20], [21]; en el análisis de sentimiento de retroalimentación en el idioma inglés [16]–[19]; y en la evaluación de imprecisiones de palabras y criterios en el idioma inglés [22], [23], [26], [27].

Esta investigación aspira apoyar en los siguientes objetivos académicos pedagógicos:

1) El estudiante logre: a) expresar de forma reflexiva y critica en la retroalimentación lo que está correcto o incorrecto del trabajo abierto y de las retroalimentaciones recibidas, y comprender los contenidos de la asignatura; b) mejorar sus trabajos y rendimiento en base a las retroalimentaciones recibidas de sus pares; c) adquirir habilidades de autorregulación y de participación de forma colaborativa; y d) mejorar la implicación en el aprendizaje.
2) El docente use el modelo: a) como estrategia de enseñanza-aprendizaje; b) para abstraer la información, con las retroalimentaciones que son de tipo positiva podrá darse cuenta en que temas el estudiante ha obtenido transferencia de aprendizaje y con las de tipo negativa podrá darse cuenta en que temas los estudiantes les falta conocimiento.
3) Las autoridades podrían adoptar el modelo para satisfacer las necesidades educativas actuales ante la tecnología y la pandemia, en donde el docente se aligere de revisión de un sinnúmero de actividades y el estudiante obtenga resultados inmediatos.

1.3. Hipótesis de partida y objetivos de la tesis

1.3.1. Hipótesis de partida

La hipótesis de partida establece que: las técnicas de procesamiento de lenguaje natural y de análisis de sentimiento pueden resultar apropiadas para apoyar el problema de evaluación entre pares; sin embargo en el desarrollo de esta investigación se pudieron develar un conjunto de variables que permitieron llegar a una hipótesis detallada la cual se soporta en la variable dependiente que es la “evaluación entre pares” y la variable independiente siendo la variable especifica de “análisis de sentimiento”, por la tanto la hipótesis de esta investigación se analizará como sigue: En la evaluación entre pares aplicando análisis de sentimiento existen diferencias estadísticamente significativas en el rendimiento estudiantil (ver Subsección 5.2.2.1).

1.3.2. Objetivos

General

Diseñar un modelo de evaluación entre pares, que coadyuve a los docentes a mejorar sus procesos de enseñanza-aprendizaje mediante métodos de análisis de sentimiento (ver Subsección 5.2.2).

Específicos

1. Analizar la situación actual de evaluación entre pares (ver Sección 2).
2. Diseñar los artefactos para validar los métodos teóricos (ver Sección 3).
3. Construir un modelo de evaluación entre pares basado en análisis de sentimiento (ver Sección 4).
4. Evaluar los resultados de la precisión del modelo (ver Sección 5).

1.4. Metodología de investigación

El proceso de investigación de este trabajo de tesis doctoral hace uso de la metodología de diseño investigación-acción (Action Design Research) [32]–[34]. Esta metodología persigue dos finalidades: resolver problemas prácticos y la creación de conocimiento a través de esos mismos problemas y en colaboración con los participantes.

Se trata de una investigación aplicada, para lo cual se cumplirá las siguientes fases: 1) formulación del problema; 2) planificación, intervención y evaluación; 3) reflexión y aprendizaje; 4) formalización de aprendizaje (Figura **1**).

Metodológicamente, dicho proceso se compone de un bucle en el que se repiten: la planificación, intervención y evaluación, donde la evaluación de una acción se convierte, a través de la reflexión, en la base para la planificación de la siguiente y así sucesivamente.

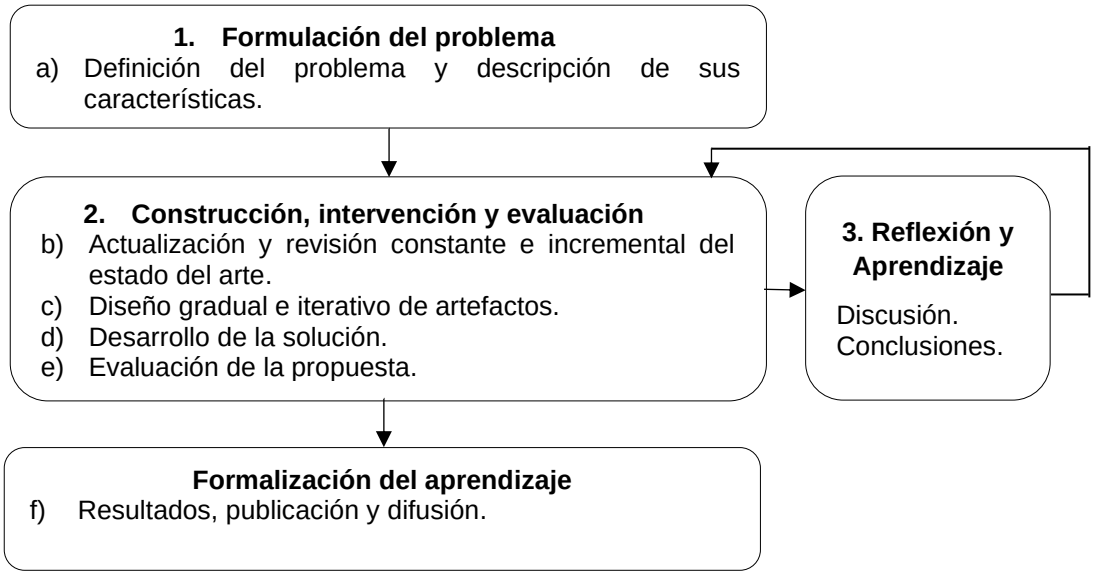


Figura 1. Fases y etapas de la metodología de investigación adaptada de [33]

Siguiendo las fases de esta metodología de investigación, se aplicaron las siguientes etapas, que serán brevemente explicadas a continuación:

a) Definición del problema y descripción de sus características

Esta actividad consiste en el estudio del problema realizando un análisis del contexto de interacción para poder realizar una definición del problema, exponiendo de una forma adecuada sus características. Además, se propone una hipótesis para solucionar parcialmente o total dicha problemática, así como plantear los objetivos para lograrlo.

La primera actividad consiste en el estudio de la situación actual de la evaluación entre pares. Se procederá a extraer aquellos elementos que participan en el proceso de evaluación entre pares basada en retroalimentación, determinando sus características, procedimientos, actividades y escenarios en los que interactúen. La información y funcionalidad general de estos

elementos serán utilizados para proponer un modelo de evaluación entre pares basado en análisis de sentimiento.

b) Actualización y revisión constante e incremental del estado del arte

Se analiza el estado del arte de la evaluación entre pares, minería de texto, técnicas computacionales y definiciones relacionadas con la presente investigación, para obtener un marco teórico sustentable que permita enriquecer el conocimiento y mejorar el proceso de desarrollo.

Se realizará una revisión sistemática de la literatura sobre la evaluación entre pares basada en retroalimentación y análisis de sentimiento de texto educativo para tratar de incorporar aquellos avances que puedan resultar de interés.

c) Diseño gradual e iterativo de artefactos

Partiendo de la información obtenida de las actividades anteriores, se diseña artefactos que integre los elementos necesarios para proponer una solución útil e innovadora a la problemática definida, siguiendo el objetivo planteado. Esta actividad consiste en dos tareas principales:

1. Diseño de artefactos de recolección de datos

   Esta tarea consiste en diseñar rúbricas o prototipó para recolectar datos, así como diseñar encuestas para recabar información sobre la percepción de los estudiantes.

2. Diseño de artefactos en la metodología de la propuesta de solución

   Esta tarea consiste en el diseño de procedimientos para desarrollar un modelo de evaluación entre pares, que pueda integrar las definiciones de evaluación entre pares, así como técnicas computacionales que participe en las interacciones.

d) Desarrollo de la solución

Esta tarea se extenderá en la mayor parte del trabajo de investigación que consiste en construir un modelo de evaluación entre pares con técnicas computacionales en tres iteraciones:

1. Prototipo de recolección de datos.
2. Inserción de control de algoritmos de captura de retroalimentación.
3. Modelo de evaluación implementado.

e) Evaluación de la propuesta

Las soluciones teóricas y computacionales, obtenidas a lo largo del trabajo de investigación, serán testeadas con diferentes procesos de evaluación entre pares (cualitativa, cuantitativa, inversa, en dos rondas y calibrada) a lo largo del marco temporal del mismo, con el objetivo de su validación y necesario refinamiento.

La finalidad de esta actividad es realizar una validación completa de los alcances, comprobando la aplicabilidad y utilidad del modelo en escenarios de educación superior (virtual asincrónico/sincrónico y presencial).

f) Resultados, publicación y difusión

A lo largo del periodo de investigación se procederá a la difusión y publicación de los resultados obtenidos en revistas y conferencias de carácter nacional e internacional de prestigio reconocido en la línea de investigación en la que se encasilla el trabajo de investigación.

En la Tabla **1**, se detalla las acciones aplicadas en cada una de las fases y etapas de la metodología de investigación.

Tabla 1. Acciones aplicadas en las fases de la metodología de investigación

Iteración	Formulación del problema	Construcción, intervención y evaluación				Reflexión y aprendizaje	Formalización del aprendizaje	
	Definición	**Estado arte**	**Diseño de artefacto**	**Desarrollo de la solución**	**Evaluación**	**Discusión y conclusiones**	**Producto**	**Publicación y Difusión**
1	¿Como aplicar evaluación entre pares cuantitativa cualitativa, inversa, en dos rondas y calibrada en escenarios de educación superior?	Se investigó sobre evaluación entre pares y rúbricas de evaluación. Se realizó una revisión sistemática de la literatura.	Se diseñó el procedimiento básico de evaluación entre pares basado en análisis de sentimiento. Se diseño un modelo de evaluación entre pares. Se diseñó rúbrica de tipo analítica y holística. Se elaboró un cuestionario de percepción de estudiantes.	Se elaboró rúbricas en Excel y Google Form para recolectar datos.	Se realizó un análisis comparativo del tipo de rúbrica que mejor se adapta para adicionar retroalimentación textual por cada criterio.	Se deduce que la rúbrica de tipo holística facilita al estudiante evaluar tareas en procesos de evaluación entre pares con puntuación numérica y retroalimentación textual.	Prototipo de recolección de datos.	Sentiment Analysis Techniques for Peer Feedback: A Review". https://ieeexplore.ieee.org/document/10122085
				Se desarrolló un prototipo de evaluación entre pares: cuantitativa, cualitativa, inversa y en dos rondas.	Se indagó la percepción de los estudiantes sobre la usabilidad del prototipo en el proceso de evaluación entre pares.	Se colige que fueron pocas las dificultades encontradas en cada una de las interfaces del prototipo (ver Apéndice A)		
2	¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?	Se investigó sobre minería de texto, NLP, análisis de sentimiento, algoritmos de clasificación.	Se diseñó el esquema metodológico de análisis sentimiento de la retroalimentación textual.	Se realizó análisis de sentimiento de la retroalimentación textual.	Se evalúo el rendimiento de algoritmos de aprendizaje automático con Bag of Words, TF-IDF.	El algoritmo de clasificación con mejor desempeño fue SVM (F-Measure de 0.870) por sus características de entrenamiento con textos cortos.	Inserción de control de algoritmos de captura de retroalimentación.	Sentiment Analysis of Peer Feedback in Higher Education. Aceptado para publicación en AIP Conference Proceedings (ver Apéndice B) Accuracy' Measures of Sentiment Analysis Algorithms for Spanish Corpus generated in Peer Assessment.

Iteración	Formulación del problema	Construcción, intervención y evaluación				Reflexión y aprendizaje	Formalización del aprendizaje	
	Definición	**Estado arte**	**Diseño de artefacto**	**Desarrollo de la solución**	**Evaluación**	**Discusión y conclusiones**	**Producto**	**Publicación y Difusión**
	¿Cómo correlacionar puntuación numérica (evaluación cuantitativa) con retroalimentación textual (evaluación cualitativa), y generar una puntuación equilibrada entre estas dos evaluaciones?	Se investigó sobre lógica difusa.	Se diseñó el esquema metodológico de detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación.	Se refinó el análisis de sentimiento y se realizó correlación entre puntuación de sentimiento y numérica en Python.	Se evalúo: Rendimiento de algoritmos de clasificación con N-Grams + TF-IDF. + Stop-Words, y métodos de desfuzzificación	Los resultados mostraron que el modelo SVM (F-Measure de 0.879) obtuvo un buen desempeño cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF. Los métodos MOM, SOM, y LOM fueron los más apropiados para generar la puntuación de la correlación entre puntuación numérica y de sentimiento.		https://dl.acm.org/doi/10.1145/3410352.3410838 Peer assessment using soft computing techniques. https://doi.org/10.1007/s12528-021-09296-w Peer Feedback Sentiment Analysis Prototype Autores: Aceptado para publicación en RISTI (ver Apéndice C)
	¿Cómo determinar un índice (rating) de confidencia de los pares evaluadores (evaluación inversa)?	Se investigó sobre evaluación inversa y en dos rondas.	Se diseñó un procedimiento de recopilación de datos de evaluación de tarea, evaluación inversa y en dos rondas.	Se refinó el análisis de sentimiento y detección de imprecisiones entre puntuación de sentimiento y numérica, y se obtuvo el rating de confianza de los evaluadores en Python. Se implementó en Python el modelo predictivo que genera la puntuación de sentimiento en CSV.	Se evalúo el rendimiento de algoritmos de clasificación de sentimiento con Word2Vec y Glove.	Los resultados con los datos de evaluación de tarea mostraron que el modelo Bi-LSTM+Glove obtuvo mejor rendimiento (F-Measure de 0.963). Los resultados con los datos de evaluación de la calidad de la evaluación mostraron que el modelo LSTM+Glove obtuvo mejor rendimiento (F-Measure de 0.980).		

Iteración	Formulación del problema	Construcción, intervención y evaluación				Reflexión y aprendizaje	Formalización del aprendizaje	
	Definición	**Estado arte**	**Diseño de artefacto**	**Desarrollo de la solución**	**Evaluación**	**Discusión y conclusiones**	**Producto**	**Publicación y Difusión**
	¿Qué método de tendencia central (media/mediana) tiene el mejor ajuste para generar una puntuación del colectivo confiable en procesos de evaluación entre pares?	Se investigo sobre medidas de cálculos.	Se diseñó el procedimiento de cálculo de puntuación de evaluación de tarea y rating de confianza del colectivo.	Cálculos de puntuación del colectivo de evaluación de tarea y rating de confianza del evaluador.	Se evaluó sí la puntuación de evaluación debe ser calculada con media o mediana.	Los resultados mostraron que la media cambie su magnitud, en contraste la mediana no se ve afectada por valores grandes o pequeños.		
	¿Cómo calibrar la puntuación de evaluación de tarea, que establezca fiabilidad en el proceso de evaluación entre pares?	Se investigó sobre calibración de evaluaciones.	Se diseñó el procedimiento de calibración de puntuación de evaluación de tarea.	Se desarrolló calibración de puntuación de evaluación de tarea en Python.	Se evaluó: Si la calibración de evaluación de tarea debe estar en función del rating de confianza y/o también del rendimiento del evaluador.	Se determinó como factores tanto la puntuación obtenida y el rating de confianza del evaluador para calibrar la puntuación de evaluación de tarea, y así establecer una mejor fiabilidad en el proceso de evaluación entre pares.		
3	¿Cuál es la incidencia de la modalidad de educación en procesos de evaluación entre pares? ¿De qué manera el modelo de evaluación entre pares basado en análisis de sentimiento podría contribuir en el proceso de enseñanza-aprendizaje?	Se investigó sobre la validez de la evaluación entre pares.	Se elaboró un cuestionario de percepción de estudiantes.	Se implementó el modelo de análisis de sentimiento y de lógica difusa al prototipo de evaluación entre pares. Se evaluó el modelo mediante estadística descriptiva.	Se evaluó: Similaridad entre la puntuación de los evaluadores con la del docente. Si existe mejora del rendimiento estudiantil en la segunda ronda del proceso de evaluación entre pares. Percepción del estudiante sobre el proceso de evaluación entre pares.	Se determinó que: Existe diferencia significativa en las 27 actividades realizadas, entre la puntuación media de la segunda y primera ronda en procesos de evaluación entre pares. Por tanto, la mayoría de los estudiantes si mejoran el rendimiento en la segunda ronda.	Modelo de evaluación implementado.	

1.5. Estructura del documento

Tras el presente capítulo de introducción, se definen los fundamentos teóricos y se presenta la revisión sistemática de la literatura realizada. En los capítulos 3, 4 y 5 se describen los componentes principales del modelo propuesto. Finalmente, el capítulo 6 recoge las conclusiones y líneas futuras. El contenido de los diferentes capítulos se describe de un modo general:

- **Capítulo 1. Introducción**

  En este capítulo de define el problema, exponiendo sus principales características. Se exponen las justificaciones que motivan la realización de este trabajo y se establecen los objetivos. Por último, se describe la metodología usada durante el proceso de investigación y se expone la estructura del documento.
- **Capítulo 2. Estado del arte**

  Este capítulo trata de recoger todas aquellas aportaciones realizadas por diferentes investigadores relacionadas con la temática recogida en esta tesis. Además, se realizó un análisis exhaustivo de los trabajos de investigación pertenecientes a las categorías de:
  - Evaluación entre pares basado en retroalimentación.
  - Análisis de sentimiento de texto educativo.
- **Capítulo 3. Metodología de la propuesta de solución**

  En este capítulo se diseña el modelo de evaluación entre pares con técnicas computacionales.
- **Capítulo 4. Experimentación**

  En este capítulo se tratan todos aquellos aspectos relacionados con el desarrollo del modelo mediante iteraciones, aplicando técnicas computacionales.
- **Capítulo 5. Modelo de evaluación implementado**

  En este capítulo se tratan todos aquellos aspectos relacionados a la implementación y evaluación del modelo en escenarios de educación superior.
- **Capítulo 6. Discusión de resultados, conclusiones, limitaciones y líneas futuras**

  En este capítulo se exponen la discusión de las preguntas de investigación, las conclusiones obtenidas en esta tesis, así como los objetivos alcanzados. Además, en este capítulo se exteriorizan las líneas futuras que permitirán la continuación de esta investigación.

2. ESTADO DEL ARTE

En este capítulo se introducen los conceptos básicos de evaluación entre pares y las disciplinas que tiene una relación directa con la minería de texto y que serán utilizadas en esta investigación, como son, el análisis de sentimiento, procesamiento del lenguaje natural, la extracción de característica y el aprendizaje automático.

2.1. Fundamentos teóricos

2.1.1. Evaluación entre pares (Peer Assessment)

Es una estrategia de aprendizaje activo que se basa teóricamente en el marco constructivista [35]. La pedagogía constructivista busca involucrar a los estudiantes en el proceso de aprendizaje para que construyan su propia comprensión [36]. La naturaleza de la evaluación entre pares es colaborativa en la que los estudiantes aprenden unos con otros [37]. Los estudiantes juzgan el desempeño de sus compañeros de manera cuantitativa y/o cualitativa [38]. Estimula a los estudiantes a reflexionar, discutir y colaborar en su proceso de aprendizaje [39]. Cuando los estudiantes evalúan a sus compañeros aprenden a niveles más altos como analizar y evaluar de acuerdo con la taxonomía de Bloom [40]. Reduce la carga de trabajo del docente [41][42].

La evaluación entre pares se ha realizado no solo en entornos de educación tradicional [43], sino también en línea [44]–[46] y en MOOC [47]–[50]. Puede incluir evaluación formativa que proporciona retroalimentación sobre el aprendizaje de los estudiantes, y evaluación sumativa que resume los logros de los estudiantes principalmente a través de la asignación de calificaciones [51]. Se recomienda múltiples revisiones para que los estudiantes estén expuestos a una variedad de trabajos de diferente calidad, y ellos mismos reciban más de una revisión por pares [52]. Un resultado del colectivo es más confiable que un resultado individual [53].

En todas las etapas de la educación se ha aplicado la evaluación entre pares, con mayor énfasis en la educación superior para ayudar a los estudiantes a evaluar su propio trabajo y el de los demás, y a adoptar una actitud más autodirigida hacia su aprendizaje en preparación para su desarrollo profesional continuo y aprendizaje permanente [54]. La eficacia depende de una variedad de factores, incluido si la evaluación entre pares se completa de forma anónima [55]–[57], el tipo de evaluación, si es basada en calificación o diálogo entre pares tanto en evaluaciones formativa como sumativa [58], uso de rúbricas [59], forma de evaluación en papel versus en línea

[45], capacitación de los revisores pares y frecuencia de la evaluación [60]. El efecto sobre el evaluador de pares y el evaluado también es diferente, y los estudiantes generalmente obtienen más a través del proceso de dar retroalimentación en lugar de recibirla [61].

2.1.1.1. Validez de la evaluación entre pares

Un desafío importante en la implementación del proceso de evaluación por pares representa la validez de las evaluaciones ofrecidas por los estudiantes [62]. La noción de validez se refiere al nivel de concordancia entre las calificaciones asignadas por los estudiantes y las otorgadas por el docente [63].

Los factores sociales pueden influir en la validez de las evaluaciones entre pares como la amistad, la aversión, la popularidad, la evitación de conflictos entre otras [64]; y aparecen particularmente cuando las actividades se llevan a cabo en un escenario educativo cara a cara, no son un factor crítico en los escenarios en línea, debido a la distancia geográfica, el anonimato e incluso por la asincronía [65]. Sin embargo, factores como la ansiedad están presentes en cualquier escenario educativo tanto para el evaluador como para el evaluado [66]. Los ambientes de anonimato y distancia de los MOOC disminuyen la subjetividad de evaluación provocada por estos factores sociales; pero factores sociales, como la inevitable simpatía hacia los compañeros, los factores económicos, el género, entre otros no se puede evitar ni controlar [67].

2.1.1.2. Ventajas y desventajas de la evaluación entre pares

La evaluación entre tiene varios beneficios: 1) comparativo, en el sentido de que los trabajos están expuestos a sus compañeros y, por lo tanto, los estudiantes están conscientes de cualquier disparidad en su propio trabajo, lo que fomenta el aprendizaje en sus propios envíos [42][68]; 2) analítico, en el sentido de que el acto de evaluación entre pares requiere que los estudiantes desarrollen la capacidad de identificar debilidades y desarrollar posibles mejoras, perfeccionando así sus habilidades para usarlas en el futuro [42]; 3) descriptivo, en el que los estudiantes deben desarrollar las habilidades para comunicar debilidades y fortalezas, debiendo así explicar sus argumentos y beneficiándose de lo que se denomina el efecto explicación [42], [69], [70].

Una de las dificultades que se presenta es involucrar a los estudiantes en la evaluación, por la creencia que tienen los estudiantes que no obtendrán ningún beneficio del proceso de evaluación entre pares porque es una estrategia de reducir la carga de trabajo del docente [71].

Los estudios han sugerido varios incentivos para la participación, como recompensa a la calificación [72], penalización de calificación [73], y calificación del evaluador en función de la calidad de su revisión [49]. Además, el aprendizaje entre pares puede estar sesgado debido a la variación en el conocimiento previo de los estudiantes, las características de los compañeros, las preferencias personales y las relaciones con los compañeros [56].

2.1.2. Retroalimentación entre pares (Peer Feedbak)

La retroalimentación es, sin duda, el mecanismo central en la evaluación entre pares para convertirse en formativa [74]. Cuando se implementa correctamente, la evaluación entre pares involucra a los estudiantes en ambos roles de retroalimentación: como evaluadores, al contribuir con ideas y comentarios a las tareas evaluadas, y como evaluados, al recibir las observaciones de los pares con comentarios constructivos para mejorar su propio trabajo [74]. Este tipo de evaluación suele coexistir con las sumativas, aunque puede presentarse de forma separadas. No obstante, se recomienda que la evaluación formativa vaya acompañada de la sumativa [75]. En este sentido, se destaca la importancia de que los estudiantes reciban retroalimentación personalizada en lugar de solo recibir puntuaciones [74].

La estrategia de retroalimentación se utiliza en el aprendizaje crítico [76], así como en el aprendizaje social [55]. Esta estrategia de aprendizaje se basa en la teoría constructivista que postula que la adquisición de conocimientos se puede lograr como resultado de la interacción, el intercambio y la reflexión de los estudiantes [35]; en tal contexto de aprendizaje, los estudiantes tienen la oportunidad de observar el desempeño de otros, conocer sus fortalezas y evitar deficiencias promoviendo la autorreflexión [77].

Los estudiantes cuando intentan dar puntuaciones o retroalimentación sobre el trabajo de sus compañeros generalmente revisan su propio trabajo y hacen comparaciones de su propio trabajo con el de otros al consultar las rúbricas proporcionadas por el docente. Esto les permite reflexionar sobre su propio trabajo y aprender los criterios para juzgar la calidad del trabajo [78]. Es decir, se produce una interacción entre los estudiantes y sus saberes, donde los estudiantes se involucran en reconceptualizar, integrar y recrear conocimientos previos [79]. Como resultado, los estudiantes pueden mejorar sus logros de aprendizaje, así como la motivación para el aprendizaje y las habilidades de pensamiento crítico [80]. Por un lado, podrían potenciar su capacidad reflexiva, y por otro, podrían encontrar sus propias carencias [79]. Como

resultado, los estudiantes pueden potenciar sus habilidades al participar en el proceso de evaluación entre pares [81].

Los estudiantes cuando dan y/o reciben retroalimentación, para tener un impacto en el aprendizaje, deben comparar esa información y generar nuevos conocimientos (retroalimentación interna) a partir de esa comparación [82][83]. Durante la revisión, los estudiantes comparan su propio trabajo con ejemplos concretos de trabajos similares. Por el contrario, la recepción de retroalimentación los involucra en la comparación de su propio trabajo con una descripción textual de lo que es bueno o deficiente en su trabajo o sobre cómo se podría mejorar ese trabajo; podría decirse que esta diferencia en la información de comparación ayuda a explicar el hallazgo de que los estudiantes aprenden cosas diferentes al revisar y al recibir retroalimentación [84].

2.1.3. Evaluación inversa

La evaluación inversa es otro factor que facilita la evaluación entre pares efectiva con el diálogo o la respuesta [85]. Este tipo de interacción y co-construcciones de comprensiones se alinean bien con las teorías de aprendizaje constructivista social [86]. Una característica innovadora del enfoque es una respuesta escrita en la que los estudiantes aprueban o refutan las retroalimentaciones de los evaluadores y explican por qué se aceptan o rechazan [85]. Este aspecto alienta a los estudiantes a procesar la información de la retroalimentación, proporcionar justificaciones para las retroalimentaciones que se han utilizado o no, y promulgar los resultados de la retroalimentación [85][87].

2.1.4. Evaluación en dos rondas

Los evaluadores proporcionan retroalimentación para que los estudiantes corrijan sus errores de acuerdo a los comentarios constructivos. En una investigación, demostraron que una vez que los estudiantes mejoraron el trabajo, las retroalimentaciones de tipo elogio aumentaron en la segunda ronda y las de tipo irrelevante disminuyeron en la segunda y tercera ronda [88].

2.1.5. Calibración de puntuaciones

Se han realizado varios estudios para para mejorar el proceso de calificación entre pares; han aplicado el método estadístico post hoc bayesiana, para corregir las distribuciones de calificaciones de toda la clase después de que se termine la calificación [89]; y han utilizado el

método de muestra donde el docente coloca calificaciones correctas, en base a ello se evalúa la confiabilidad de cada evaluador y se pondera las calificaciones que asignaron [47].

2.1.6. Trabajo colaborativo

El trabajo en equipo permite desarrollar el pensamiento crítico, la colaboración y la argumentación [90]. La formación de grupos es uno de los procesos clave en el aprendizaje colaborativo porque contar con miembros adecuados en los grupos de aprendizaje favorece las buenas interacciones colaborativas entre los miembros y es fundamental para garantizar un desempeño de aprendizaje satisfactorio [91] . Los métodos más utilizados para la formación de grupos incluyen la agrupación aleatoria, la selección por parte del docente y la selección por parte de los estudiantes [91]. En [92] se afirma que los grupos homogéneos formados por estudiantes con habilidades, experiencias e intereses similares tienden a ser mejores para lograr objetivos específicos. Sin embargo, los grupos heterogéneos se forman con el objetivo de crear equipos equilibrados de personas que tienen una variedad de habilidades, destrezas, géneros y orígenes étnicos [93].

El aprendizaje colaborativo tiene un gran potencial en el campo de la educación superior porque promueve la construcción conjunta de conocimientos, así como el desarrollo de habilidades relacionadas con la interacción que redunda en procesos de aprendizaje más esenciales [94]. Los objetivos de aprendizaje colaborativo generalmente se reconocen como la mejora de las habilidades interpersonales, el conocimiento del contenido y la capacidad de pensamiento de alto nivel [95].

2.1.7. Trabajo abierto

Los trabajos abiertos generalmente conllevan una serie de beneficios como la posibilidad de desarrollar ideas originales y un proceso de aprendizaje más productivo para el estudiante, que las actividades de respuesta cerrada [96]. Calificar trabajos abiertos es notablemente más desafiante ya que estas tareas generalmente no exhiben una única solución correcta, dada esta variabilidad, exige una cantidad significativa de tiempo y esfuerzo [97]. Ante esta situación, una alternativa es recurrir a la evaluación entre pares ya que permite retroalimentar al estudiante mediante la verificación de soluciones alternativas al mismo problema por parte de otros estudiantes [42].

2.1.8. Rúbrica

El uso de una rúbrica permite abordar la tarea de revisión en el proceso de evaluación entre pares, dado que no se requiere que los estudiantes conozcan la solución correcta del trabajo abierto, sino que evalúen un conjunto de puntos determinados en una colección de trabajos [96]. La incorporación de escalas numéricas en la revisión de trabajos abiertos produce más comentarios explicativos [98]. Se ha revelado que cuando los estudiantes otorgan puntuaciones más bajas, intentan proporcionar más explicaciones escritas [99]. Con el fin de reducir el efecto de posibles evaluaciones incorrectas, el mismo trabajo debe ser revisado por diferentes estudiantes para finalmente producir una puntuación agregada de todos ellos [96][100].

La variación en el proceso de respuesta a las preguntas de la rúbrica es un desafío, porque parte de la variación proviene de las diferencias en las habilidades y concepciones de los estudiantes [96][99]. Para explicar las respuestas a las preguntas de la rúbrica, los estudiantes deben revisar cuidadosamente los contenidos para ver si cumplen con los criterios específicos o no, y luego, mediante un profundo proceso cognitivo, articular el porqué de sus respuestas [99]. Es posible que esto no elimine las variaciones entre los estudiantes, pero ayudaría a los estudiantes a comprender mejor las respuestas de sus compañeros a las preguntas de la rúbrica [99].

2.1.9. Minería de datos educativa (Educational Data Mining)

Las técnicas de minería de datos educativa se han empleado con éxito para mejorar el aprendizaje de los estudiantes, y ayudar a los docentes a mejorar el proceso de aprendizaje [101]. Sin embargo no explora a fondo todos los recursos educativos disponibles, como preguntas abiertas y ejercicios de redacción que podrían usarse para evaluar al estudiante [102]. Para abordar este problema, se podrían adoptar técnicas de minería de textos para extraer información de alta calidad de un texto no estructurado [100][101].

2.1.10. Minería de texto (Text Mining)

Minería de textos también conocida como análisis de textos inteligente (Intelligent Text Analysis), minería de datos de textos (Text Data Mining) o descubrimiento de conocimiento en textos (Knowledge-Discovery in Text), generalmente se refiere al proceso de extraer información y conocimiento interesante de un texto no estructurado [104]–[107]. La nueva generación de

plataformas en línea podría beneficiarse de diferentes técnicas de minería de textos, como el procesamiento del lenguaje natural, la clasificación y agrupación de textos, la recuperación de información y el resumen de textos [104].

En el ámbito educativo, la minería de texto se ha centrado principalmente en analizar los contenidos de los recursos educativos [108], en especial sobre minería de textos educativo (Educational Text Mining, ETM) [109]. La aplicación de técnicas de ETM ha logrado resultados significativos, especialmente en tareas y ensayos en línea, análisis de foros y chats, comentarios, producción de textos académicos, redes sociales y blogs [104], [110], [111].

2.1.11. Procesamiento de lenguaje natural (Natural Language Processing)

El procesamiento de lenguaje natural es el campo que se basa en técnicas computacionales para analizar y representar textos que ocurren naturalmente en uno o más niveles del análisis lingüístico: fonológico, morfológico, léxico, sintáctico, semántico, del discurso y pragmático, con el propósito de obtener el procesamiento de lenguaje similar al humano para una serie de tareas o aplicaciones [112], [113].

2.1.12. Análisis de sentimiento (Sentiment Analysis)

El análisis de sentimiento también conocido como minería de opinión (Opinion Mining), es otra técnica de minería de texto muy utilizada en educación [104]. Recurre a NLP para identificar y extraer opiniones y sentimientos desde diversas fuentes de información [114]. Tiene como objetivo determinar la polaridad general de un documento, una oración o un aspecto del mismo, tratando de detectar la actitud del creador en base a las posibles emociones, juicios o evaluaciones contenidas en el documento, las etiquetas más extendidas para clasificar la polaridad son: positiva, negativa o neutra [107].

2.1.12.1. Niveles de análisis de sentimiento

El análisis de sentimiento se puede llevar a cabo a tres niveles distintos en base a la granularidad, profundidad y detalle requeridos [115], [116]. Estos niveles son:

- **Nivel de documento**

  En este nivel se analiza el sentimiento global de un documento como un todo indivisible, clasificándolo como positivo, negativo o neutro o usando otro sistema de calificación; en estos

casos, se asume que dicho documento expresa una valoración sobre una única entidad, por lo que no es aplicable en aquellos que hablen sobre varias entidades simultáneamente [117].

- **Nivel de oración**

  El análisis de sentimiento a este nivel determina si cada sentencia expresa una opinión positiva, negativa, o neutra [117]. Este nivel de análisis se relaciona con la clasificación de la subjetividad, el cual consiste en distinguir sentencias objetivas que proporcionan información fáctica de sentencias subjetivas que expresan puntos de vista y opiniones [118]. Sin embargo, es importante aclarar que la subjetividad no es equivalente a sentimientos, ya que muchas sentencias objetivas pueden implicar opiniones. Una oración puede expresar opiniones tanto positivas como negativas, también puede contener cláusulas tanto subjetivas como objetivas [119]. Con el objetivo de analizar las opiniones compuestas los investigadores determinan las cláusulas que componen una oración para determinar la polaridad [117], [120], [121].

- **Nivel de aspecto y entidad**

  Este es el nivel de análisis con mayor detalle posible, en donde una entidad está formada por distintos elementos o aspectos y sobre cada uno de ellos se expresa una opinión cuya polaridad puede ser distinta en cada caso [122]. Se basa en la idea de que una opinión consiste en un sentimiento (positivo o negativo o neutro) y un objetivo (entidad) [123]. El nivel de aspectos también ha sido formalizado como el proceso de extracción de las características del objeto comentado, la determinación de la polaridad de las características del objeto, y luego agrupar las características similares y producir un informe en forma de resumen [24], [124]–[126].

2.1.12.2. Tareas de análisis de sentimiento

En el análisis de sentimiento se puede realizar varias tareas:

- **Clasificación de la subjetividad**

  Identifica fragmentos de texto que poseen un significado o una carga subjetiva, expresada por parte de la persona que ha escrito el texto, ya sea una opinión, la expresión de un sentimiento, entre otras [127], [128].

- **Clasificación de la polaridad**

  Clasifica fragmentos de texto, que pueden ser desde documentos hasta sintagmas, en positivo o negativo o neutro dependiendo de su significado emocional [128]–[130].

- **Clasificación de la intensidad**
  Clasifica los textos de entrada de acuerdo a la intensidad emocional expresada, como por ejemplo: fuertemente positivo, positivo, neutral, negativo, fuertemente negativo [131].
- **Resumen de opinión**
  Permite extraer las características principales que son compartidas por uno o más documentos y el sentimiento acerca de estas características [132].
- **Recuperación de opinión**
  Permite extraer documentos que expresan cierta opinión sobre la consulta realizada [132].

2.1.12.3. Diferentes tipos de opiniones

Las opciones se pueden clasificar en función de cómo se expresan en el texto [132]. A continuación, se detallan:

- **Opinión regular**
  Tiene dos subtipos principales: (a) directa, una opinión expresada directamente sobre una entidad o un aspecto de la entidad y (b) indirecta, una opinión que se expresa indirectamente sobre una entidad o aspecto de una entidad en función de sus efectos en algunas otras entidades [132].
- **Opinión comparativa**
  Una opinión comparativa expresa una relación de similitudes o diferencias entre dos o más entidades y/o una preferencia del titular de la opinión basada en algunos aspectos compartidos de las entidades [132].
- **Opinión explícita**
  Una opinión explícita es una declaración subjetiva que da una opinión regular o comparativa [132].
- **Opinión implícita**
  Una opinión implícita es una declaración objetiva que implica una opinión regular o comparativa [132], [133].

2.1.12.4. Proceso de análisis de sentimiento

El proceso de análisis de sentimiento se divide en cuatro fases principales: adquisición de datos, preprocesamiento, extracción de características y clasificación de sentimiento.

2.1.12.4.1. Adquisición de datos

Durante la adquisición se debe considerar extraer datos con técnicas NLP [112] o crear el propio conjunto de datos, teniendo en cuenta ciertas características como, por ejemplo: el idioma, sentimientos, temática, entre otras. Estas características del conjunto de datos, dependerá del análisis que se vaya a realizar [134].

2.1.12.4.2. Preprocesamiento

En esta fase se aplican varias técnicas de NLP que reducen el ruido de los comentarios, también permiten la reducción de dimensiones y una selección correcta de los datos que posteriormente serán usados en la fase de extracción de características [134]–[138]. A continuación, se detallan:

- **Normalización**
  Consiste en unificar términos que representan la misma información y pueden ser escritos en distintos formatos [134]. Las técnicas más utilizadas:
  - **Transformar de mayúscula a minúscula (Lowercasing)**
    Al transformar todas las palabras de mayúsculas a minúsculas, muchas palabras se fusionarán y la dimensionalidad se reducirá, ya que una palabra será considerada como una sola entrada, independientemente si es mayúscula [139].
  - **Eliminación de números**
    Es muy frecuente la eliminación de caracteres numéricos en los comentarios, sin embargo, algunos investigadores han mencionado que los números puede mejorar la eficiencia de la clasificación de texto [134].
  - **Eliminación de palabras vacías (Stop-Words)**
    Las palabras vacías se eliminan ya que no aportan en gran medida al análisis de sentimiento y puede causar ruido al conservarlas [134][139]. En español, estas palabras son las preposiciones, los pronombres, las conjunciones y las distintas formas del verbo haber, entre otras [140].
  - **Eliminación de signos de puntuación**
    En muchos trabajos, es muy frecuente la eliminación de signos de puntuación en el procesamiento de texto. Sin embargo, se debe considerar que la presencia de signos de puntuación denota algún sentimiento [134].

- **Lematización (Stemming)**
  Es un proceso de normalización morfológica que transforma cada palabra en su lema mediante el uso de diccionarios y de un proceso de análisis morfológico [141]. Permitiendo de esta manera que las palabras se fusionen y la dimensionalidad se reduzca [142]. A modo de ejemplo, la lematización convertiría la palabra “mesas” a su lema “mesa”.
- **Radicalización (Stemmer)**
  Permite eliminar las terminaciones de las palabras y detectar la forma raíz de las mismas, al hacerlo muchas palabras se fusionan y la dimensionalidad se reduce [134]. Muchas veces diferentes tokens pueden hacer referencia al mismo concepto ya que este puede ser representado por variantes morfológicas de una misma familia de palabras [140]. Por ejemplo, podemos representar \escribo, \escribíamos y \escribimos, ya que tienen un significado similar y derivan del mismo verbo, en su raíz como \escrib".

- **Tokenización**
  Una vez completado el proceso de normalización, los textos se dividen en unidades más pequeñas llamadas tokens y que normalmente se corresponden con las palabras de cada texto; este proceso puede ser tan sencillo como separar los términos de las frases por los espacios en blanco y los caracteres de puntuación o bien considerar además que la agrupación de determinados símbolos puede contener algún tipo de información que sea útil al proceso de clasificación [143].

2.1.12.4.3. Extracción y selección de características

En esta etapa de análisis de sentimiento se aborda técnicas de NLP para la representación de los documentos u oración [136][138]. Los métodos de aprendizaje automático supervisado requieren de una representación, como es de un vector de características ponderadas [134][144]. Las técnicas más utilizadas:

- **Vector de palabras (Bag of Words, BoW)**
  Es uno de los métodos más conocidos para la creación de características a partir del texto original, sin tener en cuenta la sintaxis, el orden de las palabras y la gramática [145]. Dentro de los documentos u oración hay palabras o características, cada comentario se muestra como un vector en el espacio, cada dimensión del espacio representa una característica del comentario [121], [146], [147]. Cada texto t se define de la forma t= (w1, w2, w3,…,w|V|)

donde $|V|$ es el tamaño del vocabulario total que tiene el corpus y cada wi toma valor [0,1] si el término o palabra aparece en el texto t [148], [149].

- **N-gramas (N-gram)**

  Son representaciones de n términos contiguos los cuales en conjunto encierran una idea [134] [150]. De esta forma, si N = 2 se construyen bigramas, agrupando los términos consecutivos de dos en dos; de la misma forma, si N = 3, se construyen trigramas, y si N= 1 se obtienen unigramas, lo que es equivalente al vector de palabras [139], [151]. El rendimiento de los n–gramas no depende de si se usa un n igual a 1, 2 o 3, pero si presenta variaciones en el rendimiento si se acompaña con la utilización de presencia o frecuencia de términos [152].

- **Ponderación de las características**

  Las características extraídas pueden ser consideradas todas de igual importancia u otorgarles distintos pesos en función de algún tipo de criterio; dichos pesos determinan la relevancia de cada característica dentro del comentario al que pertenecen y, por tanto, influyen a la hora de clasificar los textos por parte de los algoritmos de aprendizaje supervisado [153], las más aplicadas son:

  - **Ponderación binaria (Binary Term Ocurrences, BTO)**

    También conocida como presencia de términos (Term Presence, TP), dada una lista con todas las características de todos los comentarios del corpus de entrenamiento, se indicará con un valor 1 aquellas características que formen parte del mismo, y con un 0 en caso contrario [153]–[155].
  - **Frecuencia absoluta (Term Ocurrences, TO)**

    Cada característica tendrá un peso igual al número de veces que aparece en un comentario dado [153]–[155].
  - **Frecuencia relativa (Term Frequency, TF)**

    Es igual que TO, pero al valor de cada característica se le aplica un proceso de normalización Euclídea que tiene en cuenta el número de características del comentario al que pertenecen y sus frecuencias absolutas [148], [149], [153]–[155]. Se calcula el número de ocurrencias de cada palabra encontrada en un documento, de este modo, a las palabras que se encuentran con frecuencia se les asigna valores de puntuación más altos, mientras que a las palabras que se encuentran con poca frecuencia se les asignan

valores de puntuación más bajos, puede ser problemático, ya que las palabras frecuentes dominarán sobre las que se encuentran con poca frecuencia, para algunas tareas en NLP, algunas palabras rara vez encontradas pueden ser palabras específicas del dominio y más informativas sobre el contexto; para eliminar los problemas asociados con la ponderación basada en TF, se puede utilizar una frecuencia de documento inversa para medir la frecuencia de palabras raras en los documentos de texto [144].

- **Frecuencia de termino-Frecuencia inversa de documento (Term Frequency-Inverse Document Frequency, TF-IDF)**

  Otorga una mayor importancia a aquellas características que aparecen un mayor número de veces en el corpus, pero en pocos comentarios del mismo [136], [150], [153], [154]. Estos términos son los que suelen ayudar a identificar con mayor facilidad las distintas clases existentes; de esta forma, se evitan los problemas que implica el uso la frecuencia absoluta o relativa en donde las características más repetidas son las que tienen mayor importancia independientemente del tipo de comentario en el que aparezca [135]. TF es el número de veces que aparece un término particular en el texto, IDF mide la frecuencia de ocurrencia de cualquier palabra en todos los documentos [140], y está dada por la siguiente formula: IDF(w_k)= $|D|$/DF((w_k), donde $|D|$ es número de documentos en la colección, DF es el número de documentos en colección en la que la palabra w_k aparece [148].

- **Incrustación de palabras (Word Embendding)**

Para la clasificación de texto, la representación basada en incrustación de palabras es un esquema eficaz, que se puede utilizar junto con algoritmos de aprendizaje automático y arquitecturas de aprendizaje profundo; el uso de la incrustación de palabras permite representar documentos de texto de forma compacta y más expresiva; la representación basada en la incrustación de palabras proporciona aprendizaje mediante expresiones distribuidas de palabras que existen en un espacio de baja dimensión [156]. Las incrustaciones de palabras se han basado en la hipótesis distributiva, según esta hipótesis, las palabras con significados similares deberían encontrarse en un contexto similar, por lo tanto, la representación basada en vectores tiene como objetivo capturar las características de los vecinos de una palabra, de esta manera, se puede capturar la similitud entre las palabras, en este esquema, se ha utilizado un gran conjunto de documentos no supervisados para extraer

el significado semántico y sintáctico entre las palabras [157]. A continuación, se describe brevemente los esquemas de representación más utilizados:

- **Word2Vec**

  Es un modelo predictivo, de aprendizaje no supervisado que tiene como objetivo capturar la relación semántica entre palabras en función de su co-ocurrencia en documentos de un corpus específico [158]. La idea principal de Word2Vec es detectar el contexto de las palabras utilizando enfoques de aprendizaje profundo, consta de dos modelos, el modelo continuo de bolsa de palabras (Continuous Bag of Words, CBoW), y el modelo continuo de Skip-Gram [159]. El modelo CBoW predice la palabra objetivo a partir de las palabras de contexto que la rodean en un tamaño de ventana de k, y el modelo Skip-Gram predice las palabras de contexto, dada la palabra objetivo [157].

- **FastText**

  Es un esquema de representación computacionalmente eficiente para aprender incrustaciones de palabras a partir de documentos de texto; en este esquema, cada palabra ha sido considerada como una bolsa de n-gramas de caracteres [160]. En comparación con Word2Vec, el esquema FastText puede producir un mayor rendimiento predictivo para lenguajes morfológicamente ricos y en palabras raras [161].

- **Vectores globales (Global Vectors, GloVe)**

  Es un modelo no supervisado basado en conteo, que busca construir una representación vectorial de palabras basada en la matriz de conteo de co-ocurrencia [157]. Sobre la base de las estadísticas globales de co-ocurrencia palabra-palabra obtenidas del corpus de texto, se ha llevado a cabo una formación, y con base en el proceso de entrenamiento, se han extraído estructuras lineales del espacio vectorial de palabras [162].

2.1.12.4.4. Clasificación de sentimiento

En esta fase se clasifica una nueva opinión como positivo, negativo o neutro mediante la implementación de enfoques de aprendizaje automático, aprendizaje profundo, basados en léxico o híbridos [107], [163].

2.1.12.4.4.1. Enfoque de aprendizaje automático (Machine Learning)

En el análisis de sentimiento basado en el aprendizaje automático, hay dos etapas principales, la extracción de características de los datos y su representación en términos de vectores de características, y el entrenamiento de los algoritmos de aprendizaje supervisado en los vectores de características para obtener el modelo de aprendizaje [144]. Se encuentra los métodos: supervisado, no supervisado y semi-supervisado.

2.1.12.4.4.1.1. Aprendizaje supervisado

Para el análisis de sentimiento se define un conjunto de entrenamiento D = {x+, x-, , xt}, donde cada registro es etiquetado a una clase; el modelo de clasificación está relacionado con las características en el registro subyacente a una de las etiquetas de clase; luego, para una instancia dada de clase desconocida, el modelo se usa para predecir una etiqueta de clase [164].

Dada una colección de documentos/comentarios D= {d1...dn} y categorías predefinidas en el conjunto C = {positivo, negativo, neutro}, la clasificación del sentimiento es clasificar cada di en D, con una etiqueta expresada en C [132].

A continuación, se detalla brevemente los algoritmos más usados en la clasificación de sentimiento:

- **Bayes ingenuo (Naïve Bayes, NB)**

  El clasificador bayesiano ingenuo es probabilístico [107]; está basado en el Teorema de Bayes; sea {A1, A2, … , An} un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero; sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|Ai)$; entonces, la probabilidad de $P(Ai|B)$ viene da por la expresión: $P(Ai|B)=P(B|Ai)*P(Ai)P(B)$ [149], [164].

  En el caso concreto de la clasificación de textos, los sucesos excluyentes y exhaustivos son las diferentes clases que se pueden asignar a un comentario, de manera que no es posible asignar más de una simultáneamente (excluyentes) y esas clases son todos los tipos que existen (exhaustivos) [149]. Los algoritmos NB suelen recibir el apelativo de “ingenuos” debido a que en sus cálculos las características seleccionadas para representar a los ejemplos de entrenamiento son estadísticamente independientes y contribuyen por igual en el proceso de clasificación [165]. Genera un modelo de probabilidades a partir de los textos de entrenamiento, en base a las frecuencias de las palabras; si un dato sin etiquetar, posee

frecuencias de palabras similares al de un dato de etiqueta conocida, es probable que ambos pertenezcan a una misma categoría [149].

- **Bayes ingenuo multinomial (Multinomial Naive Bayes, MNB)**

  El clasificador bayes ingenuo multinomial es probabilístico, se puede usar para predecir la etiqueta de un texto, calcula la probabilidad de cada etiqueta para el texto de entrada y luego genera la etiqueta con la probabilidad más alta como salida [138], [166].

  Construye un modelo de lenguaje que asume la independencia condicional entre las características lingüísticas, es simple y se puede escalar de forma endeble para un gran número de clases, a diferencia de los clasificadores discriminativos [167]. Al ser un modelo probabilístico, es muy fácil de extender para tareas de modelado estructurado, como clases de múltiples documentos y etiquetas [129], [168], [169].

- **Árboles de decisión (Decision Tree, DT)**

  Los árboles de decisión se basa en la división recursiva del conjunto de datos de entrenamiento, al cual se le aplica una serie de condiciones sobre los valores de los atributos del texto [107]. Las condiciones, en clasificación de texto, corresponden normalmente a la evaluación de presencia o ausencia de una o más palabras en el documento/oración [164]. La división del espacio de datos se realiza recursivamente hasta que los nodos hoja contengan un número mínimo de registros que se utilizan para fines de clasificación [164].

- **Máquina de vectores de soporte (Support Vector Machines, SVM)**

  Las máquinas de vectores de soporte, es un clasificador lineal, tiene como enfoque principal determinar los separadores lineales en el espacio de búsqueda que mejor pueden separar las diferentes clases [135], [170], [171]. Este tipo de algoritmos cuenta con una serie de parámetros que permiten ajustar su configuración interna y así optimizar los resultados durante el proceso de clasificación; el parámetro del kernel se utiliza cuando no es posible separar las muestras mediante una línea recta, plano o hiperplano de N dimensiones, permitiendo tal separación mediante otro tipo de funciones matemáticas como polinomios, funciones de base radial Gaussiana, Sigmoid u otras; el parámetro regularización (también conocido como “C”) permite crear un margen blando de manera que se consientan ciertos errores en la clasificación y se evite el sobreentrenamiento; y el parámetro gamma que determina la distancia máxima a partir de la cual una muestra pierde su influencia en la configuración del

vector de soporte, y margen, que es la separación entre el vector y las muestras de cada clase más cercanas al mismo [149], [164].

En la clasificación de texto se usa un conjunto de entrenamiento donde cada muestra tiene un peso y un vector asociado que separa lo más posible los casos positivos de los negativos; generalmente, los datos usados son palabras (unigramas) a las cuales se les asigna un peso durante la fase de aprendizaje con el valor $\delta \geq 0$; cada palabra etiquetada que cumpla que su peso $\delta > 0$ es llamado vectores de soporte; de esta manera los vectores de soporte separan el híperplano entre la clasificación positiva y negativa; así, las palabras que aún no han sido entrenadas, son asignadas a los vectores de soporte más cercanos de acuerdo a una ecuación que incluye la función kernel apropiada [141]. Para seleccionar las características a ocupar en SVM correctamente hay varios métodos; usualmente se ocupan palabras solas que se usen una cierta cantidad de veces en el texto a analizar; también es posible seleccionar bi-grams (dos palabras juntas), tri-grams (3 palabras juntas), la categoría gramatical de la palabra, entre otros [141], [149].

- **K-Vecinos más cercanos (K-Nearest Neighbours, K-NN)**

  El algoritmo K-Vecinos más cercanos asigna cada documento a la clase mayoritaria de sus k vecinos más cercanos donde k es un parámetro; el fundamento de la clasificación K-NN es que, según la hipótesis de contigüidad, se espera que un documento de prueba d tenga la misma etiqueta que los documentos de capacitación ubicados en la región local que rodea a d [149].

- **Regresión Logística (Logistic Regression, LR)**

  El algoritmo regresión logística, utiliza función sigmoidea que genera una probabilidad entre 0 y 1 [172], [173]. La regresión logística pertenece a la familia de clasificadores exponenciales o logarítmicos lineales, funciona extrayendo un conjunto de características ponderadas de la entrada, tomando registros y combinándolos linealmente (cada característica se multiplica por un peso y luego se suma) [174], técnicamente clasifica una observación en una de dos clases [169], [175].

- **Bosque aleatorio (Random Forest, RF)**

  El algoritmo bosque aleatorio, se denomina simplemente como una colección de árboles y cada árbol es diferente entre sí [176]. Es un método de aprendizaje conjunto para clasificación

y regresión que construye una serie de árboles de decisión en tiempo de entrenamiento y entrega la clase que es el modo de salida de clases por árboles individuales [177]. RF divide cada nodo utilizando el mejor entre un subconjunto de predictores elegidos aleatoriamente en ese nodo; se crea un nuevo conjunto de datos de entrenamiento a partir del conjunto de datos original, luego, se cultiva un árbol mediante la selección aleatoria de características [178]. En este método, muchos clasificadores se generan a partir de subconjuntos más pequeños de los datos de entrada y luego sus resultados individuales se agregan en función de un mecanismo de votación para generar la salida deseada de la entrada [176].

- **Conjunto de votación (Voting ensemble, VE)**

  El clasificador de conjunto de votación entrenan a múltiples clasificadores para resolver el mismo problema [179]. Los resultados de varios modelos se agregan por votación, pueden ser buenos o malos, por lo que el resultado de la agregación es el mejor [169]. La idea básica es combinar diferentes modelos de aprendizaje automático y predecir las etiquetas finales votando o promediando [180]. Por ejemplo, hay cinco clasificadores, clasificador1→clase2, clasificador2→clase2, clasificador3→clase3, clasificador4→clase4, clasificador5→clase5, entonces la clase de predicción final es la clase 2 [169].

2.1.12.4.4.1.2. Aprendizaje no supervisado

El aprendizaje no supervisado se basa en observaciones, donde los datos de entrada para entrenamiento no se encuentran etiquetados o clasificados; en los modelos de aprendizaje no supervisado se debe aprender las relaciones entre los elementos del conjunto de datos de entrada [164]. Uno de los algoritmos más utilizados en aprendizaje no supervisado es el de agrupamiento.

- **Agrupamiento (Clustering)**

  El objetivo del algoritmo de agrupamiento es detectar potenciales agrupaciones en el conjunto de datos de entrada de acuerdo con la similitud que exista entre los elementos del conjunto de datos, un ejemplo es el algoritmo de las k medias [140]. El clustering se realiza cuando se identifican elementos extraños, eventos u observaciones que ocurren en el conjunto de datos de entrada los cuales son significativamente diferentes a la mayoría de los datos [140].

2.1.12.4.4.1.3. Aprendizaje semi-supervisado

El aprendizaje semi-supervisado es la combinación del aprendizaje supervisado y el no supervisado; en éste se aprende con la ayuda de dos conjuntos, uno que contiene datos asociados a una clase, y el otro que contiene datos no asociados a una clase; la idea es aprender con los datos asociados a su clase y asociar una clase a los datos que no contienen asociada una clase [132], [155].

2.1.12.4.4.2. Enfoque de aprendizaje profundo (Deep Learning)

Para el análisis de sentimiento basado en el aprendizaje profundo, el corpus de texto se representa mediante esquemas de incrustación de palabras [144]. Las arquitecturas tienen como objetivo identificar modelos de aprendizaje sobre la base de múltiples capas o etapas de procesamiento de información no lineal de forma jerárquica [181].

A continuación, se detalla brevemente los algoritmos más usados en la clasificación de sentimiento:

- **Red de memoria a corto y largo plazo (Long Short-Term Memory, LSTIM)**

  Es una arquitectura de red neuronal recurrente (Recurrent Neural Networks, RNN) utilizada en el campo del aprendizaje profundo [182]–[184]. Consta de unidades o bloques de memoria en la capa oculta recurrente, que contiene celdas de memoria con autoconexiones que almacenan el estado temporal de la red; además tiene unidades multiplicativas especiales llamadas puertas para controlar el flujo de información en la red, cada bloque de memoria en la arquitectura original contiene una puerta de entrada y una puerta de salida [185].

  La puerta de entrada controla el flujo de activaciones de entrada hacia la celda de memoria y la puerta de salida controla el flujo de salida de activaciones de celda hacia el resto de la red, la puerta de olvido se agrega al bloque de memoria que escala el estado interno de la celda antes de agregarlo como entrada a la celda a través de la conexión autorrecurrente de la celda por lo tanto, olvida o restablece la memoria de la celda de manera adaptativa [182], [184], [186], [187]. Además, la arquitectura LSTM moderna también puede contener conexiones de rendija desde sus celdas internas a las puertas en la misma celda que ayudan a conocer la sincronización precisa de las salidas [188].

Las redes LSTM están diseñadas principalmente para problemas de predicción de secuencias; funcionan aprendiendo una función "f" que mapea el valor de entrada (X) en la secuencia de salida Y mediante la siguiente ecuación: Y(t)=f(X(t)) [187].

- **Red de memoria a corto y largo plazo bidireccional (Bidirectional Long Short-Term Memory, BiLSTM)**

  El LSTM bidireccional consta de una capa de memoria a largo y corto plazo hacia adelante y una capa de memoria a largo y corto plazo hacia atrás; el principio de funcionamiento es el siguiente: la capa de avance captura la información histórica de la secuencia, la capa hacia atrás captura la información futura de la secuencia; ambas capas están conectadas a la misma capa de salida, y la información de contexto de secuencia se considera completamente [189]–[191].

  BiLSTM puede resolver el problema del modelo LSTM tradicional que no puede procesar palabras relacionadas en oraciones de atrás hacia adelante, considera completamente los contextos anteriores y posteriores de las oraciones para extraer características semánticas bidireccionales [192]–[194].

2.1.12.4.4.3. Enfoque basado en el léxico

El enfoque basado en el léxico clasifica un texto según las palabras positivas, negativas y neutras que contiene; este enfoque no requiere una fase de entrenamiento, se divide en dos categorías: basados en diccionarios y en corpus [130].

- **Basado en diccionarios**

  El diccionario es un pequeño conjunto de palabras de opinión que se recopila manualmente con orientaciones conocidas o vinculadas, luego, este conjunto crece al buscar en los conocidos corpus como WordNet, SentiWordNet, entre otros sus sinónimos y antónimos [130], [132]. Las palabras recién encontradas se añaden a la lista de semillas y luego comienza la siguiente iteración, el proceso iterativo se detiene cuando no se encuentran palabras nuevas, una vez que se completa el proceso, es posible realizar una inspección para descartar o corregir los errores [130].

- **Basado en corpus**

  Este método basado en Corpus ayuda a resolver el problema de encontrar palabras de opinión con orientaciones o relaciones específicas dentro del contexto de análisis errores, los dos métodos de este enfoque son [130], [132]:

  - **Enfoque estadístico**

    Las palabras que muestran un comportamiento errático en el comportamiento positivo se considera que tienen polaridad positiva; si muestran recurrencia negativa en texto negativo, tienen polaridad negativa; si la frecuencia es igual tanto en el texto positivo como en el negativo, la palabra tiene polaridad neutra [130].

  - **Enfoque semántico**

    Este enfoque asigna valores de sentimiento a las palabras y las palabras que están semánticamente más cerca de esas palabras; esto se puede hacer encontrando sinónimos y antónimos con respecto a esa palabra [130].

2.1.12.4.4.4. Enfoque híbrido

Este enfoque combina el enfoque basado en el aprendizaje automático y el léxico; la principal ventaja es la simbiosis léxico/aprendizaje, la detección y medición del sentimiento a nivel de concepto y la menor sensibilidad a los cambios en el dominio del tema [163].

2.1.13. Lógica difusa (Fuzzy Logic)

La lógica difusa proporciona las bases del razonamiento aproximado, de manera que a partir de premisas imprecisas se formula el conocimiento [195]. Es una lógica alternativa a la lógica clásica, que es capaz de añadir cierto grado de incertidumbre, supliendo así de alguna manera la poca capacidad de expresión de la lógica clásica, por ello, un sistema de razonamiento difuso modela la incertidumbre inherente en el proceso de razonamiento humano, plasmando su conocimiento y experiencia en un conjunto de expresiones lingüísticas que manejan palabras en lugar de valores numéricos [196].

Así, las variables no asumen dos valores antagónicos (uno y cero), sino que asume grados de verdad (poco negativo, muy negativo, entre otros); la lógica difusa está construida sobre el concepto de variable lingüísticas, que es una variable cuyos valores son palabras o sentencias en un lenguaje natural o artificial, y se define mediante conjuntos difusos [197].

2.1.13.1. Sistema difuso

Un sistema difuso permite establecer reglas de combinación en las que se maneje cierta incertidumbre [196]. En general, los sistemas difusos se pueden clasificar, según su estructura en:

- **Sistema difuso puro**

  En este sistema tanto las entradas como las salidas son conjuntos difusos, se componen de una base de reglas difusas y de un mecanismo de inferencia difuso, la base de reglas difusas es un conjunto de reglas (IF-THEN) expresadas en forma lingüísticas, el mecanismo de inferencia difusa busca en la base de reglas difusas las que son aplicables a la situación actual, operando con ellas de tal forma que el espacio de entradas se proyecte en el espacio de salidas [198]–[200].

- **Sistema difuso con fuzzificador/defuzzificador**

  En este sistema tanto las entradas como las salidas son valores numéricos concretos; a la entrada es necesaria una fase de fuzzificación que se encarga de traducir la entrada a los conjuntos difusos, después pasa por un sistema difuso puro, que contiene la base de reglas y el mecanismo de inferencia, y finalmente, a la salida es necesaria una fase de defuzzificación para transformar el conjunto difuso a un valor numérico [200]–[206].

Un sistema difuso se compone de:

- **Variable lingüística**

  Es una variable que puede tomar como valor palabras del lenguaje natural o números, estas palabras suelen estar ligadas a conjuntos difusos [197], [198], [200]. Por ejemplo, la variable lingüística “altura de una persona” (Figura **2**).

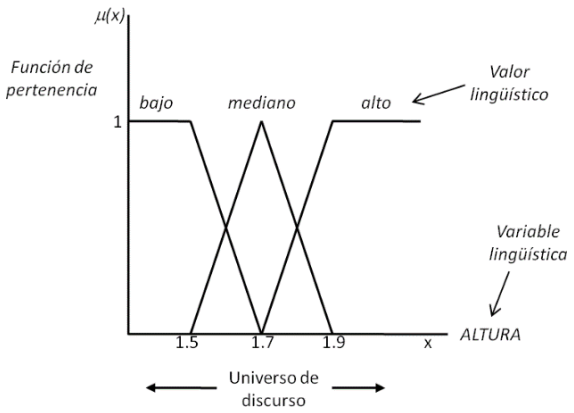


Figura 2. Variable lingüística de un sistema difuso

- **Universo de discurso**

  Rango de valores que pueden tomar los elementos que poseen la propiedad expresada por la variable lingüística [198]. En el caso de la variable lingüística “altura de una persona”, sería el conjunto de valores comprendido entre 1.4 y 2.5 m.

- **Conjunto difuso**

  Es el valor lingüístico junto a una función de pertenencia, el valor lingüístico es el “nombre” del conjunto, y la función de pertenencia se define como aquella aplicación que asocia a cada elemento del universo de discurso, el grado con que pertenece al conjunto difuso [198], [204], [207]. Un conjunto es nítido si su función de pertenencia toma valores en {0,1}, y es difuso si toma valores en [0,1] [198]. Por ejemplo, la variable lingüística “altura de una persona”, toma valores en el universo de discurso U = [1.4, 2.50]. La clasificación difusa podría ser en tres conjuntos difusos (o valores lingüísticos): bajo, mediano y alto (Figura 2).

- **Funciones de pertenencia**

  Un conjunto difuso permite describir el grado de pertenencia de un objeto a una determinada clase, dicho grado de pertenencia viene descrito por una función de pertenencia [200], [202], [208]. Las funciones L y GAMMA se usan para calificar valores lingüísticos extremos, las funciones PI y LAMBDA (triangular) se usan para describir valores intermedios; su principal diferencia reside en que la función PI implica un margen de tolerancia alrededor del valor que se toma como más representativo del valor lingüístico asociado al conjunto difuso [204][206].

En el ejemplo de la Figura 2 se puede ver que para los valores bajo y alto las funciones de pertenencia son de tipo trapezoidal, mientras que para el valor mediano la función de pertenencia es de tipo triangular.

- **Mecanismo de inferencia difusa**

  En el proceso de inferencia (Figura **3**) toda variable numérica de entrada debe ser borrosificada, así, el fuzzificador o borrosificador toma los valores numéricos provenientes del exterior y los convierte en valores que puedan ser procesados por el mecanismo de inferencia; estos valores son los niveles de pertenencia de los valores de entrada a los diferentes conjuntos difusos en los que se ha dividido el universo de discurso, es decir, a los diferentes conjuntos difusos de las variables de entrada al sistema [200], [202]–[206].

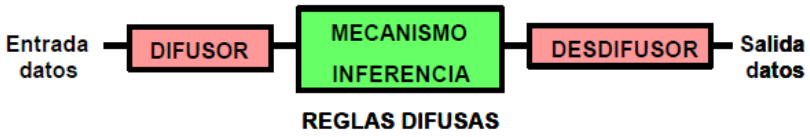


Figura 3. Esquema general de un sistema basado en lógica difusa

El proceso de inferencia difusa es el centro del sistema de razonamiento y se encarga de interpretar el conjunto de reglas IF-THEN disponibles en la base de conocimiento; mediante las reglas difusas se combinan uno o más conjuntos difusos de entrada, que son los antecedentes, a los que se les asocia un conjunto difuso de salida, que es el consecuente [203]–[206].

Dependiendo de los operadores elegidos se tendrán diferentes interpretaciones de las mismas reglas; los métodos de inferencia más comunes son los de Mamdani [209] y Takagi-Sugeno, se diferencian fundamentalmente en el formato de las reglas, ya que la salida es diferente [210].

Las reglas de tipo Mamdani es: IF u_1 es A_1 Y u_2 es A_2 Y.. Y u_n es A_n THEN v es B

Donde los u_i y v son variables lingüísticas y los A_i y B representan los valores lingüísticos que dichas variables pueden asumir, por lo tanto, en un sistema difuso tipo Mamdani tanto el antecedente como el consecuente de las reglas están dados por expresiones lingüísticas [205], [206], [209].

Las reglas de tipo Sugeno es: IF u_1 es A_1 Y u_2 es A_2 Y..Y u_n es A_n THEN v = f (u_1, u_2,. u_n)

Donde los u_i son variables lingüísticas y los A_i representan los valores lingüísticos que dichas variables pueden asumir, v es la variable de salida y f representa una función lineal de las entradas, por lo tanto, el consecuente de estas reglas ya no es una etiqueta lingüística, sino que es una función de la entrada que tenga el sistema en un momento dado [210].

La salida que genera el mecanismo de inferencia es una salida difusa, por lo que no podría ser interpretada por un elemento externo (como puede ser un controlador) que solo pueda manipular información numérica [205]. El defuzzificador o desborrosificador es el encargado de convertir la salida del sistema difuso para que pueda ser interpretada por elementos que solo procesen información numérica [29][202]–[206]. Para realizar esta desborrosificación se tienen diferentes métodos, entre los que se encuentran:

- **Criterio de máximo (MC)**
  La salida es aquella para la cual la función de membresía alcanza su máximo valor [211].
- **Método de la media de máximo (Middle of Máximum, MOM)**
  La salida es el valor medio de los valores cuyas funciones de membresía alcanzan el valor máximo [202][211][212].
- **Método del máximo más chico (Smallest of Máximum, SOM)**
  La salida es el mínimo valor de todos aquellos que generan el valor más alto de la función de membresía [202][211][212].
- **Método del máximo más grande (Largest of Máximum, LOM)**
  La salida es el máximo valor de todos aquellos que generan el valor más alto de la función de membresía [202][211][212].
- **Método de centro de área (Centroid of Area, COA) o de centro de gravedad (Center of Gravity, COG).**
  Este método proporciona un valor nítido basado en el centro de gravedad del conjunto difuso, se calcula el área y el centro de gravedad o centroide de cada subárea y luego se toma la suma de todas estas subáreas para encontrar el valor desborroso para un conjunto difuso discreto [29][204][212].
- **Bisector de área**
  La salida es el valor que separa el área bajo la curva en dos sub-áreas iguales [202].

2.2. Revisión sistemática de la literatura de evaluación entre pares basada en análisis de sentimiento de texto educativo

Para efectuar el objetivo 1, se realizó la revisión sistemática de la literatura (Systematic Literature Review) como estrategia para identificar los estudios más relevantes que tiene la evaluación entre pares basada en retroalimentación y análisis de sentimiento de texto educativo.

El proceso de esta revisión se efectuó mediante tres fases:

Fase 1: Planificación de la revisión. El objetivo de este estudio se centró en contestar las siguientes preguntas de investigación:

RQ1. ¿Cuál es la tendencia de estudio acerca de retroalimentación entre pares y análisis de sentimiento de texto educativo?

RQ2. ¿Qué dominios del conocimiento y aspectos han sido objeto de estudio en retroalimentación entre pares, y análisis de sentimiento de texto educativo?

RQ3: ¿Cuáles son las técnicas, métodos y algoritmos utilizados en análisis de sentimiento de texto educativo?

En base al enfoque de [213], en la exploración inicial se realizan búsquedas sistemáticas y resúmenes formales de la literatura para identificar y clasificar los resultados de los estudios sobre un tema en particular.

La selección de publicaciones científicas sobre la literatura de evaluación entre pares basada en retroalimentación y análisis de sentimiento se realizó en: ScienceDirect, Web of Science, SpringerLink, IEEE Xplore. Para la preselección se incluyeron artículos de publicaciones completas, y conferencias ya que es un campo poco explorado en artículos científicos y se necesitaba de más información para él análisis. Las búsquedas se realizaron en base a los términos: peer assessment, peer feedback, sentiment analysis, education que debería aparecer en todo el contenido y en el periodo 2014-2020. Se excluyeron manuscritos no revisado por pares y estudio teórico. Se localizaron 2525 publicaciones. La Tabla **2** presenta los resultados obtenidos para cada base de datos seleccionada.

Tabla 2. Número de estudios extraídos para cada base de datos

Nombre de la base de datos	Cadena de búsqueda	Cantidad de estudios
Web of Science	(TS=("peer assessment" or "peer feedback" and "sentiment analysis" and education))	612
ScienceDirect	"peer assessment" OR "peer feedback" AND "sentiment analysis" AND education	799
SpringerLink	"peer assessment" or "peer feedback" AND "sentiment analysis" AND education'	791
IEEE Xplore	((((("Full Text & Metadata":"peer feedback") OR "Full Text & Metadata":"peer assessment") OR "Full Text & Metadata":"sentiment analysis") AND "Full Text & Metadata":education)	323
	Total	2525

Fase 2: Conducción. Se ejecutó para dar respuestas a las preguntas RQ1, RQ2 y RQ3, el proceso se dio a través de la evaluación y extracción de los datos de artículos [214]. Se refinó la búsqueda incluyendo una combinación de términos secundarios: natural language processing, machine learning, deep learning, opinion mining, text mining, artificial intelligence, higher education en palabras claves o título o resumen, depurando a 167 publicaciones científicas. La Tabla **3** presenta los resultados obtenidos para cada base de datos seleccionada.

Tabla 3. Número de estudios extraídos por términos para cada base de datos

Nombre de la base de datos	Cadena de búsqueda	Cantidad de estudios
Web of Science	(TI=("peer feedback" OR "peer assessment") OR TI=("sentiment analysis" AND "education")) AND (AB=("natural language processing" OR "machine learning" OR "deep learning" OR "higher education" OR "opinion mining" OR "artificial intelligence" or "text mining"))	70
ScienceDirect	Title, abstract, keywords: "higher education" OR "natural language processing "OR "machine learning" OR "deep learning" OR "higher education" OR "opinion mining" OR "artificial intelligence" or "text mining"	
	Title: "peer feedback" OR "peer assessment" OR "sentiment analysis"	30
SpringerLink	"higher education OR natural language processing OR machine learning OR deep learning OR higher education OR opinion mining OR artificial intelligence OR text mining'"	18
IEEE Xplore	((("Document Title":"peer feedback") OR "Document Title":"peer assessment") OR "Document Title":"sentiment analysis") AND (((((("Abstract":"natural language processing") OR "Abstract":"machine learning") OR "Abstract":"deep learning") OR "Abstract":"higher education") OR "Abstract":"opinion mining") OR "Abstract": "artificial intelligence") OR "Abstract":"text mining")	49
	Total	167

Siguiendo el enfoque de Zott, se vuelve a refinar el conjunto de datos para obtener el listado final [3]. Se leyeron los resúmenes, se eliminaron las redundancias considerando

solamente los documentos relevantes para el estudio. Se obtuvo 89 estudios. La Tabla **4** presenta los resultados obtenidos para cada base de datos seleccionada.

Tabla 4. Número de estudios extraídos para cada base de datos mediante el enfoque de Zott

Nombre de la base de datos	Cantidad de estudios
Web of Science	33
ScienceDirect	10
SpringerLink	17
IEEE Xplore	29
Total	89

Fase 3: Informe de los resultados. Se realizó un proceso de validación donde se usaron criterios: inclusión y exclusión de la revisión, cobertura de los estudios pertinentes, evaluación de la calidad/validez de los estudios incluidos y descripción de los datos o estudios básicos [215]. Los criterios de inclusión y exclusión que utilizamos en estos estudios se enumeran en la Tabla **5**.

Tabla 5. Criterios de inclusión y exclusión

Criterios de inclusión	Criterios de exclusión
Publicado entre 2014 y 2020	Estudio teórico
Estudio empírico de retroalimentación entre pares o análisis de sentimiento de texto educativo en un contexto auténtico	Manuscrito no revisado por pares
Escrito en ingles	Texto completo no disponible para investigadores

A los 89 estudios seleccionados se realizó la extracción de datos y análisis de contenido, para dar respuestas a RQ1, RQ2 y RQ3.

En RQ1, se procedió a identificar la tendencia de estudio en bases de datos electrónicas, cuando el estudio captaba más de una temática, se englobó en la más influyente, se determinaron tres temáticas: retroalimentación entre pares, análisis de sentimiento de retroalimentación entre pares, y análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje.

En RQ2, se procedió a identificar los dominios, cuando el alcance del estudio comprendía más de uno, se prefirió el más influyente, por lo tanto, se establecieron dos dominios: educación y computación. Además, se analizó mediante categorías los aspectos de los estudios. En la Tabla **6** se muestra las categorías que agrupa información que se refiere al mismo aspecto.

Tabla 6. Categorías con sus aspectos asociadas a retroalimentación entre pares y análisis de sentimiento de texto educativo

Categoría	Aspectos
Aprendizaje	Aprendizaje colaborativo, aprendizaje de constructivismo social, aprendizaje reflexivo, aprendizaje basado en la web, aprendizaje basado en problemas, aprendizaje basado en casos, aprendizaje autorregulado, aprendizaje experiencial, aprendizaje electrónico adaptativo, aprendizaje inteligente, análisis de aprendizaje, aprendizaje supervisado, aprendizaje móvil, e-learning, aprendizaje colaborativo asistido por computadora, entornos de aprendizaje interactivo, aprendizaje de preferencia, avances en tecnologías de aprendizaje
Tecnología	Redes sociales, comunicación mediada por computadora, tecnología, herramientas de enseñanza en línea, sistema, web
Inteligencia artificial	Aprendizaje automático, aprendizaje profundo, procesamiento del lenguaje natural, análisis de sentimiento, minería de opiniones, minería de textos, análisis semántico latente, algoritmos, técnicas computacionales, modelos estadísticos

En RQ3, se procedió a identificar técnicas, métodos y algoritmos utilizados en análisis de sentimiento de texto educativo.

En la Figura **4** se muestra el esquema que se aplicó para la selección y clasificación de las publicaciones científicas.

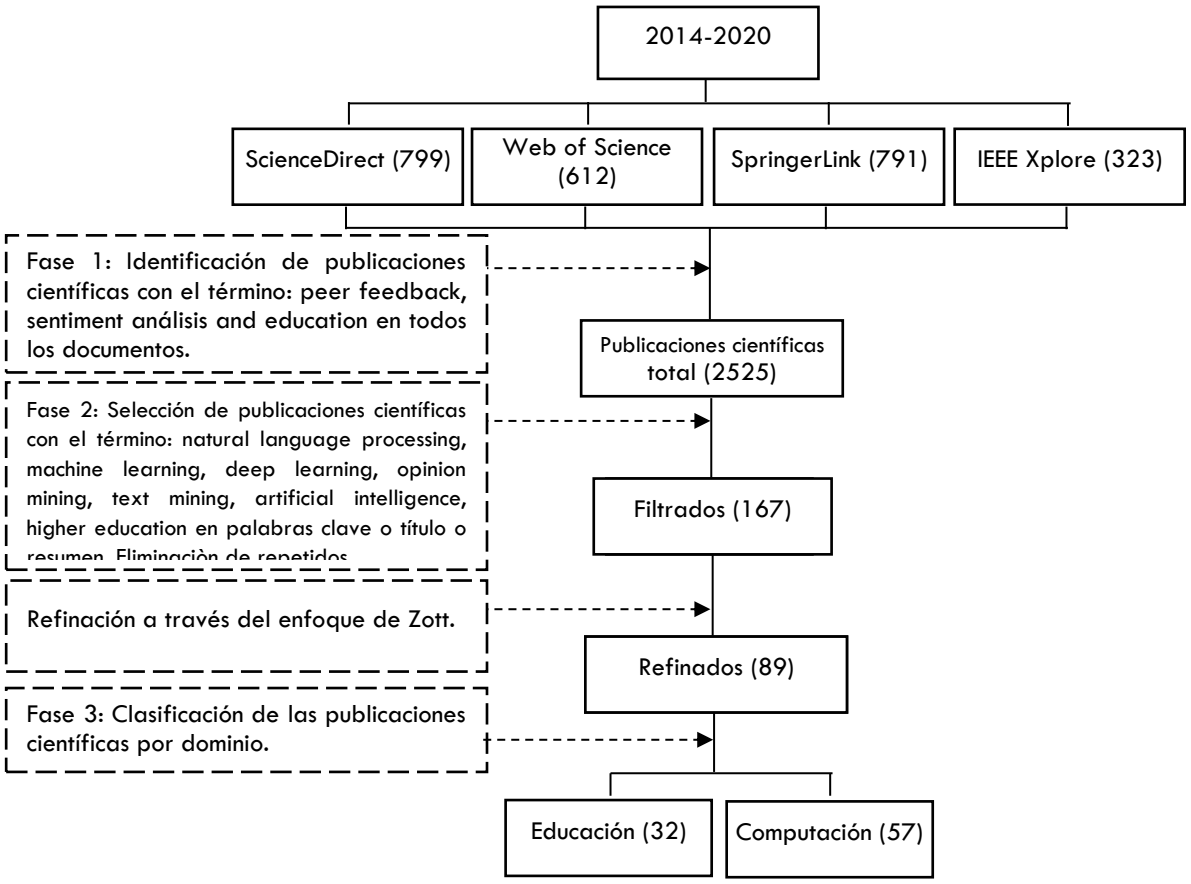


Figura 4. Esquema de selección y clasificación de las publicaciones científicas

A partir de los datos extraídos de los 89 estudios seleccionados, en esta sección se presentan los resultados obtenidos, considerando su información general y preguntas de investigación propuestas, en tres subsecciones: Tendencia de estudio de retroalimentación entre pares y análisis de sentimiento de texto educativo, dominios del conocimiento y aspectos de retroalimentación entre pares y análisis de sentimiento de texto educativo, y técnicas, métodos y algoritmos utilizados en análisis de sentimiento de texto educativo.

2.2.1. Tendencia de estudio de retroalimentación entre pares y análisis de sentimiento de texto educativo

El estudio de la tendencia en bases de datos electrónicas acerca de retroalimentación entre pares y análisis de sentimiento entre 2014-2020, exterioriza que la base de datos electrónica que ha publicado más sobre estos tópicos es Web of Science (33 estudios, 37 por ciento), siguiendo IEEE Xplore (29 estudios, 33 por ciento), SpringerLink (17 estudios, 19 por ciento), y ScienceDirect (10 estudios, 11 por ciento), y que la mayoría de los estudios se publicaron en artículos (60 estudios, 67 por ciento) con mayor énfasis en Web of Science, y conferencias (29 estudios, 33 por ciento) acentuándose en IEEE Xplore (Figura **5**).

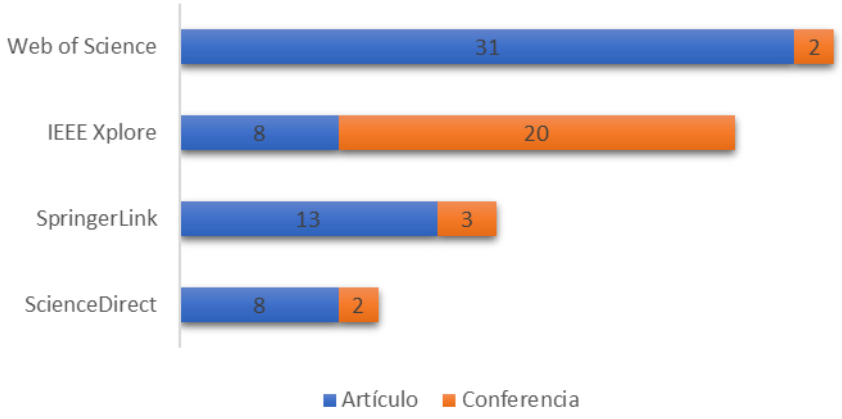


Figura 5. Número de estudios extraídos por cada base de datos

Existe un incremento sostenido de estudios entre 2018-2020. Esta evolución refleja el interés creciente de la comunidad científica por los distintos aspectos relacionados con retroalimentación entre pares y análisis de sentimiento. En donde la tendencia de estudios en análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje (SA-TL) es

del 58 por ciento (52 estudios), con crecimiento entre 2018-2020, en retroalimentación entre pares (PF) es del 37 por ciento (33 estudios), acrecentándose entre 2019-2020, y análisis de sentimiento de retroalimentación entre pares (SA-PF) es del 4 por ciento (4 estudios), reflejándose en todos los años una escasa investigación (Figura **6**).

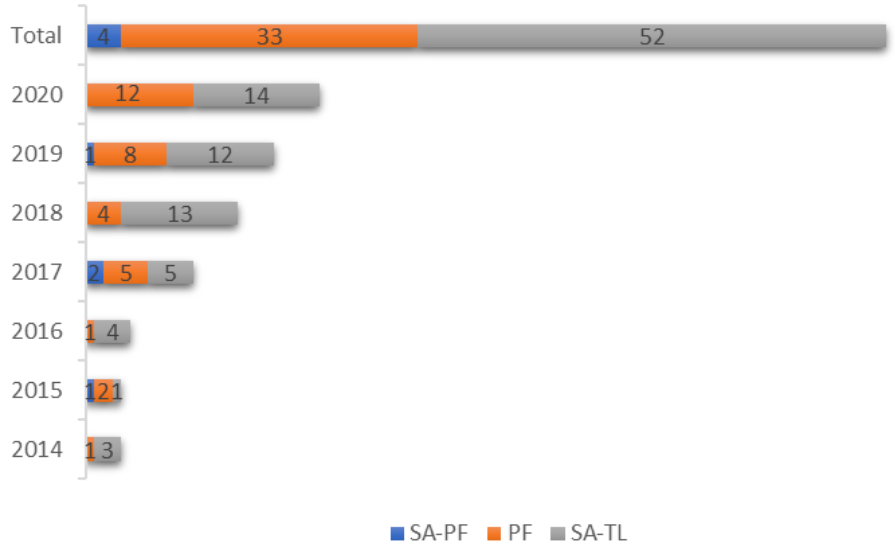


Figura 6. Número de estudios extraídos de cada temática por año

La Figura **7** muestra la tendencia de los métodos aplicados en los estudios. Retroalimentación entre pares (PF) emplea para analizar los datos una combinación de método cuantitativo y cualitativo (32 estudios, 36 por ciento), y el método de aprendizaje automático (1 estudios, 1 por ciento). Análisis de sentimiento de retroalimentación entre pares (SA-PF) utiliza para clasificar el texto el método de aprendizaje automático (4 estudios, 4 por ciento). Análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje (SA-TL) utiliza para clasificar el texto el método de aprendizaje automático (31 estudios, 35 por ciento), aprendizaje profundo (9 estudios, 10 por ciento), léxico (6 estudios, 7 por ciento) y herramientas de análisis de sentimiento (6 estudios, 7 por ciento).

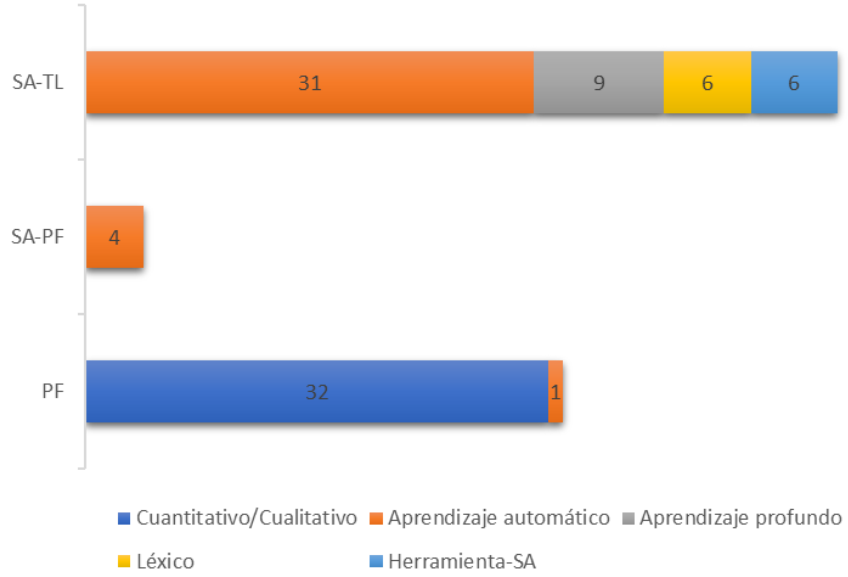


Figura 7. Número de estudios extraídos de cada método por temática

2.2.2. Dominios del conocimiento y aspectos de retroalimentación entre pares y análisis de sentimiento de texto educativo (RQ2)

Los dominios del conocimiento que han sido objeto de estudio en retroalimentación entre pares y análisis de sentimiento se encapsularon en dos dominios: educación y computación. Los estudios analizados dan cuenta de que las experiencias de retroalimentación entre pares y análisis de sentimiento, se han dado en mayor medida en el dominio de computación (57 estudios, 64 por ciento), debido a que la tecnología mejora la oportunidad de acceso a la educación mediante la colaboración de distintos actores [124], [143], [150], [216]–[220] y en segundo lugar, se destacan los estudios relacionados con el dominio de educación (32 estudios, 36 por ciento), por cuanto retroalimentación entre pares ayuda a mejorar el proceso de enseñanza-aprendizaje [221], [8], [14], revelándose entre 2019-2020 un acrecentamiento de publicaciones científicas en ambos dominios (Figura **8**).

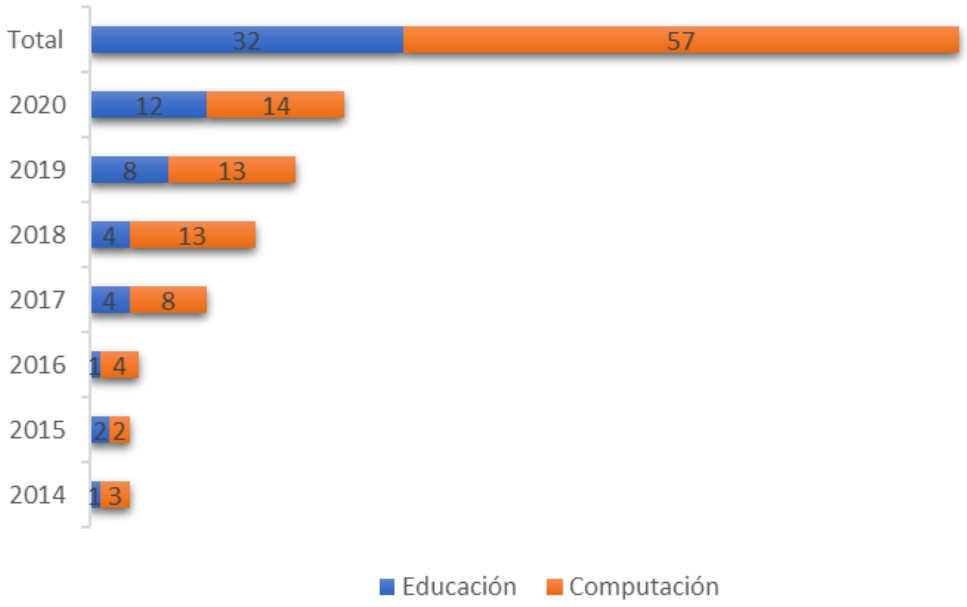


Figura 8. Número de estudios extraídos de cada dominio por año

A partir de estos datos, se recurrió al análisis de contenido mediante las categorías, distribuyéndose según su frecuencia en: Inteligencia artificial (57 estudios, 64 por ciento), aprendizaje (24 estudios, 27 por ciento) y tecnología (8 estudios, 9 por ciento), en la que se encontraron aspectos que se abordan en retroalimentación entre pares y análisis de sentimiento (Figura **9**).

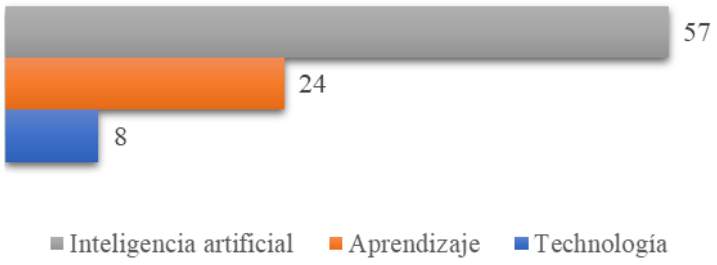


Figura 9. Número de estudios extraídos por categoría

2.2.2.1. Dominio de educación

Varios autores destacan que retroalimentación entre pares está fuertemente relacionada con la educación [8], [4], en la que han aplicado métodos cuantitativos y cualitativos para el análisis de datos. La Tabla **7** muestra aspectos que se refieren al dominio de educación.

Tabla 7. Aspectos del dominio de la educación

Descripción	Autor
Materiales Rúbricas de evaluación, cuestionarios, herramientas	[5], [222]–[224]
Mejora del proceso del aprendizaje Aspectos participativos, reflexivos, críticos, constructivista en el proceso de enseñanza-aprendizaje mediante retroalimentación entre pares.	[8], [9], [14], [221], [225], [226]
Ejemplos Andamiaje, calidad de retroalimentación en la evaluación entre pares: anonimato, niveles de retroalimentación (tarea, proceso, autorregulación).	[6], [8], [11], [227]–[231]
Revisión de escritura y contenido de retroalimentación general y/o específica (afectiva, cognitiva, metacognitiva).	[222], [232]–[236]
Educación médica aplicando retroalimentación entre pares.	[19], [221]
Desarrollo del rendimiento de la comunicación en la enseñanza de idiomas mediante retroalimentación entre pares.	[5], [10], [217]
Entornos de medios sociales y sistemas utilizados en la retroalimentación entre pares.	[217], [227], [237]–[239]

En la literatura investigada se fusiona retroalimentación entre pares con la categoría aprendizaje y tecnología.

Retroalimentación entre pares y aprendizaje

En investigaciones de retroalimentación entre pares, [8], [11] aplicaron andamiaje en la evaluación entre pares, determinando que el método promueve el aprendizaje colaborativo a los estudiantes tanto en la redacción académica, habilidades de evaluación y retroalimentación; [14] emplearon análisis factorial en la creación, validez y confiabilidad de cuatro constructos (objetividad, imparcialidad, salvaguardar represalias y retroalimentación constructiva) para determinar cómo los estudiantes perciben la evaluación de pares dentro de sus cursos y que es un componente crítico del aprendizaje experiencial; y [9], [221]–[223], [225], [226], [228], [232]–[236], [240], [241] desarrollaron experimentos que brindó a los estudiantes la oportunidad de practicar cómo dar y recibir comentarios entre pares en un escenario de aprendizaje colaborativo.

Otros autores, [6], [11], [229]–[231], [242] investigaron el efecto de retroalimentación de pares sobre: calidad de comentarios, rendimiento de los estudiantes y comparación de reflexión individual y compartida en grupos de aprendizaje basados en problemas, coligiendo que mejoró el rendimiento de los estudiantes; [243] revelaron que la combinación de la enseñanza reciproca con la instrucción explícita en el aprendizaje autorregulado mejoró el trabajo en equipo de los estudiantes y la retroalimentación; y [224] identificó si los diseños de instrucción de retroalimentación de pares influyen en las percepciones de aprendizaje de los estudiantes.

Retroalimentación entre pares y tecnología

Se efectuaron investigaciones de retroalimentación entre pares con tecnología, [4] pudieron administrar y registrar el proceso de revisión por pares, destacando que los estudiantes necesitan apoyo para alcanzar niveles de revisión reflexiva que conduzcan a una mejor práctica a través de la autogestión independiente, y determinaron que los educadores también desarrollan nuevas habilidades como mediadores en el proceso de ayudar a los estudiantes en la retroalimentación a tomar posesión de su propio aprendizaje reflexivo dentro del entorno de aprendizaje basado en la web; [5] desarrollaron una aplicación móvil para facilitar la participación de estudiantes en tareas de monitorear su producción oral, proporcionar/recibir retroalimentación correctiva y participar en diversas estrategias de comunicación, concluyeron que mejoró el rendimiento de la comunicación oral de los estudiantes, pero no el uso de la estrategia de comunicación, que es necesario aplicar expresiones prefabricadas en el proceso de retroalimentación y resalta que la evaluación entre pares puede fomentar aprendizaje constructivista-social; y [10] diseñó un sistema de retroalimentación entre pares y retroalimentación automatizada para apoyar las necesidades y prácticas del idioma inglés.

En otras investigaciones de retroalimentación entre pares, incluyendo [237] realizaron un estudio sobre las experiencias y percepciones de los estudiantes mediante herramientas de medios sociales, para proporcionar comentarios entre pares en proyectos, los resultados arrojaron que los estudiantes se beneficiaron al participar en el proceso de retroalimentación entre pares tanto en Wiki como en Facebook; [217] exploraron tres aplicaciones gratuitas: Google Docs, entorno de aprendizaje virtual de Sakai (VLE) y wiki de Sakai, el estudio informa sobre los aspectos prácticos y las experiencias de los revisores de estas aplicaciones y compara los comentarios entre las aplicaciones, los comentarios se codificaron según la cantidad, área,

naturaleza, ubicación y tipo de retroalimentación; y [239] desarrolló una aplicación en línea basada en investigaciones que apoya el aprendizaje auténtico.

Otros autores, [227] investigaron el efecto de cantidad, modalidad y satisfacción de la retroalimentación entre pares a través del sistema en línea sobre rendimiento, autoeficacia y aceptación de la tecnología, los estudiantes que enviaron sus tareas y recibieron comentarios en cantidades variables y en formas diferentes (texto o texto-video) no difirieron respectivamente en términos de puntajes en las pruebas de rendimiento, así como en las calificaciones de autoeficacia y aceptación de la tecnología; y [238] utilizaron la herramienta de aprendizaje en línea ComPAIR, que contiene un algoritmo adaptativo que utiliza los datos de comparaciones anteriores para generar pares que son cada vez más similares, los estudiantes podían reflexionar sobre sus propias respuestas y elegir la mejor de dos respuestas de emparejamientos ofrecidos en un entorno de aprendizaje colaborativo.

2.2.2.2. Dominio de computación

Incluye investigaciones de amplio espectro de retroalimentación entre pares y análisis de sentimiento. La Tabla **8** muestra aspectos que se refieren al dominio de computación.

Tabla 8. Aspectos del dominio de computación

Descripción	**Autor**
Materiales	
Algoritmos, recursos léxicos, modelos	[124], [244]–[246]
Mejora el aprendizaje	
Análisis de sentimiento proporciona información útil, a los docentes para atender las necesidades de los estudiantes, a los estudiantes en adquirir aprendizaje y a los directivos en la toma de decisiones.	[124], [134], [143], [150], [216], [218]–[220], [247]
Ejemplos	
Análisis de sentimiento mediante aprendizaje automático	[2], [16]–[19], [107], [134], [136], [139], [216], [246]–[261]
Análisis de sentimiento mediante aprendizaje profundo	[144], [218], [219], [262]–[265]
Análisis de sentimiento basado en aspecto o tópico	[123]–[126], [152], [266]–[270]
Análisis de sentimiento mediante recursos léxicos	[25], [31], [143], [150], [220], [244], [245], [271]–[273]
Análisis de sentimiento utilizando herramientas.	[274]–[277]

En la literatura investigada se fusiona retroalimentación entre pares y análisis de sentimiento con la categoría inteligencia artificial.

Retroalimentación entre pares, análisis de sentimiento e inteligencia artificial

Varios autores, aplicaron el método de machine learning en investigaciones de análisis de sentimiento de retroalimentación entre pares, para evaluar mediante una combinación de calificaciones numéricas y texto de forma libre, [19] aplicaron Support Vector Machine (SVM) y los resultados demostraron la variabilidad en la calificación de los estudiantes de medicina; [18] consideraron un detector automático de inconsistencias entre la puntuación numérica y la retroalimentación textual proporcionada por los evaluadores, a través de Natural Language Processing (NLP); con el fin de crear un ambiente de aprendizaje favorable y proteger la privacidad de los estudiantes, [278] eliminaron los comentarios negativos utilizando NLP y Naïve Bayes (NB), luego aplicaron Singular Value Decomposition (SVD), y Latent Semantic Analysis (LSA) en los comentarios positivos; y para medir la calidad de la revisión, [16] emplearon métricas: tipo de contenido, relevancia, cobertura de una presentación, el tono de opinión, el volumen y el plagio, mismas que fueron calculadas mediante NLP, SVM y Logistic Regression (LR).

En investigaciones de retroalimentación entre pares, [17] usaron NLP para evaluar automáticamente la calidad de los comentarios de los compañeros con respecto a la localización. Otros autores también aplicaron el método de machine learning en investigaciones de análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje, para intervenciones en entornos educativos, [250] desarrollaron una aplicación SentBuk para identificar la polaridad de sentimiento y cambios emocionales de los estudiantes asociados a cursos en línea (contenido, metodología, actividades, entre otras), usaron léxicos, y algoritmos J48-C4.5, SVM y NB; [261] aplicaron para clasificar las opiniones de los estudiantes NB, SVM y Artificial Neural Nework (ANN), ANN mejoro a otros en cuanto a su precisión; [249] aplicaron SVM con tres núcleos: lineal, radial y polinomial para predecir la clasificación de comentarios en positivo, negativo o neutral, y calcularon la sensibilidad, la especificidad y los valores predictivos como medidas de evaluación; [107] utilizaron: Vote Ensemble (VE), ID3, J48 y NB en análisis de sentimiento, concluyeron que el método de conjunto de votos es el modelo más eficiente; [134] emplearon SVM, Maximum Entropy (ME), NB y Complement Naive Bayes (CNB) en el análisis de sentimiento, mostrando mayor rendimiento en clases diferentes NB-CNB, y resultados aceptables en la mayoría de las combinaciones SVM-ME; [136] utilizaron Rapid Miner en el análisis de sentimiento, compararon SVM, NB, K-Nearest Neighbor (K-NN) y Neural Network classifier (NN), indicaron como resultado que NB superó a los otros algoritmos en precisión y recuperación, y K-NN en precisión; [253]

aplicaron MaxEnt, NB y ME, mostrando mejor precisión ME; [25] efectuaron un estudio de caso para verificar cómo los resultados pueden proporcionar información sobre las emociones o los patrones de los estudiantes en el MOOC, usando algoritmos supervisados: LR, SVM, Decision Trees (DT), Random Forest (RF), NB, y no supervisados basados en el léxico: Diccionaris, SentiWordNet para análisis de sentimiento, finiquitando que el más confiable fue RF; y [31] emplean un enfoque de base difusa, primero, los datos de retroalimentación de los estudiantes se preprocesan utilizando: eliminación de palabras vacías, tokenización, luego, se aplica una técnica de clasificación de sentimiento basada en el léxico mediante el cálculo de puntajes de sentimiento de palabras de opinión y cambiadores de polaridad, y un total se calcula el puntaje de sentimiento, finalmente, se aplica el sistema de lógica difusa para analizar los comentarios y la satisfacción del cliente.

En otras investigaciones de análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje, [256] se realizaron estudios de evaluación experimental que mostraron el rendimiento de varios algoritmos: NB, SVM, Multi-Layer Perceptron (MLP), Instance Bases Learning with parameter K (IBK), DT, Reduced Error Pruning Tree (REPT), LSTM y RNN; [2] adoptaron la metodología genérica de análisis de sentimiento: entrada de texto, tokenización, filtrado de palabras vacías, manejo de negaciones, derivación y lematización, clasificación y agrupamiento, y clasificación de sentimiento; [247] adoptaron el algoritmo: K-NN, gradient boosting tres (GBT), SVM, LR, NB, utilizando frecuencia de término-frecuencia inversa de documento (Term Frequency-Inverse Document Frequency, TF-IDF), el modelado lineal jerárquico para analizar las características del curso de los MOOC, el modelo GBT se desempeñó mejor que todos los demás modelos candidatos; [257] incorporan características de sintaxis pragmática, características semánticas que representan la coherencia basada en valores de similitud semántica precisos y características de polaridad basadas en sentimiento, e hicieron una comparación de diferentes modelos de aprendizaje supervisado: SVM, RF y NN; [258] aplicó dos tipos de conjuntos de datos que incluyen puntajes de calificación y comentarios textuales, usó Word Frequency (WF), K-means clustering para agrupar puntajes de calificación, y el clasificador NB para entrenar un modelo y clasificar el conjunto de datos de prueba en sentimiento negativo y positivo; [150] emplearon RF y SVM junto con recursos léxicos para realizar análisis de sentimiento a comentarios de estudiantes, el mejor modelo se logró utilizando TF-IDF y domain-specific sentiment lexicon; y [143] emplearon ID3, One R, NB, SVM, BFTree, SimpleLogistic, Logistic, BayeNet, Stacking, Ada Boost, Attribute selected classifier, Zero R, Hoeffding tree

(VFDT), y Sentiment Phrase Pattern Matching (SPPM), los patrones determinaron y filtraron la frase que aparece un nuevo término y se recopilaron en el sistema para evaluar la puntuación del sentimiento de una palabra con Teaching Senti-Lexicon.

Para intervenciones en tiempo real en las aulas, [254] desarrollaron una aplicación para conocer el estado emocional de los estudiantes, asociando información en tiempo real, utilizaron SVM, NB, CNB y ME, los resultados indicaron que SVM y CNB podrían ser utilizados para el análisis de información en tiempo real; y [139] encontraron que los mejores modelos fueron SVM y CNB; CNB puede ser una buena solución para las clases de entrenamiento diferentes y cuando no hay suficientes datos en la clase neutral.

Para evaluar Higher Education Institution (HEI), [216] emplearon Text Mining (TM) y análisis de sentimiento para identificar múltiples factores (imagen del país anfitrión, financiero, motivacional, información de la institución) que influyen en la elección de HEI, con Biterm theme model (BTM) identificaron diferentes temas discutidos por ex alumnos internacionales (IS) y Semantria para clasificar los sentimiento expresados por IS sobre su HEI, de la misma manera [248] usaron la cuenta de twitter oficial de cada universidad para establecer el ranking, generaron un léxico, luego utilizaron SVM y NB para la clasificación; y [259] aplicaron para clasificar las opiniones sobre las universidades NB, SVM, K-NN, y DT, SVM logró el mejor rendimiento y usaron Bag of Words (BoW), Part-Of-Speech (POS), TF-IDF.

Para análisis de sentimiento basado en aspectos, [266] usaron Latent Dirichlet Allocation (LDA) método estadístico para identificar aspectos de la opinión de los estudiantes de un curso; [260] usaron OpenNLP parser para etiquetado POS y sentiWordNet lexical para definir wordScore, la precisión de este sistema se mide por la precisión y el recuerdo aplicando NB en el conjunto de datos de comentarios y su opinión; [124] mediante k-mean clustering, NB, CNB, y PART, calcularon la relación semántica entre la palabra de aspecto y la oración de opinión de los estudiantes utilizando R y Weka; [125] mediante NLP, SVM y componentes Rule-Based (RB) y Dictionary-Based (DB) realizaron análisis de sentimiento basado en aspecto; [152] usaron: DT, NB, K-NN, SVM, técnicas de clasificación con una combinación de características: TFIDF, unigram, y bigram, SVM fue el mejor modelo de clasificación Multi-Class para aspectos de etiquetado; [267] apliaron LDA con diferentes temas y dispuso de un sistema de recomendación automática una vez generada la opinión; [268] categorizaron las respuestas escritas de los estudiantes a temas con LDA; [269] usaron LDA, una red de conceptos de agrupamiento de temas basada en Formal Concept Analysis (FCA) construyeron el modelo, donde el sentimiento del

tema se puede identificar midiendo sus puntajes de sentimiento; [270] aplicaron modelado de temas, análisis de sentimiento, primero, introdujeron un conjunto de Latent Dirichlet Allocation (E-LDA) modelos de temas para identificar automáticamente las características clave (temas), luego determinó la opinión de los estudiantes asociada con cada tema usando la herramienta de análisis de sentimient VADER, y [246] implementaron un sistema que recupera los datos de las redes sociales y otorga una calificación a una institución mediante análisis de sentimiento, aplicaron POS y la biblioteca StanfordCoreNLP.

Algunos autores aplicaron el método de deep learning en investigaciones de análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje, [262] aplicaron 8 clasificadores: SVM, MLP, DT, K-star, Bayes Net, Simple Logistics, Multi-class Classifier y RF, los métodos SVM y MLP, obtuvieron mejor rendimiento en comparación con los otros clasificadores; [218] integraron un módulo de Opinion Mining (OM), para detectar la polaridad de opinión de los estudiantes en relación con los ejercicios que resuelven en un entorno de aprendizaje inteligente, así como la detección de emociones, encontraron buenos resultados combinando Convolutional Neural Network (CNN) y Long Short-Term Memory (LSTM); [144] analizaron la revisión del corpus MOOC, con el uso de aprendizaje automático (K-NN, SVM, LR, NB, RF), ensemble learning (adaBoost, bagging, random subspace, voting, stacking), y deep learning (CNN, RNN, LSTM, Gated Recurrent Units (GRU), y Attention Mechanism (AM)), llegaron a la conclusión de que las arquitecturas basadas en el aprendizaje profundo superan a los otros métodos para la tarea de análisis de sentimiento en la minería de datos educativos; y [265] realizaron un sistema interactivo de múltiples agentes en el que los agentes modelan implícitamente a otros agentes, con un enfoque basado en la semiótica hacia el análisis de sentimiento, comparan los resultados con deep learning y otra técnicas de línea base, y proponen la semiótica como una alternativa a las dicotomías dominantes-basado en reglas y basados en datos dentro de la inteligencia artificial.

En otros experimentos con el mismo conjunto de datos aplicado de [253], [219] usaron N-Grams, Dependency Relation (DEP) y POS, con clasificadores tradicionales NB, ME y clasificadores de aprendizaje profundo LSTM, Bi-Directional Long Short-Term Memory (Bi-LSTM), llegaron a la conclusión que Bi-LSTM supero a los otros algoritmos; [263] analizaron el rendimiento de los modelos según la longitud de la oración, el modelo LSTM tuvo buen rendimiento en oraciones cortas, y el modelo Dependency Tree-LSTM con un clasificador de SVM fue mejor en oraciones medias y largas; y [264] usaron LSTM ATT (attention layer), multi-head

attention, las secuencias de entrada de oraciones se procesan en paralelo a través de la capa de atención de múltiples cabezas con incrustaciones de Glove y Cove y se prueban con diferentes tasas de abandono para aumentar la precisión.

Para análisis de sentimiento basado en aspecto, [123] propone un sistema de minería de opiniones basado en aspectos supervisados, el modelo LSTM de dos capas, en el que la primera capa predice los aspectos descritos dentro de la retroalimentación y luego especifica la orientación (positiva, negativa y neutral) de esos aspectos predichos; y [126] identificaron automáticamente la polaridad de opinión expresada hacia un aspecto dado relacionado con el MOOC mediante CNN y LSTM.

Mediante recursos léxico se efectuaron investigaciones de análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje, [273] formaron un corpus Wikipedia basada en el modelo Word2Vec y el software de Gensim para formar palabras chinas y encontrar similitud semántica; [271] utilizaron NLP y NRC Emotion Lexicon para clasificar los sentimientos y las emociones; [272] empleó métodos de aprendizaje automático junto con léxicos de sentimientos, encontró que el modelo de mejor rendimiento se logró utilizando TF-IDF y léxico de sentimiento específico del dominio; [245] examinaron sentimientos expresados por los estudiantes sobre la relación entre el uso de teléfonos inteligentes y el rendimiento académico, mediante recursos léxicos: Harvard general inquirer Bing Liu's Opinion Lexicon, MPQA léxico de subjetividad y AFINN léxico de afectividad, los resultados mostraron que si existe relación entre los factores; [220] usaron KNIME y SentiStrength para realizar análisis de sentimiento, demostraron como resultado que las técnicas de visualización de nubes de palabras pueden ayudar a obtener una visión del desempeño de un docente que normalmente no se observa a través de puntajes basados en Likert; y [244] aplicaron SentiStrength para análisis de sentimiento de comentarios en línea en tiempo real de estudiantes en un entorno de conferencias.

Con las herramientas de análisis de sentimiento, se realizaron investigaciones de análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje, [274] usaron la herramienta CourseObservatory para el análisis de sentimiento a los comentarios realizados por los alumnos; [275] aplicaron la herramienta VADER para todas las respuestas de texto, calificando cada respuesta con una puntuación compuesta en una escala entre +1 (sentimiento positivo) y -1 (sentimiento negativo); [276] utilizaron para clasificar la polaridad de sentimiento cuatro API de servicios en línea que ofrecen análisis de sentimiento: Amazon Comprehend, Google Natural Language, IBM Watson Natural Language Understanding y Microsoft Text

Analytics, todas las API generaron más valores positivos que negativos, un hecho que también puede explicarse por el ambiente académico; [277] aplicó la herramienta VADER para SA, realizaron prueba en diez publicaciones representativas que contenían texto normal con código de programación incrustado y se descubrió que el sentimiento positivo y negativo del texto normal no se vio afectado por el código incrustado; [255] desarrollaron Content Analyser System for edX MOOCs (edX-CAS) con técnicas de NPL para analizar los contenidos de los cursos en línea y las contribuciones de sus alumnos para mejorar el material didáctico y los procesos de enseñanza-aprendizaje de estos cursos; y para evaluar la institución; y [251] utilizó la herramienta TalkWalker y wikificación para análisis de sentimiento y así orientar la elección de una institución.

2.2.3. Técnicas, métodos y algoritmos utilizados en análisis de sentimiento de texto educativo (RQ3)

En esta sección, se analizó la tendencia de los investigadores en utilizar distintas técnicas, métodos y algoritmos utilizados en análisis de sentimiento de texto educativo, en base a los 57 estudios seleccionados en el dominio de la computación.

Los investigadores para realizar tareas de análisis de sentimiento aplicaron con mayor frecuencia: preprocesamiento (51 estudios), clasificación (50 estudios), extracción de características (26 estudios), detección de aspecto (11 estudios), detección de subjetividad (6 estudios) y detección de emoticonos (3 estudios); y con menor frecuencia aplicaron: proceso de anotación, proceso de anotación y aprendizaje en conjunto (Figura **10**). los campos relacionados a detección de aspecto y detección de emoticonos han atraído recientemente a los investigadores, ya que son campos de búsquedas emergentes.

Figura 10. Número de estudios con diferentes tareas de análisis de sentimiento

Para extraer y/o seleccionar características, los investigadores aplicaron con mayor frecuencia la técnica de POS (17 estudios), seguida N-Grams (13 estudios), TF-IDF (9 estudios), LDA y Word2Vec (7 estudios), GloVe (4 estudios), FastText (3 estudios), y BoW (2 estudios); y con menor frecuencia utilizaron: SVD, LSA, E-LDA, BTM, WF, TP, TF, TextBlob3, Doc2Vec, DEP y Cove (Figura **11**). En general, N-Grams y TF-IDF se combinaron con otras técnicas.

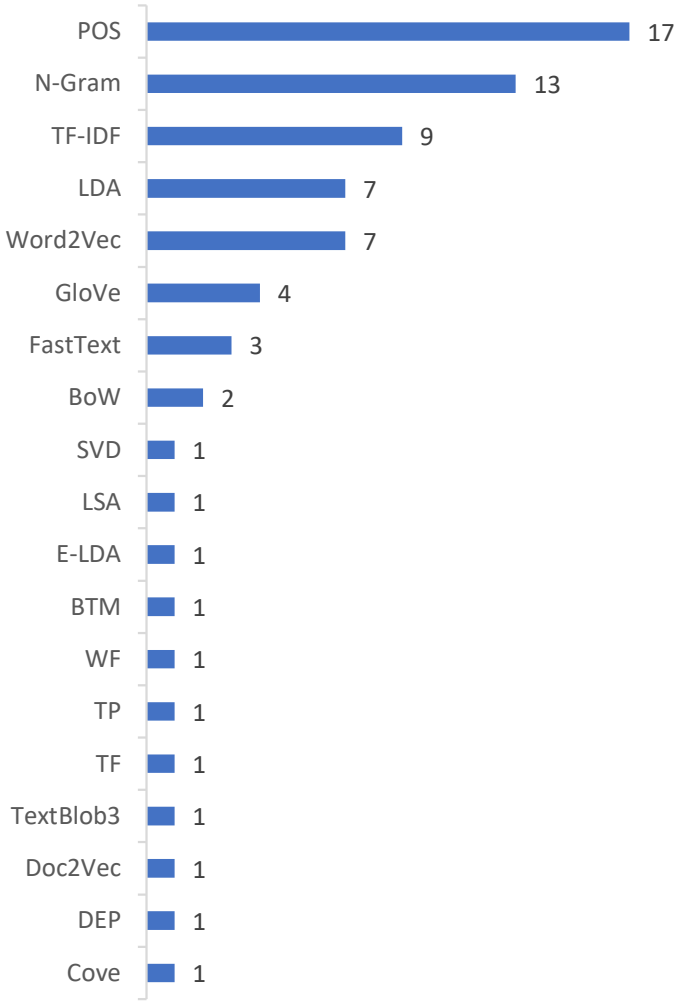


Figura 11. Número de estudios con diferentes técnicas de extracción de características

La Figura **12** ilustra los métodos de análisis de sentimiento utilizados. Los investigadores aplicaron entre 2014-2020 con mayor frecuencia el método de aprendizaje automático (35 estudios), seguido aprendizaje profundo (9 estudios) con evolución entre 2018-2020, léxico (7 estudios) y herramientas-SA (6 estudios). Lo que significa que en los últimos años, los investigadores se encaminan hacia el análisis general de textos y basado en aspectos.

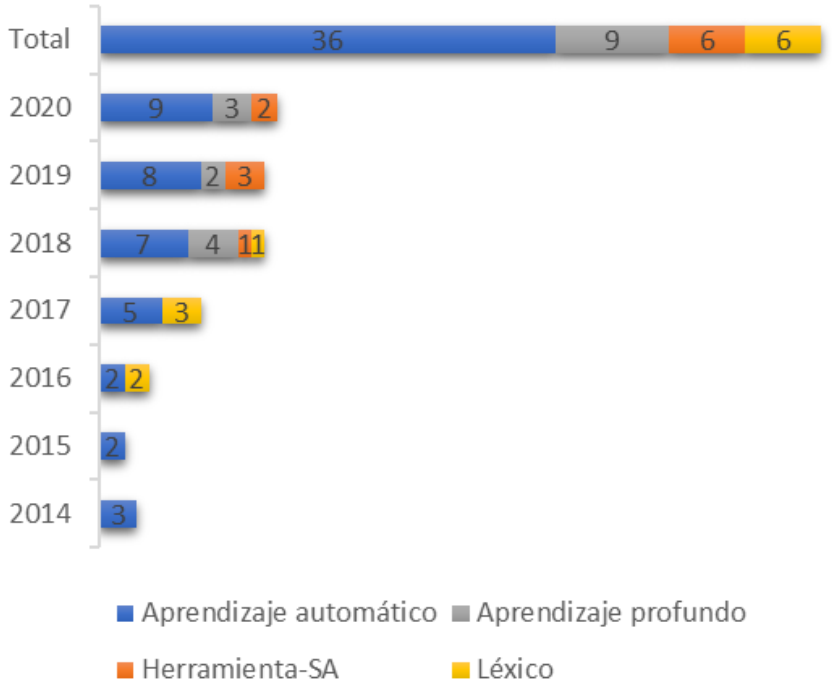


Figura 12. Número de estudios según el método de análisis de sentimiento por años

Se encontraron 42 algoritmos aplicados en los estudios (Figura **13**). Con el método de Machine Learning entre 2014-2020 los investigadores han utilizado con mayor frecuencia para clasificar sentimiento: SVM (24 estudios), seguido NB (21 estudios), K-NN y DT (7 estudios), RF (6 estudios), ME, LR y CNB (4 estudios), y NN (3 estudios); y con menor frecuencia: ANN, Attribute Selected Classifier, Bagging, BFTree, GBT, Hoeffding Tree (VFDT), K (IBK), K-Star, Logistic, MaxEnt, Multi-Class Classifier, One R, PART, Random Subspace, REPT, RL, SPPM, Zero R, AdaBoost, Bayes Net, ID3, J48, SimpleLogistic, Stacking y Voting.

Entre 2018-2020 con el método de Deep Learning los investigadores para clasificar sentimiento han aplicado con mayor énfasis: LSTM (9 estudios), seguido CNN (4 estudios) y MLP (3 estudios); y con menos énfasis: Bi-LSTM, DT-LSTM, GRU, RNN-AM y RNN.

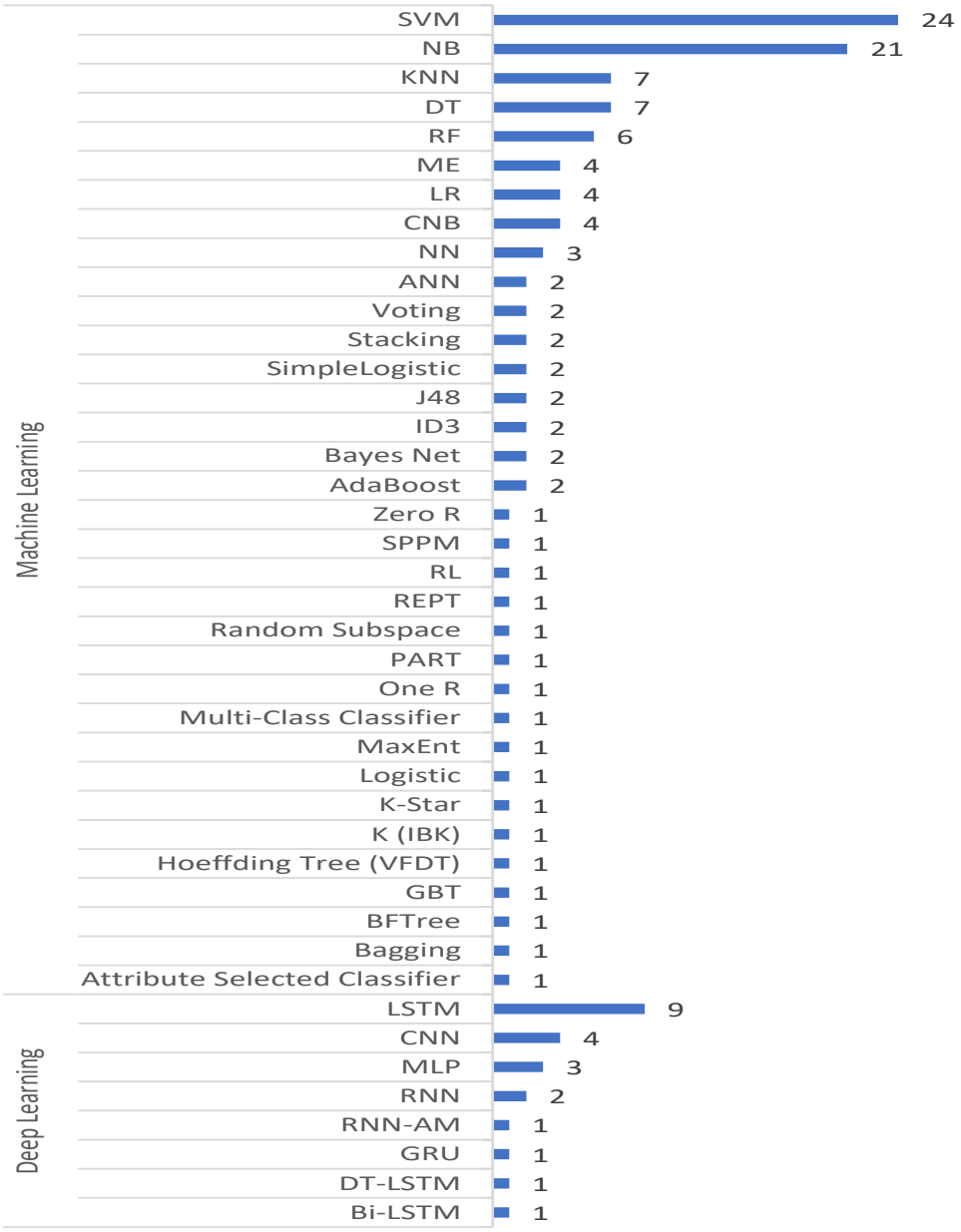


Figura 13. Número de estudios según los algoritmos de clasificación

Con el método de Lexicón aplicaron léxicos para obtener la palabra de sentimiento y la polaridad, con mayor énfasis usaron SentiWordNet (3 estudios), seguido WordNet y SentiStrength (2 estudios); y con menor énfasis: Teaching Senti, Senti-Lexicon, NRC Emotion Lexicon, MPQA, LIWC, Harvard General Inquirer y Bing Liu's Opinion (Figura **14**).

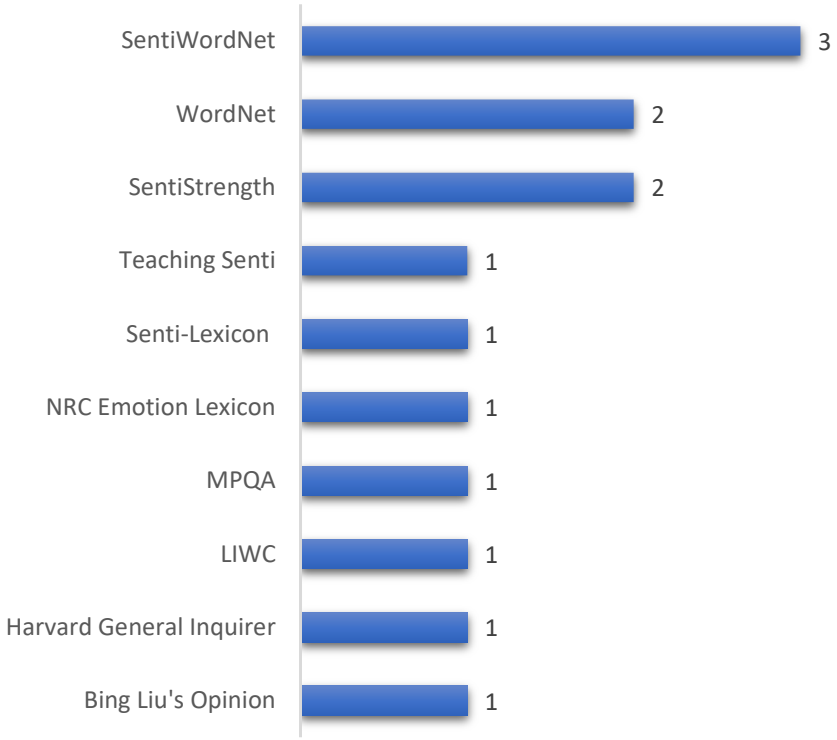


Figura 14. Número de estudios según el lexicón

Algunos estudios aplicaron herramienta para análisis de sentimiento como: TalkWalker, wikification, StandfordCoreNLP, CourseObservatory, VADER, Kanime Comprehend, Google Natural Language, IBM Watson Natural Language Understanding, Microsoft Text Analytics, y desarrollaron edX-CAS.

La Figura **15** ilustra que la tendencia de las investigaciones entre 2014-2020 ha sido clasificar pos/neg (22 estudios), lo que significa que el interés en la clasificación pos/neg continúa. En los últimos cuatro años ha sido hacer clasificación pos/neg/neu (31 estudios). Este incremento implica que el campo de polaridad de análisis de sentimiento está madurando.

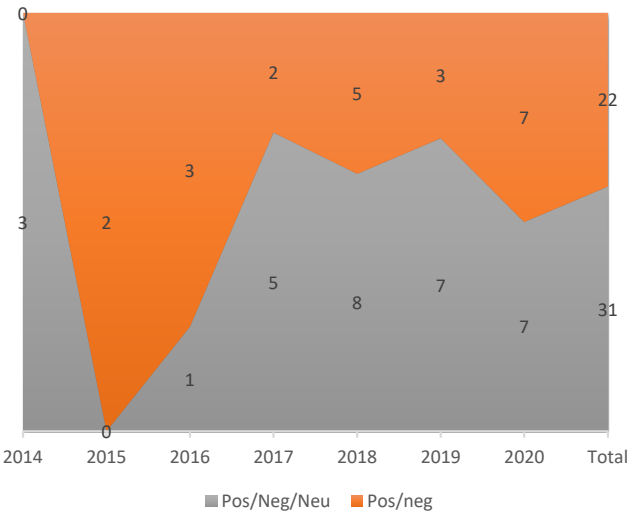


Figura 15. Número de estudios según la representación del sentimiento por año

La Figura **16** muestra que se involucraron 9 idiomas en los corpus. El idioma inglés es el más influyente (46 estudios), evolucionando en todos los años, seguido el vietnamita y español (3 estudios) con investigaciones entre 2018-2019, los 5 idiomas restantes se consideran menos influyentes ya que se encontró que se emplearon una vez en este estudio: myanmar, chína, serbio, tailandés y turco. El idioma inglés es el más utilizado por la disponibilidad de corpus, recursos léxicos y diccionarios.

Figura 16. Número de estudios según los idiomas por año

Los resultados revelaron:

- Una tendencia de investigación creciente en análisis de sentimiento de retroalimentación en procesos de enseñanza-aprendizaje en los tres últimos años, pero escasa en análisis de sentimiento de retroalimentación de evaluación entre pares en todos los años de estudios.
- Las investigaciones de retroalimentación entre pares se han realizado con mayor énfasis en la educación médica, en la enseñanza de idiomas, en la revisión de la escritura, en entornos: online, medios sociales, y sistemas.
- Los estudios de análisis de sentimiento se han enfocado en evaluación de: calidad de la enseñanza (contenido, metodología, actividades), aprendizaje (comentario, calificación numérica-comentario), instituciones (ranking), entre otras.

Estos resultados abren un abanico para que los investigadores realicen análisis de sentimiento de retroalimentación entre pares con diferentes métodos, técnicas NLP, algoritmos de clasificación, léxicos o herramientas, y un desafío en realizar análisis de sentimiento para otros idiomas.

La revisión sistemática de la literatura se presenta en el artículo: Sentiment Analysis Techniques for Peer Feedback: A Review", 2023 Ninth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 2023, pp. 1-8. https://ieeexplore.ieee.org/document/10122085

3. METODOLOGÍA DE LA PROPUESTA DE SOLUCIÓN

Esta investigación tiene como propósito diseñar un modelo de evaluación entre pares, que coadyuve a los docentes a mejorar sus procesos de enseñanza-aprendizaje mediante métodos de análisis de sentimiento.

La propuesta busca complementar y apoyar a los docentes en el proceso de enseñanza-aprendizaje, aligerar la carga de revisiones de trabajos abiertos y de control de calidad de evaluación de los pares, y al estudiante en corregir sus trabajos en base a las retroalimentaciones dadas por sus pares y mejorar el rendimiento en la segunda ronda.

3.1. Solución propuesta y esquema metodológico

Para llevar a cabo la propuesta, se realizó una revisión sistemática de la literatura descrita en la subsección 2.2, y que está publicada en [279], se reconceptualizaron las teorías y se efectúa el objetivo 2, elaborando un procedimiento para docentes y estudiantes con los estudios hallados en el dominio de educación [4]–[6], [8], [10], [11], [217], [221], [222], [227]–[236], [238], [239], y a la topología de topping [39], escenarios que ayudan a llevar a cabo la práctica de evaluación entre pares y la recolección de datos para formar el conjunto de datos.

Luego se estableció el procedimiento para que se ejecute análisis de sentimiento de retroalimentación textual mediante técnicas computacionales en base a los estudios encontrados en el dominio de computación en el que se fusiona la evaluación entre pares basada en la retroalimentación y análisis de sentimiento con la inteligencia artificial [2], [16]–[19], [25], [31], [107], [123]–[126], [134], [136], [139], [143], [144], [150], [152], [216], [218]–[220], [244]–[273].

A continuación, se detalla:

Procedimiento básico de evaluación entre pares basado en análisis de sentimiento

Docente

- Diseña escenario de evaluación entre pares, plantea actividad en un entorno de aprendizaje y la forma de recolección de datos [39], [217], [227], [237], [238].
- Establece actividad individual o en equipo de n integrantes [217], [222], [237], [238].

- Elabora artefacto de evaluación, orientado a los objetivos de la actividad o a los resultados del aprendizaje [222], y considera adquirir datos cualitativos y/o cuantitativos [217], [227], [237], [238].
- Asigna n revisiones de forma anónima o no anónima [217], [227], [238].
- Brinda andamiaje o instrucciones [8], [227].
- Realiza el esquema de codificación a la retroalimentación textual [17], [217], [273].

Estudiante

- Entrega tarea en un entorno de aprendizaje [217], [238].
- Proporciona puntuación numérica y/o retroalimentación textual a cada tarea asignada de manera objetiva y ética según el artefacto de evaluación [217], [227], [237], [238].

Análisis de sentimiento de la retroalimentación textual

- Preprocesamiento de la retroalimentación
  - Realiza limpieza y preparación del conjunto de datos: tokenización, filtrado de palabras vacías, manejo de negaciones, derivación y lematización [135].
- Extracción de características
  - Extrae características que serán las entradas de los algoritmos de clasificación después de convertirlas en los vectores de características representativas, utilizando diferentes técnicas como: POS, N-Grams, TF-IDF, LDA, Word2Vec, GloVe, FastText, BoW, SVD, LSA, E-LDA, BTM, WF, TP, TF, TextBlob3, Doc2Vec, DEP y Cove [152], [219], [246], [259], [260], [264].
- Entrenamiento de algoritmos
  - Entrena algoritmos de clasificación y obtiene un modelo predictivo, ya sea con el método de:
    - Aprendizaje automático: SVM, NB, K-NN, DT, RF, ME, LR, CNB, NN, ANN, Attribute Selected Classifier, Bagging, BFTree, GBT, Hoeffding Tree (VFDT), K (IBK), K-Star, Logistic, MaxEnt, Multi-Class Classifier, One R, PART, Random Subspace, REPT, RL, SPPM, Zero R, AdaBoost, Bayes Net, ID3, J48, SimpleLogistic, Stacking y Voting [2], [16]–[19], [25], [31], [107], [124], [125], [134], [136], [139], [143], [150], [152], [216], [246]–[250], [252]–[254], [256]–[261], [266]–[270].

- Aprendizaje profundo: LSTM, MLP, Bi-LSTM, DT, GRU y RNN [123], [126], [144], [218], [219], [262]–[265].

Una vez obtenidas las bases teóricas se desarrolló un modelo de evaluación entre pares que armonice varios tipos de evaluaciones (cualitativa, cuantitativa, inversa, en dos rondas y calibrada). A continuación, se detalla:

3.2. Modelo de evaluación entre pares

Para responder a la pregunta de investigación: **¿Como aplicar evaluación entre pares cuantitativa, cualitativa, inversa, en dos rondas y calibrada en escenarios de educación superior?,** se diseñó un modelo hibrido de evaluación entre pares. En primer lugar, para obtener la puntuación de evaluación de tarea del colectivo, los evaluadores mediante una rúbrica dividida en criterios evalúan aspectos específicos de una tarea, proporcionando por cada criterio una puntuación numérica (evaluación cuantitativa) y retroalimentación textual (evaluación cualitativa) fundamentando las razones por las que determinan tal puntaje numérico. En segundo lugar, los evaluados evalúan la calidad de la evaluación de la tarea (evaluación inversa). En tercer lugar, los grupos corrigen el trabajo en base a las retroalimentaciones dadas por los evaluadores en la primera ronda para mejorar el rendimiento (evaluación en dos rondas). Finalmente, para llevar a cabo un procedimiento fiable, se calibra la puntuación de evaluación de tarea (Figura **17**).

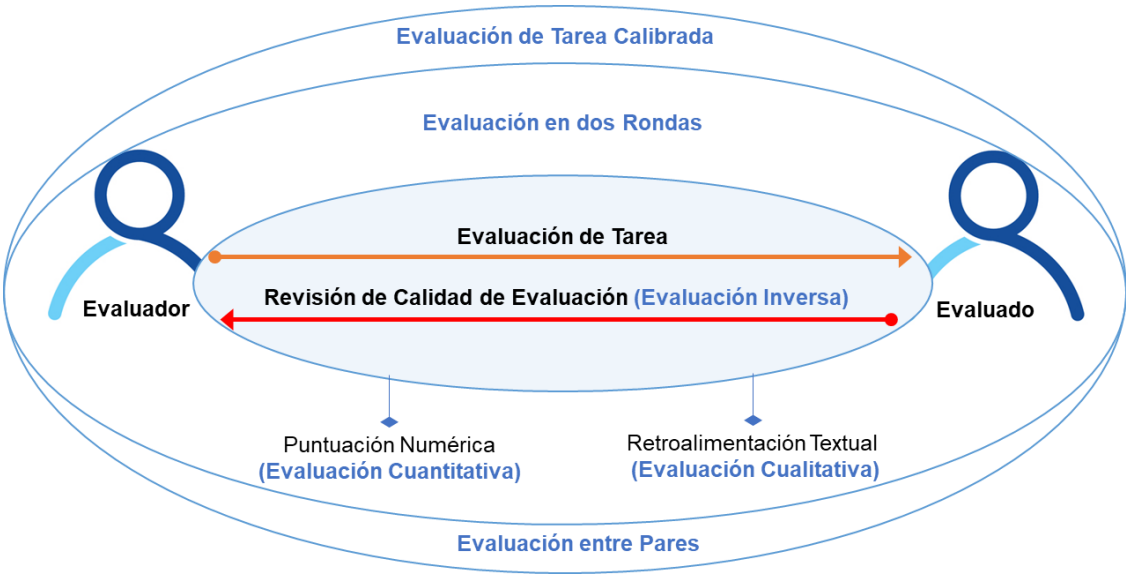


Figura 17. Modelo para evaluación entre pares cuantitativa, cualitativa, inversa, en dos rondas y calibrada

Los términos utilizados en el modelo propuesto se enlistan en la Tabla **9**.

Tabla 9. Términos y definiciones del modelo de evaluación entre pares

Términos	**Definiciones**
Roles	El estudiante desempeña dos roles, tanto el papel de evaluador como el de evaluado. El evaluador evalúa los trabajos abiertos. El evaluado evalúa la calidad de evaluación recibida de los pares evaluadores.
Colectivo	Es un grupo de revisión que consta de varios evaluadores, para que los resultados de la evaluación entre pares sean más confiables.
Tarea	El trabajo abierto lo realiza un grupo de estudiantes, con la intención que mejoren las habilidades de colaboración, reflexionen y aprendan de situaciones específicas de trabajo en equipo.
Evaluación entre pares	Se lo aplicó como una estrategia de evaluación para que los estudiantes mejoren el trabajo y el rendimiento en la segunda ronda sobre una temática especifica en un contexto social de interacción y colaboración.
Evaluación cuantitativa	El evaluador en base a una escala de Likert: 1 (Nada adecuado), 2 (Poco adecuado), 3 (Adecuado), 4 (Bastante adecuado), 5 (Totalmente adecuado), califica el trabajo de su compañero por cada criterio de la rúbrica.
Evaluación cualitativa	El evaluador dará retroalimentación al trabajo de su compañero, argumentando por cada criterio de la rúbrica, lo que detecta como correcto o incorrecto con sugerencias de posibles soluciones. con la finalidad que el estudiante revise el contenido de la temática, compare el trabajo que está revisando con su propio trabajo, desarrolle habilidades de reflexión, criticidad y asimile nuevos conocimientos. Se consideró los consejos de cómo escribir retroalimentaciones efectivas de [280] que se centran en: (1) seguir los criterios (alineación con la rúbrica), (2) ser explícito y exhaustivo (especificidad), (3) ofrecer sugerencias para mejorar, y (4) utilizando un lenguaje constructivo.
Evaluación inversa	El evaluado revisa la calidad de evaluación recibida, y por cada criterio de la rúbrica otorga puntuación numérica y comenta si está de acuerdo o no con la retroalimentación justificando las razones en base al contenido de la temática. Se consideró que los evaluados pueden estar en desacuerdo con la retroalimentación proporcionada sin tener la oportunidad de plantear sus inquietudes [281]. Una forma posible de mejorar el proceso y aumentar potencialmente su confiabilidad es cerrar el ciclo y ampliar el flujo de trabajo para permitir que el evaluado revise y brinden retroalimentación sobre la evaluación de sus pares evaluadores [280]. La refutación requiere una consideración de cada comentario y una justificación de por qué es aceptado, parcialmente aceptado o rechazado [85].
Evaluación en dos rondas	La evaluación se ejecuta en dos rondas. En la primera ronda los estudiantes evalúan los trabajos y hacen revisión de la calidad de evaluación recibida. En la segunda ronda los grupos corrigen el trabajo en base a las retroalimentaciones recibidas, el evaluador puede revisar la evaluación anterior y la aprobación/refutación del evaluado y vuelven hacer las mismas evaluaciones que la primera ronda.
Evaluación calibrada	Es la puntuación de evaluación de tarea calibrada. La calibración se realiza considerando el rendimiento y índice (rating) de confianza que obtuvo el evaluador del colectivo como punto de referencia para la decisión de dar bonificación o penalización. A la evaluación de tarea se añade o resta la proporción de varianza/desviación estándar de todas las puntuaciones dadas por los evaluadores por cada actividad, ya que cada actividad es sobre una temática diferente y en escenario diverso.

Para llevar a cabo el desarrollo del modelo, se elaboró 6 procesos (Figura **18**): a) escenario de evaluación entre pares; b) procedimiento de recopilación de datos de evaluación de tarea y evaluación de calidad de la evaluación; c) esquema metodológico de análisis de sentimiento de retroalimentación textual; d) esquema metodológico de correlación de puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación mediante la técnica de lógica difusa; e) procedimiento de cálculo de puntuación individual y del colectivo de evaluación de tarea y rating de confianza del evaluador; y f) procedimiento de calibración de puntuación de evaluación de tarea.

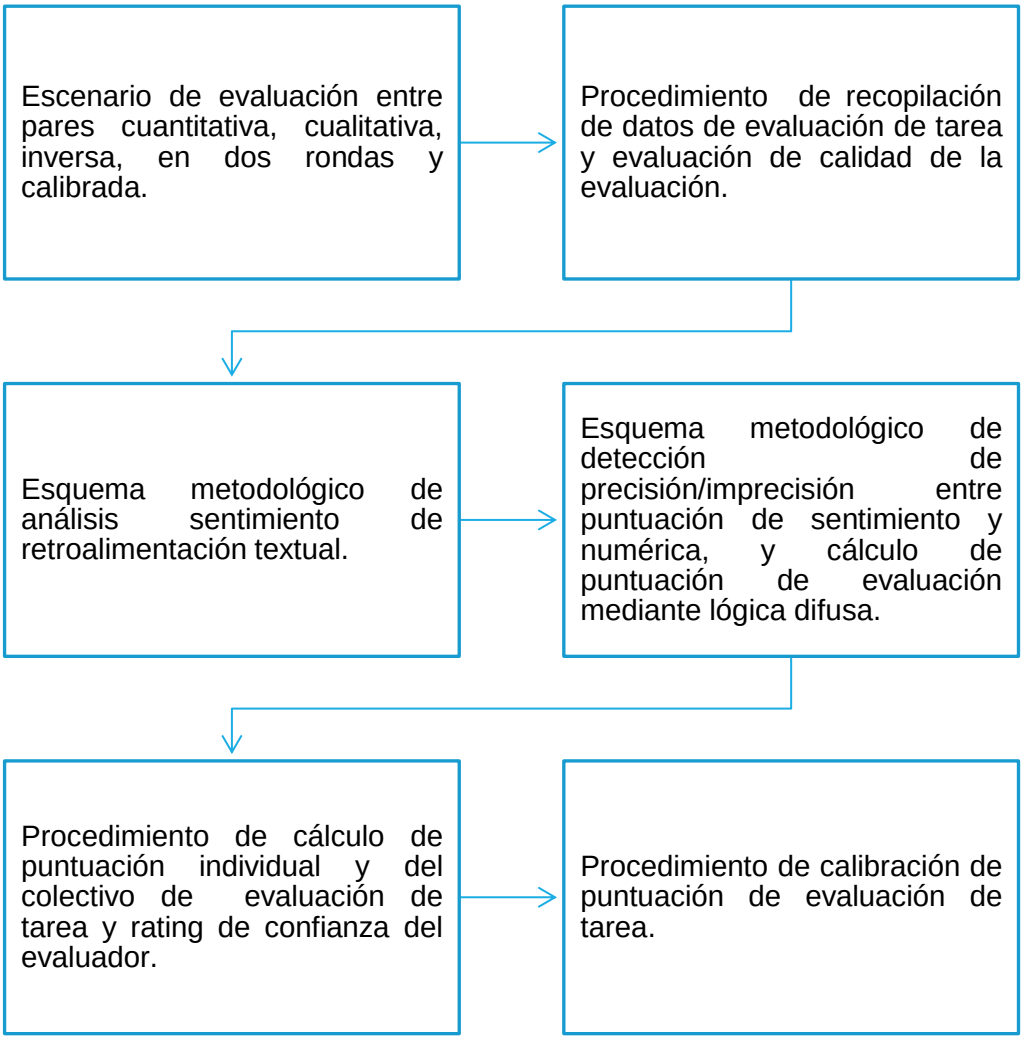


Figura 18. Procesos del modelo de evaluación entre pares

A continuación, se detallan cada uno de los procesos:

3.3. Escenario de evaluación entre pares

El escenario de evaluación entre pares se elaboró considerando la topología de topping [39] (Tabla **10**).

Tabla 10. Escenario de evaluación entre pares adaptada de [39]

Dimensión	**Rango de variación**
Área curricular/Asignatura	Asignaturas del área de informática.
Objetivos	Docente reduce la carga de revisión, y los estudiantes ven otras posibles soluciones y obtienen ganancias cognitivas.
Enfoque	Cuantitativa, cualitativa, inversa, en dos rondas y calibrada.
Producto/Salida	Trabajos abiertos.
Relación de la evaluación personal	Sustitucional.
Valor oficial	100% de la puntuación de evaluación del colectivo.
Direccionalidad	Mutua.
Privacidad	Anónima.
Contacto	La evaluación se realiza mediante un formulario en línea/prototipo de evaluación entre pares.
Año	El mismo año de estudio.
Habilidad	La evaluación está guiada por una rúbrica para obtener el máximo beneficio de las habilidades del evaluador.
Constelación del evaluador	Individual (1 o 2 o 3 tareas asignadas).
Constelación evaluada	Las tareas enviadas se realizaron de manera colaborativa.
Lugar	En clase presencial/en línea.
Tiempo	Hora de clase.
Requerimiento	Obligatorio para evaluadores/evaluados.
Bonificación	Se dará bonificación/penalización a la puntuación de evaluación en base al rendimiento y rating de confianza del evaluador.

3.4. Procedimiento de recopilación de datos de evaluación de tarea, evaluación inversa y en dos rondas

Se realizo el procedimiento para evaluar la tarea y evaluar la calidad de evaluación de los pares en dos rondas (Figura **19**). Donde el docente crea la actividad, especifica plazos. Cada grupo realiza la tarea y el líder del grupo sube la tarea. Un grupo de evaluadores designados revisa y evalúa la tarea. El evaluado evalúa la calidad de evaluación de sus pares. Los 9 pasos se resumen a continuación:

Paso 1. El docente configura grupos por uno o varios paralelos.

Paso 2. El docente diseña la rúbrica.

Paso 3. El docente configura la actividad con n asignaciones, y andamiaje por cada etapa (envió de tarea, evaluación de tarea, evaluación de calidad de evaluación).

Paso 4. El estudiante envía el trabajo abierto realizado de manera colaborativa.

Paso 5. Se asigna n tareas a cada estudiante.

Paso 6. El evaluador evalúa individualmente las tareas asignadas mediante una rúbrica con puntuación numérica (evaluación cuantitativa) y retroalimentación textual (evaluación cualitativa).

Paso 7. Se asigna las evaluaciones de las tareas a los evaluados.

Paso 8. El evaluado evalúa la retroalimentación recibida (evaluación inversa) con puntuación numérica (evaluación cuantitativa) y retroalimentación textual (evaluación cualitativa).

Paso 9. El evaluado corrige el trabajo en base a las retroalimentaciones dada por sus pares evaluadores, volviendo al paso 4 (ronda 2).

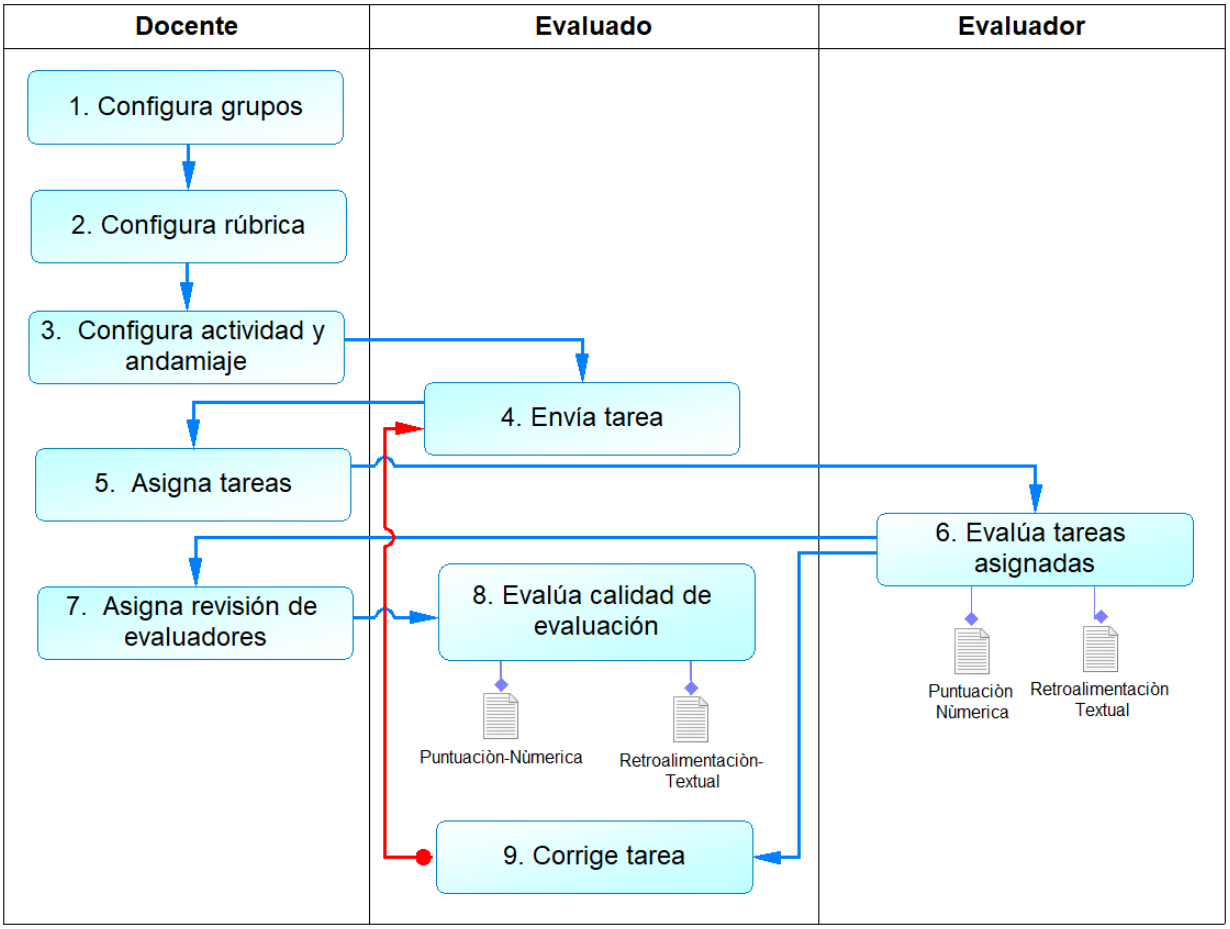


Figura 19. Procedimiento de evaluación de tarea, evaluación inversa y en dos rondas

3.5. Esquema metodológico de análisis de sentimiento de retroalimentación textual

El análisis de sentimiento es una técnica de minería de texto y una subárea de investigación de NLP ampliamente utilizada en educación [104]. Es la tarea de identificar cómo se expresan los sentimientos en los textos, y si las expresiones indican opiniones positivas (favorables) o negativas (desfavorables) hacia el tema [282]. Permite asignar una polaridad de sentimiento a un texto, en este caso a los textos generados por los estudiantes. La polaridad del sentimiento indica si el comentario tiene un sentimiento positivo, negativo o neutral [119]. Muchos estudios en la literatura utilizan los enfoques de aprendizaje automático para resolver tareas de análisis de sentimiento desde diferentes perspectivas. Dado que el rendimiento de un aprendiz de máquina

depende en gran medida de las opciones de representación de datos y extracción de características [104].

Se realizó un modelo de análisis de sentimiento utilizando el enfoque de aprendizaje automático, que recibe como entrada un corpus de texto en lenguaje natural etiquetado, y genera una puntuación de sentimiento que corresponde a una retroalimentación textual específica (Figura **20)**.

La Figura **20** muestra cinco representaciones gráficas, rectángulos para los pasos, almacenes de datos para los conjuntos de datos, archivos para las variables de entrada y salida, un icono para el modelo predictivo y la flecha de retroalimentación de color rojo para ilustrar que el entrenamiento del modelo puede regresar a ingeniería de características. Los pasos de recopilación, limpieza y etiquetado están orientados a los datos; los pasos de requerimientos, ingeniería de características, entrenamiento y evaluación están orientados al modelo. Las flechas de retroalimentación de color turquesa indican que la evaluación y el monitoreo del modelo pueden retroceder a cualquiera de los pasos anteriores.

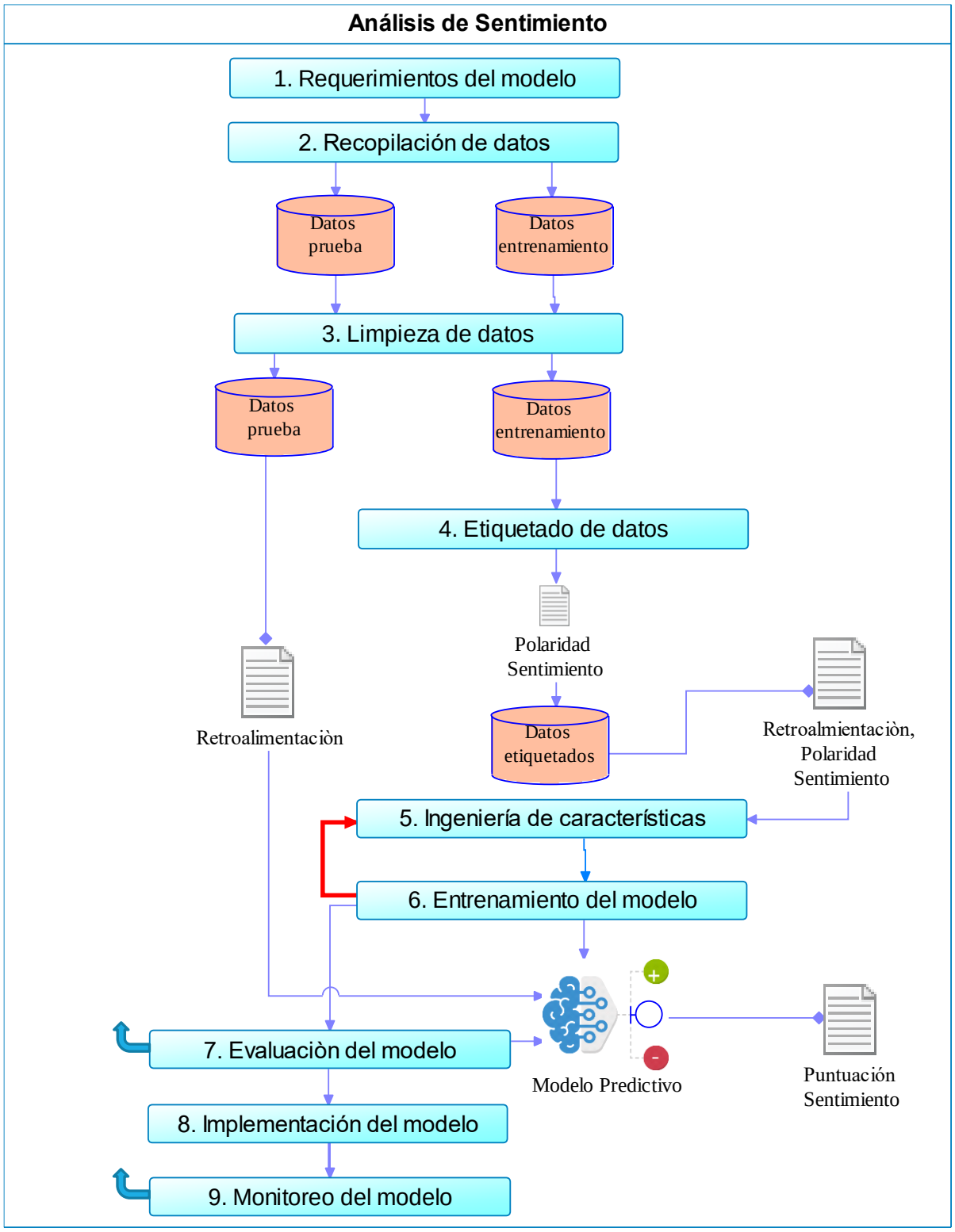


Figura 20. Esquema metodológico de análisis de sentimiento de retroalimentación textual

Los nueve pasos se resumen a continuación considerando el flujo de trabajo de [283].

Paso 1. Requerimientos del modelo

- En este paso se decide qué características son factibles de implementar con el aprendizaje automático y qué tipos de modelos son los más apropiados para el problema dado [283].
- Se aplicó el enfoque de aprendizaje supervisado, con modelos de aprendizaje automático y profundo.
- El sentimiento se evaluó a nivel de oración para probar hasta qué punto los algoritmos pueden evaluar correctamente la polaridad global asociada con la retroalimentación.

Paso 2. Recopilación de datos

- En este paso se buscan e integran conjuntos de datos disponibles o se recopilan datos propios. Se puede entrenar un modelo parcial utilizando conjuntos de datos genéricos y luego usar transferencia de aprendizaje con datos más especializados para entrenar un modelo más específico [283].
- Se recopiló datos de evaluación entre pares en el idioma español, considerando el escenario descrito en la subsección 3.3 y el procedimiento de recopilación de datos de evaluación entre pares descrito en la subsección 3.4.

Paso 3. Limpieza de datos

- En este paso se elimina registros inexactos o ruidosos del conjunto de datos [283].
- Se descartó registros vacíos de puntuación numérica y retroalimentación textual.

Paso 4. Etiquetado de datos

- En este paso se asigna etiquetas reales a cada registro. La mayoría de las técnicas de aprendizaje supervisado requieren etiquetas para poder inducir un modelo. Las etiquetas pueden ser proporcionadas por los propios expertos en el dominio o por trabajadores en plataformas en línea de colaboración abierta [283].
- En el etiquetado se consideró el esquema de análisis de contenido jerárquico [284], y el esquema de anotación de polaridad: (1) el etiquetado a nivel de oración explicita [132] e

implícita [132], [133]; (3) distinción entre palabras objetivas y subjetivas, no solo las palabras subjetivas, también las palabras objetivas pueden tener una connotación negativa o positiva [285], [286]; (4) consideración del contexto educativo y cómo los estudiantes se expresan en las retroalimentaciones. Se elaboraron las siguientes reglas para etiquetar los datos:

- Etiquetar cada retroalimentación en tres clases: positivo (+ 1), negativo (-1), neutral (0).
- Cuando la retroalimentación contiene palabras como: “bien”, “correcto”, “claro”, “me gustó”, y “adecuado”, entre otras, fueron etiquetados con una clase positiva (+ 1).
- Cuando la retroalimentación contiene palabras como: “no”, “falta”, “mal”, “incompleto”, “incorrecto”, “escaso”, “confuso”, “inadecuado”, y “errores”, entre otros, fueron etiquetados con una clase negativa (-1).
- Cuando la retroalimentación refleja opiniones mixtas igualmente positivas y negativas, se etiquetaba por la relevancia del aspecto o característica o se etiquetaba con una clase neutral (0).
- Cuando la retroalimentación refleja una combinación de opiniones positivas y negativas, se etiquetaba con la clase mayoritaria.

Paso 5. Ingeniería de características

- En este paso se realiza todas las actividades para extraer y seleccionar características informativas para los modelos de aprendizaje automático [283].
- Para abordar la tarea de preprocesamiento y extracción de características se ha aplicado diferentes técnicas tomadas de las áreas de minería de texto [140] y NLP [112], [113]. El preprocesamiento de texto no estructurado es una de las tareas más laboriosas e importantes en la construcción de un modelo de minería de texto. Para llevar a cabo una adecuada selección de rasgos se requiere realizar un proceso previo de transformación del texto [135]. Las características en el contexto de la minería de opiniones son las palabras, términos o frases que expresan fuertemente la opinión como positiva o negativa. Esto significa que tienen un mayor impacto en la orientación del texto que otras palabras en el mismo texto. Hay varias formas de evaluar la importancia de cada característica asignando un cierto peso en el texto [135].

- En el preprocesamiento (Figura **22**), se consideró realizar:
  - Normalización
    - Corrección de errores ortográficos.
    - Transformar todos los textos a minúsculas, ya que reduce el ruido en los datos, lo que generará resultados más precisos. Por ejemplo, la palabra "DIAGRAMA" y "Diagrama" se convertirá en "diagrama", para que el algoritmo pueda analizar y categorizar correctamente los datos.
    - Eliminar todos los caracteres especiales, símbolos (p. ej., puntos, punto y coma, comas, alfanuméricos, espacios en blanco y signos de interrogación).
    - Conversión de letras especiales. Las vocales con acentos y caracteres especiales, como “ñ”, “á”, “e” (parte de la gramática española) tomarán la forma de ‘n’, ‘a’, ‘e’, respectivamente.
  - Tokenización
    - Se usa para dividir una oración en palabras, frases, símbolos u otros tokens significativos mediante la eliminación de los signos de puntuación [282].
    - Las oraciones del conjunto de datos se dividirán en palabras.
  - Eliminación de Stop-Words
    - Las Stop-Words como artículos, preposiciones, conjunciones, entre otras, que no contribuyen al análisis se eliminan durante el paso de preprocesamiento [282].
    - Se creó una lista propia de Stop-Words en español en un archivo de texto con el nombre Stopwords.txt, separadas por comas, sin espacio entre las palabras, de acuerdo con el análisis del conjunto de datos. No contiene palabras como: "no", "falta", "malo", "nadie", "nunca", "sin", "nada", "ninguno", "bueno", "me gusta", entre otros, ya que son términos utilizados por los estudiantes para dar sus opiniones (Figura **21**).

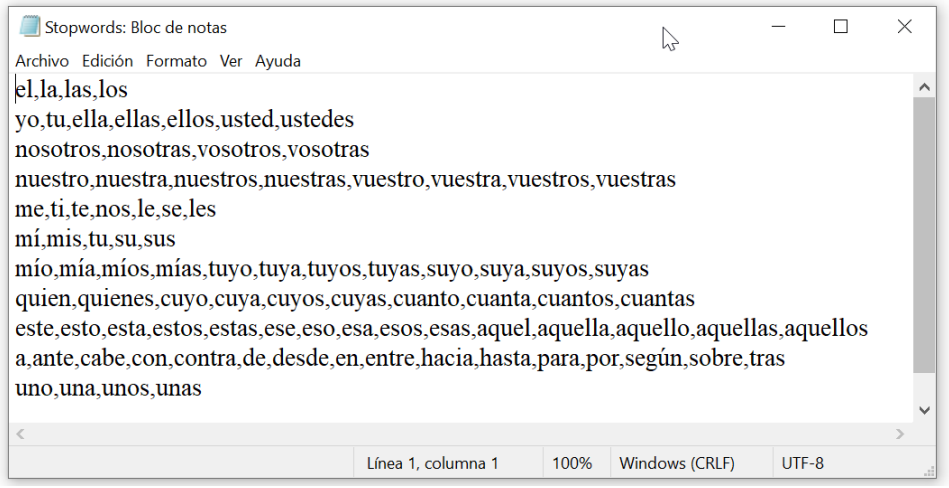


Figura 21. Ejemplo de Stop-Words en idioma español

- Partiendo de la lista de términos candidatos obtenidos en la tarea de preprocesamiento, se aplicarán técnicas de minería de texto del estado del arte, con el propósito de construir un vocabulario adecuado que consiga aumentar el rendimiento de la tarea de clasificación de sentimiento. Se han considerado distintas parametrizaciones, para determinar cuál es la mejor combinación de técnicas que hacen aumentar la eficacia del clasificador: a) Bag of Words, que transforma cada palabra en un número de forma que la entrada del algoritmo de clasificación es un vector de números, en el que cada posición es una palabra del texto a clasificar, para realizar esta transformación existen dos posibles formas: transformar en un vector de ocurrencias o transformar en un vector de frecuencias; b) TF-IDF que es un algoritmo para crear representaciones vectoriales de las palabras, aumenta proporcionalmente al número de veces que aparece una palabra en el documento, oración, frase, pero se compensa con la frecuencia de la palabra en el corpus, lo que ayuda a controlar el hecho de que algunas palabras son generalmente más comunes que otras; c) N-Grams que es una secuencia de N palabras, letras o sílabas; d) Word2Vec, que es un método predictivo para representar las palabras como vectores cortos y denso; e) Glove, que un algoritmo de aprendizaje no supervisado que obtiene representaciones de palabras en vectores a través de estadísticas de co-ocurrencia [147], [158].

- Para la extracción de características (Figura **22**), se prueba: Bag of Words (ver Subsección 4.1.1), combinaciones de N-Grams + TF-IDF con o sin Stop-Words, y Word Embeddings (ver Subsección 4.1.2), y Word2Vec y Glove pre-entrenado (ver Subsección 4.1.3).

Paso 6. Entrenamiento del modelo

- En este paso los modelos elegidos se entrenan y se ajustan en los datos limpios recopilados y sus respectivas etiquetas [283].
- Habiendo obtenido el formato legible (vector de palabras) por computadora en el paso 5, el algoritmo de aprendizaje automático clasifica el sentimiento. Los algoritmos aprenden automáticamente a identificar patrones utilizando un conjunto de entrenamiento. Esos patrones se pueden usar para la clasificación de texto con un conjunto de validación [147], [287].
- Se consideró varios algoritmos de aprendizaje automático del estado del arte para la clasificación del sentimiento (Figura **22**): SVM, NB y K-NN (IBk) (ver Subsección 4.1.1), MNB, SVM, LR, RF, DT y VE (ver Subsección 4.1.2), LSTM/Bi-LSTM (ver Subsección 4.1.2 y 4.1.3). Estos pasos se repitieron con todos los modelos, volviendo a la ingeniería de características.
- Para evaluar el rendimiento del modelo se consideró el método retención (Hold-Out), en este método, los datos se dividen en dos conjuntos separados: datos de entrenamiento y datos de prueba [288][289]. La proporción entre los datos de entrenamiento y los datos de prueba no es vinculante, pero para garantizar que la variante en el modelo no sea demasiado amplia, generalmente 2/3 de los datos se usan como datos de entrenamiento y el otro 1/3 como datos de prueba [289]. El método de validación cruzada (K-Fold), es un método para estimar el error de predicción, los datos se dividen en k subconjuntos con casi el mismo tamaño; el modelo en la clasificación se entrena y prueba hasta k veces [289]. En cada repetición, uno de los subconjuntos se utilizará como datos de prueba y los otros k-1 subconjuntos de datos servirán como datos de entrenamiento [289].

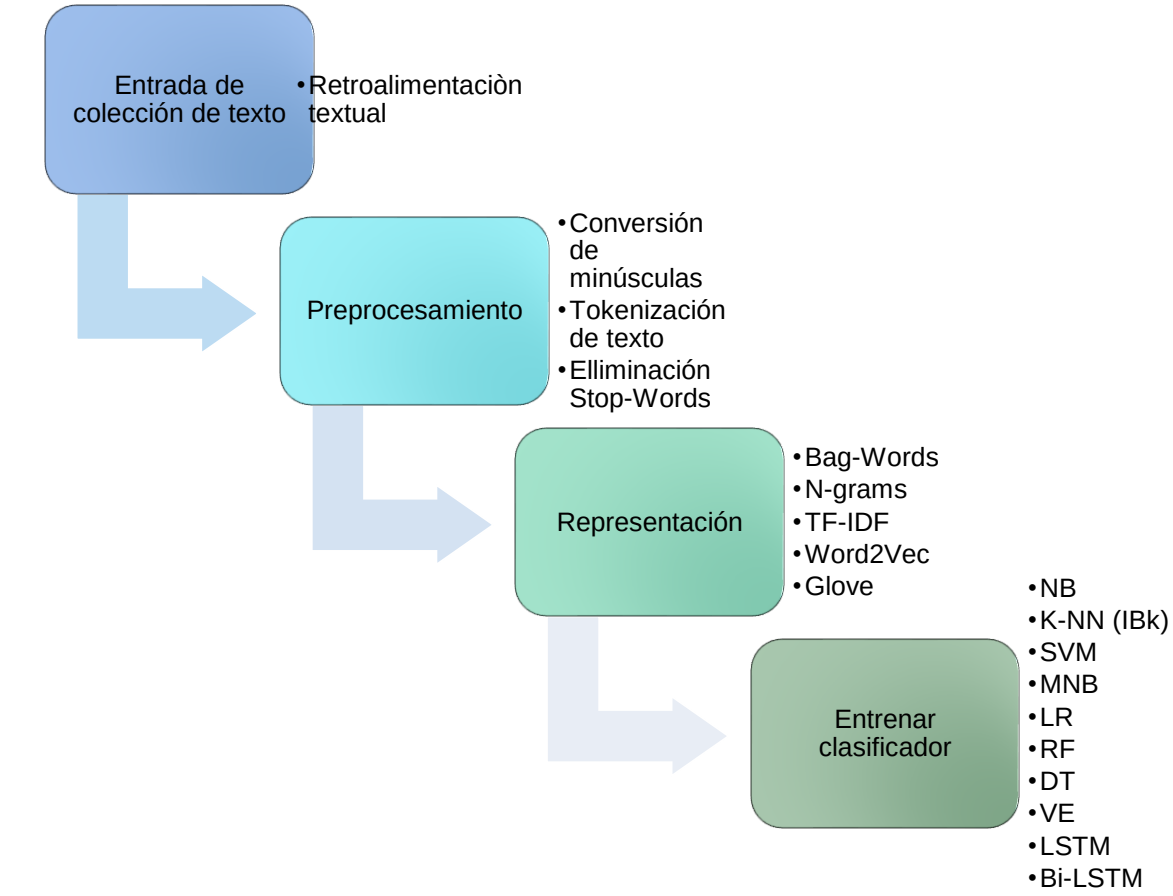


Figura 22. Tareas de ingeniería de características y entrenamiento del modelo

Paso 7. Evaluación del modelo

- En este paso se evalúa el modelo de salida en conjuntos de datos probados o de validación utilizando métricas predefinidas [283].

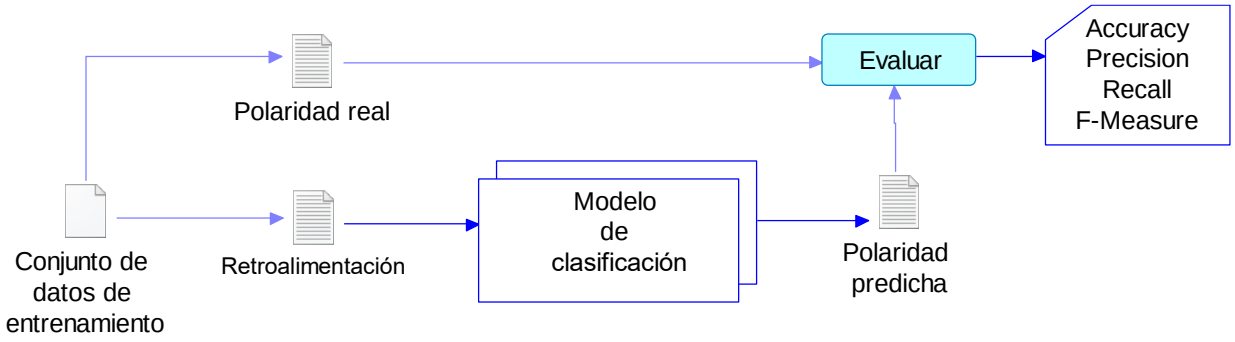


Figura 23. Esquema de evaluación del modelo mediante métricas

- Para evaluar cada modelo (Figura **23**) se aplicó las métricas de Accuracy, Precision, Recall, y F-Measure. Accuracy es el número total de instancias clasificadas correctamente (Ecuación 1). Precisión es la fracción de instancias predichas que son verdaderas para una clase *C.* Alta precisión significa que un algoritmo devuelve sustancialmente más resultados verdaderos que falsos para una clase *C* (Ecuación 2). Recall es la fracción de instancias verdaderas de una clase *C* que se predicen. Alta exhaustividad significa que un algoritmo devuelve la mayoría de los resultados verdaderos para una clase *C* (Ecuación 3). F-Measure es la media armónica de Recall y Precision (Ecuación 4) [119][147].

$$\boldsymbol{Accuracy}\ (\boldsymbol{C}) = \frac{\boldsymbol{tp}+\boldsymbol{tn}}{\boldsymbol{tp}+\boldsymbol{fp}+\boldsymbol{fn}+\boldsymbol{tn}} \qquad \text{Ecuación 1}$$

$$\boldsymbol{Precision}\ (\boldsymbol{C}) = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fp}} \frac{(\textbf{Instancias de C correctamente identificadas})}{(\textbf{Todas las instancias de C})} \qquad \text{Ecuación 2}$$

$$\boldsymbol{Recall}\ (\boldsymbol{C}) = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fn}} \frac{(\textbf{Instancias de C correctamente identificadas})}{(\textbf{Todas las instancias identificadas como C})} \qquad \text{Ecuación 3}$$

$$\boldsymbol{F-Measure}\ (\boldsymbol{C}) = \frac{\boldsymbol{2}+\boldsymbol{Precision}\ (\boldsymbol{C})+\boldsymbol{Recall}(\boldsymbol{C})}{\boldsymbol{Precision}\ (\boldsymbol{C})+\boldsymbol{Recall}\ (\boldsymbol{C})} \qquad \text{Ecuación 4}$$

- Las métricas basadas en etiquetas evalúan cada etiqueta por separado siendo posteriormente promediadas para obtener un único valor. En la Ecuación 5 se definen formalmente las métricas de evaluación F-Measure macro, donde q es el número de etiquetas, tp indica el número de patrones positivos clasificados correctamente, fp indica el número de falsos positivos, es decir, los patrones negativos que han sido correctamente asignados, fn indica el número de verdaderos negativos y por último, fn que indica el número de falsos negativos, es decir, los patrones que no han sido asignados [290].

$$F-Measure_{macro} = \frac{1}{q}\sum_{i=1}^{q} \frac{2*tp_i}{2*tp_i+fn_i+fp_i} \qquad \text{Ecuación 5}$$

- Se comparó el rendimiento de cada modelo con la métrica F-Measure macro-average en lugar de Precision, ya que es la mejor medida para describir el rendimiento del modelo cuando los datos no están equilibrados [146], [291]–[293].
- Con los modelos predictivos seleccionados, se generó la puntuación de sentimiento de cada una de las actividades de evaluación entre pares y se determinó:

- Medir la concordancia entre la puntuación de sentimiento que generó el modelo predictivo y la polaridad de sentimiento que proporcionó el anotador, utilizando los coeficientes de Kappa, Pearson y Spearman;
- Evaluar la portabilidad de los modelos predictivos entre asignaturas utilizando texto no etiquetado.

Paso 8. Implementación del modelo

- En este paso el modelo se implementa en los artefactos de destino [283].

Paso 9. Monitoreo del modelo

- En este paso se monitorea continuamente para detectar posibles errores durante la ejecución en el mundo real [283].

3.6. Esquema metodológico de detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación mediante lógica difusa

Se realizó el modelo utilizando el enfoque de Mamdani en lógica difusa, que recibe como entrada la Puntuación Numérica y de Sentimiento generada del modelo predictivo para detectar precisión o imprecisión entre ambas puntuaciones, y generar la Puntuación de Evaluación (Figura **24**).

La Figura **24** muestra cuatro representaciones gráficas, rectángulos para los pasos, almacenes de datos para los conjuntos de datos, archivos para las variables de entrada y salida, un icono para el modelo predictivo. Los pasos de fuzzificación, reglas y defuzzificación están orientados al cálculo.

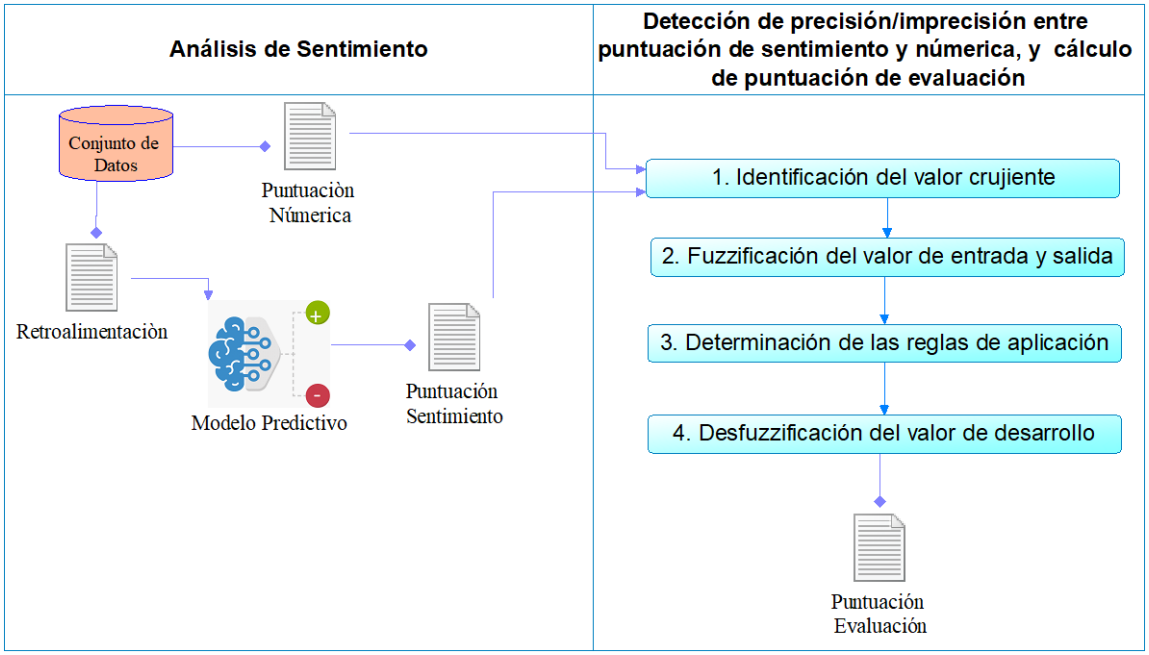


Figura 24. Esquema metodológico de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación mediante lógica difusa

Los cuatro pasos se resumen a continuación:

Paso 1. Identificación del valor nítido

- En este paso, se identifica las variables, que pueden tomar como valor palabras del lenguaje natural o números [30][31].
- Para obtener la puntuación de evaluación, se seleccionó las variables de entrada: Puntuación Numérica, y Puntuación Sentimiento.

Paso 2. Fuzzificación del valor de entrada y salida

- En este paso las variables de entrada se dividen en variables lingüísticas [22], [26]–[28], luego se forman las funciones de pertenencia, asignando el rango adecuado a las respectivas variables lingüísticas [30][31].
- Los valores nítidos de las variables de entrada (Puntuación Numérica, Puntuación Sentimiento), y salida (Puntuación Evaluación) se convierten en conjuntos difusos y se forma las funciones de pertenencia de cada variable.

Paso 3. Determinación de reglas de aplicación

- Las reglas determinan las funciones de pertenencia de entrada y salida que se utilizarán en el proceso de inferencia. Estas reglas son lingüísticas y se denominan reglas SI (antecedente) ENTONCES (consecuente) [28][30][31].
- Para determinar las funciones de pertenencia de entrada y salida a utilizar en el proceso de inferencia, se formuló reglas difusas SI antecedente (Puntuación Numérica, Puntuación Sentimiento) ENTONCES consecuente (Puntuación Evaluación) de tipo Mamdani con operador lógico Y de Zadeh (min) entre antecedentes.

Paso 4. Desfuzzificación del valor de desarrollo

- Después de completar el proceso de decisión difusa, el número difuso obtenido debe convertirse en un valor nítido [29]–[31]. Se utilizó el sistema de inferencia Mamdani, considerado por [294]. Para calcular la salida nítida (y) a partir de la entrada numérica nítida (X= x), Dada una base de reglas de sentencias en la forma de SI X es A_k ENTONCES Y es B_k, donde A_k conjunto difuso que aparece en el antecedente (Puntuación Numérica, Puntuación Sentimiento), y B_k conjunto difuso que aparece en el consecuente (Puntuación Evaluación). El grado de pertenencia de la entrada (x) en el conjunto difuso A es calculado como $\mu_{A_k}(x)$, y se activan las reglas correspondientes con grados de pertenencia. El conjunto difuso en el consecuente de cada regla se trunca al nivel del grado de pertenencia calculado previamente, formando el conjunto difuso de salida $\mu_{salida\ k\backslash x}$ por la ecuación:

  $$\boldsymbol{\mu_{output\ k\backslash x}(y) = min\left(\mu_{B_k}(y), \mu_{A_k}(x)\right)}$$ (Ecuación 6)

  Todos los conjuntos difusos truncados se agregaron para proporcionar un solo conjunto $\mu_{Mamdani\backslash x}$ que se puede definir mediante la función de pertenencia:

  $$\mu_{Mamdani\backslash x}(y) = \max_k \left[min\left(\mu_{B_k}(y), \mu_{A_k}(x)\right)\right]$$ (Ecuación 7)
- La salida nítida se calcula a partir de la defuzzificación del conjunto difuso. En cada experimento, se probó los métodos de Centroid, Bisector, SOM, MOM y LOM, considerados por [212][29].
- En este paso se obtuvo la puntuación de evaluación de la correlación de puntuación de sentimiento y numérica (Figura **25**).

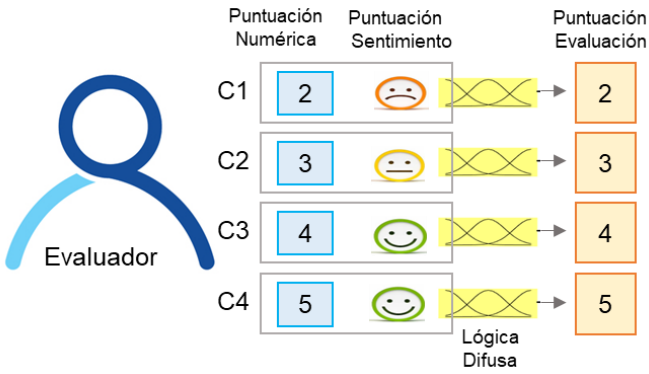


Figura 25. Ejemplo de puntuación de evaluación por cada criterio

3.7. Procedimiento de cálculo de puntuación individual y del colectivo de evaluación de tarea, y rating de confianza del evaluador

Para calcular la puntuación de evaluación de tarea y el rating de confianza del evaluador, se ejecutó los siguientes pasos (Figura **26**):

- **Cálculo de puntuación de evaluación de tarea**

  Paso 1. Cálculo de puntuación individual

  – Se calcula la media de todas las puntuaciones de evaluación de criterios que proporcionó cada evaluador.

  Paso 2. Cálculo de puntuación del colectivo

  – Se calcula la media/mediana del conjunto de puntuaciones individuales de los evaluadores pares por grupo evaluado.

- **Cálculo del índice (rating) de confianza del evaluador**

  Paso 1. Cálculo de puntuación individual

  – Se calcula la media de todas las puntuaciones de calidad de evaluación de criterios que proporcionó cada evaluado.

Paso 2. Cálculo de puntuación del colectivo

- Se calcula la media/mediana del conjunto de puntuaciones individuales de los evaluados por cada estudiante.

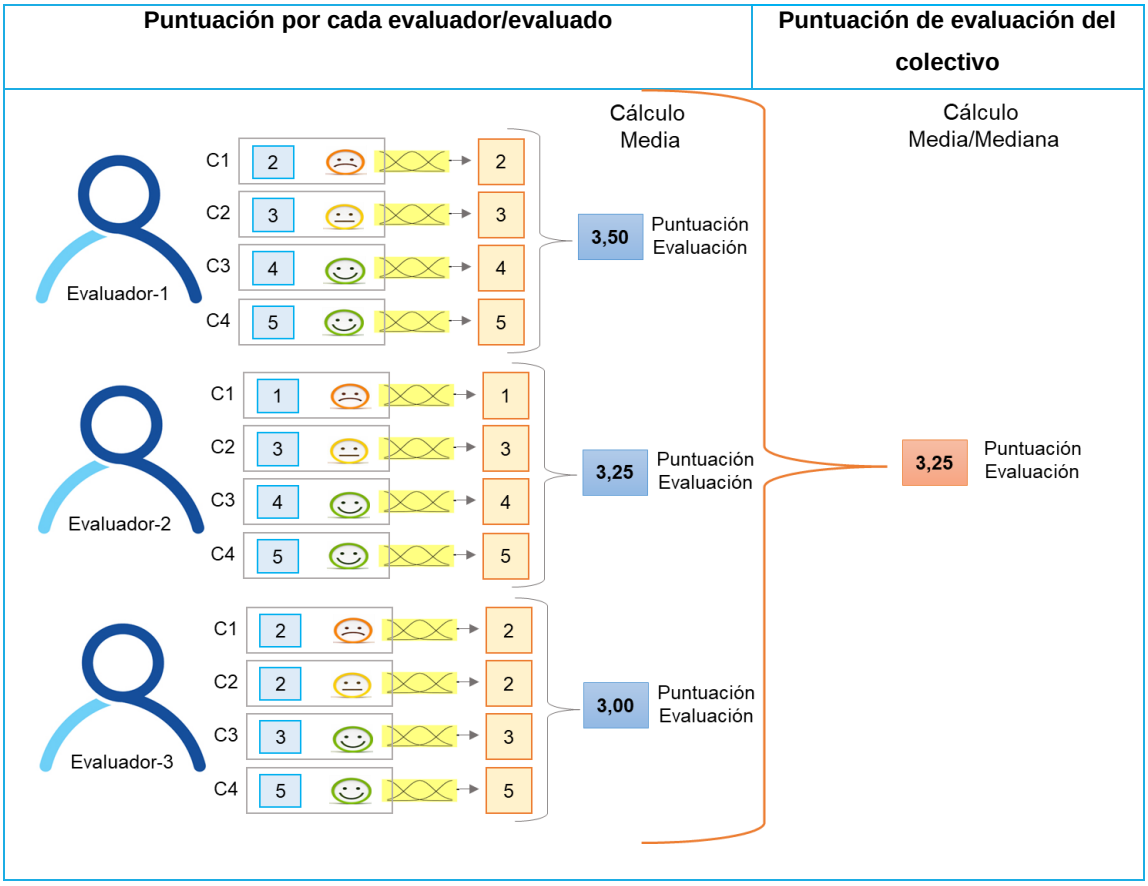


Figura 26. Ejemplo de cálculo de puntuación de evaluación del colectivo

3.8. Procedimiento de calibración de puntuación de evaluación de tarea

El principio del modelo de calibración es que pueda implementarse y ejecutarse fácilmente sin la intervención del docente, por tal razón se realizó un análisis estadístico para determinar la relación que existe entre puntuación dada con puntuación recibida del colectivo, y entre puntuación dada por los evaluadores con rating de confianza dada por los evaluados. Los resultados mostraron que no existe correlación sustancial, ni fuerte entre las variables (ver Análisis). Por tal razón se determinó como factores tanto la puntuación recibida (perfil cognitivo)

y rating de confianza del evaluador (perfil evaluador) para calibrar la puntuación de evaluación de tarea.

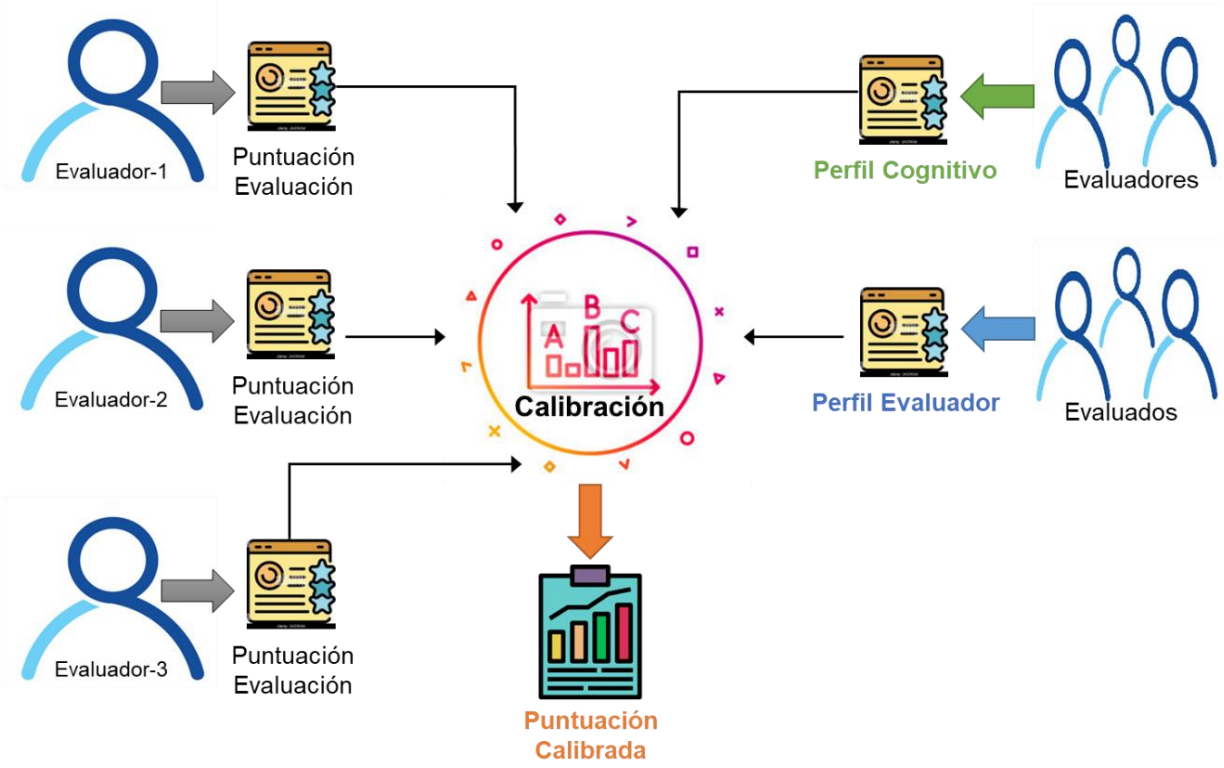


Figura 27. Esquema metodológico de calibración de puntuación de evaluación de tarea

Se realizó el modelo de calibración recibiendo como entrada la Puntuación Evaluación dada por cada evaluador, a la que se bonificará o penalizará en función del perfil cognitivo y de evaluador, por cada actividad, puesto que cada actividad es sobre una temática diferente y en escenario diverso (Figura **27**). A continuación, se detalla cada uno de los pasos:

Paso 1. Normalización de datos

- Se normaliza los datos a escala de 1.

Paso 2. Especificación del perfil del estudiante

- Con los factores: Puntuación Recibida y Rating Confianza se definió los perfiles del evaluador.
- Según las puntuaciones recibidas del colectivo, que representa el nivel de conocimiento del estudiante en esa actividad, se definió el perfil cognitivo del evaluador de la siguiente manera:

  - Si (0.8 < Puntuación Recibida < =1), entonces (Perfil Cognitivo es Alto)
  - Si (0.6 < Puntuación Recibida < =0.8), entonces (Perfil Cognitivo es Medio Alto)
  - Si (0.4 < Puntuación Recibida < =0.6), entonces (Perfil Cognitivo es Medio)
  - Si (0.2 < Puntuación Recibida < =0.4), entonces (Perfil Cognitivo es Medio Bajo)
  - Si (0.0 < Puntuación Recibida < =0.2), entonces (Perfil Cognitivo es Bajo)
- Según el rating de confianza del colectivo, que representa la capacidad del estudiante para evaluar el trabajo de sus compañeros, se definió el perfil del evaluador de la siguiente manera:
  - Si (0.8 < Rating Confianza < =1), entonces (Perfil Evaluador es Alto)
  - Si (0.6 < Rating Confianza < =0.8), entonces (Perfil Evaluador es Medio Alto)
  - Si (0.4 < Rating Confianza < =0.6), entonces (Perfil Evaluador es Medio)
  - Si (0.2 < Rating Confianza < =0.4), entonces (Perfil Evaluador es Medio Bajo)
  - Si (0.0 < Rating Confianza < =0.2), entonces (Perfil Evaluador es Bajo)

Paso 3. Cálculo de calibración de puntuación dada por cada evaluador

- Se calcula la varianza (var) y desviación estándar (std) de todas las puntuaciones dadas de los evaluadores por cada actividad, ya que cada evaluador califica en base al conocimiento que tenga sobre la temática de esa actividad. La varianza permitió obtener la variabilidad de las puntuaciones dadas de los evaluadores, cuanto se mueve cada puntuación de la media, y cuanto se aleja el comportamiento del evaluador de todos los evaluadores. Y la desviación estándar permitió obtener cuan dispersa están las puntuaciones dadas de todos los evaluadores en una determinada actividad. Un valor pequeño significa puntajes similares para diferentes evaluadores, que probablemente tuvieron el mismo conocimiento, y un valor grande significa puntajes diferentes, que probablemente no todos los evaluadores tuvieron el mismo conocimiento en esa temática.
- Se calibra la puntuación dada de cada evaluador, adicionando o restando la proporción que se obtiene de varianza multiplicada por desviación estándar, de acuerdo al perfil cognitivo y de evaluador (Tabla **11** y Figura **28**).

Tabla 11. Bonificación o penalización de acuerdo al perfil del evaluador

Perfil Cognitivo/Evaluador	Bonificación/Penalización	
Alto	var * std	Bonificación
Medio Alto	1/2 var * std	Bonificación
Medio		No se da bonificación, ni penalización
Medio Bajo	-1/2 var * std	Penalización
Bajo	-var * std	Penalización

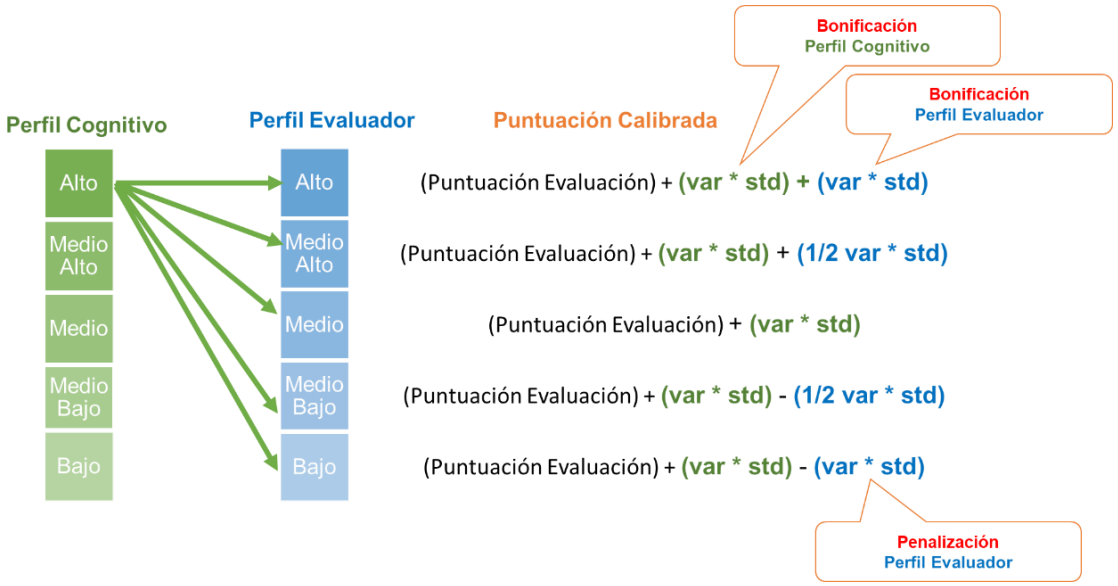


Figura 28. Ejemplo de cálculo de puntuación calibrada con perfil cognitivo “Alto” y perfil de evaluador (rating confianza) “Alto/Medio Alto/Medio/Medio Bajo/Bajo”

Paso 4. Cálculo de puntuación calibrada del colectivo

- Se recalcula la media/mediana del conjunto de puntuaciones individuales de los evaluadores pares por grupo evaluado.

4. EXPERIMENTACIÓN

En este capítulo se especifica el objetivo 3, detallando diferentes experimentos llevados a cabo para demostrar que la metodología propuesta es útil para obtener un modelo de evaluación entre pares con técnicas de computación blanda.

4.1. Ejecución de experimentos

Se realizaron cuatro experimentos (Figura **29**). En la primera iteración se realizó evaluación de diferentes algoritmos de aprendizaje automático del estado del arte con combinaciones (Bag of Words, TF-IDF, Stemmer, Stop-Words) para predecir la puntuación de sentimiento de la retroalimentación textual. En la segunda iteración se refina la primera iteración con análisis de distintos tipos de parametrizaciones (N-Grams, TFIDF, Stop-Words), y algoritmos de aprendizaje automático y profundo para lograr el mejor rendimiento predictivo en la tarea de clasificación de sentimiento, y se analizó qué método de defuzzificación en lógica difusa, resulta más apropiado para generar la puntuación de evaluación de la correlación entre puntuación de sentimiento y numérica. En la tercera iteración se refina la segunda iteración con enriquecimiento semántico de Word2Vec/Glove preentrenado y algoritmos de aprendizaje profundo, y se obtiene la puntuación de evaluación de tarea y el rating de confianza del evaluador en dos rondas. En la cuarta iteración se calibra la puntuación de evaluación de tarea considerando los perfiles del evaluador.

Otro punto importante de esta investigación es el prototipo de evaluación entre pares que permitió recopilar datos de evaluación de tarea, evaluación inversa con puntuación numérica y retroalimentación textual en dos rondas, además se implementó el modelo de análisis de sentimiento y lógica difusa.

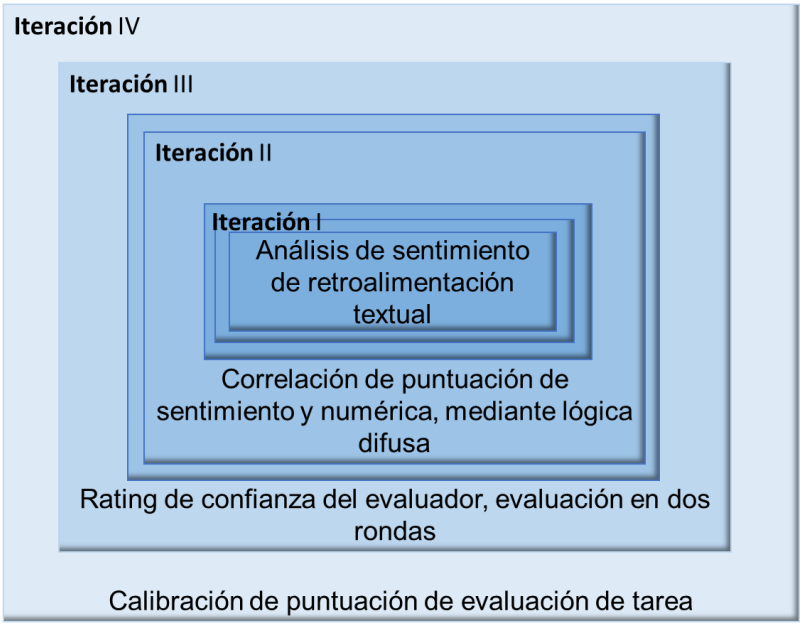


Figura 29. Configuración de los experimentos

A continuación, se detalla cada experimento con planteamientos, preguntas de investigación que se pretende responder, objetivos, resultados, conclusiones y futuro experimento.

4.1.1. Experimento I

En este experimento se consideró la evaluación cualitativa de la tarea en dos rondas (Figura **30**), del modelo de evaluación entre pares descrito en la subsección 3.2.

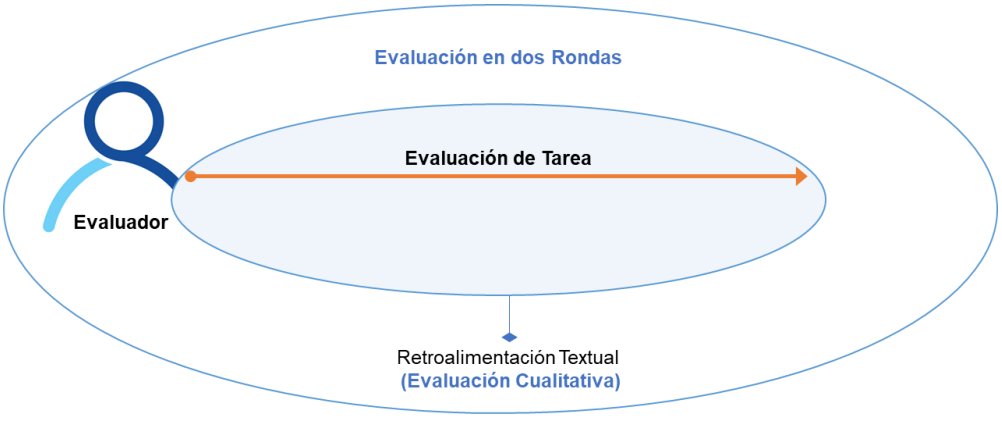


Figura 30. Evaluación entre pares cualitativa

Planteamientos, preguntas de investigación y objetivos

<table>
<tr><th colspan="2">Planteamientos</th></tr>
<tr><td colspan="2">Analizar diferentes tipos de rúbricas para la recolección de datos de evaluación entre pares.
Determinar qué algoritmo ofrece un mejor rendimiento en la tarea de clasificación de sentimiento de retroalimentación textual en español con la técnica de minería de texto Bag of Words, TF-IDF, Stemmer, Stop-Words.</td></tr>
<tr><th colspan="2">Preguntas de investigación</th></tr>
<tr><td>¿Como aplicar evaluación entre pares (cualitativa) en escenarios de educación superior?</td><td>¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?</td></tr>
<tr><th colspan="2">Objetivos</th></tr>
<tr><td colspan="2">Diseñar los artefactos para validar los métodos teóricos.
Construir un modelo de evaluación entre pares basado en análisis de sentimiento.
Evaluar los resultados de la precisión del modelo.</td></tr>
</table>

A continuación, se detalla el experimento:

Análisis de rúbricas

En primer lugar, para la recolección de datos de evaluación entre pares se analizó dos tipos de rúbricas:

Rúbrica analítica

Se diseñó rúbrica de tipo analítica en dos partes para evaluar tareas en el campo de la ingeniería de sistemas informáticos (Tabla **12**). La primera parte considera criterios, niveles de logro y descriptores de cada nivel. La segunda parte contiene criterios para que el estudiante evalúe la tarea con puntuación numérica (nivel de logro) y retroalimentación textual, y un andamiaje que variaría dependiendo de la tarea a evaluar.

Tabla 12. Rúbrica analítica

Criterios	Totalmente adecuado (5)	Bastante adecuado (4)	Adecuado (3)	Poco adecuado (2)	Nada adecuado (1)
Documento	La **carátula** consta de nombres: institución, facultad, carrera, materia, tema del trabajo, periodo académico, y el **enunciado** describe los requerimientos de lo que se desea obtener sin errores ortográficos.	La **carátula** consta de nombres: facultad, carrera, materia, tema del trabajo, y el **enunciado** describe los requerimientos de lo que se desea obtener con pocos errores ortográficos.	La **carátula** consta de nombres: carrera, materia, tema del trabajo, y el **enunciado** describe parcialmente los requerimientos de lo que se desea obtener con algunos errores ortográficos**.**	La **carátula** consta de nombres: materia, tema del trabajo, y el **enunciado** no describe los requerimientos de lo que se desea obtener con varios errores ortográficos**.**	La **carátula** consta del tema del trabajo y no contiene **enunciado**.
Estructura	La estructura se visualiza fácilmente porque contiene todos los **elementos,** la **sintaxis** es correcta, y está **ordenada.**	La estructura se visualiza fácilmente porque contiene la mayoría de los **elementos,** la **sintaxis** es correcta, y está **ordenada.**	La estructura se visualiza con dificultad porque contiene algunos **elementos,** la **sintaxis** es correcta, y no está **ordenada.**	La estructura no se visualiza fácilmente porque contiene pocos **elementos,** la **sintaxis es** correcta, y no está **ordenada.**	La estructura no se visualiza fácilmente porque contiene pocos **elementos,** la **sintaxis** no es correcta, y no está **ordenada.**
Procedimiento	La solución tiene todos los **procesos** con **secuencia lógica**.	La solución tiene la mayoría de los **procesos** con **secuencia lógica**.	La solución tiene algunos **procesos** con **secuencia lógica**.	La solución tiene pocos **procesos** sin **secuencia lógica**.	La solución no tiene procesos, ni una **secuencia lógica**.
Funcionamiento	La solución tiene una **funcionalidad** correcta de todos los **requerimientos** descritos.	La solución tiene una **funcionalidad** correcta de la mayoría de los **requerimientos** descritos.	La solución tiene una **funcionalidad** correcta de algunos de los **requerimientos** descritos.	La solución tiene una **funcionalidad** incorrecta de algunos de los **requerimientos** descritos.	La solución tiene una **funcionalidad** incorrecta de todos los **requerimientos** descritos.

Las respuestas colocarla en la siguiente matriz:

Código Actividad:

Código Evaluador:
Código Evaluado:
Código Docente:

Criterios	Puntuación numérica	Retroalimentación
Documento		
Estructura		
Procedimiento		
Funcionamiento		

Andamiaje

Documento

- Caratula (la tarea no debe llevar el nombre del grupo ni de los integrantes)
- Enunciado

  Pocos errores ortográficos: menor a 3

  Algunos errores ortográficos: menor a 10

  Varios errores ortográficos: mayor a 10

Estructura

- Elementos (actores, casos de uso, comunicación)
- Sintaxis (verbo en infinitivo, verbo+objeto, código de caso de uso, comunicación entre actor y caso de uso, comunicación entre casos de uso, sentido de las flechas de aplicación de include, extend y herencia)
- Ordenado (orden de código de casos de uso por prioridad, líneas no presentan una adecuada asociación entre actor y caso de uso, y entre casos de uso, presentando dificultad en el entendimiento del diagrama)

Procedimiento

- Procesos (aplicación de include, extend y herencia entre actores y casos de uso de forma correcta)
- Secuencia lógica (aplicación lógica y secuencial en la asociación entre actores y casos de usos de forma correcta)

Funcionamiento

- Requerimientos (incluye todos los requerimientos del caso de estudio)
- Funcionalidad (casos de uso demuestran funcionalidad, interacción con el objeto)

Rúbrica holística

Se diseñó rúbrica de tipo holística para evaluar tareas específicas, con criterios que incluye las características del nivel de logro más alto que ayude al estudiante a identificar las características de la tarea, y a designar la puntuación numérica (nivel de logro) considerando una escala tipo Likert: 1 (Nada adecuado), 2 (Poco adecuado), 3 (Adecuado), 4 (Bastante adecuado) y 5 (Totalmente adecuado), y un campo de texto para que completen sugerencias u observaciones para mejorar por cada criterio. En la Tabla **13**, se detalla un ejemplo de rúbrica tipo holística.

Tabla 13. Rúbrica holistica

Código Actividad: **Código Evaluador:** **Código Evaluado:** **Código Docente:**			
Criterios	**Características**	**Puntuación numérica**	**Retroalimentación**
Diseño	La **estructura** es correcta con actores involucrados y caso de uso relacionado.		
Actores	Los **actores** están bien especificados y son necesarios en el sistema.		
Casos de uso	La **sintaxis** es adecuada aplicando numeración ordenada, verbos en infinitivo + objeto. Todos los **casos de uso** están completos. La **funcionalidad** descrita es correcta.		
Comunicación	Se utiliza **extend, include** y **herencia** de forma correcta y justificada según el requerimiento.		

Andamiaje

La valoración por cada criterio está dada con niveles de ponderación de 1-5: los niveles (1-2) indican que el estudiante ha hecho poco o nada en la tarea y los niveles (3-5) reflejan que el criterio ha sido completado en su mayoría o en su totalidad adecuadamente.

El nivel de ponderación se designa considerando la descripción de la retroalimentación textual

Retroalimentación	Nivel de ponderación
Tiene todos los aspectos evaluados de manera incorrecta y no ha realizado los requerimientos especificados.	1 (Nada adecuado)
Tiene todos los aspectos evaluados de manera incorrecta.	2 (Poco adecuado)
Tiene aspectos evaluados de manera correcta e incorrecta.	3 (Adecuado)
Tiene aspectos evaluados de manera correcta, sin embargo, puede tener un aspecto incompleto.	4 (Bastante adecuado)
Tiene todos los aspectos evaluados de manera correcta.	5 (Totalmente adecuado)

Resultados

Se realizó una prueba piloto para evaluar tareas en procesos de evaluación entre pares. En primer lugar, se aplicó la rúbrica tipo analítica (Tabla **12**) para evaluar la tarea D01-ejercicios de sentencias SQL en la asignatura de administración de bases de datos y para evaluar la tarea FIS01-ejercicios de diagrama de casos de uso en la asignatura de fundamentos de ingeniería del software. En segundo lugar, se aplicó rúbrica tipo holística (Tabla **13**), para evaluar la tarea FIS02-

ejercicios de diagrama de casos de uso en la asignatura de fundamentos de ingeniería del software.

Posteriormente en base a las experiencias de los estudiantes en la utilización de las rúbricas tipo analítica y holística se realizó un análisis comparativo del tipo de rúbrica que mejor se adapta para adicionar retroalimentación por cada criterio. Participaron 52 estudiantes de la asignatura de fundamentos de ingeniería de software impartida en escenario de educación presencial en el periodo académico octubre 2019-febrero 2020.

A continuación, se detalla el análisis comparativo de los elementos de la rúbrica: indicadores, niveles de logro, descriptores de logro y retroalimentación (Tabla **14**).

Tabla 14. Análisis comparativo de la utilización de rúbrica tipo analítica vs holística

Elementos	**Rúbrica analítica**			**Rúbrica holística**		
	Ventajas	**Desventajas**	**Nº**	**Ventajas**	**Desventajas**	**Nº**
Indicadores		Los criterios y características están generalizados para evaluar diferentes tareas, lo que implica interpretar el andamiaje y conllevaba mucho tiempo en el proceso.	6	Los criterios y características están específicos de la temática de la tarea, fácil de interpretar para evaluar.		46
Niveles de logros	Ubicado en la parte superior de la rúbrica lo que facilita la selección.				No se encuentra en la rúbrica, se tenía que interpretar en el andamiaje.	
Descriptores de logros	Ubicado en cada nivel de logro lo que facilita la selección.				No contenía descriptores de logros, lo que dificultaba escoger el nivel de logro correcto.	
Retroalimentación		Se complicaba dar la retroalimentación ya que se tenía que leer los niveles y descriptores de logro, luego el andamiaje, el proceso es tedioso.		La descripción de las características es específica de la tarea evaluar, lo que permitió proporcionar retroalimentación en base a la temática.		

N°= Número de estudiantes que prefieren el tipo de rúbrica

Los resultados del sondeo revelan que el 11,54% de los estudiantes prefieren la rúbrica tipo analítica y el 88,46% optan por la rúbrica de tipo holística ya que la descripción de las características por cada criterio les permitía proporcionar retroalimentación en base a la temática a evaluar.

Se deduce que la rúbrica de tipo holística facilita al estudiante evaluar tareas en procesos de evaluación entre pares con puntuación numérica y retroalimentación textual. Por tanto, será utilizada en la experimentación.

Análisis de sentimiento de retroalimentación textual

En segundo lugar, se evaluaron los algoritmos de clasificación de sentimiento de retroalimentación textual en español. A continuación, se detalla:

Materiales y métodos

Participantes

Participaron 81 estudiantes de la asignatura de fundamentos de ingeniería de software de la carrera de sistemas de información, en escenario de educación virtual asincrónica en el periodo académico noviembre 2020-marzo 2021 de la Universidad Técnica de Manabí (Ecuador), divididos en 20 grupos de cuatro integrantes. La actividad grupal consistió en realizar un diagrama de clases de un caso de estudio.

Metodología de experimentación

La metodología utilizada en la experimentación se realizó en base al procedimiento básico de evaluación entre pares basado en análisis de sentimiento de la subsección 3.1 y 3.5.

Se diseñó un modelo de análisis de sentimiento, aplicando la técnica de aprendizaje automático, que recibe como entrada un corpus de texto en lenguaje natural etiquetado y clasifica el sentimiento en (positivo/negativo) que corresponde a una retroalimentación textual específica (Figura **31**).

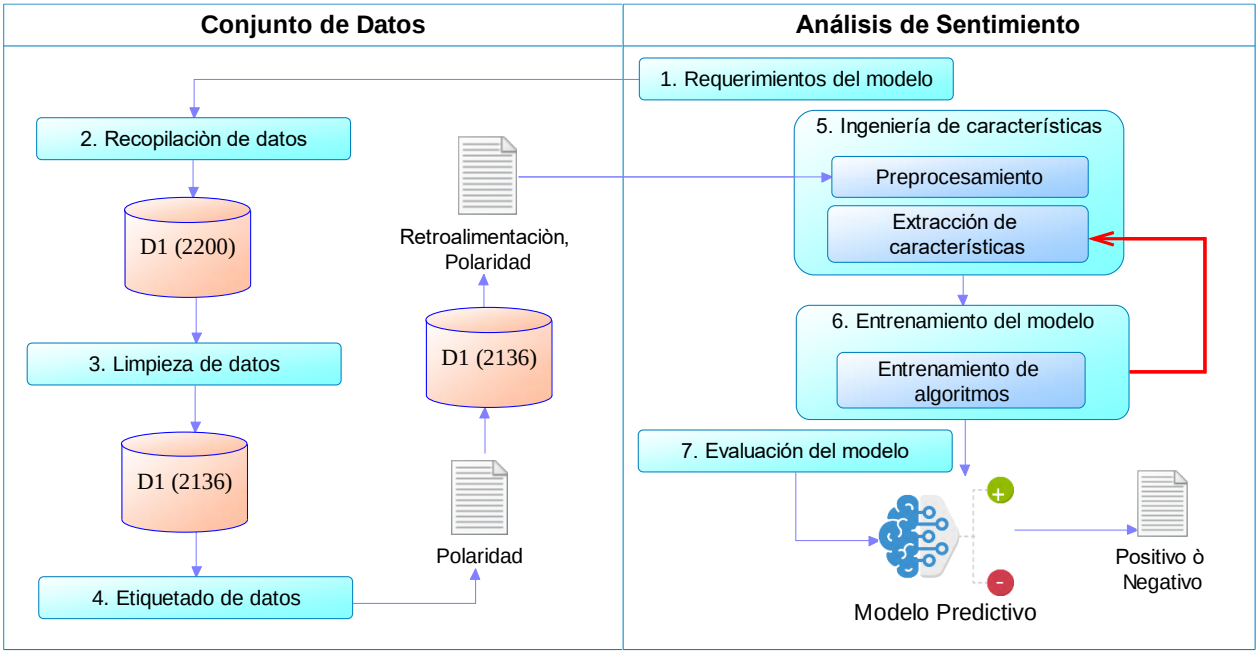


Figura 31. Metodología aplicada en el experimento I

La Figura **31** muestra tres representaciones gráficas, rectángulos para los pasos, almacenes de datos para el conjunto de datos y archivos para las variables de entrada y salida. Los pasos de recopilación, limpieza y etiquetado están orientados a los datos; los pasos de requisitos, ingeniería de características, capacitación y evaluación están orientados al análisis de sentimiento. A continuación, se detallan los pasos:

Paso 1. Requerimientos del modelo

- Se aplicó el enfoque supervisado con modelos de aprendizaje automático para evaluar el sentimiento a nivel de oración.

Paso 2. Recopilación de datos

- Se recopiló datos de evaluación por pares en el idioma español en una asignatura.

  Escenario de evaluación entre pares

  – El escenario se lo realizó considerando a la topología de topping [39] (Tabla **15**). La función de la actividad de evaluación entre pares fue de naturaleza formativa y sumativa.

- La actividad se realizó en dos rondas. En la primera ronda, los trabajos de cada grupo consistieron en la elaboración de un diagrama de clase, y los evaluadores proporcionaron retroalimentación formativa. En la segunda ronda, en base a las retroalimentaciones dadas en la primera ronda cada grupo realizó la corrección del trabajo y los evaluadores volvieron a brindar retroalimentaciones.
- A los estudiantes se les indicó que la retroalimentación que proporcionen a sus compañeros en la primera ronda no afectaría la calificación de la actividad para evitar posibles aprensiones. La intención es promover la idea de que los estudiantes pueden participar en un proceso de reflexión y mejorar las tareas. Sin embargo, para estimular el esfuerzo y justificar la inversión de tiempo en la actividad, se tuvo en cuenta el 20% de la puntuación media de los evaluadores pares de la segunda ronda.

Tabla 15. Escenario de evaluación entre pares del experimento I

Dimensión	**Rango de variación**
Área curricular/Asignatura	Fundamentos de ingeniería de software.
Objetivos	El docente reduce la carga de calificación y los estudiantes ven otras posibles soluciones y obtienen ganancias cognitivas.
Enfoque	Formativa y sumativa.
Producto/Salida	Diagrama de clases.
Relación de la evaluación personal	Parcialmente sustitucional.
Valor oficial	80% de la puntuación del docente + 20% de la puntuación media de los pares evaluadores de la segunda ronda.
Direccionalidad	Mutua.
Privacidad	Anónima.
Contacto	La evaluación se realiza apoyándose del aula virtual.
Año	Mismo año de estudio.
Habilidad	La evaluación se guía por una rúbrica para obtener el máximo beneficio de las habilidades del evaluador.
Constelación del evaluador	Individual.
Constelación evaluada	Las tareas enviadas se realizaron de manera colaborativa.
Lugar	Virtual.
Tiempo	Tiempo de clase.
Requerimiento	Obligatorio para evaluadores/evaluados.
Bonificación	Ninguna.

Rúbrica

- Se diseñó una rúbrica tipo holística de diagrama de clases con cuatro criterios y características para cada criterio (Tabla **16**).

Tabla 16. Rúbrica de evaluación de diagrama clases

Código Actividad: **Código Evaluador:** **Código Evaluado:**			
Criterio	**Descripción**	**Puntuación Numérica**	**Retroalimentación**
Clases	Los **nombres de las clases** son adecuados, son sustantivos y empieza con mayúsculas. Globalmente se detectan todas las **clases** correctamente. Para cada clase se indica la **estructura** adecuada: atributos y métodos importantes. La **información** reportada es detallada y coherente.		
Atributos	Los **nombres de los atributos** son adecuados, son sustantivos o adjetivos, **y** expresan claramente una propiedad de la clase. Se indica el **tipo de dato** y la **visibilidad** adecuada. Se han detectado todos los **atributos** correctamente.		
Métodos	Los **nombres de los métodos** son adecuados, son verbos, y expresan claramente una acción u operación de la clase. Se indica los **argumentos, tipo de dato**, **tipo de valor de retorno** y **visibilidad** adecuada**.** Los métodos indican **funcionalidad** correcta de los requerimientos descritos.		
Asociaciones	Todas las **cardinalidades** se denotan correctamente**.** La **dirección de navegación** de todas las asociaciones se especifica de forma correcta.		

- La valoración por cada criterio está dada con niveles de ponderación de 1-5: los niveles (1-2) indican que el estudiante ha hecho poco o nada en la tarea, y los niveles (3-5) reflejan que el criterio ha sido completado en su mayoría o en su totalidad adecuadamente.
- El nivel de ponderación se designa considerando la descripción de retroalimentación textual.

Retroalimentación	Nivel de ponderación (Puntuación Numérica)
Tiene todos los aspectos evaluados de manera incorrecta y no ha realizado los requerimientos especificados.	1 (Nada adecuado)
Tiene todos los aspectos evaluados de manera incorrecta.	2 (Poco adecuado)
Tiene aspectos evaluados de manera correcta e incorrecta.	3 (Adecuado)
Tiene aspectos evaluados de manera correcta, sin embargo, puede tener un aspecto incompleto.	4 (Bastante adecuado)
Tiene todos los aspectos evaluados de manera correcta.	5 (Totalmente adecuado)

Procedimiento para la evaluación por pares

Formación de grupos

- El docente establece actividad en equipos [217], [222], [238] en un entorno virtual de aprendizaje. Se formaron 20 grupos de cuatro integrantes.

Andamiaje

- El docente proporciona instrucciones [8], [227] sobre los criterios de la rúbrica y las etapas de la evaluación por pares.

Envíos de trabajo

- Cada grupo realizó el trabajo y lo envió a la plataforma virtual.

Asignación del evaluador

- El docente asigna tres revisiones de forma anónima [217], [227], [238].

Evaluación del trabajo

- Los estudiantes/docente evalúan los trabajos utilizando un artefacto de evaluación [217], [227], [238]. Cada evaluador evalúa cada trabajo asignado, proporcionando retroalimentación para cada criterio de la rúbrica.

- En este paso, se obtuvo un conjunto de datos (D1) con 2200 instancias.

Paso 3. Limpieza de datos

- Se descarta los registros vacíos en la retroalimentación, el conjunto de datos (D1) se redujo en 2136 instancias (Tabla **18**).

Paso 4. Etiquetado de datos

- El docente de la asignatura de fundamentos de ingeniería de software etiquetó cada retroalimentación del conjunto de datos (D1) considerando las reglas de la subsección 3.5. La Tabla **17** muestra ejemplos de retroalimentaciones etiquetadas en español. La retroalimentación (F1) fue etiquetada como (positiva) porque contiene las palabras "adecuada" y "correctamente". La retroalimentación (F2) fue etiquetada (negativa) porque contiene las palabras "no detectado".

Tabla 17. Esquema de etiquetado adaptado de [284] y ejemplo

Categoría	Subcategoría	Descripción	Ejemplo	
			Retroalimentación	Polaridad
Tipo de verificación	Positivo	¿Es la oración de retroalimentación una afirmación evaluativa positiva?	F1: Los nombres de las clases son adecuados, si se detectan las clases correctamente.	Positivo
	Negativo	¿Es la oración de retroalimentación una declaración evaluativa negativa?	F2: No se han detectado todos los atributos correctamente.	Negativo

- La Tabla **18** muestra el detalle del conjunto de datos para la experimentación.

Tabla 18. Detalle del conjunto de datos para el experimento I

Conjunto de datos	Asignatura	Tarea	Ronda	Instancia	N°. Retroalimentación positiva	N°. Retroalimentación negativa
D1	Fundamentos de ingeniería de software	Diagrama de clases	R1	1068	601	467
			R2	1068	692	376
		Total de instancias		2136	1293	843

Paso 5. Ingeniería de características

- Una vez que se obtuvieron los conjuntos de datos, se preparó un CSV con el conjunto de datos (D1) (Tabla **18**) con las variables de entrada (Retroalimentación y Polaridad Sentimiento) para el entrenamiento del modelo.

  Preprocesamiento

  – Se consideró realizar normalización, convertir todos los tokens a minúsculas al compararlos con las entradas del diccionario, y tokenización, para dividir una oración en palabras mediante la eliminación de los signos de puntuación [282]. Stemmer para eliminar el sufijo de la palabra al reducirlo a su forma de raíz [2]. Stop-Words que no contribuyen al análisis se eliminan durante el paso de preprocesamiento [282].

  Extracción de características

  – Se probó Bag of Words, para convertir un texto en su vector equivalente de números [259]. TF-IDF para medir qué tan importante es una palabra para un oración en un corpus [150], [247], [259], [272].

Paso 6. Entrenamiento del modelo

- Se utilizó la técnica de validación cruzada de 10 veces para evaluar los clasificadores; ya que es más adecuado para pequeños conjuntos de datos.
- Se probó los algoritmos: NB usa las probabilidades condicionales de las palabras en un texto para determinar a qué categoría pertenece [150], [247], [254], [258]. SVM construye una serie de híperplanos, un hiperplano optimo es el que maximiza las distancias entre clases, los puntos usados para definir el hiperplano se llaman vectores de soporte y dependiendo del lado del hiperplano será la clase a la que pertenece el texto [16]–[19], [134], [136], [249], [250], [261]. LibSVM implementa el algoritmo de optimización mínima secuencial (SMO) para SVM kernelizadas [295]. K-NN (IBk) usa cada instancia encontrada para clasificarla en la clase más frecuente a la que pertenezcan sus K vecinos más cercanos [136], [144], [152], [259].

Paso 7. Evaluación del modelo

- Se comparó con F-Measure porque es la mejor medida para describir el rendimiento del modelo cuando los datos no están equilibrados [146], [291]–[293].
- Evaluar el modelo mediante análisis estadístico Chi-cuadrado de McNemar’s y prueba T pareada.

Materiales

Para implementar el experimento, se usó:

- Bibliotecas de Waikato Environment for Knowledge Analysis (WEKA) para implementar algoritmos de minería de datos para el preprocesamiento que se ejecutaron en v. 3.9.5:
- Software de análisis de datos (SPSS) v. 25 para análisis descriptivo.

Resultados

Rendimiento de los algoritmos

Se obtiene un modelo predictivo con mejor rendimiento para generar el sentimiento (positivo o negativo) correspondiente a una retroalimentación textual específica (Figura **31**), aplicando minería de texto a través de WEKA.

Se utilizó el formato de archivo de relación de atributos (ARFF) porque consume menos memoria, es más rápido y mejor para el análisis, ya que incluye metadatos sobre los encabezados de las columnas. Fueron necesarias varias técnicas de preprocesamiento: conversión a minúsculas y Stemmer, para transformar la retroalimentación antes de aplicar el clasificador.

Se usó un filtro StringToWordVector para convertir atributos de cadena a numéricos y se usó WordTokenizer para dividir las retroalimentaciones en palabras/términos, construyendo un vector de palabras conocido como como Bag of Words y TF-IDF para determinar la importancia de una palabra clave.

Se realizó cada prueba con los algoritmos: NB, LibSVM y K-NN (IBk) enfocados en encontrar la combinación óptima de Bag of Words, TF-IDF, Stemmer y Stop-Words. El mejor rendimiento de cada algoritmo (Tabla **19**) se detalla a continuación:

El algoritmo NB obtuvo mejor rendimiento cuando la retroalimentación se representó con Bag of Words, Stemmer y Stop-Words. La medida ponderada de Precision informa que en un 0.834 los valores de dicha clase son realmente de dicha clase. La medida ponderada de Recall informa que los verdaderos positivos de cada una de las clases se clasificaron en un 0.833 de manera correcta. F-Measure del algoritmo es del 0.833. El valor de Kappa es 0.662 que informa que el algoritmo tuvo una buena concordancia diciendo que los datos en su clasificación concuerdan con los valores predichos.

El algoritmo LibSVM obtuvo mejor rendimiento usando una función radial y cuando la retroalimentación se representó con Bag of Words, TF-IDF y Stop-Words. La medida ponderada de Precision informa que en un 0.882 los valores de cada clase son realmente de dicha clase. La medida ponderada de Recall informa que los verdaderos positivos de cada una de las clases se clasificaron en un 0.873 de manera correcta. F-Measure del algoritmo es del 0.870 en sus clasificaciones. El valor de Kappa es de 0.750 que informa que el algoritmo tuvo una buena clasificación de instancias verdaderas y una buena predicción para cada clase.

El algoritmo K-NN (IBk) obtuvo mejor rendimiento cuando la retroalimentación se representó con Bag of Words, Stemmer y Stop-Words. La medida ponderada de Precision informa que en un 0.819 los valores de cada clase son realmente de esa clase. La medida ponderada de Recall informa que los verdaderos positivos de cada una de las clases se clasificaron en un 0.816 de manera correcta. La medida ponderada de F-Measure muestra un balance entre Precision y Recall del 0.813. El valor de Kappa es de 0.619 que informa que el algoritmo tuvo una buena concordancia entre su clasificación con los valores predichos.

Se seleccionó LibSVM (F-Measure of 0.870) para generar las predicciones de sentimiento en idioma español porque obtuvo mejores resultados que los clasificadores NB e IBk (K-NN) cuando la retroalimentación se representó con Bag of Words, TF-IDF y Stop-Words. Los resultados son similares a otras investigaciones; [19] obtuvieron 0.780 aplicando SVM para evaluar texto en idioma inglés; [248] consiguieron 0.990 usando SVM para clasificar texto en idioma turco; [18] lograron 0.42 utilizando SVM en texto en idioma inglés; [278] alcanzaron 0.829 en positivo y 0.868 en negativo aplicando SVM en clasificar texto en idioma inglés; [250] adquirieron 0.830 empleando SVM en clasificar texto en idioma inglés; [261] obtuvieron 0.600 utilizando SVM en clasificar texto en idioma inglés; [134] consiguieron 0.950 usando SVM en clasificar texto en idioma inglés; y [136] obtuvieron 0.958 usando SVM en clasificar texto en idioma inglés.

La literatura muestra que los modelos con Bag of Words utiliza frecuencias de palabras individuales discretas para representar el texto [145]. No puede capturar las relaciones semánticas entre los componentes de los documentos de texto [144]. Además, este esquema produce una representación de datos dispersos con un espacio de características de alta dimensión [159]. Sin embargo si se emplea para seleccionar las características métodos como TF-IDF, donde las ocurrencias de un término en el texto procesado están relacionadas con las ocurrencias en todos los textos del conjunto de datos se obtiene mejores resultados en la clasificación del sentimiento [150], [153], [259], [296]. Otros autores destacan que los enfoques más sofisticados que se discuten en la literatura usan secuencia de palabras (N-Grams) para clasificar el sentimiento [145], [152], [219].

Tabla 19. Comparaciones de rendimiento de NB, LibSVM y K-NN (IBk)

Models	LibSVM				NB				K-NN (IBk)			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
BoW	0.768	0.616	0.516	0.451	0.797	0.798	0.797	0.136	0.790	0.790	0.788	0.568
BoW + Stemmer	0.813	0.786	0.775	0.546	0.818	0.816	0.817	0.629	0.798	0.795	0.793	0.577
BoW + SW	0.834	0.780	0.763	0.528	0.815	0.814	0.814	0.623	0.806	0.802	0.799	0.589
BoW + Stemmer + SW	0.840	0.830	0.827	0.647	0.834	0.833	**0.833**	**0.662**	0.819	0.816	**0.813**	**0.619**
BoW + TF-IDF	0.777	0.715	0.683	0.381	0.767	0.768	0.767	0.524	0.784	0.785	0.783	0.558
BoW + TF-IDF + Stemmer	0.784	0.736	0.713	0.432	0.762	0.763	0.761	0.513	0.797	0.796	0.794	0.579
BoW + TF-IDF + SW	0.882	0.873	**0.870**	**0.750**	0.772	0.773	0.771	0.534	0.791	0.790	0.788	0.568
BoW + TF-IDF + Stemmer + SW	0.796	0.759	0.743	0.486	0.765	0.766	0.763	0.517	0.785	0.784	0.782	0.554

P=Precision, R=Recall, F1=F-Measure, K=Kappa, BoW=Bag of Words, SW=Stop-Words, TF-IDF=Term Frequency-Inverse Document Frequency

Evaluación del modelo

Se realizó un análisis estadístico mediante la prueba de McNemar's para evaluar si el modelo de análisis de sentimiento de retroalimentación textual es útil para apoyar que los estudiantes mejoren sus tareas en base a la retroalimentación formativa. La hipótesis nula y alternativa se formuló de la siguiente manera:

- H0: En la evaluación entre pares cualitativa aplicando análisis de sentimiento no existen diferencias estadísticamente significativas en la puntuación de sentimiento de la primera y segunda ronda.
- H1: En la evaluación entre pares cualitativa aplicando análisis de sentimiento existen diferencias estadísticamente significativas en la puntuación de sentimiento de la primera y segunda ronda.

El estadístico Chi-cuadrado de McNemar's es 19.51 y el valor de p es 0.000 con un grado de libertad, por lo que se rechaza la hipótesis nula (valor p<0.05) (Tabla **20** y Tabla **21**). Además, se aplicó la prueba t pareada. La puntuación media de la primera ronda fue de 0.56, mientras que la de la segunda ronda fue de 0.65 y el valor P es 0.000<0.05 (Tabla **22** y Tabla **23**). Por tanto, existe una diferencia significativa entre la puntuación media de sentimiento de la primera y segunda ronda. Se deduce que la retroalimentación brindada por los pares en la primera ronda ayudó a los estudiantes a corregir sus trabajos. Se puede determinar en este experimento que si es útil aplicar el modelo de análisis de sentimiento de retroalimentación textual en el proceso de evaluación entre pares.

Tabla 20. Aplicación del modelo de análisis de sentimiento de retroalimentación entre pares en la primera y segunda ronda

		Segunda Ronda		
		Negativo	**Positivo**	**Total**
Primera Ronda	Negativo	214	162	376
	Positivo	253	439	692
	Total	467	601	1068

Tabla 21. Pruebas de chi-cuadrado aplicada en la primera y segunda ronda

	Valor	**Significación exacta (bilateral)**	**Significación exacta (unilateral)**	**Probabilidad en el punto**
Prueba de McNemar	19.51	,000[a]	,000[a]	,000a

[a] Distribución binomial utilizada.

Tabla 22. Prueba T pareada aplicada en la primera y segunda ronda

	Media	N	Desv. Desviación	Desv. Error promedio
Primera Ronda	,56	1068	,468	,015
Segunda Ronda	,65	1068	,478	.015

Tabla 23. Correlación entre primera y segunda ronda

	N	Correlación	Sig.
Primera Ronda & Segunda Ronda	1068	,196	,000

Muestra de aplicación del modelo de análisis de sentimiento de retroalimentación de pares en la primera y segunda ronda

Para obtener la puntuación final de la actividad se equiparó el sentimiento negativo con una puntuación de 0 y el sentimiento positivo con una puntuación de 1 en cada criterio (Tabla **24**). Posteriormente se obtuvo la puntuación media del docente y de cada evaluador. Finalmente se obtiene la puntuación de cada grupo con el 80% de la puntuación media del docente y el 20 % de la puntuación media de pares (Tabla **25**). Para esta actividad se determinó que la calificación es sobre 10 puntos.

Tabla 24. Algunas muestras de análisis de sentimiento de retroalimentación de pares del grupo (Q) y evaluador (E-1)

Criterio	Primera ronda		Puntuación Sentimiento	Segunda ronda		Puntuación Sentimiento
	Retroalimentación	Sentimiento		Retroalimentación	Sentimiento	
Clases	Los nombres de las clases no son los adecuados. No se detectan las clases correctamente. La información no es detalla ni coherente.	Negativo	0	Los nombres de las clases son adecuados; son sustantivos y empiezan con mayúscula, si se detectan las clases correctamente. La información reportada es detallada y coherente.	Positivo	1
Atributos	Los atributos no son los adecuados. No se indica tipo de dato. No se han detectado todos los atributos correctamente.	Negativo	0	Los nombres de los atributos son los adecuados; expresan propiedad. Se indica tipo de dato. Los atributos son correctos.	Positivo	1

Criterio	Primera ronda		Puntuación Sentimiento	Segunda ronda		Puntuación Sentimiento
	Retroalimentación	Sentimiento		Retroalimentación	Sentimiento	
Métodos	Los nombres de los métodos no son adecuados; no son claros. No se indica tipo de dato. No indican Funcionalidad.	Negativo	0	Los métodos son adecuados; son verbos y expresan operación. Si índica el tipo de dato. Los métodos indican funcionalidad.	Positivo	1
Asociaciones	No tienen cardinalidad.	Negativo	0	Las cardinalidades son parcialmente correctas, la dirección de navegación no la añadieron.	Negativo	1
	Puntuación del evaluador (media)		0			0.75

Tabla 25. Algunas muestras del grupo (Q) en la segunda ronda

Ronda	Grupo Evaluado	Evaluador	Puntuación Sentimiento Pares	Docente	Puntuación Grupo	Puntuación Final
Segunda	Q	P		0.5	0.58	5.81
		E-1	0.75	(80%) 0.4		
		E-2	1			
		E-3	0.75			
		E-4	0.75			
		E-5	1			
		E-6	1			
		E-7	1			
		E-8	1			
		E-9	1			
		E-10	1			
		E-11	1			
		E-12	0.75			
		E-13	0.75			
			0.90			
			(20%) 0.18			

P=Professor; E=Estudiante

Percepciones de los estudiantes sobre la retroalimentación de los compañeros

Al finalizar la actividad, se solicitó a los estudiantes que respondieran un cuestionario diseñado a partir de la propuesta de [222] para medir percepciones con las siguientes categorías: (1) percepción general respecto a la actividad realizada; (2) evaluación de la retroalimentación recibida; y (3) evaluación sobre la intención de la retroalimentación dada.

Desde la percepción general del estudiante, se observó que todos los aspectos considerados han recibido una puntuación media muy positiva (Tabla **26**). Los participantes consideran que la retroalimentación ha aumentado su participación en las actividades (4.38), el

andamiaje recibido de la actividad ha sido adecuado (4.35), la frecuencia de la retroalimentación ha sido adecuada (4.10) y la carga de trabajo ha sido aceptable (3.94).

Tabla 26. Percepción de los estudiantes sobre la utilidad de la retroalimentación

Elementos	Percepción de los estudiantes (medias)
La retroalimentación ha aumentado su participación en las actividades	4.38
El andamiaje recibido de la actividad ha sido adecuado	4.35
La frecuencia de la retroalimentación ha sido adecuada	4.10
La carga de trabajo ha sido aceptable	3.94

La Figura **32** detalla los resultados obtenidos en los aspectos referidos a la valoración de la retroalimentación recibida por parte de los evaluadores y la valoración de la retroalimentación dada. Se puede observar que todos los aspectos considerados han recibido una puntuación media muy positiva. Los estudiantes han considerado que la retroalimentación les ha ayudado a mejorar sus habilidades (4.36), a mejorar el proceso de aprendizaje (4.27), a mejorar sus trabajos (4.25), a mejorar su implicación en el aprendizaje (4.25) y a mejorar la adquisición de contenidos (4.18). Además, se destaca que los estudiantes tendieron a valorar más la retroalimentación que proporcionan a sus compañeros en todos los indicadores propuestos.

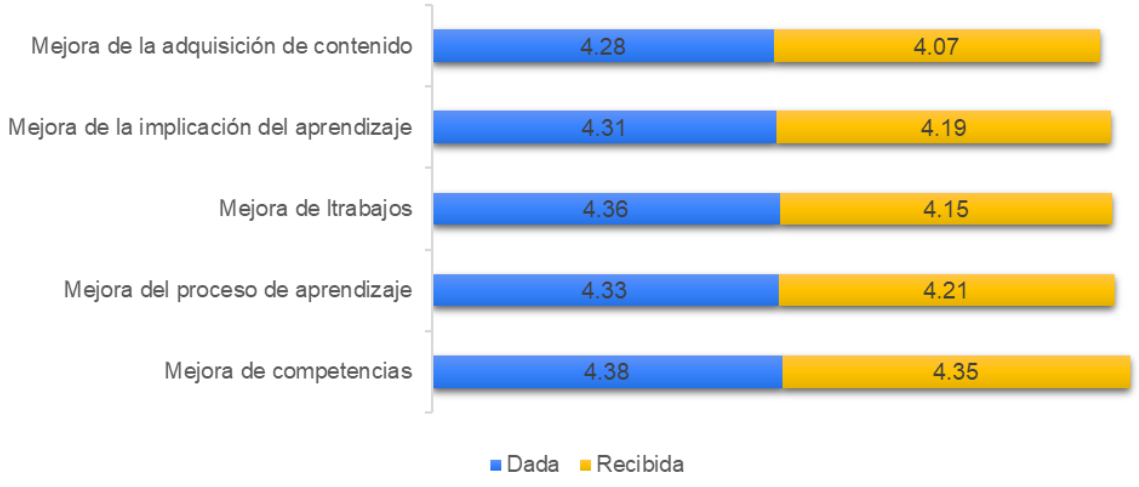


Figura 32. Puntajes promedio dados y recibidos de las retroalimentaciones

Finalmente, las respuestas abiertas de los estudiantes destacaron las potencialidades y los puntos para mejorar la retroalimentación y fueron polarizadas por el modelo en positivas y negativas (Figura **33**).

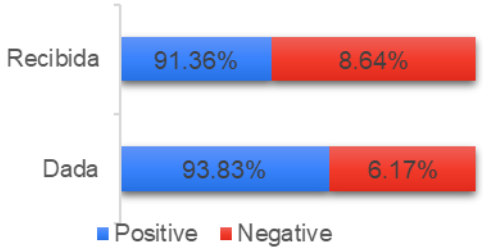


Figura 33. Percepción de los estudiantes.

El 91,36 % de los participantes destacó el papel proactivo de la retroalimentación recibida; un estudiante comentó: “La retroalimentación ha sido de gran ayuda para aceptar críticas constructivas y saber, desde el punto de vista de mis compañeros, qué podría mejorar en el trabajo”. El 8,64% afirma que la retroalimentación recibida no ha sido beneficiosa; un estudiante comentó: "Algunos evaluadores no especifican adecuadamente dónde está el error en la retroalimentación; sin embargo, otros me ayudaron a mejorar en aspectos que no dominaba en la asignatura". El 93,83% destaca el papel proactivo de la retroalimentación dada, y un estudiante comentó: "La retroalimentación que proporcione me ayudó a mejorar mi reflexión y a aprender de los errores de los demás para aplicar correcciones a mi trabajo". El 6,17% afirmó que no fue útil; un estudiante comentó: "Me resultaba difícil dar retroalimentación a un grupo porque no se entendía el diagrama y me absorbía mucho tiempo".

Conclusiones, limitaciones y futuro experimento

- Se analizaron dos tipos de rúbricas: analítica y holística. Los resultados mostraron que la rúbrica tipo holística facilita al estudiante evaluar tareas en procesos de evaluación entre pares con puntuación numérica y retroalimentación textual.
- Se entrenó los algoritmos: NB, LibSVM, y K-NN (IBk) con Bag of Words, TF-IDF, Stemmer y Stop-Words, utilizando un conjunto de datos en español. El algoritmo de clasificación con mejor rendimiento fue LibSVM (F-Measure de 0.870) con representación de Bag of Words+TF-IDF+Stop-Words por sus características de entrenamiento con textos cortos y la utilización de frecuencias de términos para representar el texto a clasificar capturó mejor la

polaridad de las oraciones en una colección cuyo vocabulario se repitió a través de las distintas instancias.

- El reemplazar los términos del conjunto de datos con sus formas \stemizadas no representó una mejora en el rendimiento de los algoritmos.
- En la comprobación de la hipótesis, el nivel de significancia fue inferior a 0.05. Por lo tanto, se deduce que la retroalimentación brindada por los compañeros en la primera ronda ayudó a los estudiantes a corregir su trabajo y mejorar su rendimiento en la segunda ronda.
- Las limitaciones que se tuvieron en este experimento fueron:
  - No existe colección de datos de retroalimentación entre pares en español, se tuvo que crear el conjunto de datos, por lo que el tamaño de la muestra fue pequeño.
  - La herramienta WEKA no facilitó la combinación de modelos.
- En el trabajo futuro, se planea:
  - Mejorar el escenario de evaluación por pares con evaluación cuantitativa para mejorar la confiabilidad del modelo.
  - Expandir el conjunto de datos.
  - Analizar otras combinaciones de técnicas de extracción de características como N-Grams, TF-IDF, word embedding, y algoritmos de aprendizaje automático y profundo que clasifiquen la retroalimentación como positivo/negativo.
  - Investigar y aplicar una técnica de computación blanda que permita correlacionar la evaluación cualitativa y cuantitativa.
  - Utilizar otra herramienta que facilite la combinación de modelos.
- Estos resultados se presentan:
  - Artículo: Sentiment Analysis of Peer Feedback in Higher Education. Aceptado para publicación en AIP (American Institute of Physics) Conference Proceedings (ver Apéndice B)

4.1.2. Experimento II

En este experimento se consideró la evaluación cualitativa y cuantitativa de la tarea (Figura **34**), del modelo de evaluación entre pares descrito en la subsección 3.2.

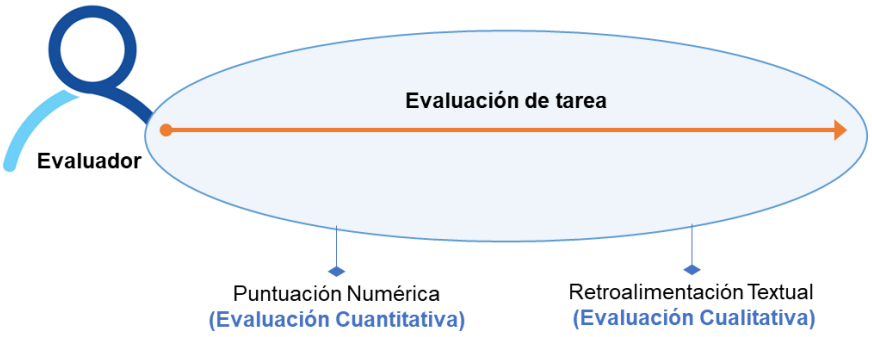


Figura 34. Evaluación entre pares cualitativa y cuantitativa

Planteamientos, preguntas de investigación y objetivos

Planteamientos		
• Evaluar diferentes técnicas de minería de texto (N-Grams, TFIDF, Stop-Words) para formar los distintos vocabularios y establecer que combinación hace mejorar la tarea de clasificación de sentimiento y determinar qué algoritmo arroja un mejor rendimiento en la tarea de clasificación de sentimiento de retroalimentación textual en español. • Determinar qué método de defuzzificación en lógica difusa, resulta más apropiado para generar la puntuación de evaluación de la correlación entre puntuación de sentimiento y numérica, por cada criterio o evaluador o grupo.		
Preguntas de investigación		
• ¿Como aplicar evaluación entre pares (cualitativa, cuantitativa) en escenarios de educación superior?	• ¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?	• ¿Cómo correlacionar puntuación numérica (evaluación cuantitativa) con retroalimentación textual (evaluación cualitativa), y generar una puntuación equilibrada entre estas dos evaluaciones?

Objetivos

- Diseñar los artefactos para validar los métodos teóricos.
- Construir un modelo de evaluación entre pares basado en análisis de sentimiento.
- Evaluar los resultados de la precisión del modelo.

A continuación, se detalla el experimento:

Materiales y métodos

Participantes

Participaron 68 estudiantes de cuarto nivel de la asignatura de fundamentos de ingeniería de software y 63 estudiantes de tercer nivel de la asignatura de administración de bases de datos, en escenario de educación presencial del período académico octubre 2019-febrero 2020, todos pertenecientes a la carrera de ingeniería de sistemas informáticos, de la Universidad Técnica de Manabí (Ecuador).

Metodología de experimentación

La metodología utilizada en la experimentación se realizó en base a la subsección 3.5 y 3.6, que constó de dos fases (Figura **35**). En la primera fase, se realizó análisis de sentimiento, aplicando la técnica de aprendizaje automático, que recibe como entrada un corpus de texto en lenguaje natural etiquetado y genera una puntuación de sentimiento (1 positivo/-1 negativo) que corresponde a retroalimentación textual específica. En la segunda fase, se realizó la detección de precisión/imprecisión entre la puntuación de sentimiento y numérica, y se calculó la puntuación de evaluación utilizando la técnica de lógica difusa.

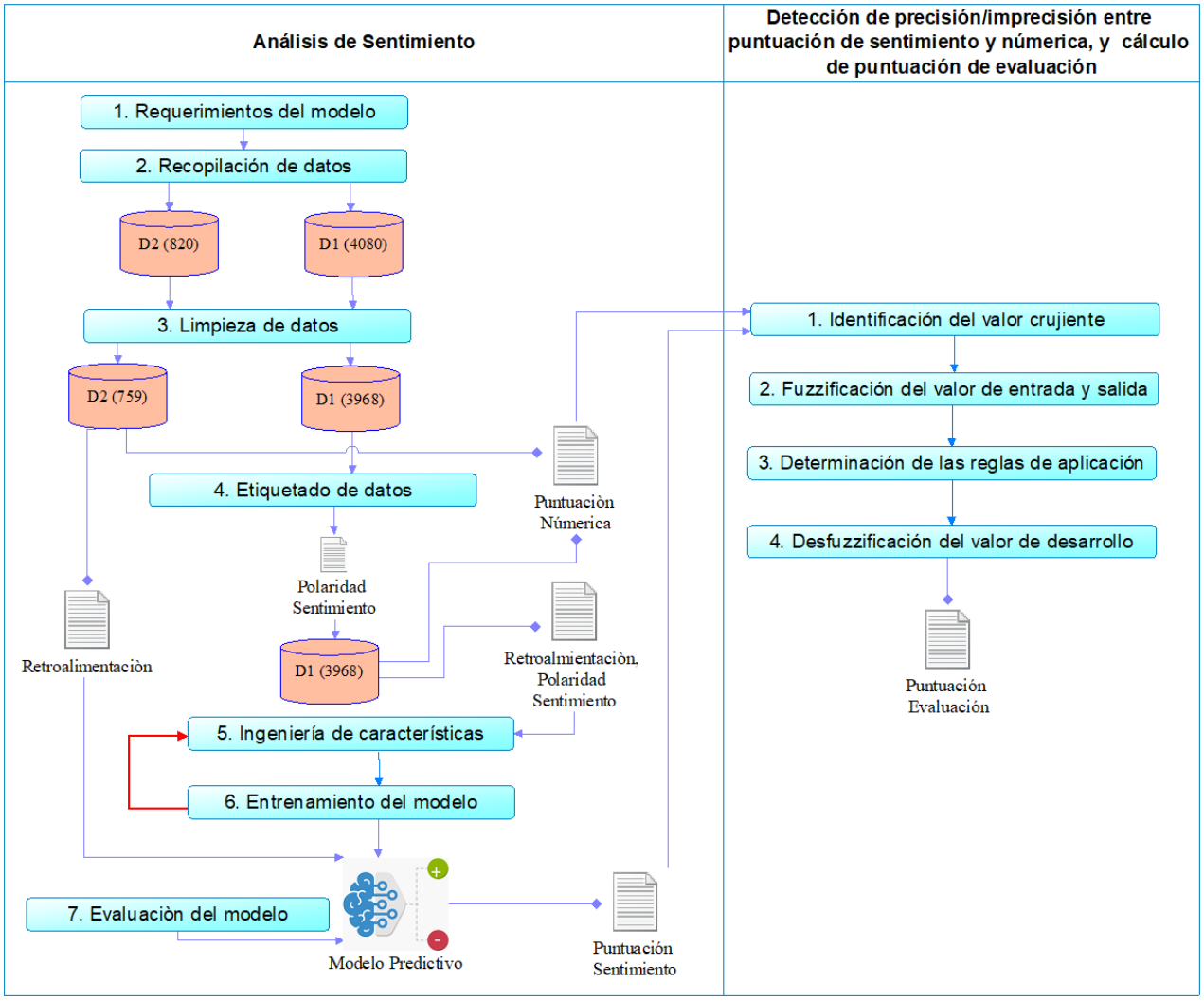


Figura 35. Metodología aplicada en el experimento II

La Figura **35** muestra cinco representaciones gráficas, rectángulos para los pasos, almacenes de datos para los conjuntos de datos, archivos para las variables de entrada y salida, un icono para el modelo predictivo y la flecha de retroalimentación de color rojo para ilustrar que el entrenamiento del modelo puede regresar a ingeniería de características. Los pasos de recopilación, limpieza y etiquetado están orientados a los datos; los pasos de requerimientos, ingeniería de características, entrenamiento y evaluación están orientados al modelo; y los pasos de fuzzificación, reglas y defuzzificación están orientados al cálculo.

A continuación, se describe con detalle el procedimiento llevado a cabo en cada una de las fases:

Primera fase: Análisis de sentimiento

Esta fase considera los pasos del proceso de minería de datos descrito en la subsección 3.5, para obtener el modelo predictivo que genera la puntuación de sentimiento a partir de la retroalimentación textual. Los siete pasos se resumen a continuación:

Paso 1. Requerimientos del modelo

- En la experimentación, se aplicó el enfoque de aprendizaje supervisado con modelos de aprendizaje automático y aprendizaje profundo.
- El sentimiento se evaluó a nivel de oración para probar hasta qué punto los algoritmos pueden evaluar correctamente la polaridad global asociada con la retroalimentación.

Paso 2. Recopilación de datos

- A través de un escenario de evaluación por pares descrito en la subsección 3.3, se recolectó datos en el idioma español en dos asignaturas (Tabla **27**).

Tabla 27. Escenario de evaluación entre pares del experimento II

Dimensión	**Rango de variación**
Área curricular/Asignatura	Fundamentos de ingeniería de software, administración de bases de datos.
Objetivos	Docente reduce la carga de calificación y los estudiantes ven otras posibles soluciones y obtienen ganancias cognitivas.
Enfoque	Cuantitativa y cualitativa.
Producto/Salida	Diagrama de: casos de uso, clases, actividad y secuencia.
Relación de la evaluación personal	Sustitucional.
Valor oficial	100% de la puntuación de evaluación del colectivo.
Direccionalidad	Mutua.
Privacidad	Anónima.
Contacto	La evaluación se realiza mediante rúbricas en Excel.
Año	El mismo año de estudio.
Habilidad	La evaluación está guiada por una rúbrica para obtener el máximo beneficio de las habilidades del evaluador.

Dimensión	Rango de variación
Constelación del evaluador	Individual (3 tareas asignadas).
Constelación evaluada	Las tareas enviadas se realizaron de manera colaborativa.
Lugar	En clase presencial.
Tiempo	Hora de clase.
Requerimiento	Obligatorio para evaluadores/evaluados.
Bonificación	Cuando la correlación entre puntación de sentimiento y numérica es imprecisa se penaliza dando la puntuación más baja (1-Nada adecuado)

- Se aplicó los pasos del procedimiento de evaluación de tarea, descrito en la subsección 3.4.

Formación de grupos y diseño de artefactos

- En la asignatura de fundamentos de la ingeniería del software se formaron grupos de cuatro estudiantes y se diseñaron cinco actividades que consistieron en la resolución de ejercicios de diagrama de casos de uso, diagrama de clases, diagrama de secuencia y diagrama de actividades. En la asignatura de administración de bases de datos se formaron grupos de cinco estudiantes y se diseñó una actividad que consistió en la resolución de ejercicios de sentencias SQL (Tabla **29**).
- Para cada actividad se diseñaron rúbricas considerando los parámetros de rúbrica tipo holística, divididas en criterios, cada criterio incluía un campo para recolectar la puntuación numérica (1 a 5) y un campo de texto para que los evaluadores completen retroalimentación con sugerencias u observaciones para mejorar (Tabla **28**):

Tabla 28. Rúbrica de evaluación de ejercicios de diagrama de casos de uso

Código Actividad: **Código Evaluador:** **Código Evaluado:**			
Criterio	**Descripción**	**Puntuación Numérica**	**Retroalimentación**
Diseño	La **estructura** es correcta con actores involucrados y caso de uso relacionado.		
Actores	Los **actores** están bien especificados y son necesarios en el sistema.		
Casos de uso	La **sintaxis** es adecuada aplicando numeración ordenada, verbos en infinitivo + objeto. Todos los **casos de uso** están completos. La **funcionalidad** descrita es correcta.		
Comunicación	Se utiliza **extend, include** y **herencia** de forma correcta y justificada según el requerimiento.		

Andamiaje

La valoración por cada criterio está dada con niveles de ponderación de 1-5: los niveles (1-2) indican que el estudiante ha hecho poco o nada en la tarea, y los niveles (3-5) reflejan que el criterio ha sido completado en su mayoría o en su totalidad adecuadamente.

El nivel de ponderación se designa considerando la descripción de retroalimentación textual

Retroalimentación	Nivel de ponderación (Puntuación Numérica)
Tiene todos los aspectos evaluados de manera incorrecta y no ha realizado los requerimientos especificados.	1 (Nada adecuado)
Tiene todos los aspectos evaluados de manera incorrecta.	2 (Poco adecuado)
Tiene aspectos evaluados de manera correcta e incorrecta.	3 (Adecuado)
Tiene aspectos evaluados de manera correcta, sin embargo, puede tener un aspecto incompleto.	4 (Bastante adecuado)
Tiene todos los aspectos evaluados de manera correcta.	5 (Totalmente adecuado)

- En cada actividad, el docente explicó la tarea, los tiempos de cada etapa de la evaluación entre pares, los criterios de la rúbrica y el andamiaje.

Envío de tareas

- Cada grupo realizó la tarea y la envió a la plataforma virtual.

Asignación del evaluador

- Se utilizaron revisiones anónimas, el docente codificó cada tarea (código evaluado) con las siglas de la asignatura + código de actividad + letra del alfabeto (p. ej., FIS02M), luego asignó tres tareas a cada estudiante, verificando que ningún grupo se evaluara a sí mismo.

Evaluación de tareas

- Cada estudiante evaluó individualmente las tres tareas asignadas.
- El estudiante y el docente evaluaron cada tarea de manera objetiva y ética, de acuerdo con la rúbrica dada y proporcionaron puntuación numérica y retroalimentación textual para cada criterio.
- El estudiante envió cada evaluación (rúbrica) con el código evaluado + código estudiante (ej. FIS02M-AR9371) a la plataforma virtual.

- En este paso se obtuvo un conjunto de datos (D1) con 4088 instancias y un conjunto de datos (D2) con 820 instancias (Tabla **29**).

Tabla 29. Datos por asignaturas

Conjunto de datos	Asignatura	Grupo	Actividad	Tarea	Total de instancias			
					Recopilación de datos		Limpieza de datos	
D1	Fundamentos de ingeniería de software	17	1	Ejercicios de diagrama de casos de uso	924	4,088	912	3,968
		16	2	Ejercicios de diagrama de casos de uso	794		764	
		16	3	Ejercicios de diagrama clases	790		764	
		16	4	Ejercicios de diagrama de secuencia	790		764	
		16	5	Ejercicios de diagrama de actividad	790		764	
D2	Administración de base de datos	13	6	Ejercicios de sentencias SQL	820	820	759	759

Paso 3. Limpieza de datos

- Se descartó registros duplicados y vacíos en puntuación numérica o retroalimentación de los conjuntos de datos.
- En este paso, se redujo el conjunto de datos (D1) en 3968 instancias para el entrenamiento del modelo y el conjunto de datos (D2) en 759 instancias para la evaluación del modelo (Tabla **29**).

Paso 4. Etiquetado de datos

- Un anotador (docente de la asignatura de fundamentos de ingeniería del software) etiquetó cada retroalimentación del conjunto de datos (D1) (Tabla **29**), considerando las reglas de la subsección 3.5. En este paso, se obtuvo para el conjunto de datos (D1) una nueva variable (Polaridad Sentimiento). La Tabla **30** muestra ejemplos de retroalimentación en español etiquetadas. La retroalimentación (F1) fue etiquetada (-1, negativa) porque contiene las palabras “poco adecuada”. La retroalimentación (F2) se etiquetó (1, positivo) porque contiene

"especificados correctamente". Las retroalimentaciones (F3 y F4) se etiquetaron (-1, negativo) porque contiene las palabras "no es correcta".

Tabla 30. Ejemplos de retroalimentación en español etiquetadas de la actividad-2 (ejercicios de diagrama de casos de uso)

Evaluado	Evaluador	Criterio	Retroalimentación	Polaridad Sentimiento
FIS02M	AI7915	Diseño	F1: La estructura es *poco adecuada* para la relación entre actores y casos de uso	-1
		Actores	F2: Los actores están *especificados correctamente* en el sistema y son necesarios para su desarrollo	1
		Casos de uso	F3: La funcionalidad *no* es correcta en algunos casos de uso. Los casos de uso *no* tienen numeración y no están completos	-1
		Comunicación	F4: El extend en un caso de uso *no* es correcto	-1

- La Tabla **31** muestra el detalle de los conjuntos de datos para la experimentación.

Tabla 31. Detalle de conjuntos de datos para el experimento II

Conjunto de datos	Asignatura	Total de instancias	N° de retroalimentaciones positivas	N° de retroalimentaciones negativas	
D1	Fundamentos de ingeniería de software	3,968	2,881	1,687	Conjunto de datos para entramiento del modelo
D2	Administración de base de datos	759	-	-	Conjunto de datos para la evaluación del modelo

Paso 5. Ingeniería de características

- Una vez que se obtuvieron los conjuntos de datos, se preparó un CSV con el conjunto de datos (D1) (Tabla **30**) con las variables de entrada (Retroalimentación y Polaridad Sentimiento) para el entrenamiento del modelo (Tabla **32**).

Tabla 32. Algunos ejemplos del conjunto de datos (D1) en idioma español para el entrenamiento del modelo

Código-Actividad	Código-Evaluador	Código-Evaluado	Criterio	Puntuación Numérica	Retroalimentación	Polaridad Sentimiento
A02	AI7915	FIS02M	Diseño	2	La estructura es poco adecuada para la relación entre actores y casos de uso.	-1
A02	AI7915	FIS02M	Actores	5	Los actores están especificados en el sistema y son necesarios para su desarrollo.	1
A02	AI7915	FIS02M	Casos de uso	3	La funcionalidad no es correcta en algunos casos de uso. Los casos de uso no tienen numeración y no están completos.	-1
A02	AI7915	FIS02M	Comunicación	2	El extend en un caso de uso no es correcto.	-1
A02	AR9371	FIS02M	Diseño	2	La estructura no es correcta.	-1
A02	AR9371	FIS02M	Actores	5	Los actores están bien especificados.	1
A02	AR9371	FIS02M	Casos de uso	1	La funcionalidad descrita no es correcta, la sintaxis no está bien especificada y faltan casos de uso.	-1
A02	AR9371	FIS02M	Comunicación	1	La forma en que se usa el extend, include y herencia no es de forma correcta y no justifica totalmente las funcionalidades.	-1

- En el preprocesamiento, se consideró la tokenización y la normalización.
- Para la extracción de características, con aprendizaje automático, se usó combinaciones de N-Grams + TF-IDF con o sin Stop-Words; y con el aprendizaje profundo, usamos Word Embeddings del mismo conjunto de datos.
  - Los modelos de N-Grams se formaron de diferentes ordenes (1-g, 2-g, 3-g y 4-g), y combinaciones (1-g + 2-g, 1-g + 2-g + 3-g, 1-g + 2-g + 3-g + 4-g, 2-g + 3-g, 2-g + 3-g + 4-g, 3-g + 4-g).
  - Se aplicó el archivo Stopwords.txt, creado en esta investigación (Figura **21)**.
  - Se creó representaciones vectoriales basadas en el peso TF-IDF.
- En este paso, se obtuvo un vector de palabras para ejecutar los algoritmos.

Paso 6. Entrenamiento modelo

- Para cada experimento, se utilizó el método de retención. El corpus se dividió aleatoriamente, el 80% de los datos para el conjunto de entrenamiento y el 20% restante para el conjunto de pruebas.
- Se ejecutó algoritmos de aprendizaje automático (MNB, SVM, LR, RF, DT y VE) y aprendizaje profundo (LSTM y Bi-LSTM), y se ajustó parámetros para mejorar el rendimiento de los modelos. Estos pasos se repitieron con todos los modelos, volviendo a la ingeniería de características. La Tabla **33** muestra la configuración de los parámetros de los algoritmos de aprendizaje automático. La Tabla **34** muestra la configuración de los parámetros de los algoritmos de aprendizaje profundo, y la Tabla **35** muestra la arquitectura aplicada en el algoritmo Bi-LSTM. Las mejores configuraciones se resaltan en color azul claro.

Tabla 33. Configuración de parámetros de algoritmos de aprendizaje automático

Parámetro	**Modelo-1**						**Modelo-2**					
	MNB	**SVM**	**LR**	**RF**	**DT**	**VE**	**MNB**	**SVM**	**LR**	**RF**	**DT**	**VE**
Alpha	1.0						0.1					
C		1.0	1.0					1.0	2.0			
Kernel		Sigmoid						Linear				
Probability		True						True				
Solver			lbfgs						SAGA			
N Jobs			4	4					4	4		
N Estimators				100						100		
Criterion				Gini	Entropy					Entropy	Gini	
Max Depth				50	10					50	10	
Splitter					Best						Best	
Voting						Hard/ Soft						Hard/ Soft

Tabla 34. Configuración de parámetros de algoritmos de aprendizaje profundo

Parámetro	Valor
Model	Sequential
Num_words	10000
Maxlen	130
Units	16
Dropout	0.4
Dense	2
Activation	Softmax
Optimizer	Adam
Loss	Binary_crossentropy
Epochs	15
Learning rate	0.001
Batch size	15
Validation_split	0.2

Tabla 35. Arquitectura Bi-LSTM

Model: "sequential"		
Layer (type)	Output Shape	Param #
Embedding	(None, 130, 100)	273200
Bidirectional	(None, 32)	14976
Dropout	(None, 32)	0
Dense	(None, 2)	66
Total params: 288,242 Trainable params: 15,042 Non-trainable params: 273,200		

Paso 7. Evaluación del modelo

- Al ejecutar cada modelo se obtuvieron los valores de las métricas de evaluación (Precision, Recall, F-Measure y Accuracy).
- En este paso, se obtuvo el modelo predictivo con mejor rendimiento para generar la puntuación de sentimiento de cada actividad de evaluación por pares.
- Con el modelo predictivo, se genera la puntuación de sentimiento de cada una de las actividades de evaluación por pares y se determinó:
  - Medir la concordancia entre la puntuación de sentimiento que generó el modelo predictivo y la polaridad de sentimiento que proporcionó el anotador, utilizando los coeficientes de Kappa, Pearson y Spearman;
  - Evaluar la portabilidad del modelo predictivo entre asignaturas utilizando el conjunto de datos (D2) de texto no etiquetado (Tabla **31**).

- Medir la similitud entre la puntuación numérica y de sentimiento utilizando los coeficientes de Pearson y Spearman.

Segunda fase: Detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo puntuación de evaluación por cada criterio, de todos los criterios por evaluador o grupo mediante lógica difusa

Una vez que se obtuvo el modelo predictivo para generar la puntuación de sentimiento (1-positivo ó -1-negativo), se realizó la segunda fase considerando la subsección 3.6, para detectar precisión/imprecisión entre puntuación de sentimiento y numérica. Luego se obtuvo la puntuación de evaluación para cada criterio, con todos los criterios por evaluador (puntuación individual) o grupo (puntuación del colectivo), considerando el enfoque de Mamdani en lógica difusa descrito. Los cuatro pasos se resumen a continuación:

Paso 1. Identificación del valor crujiente

- Para obtener la puntuación de evaluación por cada criterio, se seleccionó la Puntuación Numérica de las variables de entrada que contiene valores de 1 a 5, y la Puntuación Sentimiento que contiene valores 1 o -1, generados por el modelo predictivo (Tabla **36**). Para obtener la puntuación de evaluación con todos los criterios por evaluador o grupo, se consideró como variable de entrada la media de las puntuaciones tanto numéricas como de sentimiento. En este caso, la media de una puntuación de sentimiento generó un 0-neutral.

Tabla 36. Algunos ejemplos del conjunto de datos (D1) en idioma español con puntuación numérica y de sentimiento generado por el modelo predictivo a partir de la retroalimentación textual para la detección de precisión/imprecisión y cálculo de puntuación de evaluación

Código-Actividad	Código-Evaluador	Código-Evaluado	Criterio	Puntuación Numérica	Puntuación Sentimiento
A02	FIS02M	AI7915	Diseño	2	-1
			Actores	5	1
			Casos de uso	3	-1
			Comunicación	2	-1
		AR9371	Diseño	2	-1
			Actores	5	1
			Casos de uso	1	-1
			Comunicación	1	-1
		AS1826	Diseño	5	1
			Actores	5	1
			Casos de uso	4	-1
			Comunicación	4	-1

Paso 2. Fuzzificación del valor de entrada y salida

- En el proceso de fuzzificación, los valores nítidos de las variables de entrada y salida se convierten en conjuntos difusos. Por tanto, para la variable de entrada (Puntuación Numérica) y la variable de salida (Puntuación Evaluación), se definió el universo de discurso con rangos entre 1-5, y se dividió en términos lingüísticos: 1-Nada adecuado, 2-Poco adecuado, 3-Adecuado, 4-Bastante adecuado y 5-Totalmente adecuado; para la variable de entrada (Puntuación Sentimiento), se definió el universo de discurso con rangos entre 1 y -1, y se dividió en términos lingüísticos: 1-positivo, -1-negativo y 0-neutral. Luego, se formaron las funciones de pertenencia, asignando los parámetros apropiados a los respectivos términos lingüísticos (Tabla **44** y Figura **38**).

Paso 3. Determinación de las reglas de aplicación

- Para obtener la puntuación de evaluación por cada criterio, se formuló reglas difusas (Figura **39** y Figura **40**):

  Detección de precisión y generación de puntuación de evaluación

  1. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
  2. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es poco adecuado)
  3. Si (puntuación numérica es adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es adecuado)
  4. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es bastante adecuado)
  5. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es totalmente adecuado)

  Detección de imprecisión y generación de puntuación de evaluación

  6. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es nada adecuado)
  7. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es nada adecuado)

8. Si (puntuación numérica es adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
9. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
10. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)

- Para obtener la puntuación de evaluación con todos los criterios por evaluador (puntuación individual) o grupo (puntuación del colectivo), se formuló reglas difusas:

  Detección de precisión y generación de puntuación de evaluación.

  1. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
  2. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es poco adecuado)
  3. Si (puntuación numérica es adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es adecuado)
  4. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es bastante adecuado)
  5. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es totalmente adecuado)

  Detección de imprecisión y generación de puntuación de evaluación.

  6. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es nada adecuado)
  7. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es nada adecuado)
  8. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es nada adecuado)
  9. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es nada adecuado)
  10. Si (puntuación numérica es adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
  11. Si (puntuación numérica es adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es nada adecuado)

12. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
13. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es nada adecuado)
14. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
15. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es nada adecuado)

Paso 4. Desfuzzificación del valor de desarrollo

- En cada experimento, se probó los métodos de Centroid, Bisector, SOM, MOM y LOM (Figura **41**, Tabla **45**, Figura **42**, Tabla **46** y Figura **43**).
- En este paso se obtuvo la puntuación de evaluación de cada criterio, de todos los criterios por evaluador o grupo.

Materiales

Para implementar el experimento, se usó:

- Bibliotecas de Python que se ejecutaron en Jupyter Notebook v. 6.1.4:
- NumPy para la vectorización de columnas en forma de panda.
- Pandas para leer CSV.
- Métricas de SK-Learn para comparación de modelos.
- NLTK para procesamiento de texto.
- Pickle para guardar y cargar modelos de aprendizaje automático.
- Cadena, unidecode y puntuación para agregar vocabulario Stop-Words.
- Skfuzzy para aplicar lógica difusa.
- Software de análisis de datos (SPSS) v. 25 para análisis descriptivo.

Resultados

En esta sección se muestran los resultados obtenidos en cada fase.

Análisis de sentimiento

En esta fase, se obtuvo un modelo predictivo con el mejor rendimiento para generar una puntuación de sentimiento (1-positivo ó -1-negativo) que corresponde a retroalimentación textual específica (Figura **35**), utilizando el enfoque de aprendizaje supervisado con algoritmos de aprendizaje automático. (MNB, SVM, LR, RF, DT y VE) y aprendizaje profundo (LSTM y Bi-LSTM) a través de bibliotecas de Python.

Ingeniería de características, entrenamiento de modelos

Para el entrenamiento de los modelos, se utilizó el conjunto de datos (D1) (Tabla **31**). Se preprocesó la retroalimentación, aplicando tokenización y normalización (corrección de errores ortográficos, conversión a minúsculas y tratamiento de caracteres especiales UTF-8).

En primer lugar, se ejecutó modelos de aprendizaje automático. Se probó el rendimiento de MNB, SVM, LR y RF, los resultados tempranos mostraron que el modelo SVM con función lineal tenía el mejor rendimiento, que ha sido publicado en [297]. Posteriormente, se evaluó los mismos algoritmos y los algoritmos DT y VE, ajustando parámetros para mejorar el rendimiento de los modelos (Tabla **33**), y se realizó con cada algoritmo veinte pruebas, enfocadas en encontrar la combinación óptima de N-Grams con TF-IDF y Stop-Words (Tabla **37**, Tabla **38**).

Tabla 37. Rendimiento de algoritmos de aprendizaje automático con configuración de parámetros del modelo-1 (Tabla 28)

Prueba	Modelo	MNB				SVM				LR				RF			
		Macro promedio															
		P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A
1	1-g+TF-IDF	0.756	0.711	0.716	0.743	0.859	0.863	0.861	0.864	0.867	0.866	0.866	0.870	0.859	0.858	0.859	0.863
2	2-g+TF-IDF	0.814	0.738	0.745	0.775	0.843	0.840	0.841	0.846	0.839	0.827	0.831	0.839	0.801	0.776	0.782	0.796
3	3-g+TF-IDF	0.777	0.700	0.702	0.741	0.786	0.782	0.783	0.791	0.769	0.732	0.738	0.759	0.744	0.648	0.638	0.698
4	4-g+TF-IDF	0.729	0.650	0.643	0.696	0.725	0.693	0.696	0.722	0.700	0.632	0.623	0.679	0.713	0.568	0.515	0.636
5	1-g+2-g+TF-IDF	0.795	0.719	0.723	0.757	0.874	0.876	0.875	0.878	0.867	0.865	0.866	0.870	0.844	0.835	0.838	0.845
6	1-g+2-g+3-g+TF-IDF	0.816	0.724	0.729	0.764	0.870	0.871	0.870	0.874	0.859	0.855	0.857	0.861	0.834	0.821	0.826	0.834
7	1-g+2-g+3-g+4-g+TF-	0.807	0.707	0.709	0.751	0.871	0.873	0.872	0.875	0.851	0.847	0.849	0.854	0.813	0.793	0.799	0.810
8	2-g+3-g+TF-IDF	0.811	0.726	0.732	0.766	0.840	0.841	0.841	0.845	0.830	0.813	0.818	0.827	0.796	0.762	0.769	0.786
9	2-g+3-g+4-g+TF-IDF	0.803	0.716	0.720	0.757	0.834	0.834	0.834	0.839	0.815	0.791	0.798	0.810	0.790	0.740	0.747	0.771
10	3-g+4-g+TF-IDF	0.766	0.689	0.690	0.730	0.784	0.773	0.777	0.787	0.759	0.715	0.720	0.746	0.734	0.634	0.620	0.686
11	1-g+TF-IDF+SW	0.749	0.713	0.718	0.742	0.856	0.858	0.857	0.860	0.867	0.869	0.868	0.872	0.868	0.866	0.867	0.872
12	2-g+TF-IDF+SW	0.797	0.763	0.770	0.787	0.805	0.803	0.804	0.810	0.817	0.801	0.806	0.816	0.772	0.721	0.726	0.753
13	3-g+TF-IDF+SW	0.758	0.703	0.707	0.738	0.771	0.740	0.746	0.764	0.761	0.694	0.696	0.733	0.709	0.589	0.554	0.651
14	4-g+TF-IDF+SW	0.723	0.624	0.607	0.678	0.720	0.645	0.638	0.691	0.747	0.601	0.567	0.664	0.770	0.542	0.457	0.618
15	1-g+2-g+TF-IDF+SW	0.808	0.762	0.770	0.790	0.871	0.874	0.872	0.875	0.860	0.859	0.860	0.864	0.844	0.838	0.841	0.846
16	1-g+2-g+3-g+TF-	0.808	0.753	0.760	0.783	0.873	0.876	0.875	0.878	0.852	0.849	0.850	0.855	0.831	0.809	0.816	0.826
17	1-g+2-g+3-g+4-g+TF-	0.812	0.746	0.753	0.780	0.867	0.871	0.869	0.872	0.850	0.850	0.850	0.854	0.825	0.798	0.805	0.817
18	2-g+3-g+TF-IDF+SW	0.808	0.770	0.777	0.795	0.826	0.822	0.824	0.830	0.807	0.786	0.792	0.804	0.766	0.702	0.705	0.739
19	2-g+3-g+4-g+TF-	0.794	0.750	0.758	0.778	0.832	0.829	0.831	0.836	0.799	0.773	0.779	0.793	0.759	0.691	0.693	0.730
20	3-g+4-g+TF-IDF+SW	0.744	0.679	0.680	0.719	0.759	0.720	0.726	0.749	0.756	0.666	0.662	0.713	0.736	0.584	0.540	0.650

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, SW= Stop-Words

*En cada algoritmo, las mejores pruebas se resaltan en azul claro, las peores en verde y los mejores resultados en naranja.

Continuación de Tabla 37

Prueba	Modelo	DT				VE (H)				VE (S)			
		Macro promedio											
		P	R	F1	A	P	R	F1	A	P	R	F1	A
1	1-g+TF-IDF	0.830	0.823	0.826	0.832	0.867	0.865	0.866	0.870	0.869	0.869	0.869	0.873
2	2-g+TF-IDF	0.739	0.677	0.677	0.717	0.838	0.814	0.821	0.831	0.827	0.810	0.816	0.825
3	3-g+TF-IDF	0.710	0.534	0.448	0.611	0.775	0.714	0.718	0.749	0.791	0.764	0.771	0.786
4	4-g+TF-IDF	0.670	0.510	0.396	0.592	0.715	0.633	0.622	0.683	0.718	0.681	0.683	0.713
5	1-g+2-g+TF-IDF	0.840	0.837	0.839	0.844	0.862	0.854	0.857	0.863	0.867	0.862	0.864	0.869
6	1-g+2-g+3-g+TF-IDF	0.823	0.821	0.822	0.827	0.867	0.856	0.861	0.866	0.867	0.862	0.864	0.869
7	1-g+2-g+3-g+4-g+TF-IDF	0.802	0.802	0.802	0.807	0.854	0.843	0.847	0.854	0.859	0.854	0.856	0.861
8	2-g+3-g+TF-IDF	0.747	0.682	0.682	0.722	0.841	0.810	0.818	0.830	0.834	0.814	0.820	0.830
9	2-g+3-g+4-g+TF-IDF	0.749	0.689	0.691	0.727	0.824	0.788	0.796	0.811	0.835	0.813	0.820	0.830
10	3-g+4-g+TF-IDF	0.752	0.538	0.450	0.615	0.763	0.697	0.700	0.736	0.771	0.749	0.754	0.770
11	1-g+TF-IDF+SW	0.855	0.840	0.845	0.853	0.868	0.867	0.868	0.872	0.874	0.871	0.872	0.877
12	2-g+TF-IDF+SW	0.721	0.650	0.645	0.695	0.814	0.797	0.802	0.812	0.818	0.805	0.810	0.819
13	3-g+TF-IDF+SW	0.726	0.544	0.465	0.618	0.757	0.686	0.687	0.727	0.769	0.730	0.736	0.758
14	4-g+TF-IDF+SW	0.794	0.509	0.388	0.592	0.748	0.598	0.561	0.661	0.727	0.660	0.657	0.703
15	1-g+2-g+TF-IDF+SW	0.837	0.837	0.837	0.841	0.869	0.865	0.867	0.872	0.876	0.875	0.875	0.879
16	1-g+2-g+3-g+TF-IDF+SW	0.829	0.830	0.829	0.834	0.855	0.851	0.853	0.858	0.873	0.873	0.873	0.877
17	1-g+2-g+3-g+4-g+TF-IDF+SW	0.829	0.835	0.831	0.834	0.853	0.850	0.852	0.856	0.870	0.870	0.870	0.874
18	2-g+3-g+TF-IDF+SW	0.731	0.659	0.655	0.703	0.810	0.785	0.792	0.805	0.825	0.809	0.815	0.824
19	2-g+3-g+4-g+TF-IDF+SW	0.729	0.652	0.646	0.698	0.796	0.766	0.772	0.788	0.823	0.809	0.814	0.822
20	3-g+4-g+TF-IDF+SW	0.729	0.545	0.468	0.620	0.746	0.656	0.650	0.704	0.752	0.714	0.719	0.743

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, SW= Stop-Words

*En cada algoritmo, las mejores pruebas se resaltan en azul claro, las peores en verde y los mejores resultados en naranja.

Tabla 38. Rendimiento de algoritmos de aprendizaje automático con configuración de parámetros del modelo-2 (Tabla 28)

Prueba	Modelo	MNB				SVM				LR				RF			
		Macro promedio															
		P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A
1	1-g+TF-IDF	0.721	0.712	0.715	0.728	0.863	0.866	0.865	0.868	0.869	0.870	0.869	0.873	0.866	0.869	0.867	0.870
2	2-g+TF-IDF	0.771	0.764	0.766	0.776	0.844	0.842	0.843	0.848	0.845	0.837	0.840	0.846	0.805	0.775	0.782	0.797
3	3-g+TF-IDF	0.762	0.750	0.754	0.766	0.786	0.782	0.783	0.791	0.795	0.778	0.783	0.795	0.732	0.633	0.618	0.685
4	4-g+TF-IDF	0.697	0.670	0.673	0.699	0.717	0.695	0.698	0.719	0.713	0.670	0.672	0.705	0.721	0.579	0.533	0.645
5	**1-g+2-g+TF-IDF**	0.778	0.769	0.773	0.782	0.877	0.881	**0.879**	0.882	0.870	0.871	0.871	0.874	0.844	0.841	0.842	0.848
6	1-g+2-g+3-g+TF-IDF	0.787	0.779	0.782	0.791	0.872	0.875	0.873	0.877	0.870	0.868	0.869	0.873	0.838	0.828	0.832	0.839
7	1-g+2-g+3-g+4-g+TF-IDF	0.792	0.782	0.785	0.795	0.875	0.879	0.876	0.879	0.854	0.852	0.853	0.858	0.823	0.797	0.804	0.816
8	2-g+3-g+TF-IDF	0.789	0.780	0.783	0.792	0.847	0.848	0.847	0.851	0.834	0.824	0.828	0.835	0.792	0.751	0.758	0.778
9	2-g+3-g+4-g+TF-IDF	0.785	0.775	0.778	0.788	0.837	0.837	0.837	0.841	0.825	0.812	0.817	0.825	0.792	0.740	0.746	0.771
10	3-g+4-g+TF-IDF	0.755	0.741	0.745	0.758	0.796	0.789	0.792	0.800	0.769	0.744	0.749	0.766	0.729	0.623	0.604	0.678
11	1-g+TF-IDF+SW	0.721	0.713	0.716	0.728	0.871	0.873	0.872	0.875	0.875	0.877	0.876	0.879	0.867	0.869	0.868	0.872
12	2-g+TF-IDF+SW	0.767	0.763	0.765	0.773	0.812	0.811	0.812	0.817	0.824	0.818	0.821	0.827	0.768	0.722	0.728	0.753
13	3-g+TF-IDF+SW	0.753	0.732	0.737	0.753	0.774	0.754	0.759	0.773	0.763	0.721	0.726	0.751	0.733	0.594	0.558	0.657
14	4-g+TF-IDF+SW	0.710	0.657	0.655	0.696	0.716	0.662	0.661	0.702	0.734	0.643	0.633	0.693	0.752	0.538	0.450	0.615
15	1-g+2-g+TF-IDF+SW	0.781	0.780	0.780	0.787	0.870	0.873	0.871	0.874	0.867	0.869	0.868	0.872	0.852	0.849	0.850	0.855
16	1-g+2-g+3-g+TF-IDF+SW	0.795	0.790	0.792	0.800	0.866	0.868	0.867	0.870	0.863	0.866	0.865	0.868	0.819	0.805	0.809	0.819
17	1-g+2-g+3-g+4-g+TF-IDF+SW	0.801	0.794	0.797	0.805	0.864	0.869	0.866	0.869	0.857	0.860	0.858	0.861	0.831	0.801	0.809	0.821
18	2-g+3-g+TF-IDF+SW	0.789	0.783	0.785	0.793	0.829	0.828	0.828	0.834	0.830	0.820	0.824	0.831	0.761	0.701	0.705	0.738
19	2-g+3-g+4-g+TF-IDF+SW	0.789	0.783	0.785	0.793	0.826	0.823	0.824	0.830	0.819	0.807	0.811	0.820	0.769	0.688	0.688	0.730
20	3-g+4-g+TF-IDF+SW	0.757	0.736	0.741	0.757	0.761	0.738	0.743	0.759	0.755	0.700	0.704	0.736	0.744	0.588	0.546	0.654

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, SW= Stop-Words

*En cada algoritmo, las mejores pruebas se resaltan en azul claro, las peores en verde y los mejores resultados en naranja.

Continuación de Tabla 38

Prueba	Modelo	DT				VE(H)				VE(S)			
		Macro promedio											
		P	R	F1	A	P	R	F1	A	P	R	F1	A
1	1-g+TF-IDF	0.831	0.833	0.832	0.836	0.863	0.866	0.864	0.868	0.872	0.872	0.872	0.875
2	2-g+TF-IDF	0.747	0.686	0.687	0.724	0.839	0.827	0.831	0.839	0.832	0.821	0.825	0.832
3	3-g+TF-IDF	0.699	0.538	0.458	0.613	0.783	0.752	0.758	0.776	0.801	0.785	0.790	0.801
4	4-g+TF-IDF	0.660	0.518	0.416	0.597	0.717	0.670	0.671	0.707	0.712	0.683	0.686	0.712
5	1-g+2-g+TF-IDF	0.848	0.842	0.844	0.850	0.873	0.873	0.873	0.877	0.882	0.878	0.880	0.884
6	1-g+2-g+3-g+TF-IDF	0.829	0.828	0.829	0.834	0.871	0.870	0.870	0.874	0.874	0.874	0.874	0.878
7	1-g+2-g+3-g+4-g+TF-IDF	0.835	0.838	0.836	0.840	0.869	0.866	0.867	0.872	0.865	0.866	0.865	0.869
8	2-g+3-g+TF-IDF	0.749	0.687	0.688	0.725	0.837	0.826	0.830	0.838	0.837	0.829	0.832	0.839
9	2-g+3-g+4-g+TF-IDF	0.765	0.693	0.694	0.733	0.832	0.816	0.821	0.830	0.840	0.828	0.833	0.840
10	3-g+4-g+TF-IDF	0.696	0.537	0.455	0.612	0.773	0.739	0.745	0.764	0.786	0.768	0.773	0.786
11	1-g+TF-IDF+SW	0.849	0.847	0.848	0.853	0.869	0.869	0.869	0.873	0.873	0.873	0.873	0.877
12	2-g+TF-IDF+SW	0.728	0.659	0.656	0.703	0.817	0.807	0.811	0.819	0.813	0.805	0.809	0.816
13	3-g+TF-IDF+SW	0.731	0.547	0.471	0.621	0.755	0.708	0.713	0.741	0.770	0.747	0.753	0.768
14	4-g+TF-IDF+SW	0.796	0.514	0.398	0.596	0.731	0.636	0.624	0.688	0.711	0.660	0.659	0.699
15	**1-g+2-g+TF-IDF+SW**	0.840	0.843	0.841	0.845	0.875	0.876	0.876	0.879	0.881	0.884	**0.882**	0.885
16	1-g+2-g+3-g+TF-IDF+SW	0.835	0.841	0.837	0.840	0.863	0.865	0.864	0.868	0.867	0.870	0.868	0.872
17	1-g+2-g+3-g+4-g+TF-IDF+SW	0.824	0.832	0.826	0.829	0.866	0.868	0.867	0.870	0.862	0.865	0.863	0.866
18	2-g+3-g+TF-IDF+SW	0.728	0.661	0.659	0.704	0.824	0.810	0.815	0.824	0.831	0.822	0.825	0.832
19	2-g+3-g+4-g+TF-IDF+SW	0.729	0.669	0.668	0.709	0.816	0.800	0.805	0.815	0.821	0.814	0.817	0.824
20	3-g+4-g+TF-IDF+SW	0.733	0.548	0.473	0.622	0.750	0.692	0.695	0.729	0.764	0.738	0.743	0.761

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, SW= Stop-Words

El mejor rendimiento de cada uno de los algoritmos se detalla a continuación (Tabla **39**). El algoritmo MNB aumentó en F-Measure (0.797), Recall (0.794) y Precision (0.805) con un factor alfa de 0.1, y cuando la retroalimentación se representó como un vector de 1-g, 2-g, 3-g, y 4-g, cuya relevancia se ponderó con TF-IDF y se eliminó las Stop-Words, la Precision (0.801) fue menor (0.015) que con factor alfa de 1.0.

El algoritmo SVM obtuvo el mejor desempeño en F-Measure (0.879), Precision (0.877), Recall (0.881) y Accuracy (0.882), usando una función lineal con C (1.0), y cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF.

El algoritmo LR tuvo un mayor desempeño en F-Measure (0.876), Precision (0.875), Recall (0.877) y Accuracy (0.879), usando C (2.0), solver saga, y cuando la retroalimentación se representó como un vector de 1-g, cuya relevancia se ponderó con TF-IDF, y se eliminó las Stop-Words.

El algoritmo RF logró el mejor rendimiento en F-Measure (0.868), Recall (0.869) y Accuracy (0.872) cuando aplicamos un estimador de 100, criterio de impureza de entropía, un máximo de 50 niveles de profundidad y cuando se representó la retroalimentación como vector de 1-g, cuya relevancia se ponderó con TF-IDF, y fueron eliminados las Stop-Words, pero la Precision (0.867) fue menor (0.001) que con el criterio de impureza de gini.

El algoritmo DT obtuvo el mejor rendimiento tanto en F-Measure (0.848), Recall (0.847) y Accuracy (0.853), utilizando gini para medir la pureza de los nodos que contienen en su mayoría elementos de una sola clase y con un máximo de 10 niveles de profundidad, y cuando la retroalimentación se representó como un vector de 1-g, cuya relevancia se ponderó con TF-IDF, y se eliminaron las Stop-Words, pero la Precision (0.849) fue menor (0.006) que con el criterio de impureza de entropía.

El algoritmo VE con votación fuerte (VE (H)) obtuvo el mejor rendimiento en F-Measure (0.876), Precision (0.875), Recall (0.876) y Accuracy (0.879) cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF, y fueron eliminados las Stop-Words. El voto suave (VE(S)) obtuvo el mejor desempeño en Precision (0.882) cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF, pero el rendimiento aumentó en F-Measure (0.882), Recall (0.884) y Precision (0.885) también cuando se eliminó las Stop-Words. Este método de conjunto aumentó el rendimiento en comparación con el rendimiento individual de los cinco algoritmos base: MNB,

SVM, LR, RF y DT, pero con altos costos computacionales porque cada algoritmo base tiene un sesgo específico.

La Figura **36** detalla el rendimiento (F-Measure) de los algoritmos. La progresión de cada línea permite apreciar en qué medida el resultado de cada algoritmo aumenta o disminuye con las pruebas aplicadas (Tabla **38**). Las pruebas 1 a 4 usan una ponderación TF-IDF de 1 a 4-g, en el intervalo 5 a 10, la ponderación fue TF-IDF con combinaciones de 1 a 4-g, las pruebas 11 a 14 tienen una ponderación TF-IDF con 1 a 4-g y reducción de características, y en las pruebas 15 a 20, la ponderación fue TF-IDF con combinaciones de 1 a 4-g y reducción de características.

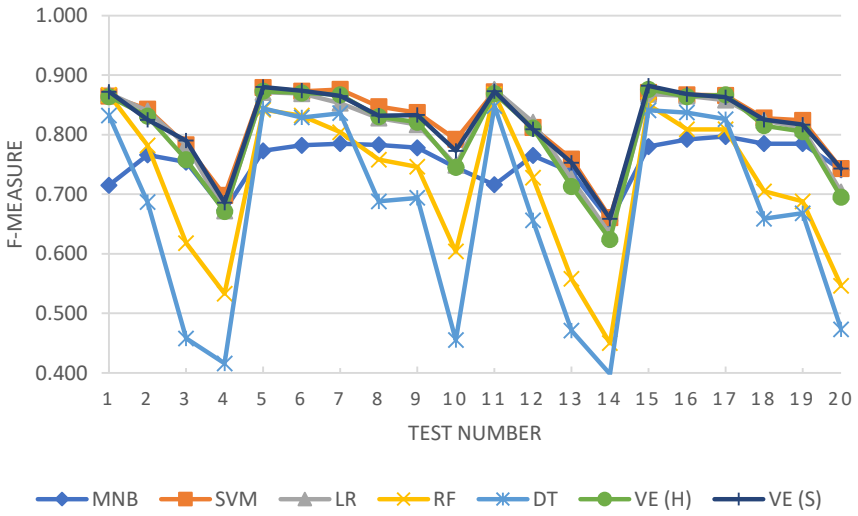


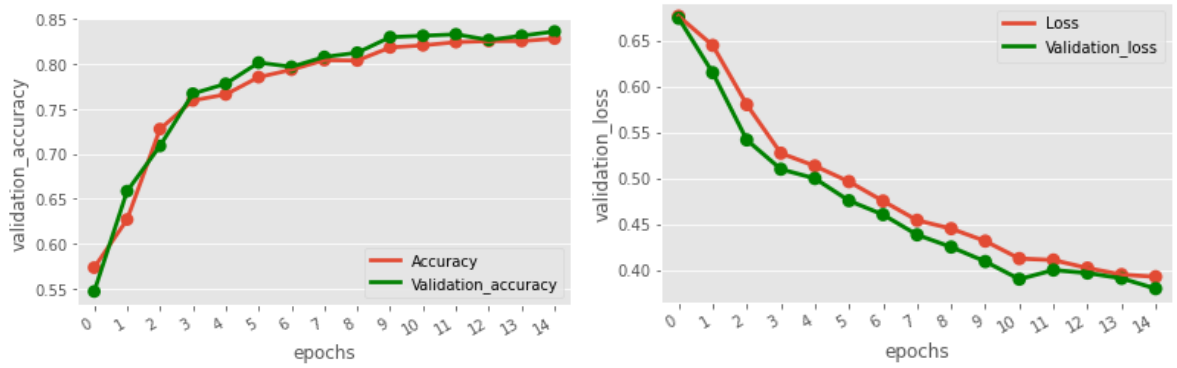
Figura 36. Rendimiento de modelos de aprendizaje automático

En las pruebas 11 a 18, los algoritmos que obtuvieron el mejor rendimiento al reducir las características fueron MNB, LR, RF, DT, VE (H) y VE (S), mientras que el rendimiento de SVM disminuyó. Esta diferencia se recuperó con el mejor uso de los recursos computacionales porque al eliminar las Stop-Words se redujo considerablemente la dimensión del espacio en el que se habían representado las opiniones [298].

Los mejores modelos se lograron en las pruebas 5 y 15 aplicando las configuraciones 1-g y 2-g, cuya relevancia se ponderó con TF-IDF, que mejor representó la información de retroalimentación en los algoritmos SVM y VE(S).

En las pruebas 4, 10, 13, 14 y 20 se obtuvieron peores resultados con el patrón de 3 a 4-g de todos los algoritmos.

En segundo lugar, se ejecutó modelos de aprendizaje profundo, que definieron el modelo de tipo secuencial (Tabla **34** y Tabla **35**), se usó una capa de incrustación con el conjunto de datos etiquetado, un máximo de 10 000 palabras en el vocabulario, una longitud máxima de 130, y una representación vectorial de 100. Los datos que se ingresaron a la capa LSTM/Bi-LSTM de 16 unidades, obtuvieron la capa de clasificación, compuesta por dos neuronas y softmax en función de la activación. Para evitar el sobreentrenamiento se aplicó un dropout de 0.4, optimizador (Adam), función de error (Binary Cross-Entropy) y una tasa de aprendizaje de 0.001 para 15 épocas (15 iteraciones en todas las muestras en mini-lotes de 15 muestras). Simultáneamente, también se monitoreó la pérdida y precisión de los datos de validación (Figura **37**). Los resultados mostraron que Bi-LSTM superó a LSTM, en F-Measure (0.837), Precision (0.853), Recall (0.831) y Accuracy (0.846), con altos costos computacionales (Tabla **39**).



(a) Precisión del entrenamiento vs validación (b) Pérdida del entrenamiento vs validación

Figura 37. Rendimiento del modelo de aprendizaje profundo (Bi-LSTM)

La Figura **37** muestra el entrenamiento versus el proceso de validación hasta alcanzar un alto nivel de precisión en diferentes momentos, y muestra la evolución del entrenamiento visto desde la perspectiva de la pérdida de información, tendiendo a cero en las épocas finales; es decir, inicialmente hubo altas tasas de pérdida en Bi-LSTM.

Los resultados experimentales muestran que tanto el modelo de aprendizaje automático como el de aprendizaje profundo tienen un aprendizaje igual o superior al 80% (Tabla **39**). Seleccionamos SVM (F-Measure de 0.879) como modelo predictivo porque obtuvo un buen desempeño con bajo costo computacional cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF. La diferencia no fue significativa

(0.003) con respecto a VE (F-Measure de 0.882), que obtuvo mejor rendimiento, pero con altos costos computacionales.

Tabla 39. Comparación del rendimiento de los modelos de aprendizaje automático y aprendizaje profundo

Algoritmo		**Macro promedio**			
		Precision	**Recall**	**Accuracy**	***F-Measure***
Aprendizaje automático	MNB	0.801	0.794	0.805	0.797
	SVM	0.877	0.881	0.882	***0.879***
	LR	0.875	0.877	0.879	0.876
	RF	0.867	0.869	0.872	0.868
	DT	0.849	0.847	0.853	0.848
	VE (H)	0.875	0.876	0.879	0.876
	VE (S)	0.882	0.884	0.885	0.882
Aprendizaje profundo	LSTM	0.841	0.828	0.840	0.832
	Bi-LSTM	0.853	0.831	0.846	0.837

Evaluación del modelo

En la evaluación del modelo, se comparó la concordancia entre la puntuación que el modelo SVM obtuvo automáticamente (puntuación de sentimiento) con la que el anotador (polaridad de sentimiento) dio a cada una de las retroalimentaciones del conjunto de datos (D1) (Tabla **31**) a través de coeficientes de Kappa, Pearson y Spearman. Los resultados mostraron que el modelo es confiable porque se obtuvo una concordancia igual o superior al 80% en todas las actividades, en todos los coeficientes (Tabla **40**). Es fundamental resaltar que el ser humano tiene un 80% de precisión en la clasificación de textos en análisis de sentimiento; esto dificulta medir la precisión de los sistemas de clasificación; la clasificación se considera aceptable cuando el porcentaje se acerca o supera lo que la humanidad puede hacer [299].

Tabla 40. Concordancia entre la puntuación de sentimiento que generó el modelo SVM y la polaridad de sentimiento que proporcionó el anotador

Conjunto de datos	**Asignatura**	**Actividad**	**Kappa**	**Pearson**	**Spearman**
D1	Fundamentos de ingeniería de software	1	0.877	0.878	0.878
		2	0.936	0.936	0.936
		3	0.892	0.893	0.893
		4	0.916	0.917	0.917
		5	0.922	0.922	0.922

Además, se evaluó la portabilidad del modelo predictivo entre asignaturas, con la actividad-6 del conjunto de datos (D2) (Tabla **31**). La verificación se realizó con las reglas de la subsección 3.5. La puntuación (−1, negativo) de la retroalimentación (F1) es correcta porque contiene la palabra “falta” y “errores”. La puntuación (−1, negativo) de la retroalimentación (F2 y F4) es correcta porque contiene la palabra "no hay", “no muestra”, “no tiene”. La puntuación (1, positivo) de la retroalimentación (F3) es correcto porque contiene la palabra “tiene” (Tabla **41**). Los resultados muestran que el modelo predictivo generó la puntuación de sentimiento correcto, por lo tanto, el modelo es portable para otra asignatura.

Finalmente, se analizó la similitud entre las puntuaciones numéricas y de sentimiento, que equivalían a una puntuación de retroalimentación (−1, negativo) con una puntuación numérica de 1 o 2, y una puntuación de retroalimentación (1, positivo) con una puntuación numérica de 3, 4, o 5. Los resultados muestran que las puntuaciones numéricas del evaluador no se corresponden con la retroalimentación textual (Tabla **41**).

Tabla 41. Algunos ejemplos de puntuación de sentimiento que generó el modelo SVM e inexactitud entre la puntuación numérica y de sentimiento detectadas en la actividad-6

Evaluado	Evaluador	Criterio	Retroalimentación	Puntuación Numérica	Puntuación Sentimiento
ABD01D	BL8705	Documento	F1: Falta el nombre de la materia y hay errores ortográficos	4	-1
		Estructura	F2: No hay un ordenamiento en todos los procesos presentados	3	-1
		Proceso	F3: Tiene secuencia lógica en algunos de los procesos	2	1
		Funcionalidad	F4: No muestra los resultados. Las soluciones no tienen una correcta funcionalidad de los requerimientos solicitados	2	-1

Del mismo modo, se realizó una correlación entre las puntuaciones numéricas y de sentimiento en todas las actividades utilizando los coeficientes de Pearson y Spearman. Los resultados muestran algunas imprecisiones entre las puntuaciones numéricas y de sentimiento en todas las actividades (Tabla **42**). Los evaluadores fueron más imprecisos al asignar una puntuación numérica de 3 o 4 a la retroalimentación textual argumentada (Tabla **43**).

Tabla 42. Correlación entre puntuación numérica y de sentimiento en todas las actividades

Conjunto de datos	Asignatura	Actividad	Kappa	Pearson	Spearman
D1	Fundamentos de ingeniería de software	1	0.877	0.878	0.878
		2	0.936	0.936	0.936
		3	0.892	0.893	0.893
		4	0.916	0.917	0.917
		5	0.922	0.922	0.922
D2	Administración de base de datos	6	0.877	0.878	0.878

Tabla 43. Inexactitudes detectadas entre puntuación numérica y de sentimiento en todas las actividades

	Actividad											
	1		**2**		**3**		**4**		**5**		**6**	
Puntuación Numérica	**Puntuación Sentimiento**											
	-1	**1**	**-1**	**1**	**-1**	**1**	**-1**	**1**	**-1**	**1**	**-1**	**1**
1	27	2	6	0	51	1	43	1	22	0	30	7
2	110	17	54	11	91	10	100	1	59	5	38	12
3	164	*59*	103	*58*	149	*66*	145	*63*	104	*48*	121	*35*
4	*140*	158	52	147	*73*	176	*101*	116	*75*	185	*146*	110
5	4	231	6	327	0	147	2	192	2	264	10	250
Total	445	467	221	543	364	400	391	373	262	502	345	414

Detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación por cada criterio, de todos los criterios por evaluador o grupo mediante lógica difusa

Para mejorar la confiabilidad del procedimiento de evaluación por pares cuando las evaluaciones son inconsistentes entre sí (por ejemplo, un evaluador asigna una puntuación numérica alta con retroalimentación que detallan que todo es incorrecto o viceversa), en esta fase, se detecta precisión/imprecisión entre la puntuación numérica y puntuación de sentimiento generada del modelo predictivo a partir de la retroalimentación textual, y se genera la puntuación de evaluación final (Figura **35**), utilizando el enfoque de Mamdani en lógica difusa a través de las bibliotecas de Python.

Valor nítido (datos)

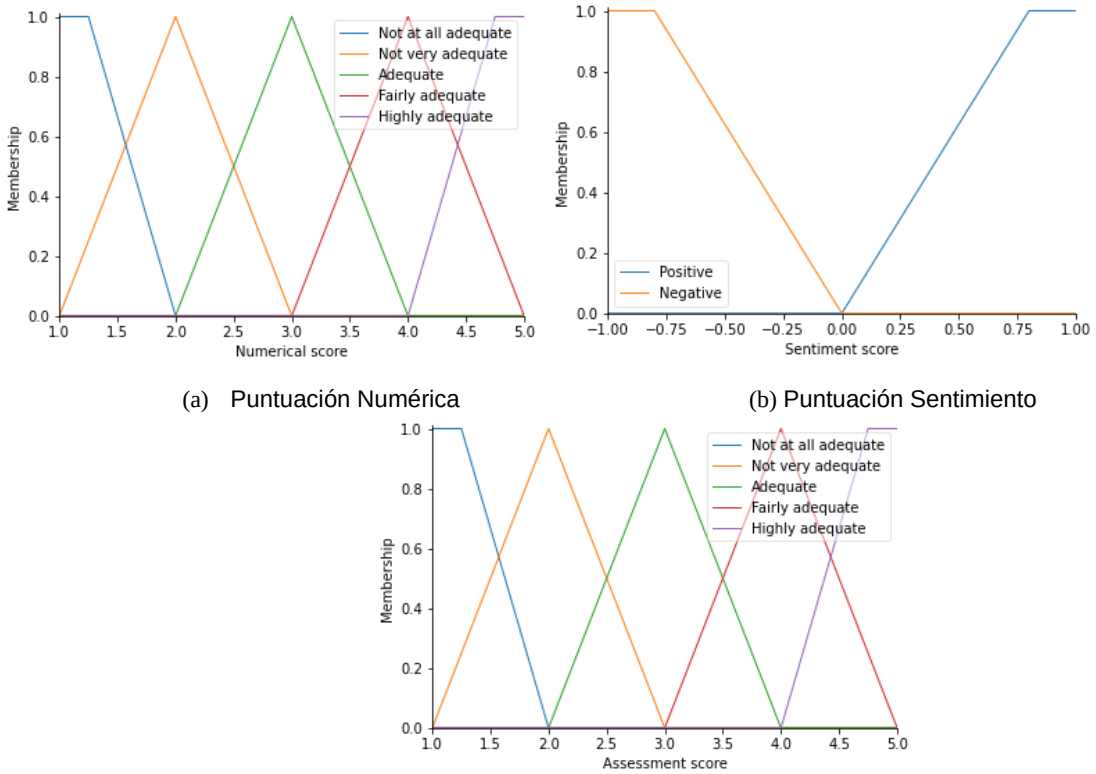
Para construir la simulación, se utilizó las variables de entrada: Puntuación Numérica y Puntuación Sentimiento obtenido del modelo predictivo (Tabla **36**).

Fuzzificación (valor difuso de entrada y salida)

Las variables de entrada (Puntuación Numérica y Puntuación Sentimiento) y la variable de salida (Puntuación Evaluación) se dividieron en términos lingüísticos y se combinaron en un par de funciones de membresía, gamma invertida y gamma, solo para los bordes y función triangular para valores medios. Esta combinación permite cubrir todas las escalas (1…5) y (1…−1). En la Tabla **44** se enumeran los términos lingüísticos con parámetros asignados a cada variable lingüística y en la Figura **38** se presenta los gráficos de las funciones de pertenencia asociadas para obtener la puntuación de evaluación de cada criterio.

Tabla 44. Variables de entrada y salida en términos lingüísticos

Representación de variable	Variable lingüística	Término lingüístico	Función de membresía	Parámetros
× 1	Puntuación Numérica	Nada adecuado	Trapezoide	[1, 1,1.25, 2]
		No muy adecuado	Triangular	[1,2,3]
		Adecuado	Triangular	[2,3,4]
		Bastante adecuado	Triangular	[3,4,5]
		Muy adecuado	Trapezoide	[4,4.75,5,5]
× 2	Puntuación Sentimiento	Positivo	Trapezoide	[0,0.8,1,1]
		Neutral	Triangular	[-1,0,1]
		Negativo	Trapezoide	[-1,-1,-0.8,0]
y	Puntuación Evaluación	Nada adecuado	Trapezoide	[1, 1,1.25, 2]
		No muy adecuado	Triangular	[1,2,3]
		Adecuado	Triangular	[2,3,4]
		Bastante adecuado	Triangular	[3,4,5]
		Muy adecuado	Trapezoide	[4,4.75,5,5]



(a) Puntuación Numérica

(b) Puntuación Sentimiento

(c) Puntuación Evaluación

Figura 38. Ejemplos de funciones de membresía aplicadas: (a) Puntuación Numérica, (b) Puntuación Sentimiento y (c) Puntuación Evaluación para obtener la puntuación de evaluación para cada criterio

Regla difusa

Las reglas lingüísticas se formularon a partir de si había imprecisión entre Puntuación Numérica y Puntuación Sentimiento, se penaliza y se asigna a la Puntuación Evaluación una calificación de (Nada adecuado), y si había precisión, se asigna una calificación de (Nada adecuado, Poco adecuado, Adecuado, Bastante adecuado y Totalmente adecuado). La Figura **39** muestra las reglas difusas para obtener la Puntuación Evaluación de cada criterio y la Figura **40** muestra las reglas difusas para obtener la Puntuación Evaluación de todos los criterios por evaluador o grupo.

```
# Detecting accuracy and generating the assessment score
R1 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R2 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
R3 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Positive'], Assessment_score['Adequate'])
R4 = ctrl.Rule(Numerical_score['Fairly adequate'] & Sentiment_score['Positive'], Assessment_score['Fairly adequate'])
R5 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Positive'], Assessment_score['Highly adequate'])
# Detecting inaccuracy and generating the assessment score
R6 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Positive'], Assessment_score['Not at all adequate'])
R7 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Positive'], Assessment_score['Not at all adequate'])
R8 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R9 = ctrl.Rule(Numerical_score['Fairly adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R10 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
```

Figura 39. Reglas difusas para obtener la puntuación de evaluación de cada criterio desarrollado en Python

```
# Detecting accuracy and generating the assessment score
R1 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R2 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
R3 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
R4 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Positive'], Assessment_score['Mainly adequate'])
R5 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Positive'], Assessment_score['Highly adequate'])
# Detecting inaccuracy and generating the assessment score
R6 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Neutral'], Assessment_score['Not at all adequate'])
R7 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Positive'], Assessment_score['Not at all adequate'])
R8 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Neutral'], Assessment_score['Not at all adequate'])
R9 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Positive'], Assessment_score['Not at all adequate'])
R10 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R11 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Positive'], Assessment_score['Not at all adequate'])
R12 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R13 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Neutral'], Assessment_score['Not at all adequate'])
R14 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
R15 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Neutral'], Assessment_score['Not at all adequate'])
```

Figura 40. Reglas difusas para obtener la puntuación de evaluación de todos los criterios por evaluador (puntuación individual) o grupo (puntuación del colectivo) desarrollado en Python

Defuzzificación

Una vez que se activaron las reglas adecuadas en el sistema de control difuso, el grado de pertenencia de la variable difusa de salida (Puntuación Evaluación) se determina mediante la codificación de los subconjuntos difusos antecedentes (Puntuación Numérica y Puntuación Sentimiento). Se utiliza el método de inferencia max-min (Ecuación 8), donde la función de pertenencia de salida final para cada regla es el conjunto difuso asignado a esa salida mediante el recorte del grado de valores de verdad de las funciones de pertenencia de los antecedentes asociados. Una vez que se determina el grado de pertenencia de la variable difusa de salida, todas las reglas que se activan se combinan y la salida nítida real se obtiene mediante la defuzzificación.

$$\mathbf{Y} = \underset{k}{\boldsymbol{max}}\left[\boldsymbol{min}\left(\boldsymbol{\mu}_{B_1^k}(\boldsymbol{y}), \boldsymbol{\mu}_{A^k}(\boldsymbol{x1}), \boldsymbol{\mu}_{A_2^k}(\boldsymbol{x2})\right)\right] \mathbf{k} = \mathbf{1}, \mathbf{2}, \ldots \mathbf{n} \quad \text{(Ecuación 8)}$$

Dónde:

× 1 = Puntuación Numérica (variable de entrada).

× 2 = Puntuación Sentimiento (variable de entrada).

k = 1 hasta n reglas.

Y = Resultado de la defuzzificación, es la función de pertenencia, (y) es la variable de salida (Puntuación Evaluación), min es el límite inferior, y max es el límite máximo para la defuzzificación

En la defuzzificación, se probó los métodos Centroid, Bisector, SOM, MOM y LOM (Figura **41**). Cuando se calculó la puntuación de evaluación para cada criterio, se obtuvieron buenos resultados con todos los métodos (Tabla **45** y Figura **42**), mientras que cuando se calculó la puntuación de evaluación con todos los criterios por evaluador o grupo, los mejores resultados se obtuvieron con los métodos SOM, MOM y LOM (Tabla **46** y Figura **43**). Por lo tanto, se colige que los métodos más apropiados para este estudio son SOM, MOM y LOM porque calculan el resultado más plausible.

```
from skfuzzy import control as ctrl
#Universe of discourse and defuzzify_method
Numerical_score= ctrl.Antecedent(np.arange(1,5.25,0.25),'Numerical score')
Sentiment_score = ctrl.Antecedent(np.arange(-1,1.2,0.2),'Sentiment score')
Assessment_score= ctrl.Consequent(np.arange(1,5.25,0.25),'Assessment score', defuzzify_method='LOM')
```

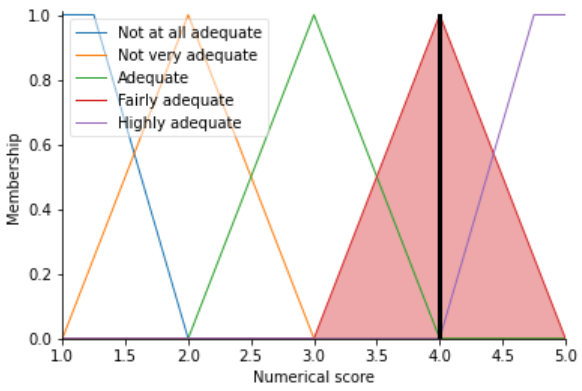
```
#Control System
tipping_ctrl = ctrl.ControlSystem([R1,R2,R3,R4,R5,R6,R7,R8,R9,R10])
tipping = ctrl.ControlSystemSimulation(tipping_ctrl)
```

```
#Input
tipping.input['Numerical score']=Numerical_score
tipping.input['Sentiment score']=Sentiment_score
#Output fuzzy-Implement rule according to Mamdani inference
tipping.compute()
#Fuzzy computed value
Numerical_score.view(sim=tipping)
Sentiment_score.view(sim=tipping)
Assessment_score.view(sim=tipping)
```

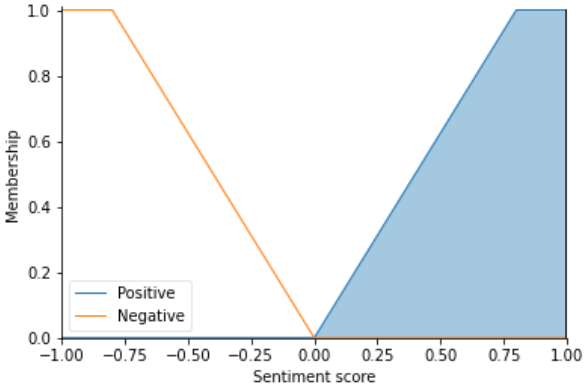
Figura 41. Muestra del sistema de control difuso para el cómputo de la puntuación de evaluación de cada criterio con el método de defuzzificación LOM, desarrollado en Python

Tabla 45. Ejemplos de comparación de puntuación de evaluación por cada criterio con diferentes métodos de defuzzificación

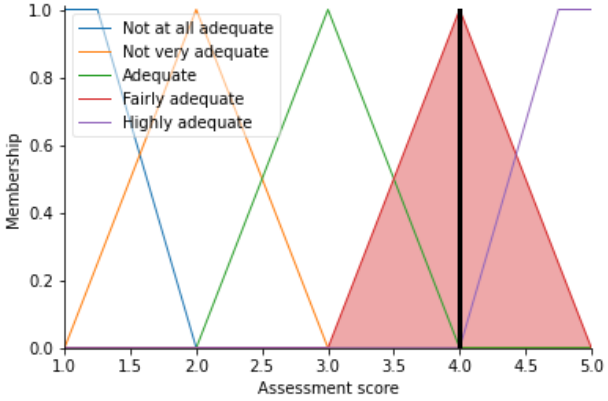
Actividad	Grupo Evaluado	Evaluador	Criterio	Puntuación Numérica	Puntuación Sentimiento	Puntuación Evaluación				
						Centroid	Bisector	SOM	MOM	LOM
A01	FIS01H	VA5740	Diseño	1	1	1.13	1.32	1.00	1.13	1.25
A02	FIS02M	AR9371	Comunicación	1	-1	1.13	1.32	1.00	1.13	1.25
		CH5277	Comunicación	2	1	1.13	1.32	1.00	1.13	1.25
		AW4646	Casos de uso	2	-1	2.00	2.00	2.00	2.00	2.00
		SJ314	Comunicación	3	1	3.00	3.00	3.00	3.00	3.00
		LB9473	Comunicación	3	-1	1.13	1.32	1.00	1.13	1.25
		TF2879	Diseño	4	1	4.00	4.00	4.00	4.00	*4.00*
		MY7491	Comunicación	4	-1	1.13	1.32	1.00	1.13	1.25
		MY7491	Diseño	5	1	4.88	4.68	4.75	4.88	5.00
	FIS02I	LJ8822	Actores	5	-1	1.13	1.32	1.00	1.13	1.25



(a) Puntuación Numérica



(b) Puntuación Sentimiento

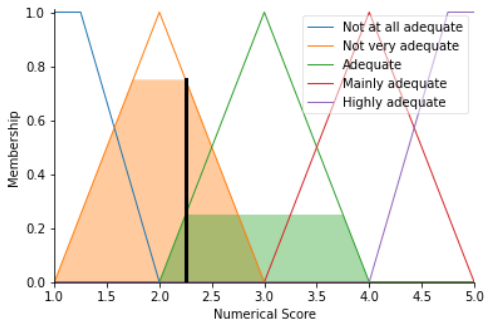


(c) Puntuación Evaluación

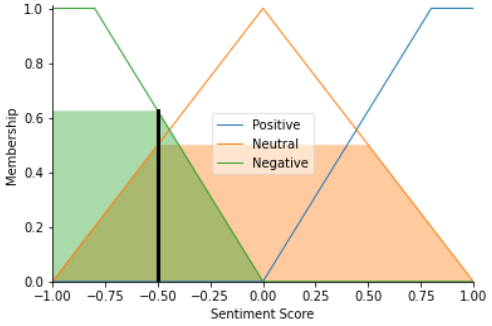
Figura 42. Ejemplo de respuesta del sistema de control difuso para el cálculo de (c) puntuación de evaluación del criterio (Diseño) del calificador (TF2879). Los conjuntos difusos resultantes se derivan del área de superficie roja resaltada, y la línea negra indica el valor de puntuación de evaluación típico óptimo después del proceso de defuzzificación con el método LOM

Tabla 46. Ejemplos de comparación de la puntuación de evaluación de todos los criterios por cada evaluador (puntuación individual) con diferentes métodos de defuzzificación

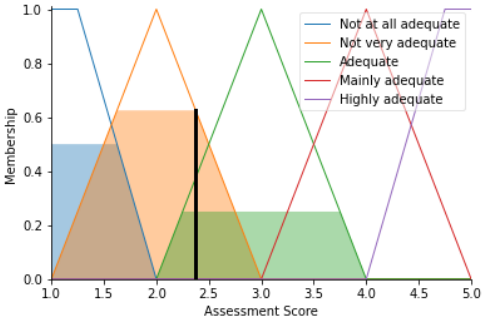
Actividad	Grupo Evaluado	Evaluador	Puntuación Numérica	Puntuación Sentimiento	Puntuación Evaluación Centroid	Bisector	SOM	MOM	LOM
A02	FIS02M	AI7915	3.00	-0.50	2.37	2.52	1.00	1.32	1.53
		AR9371	*2.25*	*-0.50*	2.19	2.10	1.63	2.00	*2.38*
		AS1826	4.50	0.00	1.50	1.38	1.00	1.25	1.50
		AW4646	2.25	-0.50	2.19	2.10	1.63	2.00	2.38
		CL6082	2.75	-0.50	2.36	2.46	1.00	1.32	1.53
		CH5277	3.25	0.00	2.70	2.85	2.75	3.00	3.25
		LB9473	3.75	0.00	2.10	1.72	1.00	1.23	1.44
		SJ314	3.75	1.00	3.50	3.85	3.75	4.00	4.25
		TA251	4.25	0.00	1.38	1.36	1.00	1.23	1.44
		TF2879	3.25	0.00	2.70	2.85	2.75	3.00	3.25
		ZJ1217	4.25	0.50	3.25	3.73	3.63	4.00	4.38
		MY7491	4.00	0.00	1.35	1.32	1.00	1.17	1.25



(a) Puntuación Numérica



(b) Puntuación Sentimiento



(c) Puntuación Evaluación

Figura 43. Ejemplo de respuesta del sistema de control difuso para el cálculo de (c) puntuación de evaluación de todos los criterios del evaluador (AR9371). Los conjuntos difusos resultantes se derivan de las áreas de superficie resaltadas en azul, naranja y verde, y la línea negra indica el valor de puntuación de evaluación típico óptimo después del proceso de defuzzificación con el método LOM

Los resultados experimentales muestran que mediante la técnica de lógica difusa se logró correlacionar la puntuación de sentimiento y numérica, y generar la puntuación de evaluación por cada criterio, con todos los criterios por evaluador (puntuación individual) o grupo (puntuación del colectivo). Se determinó que la correlación por cada criterio es la más apropiada, puesto que la correspondencia es la predicción de puntuación de sentimiento de cada retroalimentación con su puntuación numérica.

Conclusiones, limitaciones y futuro experimento

- Los resultados de los modelos predictivos mostraron que con aprendizaje automático clásico (MNB, SVM, LR, RF y DT), SVM (F-Measure de 0.879) fue el mejor modelo con función lineal. Con aprendizaje automático moderno, VE (F-Measure de 0.882) superó a SVM, pero con costos computacionales más altos; ambos modelos demostraron su eficacia cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF; y con aprendizaje profundo LSTM/Bi-LSTM (F-Measure de 0.832/0.837) respectivamente el rendimiento fue más débil, por lo que el tamaño de la muestra fue pequeño, en trabajos futuros se seguirá realizando pruebas con una muestra más grande.
- Se seleccionó SVM (F-Measure de 0.879) con representación de 1-g y 2-g+TFIDF como modelo base para predecir la puntación del sentimiento por lo que obtuvo un buen rendimiento con bajo costo computacional, y superó a LibSVM (F-Measure de 0.870) con representación de Bag of Words+TF-IDF+Stop-Words del experimento-1.
- Los resultados demostraron un rendimiento superior para las combinaciones de 2 palabras para la representación de textos, por lo que son más expresivas y capturan mejor el contexto que una sola palabra, ya que puede causar confusión y es muy imprecisa para expresar el significado. Esto es consistente con los hallazgo de [146] que obtuvieron mejor rendimiento con 2-g, que con 3-g, por lo que 3-g sufren una gran cantidad de combinaciones que provocan una rápida disminución de las frecuencias reales por característica.
- Los resultados mostraron que el modelo predictivo es confiable en las 5 actividades ejecutadas, porque se obtuvo una concordancia entre la puntuación de sentimiento generada y la polaridad de sentimiento etiquetada por el anotador igual o superior al 80%, con los coeficientes Kappa, Pearson y Spearman.

- Los resultados mostraron que el modelo predictivo es portable en otra asignatura impartida en escenario de educación tradicional.
- En la obtención de la puntuación de evaluación con todos los criterios por evaluador (puntuación individual) o grupo (puntuación del colectivo) se generó una nueva puntuación de sentimiento (0-neutral), para dar solución se añadió un nuevo termino lingüístico en la fuzzificación, y reglas difusas que correlacione la puntuación de sentimiento neutral con puntuación numérica, por tanto, se debería agregar la etiqueta neutral al conjunto de datos en trabajos futuros.
- Los resultados mostraron que los métodos de desfuzzificación SOM, MOM y LOM en lógica difusa fueron lo más precisos para generar la puntuación de la evaluación de la correlación entre puntuación de sentimiento y numérica por cada criterio, con todos los criterios por evaluador o grupo. Y se determinó que la correlación por cada criterio es la más apropiada, por tanto, debería seguir siendo probada en otros experimentos.
- Las limitaciones que se tuvieron en este experimento fueron:
  - Pequeño tamaño de la muestra.
  - El artefacto de la rúbrica realizada en Excel no facilitó la recolección de datos a gran escala.
- En el trabajo futuro, se planea:
  - Expandir el conjunto de datos.
  - Optimizar el artefacto de rúbrica con un prototipo de recolección de datos.
  - Ampliar el escenario de evaluación entre pares con evaluación inversa para cerrar el ciclo y permitir que el evaluado evalúe la calidad de evaluación de sus pares evaluadores. También que la evaluación se realice en dos rondas para que el estudiante corrija el trabajo y mejore el rendimiento en la segunda ronda.
  - Mejorar el etiquetado de cada retroalimentación, incrementar la etiqueta neutral para obtener mejores resultados en la clasificación de sentimiento.
  - Aplicar Word Embedding con algoritmos de aprendizaje profundo para comparar rendimientos con el modelo base SVM, y que clasifique la retroalimentación como positivo/neutral/negativo.
  - Utilizar un mayor número de asignaturas para mejorar la portabilidad del modelo.
  - Perfeccionar la correlación entre puntuación de sentimiento y numérica por cada criterio para obtener una puntuación equilibrada sin penalización.

- Probar otras medidas de cálculo para generar la puntuación individual y del colectivo.
- Reformar el proceso de penalización e incrementar bonificación, para obtener una puntuación de evaluación calibrada, sin que el docente interactúe.

- Estos resultados se publicaron en:
  - Accuracy' Measures of Sentiment Analysis Algorithms for Spanish Corpus generated in Peer Assessment. https://dl.acm.org/doi/pdf/10.1145/3410352.3410838
  - Peer assessment using soft computing techniques. https://doi.org/10.1007/s12528-021-09296-w

4.1.3. Experimento III

En este experimento se consideró la evaluación de tarea y evaluación inversa con puntuación numérica y retroalimentación textual en dos rondas (Figura **44**) del modelo de evaluación entre pares descrito en la subsección 3.2.

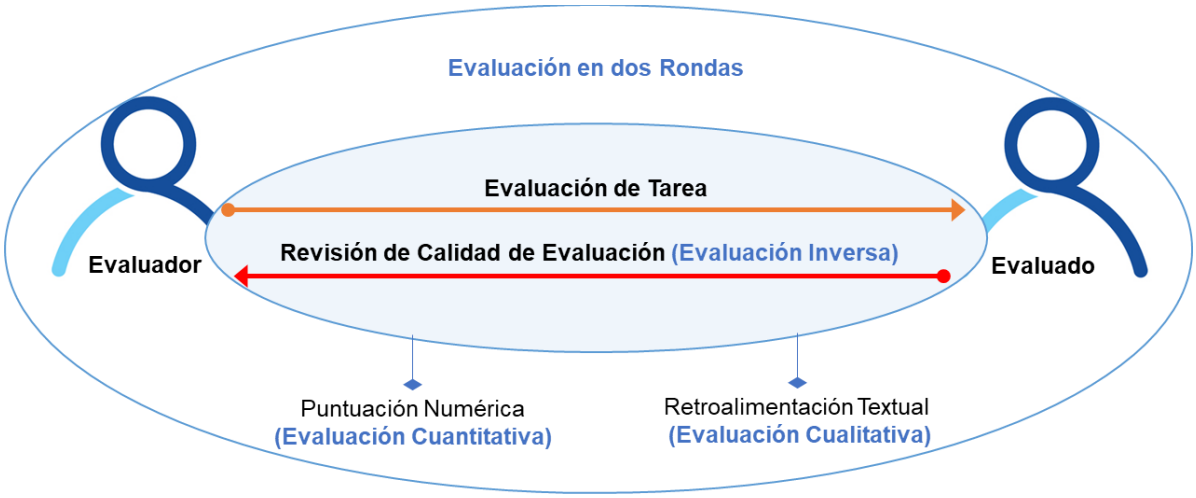


Figura 44. Evaluación entre pares cualitativa, cuantitativa, inversa y en dos rondas

Planteamientos, preguntas de investigación y objetivos

Planteamientos				
• Determinar si el enriquecimiento semántico de Word2Vec/Glove preentrenado, mejora el rendimiento de la tarea de clasificación del sentimiento de retroalimentación textual en español. • Determinar el rating de confianza de los pares evaluadores. • Determinar si la puntuación del colectivo debe ser calculada con media o mediana.				
Preguntas de investigación				
• ¿Como aplicar evaluación entre pares (cualitativa, cuantitativa, inversa y en dos rondas) en escenarios de educación superior?	• ¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?	• ¿Cómo correlacionar puntuación numérica (evaluación cuantitativa) con retroalimentación textual (evaluación cualitativa), y generar una puntuación equilibrada entre estas dos evaluaciones?	• ¿Cómo determinar un índice (rating) de confidencia de los pares evaluadores (evaluación inversa)?	• ¿Qué método de tendencia central (media/mediana) tiene el mejor ajuste para generar una puntuación del colectivo confiable en procesos de evaluación entre pares?
Objetivos				
• Diseñar los artefactos para validar los métodos teóricos. • Construir un modelo de evaluación entre pares basado en análisis de sentimiento. • Evaluar los resultados de la precisión del modelo.				

A continuación, se detalla el experimento:

Materiales y métodos

Participantes

Participaron 712 estudiantes de las asignaturas: fundamentos de ingeniería de software, ingeniería de software, fundamentos de ofimática, gestión de procesos de negocios-sistemas empresariales, y fundamentos de programación en escenario de educación presencial, virtual asincrónico y virtual sincrónico ante la pandemia COVID-19, de las carreras: ingeniería de

sistemas informáticos, tecnologías de la información, y pedagogía de la química, en los periodos académicos: octubre 2019-febrero 2020, mayo-octubre 2020, noviembre 2020-marzo 2021, mayo-septiembre 2021, octubre 2021-febrero 2022 y mayo-septiembre 2022 de la Universidad Técnica de Manabí del Ecuador.

Metodología de experimentación

La metodología utilizada en la experimentación se realizó en base a la subsección 3.5, 3.6 y 3.7, que constó de tres fases (Figura **45**). En la primera fase, se realizó análisis de sentimiento de evaluación de tarea y evaluación de calidad de evaluación, aplicando la técnica de aprendizaje automático, que recibe como entrada un corpus de texto en lenguaje natural etiquetado y genera una puntuación de sentimiento 1 (positivo), -1 (negativo), 0 (neutral), que corresponde a retroalimentación textual específica. En la segunda fase, se realizó la detección de precisión/imprecisión entre la puntuación de sentimiento y numérica, y se calculó la puntuación de evaluación por cada criterio utilizando la técnica de lógica difusa. En la tercera fase se obtiene la puntuación individual por cada evaluador, posteriormente se obtiene la puntuación del colectivo de evaluación de tarea, y rating de confianza del evaluador.

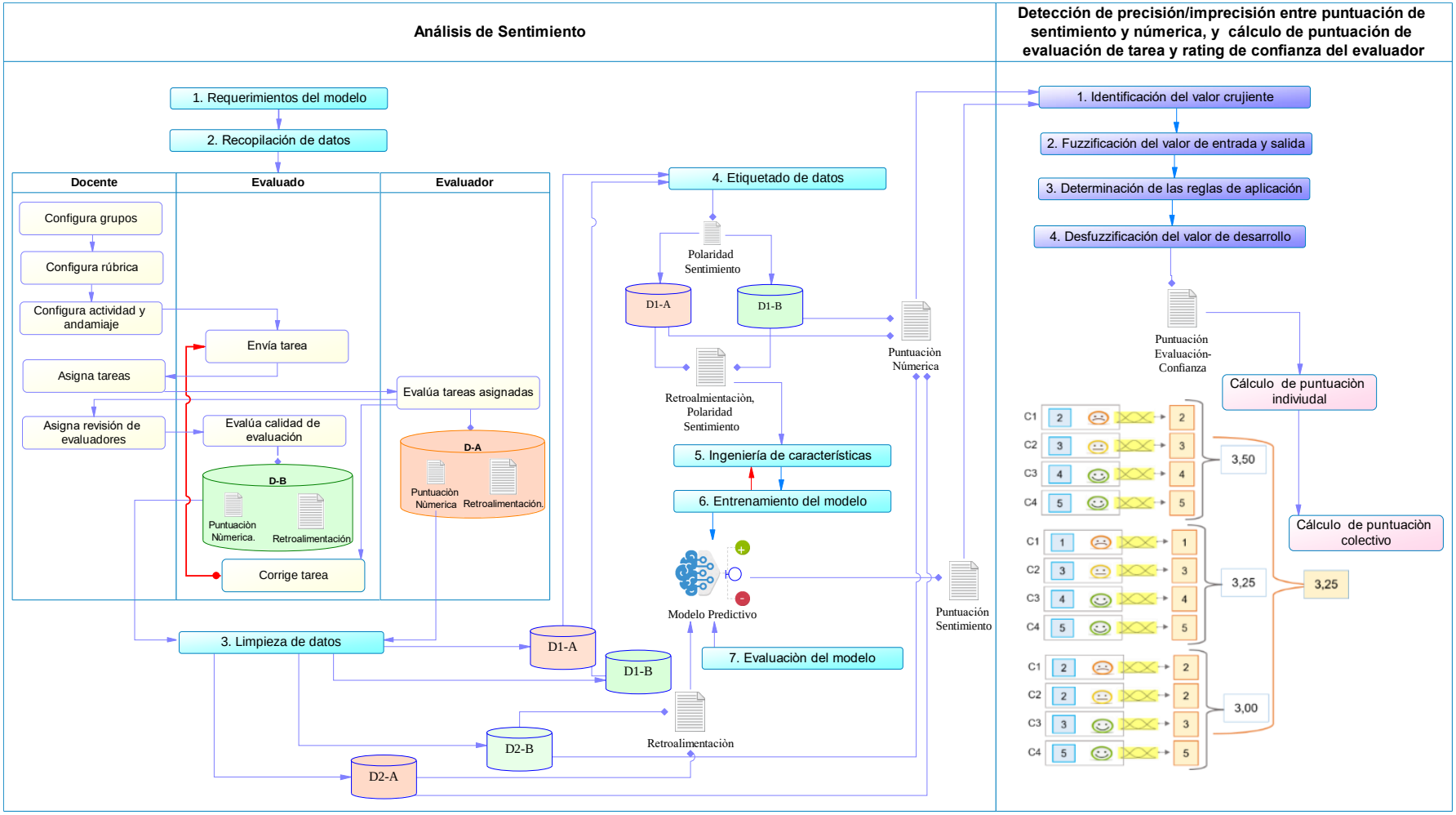


Figura 45. Metodología aplicada en el experimento

La Figura **45** muestra cinco representaciones gráficas, rectángulos para los pasos (pasos de análisis de sentimiento (color celeste), pasos de recolección de datos (color amarillo), pasos de lógica difusa (color morado), pasos de cálculo (color rosado), almacenes de datos para los conjuntos de datos (conjunto de datos de evaluación de tarea (color naranja) y conjunto de datos de evaluación de la calidad de la evaluación (color verde), archivos para las variables de entrada y salida, un icono para el modelo predictivo y la flecha de retroalimentación de color rojo para ilustrar que el entrenamiento del modelo puede regresar a ingeniería de características. Los pasos de recopilación, limpieza y etiquetado están orientados a los datos; los pasos de requerimientos, ingeniería de características, entrenamiento y evaluación están orientados al modelo predictivo; y los pasos de fuzzificación, reglas, defuzzificación, y computo de puntuación individual y del colectivo están orientados al cálculo.

A continuación, se describe con detalle el procedimiento llevado a cabo en cada una de las fases:

Primera fase: Análisis de sentimiento

- Se refinó esta fase, considerando el proceso de minería de texto descrito en la subsección 3.5, para obtener un modelo predictivo de evaluación de tarea, y otro de evaluación de la calidad de la evaluación (Figura **45)**. A continuación, se detallan los pasos:

Paso 1. Requerimientos del modelo

- El sentimiento se evaluó a nivel de oración. Se utilizó el enfoque de aprendizaje supervisado con modelos de aprendizaje profundo (LSTM/BiLSTM) con enriquecimiento semántico (Word2Vec/Glove), y se comparó con los resultados previo del modelo Baseline de esta investigación que ha sido publicado en [300].

Paso 2. Recopilación de datos

- Se realizó el escenario de evaluación entre pares (Tabla **47**), en base a la subsección 3.3.

Tabla 47. Escenario de evaluación entre pares del experimento III

Dimensión	Rango de variación
Área curricular/Asignatura	Fundamentos de ingeniería de software, ingeniería de software, fundamentos de ofimática, gestión de procesos de negocios-sistemas empresariales, y fundamentos de programación.
Objetivos	Docente reduce la carga de calificación y los estudiantes ven otras posibles soluciones y obtienen ganancias cognitivas.
Enfoque	Cuantitativa, cualitativa, inversa, y en dos rondas
Producto/Salida	Diagrama de: casos de uso, clases, actividad y secuencia, diseño de componentes, documento de texto, hoja de cálculo, diseño de aula virtual, proceso-gestión-empresa, y micro-sistema.
Relación de la evaluación personal	Sustitucional.
Valor oficial	100% de la puntuación de evaluación del colectivo.
Direccionalidad	Mutua.
Privacidad	Anónima.
Contacto	La evaluación se realiza mediante un formulario en línea/prototipo de evaluación entre pares.
Año	El mismo año de estudio.
Habilidad	La evaluación está guiada por una rúbrica para obtener el máximo beneficio de las habilidades del evaluador.
Constelación del evaluador	Individual (1/2/3 tareas asignadas).
Constelación evaluada	Las tareas enviadas se realizaron de manera colaborativa.
Lugar	Aula de clase/Virtual.
Tiempo	Hora de clase.
Requerimiento	Obligatorio para evaluadores/evaluados.
Bonificación	Ninguna

- Se recolectó datos de evaluación de tarea, evaluación inversa y en dos rondas (Figura **45**), considerando subsección 3.4.

Configuración de grupos

  - Se formaron grupos de cuatro/cinco/seis estudiantes según la asignatura. En la asignatura de fundamentos de ingeniería de software se formaron grupos considerando varios paralelos, en la asignatura de ingeniería de software, fundamentos de ofimática, gestión de procesos de negocios-sistemas empresariales, y fundamentos de programación se formaron grupos con un solo paralelo.

Configuración de rúbrica

- Se diseñaron rúbricas para cada actividad (Tabla **48**), considerando los parámetros de rúbrica tipo holística.

Tabla 48. Ejemplo de rúbrica (evaluación de diagrama de actividad)

Código Actividad: **Código Evaluador:** **Código Evaluado:**		**Nivel de ponderación (Puntuación Numérica)** 5 (Totalmente adecuado) 4 (Bastante adecuado) 3 (Adecuado) 2 (Poco adecuado) 1 (Nada adecuado)
Criterio	**Descripción**	**Puntuación Numérica** **Retroalimentación**
Diseño	El **nombre del diagrama** es adecuado. La **distribución** de carriles es ordenada, lo que permite que sea fácil de comprender.	
Nodo	El **nodo inicial** y **nodo final** es incluido correctamente en el diagrama, lo que permite saber donde inician y donde terminan los procesos.	
Actividad	El **flujo base** y **flujo alterno** es representado perfectamente y tiene la **funcionalidad** según lo descrito en la especificación de casos de uso.	
Flujo de objetos	El **flujo de contro**l y **control de decisión** se incluye donde se requieren y de manera correcta.	

Configuración de actividad y andamiaje

- Se configuró cada actividad grupal con n asignaciones y andamiaje. En la asignatura de fundamentos de ingeniería de software e ingeniería de software se diseñaron actividades que consistieron en resolución de ejercicios de diagrama de casos de uso, diagrama de clases, diagrama de secuencia, diagrama de actividades y diseño de componentes. En la asignatura de fundamentos de ofimática se diseñaron actividades que consistieron en realizar documento de texto, hoja de cálculo y diseño de aulas virtuales. En la asignatura de gestión de procesos de negocios y sistemas empresariales se diseñaron actividades de ejercicio de proceso de gestión empresarial y en la asignatura de fundamentos de programación se diseñaron ejercicios de micro sistema.
- Se diseñó el andamiaje por cada etapa de evaluación entre pares (envió de tarea, evaluación de tarea, evaluación de calidad de la evaluación) con texto y video.

Envío de tarea

- Cada grupo realizó el trabajo de manera colaborativa y subió el enlace en el prototipo de evaluación entre pares.

Asignación de tareas/Evaluación de calidad de la evaluación

- Se utilizaron revisiones anónimas, el docente asignó una/dos/tres tareas a cada estudiante, también asignó las revisiones de la calidad de la evaluación.

Evaluación de tarea/Evaluación de la calidad de la evaluación

- Cada evaluador evaluó individualmente las tareas/evaluación de la calidad de la evaluación asignadas de manera objetiva y ética, de acuerdo con la rúbrica dada, y proporcionaron puntuación numérica y retroalimentación textual por cada criterio.

- En este paso se obtuvo un conjunto de datos (D1) con 22136 instancias de evaluación de tarea y 13788 instancias de evaluación de calidad de evaluación (Tabla **49**), y un conjunto de datos (D2) con 20322 instancias de evaluación de tarea y 19997 instancias de evaluación de calidad de evaluación (Tabla **50**).

Tabla 49. Conjunto de datos para entrenamiento de los modelos del experimento III

C	E	P	Id-A	A	N°	G	Id-Act	Id-T	Recopilación de datos				Limpieza de datos			
									Evaluación de tarea		Evaluación de calidad de la evaluación		Evaluación de tarea		Evaluación de calidad de la evaluación	
									R1	R 2	R1	R2	R1	R2	R1	R2
D1	Presencial	Oct 2019-feb 2020	FIS	Rúbrica (Excel)	68 (4)	17	FIS01	A1	912	-	-	-	912	-	-	-
						16	FIS02	A1	794	-	-	-	764	-	-	-
							FIS03	A4	790	-	-	-	764	-	-	-
							FIS04	A2	790	-	-	-	764	-	-	-
							FIS05	A3	790	-	-	-	764	-	-	-
	Virtual Asincrónica	May-oct 2020			161 (6)	26	FIS06	A1	1288	-	-	-	800	-	-	-
		Nov 2020-mar 2021		Rúbrica (Google Form)	104 (4)	24	FIS07	A1	1072	1084	856	892	1068	1080	852	888
							FIS08	A4	1068	1072	840	852	1064	1068	836	852
							FIS09	A2	1068	1068	856	856	1064	1064	828	856
							FIS10	A3	1060	1068	860	800	1056	1064	852	800
		May-sep 2021		Prototipo	97 (5)	20	FIS11	A1	1244	968	882	852	1016	968	882	852
						17	FIS12	A4	1016	968	852	918	1016	968	852	918
							FIS13	A3	1016	992	884	852	1016	992	884	852
							FIS14	A2	1016	992	884	852	1016	992	884	852
						Total de instancias			13924	8212	6914	6874	13084	8196	6870	6870
									22136		13788		21280		13740	

Tabla 50. Conjunto de datos para evaluación de los modelos del experimento III

C	E	P	Id-A	A	N°	G	Id-Act	Id-T	Recopilación de datos				Limpieza de datos			
									Evaluación de tarea		Evaluación de calidad de la evaluación		Evaluación de tarea		Evaluación de calidad de la evaluación	
									R1	R 2	R1	R2	R1	R2	R1	R2
D2	Virtual Asincrónica	Oct 2021 -feb 2022	FIS	Prototipo	66 (5)	14	FIS15	A1	848	848	848	848	848	848	848	848
							FIS16	A2	848	824	816	732	848	824	816	732
							FIS17	A3	824	824	824	812	824	824	824	812
							FIS18	A4	824	824	824	824	824	824	824	824
			IS		36 (4-5)	8	IS19	A1	476	476	476	476	476	476	476	476
							IS20	A2	476	464	464	452	476	464	464	452
							IS21	A3	464	464	464	464	464	464	464	464
							IS22	A4	464	464	464	464	464	464	464	464
	Virtual Sincrónica	May 2022 -sept 2022	FIS		76 (5)	16	FIS23	A1	672	662	664	628	672	662	664	628
							FIS24	A2	640	624	640	624	640	624	640	624
							FIS25	A3	632	600	632	580	632	600	632	580
							FIS26	A4	608	592	600	549	608	592	600	549
							FIS27	A5	592	592	592	540	592	592	592	540
			FO		35 (5)	7	FO28	A6	292	292	292	292	292	292	292	292
							FO29	A7	300	300	300	300	300	300	300	300
							FO30	A8	308	308	308	308	308	308	308	308
	Presencial		IS		17 (6)	3	FIS31	A1	148	148	148	148	148	148	148	148
							FIS32	A2	148	148	148	148	148	148	148	148
							FIS33	A3	148	148	148	148	148	148	148	148
							FIS34	A4	148	148	148	148	148	148	148	148
							FIS35	A5	148	148	148	148	148	148	148	148
			GP		27 (4-5)	6	GP36	A9	108	108	108	108	108	108	108	108
			FP		25 (6)	4	FP37	A10	100	100	100	100	100	100	100	100
							Total de instancias		10216	10106	10156	9841	10216	10106	10156	9841
									20322		19997		20322		19997	

Donde:

C= Conjunto de datos
E= Escenario de educación
P= Periodo académico
Id-A = Código de asignatura
A=Artefacto de recolección de datos
N°= Número de estudiantes (Número de integrantes de grupo)
G= Grupos
Id-Act = Numero de Actividad
Id-T = Código de tarea
R1= Ronda-1
R2= Ronda-2

Id-A	Asignatura/Carrera
FIS	Fundamentos de ingeniería de software/ Carrera de ingeniería de sistemas informáticos
IS	Ingeniería de software/ Carrera de tecnologías de la información
FO	Fundamentos de ofimática/ Carrera de pedagogía de la química
GP	Gestión de procesos de negocios y sistemas empresariales/ Carrera de ingeniería de sistemas informáticos
FP	Fundamentos de programación/ Carrera de tecnologías de la información

Id-T	Tarea
A1	Ejercicios de diagrama de casos de uso
A2	Ejercicios de diagrama de actividad
A3	Ejercicios de diagrama de secuencia
A4	Ejercicios de diagrama de clases
A5	Ejercicio de diseño de componentes
A6	Documento de Texto
A7	Hoja de Cálculo
A8	Diseño de aulas virtuales
A9	Ejercicio de proceso-gestión-empresa
A10	Ejercicios de micro-sistema

Paso 3. Limpieza de datos

- Se descartó registros vacíos de puntuación numérica o retroalimentación textual de los conjuntos de datos.
- En este paso, se redujo el conjunto de datos (D1) para el entrenamiento del modelo en 21280 instancias de evaluación de tarea y 13740 instancias de evaluación de calidad de la evaluación (Tabla **53**).

Paso 4. Etiquetado de datos

- Se etiquetó cada retroalimentación textual del conjunto de datos (D1-A) y (D1-B) en positivo, negativo o neutral considerando las reglas de la subsección 3.5. En este paso, se obtuvo una nueva variable (Polaridad Sentimiento) para el conjunto de datos (D1-A) y (D1-B).
- La Tabla **51** muestra ejemplos de retroalimentación de evaluación de tarea etiquetadas. La retroalimentación (F1) fue etiquetada (1, positivo) porque contiene las palabras “adecuado, ordenada”. La retroalimentación (F2) se etiquetó (0, neutral) porque contiene las palabras “correcto, no”. La retroalimentación (F3) se etiquetó (-1, negativo) porque contiene la palabra "no”. La retroalimentación (F4) se etiquetó (0, neutral) porque contiene las palabras “parcialmente correcto, "no”.

Tabla 51. Ejemplos de retroalimentación en español etiquetadas de evaluación de tarea de la actividad-FIS14 (ejercicios de diagrama de actividad)

Evaluado	Evaluador	Criterio	Retroalimentación	Polaridad Sentimiento
Grupo-7	1312851411	Diseño	F1: El nombre del diagrama es adecuado. La distribución de carriles es ordenada, lo que permite que sea fácil de comprender.	1
		Nodo	F2: Nodo inicial correcto, no tienen nodo final algunos subprocesos.	0
		Actividad	F3: No es representando en su totalidad el flujo base y alterno especificado en la plantilla.	-1
		Flujo de objetos	F4: El flujo de control es parcialmente correcto, algunos controles de decisión no están descritos de manera correcta.	0

- La Tabla **52** muestra ejemplos de retroalimentación de evaluación de calidad de la evaluación etiquetadas. La retroalimentación (F1 y F2) fue etiquetada (1, positivo) porque contiene la palabra “correcto”. La retroalimentación (F3) se etiquetó (0, neutral) porque contiene las

palabras “adecuado, falta”. La retroalimentación (F4) se etiquetó (-1, negativo) porque contiene la palabra "falta”.

Tabla 52. Ejemplos de retroalimentación en español etiquetadas de evaluación de calidad de la actividad-FIS14 (ejercicios de diagrama de actividad)

Evaluado	Evaluador	Criterio	Retroalimentación	Polaridad Sentimiento
1312851411	1313691840	Diseño	El comentario es correcto.	1
		Nodo	Es correcto lo indicado sobre el nodo inicial, es correcto lo indicado sobre el nodo final, por lo que lo tomaremos en cuenta al momento de realizar la corrección del trabajo	1
		Actividad	El comentario es adecuado, pero falta especificar en donde no está representado.	0
		Flujo de objetos	Falta detallar cuál control de decisión no está descrito correctamente.	-1

- La Tabla **53** muestra el detalle de los conjuntos de datos subdivididos para la experimentación: (D1-A) con 21280 instancias para el entrenamiento del modelo de evaluación de tarea, (D1-B) con 13740 instancias para el entrenamiento del modelo de evaluación de calidad de evaluación, (D2-A) con 20322 instancias para la evaluación del modelo de evaluación de tarea, y (D2-B) con 19997 instancias para la evaluación del modelo de evaluación de calidad de evaluación.

Tabla 53. Detalle de conjuntos de datos para el experimento III

Conjunto de datos			Total de instancias	N° de retroalimentación positiva	N° de retroalimentación neutral	N° de retroalimentación negativa
Entrenamiento del modelo	D1-A	Evaluación de Tarea	21280	8647	6303	6330
	D1-B	Evaluación de calidad de la evaluación	13740	6327	3700	3713
Evaluación del modelo	D2-A	Evaluación de Tarea	20322	-	-	-
	D2-B	Evaluación de calidad de la evaluación	19997	-	-	-

Paso 5. Ingeniería de características

En este paso se realizaron varios subprocesos (Figura **46**). A continuación, se detallan:

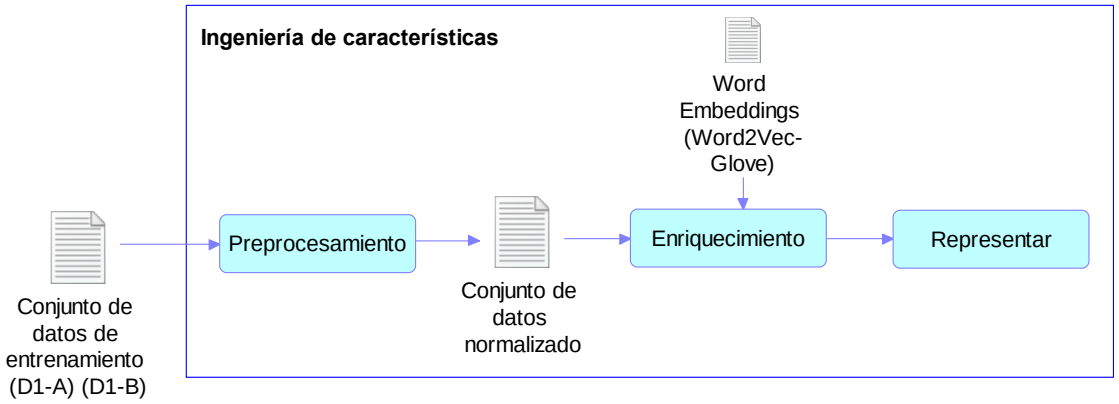


Figura 46. Ingeniería de características con Word2Vec/Glove

- Se preparó dos CSV, uno con el conjunto de datos de entrenamiento de evaluación de tarea (D1-A) y otro con el conjunto de datos de entrenamiento de evaluación de calidad de la evaluación (D1-B) con las variables de entrada (Retroalimentación y Polaridad Sentimiento), puesto que el contexto de los conjuntos de datos es disímil, se decidió después de monitorear que deben ser entrenados por separado.
- En el preprocesamiento del texto, se consideró normalización (corrección de errores ortográficos, conversión a minúsculas y tratamiento de caracteres especiales UTF-8), eliminación de Stop-Words, y tokenización.
- Cada conjunto de datos normalizado fue utilizado para alimentar el algoritmo que crea las representaciones vectoriales de las palabras. Para las mismas se utilizó transferencia de aprendizaje con los métodos Word2Vec/Glove preentrenado de word embedings construido a partir del corpus Spanish Billions Words (SBW) que contiene textos internacionales en español de [301]. Los modelos se extrajeron en (https://github.com/dccuchile/spanish-word-embeddings)
  - El conjunto de datos de Word2Vec tiene un peso de 2.67 GB y se compone de 1,000,653 vectores (Tabla **54**).

Tabla 54. Descripción de hiperparámetro utilizados en Word2Vec para SBW

Hiperparámetro	Valor
Dimensión	300
Ventana	5
Frecuencia mínima	5
Eliminar los primeros más comunes	273
Muestreo negativo	20
Arquitectura	skip-gram

- El conjunto de datos de Glove tiene un peso de 2.27 GB y se compone de 855,380 vectores (Tabla **55**).

Tabla 55. Descripción de hiperparámetro utilizados en Glove para SBW

Hiperparámetro	Valor
Dimensión	300
iter	5
Frecuencia mínima	5

- El objetivo de estos métodos es que en lugar de contar la frecuencia con la que una palabra w aparece cerca de otras, se entrena un clasificador sobre una predicción binaria, basada en una palabra objetivo w y una palabra c, la cual es: ¿La palabra c aparece en el contexto de w?. Una vez entrenado el clasificador se tomarán los pesos de este, como los valores que representan el vector de la palabra w [301].

- Se representó cada retroalimentación previamente enriquecida como una concatenación o promedio de los vectores de cada una de las palabras (Figura **47**).

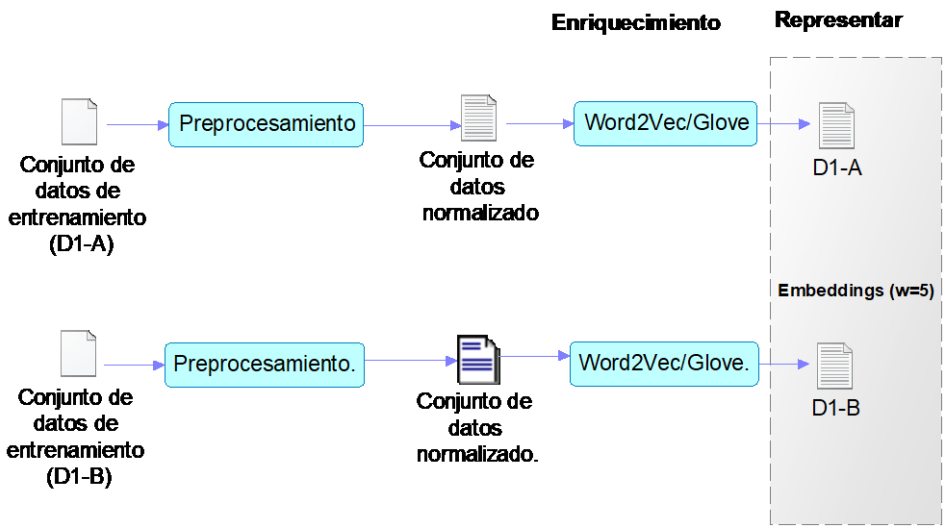


Figura 47. Resultados de representaciones vectoriales de palabras

Paso 6. Entrenamiento del modelo

- En este paso se realizaron dos entrenamientos, uno con el conjunto de datos de evaluación de tarea (D1-A) y otro con el conjunto de datos de evaluación de calidad de la evaluación (D1-B) (Figura **48**).

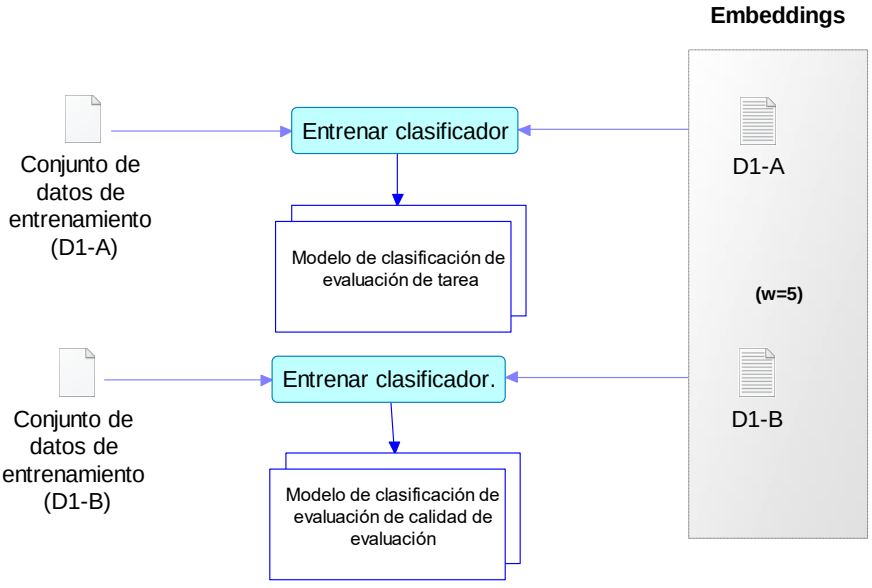


Figura 48. Modelos de clasificación de evaluación de tarea y de evaluación de calidad de la evaluación como resultado del entrenamiento

- Los conjuntos de datos se dividieron: 80% para entrenamiento y 20% para pruebas.
- Se ejecutó algoritmos de aprendizaje profundo LSTM y Bi-LSTM. Se repitieron estos pasos con todos los modelos, volviendo a la ingeniería de características. La Tabla **56** muestra la configuración de los parámetros de los algoritmos LSTM/Bi-LSTM.
- En este paso se obtuvieron modelos de clasificación para evaluación de tarea y para evaluación de calidad de la evaluación.

Tabla 56. Configuración de parámetros de algoritmos de aprendizaje profundo LSTM y Bi-LSTM de los conjuntos de datos (D1-A) y (D1-B)

Parámetro	**Valor**	
	LSTM/ Bi-LSTM (D1-A)	**LSTM/ Bi-LSTM (D1-B)**
Model	Sequential	Sequential
Num_words	10000	10000
Maxlen	200-240	150-200
Units	64	32
Dropout	0.8	0.5
Dense	3	3
Activation	Softmax	Softmax
Optimizer	Adam	Adam
Loss	Categorical _crossentropy	Categorical _crossentropy
Epochs	50	50
Learning rate	0.0001	0.0001
Batch size	50	50
Validation_split	0.2	0.2

Paso 7. Evaluación del modelo

- Al ejecutar cada modelo se obtuvieron los valores de las métricas de evaluación (Precision, Recall, F-Measure y Accuracy).
- En este paso se seleccionaron dos modelos predictivos con mejor rendimiento para generar la puntuación de sentimiento, uno para evaluación de tarea, y otro para evaluación de calidad de evaluación.
- Con los modelos predictivos seleccionados, se generó la puntuación de sentimiento de evaluación de tarea y evaluación de calidad de evaluación, en cada una de las actividades de evaluación entre pares.
- Se evaluó la portabilidad de los modelos predictivos entre asignaturas utilizando el conjunto de datos (D2-A) y (D2-B) de texto no etiquetado.

Paso 8. Implementación del modelo

- Se implementó los modelos de análisis de sentimiento en el prototipo de evaluación entre pares (ver Subsección 5.1.3).

Paso 9. Monitoreo del modelo

- Se monitoreo los modelos en los periodos académicos: mayo-septiembre 2021, octubre 2021-febrero 2022 y mayo-septiembre 2022.

Segunda fase: Detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación por cada criterio

Se refinó la fase de esta investigación que ha sido publicado en [300], considerando el proceso descrito en la subsección 3.6, para obtener un modelo que correlacione la puntuación de sentimiento y numérica sin penalización, y genere por cada criterio la puntuación de evaluación equilibrada de evaluación tarea y de calidad de evaluación (Figura **45**). A continuación, se detallan los pasos:

Paso 1. Identificación del valor nítido

- Se seleccionó las variables de entrada: puntuación numérica que contienen valores de 1 a 5, y puntuación de sentimiento que contiene valores 1 (positivo), -1 (negativo), 0 (neutral), generados por el modelo predictivo.

Paso 2. Fuzzificación del valor de entrada y salida

- A la variable de entrada (Puntuación Numérica) y variable de salida (Puntuación Evaluación) de tarea, se definió el universo de discurso con rangos entre 1 y 5 y se dividió en términos lingüísticos: 1-nada adecuado, 2-poco adecuado, 3-adecuado, 4-bastante adecuado y 5-totalmente adecuado; a la variable de entrada (Puntuación Sentimiento), se definió el universo de discurso con rangos entre 1 a -1, y se dividió en términos lingüísticos: 1-positivo, 0-neutro y -1-negativo; y a la variable de salida (Puntuación Confianza) de evaluación de calidad de la evaluación se definió el universo de discurso con rangos entre 0.2 y 1; y se dividió en términos lingüísticos: 0.2-bajo, 0.4-medio bajo, 0-6-medio, 0-8-medio alto, 1-alto.

- Luego, se formó las funciones de pertenencia, asignando los parámetros apropiados a los respectivos términos lingüísticos.

Paso 3. Determinación de reglas de aplicación

- Se determinaron reglas sin penalización, y así adquirir una puntuación equitativa entre puntuación de sentimiento y numérica.
- Para obtener la **puntuación de evaluación** de tarea por cada criterio, se formuló reglas difusas:

  – **Detección de precisión y generación de puntuación de evaluación**

  1. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es nada adecuado)
  2. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es poco adecuado)
  3. Si (puntuación numérica es adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es adecuado)
  4. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es bastante adecuado)
  5. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es totalmente adecuado)

  – **Detección de imprecisión y generación de puntuación de evaluación**

  6. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es poco adecuado)
  7. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es poco adecuado)
  8. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es positivo), entonces (puntuación evaluación es poco adecuado)
  9. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es poco adecuado)
  10. Si (puntuación numérica es adecuado) y (puntuación sentimiento es positivo), entonces (puntuación de evaluación es poco adecuado))

11. Si (puntuación numérica es adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es poco adecuado)
12. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es neutral), entonces (puntuación evaluación es adecuado)
13. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es negativo), entonces (puntuación evaluación es poco adecuado)
14. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es neutral), entonces (puntuación d evaluación es adecuado)
15. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es negativo), entonces (puntuación de evaluación es poco adecuado)

- Para obtener la **puntuación de evaluación de calidad de evaluación** por cada criterio, se formuló reglas difusas:

  – **Detección de precisión y generación de puntuación de confianza**

1. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es negativo), entonces (puntuación confianza es bajo)
2. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es negativo), entonces (puntuación de confianza es medio bajo)
3. Si (puntuación numérica es adecuado) y (puntuación sentimiento es neutral), entonces (puntuación confianza es medio)
4. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es positivo), entonces (puntuación confianza es medio alto)
5. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es positivo), entonces (puntuación confianza es alto)

  – **Detección de imprecisión y generación de puntuación de confianza**

6. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es positivo), entonces (puntuación confianza es medio bajo)
7. Si (puntuación numérica es nada adecuado) y (puntuación sentimiento es neutral), entonces (puntuación confianza es medio bajo)

8. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es positivo), entonces (puntuación confianza es medio bajo)
9. Si (puntuación numérica es poco adecuado) y (puntuación sentimiento es neutral), entonces (puntuación confianza es medio bajo)
10. Si (puntuación numérica es adecuado) y (puntuación sentimiento es positivo), entonces (puntuación confianza es medio bajo)
11. Si (puntuación numérica es adecuado) y (puntuación sentimiento es negativo), entonces (puntuación confianza es medio bajo)
12. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es neutral), entonces (puntuación confianza es medio)
13. Si (puntuación numérica es bastante adecuado) y (puntuación sentimiento es negativo), entonces (puntuación confianza es medio bajo)
14. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es neutral), entonces (puntuación confianza es medio)
15. Si (puntuación numérica es totalmente adecuado) y (puntuación sentimiento es negativo), entonces (puntuación confianza es medio bajo)

Paso 4. Desfuzzificación del valor de desarrollo

- Se utilizó el método LOM, para convertir el número difuso obtenido en un valor nítido.
- En este paso se obtuvo la puntuación de evaluación, y puntuación de confianza de cada criterio.

Tercera fase: Cálculo de puntuación individual y del colectivo de evaluación de tarea, y rating de confianza del evaluador

En esta fase se realizaron los cálculos de puntuación de evaluación de tarea y rating de confianza del colectivo en base al procedimiento descrito en la subsección 5.1.4.

- **Cálculo de la puntuación de evaluación de tarea**

1. Se calculó la media de puntuación de evaluación de todos los criterios por cada evaluador para obtener la puntuación individual.

2. Se calculó media/mediana de la puntuación de todos los evaluadores por grupo evaluado para obtener la puntuación del colectivo.

- **Cálculo del rating de confianza del evaluador**

1. Se calculó la media de puntuación de confianza de todos los criterios por cada evaluador para obtener la puntuación individual.
2. Se calculó media/mediana de la puntuación de todos los evaluadores por estudiante.

Posteriormente se evaluó la normalidad de los datos mediante las pruebas de Saphiro y los términos de simetría y curtosis para determinar si la puntuación de evaluación del colectivo (Puntuación Recibida, Rating de confianza) debe ser calculada con media o mediana.

Los términos de simetría y curtosis van a indicar la desviación en términos de forma que la distribución que se evalúa tiene respecto a una distribución normal o gaussiana. La distribución es normal si los valores de simetría y curtosis se aproximan a cero, y que si p-valor es mayor a 0.05 la distribución será normal.

Materiales

Para implementar el experimento, se usó:

- Bibliotecas de Python que se ejecutaron en Jupyter Notebook v. 6.1.4:
- NumPy para la vectorización de columnas en forma de panda.
- Pandas para leer CSV.
- Métricas de SK-Learn para comparación de modelos.
- NLTK para procesamiento de texto.
- Pickle para guardar y cargar modelos de aprendizaje automático.
- Cadena, unidecode y puntuación para agregar vocabulario de Stop-Words.
- Skfuzzy para aplicar lógica difusa.
- Software de análisis de datos (SPSS) v. 25 para análisis descriptivo.

Resultados

A continuación, se detalla los resultados de cada fase:

Análisis de sentimiento

En esta fase se refinó lo publicado de esta investigación en [300], y se obtuvo dos modelos predictivos con el mejor rendimiento para generar una puntuación de sentimiento (1-positivo/0-neutral/-1-negativo) que corresponde a una retroalimentación textual especifica de evaluación de tarea o de evaluación de calidad de la evaluación, utilizando el enfoque de aprendizaje supervisado con algoritmos de aprendizaje profundo LSTM y Bi-LSTM a través de bibliotecas de Python.

Ingeniería de características, entrenamiento y evaluación del modelo

Para el entrenamiento del modelo de evaluación de tarea, se utilizó el conjunto de datos (D1-A), y para el entramiento del modelo de evaluación de la calidad de la evaluación se utilizó el conjunto de datos (D1-B) (Tabla **53**).

En ambos modelos se procesó cada retroalimentación, aplicando conversión a minúsculas y tratamiento de caracteres especiales con la biblioteca Unidecode, se eliminaron palabras vacías utilizando un diccionario de Stop-Words creado llamado Stopwords.txt y se dividieron las retroalimentaciones en palabras o tokens mediante la herramienta \nltk.tokenize del framework NLTK.

Una vez que las retroalimentaciones fueron normalizadas se aplicó word embeding con el conjunto de datos etiquetados (Baseline), Word2Vec/Glove preentrenado (Enriquecimiento Semántico) de [301] para representar y entrenar los algoritmos de aprendizaje profundo (LSTM/Bi-LSTM).

Para ambos conjuntos de datos se definió modelos de tipo secuencial, se usó, un máximo de 10000 palabras en el vocabulario. La capa de entrada tuvo una dimensión de 150-240 x 300, la cual representó las palabras (150-240) de cada retroalimentación, con una representación vectorial de tamaño 300.

Seguidamente se obtuvo la capa LSTM/Bi-LSTM, cada una contó con 32/64 neuronas. Cada una de estas capas recibió la capa de entrada con representación de Word2Vec/Glove.

Consecutivamente se aplicó una capa Flatten para abrir la matriz 2D y representarla como un vector 1D. Por último, se tuvo la capa de clasificación, compuesta por tres unidades neuronas y softmax como la función de activación.

Para evitar sobre entrenamiento se aplicó una técnica dropout con un valor de 0.5/0.8. El optimizador del modelo fue Adam y la función de error fue categorical_crossentropy y una tasa de aprendizaje de 0.0001. Se aplicaron 50 épocas de entrenamiento con lotes de tamaño 50. Simultáneamente, también se monitoreó la pérdida y precisión de los datos de validación (Figura **49** y Figura **52**).

La salida de esta red neuronal fue un vector disperso de longitud 3, por ejemplo [0,0,1]. El índice donde se encuentra el uno indica la clase predicha por la red.

Con el conjunto de datos de evaluación de tarea (D1-A) (Tabla **57**) se obtuvieron los siguientes resultados:

Tabla 57. Rendimiento de modelos de aprendizaje profundo (LSTM/Bi-LSTM) aplicados al conjunto de datos de evaluación de tarea (D1-A)

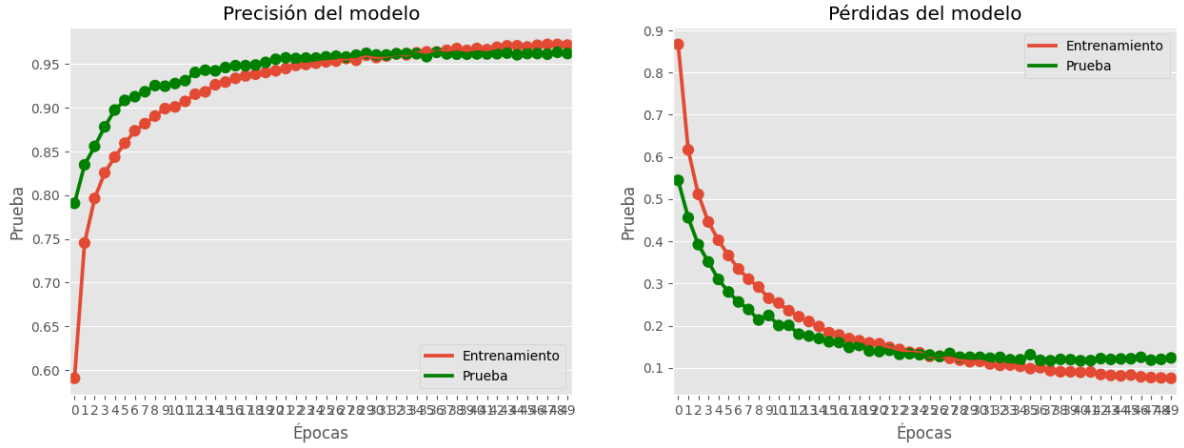
Word Embenddings	Entrada	Algoritmo	Unidad	Droup	Tasa de aprendizaje	Épocas (tamaño del lote)	Entrenamiento		Prueba				
							A	L	A	L	P	R	F1
Baseline	220	LSTM	64	0,8	0,0001	50 (50)	0,926	0,193	0,931	0,176	0,925	0,932	0,928
Baseline	230	LSTM	64	0,8	0,0001	50 (50)	0,931	0,182	0,929	0,184	0,925	0,931	0,927
Baseline	240	LSTM	64	0,8	0,0001	50 (50)	0,941	0,152	0,935	0,160	0,931	0,937	0,933
Baseline	220	Bi-LSTM	64	0,8	0,0001	50 (50)	0,934	0,173	0,935	0,167	0,931	0,935	0,933
Baseline	230	Bi-LSTM	64	0,8	0,0001	50 (50)	0,941	0,158	0,940	0,157	0,936	0,940	0,938
Baseline	240	Bi-LSTM	64	0,8	0,0001	50 (50)	0,940	0,161	0,938	0,158	0,934	0,939	0,936
Word2Vec	220	LSTM	64	0.8	0.0001	50 (50)	0.982	0.052	0.960	0.135	0.958	0.957	0.957
Word2Vec	230	LSTM	64	0.8	0.0001	50 (50)	0.984	0.051	0.960	0.132	0.957	0.958	0.957
Word2Vec	240	LSTM	64	0.8	0.0001	50 (50)	0.981	0.056	0.958	0.130	0.955	0.956	0.955
Word2Vec	220	Bi-LSTM	64	0.8	0.0001	50 (50)	0.986	0.043	0.954	0.154	0.951	0.951	0.951
Word2Vec	230	Bi-LSTM	64	0.8	0.0001	50 (50)	0.986	0.041	0.957	0.154	0.954	0.954	0.954
Word2Vec	240	Bi-LSTM	64	0.8	0.0001	50 (50)	0.985	0.042	0.954	0.167	0.952	0.951	0.952
Glove	220	LSTM	64	0.8	0.0001	50 (50)	0.961	0.107	0.962	0.114	0.959	0.960	0.960
Glove	230	LSTM	64	0.8	0.0001	50 (50)	0.963	0.105	0.961	0.116	0.958	0.958	0.958
Glove	240	LSTM	64	0.8	0.0001	50 (50)	0.962	0.106	0.960	0.114	0.958	0.957	0.957
Glove	220	Bi-LSTM	64	0.8	0.0001	50 (50)	0.973	0.071	0.965	0.117	0.963	0.964	0.963
Glove	230	Bi-LSTM	64	0.8	0.0001	50 (50)	0.973	0.072	0.962	0.121	0.960	0.961	0.960
Glove	240	Bi-LSTM	64	0.8	0.0001	50 (50)	0.975	0.072	0.963	0.119	0.961	0.960	0.960

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, L=Loss
P, R, F1, A: métricas a maximizar
L: métrica a minimizar
*Las mejores pruebas se resaltan en azul claro, las peores en verde y los mejores resultados en naranja

De un primer análisis con el conjunto de datos de evaluación de tarea (D1-A), el algoritmo Bi-LSTM obtuvo el mejor rendimiento en F-Measure (0.963), Precison (0.963), Recall (0.964), Accuracy (0.965) y Loss (0.117), utilizando representación de Glove, con una entrada de 220, 64 neuronas, droupt de 0.8, y una tasa de pérdida de 0.0001 (Tabla **58** y Figura **49**).

Tabla 58. Arquitectura Bi-LSTM con Glove del conjunto de datos D1-A

Model: "sequential"		
Layer (type)	**Output Shape**	**Param #**
Embedding	(None, 220, 300)	3000000
Bidirectional	(None, 220, 128)	186880
Flatten	(None, 28160)	0
Dense	(None, 3)	84483
Total params: 3,271,363 Trainable params: 3,271,363 Non-trainable params: 0		



(a) Precisión del entrenamiento vs validación (b) Pérdida del entrenamiento vs validación

Figura 49. Rendimiento del modelo Bi-LSTM con Glove del conjunto de datos D1-A

Posteriormente se aplicó un análisis estadístico basado en la prueba de Friedman [302][303] considerando todas las métricas para poder rechazar la hipótesis nula: no existe diferencia significativa en el rendimiento de los algoritmos de aprendizaje profundo.

Tabla 59. Prueba de Friedman para determinar si existen diferencias estadísticamente significativas entre los algoritmos de clasificación aplicados al conjunto de datos (D1-A)

Parámetros	Valor
N	36
Chi-cuadrado	4.000
gl	1
Sig. asintótica	0.046

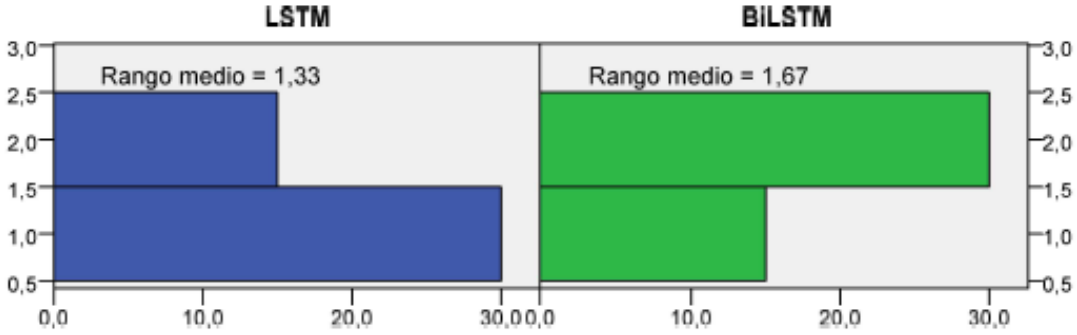


Figura 50. Rango promedio de algoritmos de aprendizaje profundo aplicados al conjunto de datos (D1-A)

El estadístico Chi-cuadrado es 4.000 y el valor de P es 0.046 con un grado de libertad, por lo que se rechaza la hipótesis nula (valor p<0.05) (Tabla **59** y Figura **50**). Por lo tanto, existe diferencia significativa en el rendimiento de los algoritmos de aprendizaje profundo.

Para llevar a cabo un análisis más detallado en relación con las parametrizaciones, se realizó un análisis estadístico (Tabla **60**,Tabla **61**) considerando todas las métricas para poder rechazar la hipótesis nula: no hay diferencias entre las distintas parametrizaciones llevadas a cabo para formar el vocabulario de cada conjunto de datos.

Tabla 60. Rango promedio de las parametrizaciones del conjunto de datos (D1-A)

N°	Algoritmo/Parámetro	Rango promedio
1	Bi-LSTM-Glove	6,00
2	LSTM-Glove	5,00
3	LSTM-Word2vec	4,00
4	Bi-LSTM-Word2vec	3,00
5	Bi-LSTM-Baseline	2,00
6	LSTM-Baseline	1,00

Tabla 61. Prueba de Friedman para determinar si existen diferencias estadísticamente significativas entre las distintas parametrizaciones aplicadas al conjunto de datos (D1-A)

Parámetros	Valor
N	12
Chi-cuadrado	60,000
gl	5
Sig. asintótica	,000

Al realizar la prueba de Friedman se observó que considerando una confianza del 95%, el valor P es 0.000<0.05 (Tabla **60**,Tabla **61**) se rechaza la hipótesis nula. Por lo tanto, existe diferencias estadísticamente significativas entre las distintas parametrizaciones llevadas a cabo para la construcción de los conjuntos de datos. Una vez confirmado que existen diferencias significativas, se aplicó Friedman para ANOVA de dos factores para determinar entre cuales parametrizaciones existe diferencia significativa (Tabla **62** y Figura **51**).

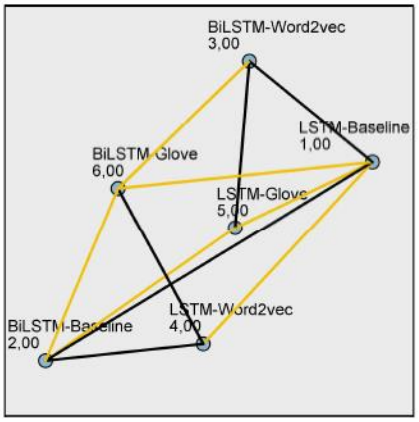


Figura 51. Comparativa entre parametrizaciones aplicadas al conjunto de datos (D1-A)

Tabla 62. Comparativa de parametrizaciones mediante prueba de Friedman para ANOVA de dos factores aplicadas al conjunto de datos (D1-A)

Muestra 1-Muestra 2	Estadístico de contraste	Error Error	Desv. Estadístico de contraste	Sig.	Sig. ajust.
LSTM-Baseline-BiLSTM-Baseline	-1,000	,764	-1,309	,190	1,000
LSTM-Baseline-BiLSTM-Word2vec	-2,000	,764	-2,619	,009	,132
LSTM-Baseline-LSTM-Word2vec	3,000	,764	3,928	,000	,001
LSTM-Baseline-LSTM-Glove	4,000	,764	5,237	,000	,000
LSTM-Baseline-BiLSTM-Glove	-5,000	,764	-6,547	,000	,000
BiLSTM-Baseline-BiLSTM-Word2vec	1,000	,764	1,309	,190	1,000
BiLSTM-Baseline-LSTM-Word2vec	2,000	,764	2,619	,009	,132
BiLSTM-Baseline-LSTM-Glove	3,000	,764	3,928	,000	,001
BiLSTM-Baseline-BiLSTM-Glove	4,000	,764	5,237	,000	,000
BiLSTM-Word2vec-LSTM-Word2vec	1,000	,764	1,309	,190	1,000
BiLSTM-Word2vec-LSTM-Glove	2,000	,764	2,619	,009	,132
BiLSTM-Word2vec-BiLSTM-Glove	3,000	,764	3,928	,000	,001
LSTM-Word2vec-LSTM-Glove	1,000	,764	1,309	,190	1,000

*Las comparaciones resaltadas de color amarillo son las que tienen diferencias significativas entre sí.

Se observa que las parametrizaciones construidas bajo la metodología de enriquecimiento semántico (Word2Vec/Glove) tienen diferencias significativas con las parametrizaciones construidas bajo la metodología Baseline. Según el rango promedio obtenido en la prueba de Friedman, el enriquecimiento semántico es el más adecuado para llevar a cabo la tarea de clasificación de sentimiento de una retroalimentación textual. También se obtuvo diferencia

significativa entre la representación de Glove y Word2Vec aplicando el algoritmo Bi-LSTM, corroborando el primer análisis que con representación de Glove se obtiene un mejor rendimiento.

Con el conjunto de datos de evaluación de calidad de evaluación (D1-B) (Tabla **63**) se obtuvieron los siguientes resultados:

Tabla 63. Rendimiento de modelos de aprendizaje profundo (LSTM/Bi-LSTM) aplicados al conjunto de datos de evaluación de calidad de evaluación (D1-B)

Word Embenddings	Entrada	Algoritmo	Unidad	Droup	Tasa de aprendizaje	Épocas (tamaño del lote)	Entrenamiento		Prueba				
							A	L	A	L	P	R	F1
Baseline	150	LSTM	32	0.5	0.0001	50 (50)	0.946	0.174	0.947	0.179	0.956	0.958	0.957
Baseline	180	LSTM	32	0.5	0.0001	50 (50)	0.935	0.206	0.948	0.196	0.957	0.959	0.958
Baseline	200	LSTM	32	0.5	0.0001	50 (50)	0.952	0.162	0.952	0.152	0.958	0.960	0.959
Baseline	150	Bi-LSTM	32	0.5	0.0001	50 (50)	0.936	0.196	0.949	0.171	0.957	0.959	0.958
Baseline	180	Bi-LSTM	32	0.5	0.0001	50 (50)	0.956	0.149	0.955	0.154	0.966	0.967	0.966
Baseline	200	Bi-LSTM	32	0.5	0.0001	50 (50)	0.945	0.189	0.948	0.173	0.954	0.956	0.955
Word2Vec	150	LSTM	32	0.5	0.0001	50 (50)	0.998	0.005	0.978	0.119	0.976	0.977	0.976
Word2Vec	180	LSTM	32	0.5	0.0001	50 (50)	0.999	0.005	0.977	0.120	0.975	0.976	0.975
Word2Vec	200	LSTM	32	0.5	0.0001	50 (50)	0.999	0.005	0.976	0.125	0.974	0.975	0.975
Word2Vec	150	Bi-LSTM	32	0.5	0.0001	50 (50)	0.999	0.005	0.973	0.146	0.973	0.971	0.972
Word2Vec	180	Bi-LSTM	32	0.5	0.0001	50 (50)	0.999	0.004	0.974	0.153	0.973	0.972	0.973
Word2Vec	200	Bi-LSTM	32	0.5	0.0001	50 (50)	0.999	0.005	0.971	0.155	0.969	0.970	0.969
Glove	150	LSTM	32	0.5	0.0001	50 (50)	0.993	0.024	0.977	0.080	0.976	0.975	0.975
Glove	180	LSTM	32	0.5	0.0001	50 (50)	0.992	0.025	0.981	0.077	0.980	0.980	0.980
Glove	200	LSTM	32	0.5	0.0001	50 (50)	0.991	0.026	0.977	0.081	0.976	0.975	0.976
Glove	150	Bi-LSTM	32	0.5	0.0001	50 (50)	0.995	0.015	0.978	0.090	0.975	0.978	0.977
Glove	180	Bi-LSTM	32	0.5	0.0001	50 (50)	0.995	0.014	0.978	0.095	0.976	0.977	0.977
Glove	200	Bi-LSTM	32	0.5	0.0001	50 (50)	0.995	0.015	0.979	0.097	0.977	0.977	0.977

P=Precision, R=Recall, F1=F-Measure, A=Accuracy, L=Loss
P, R, F1, A: métricas a maximizar
L: métrica a minimizar
*Las mejores pruebas se resaltan en azul claro, las peores en verde y los mejores resultados en naranja.

De un primer análisis con el conjunto de datos de evaluación de calidad de la evaluación (D1-B), el algoritmo LSTM obtuvo el mejor rendimiento en F-Measure (0.980), Precison (0.980), Recall (0.980), Accuracy (0.981) y Loss (0.077), utilizando representación de Glove, con una entrada de 180, 32 neuronas, droupt de 0.5, y una tasa de pérdida de 0.0001 (Tabla **64** y Figura **52**).

Tabla 64. Arquitectura LSTM con Glove del conjunto de datos D1-B

Model: "sequential"		
Layer (type)	**Output Shape**	**Param #**
Embedding	(None, 180, 300)	3000000
LSTM	(None, 180, 32)	42624
Flatten	(None, 5760)	0
Dense	(None, 3)	17283
Total params: 3,059,907 Trainable params: 3,059,907 Non-trainable params: 0		

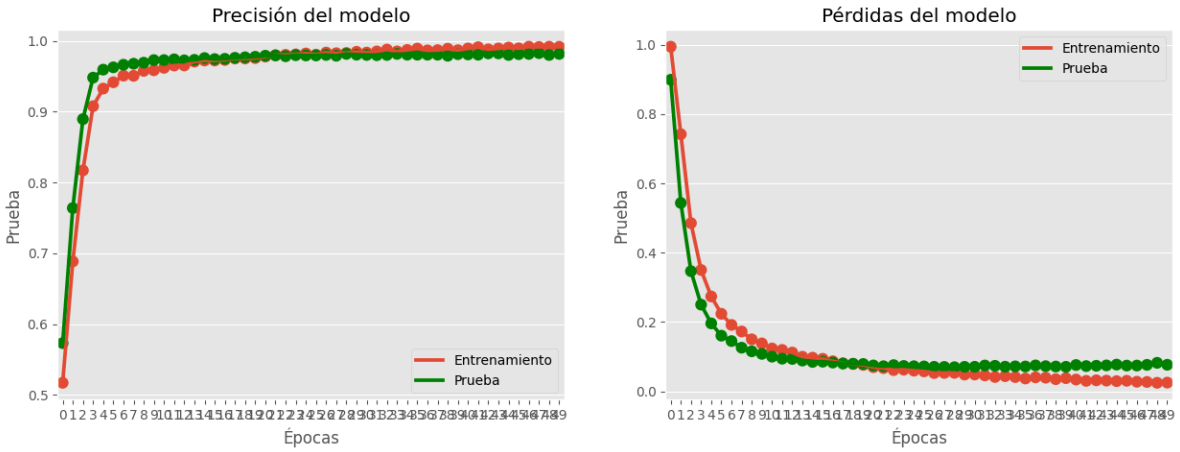


Figura 52. Rendimiento del modelo LSTM con Glove del conjunto de datos D1-B

Posteriormente se aplicó un análisis estadístico basado en la prueba de Friedman [302][303] considerando todas las métricas para poder rechazar la hipótesis nula: no existe diferencia significativa en el rendimiento de los algoritmos de aprendizaje profundo.

Tabla 65. Prueba de Friedman para determinar si existen diferencias estadísticamente significativas entre los algoritmos de clasificación aplicados al conjunto de datos (D1-B)

Parámetros	**Valor**
N	36
Chi-cuadrado	1.000
gl	1
Sig. asintótica	0.317

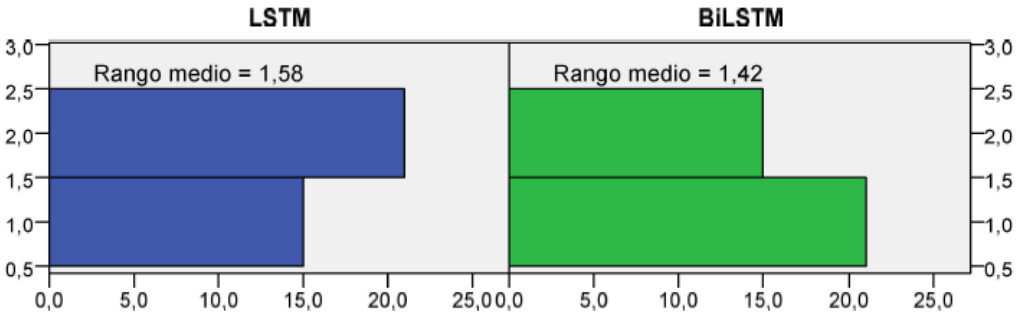


Figura 53. Rango promedio de algoritmos de aprendizaje profundo aplicados al conjunto de datos (D1-B)

El estadístico Chi-cuadrado es 1.000 y el valor de P es 0.317 con un grado de libertad, por lo que se retiene la hipótesis nula (valor p>0.05) (Tabla **65** y Figura **53**), Por lo tanto, al conjunto de datos (D1-B) se le puede aplicar LSTM o Bi-LSTM.

Para llevar a cabo un análisis más detallado en relación con las parametrizaciones, se realizó un análisis estadístico considerando todas las métricas para poder rechazar la hipótesis nula: no hay diferencias entre las distintas parametrizaciones llevadas a cabo para formar el vocabulario de cada conjunto de datos (Tabla **66**, Tabla **67**).

Tabla 66. Rango promedio de las parametrizaciones del conjunto de datos (D1-B)

N°	Algoritmo/Parámetro	Rango promedio
1	Bi-LSTM-Glove	5.46
2	LSTM-Glove	5.08
3	LSTM-Word2vec	4.46
4	Bi-LSTM-Word2vec	3.00
5	Bi-LSTM-Baseline	1.67
6	LSTM-Baseline	1.33

Tabla 67. Prueba de Friedman para determinar si existen diferencias estadísticamente significativas entre las distintas parametrizaciones aplicadas al conjunto de datos (D1-B)

Parámetros	Valor
N	12
Chi-cuadrado	53.496
gl	5
Sig. asintótica	0.000

Al realizar la prueba de Friedman se observó que considerando una confianza del 95%, el valor P es 0.000<0.05 (Tabla **67**), Por lo tanto, se confirma que existen diferencias estadísticamente significativas entre las distintas parametrizaciones llevadas a cabo para la construcción de los conjuntos de datos. Una vez confirmado que existen diferencias significativas, se aplicó Friedman para ANOVA de dos factores para determinar entre cuales parametrizaciones existe diferencia significativa (Tabla **68**, Figura **54**).

Tabla 68. Comparativa de parametrizaciones mediante prueba de Friedman para ANOVA de dos factores aplicadas al conjunto de datos (D1-B)

Muestra 1-Muestra 2	Estadístico de contraste	Error Error	Desv. Estadístico de contraste	Sig.	Sig. ajust.
LSTM-Baseline-BiLSTM-Baseline	-,333	,764	-,436	,663	1,000
LSTM-Baseline-BiLSTM-Word2vec	-1,667	,764	-2,182	,029	,436
LSTM-Baseline-LSTM-Word2vec	-3,125	,764	-4,092	,000	,001
LSTM-Baseline-LSTM-Glove	-3,750	,764	-4,910	,000	,000
LSTM-Baseline-BiLSTM-Glove	-4,125	,764	-5,401	,000	,000
BiLSTM-Baseline-BiLSTM-Word2vec	-1,333	,764	-1,746	,081	1,000
BiLSTM-Baseline-LSTM-Word2vec	-2,792	,764	-3,655	,000	,004
BiLSTM-Baseline-LSTM-Glove	-3,417	,764	-4,473	,000	,000
BiLSTM-Baseline-BiLSTM-Glove	-3,792	,764	-4,964	,000	,000
BiLSTM-Word2vec-LSTM-Word2vec	1,458	,764	1,909	,056	,843
BiLSTM-Word2vec-LSTM-Glove	-2,083	,764	-2,728	,006	,096
BiLSTM-Word2vec-BiLSTM-Glove	-2,458	,764	-3,219	,001	,019
LSTM-Word2vec-LSTM-Glove	-,625	,764	-,818	,413	1,000
LSTM-Word2vec-BiLSTM-Glove	-1,000	,764	-1,309	,190	1,000
LSTM-Glove-BiLSTM-Glove	-,375	,764	-,491	,623	1,000

*Las comparaciones resaltadas son las que tienen diferencias entre sí.

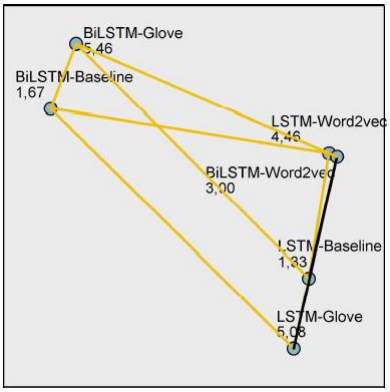


Figura 54. Comparativa entre parametrizaciones aplicadas al conjunto de datos (D1-B)

Se observa que las parametrizaciones construidas bajo la metodología de enriquecimiento semántico (Word2Vec/Glove) tienen diferencias significativas con las parametrizaciones construidas bajo la metodología Baseline (Tabla **68**, Figura **54**). Según el rango promedio obtenido en la prueba de Friedman, el enriquecimiento semántico es el más adecuada para llevar a cabo la tarea de clasificación de sentimiento de una retroalimentación textual. También se obtuvo diferencia significativa entre la representación de Glove y Word2Vec aplicando el algoritmo Bi-LSTM, corroborando el primer análisis que con representación de Glove se obtiene un mejor rendimiento.

Se evaluó la portabilidad del modelo predictivo de evaluación de tarea entre asignaturas, con la actividad FO30 del conjunto de datos (D2-A) (Tabla **53**). La verificación se realizó con las reglas de la subsección 3.5. La puntuación (1, positivo) de la retroalimentación (F1) es correcta porque contiene la palabra “si”. La puntuación (0, neutral) de la retroalimentación (F2) es correcta porque contiene la palabra "si” y “mejorar”. La puntuación (0, neutral) de la retroalimentación (F3) es correcta porque contiene la palabra "parcialmente correcta” y “falta”. La retroalimentación (-1, negativo) de la retroalimentación (F4) es correcto porque contiene la palabra “no” y “falta (Tabla **69**). Los resultados muestran que el modelo predictivo generó la puntuación de sentimiento correcta.

Tabla 69. Ejemplos de puntuación de sentimiento generada por el modelo LSTM (Glove) de la actividad-FO30 (Diseño de aulas virtuales)

Evaluador	Evaluado	Criterio	Retroalimentación	Puntuación Sentimiento
E-1	Grupo-3	Datos educativos	F1: El tema de la clase está en concordancia con la asignatura. Los objetivos y los resultados, si responden con la asignatura.	1
		Materiales de clase	F2: En el material de clases, si se evidencia la aplicación del sitio web, presentaciones, y organizador gráfico, se recomienda mejorar con títulos referentes al tema de clases	0
		Diseño de actividades	F3: En las actividades se evidencia propuesta de trabajo colaborativo con herramientas de manera parcialmente correcta, falta que todos los integrantes del grupo participen, deben estar registrados como profesor.	0
		Test de evaluación	F4: En el test se evidencia la aplicación de herramientas en línea de manera parcialmente correcta, no genera calificación, falta que el estudiante registre sus datos.	-1

Detección de precisión/imprecisión entre puntuación de sentimiento y numérica, y cálculo de puntuación de evaluación/puntuación de confianza por cada criterio mediante lógica difusa

En esta fase se refinó lo publicado de esta investigación en [300]**,** con la finalidad de detectar precisión/imprecisión entre puntuación de sentimiento y numérica, y cuando exista imprecisión equilibrar estas dos puntuaciones, no penalizar para obtener por cada criterio puntuación de evaluación de tarea y puntuación de confianza de evaluación de calidad de la evaluación, utilizando el enfoque de Mamdani en lógica difusa a través de las bibliotecas de Python.

Valor nítido (datos)

Se utilizó las variables de entrada: Puntuación Numérica y Puntuación Sentimiento.

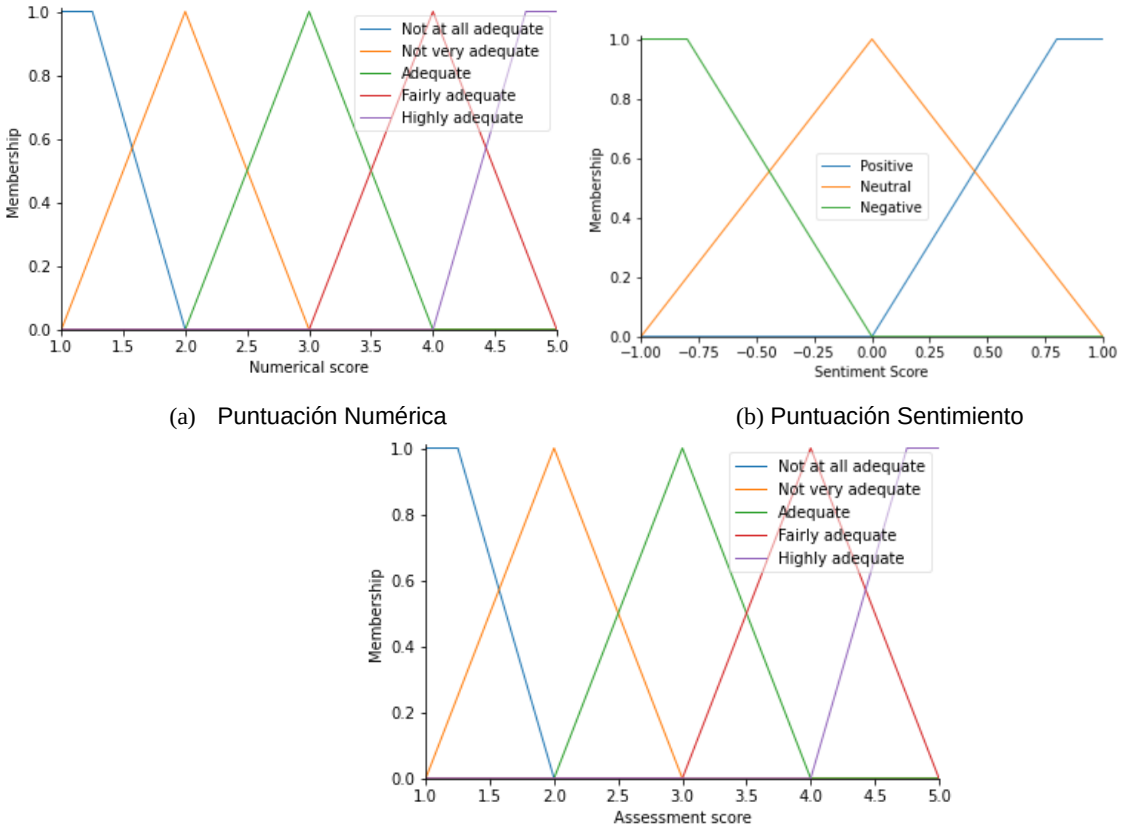
Fuzzificación (valor difuso de entrada y salida)

Las variables de entrada (Puntuación Numérica y Puntuación Sentimiento) y la variable de salida (Puntuación Evaluación/Puntuación Confianza) se dividieron en términos lingüísticos y se

combinaron en un par de funciones de membresía: gamma invertida y gamma para los bordes, y función triangular para valores medios. Esta combinación permite cubrir todas las escalas (1…5), (1…−1) y (0.2…1). En la Tabla **70** se enlista los términos lingüísticos con parámetros asignados a cada variable lingüística, y en la Figura **55** y Figura **56** se presenta los gráficos de las funciones de pertenencia asociadas para obtener la Puntuación Evaluación/Puntuación Confianza de cada criterio.

Tabla 70. Variables de entrada y salida en términos lingüísticos

Representación de variable	**Variable lingüística**	**Término lingüístico**	**Función de membresía**	**Parámetros**
× 1	Puntuación Numérica	Nada adecuado	trapezoide	[1, 1,1.25, 2]
		No muy adecuado	Triangular	[1,2,3]
		Adecuado	Triangular	[2,3,4]
		Bastante adecuado	Triangular	[3,4,5]
		Muy adecuado	trapezoide	[4,4.75,5,5]
× 2	Puntuación Sentimiento	Positivo	trapezoide	[0,0.8,1,1]
		Neutral	Triangular	[-1,0,1]
		Negativo	trapezoide	[-1,-1,-0.8,0]
y	Puntuación Evaluación	Nada adecuado	trapezoide	[1, 1,1.25, 2]
		No muy adecuado	Triangular	[1,2,3]
		Adecuado	Triangular	[2,3,4]
		Bastante adecuado	Triangular	[3,4,5]
		Muy adecuado	trapezoide	[4,4.75,5,5]
y	Puntuación Confianza	Bajo	trapezoide	[0,0,0.2,0.4]
		Medio bajo	Triangular	[0.2,0.4,0.6]
		Medio	Triangular	[0.4,0.6,0.8]
		Medio alto	Triangular	[0.6,0.8,1]
		Alto	trapezoide	[0.8,.9,1,1])

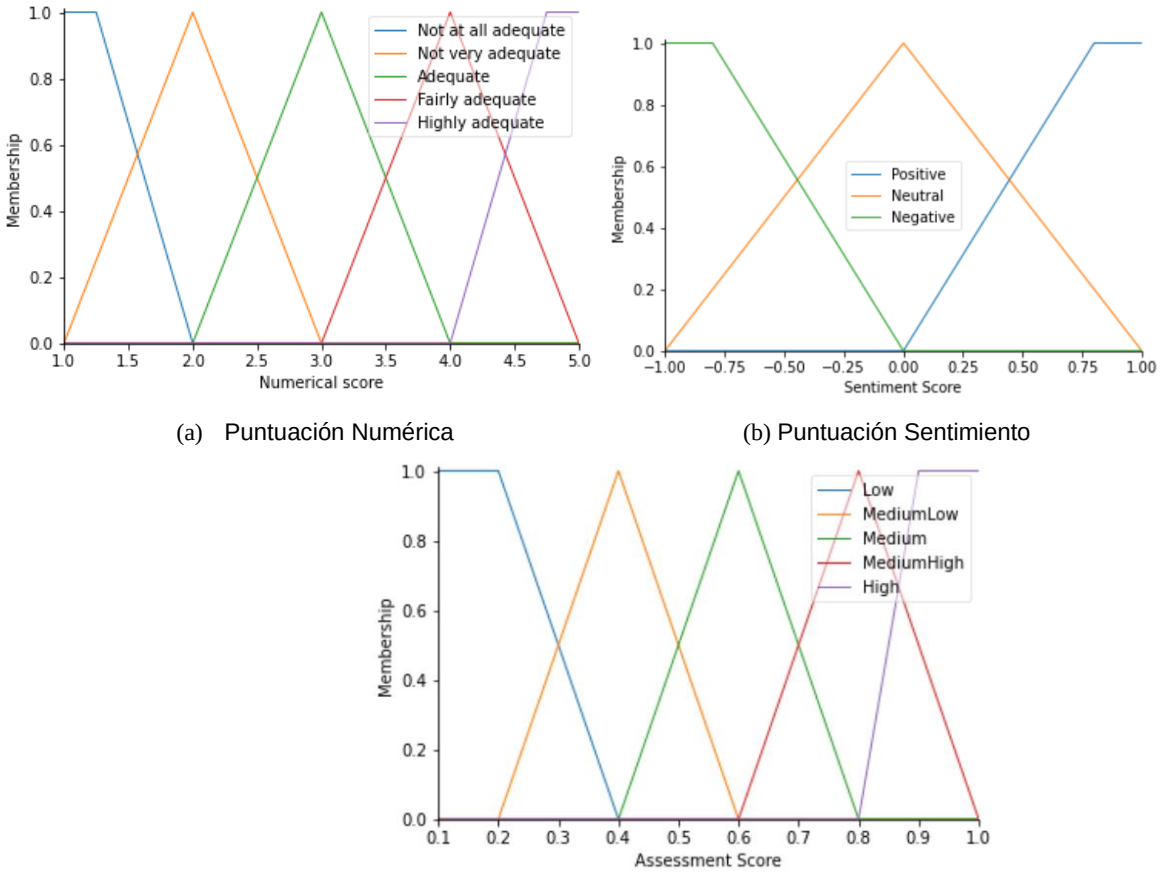


(a) Puntuación Numérica

(b) Puntuación Sentimiento

(c) Puntuación Evaluación

Figura 55. Ejemplos de funciones de membresía aplicadas: (a) Puntuación Numérica, (b) Puntuación Sentimiento para obtener (c) Puntuación Evaluación de tarea por cada criterio



(a) Puntuación Numérica

(b) Puntuación Sentimiento

(c) Puntuación Evaluación

Figura 56. Ejemplos de funciones de membresía aplicadas: **(**a**)** Puntuación Numérica, **(**b**)** Puntuación Sentimiento para obtener **(**c**)** Puntuación Confianza de evaluación de calidad de evaluación por cada criterio

Reglas difusas

Las reglas lingüísticas se formularon a partir de que, si se detecta precisión o imprecisión entre puntuación de sentimiento y numérica, se calcula la puntuación de evaluación/confianza descrita en el paso 3 de la segunda fase de la sección de metodología de experimentación. La Figura **57** muestra las reglas difusas para obtener la puntuación de evaluación de tarea por cada criterio, y la Figura **58** muestra las reglas difusas para obtener la puntuación de confianza del evaluador por cada criterio.

```
# Rules
    # Detecting accuracy and generating the assessment score
    R1 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Negative'], Assessment_score['Not at all adequate'])
    R2 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
    R3 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
    R4 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Positive'], Assessment_score['Mainly adequate'])
    R5 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Positive'], Assessment_score['Highly adequate'])
    # Detecting inaccuracy and generating the assessment score
    R6 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Neutral'], Assessment_score['Not very adequate''])
    R7 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Positive'], Assessment_score['Not very adequate''])
    R8 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Neutral'], Assessment_score['Not very adequate''])
    R9 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Positive'], Assessment_score['Not very adequate''])
    R10 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
    R11 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Positive'], Assessment_score['Adequate'])
    R12 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
    R13 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
    R14 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Negative'], Assessment_score['Not very adequate'])
    R15 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
```

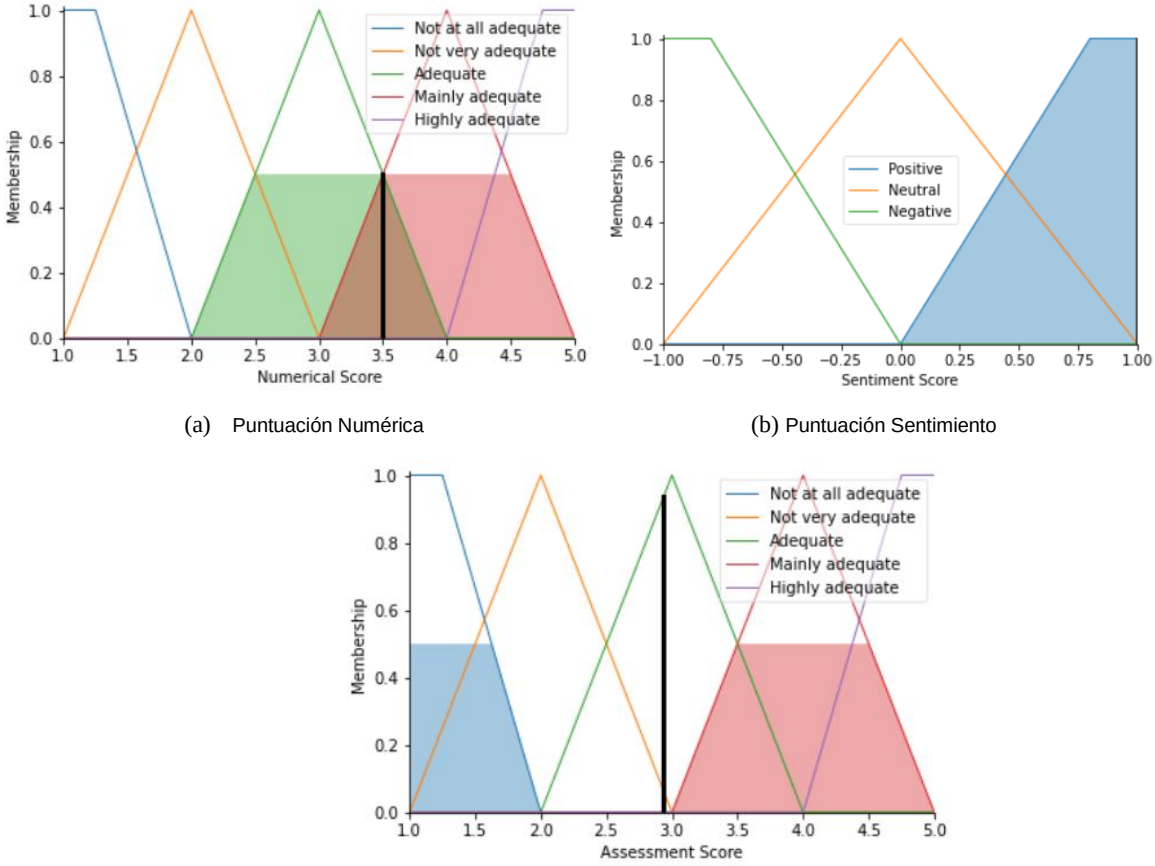
Figura 57. Reglas difusas para obtener la puntuación de evaluación de tarea por cada criterio desarrollado en Python

```
' Rules
  # Rules
    # Detecting accuracy and generating the assessment score
    R1 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Negative'], Assessment_score['Low'])
    R2 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Negative'], Assessment_score['MediumLow'])
    R3 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Neutral'], Assessment_score['Medium'])
    R4 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Positive'], Assessment_score['MediumHigh'])
    R5 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Positive'], Assessment_score['High'])
    # Detecting inaccuracy and generating the assessment score
    R6 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Neutral'], Assessment_score['MediumLow'])
    R7 = ctrl.Rule(Numerical_score['Not at all adequate'] & Sentiment_score['Positive'], Assessment_score['MediumLow'])
    R8 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Neutral'], Assessment_score['MediumLow'])
    R9 = ctrl.Rule(Numerical_score['Not very adequate'] & Sentiment_score['Positive'], Assessment_score['MediumLow'])
    R10 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Negative'], Assessment_score['MediumLow'])
    R11 = ctrl.Rule(Numerical_score['Adequate'] & Sentiment_score['Positive'], Assessment_score['Adequate'])
    R12 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Negative'], Assessment_score['MediumLow'])
    R13 = ctrl.Rule(Numerical_score['Mainly adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
    R14 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Negative'], Assessment_score['MediumLow'])
    R15 = ctrl.Rule(Numerical_score['Highly adequate'] & Sentiment_score['Neutral'], Assessment_score['Adequate'])
```

Figura 58. Reglas difusas para obtener la puntuación de confianza del evaluador por cada criterio desarrollado en Python

Defuzzificación

Una vez que se activan las reglas adecuadas en el sistema de control difuso, el grado de pertenencia de la variable difusa de salida (Puntuación Evaluación, Puntuación Confianza) se determina mediante la codificación de los subconjuntos difusos antecedentes (Puntuación Numérica y Puntuación Sentimiento). Se utilizó el método de inferencia max-min (Ecuación 8). Se aplicó el método LOM porque calcula el resultado más plausible.

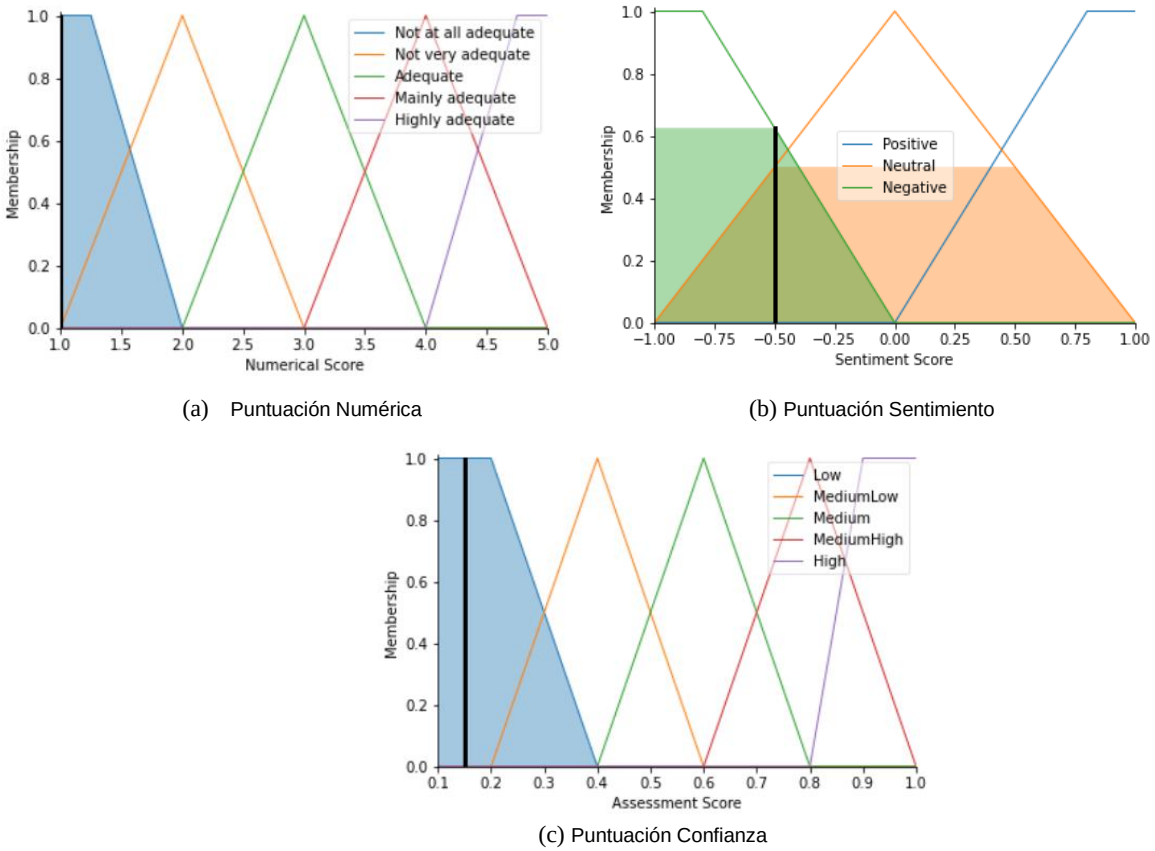


(a) Puntuación Numérica

(b) Puntuación Sentimiento

(c) Puntuación Evaluación

Figura 59. Ejemplo de respuesta del sistema de control difuso para el cálculo de (c) puntuación de evaluación de cada criterio de tarea. Los conjuntos difusos resultantes se derivan de las áreas de superficie resaltadas en azul, naranja y verde, y la línea negra indica el valor de puntuación de evaluación típico óptimo después del proceso de defuzzificación con el método LOM



(a) Puntuación Numérica

(b) Puntuación Sentimiento

(c) Puntuación Confianza

Figura 60. Ejemplo de respuesta del sistema de control difuso para el cálculo de (c) puntuación de confianza de cada criterio de evaluación de calidad de evaluación. Los conjuntos difusos resultantes se derivan de las áreas de superficie resaltadas en azul, naranja y verde, y la línea negra indica el valor de puntuación de confianza típico óptimo después del proceso de defuzzificación con el método LOM

Cálculo de puntuación individual y del colectivo de evaluación de tarea, e índice (rating) de confianza del evaluador

Se calculo la puntuación individual por cada evaluador, posteriormente se calcula la puntuación del colectivo de evaluación de tarea, e índice (rating) de confianza del evaluador con media y mediana. Finalmente se aplicó los coeficientes de curtosis, simetría y prueba de Saphiro. Los resultados mostraron que la distribución de las variables (Puntuación Recibida, Rating confianza) presenta una asimetría, deduciendo que la distribución de los datos no es normal, la media se

desplazará, mientras que la mediana al ser una medida mucho más robusta tenderá a concentrarse en donde los valores son los correctos. Esta afirmación se corroboró con los valores del p-valor que son menores a 0.05 en todas las asignaturas y periodos académicos (Tabla **71**) con lo cual se concluye que las distribuciones no son normales.

Tabla 71. Descripción de coeficiente curtosis, simetría y prueba de Saphiro aplicado al conjunto de datos por periodo académico, escenario de educación y asignatura

Escenario de educación	Periodo académico	Asignatura	Variables	Curtosis	Simetría	p-valor
Virtual asincrónico	Mayo-septiembre 2021	Fundamentos de ingeniería de software	Puntuación Recibida Media)	0.404	-0.516	1.786e-14
			Puntuación Recibida (mediana)	0.017	-0.230	3.232e-26
			Rating Confianza (Media)	-0.110	-0.046	3.798e-05
			Rating Confianza (mediana)	-0.581	-0.017	1.649e-23
	Octubre 2021-febrero 2022	Fundamentos de ingeniería de software	Puntuación Recibida (Media)	0.845	0.839	6.455e-22
			Puntuación Recibida (mediana)	0.020	-0.618	2.944e-26
			Rating Confianza (Media)	1.355	-0.792	3.026e-19
			Rating Confianza (mediana)	0.065	-0.776	2.040e-33
		Ingeniería de software	Puntuación Recibida (Media)	0.188	-0.737	2.895e-16
			Puntuación Recibida (mediana)	-0.158	-0.378	1.340e-19
			Rating Confianza (Media)	-0.845	-0.003	5.35e-10
			Rating Confianza (mediana)	-0.971	-0.480	8.615e-27
Virtual sincrónico	Mayo-septiembre 2022	Fundamentos de ingeniería de software	Puntuación Recibida (Media)	-0.233	-0.423	9.562e-14
			Puntuación Recibida (mediana)	-0.88	-0.178	4.930e-25
			Rating Confianza (Media)	-0.104	-0.412	8.108e-15
			Rating Confianza (mediana)	-0.505	-0.406	2.488e-21
		Fundamentos de ofimática	Puntuación Recibida (Media)	2.064	-1.300	1.529e-16
			Puntuación Recibida (mediana)	0.496	-0.993	9.277e-21
			Rating Confianza (Media)	-0.290	-0.629	1.834e-12
			Rating Confianza (mediana)	-0.466	-0.744	4.059e-18
Presencial	Mayo-septiembre 2022	Ingeniería de software	Puntuación Recibida (Media)	0.160	-0.600	3.373e-09
			Puntuación Recibida (mediana)	-0.769	-0.052	2.239e-10
			Rating Confianza (Media)	-0.545	-0.241	1.362e-07
			Rating Confianza (mediana)	-0.994	-0.194	8.577e-12
		Fundamentos de programación	Puntuación Recibida (Media)	-1.629	0.147	1.309e-05
			Puntuación Recibida (mediana)	-0.701	-0.790	6.680e-07
			Rating Confianza (Media)	0.954	-1.307	2.801e-06
			Rating Confianza (mediana)	0.565	-1.308	8.213e-08
		Gestión de procesos de negocios y sistemas empresariales	Puntuación Recibida (Media)	-0.717	0.591	0.000
			Puntuación Recibida (mediana)	-1.409	-0.144	4.418e-06
			Rating Confianza (Media)	-1.107	-0.212	0.000
			Rating Confianza (mediana)	-1.122	-0.333	3.437e-05

En la Figura **61** y Figura **62**, se presentan un ejemplo de histogramas y diagramas de caja de las variables consideradas dentro del análisis de la asignatura de fundamentos de ingeniería de software del periodo académico octubre 2021-febrero 2022. En los histogramas, las distribuciones no siguen una tendencia Gaussiana, además presentan una gran cantidad de

valores extremos que hacen que la media cambie su magnitud, en contraste la mediana no se ve afectada por valores muy grandes o pequeños, por lo tanto, la mediana es robusta y filtra los valores extremos.

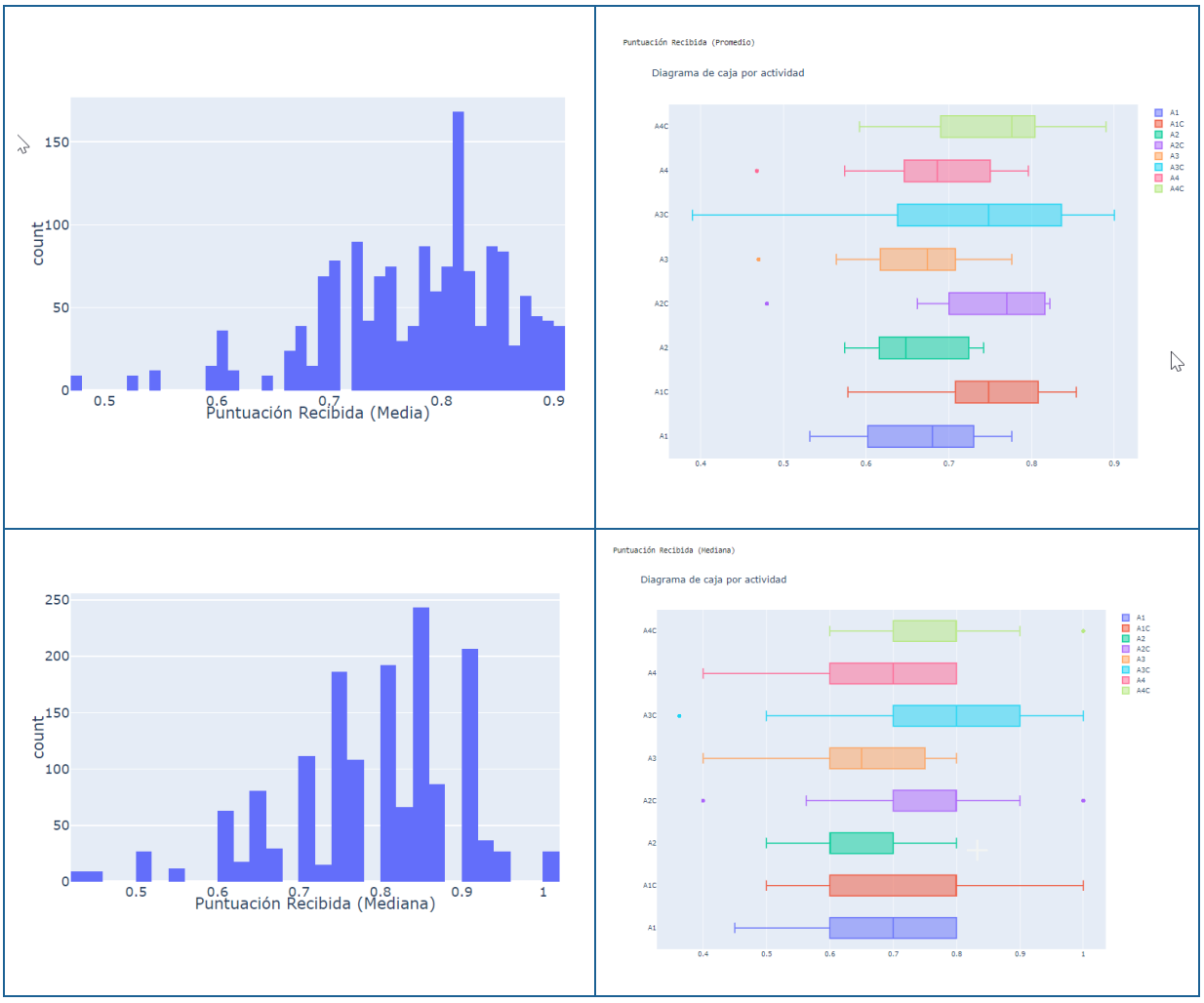


Figura 61. Ejemplo de histogramas y diagramas de caja de Puntuación Recibida (media/mediana) de la asignatura de fundamentos de ingeniería de software del periodo académico octubre 2021-febrero 2022

Figura 62. Ejemplo de histogramas y diagramas de caja de Rating Confianza (media/mediana) de la asignatura de fundamentos de ingeniería de software del periodo académico octubre 2021-febrero 2022

Conclusiones, limitaciones y futuro experimento

- Se entrenó los algoritmos LSTM/Bi-LSTM con enriquecimiento semántico Word2Vec/Glove preentrenado. Los resultados revelaron que el modelo que predice la puntuación de sentimiento de evaluación de tarea con mejor rendimiento fue Bi-LSTM con Glove (F-Measure de 0.963), y que el modelo que predice la puntuación de sentimiento de evaluación de calidad de evaluación con mejor rendimiento fue LSTM con Glove (F-Measure de 0.980). Por lo tanto,

las representaciones con Word2Vec/Glove superaron al modelo baseline SVM (F-Measure de 0.879) con representación de 1-g y 2-g+TF-IDF de esta investigación, que han sido publicados en [300]. Deduciendo que la extracción de características si influye en el rendimiento de los algoritmos.

- Además, se evidenció que el tamaño de la muestra si influye en el rendimiento de los algoritmos de aprendizaje profundo, los algoritmos Bi-LSTM (F-Measure de 0.963) con una muestra de 21280 estancias de evaluación de tarea, supero a LSTM (F-Measure de 0.832) con una muestra de 3,968 estancias de evaluación de tarea de esta investigación, que han sido publicados en [300].
- Se asevera que el modelo predictivo es portable a los resultados previos de esta investigación, que han sido publicados en [300], ya que generó la puntuación de sentimiento correcta con datos de otras asignaturas, y escenarios de educación virtual asincrónico, sincrónico y presencial.
- El pequeño tamaño de la muestra limita los resultados de este experimento.
- En el trabajo futuro, se planea:
  - Aplicar otros algoritmos que mejore la detección de la negación y realice la predicción del sentimiento atendiendo a la intensidad de su polaridad en diferentes clases (fuertemente negativo, negativo, neutro, positivo y fuertemente positivo), y la clasificación sentimental mediante tópicos o características del texto, donde los parámetros que ponderan la clasificación están basados en los tópicos o características de los temas tratados en los textos.
  - Calibrar la puntuación de evaluación de la tarea, otorgando bonificación/penalización.
  - Seguimiento del comportamiento de la aplicación de media/mediana en el cálculo de puntuación del colectivo, en escenarios de educación superior.

4.1.4. Experimento IV

En este experimento se consideró evaluación calibrada (Figura **63**) del modelo de evaluación entre pares descrito en la subsección 3.2.

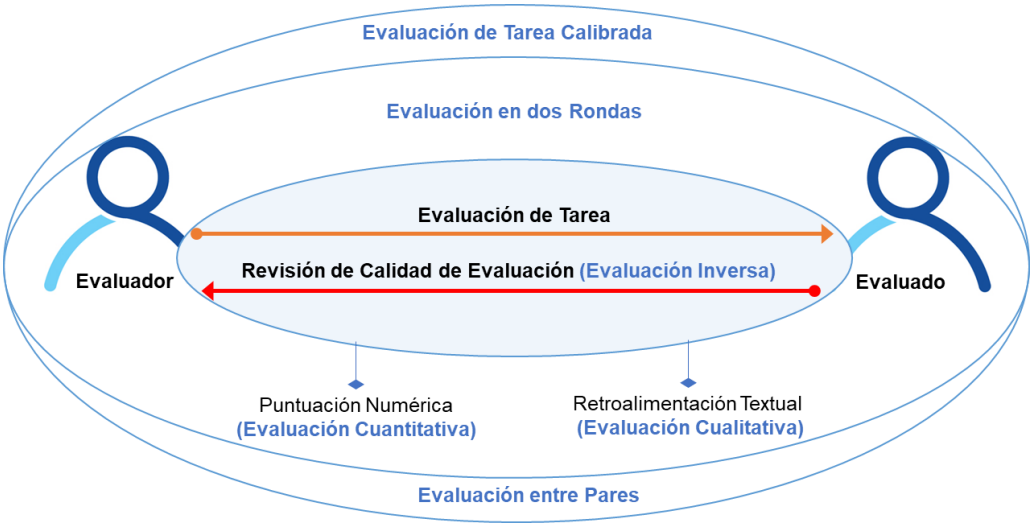


Figura 63. Evaluación entre pares calibrada

Planteamiento, pregunta de investigación y objetivos

Planteamiento
• Evaluar qué factores se pueden aplicar para calibrar la puntuación de evaluación de tarea. • Calibrar la puntuación de evaluación de tarea considerando los factores determinados.
Pregunta de investigación
• ¿Cómo calibrar la puntuación de evaluación de tarea, que establezca fiabilidad en el proceso de evaluación entre pares?
Objetivos
• Diseñar los artefactos para validar los métodos teóricos. • Construir un modelo de evaluación entre pares basado en análisis de sentimiento. • Evaluar los resultados de la precisión del modelo.

A continuación, se detalla el experimento:

Materiales y métodos

Conjunto de datos

Se consideró los datos recolectados en dos rondas de los periodos académicos: mayo-septiembre 2021 y octubre 2021-febrero 2022 en escenario de educación virtual asincrónica ante la pandemia COVID-19, y mayo-septiembre 2022 en escenario de educación virtual sincrónica y presencial.

Metodología de experimentación

La metodología aplicada en la experimentación se realizó en base a la subsección 3.1.7, que constó de dos fases. En la primera fase, se determina si la puntuación recibida proporcionada por los evaluadores, y rating de confianza dada por los evaluados son factores de calibración. En la segunda fase, se realizó la calibración de puntuación de evaluación de tarea en función de los factores determinados (Figura **64**).

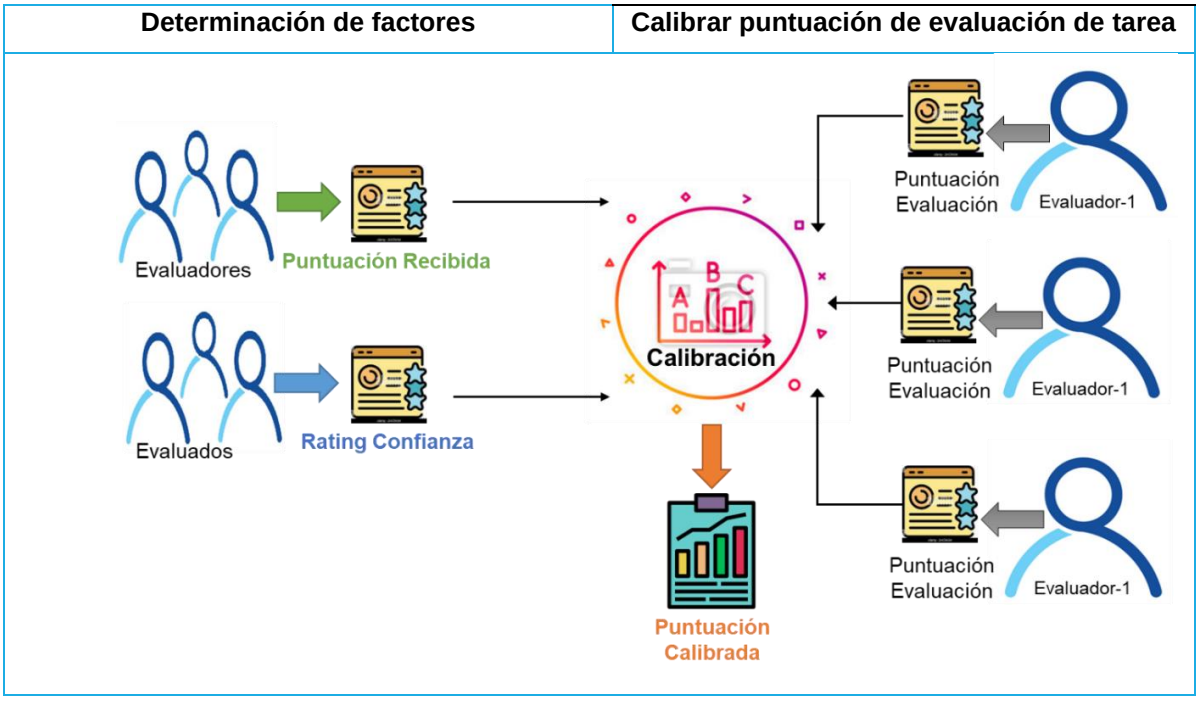


Figura 64. Metodología aplicada en la experimentación IV

A continuación, se describe con detalle el procedimiento llevado a cabo en cada una de las fases:

Primera fase: Valoración de factores para calibrar la puntuación de evaluación de tarea

En esta fase se evaluó la similaridad entre Puntuación Dada con Puntuación Recibida del colectivo, y entre Puntuación Dada por los evaluadores con Rating Confianza dada por los evaluados para determinar los factores de calibración. Se aplicó correlación de Pearson para determinar si existe relación lineal entre las variables (Puntuación Dada, Puntuación Recibida, Rating Confianza) por escenario de educación y asignatura (Tabla **74**). Dependiendo de si el valor es positivo o negativo, se establece si la relación es directa o inversa. Mientras que, si el coeficiente es cero, no existirá una relación entre las variables. Para interpretar los valores se utilizó los siguientes criterios; perfecta (IRI = 1), fuerte (1< IRI > = 0.7), sustancial (0.7< IRI >= 0.5), moderada (0.5 < IRI >= 0.3), débil (0.3 < IRI >= 0.1) o escasa/ninguna (0.1 < IRI >= 0.0).
Donde:
$\bar{X}$ = Media
Σ = Desviación Estándar
r = Correlación Pearson

Segunda fase: Calibración de puntuación de evaluación de tarea

Esta fase considera los pasos descritos en subsección 3.1.7, para obtener el modelo de calibración que recibe como entrada la Puntuación Dada por cada evaluador, a la que se bonificará o penalizará en función de cada perfil del estudiante (Figura **64**). Los cuatro pasos se resumen a continuación.

Paso 1. Normalización de datos

- Se normaliza los datos a escala de 1.

Paso 2. Especificación del perfil del estudiante

- Se categoriza los factores: Puntuación Recibida (Perfil Cognitivo) y Rating Confianza (Perfil Evaluador) que obtuvo el evaluador del colectivo (Tabla **72**).

Tabla 72. Categorización de los factores

Puntuación Recibida/Rating Confianza	Perfil Cognitivo/Evaluador	Categoría
0.8 < puntuación < =1	Alto	A
0.6 < puntuación <= 0.8	Medio alto	B
0.4 < puntuación <= 0.6	Medio	C
0.2 < puntuación <= 0.4	Medio bajo	D
0.0 < puntuación <= 0.2	Bajo	E

Paso 3. Cálculo de calibración de puntuación dada por cada evaluador

- Se calibra la Puntuación Dada de cada evaluador, adicionando o restando la proporción que se obtiene de varianza multiplicada por desviación estándar, de acuerdo al Perfil Cognitivo y de Evaluador (Tabla **73**).

Tabla 73. Bonificación o penalización de acuerdo a la categoría

Perfil Cognitivo/Evaluador	Categoría	Bonificación/Penalización	
Alto	A	var * std	Bonificación
Medio Alto	B	1/2 var * std	Bonificación
Medio	C		No se da bonificación, ni penalización
Medio Bajo	D	-1/2 var * std	Penalización
Bajo	E	-var * std	Penalización

A continuación, se detalla un ejemplo de puntuación calibrada con Perfil Cognitivo “A” y Perfil Evaluador de “A” hasta “E”**:**

- Si (Perfil Cognitivo es A) y (Perfil Evaluador es A), entonces (Puntuación Calibrada es Puntuación Evaluación + (var*std) + (var*std))
- Si (Perfil Cognitivo es A) y (Perfil Evaluador es B), entonces (Puntuación Calibrada es Puntuación Evaluación + (var*std) + (1/2 var*std))
- Si (Perfil Cognitivo es A) y (Perfil Evaluador es C), entonces Puntuación Calibrada es Puntuación Evaluación + (var*std))
- Si (Perfil Cognitivo es A) y (Perfil Evaluador es D), entonces (Puntuación Calibrada es Puntuación Evaluación + (var*std) - (1/2 var*std))
- Si (Perfil Cognitivo es A) y (Perfil Evaluador es E), entonces (Puntuación Calibrada es Puntuación Evaluación + (var*std) - (var*std))

Paso 4. Cálculo de puntuación calibrada del colectivo

- Se recalcula la media/mediana del conjunto de puntuaciones individuales de los evaluadores pares por grupo evaluado.

Materiales

Para implementar el experimento, se usó:

- Bibliotecas de Python que se ejecutaron en Jupyter Notebook v. 6.1.4:
- Software de análisis de datos (SPSS) v. 25 para análisis descriptivo.

Resultados

Valoración de factores para calibrar la puntuación de evaluación de tarea

Se calcularon los parámetros estadísticos para determinar la relación que existe entre las variables (Puntuación Dada, Puntuación Recibida, Rating Confianza) por escenario de educación y asignatura (Tabla **74**).

En las Figura **65**, Figura **66** y Figura **67**, se puede identificar similaridad entre las curvas que identifican a la puntuación que proporcionó y obtuvo del colectivo el estudiante. Se puede observar que, aunque muestran la misma tendencia de incrementar y reducir una en función de la otra, las magnitudes mantienen una aparente igualdad que se definió empleando la media (Tabla **74**).

Tabla 74. Descripción de parámetros estadísticos de las variables (Puntuación Dada, Puntuación Recibida y Rating Confianza) del conjunto de datos por periodo académico, escenario de educación y asignatura

Escenario de educación	Periodo académico	Asignatura	Variables	$\bar{X}$	Σ	r Puntuación Dada-Recibida	r Puntuación Dada-Rating Confianza
Virtual asincrónica	Mayo-septiembre 2021	Fundamentos de ingeniería de software	Puntuación Dada	0.707	0.195		
			Puntuación Recibida (Media)	0.708	0.092	0.009	
			Puntuación Recibida (mediana)	0.710	0.110	0.032	
			Rating Confianza (Media)	0.700	0.126		0.327
			Rating Confianza (mediana)	0.708	0.144		0.302
	Octubre 2021-	Fundamentos de ingeniería de software	Puntuación Dada	0.774	0.171		
			Puntuación Recibida (Media)	0.777	0.081	0.012	
			Puntuación Recibida (mediana)	0.788	0.107	0.024	

Escenario de educación	Periodo académico	Asignatura	Variables	$\bar{X}$	Σ	r Puntuación Dada-Recibida	r Puntuación Dada-Rating Confianza
	febrero 2022		Rating Confianza (Media)	0.785	0.118		0.348
			Rating Confianza (mediana)	0.799	0.142		0.340
		Ingeniería de software	Puntuación Dada	0.768	0.172		
			Puntuación Recibida (Media)	0.768	0.088	0.048	
			Puntuación Recibida (mediana)	0.769	0.104	0.043	
			Rating Confianza (Media)	0.782	0.119		0.444
			Rating Confianza (mediana)	0.796	0.145		0.443
Virtual sincrónico	Mayo-septiembre 2022	Fundamentos de ingeniería de software	Puntuación Dada	0.741	0.179		
			Puntuación Recibida (Media)	0.755	0.111	0.166	
			Puntuación Recibida (mediana)	0.757	0.133	0.182	
			Rating Confianza (Media)	0.773	0.141		0.290
			Rating Confianza (mediana)	0.773	0.141		0.290
		Fundamentos de ofimática	Puntuación Dada	0.700	0.190		
			Puntuación Recibida (Media)	0.777	0.143	0.179	
			Puntuación Recibida (mediana)	0.784	0.164	0.165	
			Rating Confianza (Media)	0.833	0.135		0.477
			Rating Confianza (mediana)	0.833	0.135		0.477
Presencial	Mayo-septiembre 2022	Ingeniería de software	Puntuación Dada	0.747	0.162		
			Puntuación Recibida (Media)	0.744	0.120	0.454	
			Puntuación Recibida (mediana)	0.734	0.132	0.470	
			Rating Confianza (Media)	0.758	0.148		0.417
			Rating Confianza (mediana)	0.758	0.148		0.417
		Fundamentos de Programación	Puntuación Dada	0.805	0.183		
			Puntuación Recibida (Media)	0.815	0.089	-0.044	
			Puntuación Recibida (mediana)	0.831	0.122	-0.069	
			Rating Confianza (Media)	0.837	0.185		-0.065
			Rating Confianza (mediana)	0.837	0.185		-0.065
		Gestión de procesos de negocios y sistemas empresariales	Puntuación Dada	0.800	0.174		
			Puntuación Recibida (Media)	0.810	0.110	0.132	
			Puntuación Recibida (mediana)	0.808	0.136	0.056	
			Rating Confianza (Media)	0.777	0.181		0.352
			Rating Confianza (mediana)	0.777	0.181		0.352

Escenario de educación virtual asincrónico

En la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2021, la puntuación que proporciona el estudiante ($\bar{X}$=0.707) se aproxima a la puntuación que recibe del colectivo (media ($\bar{X}$=0.708), mediana ($\bar{X}$=0.710)), y del rating de confianza (media ($\bar{X}$=0.700), mediana ($\bar{X}$=0.708)). Existe una relación directa escasa (media (r=0.009), mediana (r=0.032)) entre puntuación dada y recibida del colectivo, y una relación directa moderada (media (r=0.327), mediana (r=0.302)) entre puntuación dada y rating de confianza (Tabla **74** y Figura **65**).

La misma tendencia se identifica en el periodo académico octubre 2021-febrero 2022, la puntuación que da el estudiante ($\bar{X}$=0.774) se acerca a la puntuación que recibe del colectivo (media ($\bar{X}$=0.777), mediana ($\bar{X}$=0.788)), y del rating de confianza (media ($\bar{X}$=0.785), mediana ($\bar{X}$=0.799)); la relación entre puntuación dada y recibida del colectivo es directa escasa (media (r=0.012), mediana (r=0.024)), y la relación entre puntuación dada y rating de confianza es directa moderada (media (r=0.348), mediana (r=0.340)). Y en la asignatura de ingeniería de software, la puntuación que proporciona el estudiante ($\bar{X}$=0.768) se aproxima a la puntuación que recibe del colectivo (media ($\bar{X}$=0.768), mediana ($\bar{X}$=0.769)), y del rating de confianza (media ($\bar{X}$=0.782), mediana ($\bar{X}$=0.796)); la relación entre puntuación dada y recibida del colectivo es directa escasa (media (r=0.048), mediana (r=0.043)), y la relación entre puntuación dada y rating de confianza es directa moderada (media (r=0.444), mediana (r=0.443)) (Tabla **74**).

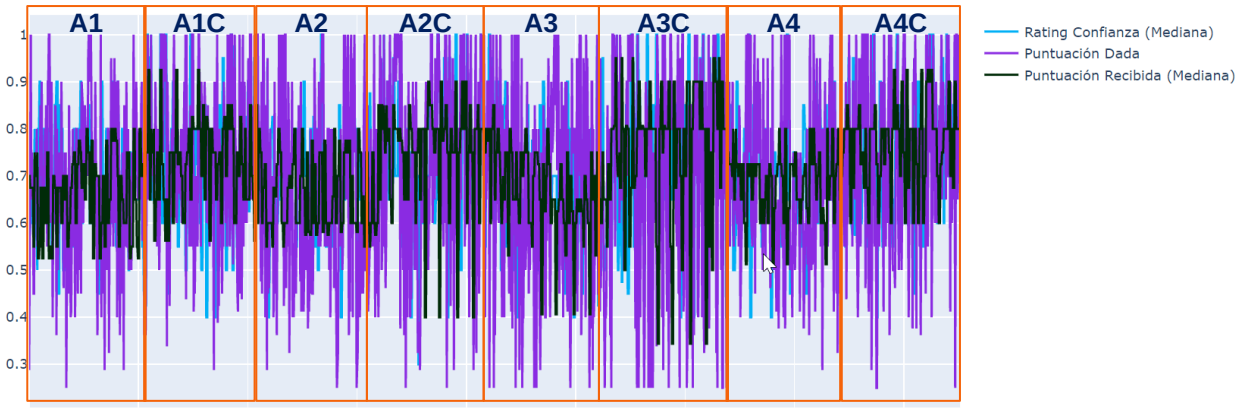


Figura 65. Comparación entre Puntuación Dada, Recibida y Rating Confianza, por cada actividad de la asignatura de fundamentos de ingeniería en software impartida en escenario de educación virtual asincrónico del periodo académico mayo-septiembre 2021

Escenario de educación virtual sincrónico

En la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022, continua la misma tendencia, la puntuación que otorga el estudiante ($\bar{X}$=0.741) se allega a la puntuación que recibe del colectivo (media ($\bar{X}$=0.755), mediana ($\bar{X}$=0.757)), y del rating de confianza (media y mediana ($\bar{X}$=0.773)); la relación entre puntuación dada y recibida del colectivo es directa débil (media (r=0.166), mediana (r=0.182)), y la relación entre puntuación dada y rating de confianza es directa débil (media y mediana (r=0.290)) (Tabla **74** y Figura **66**).

De igual manera, en la asignatura de fundamentos de ofimática, la puntuación que da el estudiante ($\bar{X}$=0.700) se aproxima a la puntuación que recibe del colectivo (media ($\bar{X}$=0.777), mediana ($\bar{X}$=0.784)), y del rating de confianza (media y mediana ($\bar{X}$=0.833)); la relación entre puntuación dada y recibida del colectivo es directa débil (media (r=0.179), mediana (r=0.165)), y la relación entre puntuación dada y rating de confianza es directa moderada (media y mediana ($\bar{X}$=0.477)) (Tabla **74**).

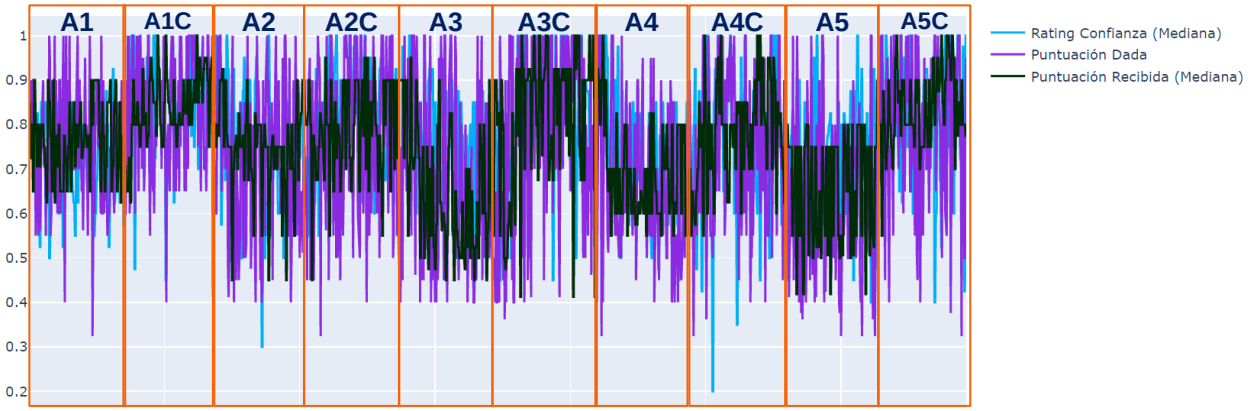


Figura 66. Comparación entre Puntuación Dada, Recibida y Rating Confianza, por cada actividad de la asignatura de fundamentos de ingeniería en software impartida en escenario de educación virtual asincrónico del periodo académico mayo-septiembre 2022

Escenario de educación presencial

En la asignatura de ingeniería de software, la puntuación que otorga el estudiante ($\bar{X}$=0.747) se aproxima a la puntuación que recibe del colectivo (media ($\bar{X}$=0.744), mediana ($\bar{X}$=0.734)), y del rating de confianza (media y mediana ($\bar{X}$=0.758)); la relación entre puntuación dada y recibida del colectivo es directa moderada (media (r=0.454), mediana (r=0.470)), y la relación entre puntuación dada y rating de confianza es directa moderada (media y mediana ($\bar{X}$=0.417)) (Tabla **74** y Figura **67**).

La misma tendencia, en la asignatura de fundamentos de programación, la puntuación que proporciona el estudiante ($\bar{X}$=0.805) se acerca a la puntuación que recibe del colectivo (media ($\bar{X}$=0.815), mediana ($\bar{X}$=0.831)), y del rating de confianza (media y mediana ($\bar{X}$=0.837)); la relación entre puntuación dada y recibida del colectivo es inversa escasa (media (r=-0.044), mediana (r=-

0.069)), y la relación entre puntuación dada y rating de confianza es directa escasa (media (r=-0.044), mediana (r=-0.069)), (r=-0.065)) (Tabla **74**).

Así mismo en la asignatura de gestión de procesos de negocios y sistemas empresariales, la puntuación que otorga el estudiante ($\overline{X}$=0.800) se avecina a la puntuación que recibe del colectivo (media ($\overline{X}$=0.810), mediana ($\overline{X}$=0.808)), y del rating de confianza (media y mediana ($\overline{X}$=0.777)); la relación entre puntuación dada y recibida del colectivo es directa débil (media (r=-0.132), mediana (r=-0.056)), y la relación entre puntuación dada y rating de confianza es directa moderada (media y mediana (r=0.352)) (Tabla **74**)

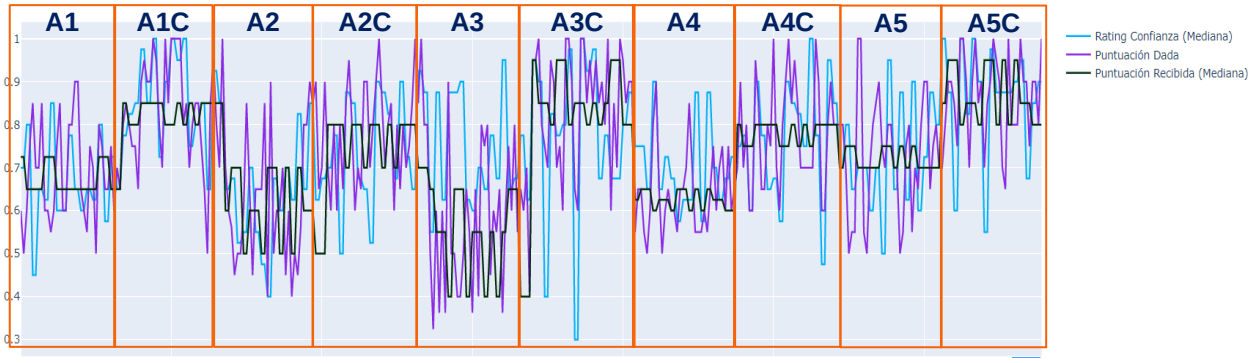


Figura 67. Comparación entre Puntuación Dada, Recibida y Rating Confianza, por cada actividad de la asignatura de ingeniería en software impartida en escenario de educación presencial del periodo académico mayo-septiembre 2022

Una vez realizado el análisis estadístico descriptivo, se estableció que:

- Entre la puntación que da el evaluador con la que recibe del colectivo, existe similaridad directa escasa en asignaturas impartidas en escenario de educación virtual asincrónico, similaridad directa débil en asignaturas impartidas en escenario de educación virtual sincrónico, y similaridad inversa escasa, directa escasa, débil y moderada en asignaturas impartidas en escenario de educación presencial. Coligiendo que en algunos casos las puntuaciones que los estudiantes proporcionan y que reciben se aproximan.
- Entre la puntación que da el evaluador con el rating de confianza dado por los evaluados, existe similaridad directa moderada en las asignaturas impartidas en escenario de educación virtual asincrónico, similaridad directa débil y moderada en asignaturas impartidas en escenario de educación virtual sincrónico y similaridad inversa escasa y directa moderada en asignaturas impartidas en escenario de educación presencial. Deduciendo que, en algunos

casos, la puntuación que da el evaluador a la tarea está asociado al rating de confianza; es decir, si el evaluador da una puntuación considerada como alta o baja, la respuesta del evaluado en la confianza es la misma.

- En escenario de educación virtual asincrónico, en la asignatura de fundamentos de ingeniería de software las puntuaciones recibidas del colectivo calculadas con media alcanzan mayor relación con la puntación dada, mientras que en la asignatura de ingeniería de software alcanza mayor relación con la puntuación calculada con mediana. En virtual sincrónico, en la asignatura de fundamentos de ingeniería de software alcanza mayor relación con la puntuación calculada con mediana, mientras que en la asignatura de fundamento de ofimática alcanza mayor relación con la puntuación calculada con media. En presencial en la asignatura de ingeniería de software y fundamentos de programación alcanza mayor relación con la puntuación calculada con mediana, mientras que en la asignatura de gestión de procesos de negocios y sistemas empresariales alcanza mayor relación con la puntuación calculada con media.
- En escenario de educación virtual asincrónico, el rating de confianza obtenido del colectivo calculadas con media alcanza mayor relación con la puntación dada, tanto en la asignatura de fundamentos de ingeniería de software e ingeniería de software. En virtual sincrónico y presencial, en todas las asignaturas se obtuvo la misma relación tanto con cálculos de media y mediana, por cuanto cada evaluador se le asignó 1 o 2 tareas de revisión.

Por lo tanto, como la correlación es inversa escasa/directa escasa/directa débil/directa moderada (de -0.044 a 0.470) entre la puntuación que da el evaluador con la que recibe del colectivo, e inversa escasa/directa débil/directa moderada (de -0.065 a 0.477) entre la puntuación que da el evaluador con el rating de confianza del colectivo, se determinó como factores tanto la puntuación recibida y el rating de confianza del evaluador para calibrar la puntuación de evaluación de tarea.

Calibración de puntuación de evaluación de tarea

Se realizó en Python la calibración de Puntuación Dada de cada evaluador en función de dos factores: Puntuación Recibida (Perfil Cognitivo) y Rating Confianza (Perfil Evaluador) que obtuvo el estudiante del colectivo, por cada actividad (Figura **64**).

Normalización de datos

- Se normalizó las variables (Puntuación Dada y Puntuación Recibida) a escala de 1.

Especificación del perfil del estudiante

- En la Figura **68** se muestra la categorización de los factores: Puntuación Recibida (Perfil Cognitivo) y Rating Confianza (Perfil Evaluador) que obtuvo el evaluador del colectivo.

```
#Paso 2. Categorización de variables: rendimiento que obtuvo el evaluador del colectivo, se ejemplificó con mediana
def categorise_2(row):
    if row['Mediana-Nota-Obtuvo (colectivo)'] > 0 and row['Mediana-Nota-Obtuvo (colectivo)'] <= 0.2:
        return 'E'
    elif row['Mediana-Nota-Obtuvo (colectivo)'] > 0.2 and row['Mediana-Nota-Obtuvo (colectivo)'] <= 0.4:
        return 'D'
    elif row['Mediana-Nota-Obtuvo (colectivo)'] > 0.4 and row['Mediana-Nota-Obtuvo (colectivo)'] <= 0.6:
        return 'C'
    elif row['Mediana-Nota-Obtuvo (colectivo)'] > 0.6 and row['Mediana-Nota-Obtuvo (colectivo)'] <= 0.8:
        return 'B'
    elif row['Mediana-Nota-Obtuvo (colectivo)'] > 0.8  and row['Mediana-Nota-Obtuvo (colectivo)'] <= 1:
        return 'A'
```

```
#Paso 2. Categorización de variables: rating confianza que obtuvo el evaluador del colectivo, se ejemplificó con mediana
def categorise_1(row):
    if row['Mediana-Confianza (Colectivo)'] > 0 and row['Mediana-Confianza (Colectivo)'] <= 0.2:
        return 'E'
    elif row['Mediana-Confianza (Colectivo)'] > 0.2 and row['Mediana-Confianza (Colectivo)'] <= 0.4:
        return 'D'
    elif row['Mediana-Confianza (Colectivo)'] > 0.4 and row['Mediana-Confianza (Colectivo)'] <= 0.6:
        return 'C'
    elif row['Mediana-Confianza (Colectivo)'] > 0.6 and row['Mediana-Confianza (Colectivo)'] <= 0.8:
        return 'B'
    elif row['Mediana-Confianza (Colectivo)'] > 0.8  and row['Mediana-Confianza (Colectivo)'] <= 1:
        return 'A'
```

```
#Paso 2. Categorización de variables
df_octfeb2022_FIS['Mediana-Confianza-Cat (Colectivo)']=df_octfeb2022_FIS.apply(lambda row: categorise_1(row), axis=1)
df_octfeb2022_FIS['Mediana-Nota-Obtuvo-Cat (colectivo)']=df_octfeb2022_FIS.apply(lambda row: categorise_2(row), axis=1)
```

Figura 68. Categorización de factores: Puntuación Recibida (Perfil Cognitivo) y Rating Confianza (Perfil Evaluador)

Cálculo de calibración de puntuación dada por cada evaluador

- Se calculó la varianza (var) y desviación estándar (std), considerando todas las puntuaciones de evaluación de los evaluadores por cada actividad (Figura **70**).
- Se calibró la Puntuación Dada por cada evaluador, otorgando bonificación o penalización (Tabla **73**) en función de los perfiles del evaluador.

- En la Figura **69**, se detalla un ejemplo de cálculo de puntuación calibrada con Perfil Cognitivo “A” y Perfil Evaluador de “A” hasta “E”.

```
#Paso 3.
def grade_correct(row):
# En esta sección se evalúan la categoría "A" que corresponde al rendimiento
# y las categorías que corresponden a rating confianza de "A" hasta "E"

    if row['Mediana-Nota-Obtuvo-Cat (colectivo)']=='A':
      if row['Mediana-Confianza-Cat (Colectivo)'] =='A':
        if row['Nota-Calificó']+(var+var)*std > 1:
          return row['Nota-Calificó']
        else:
          return row['Nota-Calificó']+(var+var)*std
      elif row['Mediana-Confianza-Cat (Colectivo)']=='B':
        if row['Nota-Calificó']+(1/2*var+var)*std > 1:
          return row['Nota-Calificó']
        else:
          return row['Nota-Calificó']+(1/2*var+var)*std
      elif row['Mediana-Confianza-Cat (Colectivo)']=='C':
          return row['Nota-Calificó']+var*std
      elif row['Mediana-Confianza-Cat (Colectivo)']=='D':
        if row['Nota-Calificó']+(-1/2*var+var)*std< 0:
          return row['Nota-Calificó']
        else:
          return row['Nota-Calificó']+(-1/24+2/24)*std
      elif row['Mediana-Confianza-Cat (Colectivo)']=='E':
          return row['Nota-Calificó']
```

Figura 69. Ejemplo de cálculo de puntuación calibrada con Perfil Cognitivo “A” y Perfil Evaluador de “A” hasta “E”

```
#Paso 3. Se calcula la varianza (var) y la desviación estándar (std),
#considerando todas las puntuaciones de evaluación "Nota-Calificó" de los evaluadores por cada actividad
activities_1 = df_octfeb2022_FIS['CodigoActividad'].unique()
results_list_1=[]
for activity in activities_1:
    std=df_octfeb2022_FIS.loc[df_octfeb2022_FIS['CodigoActividad']==activity]['Nota-Calificó'].std()
    var=df_octfeb2022_FIS.loc[df_octfeb2022_FIS['CodigoActividad']==activity]['Nota-Calificó'].var()
    results_list_1.extend(df_octfeb2022_FIS.loc[df_octfeb2022_FIS['CodigoActividad']==activity].apply(lambda row: grade_correct(row), axis=1))

#Paso 3. Se asigna los resultados de puntuación calibrada "Nota Corregida"
df_octfeb2022_FIS['Nota Corregida']=results_list_1
```

Figura 70. Cálculo de varianza y desviación estándar de todas las puntuaciones de los evaluadores por cada actividad y asignación de los resultados de puntuación calibrada

Paso 4. Cálculo de puntuación calibrada del colectivo

- Se calculó la puntuación calibrada de todos los evaluadores por grupo evaluado (Figura **71**).

```
#Paso 4. Cálculo de puntuación calibrada del colectivo
grupos = df_octfeb2022_FIS['Grupo-Pertenece'].unique()
actividades = df_octfeb2022_FIS['CodigoActividad'].unique()
for grupo in grupos:
  for actividad in actividades:
    df_octfeb2022_FIS.loc[(df_octfeb2022_FIS['Grupo-Pertenece']==grupo) & (df_octfeb2022_FIS['CodigoActividad']==actividad),'Nota Obtenida Corregida'
```

Figura 71. Cálculo de puntuación calibrada del colectivo

Conclusiones, limitaciones y futuro experimento

- Se realizó la calibración de la puntuación de la evaluación de la tarea en función del rendimiento y rating de confianza del evaluador, ya que la correlación no fue fuerte de estas variables, y así obtener una mejor fiabilidad en el proceso de evaluación entre pares, evitando que los estudiantes propendan a evaluar con un puntaje alto el rating de confianza para recibir bonificación.
- En función del rendimiento y rating de confianza del evaluador se asignó bonificación o penalización utilizando como medidas la varianza y desviación estándar de todas las puntuaciones que dieron los evaluadores por cada actividad, ya que cada actividad es sobre una temática diferente e impartida en escenario diverso, tanto el evaluador y evaluado realizan trabajos similares, y es evaluada con una rúbrica especifica basada en la temática.
- El pequeño tamaño de la muestra limita los resultados de este experimento.
- En el trabajo futuro, se planea:
  - Estudiar otro tipo de medidas para detectar puntuaciones sesgadas entre todos los evaluadores en las puntuaciones por cada criterio.
  - Evaluar la validez del modelo de evaluación entre pares mediane la correlación de la puntuación del colectivo de los pares evaluadores con la del docente.
  - Evaluar el rendimiento estudiantil en el proceso de evaluación entre pares.
  - Evaluar la utilidad del modelo en el proceso de evaluación entre pares.

5. IMPLEMENTACIÓN Y EVALUACIÓN DEL MODELO

En este capítulo se efectúa el objetivo 4, detallando el prototipo de evaluación entre pares, la implementación del modelo de análisis de sentimiento, y de lógica difusa. Además, se analiza los resultados, y se comprueba si los resultados obtenidos son estadísticamente significativos.

5.1. Prototipo de evaluación entre pares

Se construyó la aplicación de evaluación entre pares con puntuación cuantitativa, cualitativa, inversa y en dos rondas. A continuación, se detalla:

5.1.1. Especificación de requerimientos

Se utilizó la herramienta Balsamiq para identificar requerimientos de configuración de rúbricas, envío de tarea, evaluación de tarea y evaluación de calidad de la evaluación en procesos de evaluación entre pares en dos rondas. En la Figura **72** se muestra un ejemplo.

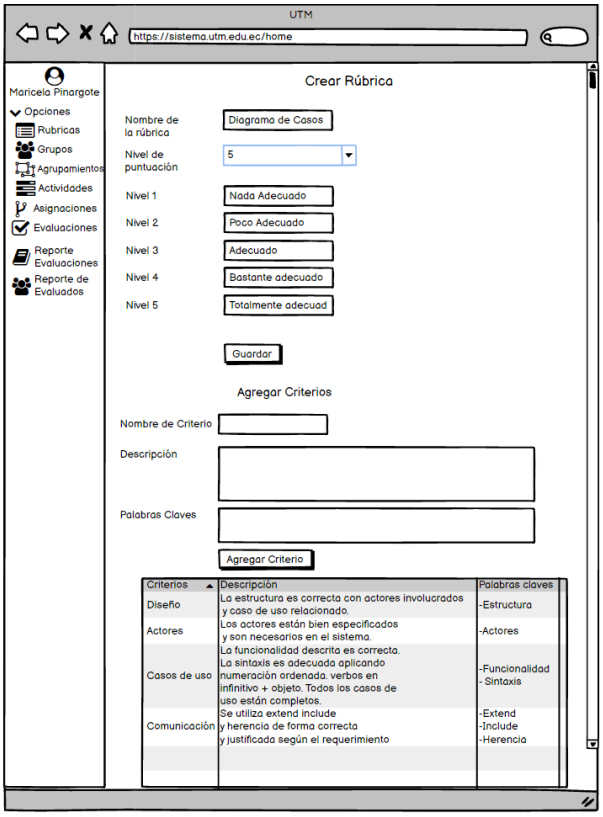


Figura 72: Ejemplo del requerimiento funcional (RF-02)

Requerimientos funcionales y no funcionales

En la Tabla **75** se describe los requerimientos funcionales y en la Tabla **76** los no funcionales identificados.

Tabla 75. Requerimientos funcionales

Requerimiento	Descripción
RF-01	La aplicación deberá permitir la autenticación de los usuarios (docente, estudiante).
RF-02	El usuario (docente) podrá configurar rúbrica con: criterios, descripción, y nivel de puntuación.
RF-03	El usuario (docente) podrá configurar grupos.
RF-04	El usuario (docente) podrá configurar actividades grupales en diferentes fases: envió de tarea, evaluación de tarea, evaluación de calidad de la evaluación en dos rondas con nombre, descripción, enlace de instrucciones, fecha de inicio y finalización.
RF-05	El usuario (docente) podrá configurar asignaciones de tareas enviadas.
RF-06	El usuario (docente) podrá configurar asignaciones de evaluación de tarea y evaluación de calidad de la evaluación.
RF-07	El usuario (docente) podrá generar reportes.
RF-08	El usuario (docente) podrá listar los estudiantes de los cursos
RF-09	El usuario (docente) podrá configurar periodos académicos, paralelos y escalas de valoración.
RF-10	El usuario (docente y estudiante) podrán seleccionar asignaturas de acuerdo al periodo académico.
RF-11	El usuario (estudiante) podrá seleccionar grupo.
RF-12	El usuario (estudiante) podrá colocar el enlace de tarea compartida en el drive.
RF-13	El usuario (estudiante, docente) podrán evaluar la tarea mediante la rúbrica configurada con puntuación numérica y retroalimentación textual, y podrán visualizar la puntuación de sentimiento de la retroalimentación.
RF-14	El usuario (estudiante) podrá evaluar la calidad de la evaluación con puntuación numérica y retroalimentación textual, además podrá visualizar la valoración de sentimiento de la retroalimentación.
RF-15	El usuario (estudiante) podrá visualizar en reportes por cada actividad la puntuación numérica, retroalimentación textual y puntuación de sentimiento proporcionada por los evaluadores.

Tabla 76. Requerimientos no funcionales

Requerimiento	Descripción
RNF-01	Usabilidad: La aplicación es de fácil entendimiento, esto con la finalidad de que cualquier usuario (persona) pueda manejarlo sin dificultad alguna.
RNF -02	Disponibilidad de información: El sistema está disponible para cualquier dispositivo que tenga conexión a internet.
RNF -03	Eficiencia: La aplicación muestra mensajes de error validando cada una de las funciones internas.

Arquitectura

En la Figura **73**, se detalla la arquitectura del prototipo.

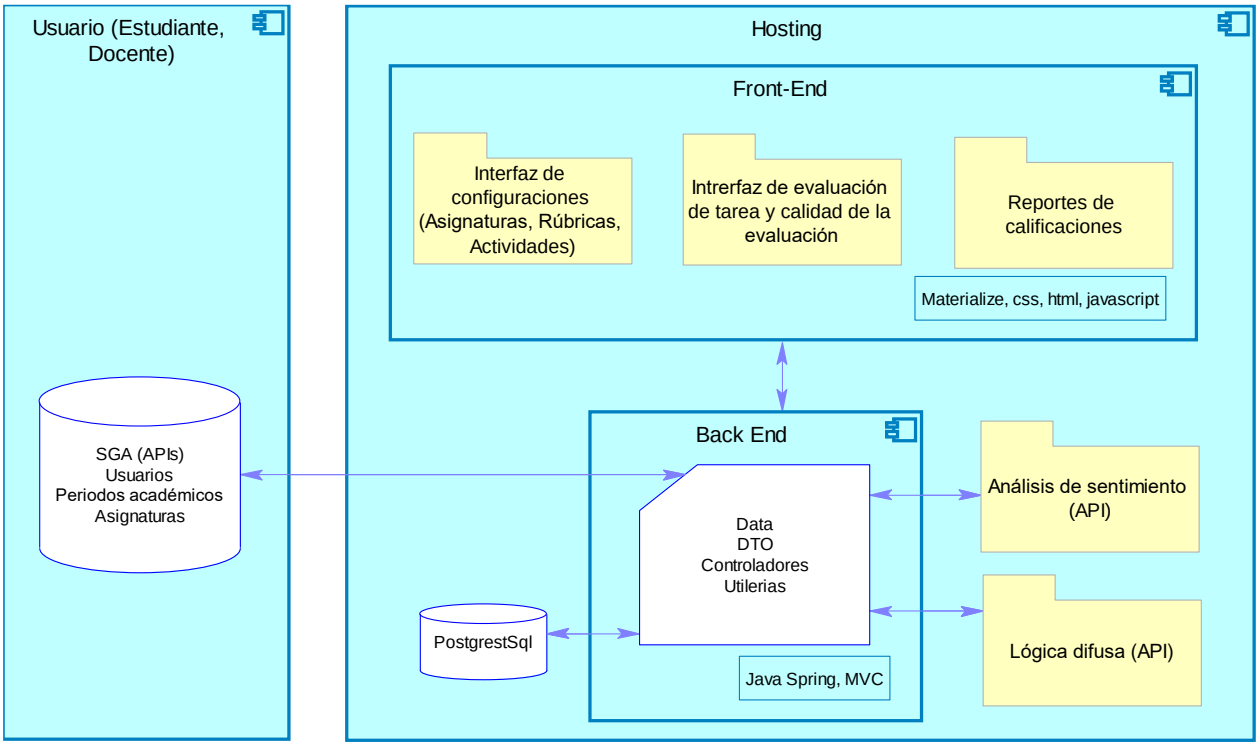


Figura 73. Arquitectura del prototipo

Para obtener los datos del Sistema de Gestión Académica (SGA) de la UTM se utilizó APIs:

- API de inicio de sesión, que recibe como parámetro: usuario, clave y X-Api-Key (Figura **74**).

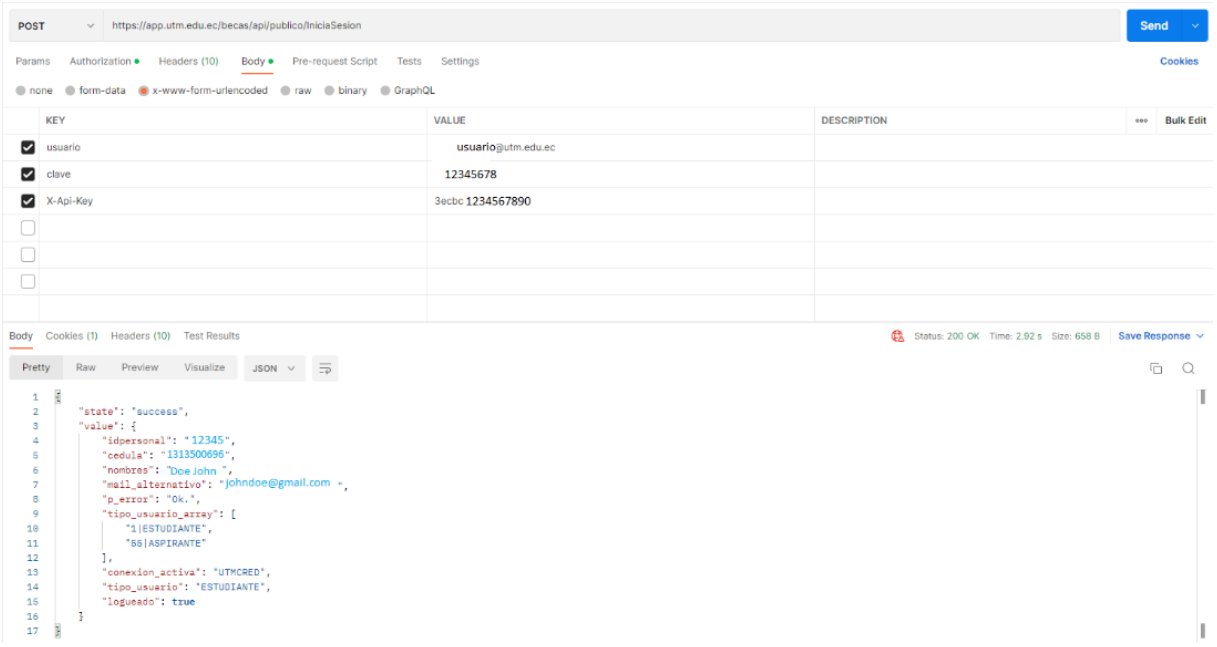


Figura 74. API de inicio de sesión

- API de obtención del departamento al que pertenece un docente, que recibe como parámetro: cedula, idperiodo y X-Api-Key (Figura **75**).

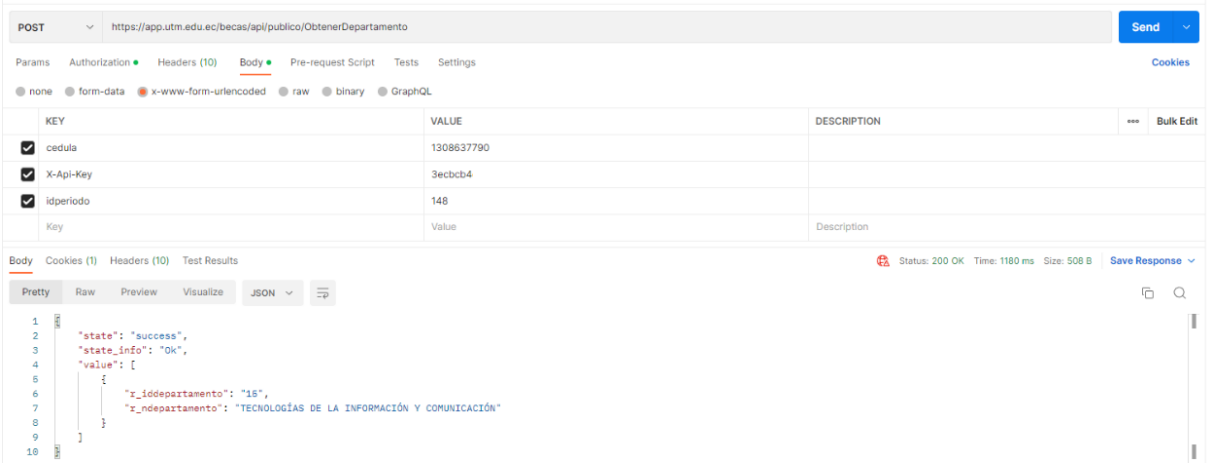


Figura 75. API de obtención del departamento al que pertenece un docente

- API de obtención de periodos académicos a los que se ha matriculado un estudiante o un docente tiene carga horaria, que recibe como parámetro: cedula, X-Api-Key (Figura **76**).

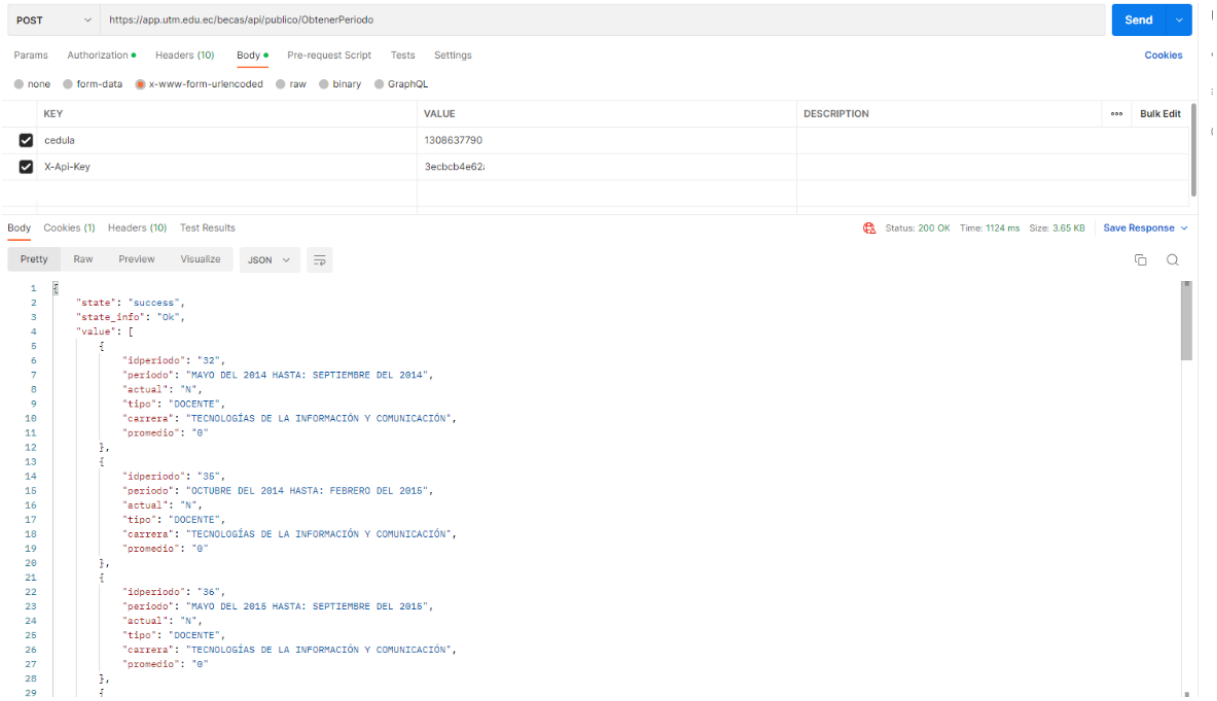


Figura 76. API de obtención de periodos académicos

- API de obtención de asignaturas que imparte el docente, que recibe como parámetro: Idperiodo, idpersonal, iddepartamento, X-Api-Key (Figura **77**).

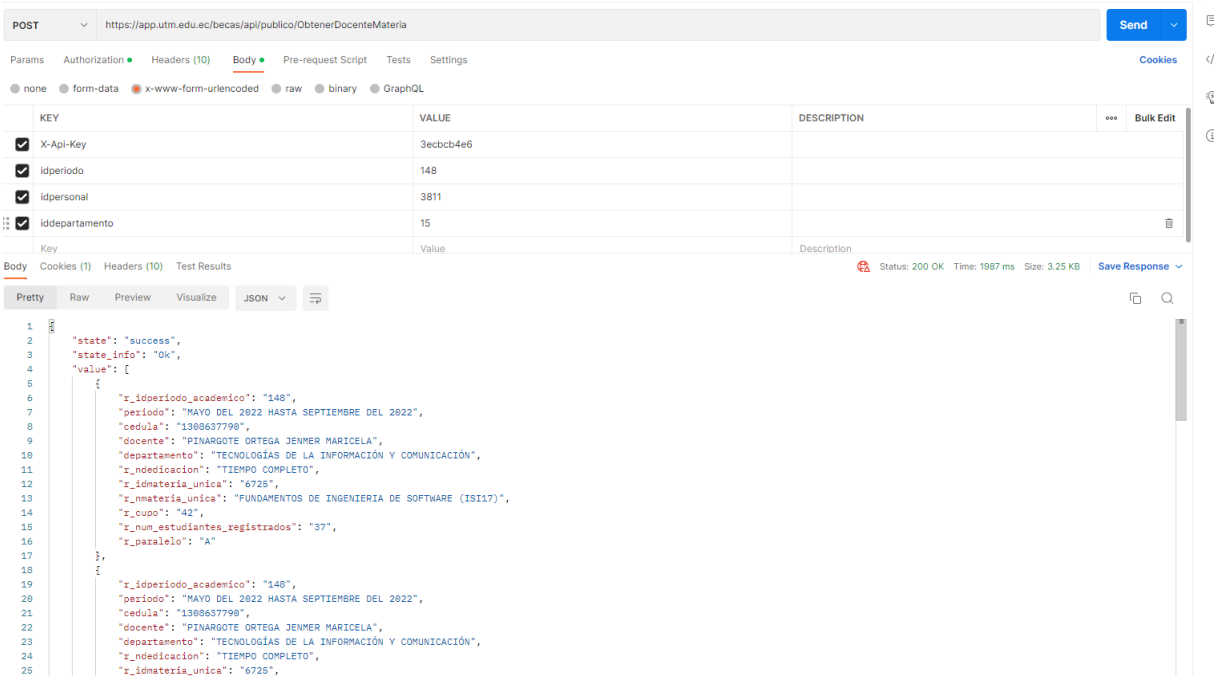


Figura 77. API de obtención de asignaturas que imparte el docente

- API de obtención de asignaturas que el estudiante recibe, que recibe como parámetro: cedula, idperiodo, X-Api-Key (Figura **78**)

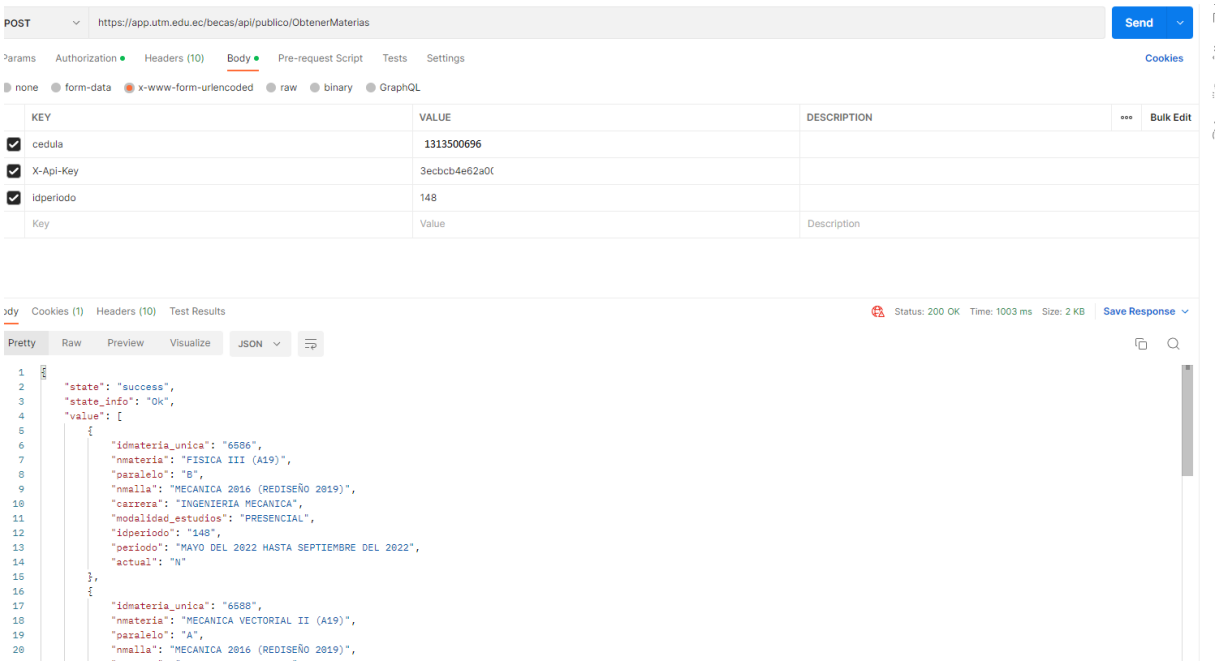


Figura 78. API de obtención de asignaturas que recibe el estudiante

- API de obtención de listado de estudiante de una asignatura, que recibe como parámetro: idmateria, idperiodo, paralelo, X-Api-Key (Figura **79**).

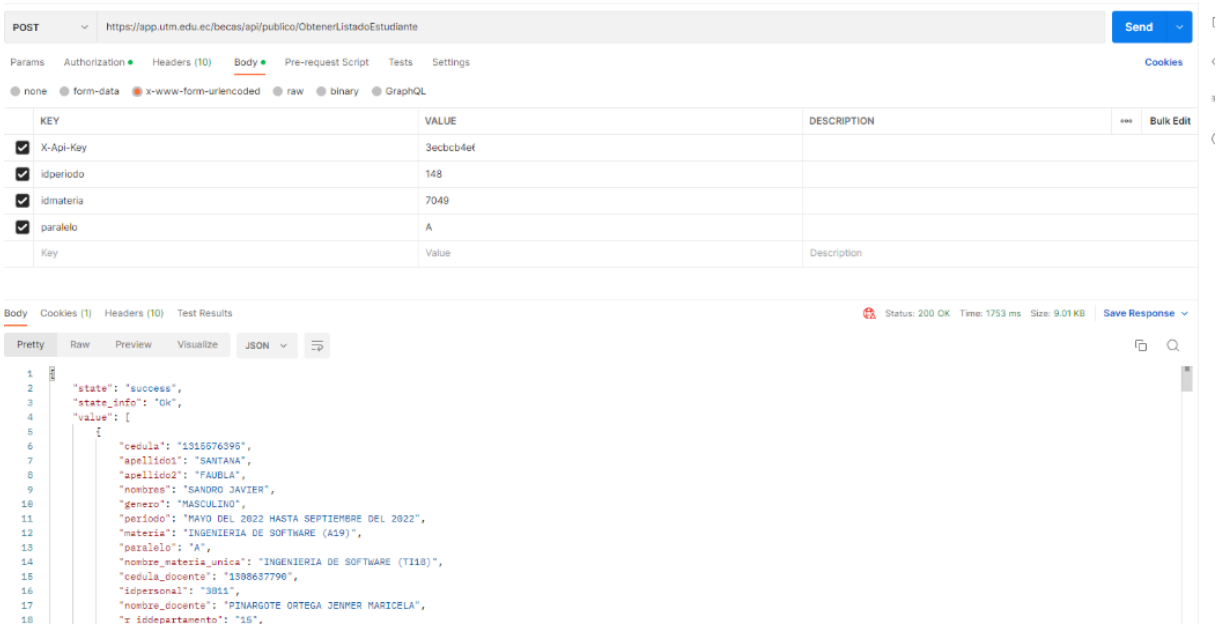


Figura 79. API de listado de estudiantes de una asignatura

Para la elaboración de la aplicación web se usó Java con el framework Spring MVC (Modelo Vista Controlador), se utilizó la librería Materialize para los estilos (CSS) y Jquery en el front-end permitiendo hacer las interfaces más dinámicas y PostgreSQL para almacenar los datos. A continuación, se detalla las funcionalidades:

5.1.2. Funcionalidades generales

A continuación, se detalla las funcionalidades principales implementada en SEP)-UTM (evaluacionpares.herokuapp.com)

- **Inicio de sesión y menú**

  El inicio de sesión, muestra dos campos uno para el usuario y el otro para la contraseña (Figura **80**). Una vez iniciada la sesión se muestra un menú con secciones y subsecciones.

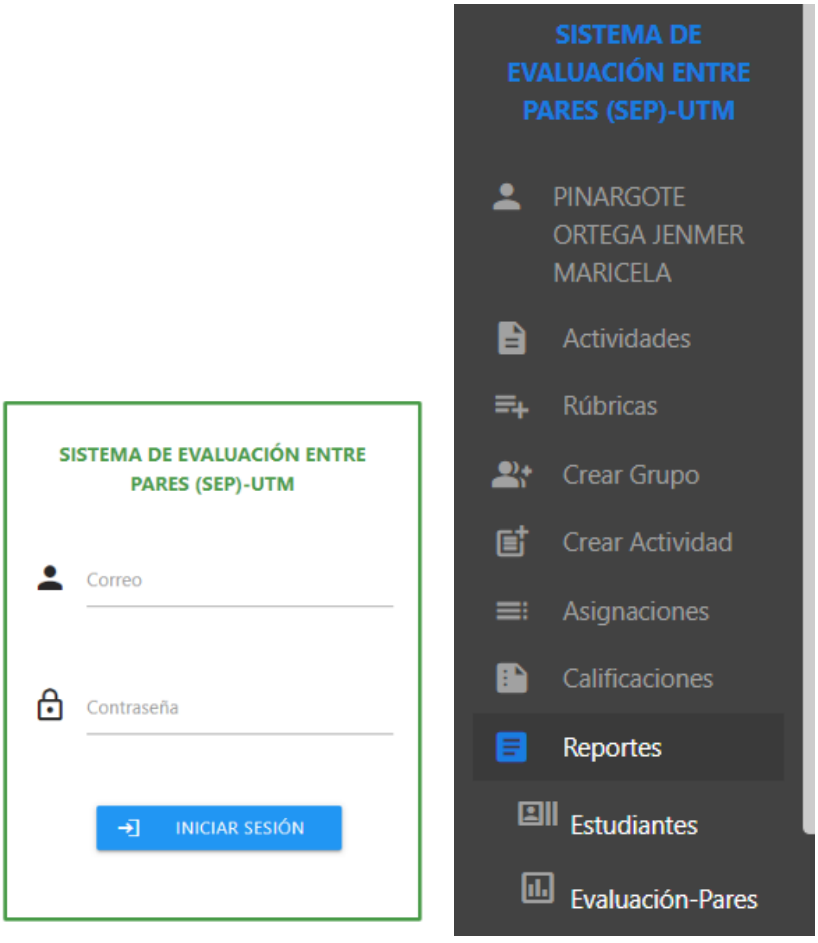


Figura 80. Interfaz de inicio de sesión y de menú

- **Asignaturas**

  El estudiante o el docente podrá escoger la asignatura que desee en el menú o en la interfaz principal (Figura **81**).

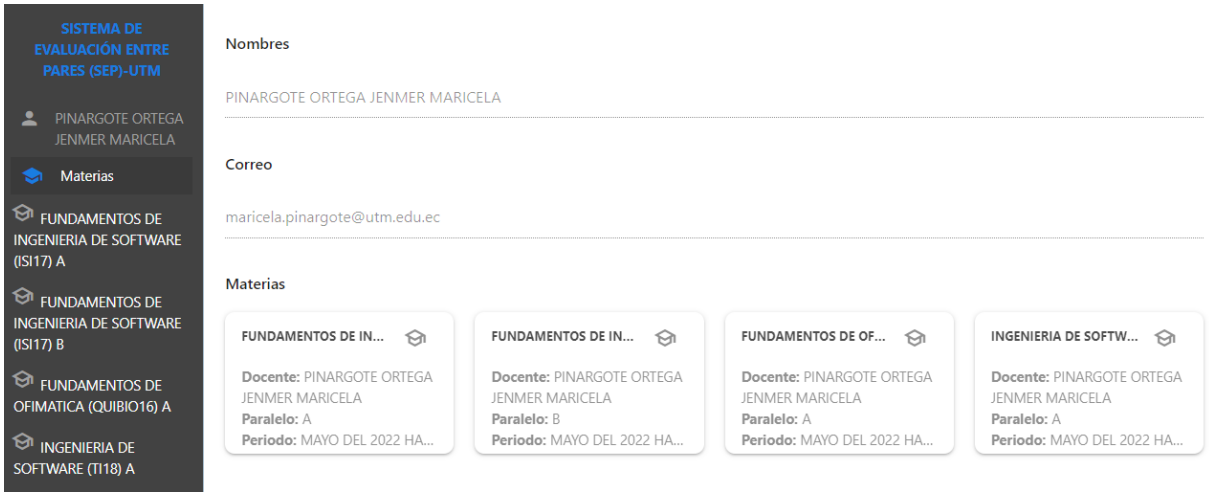


Figura 81. Interfaz para seleccionar la asignatura

- **Rúbricas**

  El docente podrá crear/editar rúbrica con: criterios, descripción, y escala de valoración (Figura **82**).

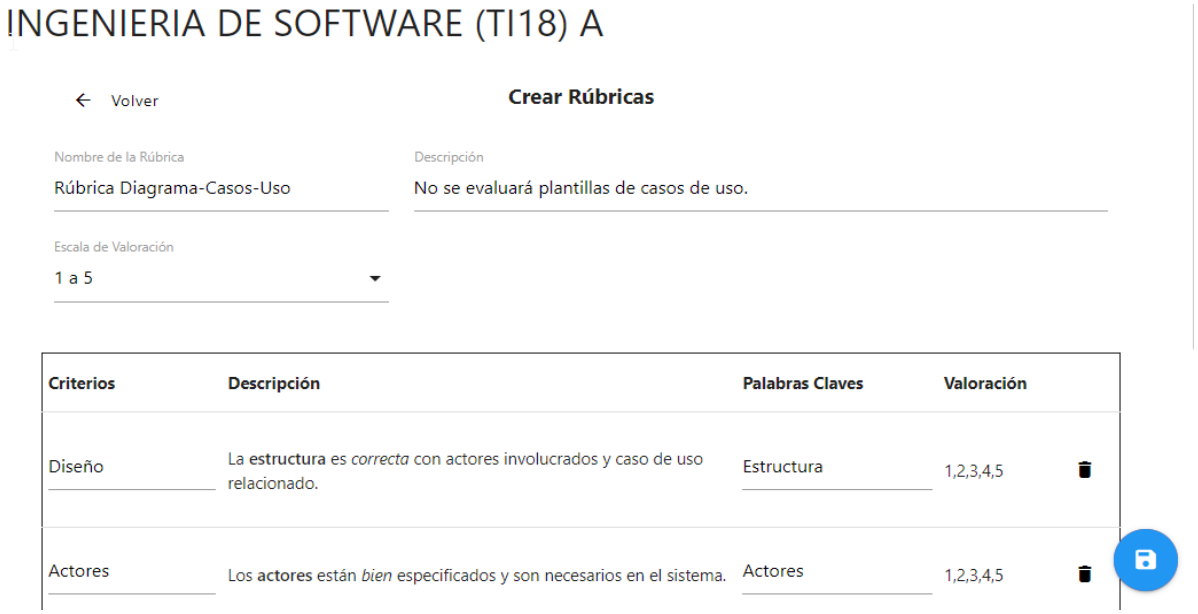


Figura 82. Interfaz de configuración de rúbricas

- **Configuración de actividad**

  El docente podrá crear/editar las actividades en diferentes fases como: envío de tareas, evaluación de las tareas, evaluación de calidad de la evaluación en dos rondas con fechas determinadas (Figura **83**).

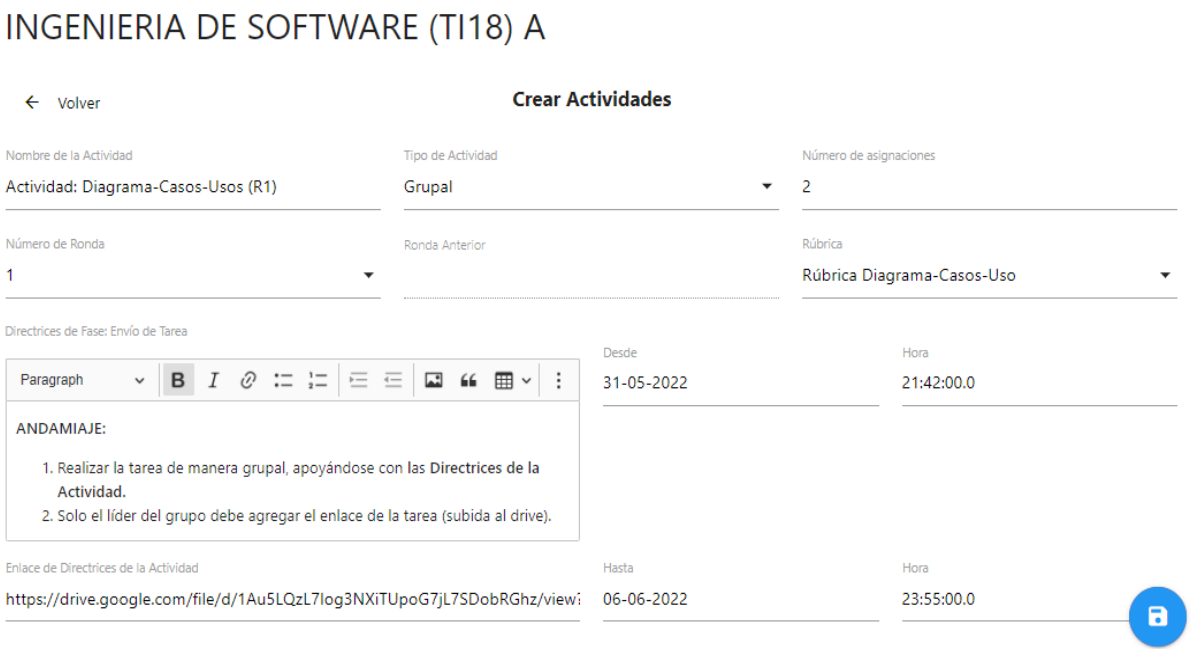


Figura 83. Interfaz de configuración de actividades

- **Configuración de asignaciones**

  Se podrá seleccionar los paralelos configurados y la actividad que se quiere realizar la asignación. Una vez realizada esa acción se visualizan los grupos que ya han realizado el envío de tarea de manera exitosa y también muestran los que no la han completado. El docente podrá asignar de manera automática los grupos para empezar a realizar las evaluaciones de las tareas (Figura **84**).

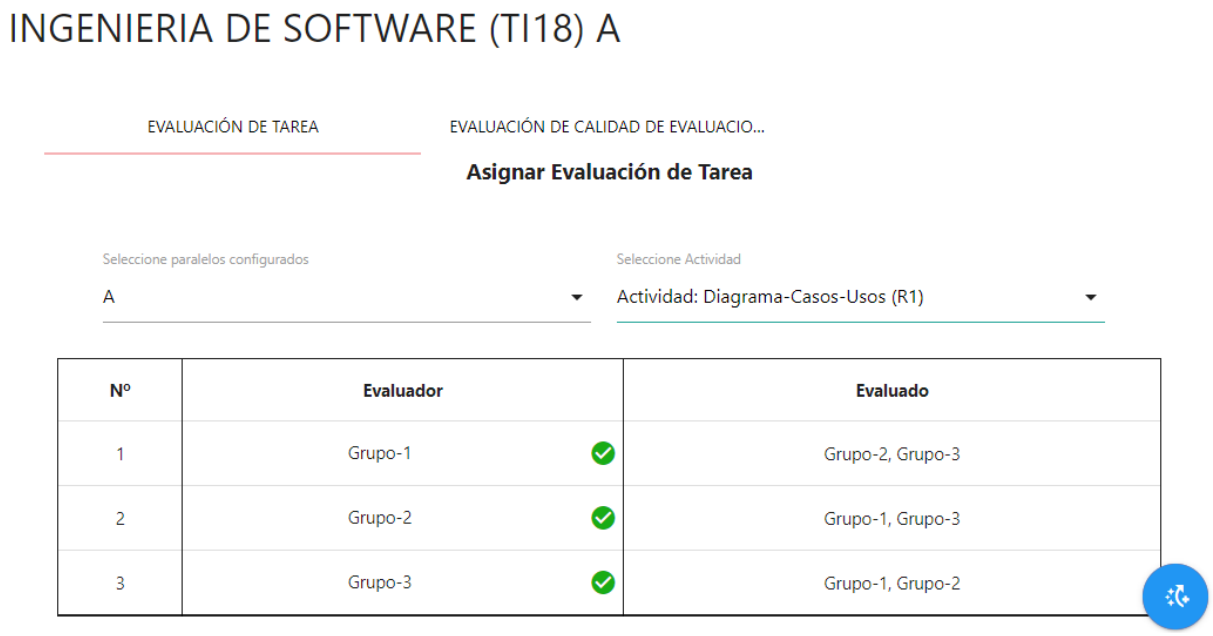


Figura 84. Interfaz de configuración de asignación de tareas para la evaluación

El docente también podrá asignar de manera automática las evaluaciones de calidad de la evaluación (Figura **85** y Figura **86**).

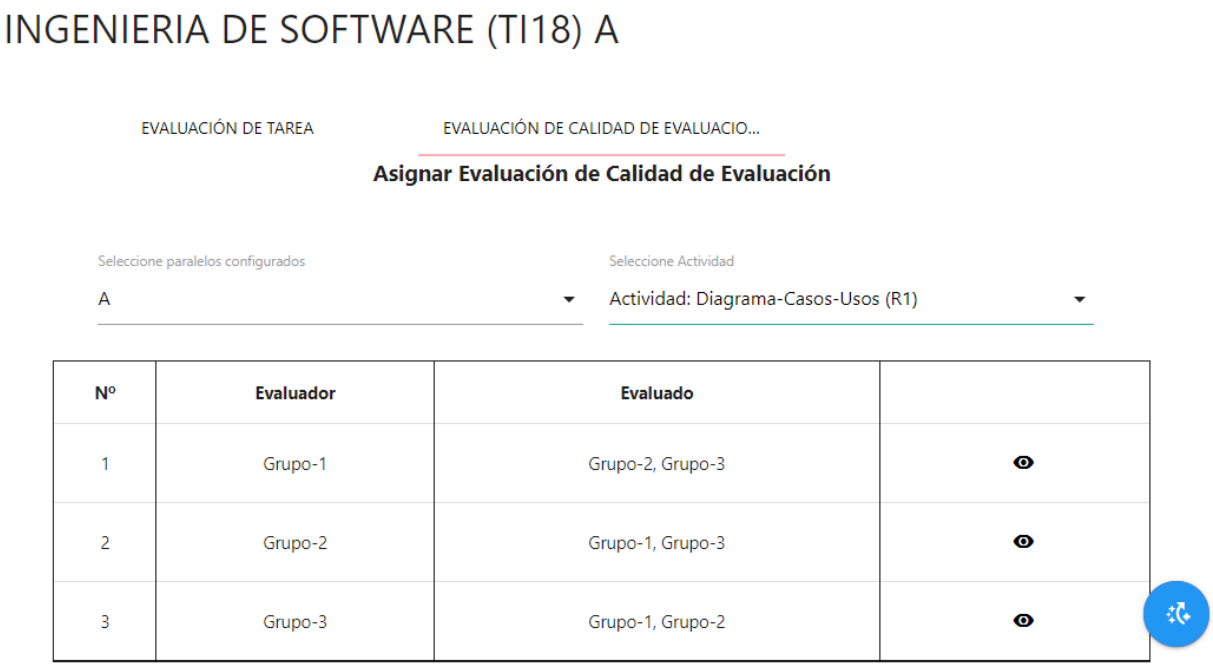


Figura 85. Interfaz de configuración de asignación de evaluación de calidad de la evaluación

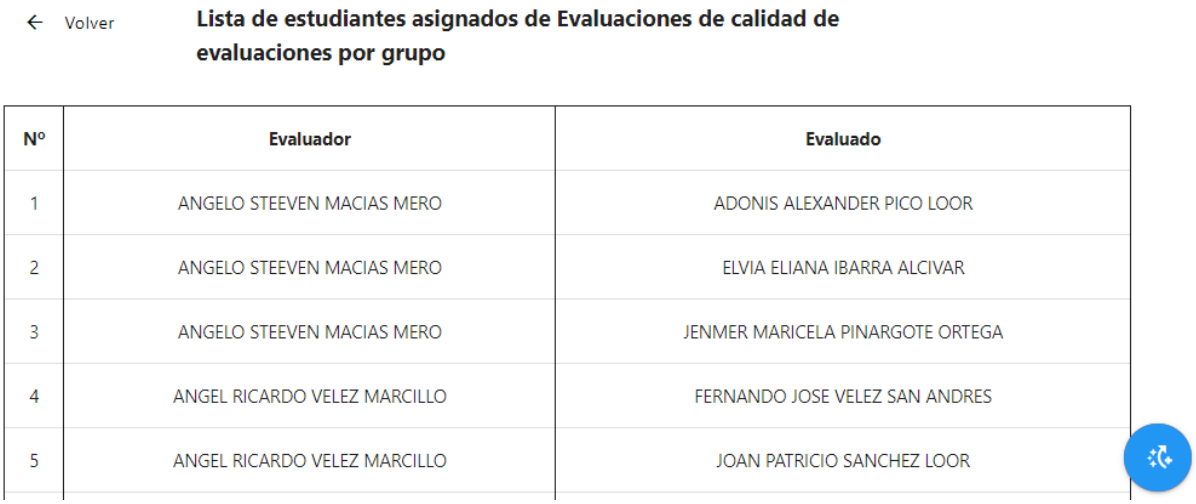


Figura 86. Interfaz para enlistar a los estudiantes asignados de evaluación de calidad de la evaluación

- **Configuración de paralelos**

  El docente podrá configurar juntos o separados los paralelos de la asignatura, para poder utilizarlo en los “Grupos” y “Actividades” (Figura **87**)

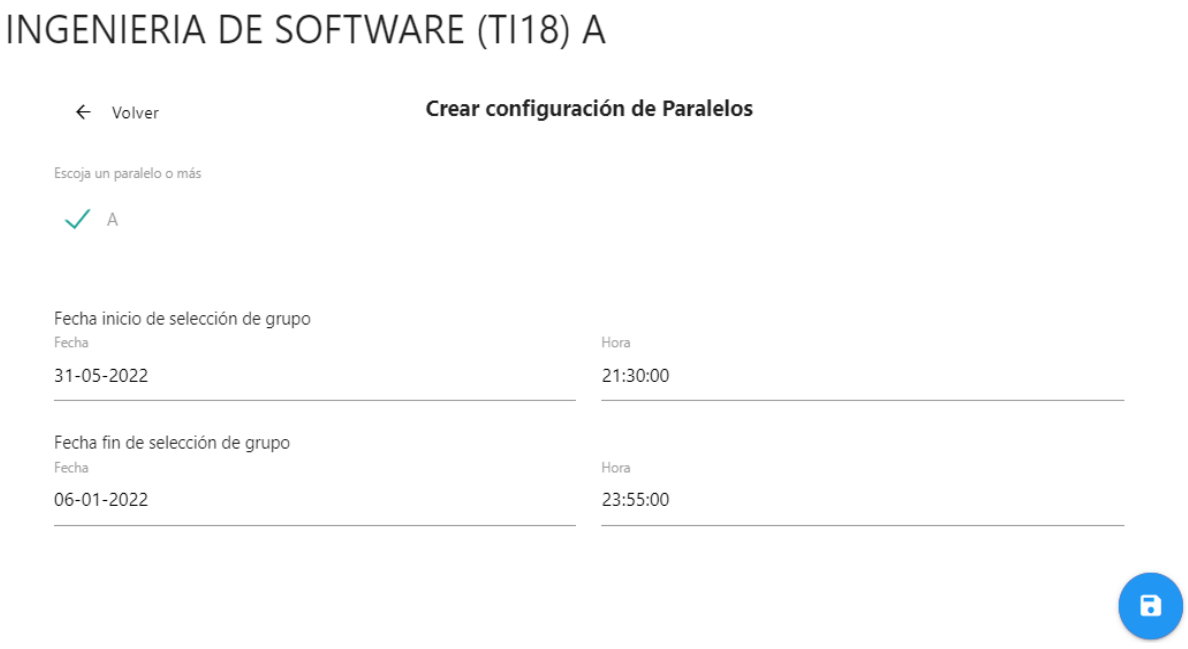


Figura 87. Interfaz de configuración de paralelos

- **Configuración de grupos**

  El docente podrá crear/editar grupos para que posteriormente el estudiante seleccione. Esto con la finalidad de que los estudiantes de uno o varios paralelos puedan ingresar en base a la configuración realizada por el docente (Figura **88**).

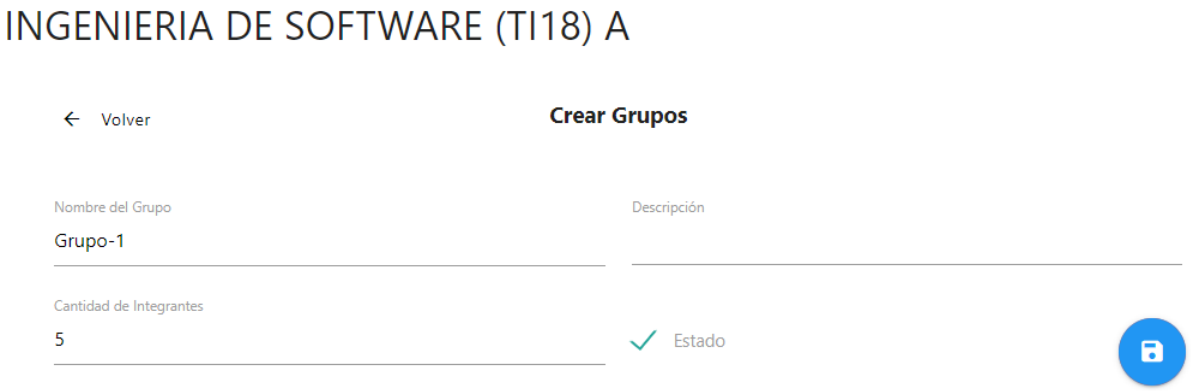


Figura 88. Interfaz de configuración de grupos

- **Selección de grupos**

  El estudiante podrá seleccionar el grupo que desee, teniendo una interfaz bastante amigable con una paginación dependiendo de la cantidad de grupos creados por el docente (Figura **89**).

Figura 89. Interfaz para seleccionar el grupo

- **Envío de tarea**

  El estudiante podrá colocar el enlace de la tarea (compartida en el drive) para que así posteriormente sea evaluada, esta interfaz está dentro de la sección “Actividades” (Figura **90**).

Figura 90. Interfaz de envío de tarea

- **Reportes**

  El docente podrá generar los reportes necesarios. En la subsección “Estudiantes” podrá visualizar a todos los estudiantes del curso seleccionado (Figura **91**).

Figura 91. Interfaz para enlistar a los estudiantes

5.1.3. Funcionalidades de análisis de sentimiento

Se implementó el modelo de análisis de sentimiento en el prototipo mediante una API. Se creo un archivo app.py, donde se importó el flask que permite hacer la API, jsonify, ya que se recibe un json y se devuelve un json. En la ruta predict comments, se recibió el json por método post. Se preprocesa, se hace token, y se hace pad_sequence. Al final se usa el predict para obtener las polaridades (Figura **92** y Figura **93**).

```
model.load_weights('modelo_weights24.hdf5')
```

Figura 92. Modelo entrenado en Python

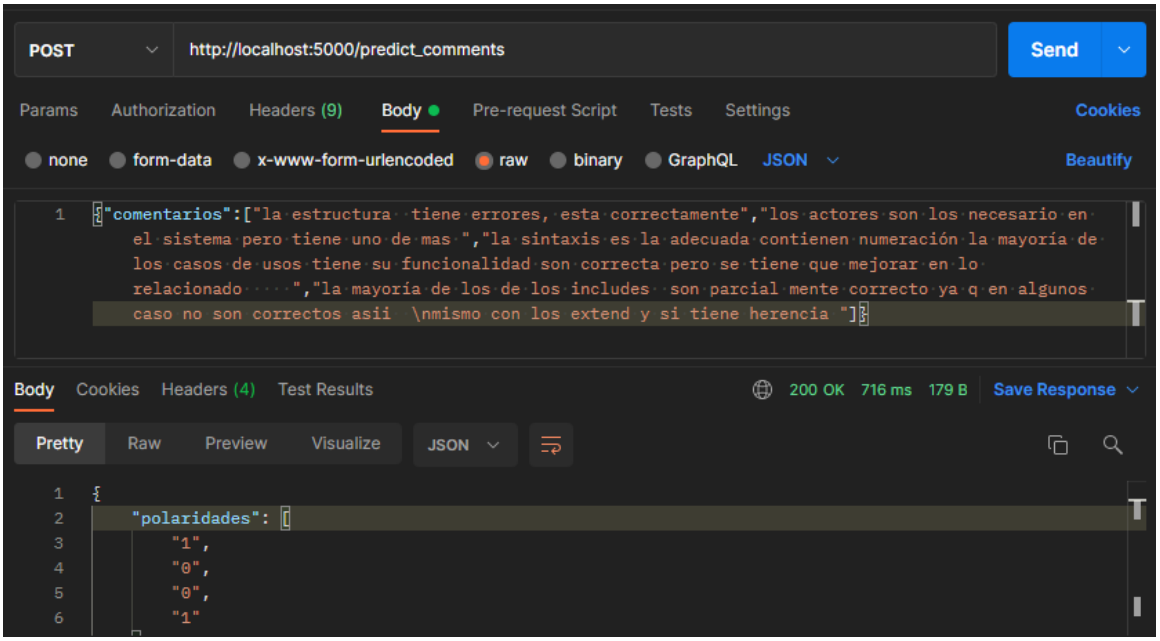


Figura 93. API de análisis de sentimiento en postman

A continuación, se detalla las funcionalidades aplicando análisis de sentimiento:

- **Evaluación de tarea con enfoque cuantitativo y cualitativo**
  El evaluador evalúa la tarea con puntuación numérica y retroalimentación textual, y se genera la puntuación de sentimiento de la retroalimentación proporcionada (Figura **94**).

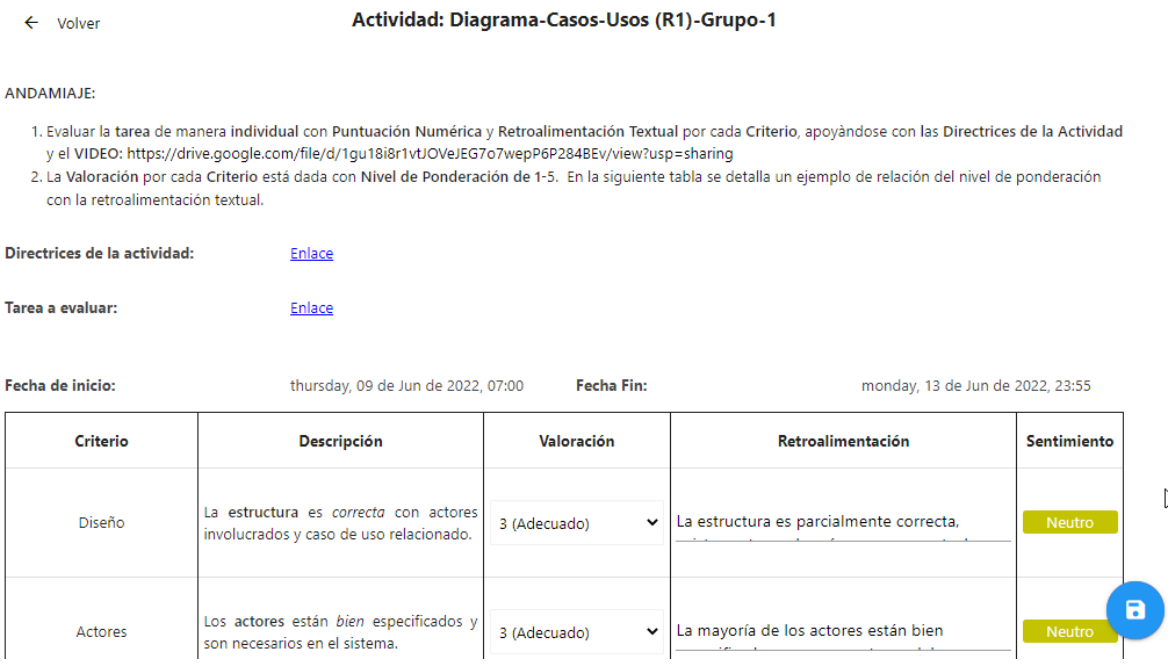


Figura 94. Evaluación de tarea

- Estos resultados se presentan:
  - Artículo: Peer Feedback Sentiment Analysis Prototype. Aceptado para publicación en RISTI (ver Apéndice C)

- **Evaluación de la calidad de la evaluación (evaluación inversa)**

  El evaluado podrá visualizar la puntuación numérica y retroalimentación que proporcionó el evaluador para evaluar la calidad de la evaluación con puntuación numérica y retroalimentación textual, de la misma manera se genera la puntuación de sentimiento de la retroalimentación proporcionada (Figura **95**).

Figura 95. Evaluación de la calidad de la evaluación

- **Reportes**

  Una vez que la tarea haya sido evaluada, el estudiante podrá visualizar en la sección “Reportes” todas las evaluaciones que realizaron sus pares, con puntuación numérica, retroalimentación y puntuación de sentimiento, y corregir el trabajo en base a las retroalimentaciones dadas (Figura **96**).

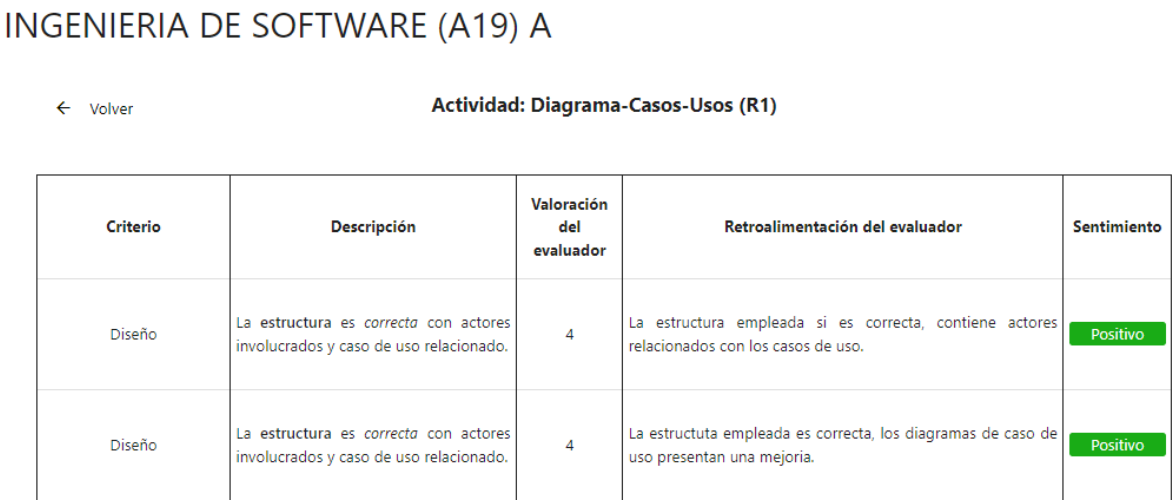


Figura 96. Reporte de evaluaciones de los pares

- **Evaluación de tarea corregida (segunda ronda)**

  El evaluador podrá visualizar la puntuación numérica y la retroalimentación que proporcionó en la primera ronda, y el comentario del evaluado, para evaluar la tarea corregida con puntuación numérica y retroalimentación textual, de la misma manera se genera la puntuación de sentimiento de la nueva retroalimentación proporcionada (Figura **97**).

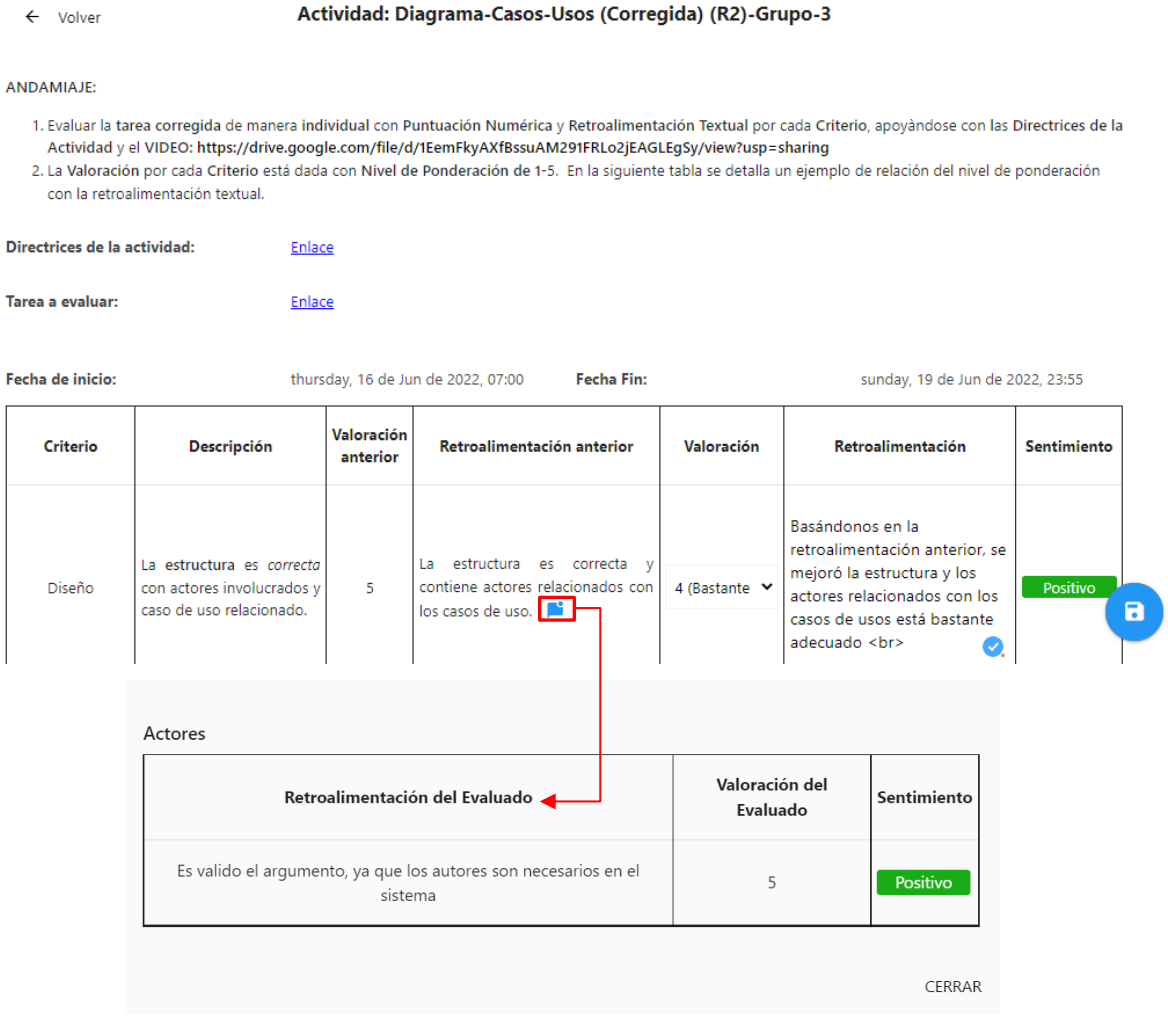


Figura 97. Evaluación de la tarea en la segunda ronda

5.1.4. Funcionalidades de cálculo

Se implementó el modelo de lógica difusa en el prototipo mediante una API, que recibe como entrada, el id de la actividad, la puntuación numérica y la puntuación sentimiento. Dentro de los procesos que corresponden a este EndPoint tenemos: generar la puntuación de evaluación y de

confianza, a través de la aplicación del modelo de la lógica difusa. Posteriormente se calculó la media de todos los criterios. Seguidamente se calculó media/mediana de todas las puntuaciones individuales, y se almacenó los datos (Figura **98**).

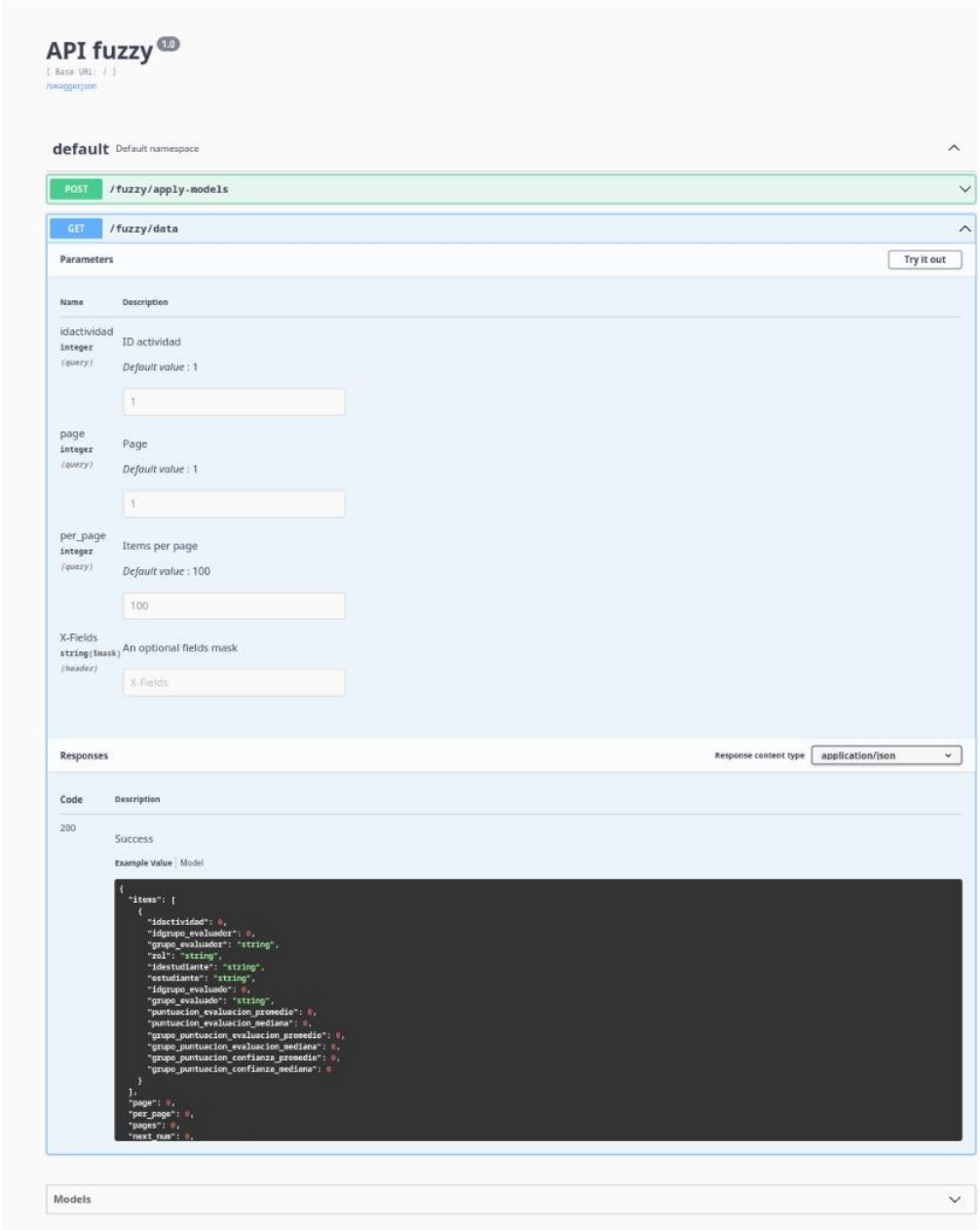


Figura 98. Interfaz de “Api Fuzzy”, datos post/fuzzy/apply-models

A continuación, se detallan las funcionalidades:

- **Generación de calificaciones**

  El docente, en la sección “Calificaciones”, podrá visualizar las calificaciones finales, seleccionando el paralelo, la actividad, el tipo de reporte y el tipo de medida a calcular (Figura **99**).

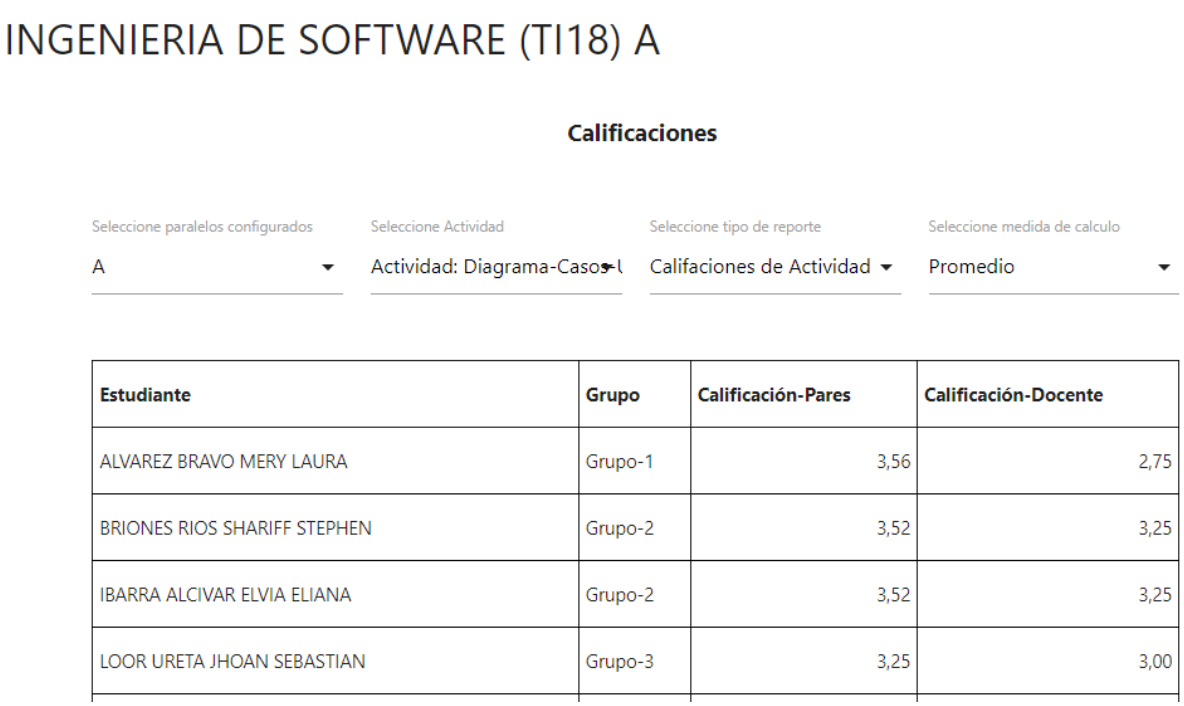
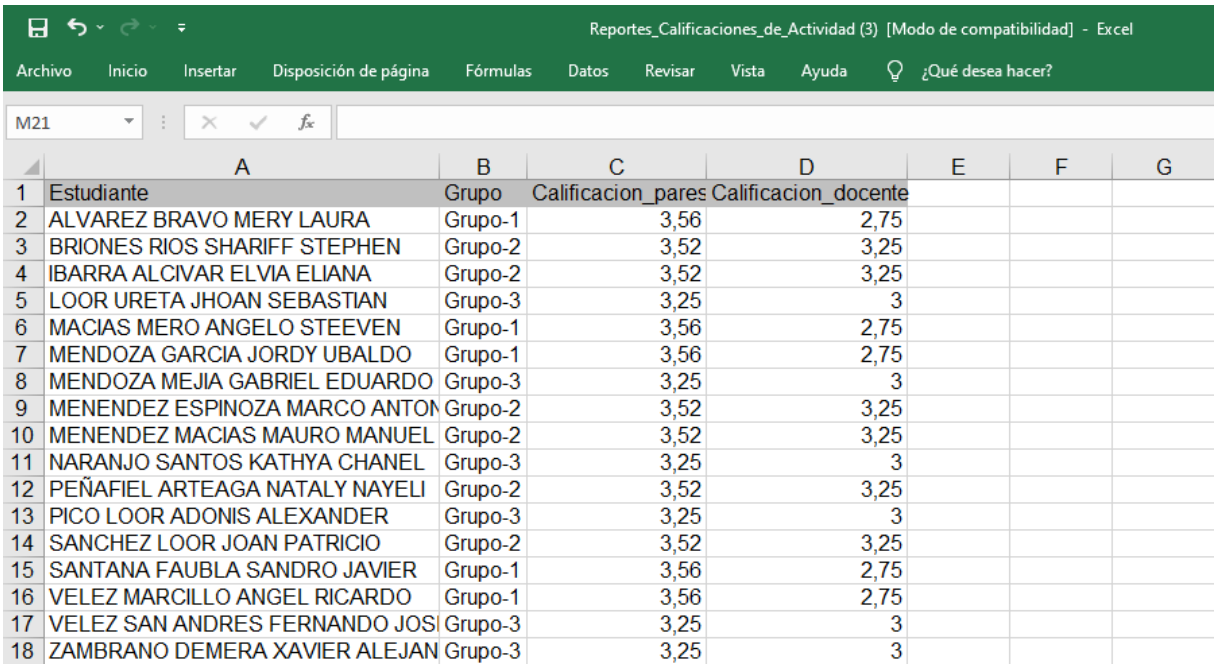


Figura 99. Interfaz de generación de calificaciones de estudiantes con media/mediana por actividad

Además, podrá descargar un archivo en Excel, con los mismos datos que se visualizan (Figura **100**).



	A	B	C	D	E	F	G
1	Estudiante	Grupo	Calificacion_pares	Calificacion_docente			
2	ALVAREZ BRAVO MERY LAURA	Grupo-1	3,56	2,75			
3	BRIONES RIOS SHARIFF STEPHEN	Grupo-2	3,52	3,25			
4	IBARRA ALCIVAR ELVIA ELIANA	Grupo-2	3,52	3,25			
5	LOOR URETA JHOAN SEBASTIAN	Grupo-3	3,25	3			
6	MACIAS MERO ANGELO STEEVEN	Grupo-1	3,56	2,75			
7	MENDOZA GARCIA JORDY UBALDO	Grupo-1	3,56	2,75			
8	MENDOZA MEJIA GABRIEL EDUARDO	Grupo-3	3,25	3			
9	MENENDEZ ESPINOZA MARCO ANTON	Grupo-2	3,52	3,25			
10	MENENDEZ MACIAS MAURO MANUEL	Grupo-2	3,52	3,25			
11	NARANJO SANTOS KATHYA CHANEL	Grupo-3	3,25	3			
12	PEÑAFIEL ARTEAGA NATALY NAYELI	Grupo-2	3,52	3,25			
13	PICO LOOR ADONIS ALEXANDER	Grupo-3	3,25	3			
14	SANCHEZ LOOR JOAN PATRICIO	Grupo-2	3,52	3,25			
15	SANTANA FAUBLA SANDRO JAVIER	Grupo-1	3,56	2,75			
16	VELEZ MARCILLO ANGEL RICARDO	Grupo-1	3,56	2,75			
17	VELEZ SAN ANDRES FERNANDO JOS	Grupo-3	3,25	3			
18	ZAMBRANO DEMERA XAVIER ALEJAN	Grupo-3	3,25	3			

Figura 100. Archivo de Excel con datos calculados con media/mediana por actividad

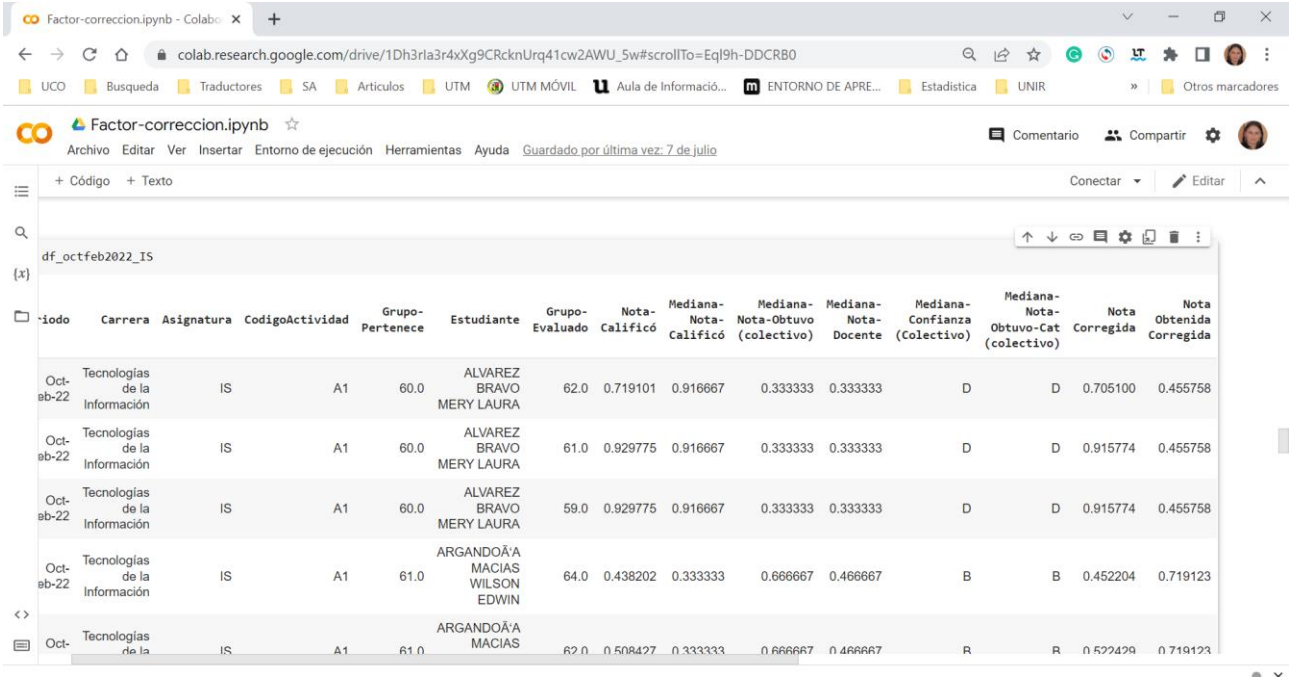
El estudiante, en la sección “Calificaciones”, también podrá ver sus puntuaciones, seleccionando la actividad (Figura **101**).

Estudiante	Grupo	Calificación-Pares	Calificación-Docente
SANTANA FAUBLA SANDRO JAVIER	Grupo-1	3,56	2,75

Figura 101. Interfaz de visualización de puntuación de evaluación del estudiante

5.1.5. Calibración de puntuación de evaluación de tarea

Se implementó en Python el modelo de calibración y genera la puntuación de la evaluación calibrada en CSV (Figura **102**).



iodo	Carrera	Asignatura	CodigoActividad	Grupo-Pertenece	Estudiante	Grupo-Evaluado	Nota-Calificó	Mediana-Nota-Calificó	Mediana-Nota-Obtuvo (colectivo)	Mediana-Nota-Docente	Mediana-Confianza (Colectivo)	Mediana-Nota-Obtuvo-Cat (colectivo)	Nota Corregida	Nota Obtenida Corregida
Oct-ab-22	Tecnologías de la Información	IS	A1	60.0	ALVAREZ BRAVO MERY LAURA	62.0	0.719101	0.916667	0.333333	0.333333	D	D	0.705100	0.455758
Oct-ab-22	Tecnologías de la Información	IS	A1	60.0	ALVAREZ BRAVO MERY LAURA	61.0	0.929775	0.916667	0.333333	0.333333	D	D	0.915774	0.455758
Oct-ab-22	Tecnologías de la Información	IS	A1	60.0	ALVAREZ BRAVO MERY LAURA	59.0	0.929775	0.916667	0.333333	0.333333	D	D	0.915774	0.455758
Oct-ab-22	Tecnologías de la Información	IS	A1	61.0	ARGANDOÃ'A MACIAS WILSON EDWIN	64.0	0.438202	0.333333	0.666667	0.466667	B	B	0.452204	0.719123
Oct-	Tecnologías de la	IS	A1	61.0	ARGANDOÃ'A MACIAS	62.0	0.508427	0.333333	0.666667	0.466667	B	B	0.522429	0.719123

Figura 102. Archivo de Excel con datos calibrados por actividad

5.2. Evaluación del modelo propuesto y validación de hipótesis

Se evaluó la validez del modelo propuesto y su utilidad en procesos de enseñanza-aprendizaje y se validó la hipótesis.

5.2.1. Validez del modelo propuesto

Para responder a la pregunta de investigación: **¿Cuál es la incidencia de la modalidad de educación en procesos de evaluación entre pares?,** se probó la validez del modelo propuesto en 3 escenarios de educación: virtual asincrónico en los periodos académicos mayo-septiembre 2021, y octubre 2021-febrero 2022 ante la pandemia COVID-19, virtual sincrónico y presencial en el periodo académico mayo-septiembre 2022.

Se realizó cálculo de parámetros estadísticos para determinar la relación que existe entre las variables: Puntuación Docente con Puntuación Dada, Puntuación Recibida (media/mediana), Puntuación Dada Calibrada, Puntuación Recibida Calibrada (media/mediana). La correlación de Pearson determina que puede existir una relación lineal entre las variables correlacionadas. Dependiendo de si el valor es positivo o negativo, se establece si la relación es directa o inversa.

Mientras que, si el coeficiente es cero, no existirá una relación entre las variables que se plantea evaluar. Para interpretar los valores se utilizó los siguientes criterios: perfecta (IRI = 1), fuerte (1< IRI > = 0.7), sustancial (0.7< IRI >= 0.5), moderada (0.5 < IRI >= 0.3), débil (0.3 < IRI >= 0.1) o escasa/ninguna (0.1 < IRI >= 0.0).

Donde:

$\bar{X}$ = Media

Σ = Desviación Estándar

r = Correlación Pearson

Id	Tarea
A1	Ejercicios de diagrama de casos de uso
A2	Ejercicios de diagrama de actividad
A3	Ejercicios de diagrama de secuencia
A4	Ejercicios de diagrama de clases
A5	Ejercicio de diseño de componentes
A6	Documento de Texto
A7	Hoja de Cálculo
A8	Diseño de aulas virtuales

Celda resaltada de color celeste-claro = dato con mayor relación

Celda resaltada de color verde-claro = tiende a subir la relación con la calibración

Celda resaltada de color naranja-claro = tiende a bajar la relación con la calibración

Celda resaltada de color gris-claro = se mantiene la relación con la calibración

En la Figura **103**, Figura **104**, Figura **105**, Figura **106**, Figura **107**, y Figura **108** se puede identificar una gran diferencia entre las curvas que identifican a la puntuación que proporcionó y recibió el estudiante, a la puntuación que corresponde al docente. Aunque muestran la misma tendencia de incrementar y reducir una en función de la otra, las magnitudes mantienen una aparente diferencia que se definió empleando la media. A continuación, se detalla los resultados por cada escenario de educación:

Escenario de educación virtual asincrónico

Análisis por asignatura

En el escenario de educación virtual asincrónico se realizaron pruebas en 24 actividades. En la asignatura de fundamentos de ingeniería de software se ejecutaron 8 actividades, y en la asignatura de ingeniería de software se ejecutaron 4 actividades en dos rondas (Tabla **77**).

Tabla 77. Relación de las puntuaciones de los pares con la del docente por periodo académico y asignatura en escenario de educación virtual asincrónico

Variables	X̄	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2021 Asignatura: Fundamentos de ingeniería de software						
Puntuación Docente	0.516	0.158				
Puntuación Dada	0.707	0.195	0.100			
Puntuación Recibida (Media)	0.708	0.092			0.589	
Puntuación Recibida (Mediana)	0.710	0.110			0.562	
Puntuación Dada Calibrada	0.711	0.193		0.106		
Puntuación Recibida Calibrada (Media)	0.716	0.110				0.592
Puntuación Dada Calibrada	0.713	0.194		0.106		
Puntuación Recibida Calibrada (Mediana)	0.714	0.092				0.567
Periodo académico: octubre 2021 - febrero 2022 Asignatura: Fundamentos de ingeniería de software						
Puntuación Docente	0.637	0.182				
Puntuación Dada	0.774	0.171	0.096			
Puntuación Recibida (Media)	0.777	0.081			0.438	
Puntuación Recibida (Mediana)	0.788	0.107			0.360	
Puntuación Dada Calibrada	0.780	0.170		0.094		
Puntuación Recibida Calibrada (Media)	0.783	0.080				0.433
Puntuación Dada Calibrada	0.780	0.170		0.095		
Puntuación Recibida Calibrada (Mediana)	0.795	0.107				0.356
Asignatura: Ingeniería de software						
Puntuación Docente	0.593	0.142				
Puntuación Dada	0.768	0.172	0.023			
Puntuación Recibida (Media)	0.768	0.088			0.325	
Puntuación Recibida (Mediana)	0.769	0.104			0.312	
Puntuación Dada Calibrada	0.773	0.171		0.022		
Puntuación Recibida Calibrada (Media)	0.773	0.088				0.322
Puntuación Dada Calibrada	0.773	0.171		0.023		
Puntuación Recibida Calibrada (Mediana)	0.774	0.104				0.311

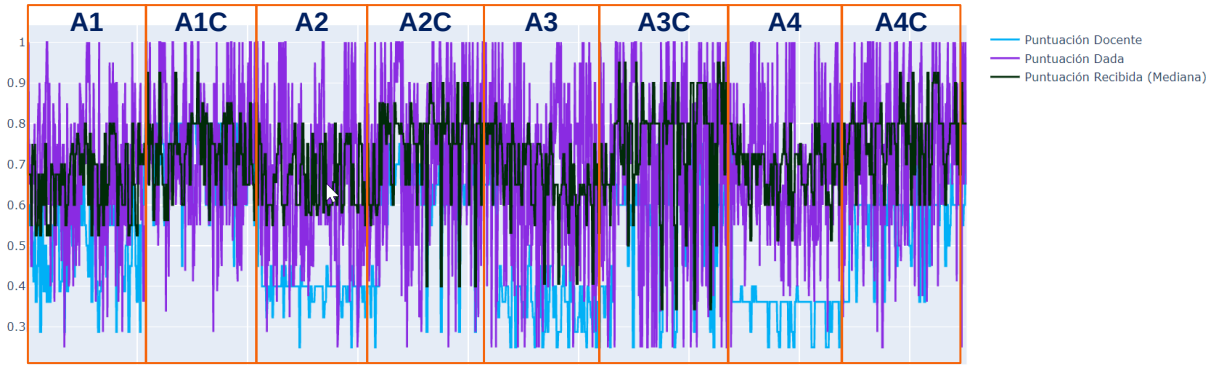
En la asignatura de fundamentos de ingeniería de software en el periodo académico mayo-septiembre 2021, de manera general los estudiantes tienden a proporcionar puntuaciones ($\bar{X}$=0.707), y recibir puntuaciones del colectivo ($\bar{X}$=0.708-0.710), mayores que la del docente ($\bar{X}$=0.516). La relación entre puntuación dada y puntuación docente es directa débil (r=0.100), y la relación entre puntuación recibida del colectivo y puntuación docente es directa sustancial (media (r=0.589), mediana (r=0.562)). Mientras que, con la calibración de la puntuación dada ($\bar{X}$=0.711-0.713), y de la puntuación recibida del colectivo ($\bar{X}$=0.714-0.716), tiende a subir la relación entre puntuación dada y puntuación docente (r=0.106), y entre puntuación recibida del

colectivo y puntuación docente (media (r=0.592), mediana (r=0.567)), teniendo mayor relación con la media. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.092-0.110) es menor en relación a la del docente (Σ=0.158) (Tabla **77** y Figura **103**).

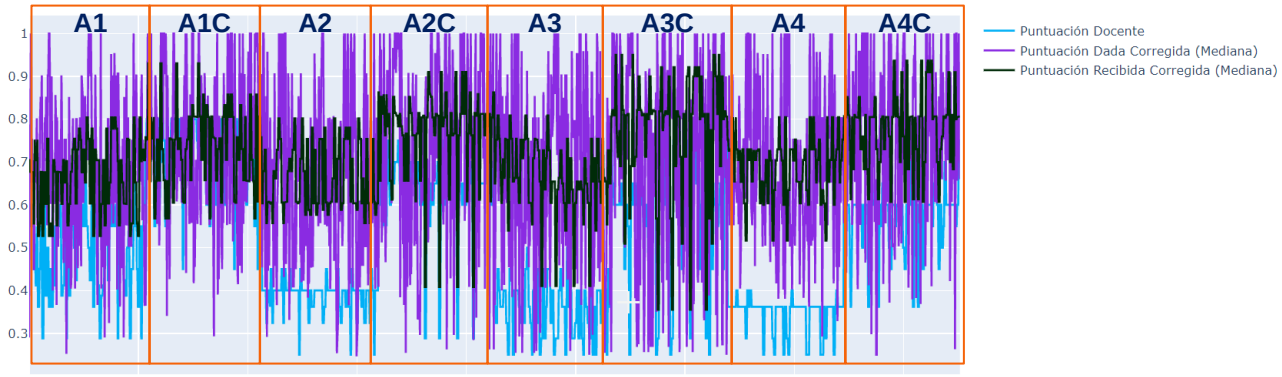
Similar propensión se determina en el periodo académico octubre 2021-febrero 2022, los estudiantes tienden a otorgar puntuaciones ($\overline{X}$=0.774), y recibir puntuaciones del colectivo ($\overline{X}$=0.777-0.788), mayores que la del docente ($\overline{X}$=0.637). La relación entre puntuación dada y puntuación docente es directa escasa (r=0.096), y la relación entre puntuación recibida del colectivo y puntuación docente es directa moderada (media (r=0.438), mediana (r=0.360)). Con la calibración de la puntuación dada ($\overline{X}$= 0.780) y de la puntuación recibida del colectivo ($\overline{X}$=0.783-0-795), tiende a bajar la relación entre puntuación dada y puntuación docente (media (r=0.094), mediana (r=0.095)), y entre puntuación recibida del colectivo y puntuación docente (media (r=0.433), mediana (r=0.356)), teniendo mayor relación con la media. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.080-0.107) es menor en relación a la del docente (Σ=0.182) (Tabla **77**).

Así mismo, en la asignatura de ingeniería de software, los estudiantes tienden a otorgar puntuaciones ($\overline{X}$=0.768), y obtener puntuaciones del colectivo ($\overline{X}$=0.768-0.769), mayores que la del docente ($\overline{X}$=0.593). La relación entre puntuación dada y puntuación docente es directa escasa (r=0.023), y la relación entre puntuación recibida del colectivo y puntuación docente es directa moderada (media (r=0.325), mediana (r=0.312)). Con la calibración, la puntuación dada ($\overline{X}$=0.773) y la puntuación recibida del colectivo ($\overline{X}$=0.773-0.774), tiende a bajar la relación entre puntuación dada y puntuación docente (media (r=0.022)), y entre puntuación recibida del colectivo y puntuación docente (media (r=0.322), mediana (r=0.311)), teniendo mayor relación con la media. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.088-0.104) es menor en relación a la del docente (Σ=0.142) (Tabla **77**).

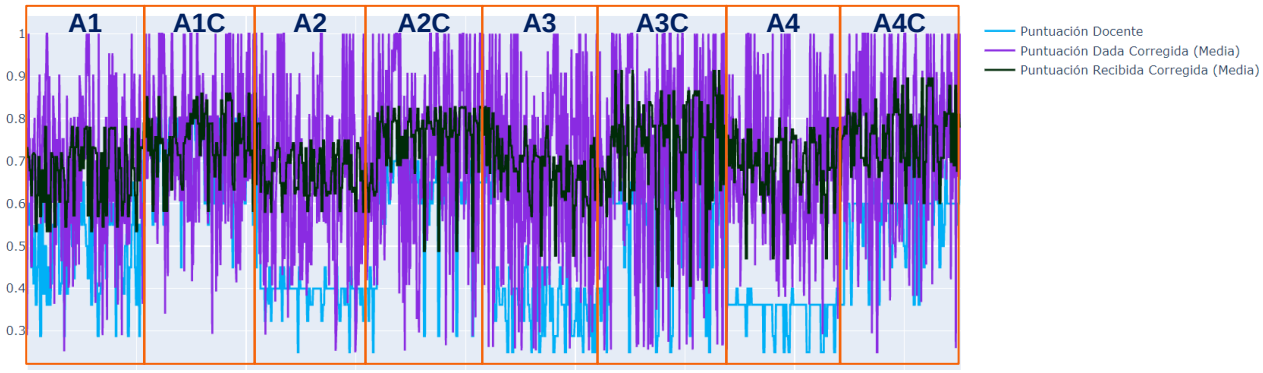
En la Figura **103** se presenta un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, por cada actividad de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2021.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 103. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2021 en escenario de educación virtual asincrónico

Análisis por cada actividad

En la relación entre puntuación dada y puntuación docente se obtuvo: 4 actividades (r=0.002 a 0.063) directa escasa, 5 actividades (r=-0.045 a -0.098) inversa escasa, 4 actividades (r=0.111 a 0.142) directa débil, 10 actividades (r=-0.103 a -0.185) inversa débil y 1 actividad (r=-0.311) inversa moderada (Tabla **78**).

En la relación entre puntuación recibida del colectivo y puntuación docente se obtuvo: aplicando media, 1 actividad (r=0.023) directa escasa, 5 actividades (r=0.108 a 0.209) directa débil, 2 actividades (r=-0.115 a -0.226) inversa débil, 3 actividades (r=0.357 a 0.459) directa moderada, 1 actividad (r=-0.347) inversa moderada, 10 actividades (r=0.517 a 0.664) directa sustancial, y 2 actividades (r=0.718-0.790) directa fuerte. Aplicando mediana, 1 actividad (r=0.034) directa escasa, 7 actividades (r=0.100 a 0.285) directa débil, 3 actividades (r=0.-108 a 0.257) inversa débil, 1 actividad (r=0.407) directa moderada, 1 actividad (r=-0.433) inversa moderada, 9 actividades (r=0.529 a 0.688) directa sustancial, y 2 actividades con (r=0.736-0.768) directa fuerte (Tabla **78**).

En la calibración aplicando la media, la relación entre puntuación dada y puntuación docente, 18 actividades tendieron a bajar, y 6 tendieron a subir; y en la relación entre puntuación recibida del colectivo y puntuación docente, 11 actividades tendieron a bajar, 2 se mantuvieron con el mismo valor y 11 tendieron a subir. Aplicando la mediana, la relación entre puntuación dada y puntuación docente, 13 actividades tendieron a bajar, 4 se mantuvieron con el mismo valor y 7 tendieron a subir; y en la relación entre puntuación recibida del colectivo y puntuación docente, 11 actividades tendieron a bajar, 2 se mantuvieron con el mismo valor y 11 tendieron a subir (Tabla **78**).

Tabla 78. Relación de las puntuaciones de los pares con la del docente por periodo académico y actividad de cada asignatura en escenario de educación virtual asincrónico

Actividad	Variables	X̄	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2021							
Asignatura: Fundamentos de ingeniería de software							
A1	Puntuación Docente	0.498	0.103				
	Puntuación Dada	0.678	0.162	-0.045			
	Puntuación Recibida (Media)	0.671	0.082			0.575	
	Puntuación Recibida (Mediana)	0.663	0.087			0.529	
	Puntuación Dada Calibrada	0.682	0.162		-0.043		
	Puntuación Recibida Calibrada (Media)	0.674	0.082				0.573
	Puntuación Dada Calibrada	0.681	0.162		-0.041		
	Puntuación Recibida Calibrada (Mediana)	0.666	0.087				0.527
A1C	Puntuación Docente	0.733	0.093				
	Puntuación Dada	0.749	0.173	-0.166			
	Puntuación Recibida (Media)	0.744	0.079			0.400	
	Puntuación Recibida (Mediana)	0.733	0.095			0.285	
	Puntuación Dada Calibrada	0.754	0.172		-0.164		
	Puntuación Recibida Calibrada (Media)	0.748	0.078				0.400
	Puntuación Dada Calibrada	0.753	0.172		-0.165		
	Puntuación Recibida Calibrada (Mediana)	0.739	0.095				0.279
A2	Puntuación Docente	0.390	0.041				
	Puntuación Dada	0.658	0.194	-0.062			
	Puntuación Recibida (Media)	0.663	0.053			0.357	
	Puntuación Recibida (Mediana)	0.660	0.080			0.226	
	Puntuación Dada Calibrada	0.664	0.193		-0.058		
	Puntuación Recibida Calibrada (Media)	0.669	0.054				0.352
	Puntuación Dada Calibrada	0.663	0.193		-0.060		
	Puntuación Recibida Calibrada (Mediana)	0.665	0.079				0.220
A2C	Puntuación Docente	0.632	0.095				
	Puntuación Dada	0.747	0.202	0.025			
	Puntuación Recibida (Media)	0.749	0.083			0.664	
	Puntuación Recibida (Mediana)	0.755	0.114			0.673	
	Puntuación Dada Calibrada	0.755	0.201		0.029		
	Puntuación Recibida Calibrada (Media)	0.757	0.082				0.674
	Puntuación Dada Calibrada	0.754	0.201		0.030		
	Puntuación Recibida Calibrada (Mediana)	0.765	0.115				0.679
A3	Puntuación Docente	0.364	0.072				
	Puntuación Dada	0.658	0.205	-0.052			
	Puntuación Recibida (Media)	0.661	0.070			0.531	
	Puntuación Recibida (Mediana)	0.658	0.092			0.663	
	Puntuación Dada Calibrada	0.665	0.205		-0.048		
	Puntuación Recibida Calibrada (Media)	0.668	0.070				0.532
	Puntuación Dada Calibrada	0.664	0.205		-0.047		
	Puntuación Recibida Calibrada (Mediana)	0.663	0.092				0.663

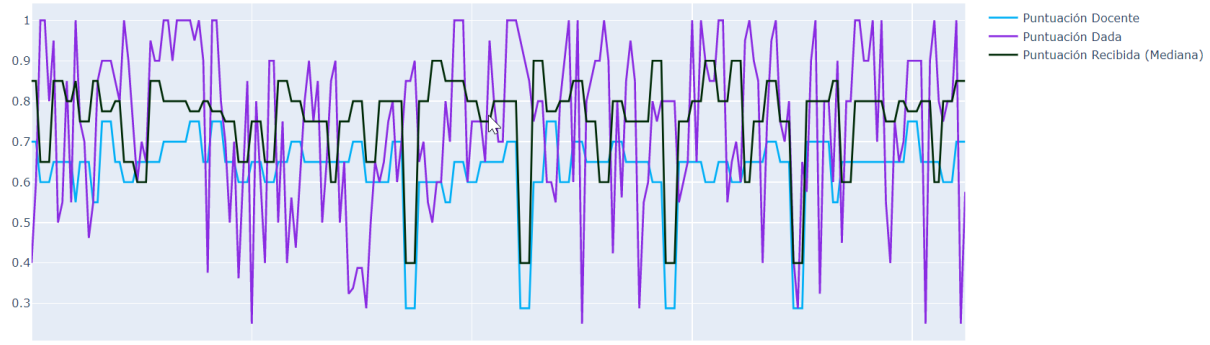
Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
A3C	Puntuación Docente	0.585	0.138				
	Puntuación Dada	0.725	0.230	-0.133			
	Puntuación Recibida (Media)	0.733	0.123			0.718	
	Puntuación Recibida (Mediana)	0.759	0.150			0.736	
	Puntuación Dada Calibrada	0.736	0.228		-0.119		
	Puntuación Recibida Calibrada (Media)	0.743	0.121				0.713
	Puntuación Dada Calibrada	0.735	0.228		-0.118		
	Puntuación Recibida Calibrada (Mediana)	0.770	0.148				0.748
A4	Puntuación Docente	0.346	0.042				
	Puntuación Dada	0.686	0.166	-0.103			
	Puntuación Recibida (Media)	0.688	0.081			0.632	
	Puntuación Recibida (Mediana)	0.684	0.071			0.610	
	Puntuación Dada Calibrada	0.689	0.166		-0.100		
	Puntuación Recibida Calibrada (Media)	0.692	0.081				0.631
	Puntuación Dada Calibrada	0.689	0.166		-0.101		
	Puntuación Recibida Calibrada (Mediana)	0.688	0.072				0.616
A4C	Puntuación Docente	0.581	0.086				
	Puntuación Dada	0.751	0.187	-0.116			
	Puntuación Recibida (Media)	0.754	0.081			0.533	
	Puntuación Recibida (Mediana)	0.769	0.092			0.603	
	Puntuación Dada Calibrada	0.758	0.186		-0.114		
	Puntuación Recibida Calibrada (Media)	0.761	0.080				0.530
	Puntuación Dada Calibrada	0.757	0.186		-0.110		
	Puntuación Recibida Calibrada (Mediana)	0.776	0.092				0.586
Periodo académico: octubre 2021 - febrero 2022							
Asignatura: Fundamentos de ingeniería de software							
A1	Puntuación Docente	0.724	0.111				
	Puntuación Dada	0.744	0.142	0.002			
	Puntuación Recibida (Media)	0.745	0.079			0.591	
	Puntuación Recibida (Mediana)	0.738	0.089			0.535	
	Puntuación Dada Calibrada	0.748	0.142		0.007		
	Puntuación Recibida Calibrada (Media)	0.748	0.079				0.595
	Puntuación Dada Calibrada	0.747	0.142		0.008		
	Puntuación Recibida Calibrada (Mediana)	0.741	0.088				0.538
A1C	Puntuación Docente	0.792	0.077				
	Puntuación Dada	0.828	0.147	0.063			
	Puntuación Recibida (Media)	0.829	0.068			0.459	
	Puntuación Recibida (Mediana)	0.845	0.091			0.260	
	Puntuación Dada Calibrada	0.833	0.146		0.064		
	Puntuación Recibida Calibrada (Media)	0.833	0.068				0.458
	Puntuación Dada Calibrada	0.833	0.146		0.064		
	Puntuación Recibida Calibrada (Mediana)	0.851	0.091				0.258
A2	Puntuación Docente	0.436	0.099				
	Puntuación Dada	0.742	0.180	-0.078			
	Puntuación Recibida (Media)	0.742	0.074			0.108	
	Puntuación Recibida (Mediana)	0.754	0.096			0.121	

Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
	Puntuación Dada Calibrada	0.748	0.179		-0.077		
	Puntuación Recibida Calibrada (Media)	0.748	0.073				0.103
	Puntuación Dada Calibrada	0.747	0.180		-0.077		
	Puntuación Recibida Calibrada (Mediana)	0.760	0.097				0.127
A2C	Puntuación Docente	0.721	0.185				
	Puntuación Dada	0.789	0.166	-0.129			
	Puntuación Recibida (Media)	0.789	0.066			0.517	
	Puntuación Recibida (Mediana)	0.810	0.087			0.534	
	Puntuación Dada Calibrada	0.794	0.165		-0.126		
	Puntuación Recibida Calibrada (Media)	0.795	0.066				0.521
	Puntuación Dada Calibrada	0.794	0.164		-0.126		
	Puntuación Recibida Calibrada (Mediana)	0.817	0.086				0.527
A3	Puntuación Docente	0.449	0.109				
	Puntuación Dada	0.747	0.200	0.111			
	Puntuación Recibida (Media)	0.753	0.090			0.183	
	Puntuación Recibida (Mediana)	0.771	0.131			0.240	
	Puntuación Dada Calibrada	0.755	0.198		0.110		
	Puntuación Recibida Calibrada (Media)	0.761	0.090				0.188
	Puntuación Dada Calibrada	0.755	0.199		0.113		
	Puntuación Recibida Calibrada (Mediana)	0.780	0.130				0.237
A3C	Puntuación Docente	0.654	0.126				
	Puntuación Dada	0.777	0.201	-0.098			
	Puntuación Recibida (Media)	0.784	0.102			0.570	
	Puntuación Recibida (Mediana)	0.799	0.153			0.529	
	Puntuación Dada Calibrada	0.787	0.198		-0.094		
	Puntuación Recibida Calibrada (Media)	0.794	0.100				0.569
	Puntuación Dada Calibrada	0.786	0.198		-0.093		
	Puntuación Recibida Calibrada (Mediana)	0.811	0.151				0.532
A4	Puntuación Docente	0.529	0.097				
	Puntuación Dada	0.768	0.151	0.133			
	Puntuación Recibida (Media)	0.770	0.066			0.169	
	Puntuación Recibida (Mediana)	0.773	0.081			0.034	
	Puntuación Dada Calibrada	0.772	0.151		0.132		
	Puntuación Recibida Calibrada (Media)	0.774	0.065				0.175
	Puntuación Dada Calibrada	0.771	0.151		0.133		
	Puntuación Recibida Calibrada (Mediana)	0.777	0.081				0.037
A4C	Puntuación Docente	0.788	0.132				
	Puntuación Dada	0.801	0.154	0.142			
	Puntuación Recibida (Media)	0.805	0.047			0.602	
	Puntuación Recibida (Mediana)	0.817	0.056			0.407	
	Puntuación Dada Calibrada	0.805	0.153		0.147		
	Puntuación Recibida Calibrada (Media)	0.809	0.046				0.604
	Puntuación Dada Calibrada	0.805	0.153		0.147		
	Puntuación Recibida Calibrada (Mediana)	0.822	0.055				0.397

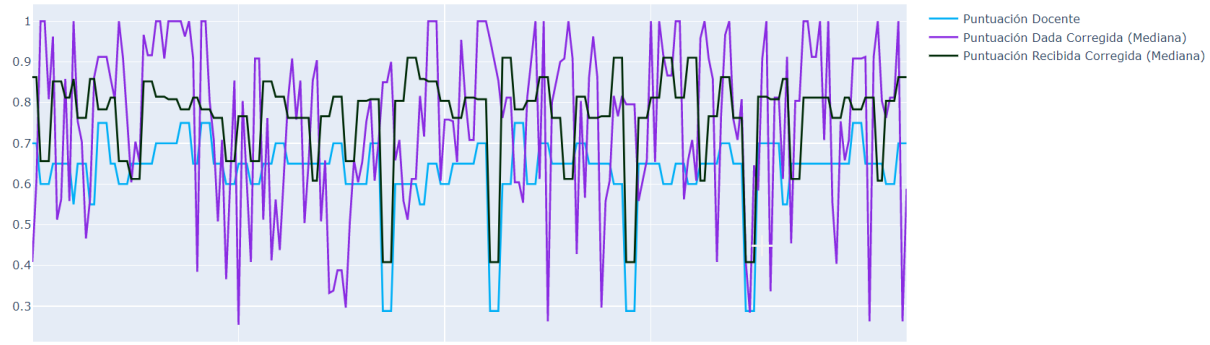
Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Asignatura: Ingeniería de software							
A1	Puntuación Docente	0.555	0.065				
	Puntuación Dada	0.699	0.172	-0.311			
	Puntuación Recibida (Media)	0.695	0.112			0.790	
	Puntuación Recibida (Mediana)	0.691	0.128			0.688	
	Puntuación Dada Calibrada	0.703	0.171		-0.305		
	Puntuación Recibida Calibrada (Media)	0.700	0.112				0.788
	Puntuación Dada Calibrada	0.703	0.172		-0.305		
	Puntuación Recibida Calibrada (Mediana)	0.696	0.130				0.684
A1C	Puntuación Docente	0.665	0.072				
	Puntuación Dada	0.812	0.143	-0.285			
	Puntuación Recibida (Media)	0.813	0.057			0.621	
	Puntuación Recibida (Mediana)	0.814	0.070			0.768	
	Puntuación Dada Calibrada	0.816	0.142		-0.284		
	Puntuación Recibida Calibrada (Media)	0.817	0.057				0.623
	Puntuación Dada Calibrada	0.816	0.142		-0.282		
	Puntuación Recibida Calibrada (Mediana)	0.819	0.070				0.767
A2	Puntuación Docente	0.483	0.050				
	Puntuación Dada	0.730	0.192	0.108			
	Puntuación Recibida (Media)	0.735	0.065			0.153	
	Puntuación Recibida (Mediana)	0.748	0.095			0.100	
	Puntuación Dada Calibrada	0.737	0.190		0.106		
	Puntuación Recibida Calibrada (Media)	0.742	0.064				0.153
	Puntuación Dada Calibrada	0.737	0.190		0.111		
	Puntuación Recibida Calibrada (Mediana)	0.754	0.094				0.092
A2C	Puntuación Docente	0.668	0.081				
	Puntuación Dada	0.837	0.137	0.015			
	Puntuación Recibida (Media)	0.835	0.042			-0.347	
	Puntuación Recibida (Mediana)	0.844	0.046			-0.257	
	Puntuación Dada Calibrada	0.840	0.136		0.020		
	Puntuación Recibida Calibrada (Media)	0.838	0.043				-0.345
	Puntuación Dada Calibrada	0.840	0.136		0.019		
	Puntuación Recibida Calibrada (Mediana)	0.849	0.045				-0.261
A3	Puntuación Docente	0.478	0.087				
	Puntuación Dada	0.749	0.175	-0.277			
	Puntuación Recibida (Media)	0.746	0.091			0.209	
	Puntuación Recibida (Mediana)	0.738	0.087			0.254	
	Puntuación Dada Calibrada	0.755	0.173		-0.279		
	Puntuación Recibida Calibrada (Media)	0.752	0.091				0.213
	Puntuación Dada Calibrada	0.755	0.174		-0.276		
	Puntuación Recibida Calibrada (Mediana)	0.745	0.087				0.256
A3C	Puntuación Docente	0.692	0.128				
	Puntuación Dada	0.835	0.146	-0.274			
	Puntuación Recibida (Media)	0.835	0.053			-0.115	
	Puntuación Recibida (Mediana)	0.846	0.069			-0.203	
	Puntuación Dada Calibrada	0.839	0.144		-0.274		

Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
	Puntuación Recibida Calibrada (Media)	0.838	0.052				-0.112
	Puntuación Dada Calibrada	0.839	0.144		-0.274		
	Puntuación Recibida Calibrada (Mediana)	0.851	0.068				-0.209
A4	Puntuación Docente	0.422	0.074				
	Puntuación Dada	0.739	0.178	-0.268			
	Puntuación Recibida (Media)	0.741	0.093			0.023	
	Puntuación Recibida (Mediana)	0.722	0.119			-0.108	
	Puntuación Dada Calibrada	0.744	0.176		-0.264		
	Puntuación Recibida Calibrada (Media)	0.746	0.092				0.025
	Puntuación Dada Calibrada	0.742	0.177		-0.268		
	Puntuación Recibida Calibrada (Mediana)	0.727	0.118				-0.108
A4C	Puntuación Docente	0.781	0.049				
	Puntuación Dada	0.745	0.176	-0.273			
	Puntuación Recibida (Media)	0.747	0.050			-0.226	
	Puntuación Recibida (Mediana)	0.751	0.071			-0.433	
	Puntuación Dada Calibrada	0.750	0.175		-0.269		
	Puntuación Recibida Calibrada (Media)	0.752	0.049				-0.229
	Puntuación Dada Calibrada	0.750	0.175		-0.270		
	Puntuación Recibida Calibrada (Mediana)	0.757	0.071				-0.434

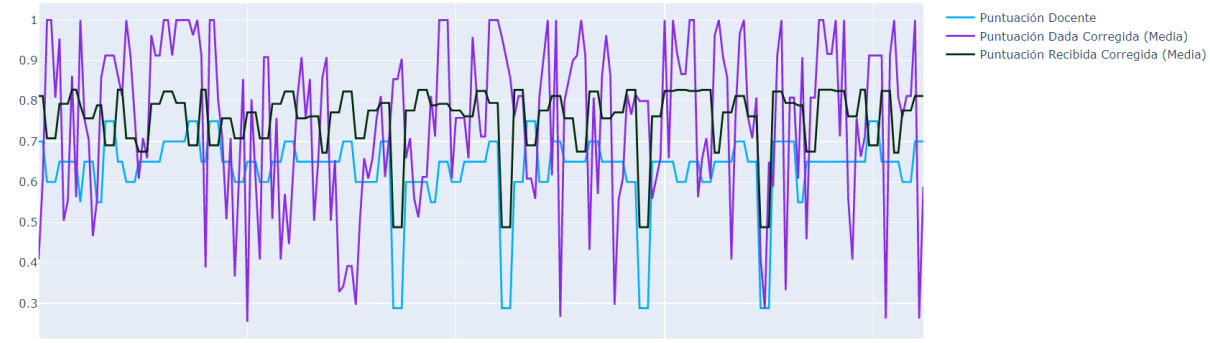
En la Figura **104** se muestra un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, de la actividad-1 de la asignatura de fundamentos de ingeniería de software.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 104. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la actividad-1 de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2021 en escenario de educación virtual asincrónico

Escenario de educación virtual sincrónico

Análisis por asignatura

En el escenario de educación virtual sincrónico se realizaron pruebas en 16 actividades. En la asignatura de fundamentos de ingeniería de software se ejecutaron 5 actividades, y en la asignatura de fundamentos de ofimática se ejecutaron 3 actividades en dos rondas (Tabla **79**).

Tabla 79. Relación de las puntuaciones de los pares con la del docente por periodo académico y asignatura en escenario de educación virtual sincrónico

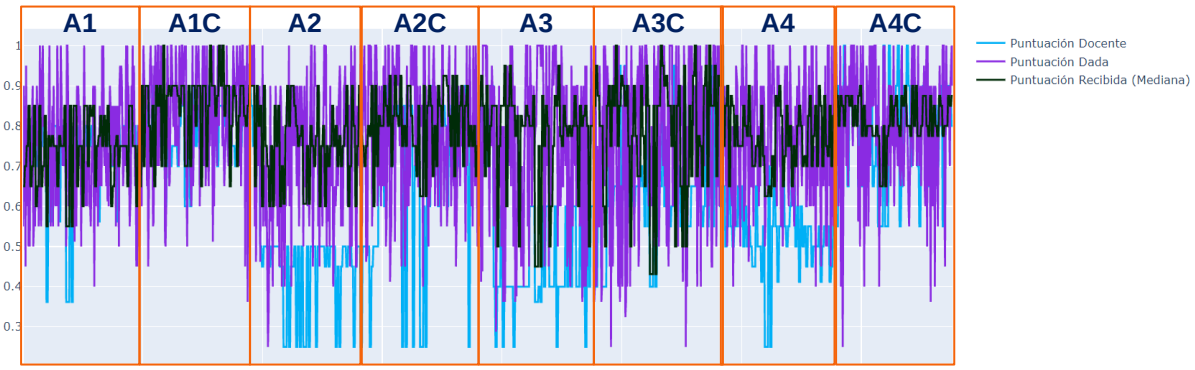
Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2022						
Asignatura: Fundamentos de ingeniería de software						
Puntuación Docente	0.665	0.174				
Puntuación Dada	0.741	0.179	0.248			
Puntuación Recibida (Media)	0.755	0.111			0.676	
Puntuación Recibida (Mediana)	0.757	0.133			0.642	
Puntuación Dada Calibrada	0.746	0.178		0.252		
Puntuación Recibida Calibrada (Media)	0.760	0.110				0.681
Puntuación Dada Calibrada	0.746	0.178		0.253		
Puntuación Recibida Calibrada (Mediana)	0.762	0.133				0.646
Asignatura: Fundamentos de ofimática						
Puntuación Docente	0.703	0.157				
Puntuación Dada	0.770	0.190	0.138			
Puntuación Recibida (Media)	0.777	0.143			0.681	
Puntuación Recibida (Mediana)	0.784	0.164			0.698	
Puntuación Dada Calibrada	0.776	0.188		0.143		
Puntuación Recibida Calibrada (Media)	0.783	0.141				0.682
Puntuación Dada Calibrada	0.776	0.188		0.143		
Puntuación Recibida Calibrada (Mediana)	0.791	0.162				0.699

En la asignatura de fundamentos de ingeniería de software de manera general los estudiantes tienden a otorgar puntuaciones ($\bar{X}$=0.741), y obtener puntuaciones del colectivo ($\bar{X}$=0.755-0.757), mayores que la del docente ($\bar{X}$=0.665). La relación entre puntuación dada y puntuación docente es directa débil (r=0.248), y la relación entre puntuación recibida del colectivo y puntuación docente es directa sustancial (media (r=0.676), mediana (r=0.642)). Con la calibración de la puntuación dada ($\bar{X}$=0.746), y de la puntuación recibida del colectivo ($\bar{X}$=0.760-0.762), tiende a subir la relación entre puntuación dada y puntuación docente (media (r=0.252), mediana (r=0.253)), y entre puntuación recibida del colectivo y puntuación docente (media

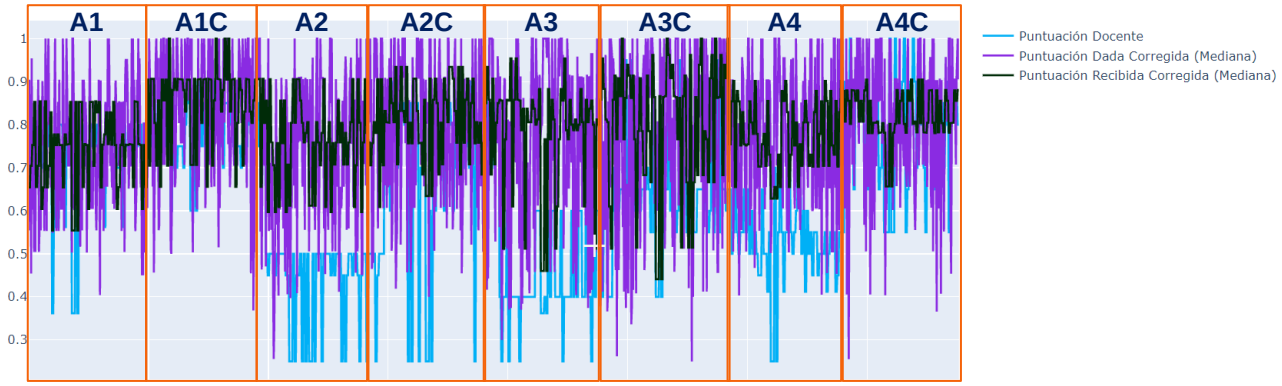
(r=0.681), mediana (r=0.646)), teniendo mayor relación con la media. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.110-0.133) es menor en relación a la del docente (Σ=0.174) (Tabla **79** y Figura **105**).

De igual forma en la asignatura de fundamentos de ofimática, los estudiantes tienden a otorgar puntuaciones ($\bar{X}$=0.770), y obtener puntuaciones del colectivo ($\bar{X}$=0.777-0.784) mayores que la del docente ($\bar{X}$=0.703). La relación entre puntuación dada y puntuación docente es directa débil (r=0.138), y la relación entre puntuación recibida del colectivo y puntuación docente es directa sustancial (media (r=0.681), mediana (r=0.698)). Con la calibración de la puntuación dada ($\bar{X}$=0.776), y de la puntuación recibida del colectivo ($\bar{X}$=0.783-0.791), tiende a subir la relación entre puntuación dada y puntuación docente (r=0.143), y entre puntuación recibida del colectivo y puntuación docente (media (r=0.682), mediana (r=0.699)), teniendo mayor relación con la mediana. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.141-0.162) es menor en relación a la del docente (Σ=0.157) (Tabla **79**).

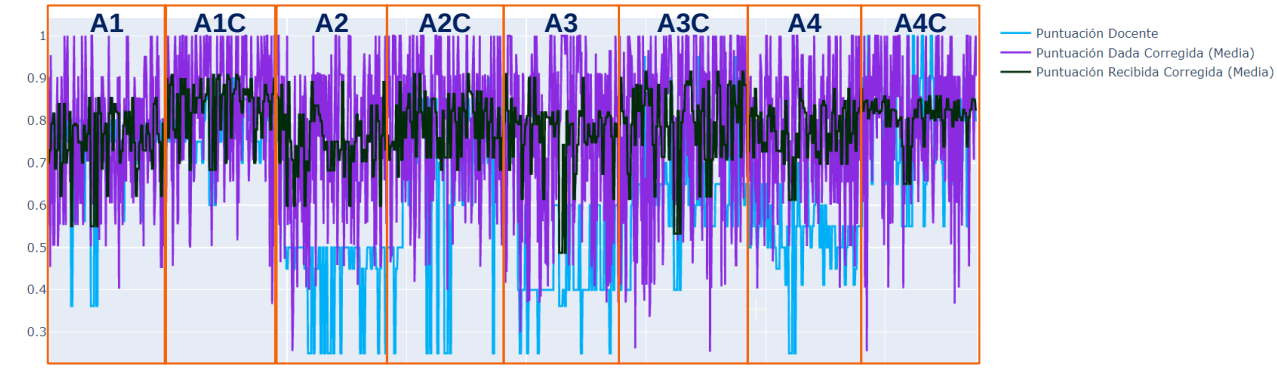
En la Figura **105** se presenta un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, por cada actividad de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 105. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022 en escenario de educación virtual sincrónico

Análisis por cada actividad

En la relación entre puntuación dada y puntuación docente se obtuvo: 3 actividades (r=0.016 a 0.038) directa escasa, 5 actividades (r=-0.013 a -0.088) inversa escasa, 1 actividad (r=0.197) directa débil, 6 actividades (r=-0.107 a -0.270) inversa débil y 1 actividad (r=-0.307) inversa moderada (Tabla **80**).

En la relación entre puntuación recibida del colectivo y puntuación docente se obtuvo: aplicando la media, 1 actividad (r=0.054) directa escasa, 5 actividades (r=0.111 a 0.269) directa débil, 1 actividad (r=-0.255) inversa débil, 2 actividades (r=0.347 a 0.475) directa moderada, 3 actividades (r=0.507 a 0.635) directa sustancial, y 4 actividades (r=0.743 a 0.951) directa fuerte. Aplicando mediana, 3 actividades (r=0.062 a 0.082) directa escasa, 1 actividad (r=-0.055) inversa escasa, 3 actividades (r=0.134 a 0.235) directa débil, 1 actividad (r=-0.217) inversa débil, 1 actividad (r=0.330) directa moderada, 3 actividades (r=0.521 a 0.694) directa sustancial, y 4 actividades (r=0.741 a 0.971) directa fuerte (Tabla **80**).

En la calibración aplicando la media, 9 actividades tendieron a bajar, 1 se mantuvo con el mismo valor y 6 tendieron a subir en la relación entre puntuación dada y puntuación docente; y, 4 actividades tendieron a bajar, 2 se mantuvieron con el mismos valor y 10 actividades tendieron a subir en la relación entre puntuación recibida del colectivo y puntuación docente. Aplicando la mediana, 10 actividades tendieron a bajar y 6 tendieron a subir en la relación entre puntuación dada y puntuación docente; y 4 actividades tendieron a bajar, 1 actividad se mantuvo con el mismo valor y 11 actividades tendieron a subir en la relación entre puntuación recibida del colectivo y puntuación docente (Tabla **80**).

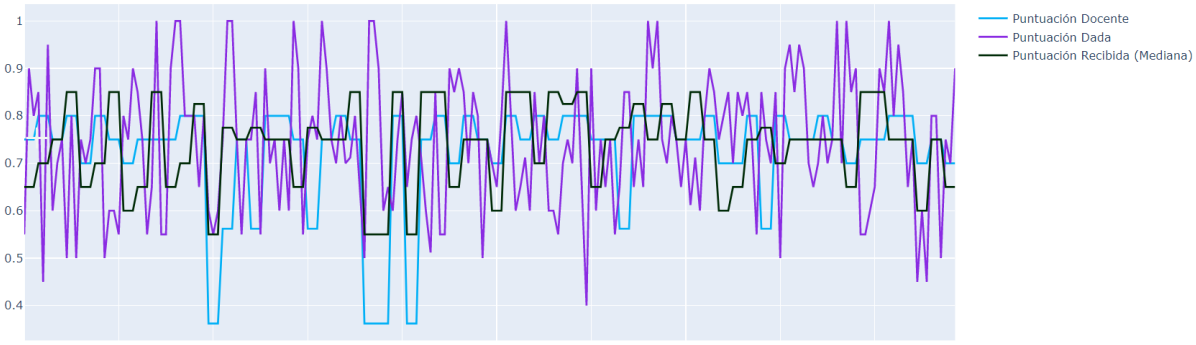
Tabla 80. Relación de las puntuaciones de los pares con la del docente por periodo académico y actividad de cada asignatura en escenario de educación virtual sincrónico

Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2022							
Asignatura: Fundamentos de ingeniería de software							
A1	Puntuación Docente	0.624	0.054				
	Puntuación Dada	0.758	0.132	-0.013			
	Puntuación Recibida (Media)	0.760	0.071			0.184	
	Puntuación Recibida (Mediana)	0.761	0.087			0.235	
	Puntuación Dada Calibrada	0.760	0.132		-0.014		
	Puntuación Recibida Calibrada (Media)	0.763	0.071				0.182
	Puntuación Dada Calibrada	0.760	0.132		-0.014		
	Puntuación Recibida Calibrada (Mediana)	0.763	0.087				0.236
A1C	Puntuación Docente	0.802	0.122				
	Puntuación Dada	0.830	0.139	0.197			
	Puntuación Recibida (Media)	0.834	0.049			0.111	
	Puntuación Recibida (Mediana)	0.850	0.069			-0.055	
	Puntuación Dada Calibrada	0.833	0.138		0.194		
	Puntuación Recibida Calibrada (Media)	0.837	0.048				0.119
	Puntuación Dada Calibrada	0.833	0.138		0.195		
	Puntuación Recibida Calibrada (Mediana)	0.854	0.068				-0.048
A2	Puntuación Docente	0.520	0.090				
	Puntuación Dada	0.713	0.176	-0.145			
	Puntuación Recibida (Media)	0.719	0.083			0.177	
	Puntuación Recibida (Mediana)	0.735	0.102			0.165	
	Puntuación Dada Calibrada	0.718	0.175		-0.145		
	Puntuación Recibida Calibrada (Media)	0.724	0.082				0.183
	Puntuación Dada Calibrada	0.718	0.175		-0.144		
	Puntuación Recibida Calibrada (Mediana)	0.740	0.102				0.161
A2C	Puntuación Docente	0.790	0.175				
	Puntuación Dada	0.783	0.177	-0.053			
	Puntuación Recibida (Media)	0.798	0.087			0.475	
	Puntuación Recibida (Mediana)	0.799	0.109			0.330	
	Puntuación Dada Calibrada	0.790	0.175		-0.054		
	Puntuación Recibida Calibrada (Media)	0.805	0.087				0.474
	Puntuación Dada Calibrada	0.790	0.175		-0.054		
	Puntuación Recibida Calibrada (Mediana)	0.808	0.108				0.333
A3	Puntuación Docente	0.540	0.126				
	Puntuación Dada	0.660	0.198	0.038			
	Puntuación Recibida (Media)	0.667	0.105			0.743	
	Puntuación Recibida (Mediana)	0.626	0.133			0.741	
	Puntuación Dada Calibrada	0.667	0.197		0.046		
	Puntuación Recibida Calibrada (Media)	0.673	0.104				0.744
	Puntuación Dada Calibrada	0.666	0.197		0.046		
	Puntuación Recibida Calibrada (Mediana)	0.634	0.133				0.734

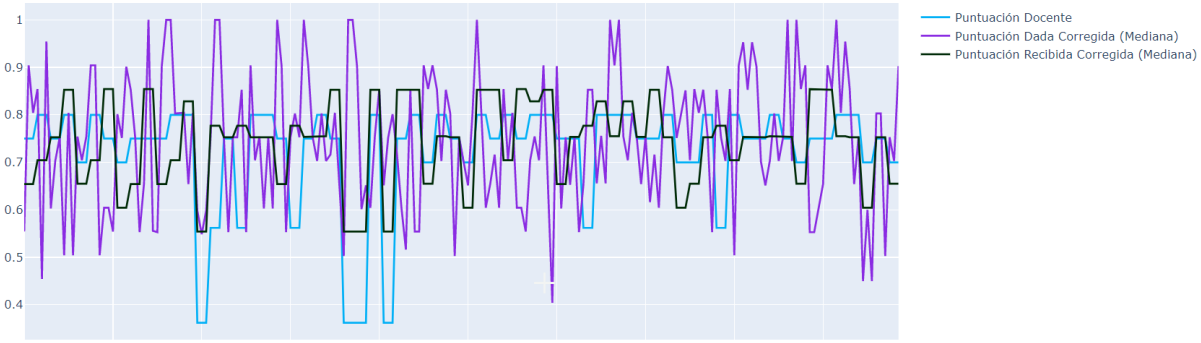
Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
A3C	Puntuación Docente	0.887	0.170				
	Puntuación Dada	0.821	0.180	-0.088			
	Puntuación Recibida (Media)	0.835	0.120			0.778	
	Puntuación Recibida (Mediana)	0.848	0.139			0.777	
	Puntuación Dada Calibrada	0.827	0.177		-0.084		
	Puntuación Recibida Calibrada (Media)	0.842	0.119				0.782
	Puntuación Dada Calibrada	0.827	0.177		-0.084		
	Puntuación Recibida Calibrada (Mediana)	0.856	0.137				0.783
A4	Puntuación Docente	0.526	0.052				
	Puntuación Dada	0.654	0.149	-0.245			
	Puntuación Recibida (Media)	0.671	0.087			0.054	
	Puntuación Recibida (Mediana)	0.665	0.084			0.082	
	Puntuación Dada Calibrada	0.657	0.149		-0.247		
	Puntuación Recibida Calibrada (Media)	0.673	0.087				0.053
	Puntuación Dada Calibrada	0.657	0.149		-0.247		
	Puntuación Recibida Calibrada (Mediana)	0.668	0.085				0.082
A4C	Puntuación Docente	0.687	0.055				
	Puntuación Dada	0.757	0.169	-0.270			
	Puntuación Recibida (Media)	0.790	0.089			0.240	
	Puntuación Recibida (Mediana)	0.796	0.109			0.062	
	Puntuación Dada Calibrada	0.762	0.169		-0.268		
	Puntuación Recibida Calibrada (Media)	0.795	0.089				0.238
	Puntuación Dada Calibrada	0.762	0.169		-0.268		
	Puntuación Recibida Calibrada (Mediana)	0.802	0.108				0.070
A5	Puntuación Docente	0.498	0.114				
	Puntuación Dada	0.640	0.171	0.024			
	Puntuación Recibida (Media)	0.662	0.109			0.507	
	Puntuación Recibida (Mediana)	0.650	0.124			0.531	
	Puntuación Dada Calibrada	0.644	0.171		0.028		
	Puntuación Recibida Calibrada (Media)	0.666	0.109				0.515
	Puntuación Dada Calibrada	0.644	0.171		0.027		
	Puntuación Recibida Calibrada (Mediana)	0.654	0.124				0.535
A5C	Puntuación Docente	0.789	0.095				
	Puntuación Dada	0.795	0.169	-0.037			
	Puntuación Recibida (Media)	0.821	0.075			0.577	
	Puntuación Recibida (Mediana)	0.839	0.089			0.521	
	Puntuación Dada Calibrada	0.801	0.167		-0.034		
	Puntuación Recibida Calibrada (Media)	0.826	0.074				0.579
	Puntuación Dada Calibrada	0.801	0.167		-0.034		
	Puntuación Recibida Calibrada (Mediana)	0.845	0.088				0.522
Asignatura: Fundamentos de ofimática							
A6	Puntuación Docente	0.618	0.124				
	Puntuación Dada	0.648	0.181	-0.107			
	Puntuación Recibida (Media)	0.659	0.144			0.872	
	Puntuación Recibida (Mediana)	0.666	0.167			0.874	

Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
	Puntuación Dada Calibrada	0.654	0.182		-0.093		
	Puntuación Recibida Calibrada (Media)	0.665	0.145				0.872
	Puntuación Dada Calibrada	0.654	0.182		-0.091		
	Puntuación Recibida Calibrada (Mediana)	0.673	0.168				0.878
A6C	Puntuación Docente	0.853	0.072				
	Puntuación Dada	0.780	0.188	-0.079			
	Puntuación Recibida (Media)	0.793	0.139			0.635	
	Puntuación Recibida (Mediana)	0.805	0.161			0.694	
	Puntuación Dada Calibrada	0.788	0.187		-0.071		
	Puntuación Recibida Calibrada (Media)	0.800	0.137				0.637
	Puntuación Dada Calibrada	0.788	0.187		-0.071		
	Puntuación Recibida Calibrada (Mediana)	0.813	0.159				0.697
A7	Puntuación Docente	0.636	0.179				
	Puntuación Dada	0.706	0.220	-0.197			
	Puntuación Recibida (Media)	0.709	0.180			0.951	
	Puntuación Recibida (Mediana)	0.691	0.198			0.971	
	Puntuación Dada Calibrada	0.716	0.217		-0.179		
	Puntuación Recibida Calibrada (Media)	0.720	0.178				0.951
	Puntuación Dada Calibrada	0.716	0.217		-0.179		
	Puntuación Recibida Calibrada (Mediana)	0.702	0.196				0.967
A7C	Puntuación Docente	0.810	0.054				
	Puntuación Dada	0.865	0.142	-0.117			
	Puntuación Recibida (Media)	0.865	0.049			0.347	
	Puntuación Recibida (Mediana)	0.906	0.063			0.134	
	Puntuación Dada Calibrada	0.868	0.140		-0.114		
	Puntuación Recibida Calibrada (Media)	0.868	0.049				0.357
	Puntuación Dada Calibrada	0.868	0.140		-0.114		
	Puntuación Recibida Calibrada (Mediana)	0.909	0.061				0.138
A8	Puntuación Docente	0.526	0.086				
	Puntuación Dada	0.745	0.170	0.016			
	Puntuación Recibida (Media)	0.754	0.092			0.269	
	Puntuación Recibida (Mediana)	0.740	0.100			0.072	
	Puntuación Dada Calibrada	0.751	0.169		0.015		
	Puntuación Recibida Calibrada (Media)	0.760	0.092				0.273
	Puntuación Dada Calibrada	0.751	0.169		0.013		
	Puntuación Recibida Calibrada (Mediana)	0.747	0.100				0.086
A8C	Puntuación Docente	0.779	0.053				
	Puntuación Dada	0.870	0.123	-0.307			
	Puntuación Recibida (Media)	0.876	0.062			-0.255	
	Puntuación Recibida (Mediana)	0.894	0.076			-0.217	
	Puntuación Dada Calibrada	0.872	0.122		-0.308		
	Puntuación Recibida Calibrada (Media)	0.878	0.061				-0.257
	Puntuación Dada Calibrada	0.872	0.122		-0.308		
	Puntuación Recibida Calibrada (Mediana)	0.896	0.074				-0.223

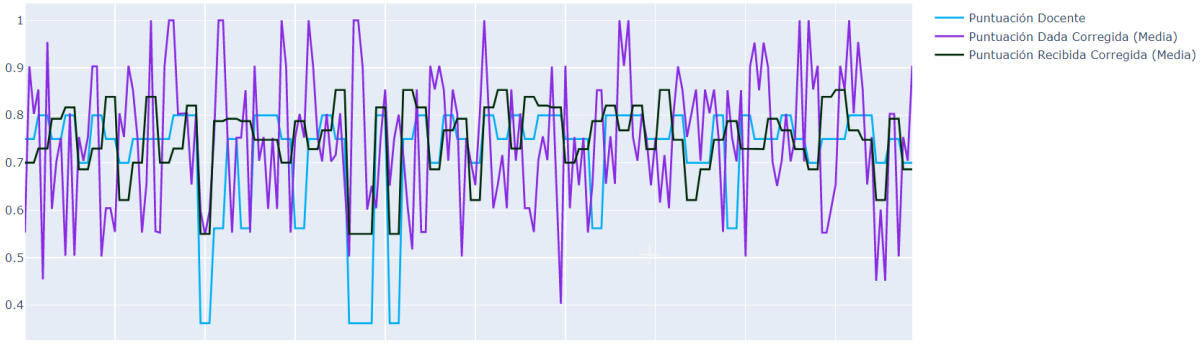
En la Figura **106** se muestra un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, de la actividad-1 de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 106. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la actividad-1 de la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022 en escenario de educación virtual sincrónico

Escenario de educación presencial

Análisis por asignatura

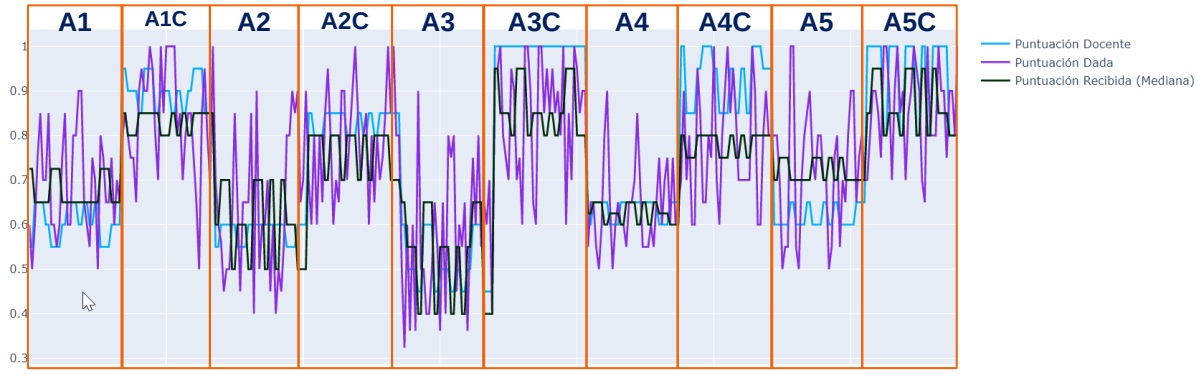
En el escenario de educación presencial se realizaron pruebas en 5 actividades en la asignatura de ingeniería de software en dos rondas (Tabla **81**).

Tabla 81. Relación de las puntuaciones de los pares con la del docente por periodo académico y asignatura en escenario de educación presencial

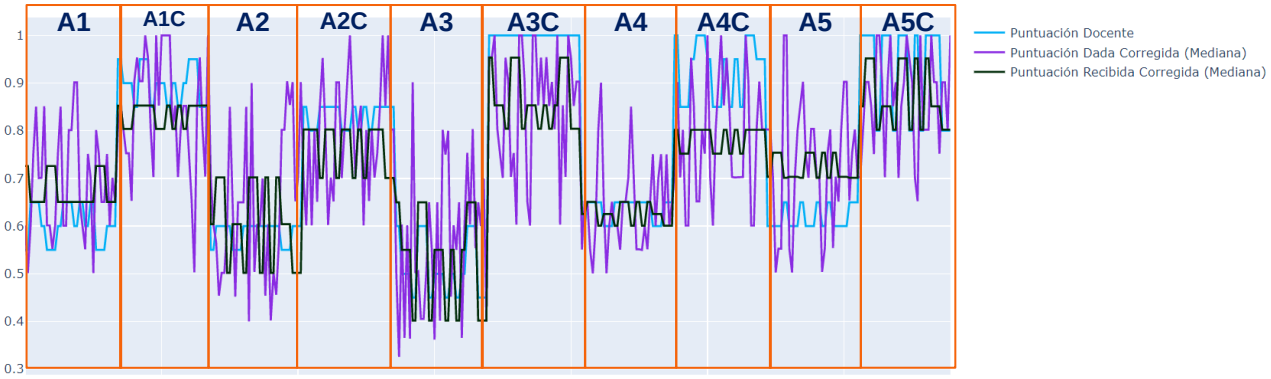
Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2022 Asignatura: Ingeniería de software						
Puntuación Docente	0.754	0.177				
Puntuación Dada	0.738	0.159	0.538			
Puntuación Recibida (Media)	0.736	0.113			0.838	
Puntuación Recibida (Mediana)	0.725	0.124			0.852	
Puntuación Dada Calibrada	0.740	0.159		0.540		
Puntuación Recibida Calibrada (Media)	0.738	0.113				0.839
Puntuación Dada Calibrada	0.740	0.159		0.540		
Puntuación Recibida Calibrada (Mediana)	0.727	0.125				0.852

En la asignatura de ingeniería de software, de manera general los estudiantes tienden a otorgar puntuaciones ($\bar{X}$=0.724), y obtener puntuaciones del colectivo ($\bar{X}$=0.709-0.722), menores que la del docente ($\bar{X}$=0.725). La relación entre puntuación dada y puntuación docente es directa sustancial (r=0.564), y la relación entre puntuación recibida y puntuación docente es directa fuerte (media (r=0.855), mediana (r=0.862)). Con la calibración de la puntuación dada ($\bar{X}$=0.726), y de la puntuación recibida del colectivo ($\bar{X}$=0.711-0.724), tiende a subir la relación entre puntuación dada y puntuación docente (media (r=0.565), mediana (r=0.566)), y entre puntuación recibida del colectivo y puntuación docente (media (r=0.856)), mediana (r=0.852)), teniendo mayor relación con la mediana. La desviación estándar de las puntuaciones recibidas calibradas del colectivo (Σ=0.116-0.127) es menor en relación a la del docente (Σ=0.169) (Tabla **81** y Figura **107**)

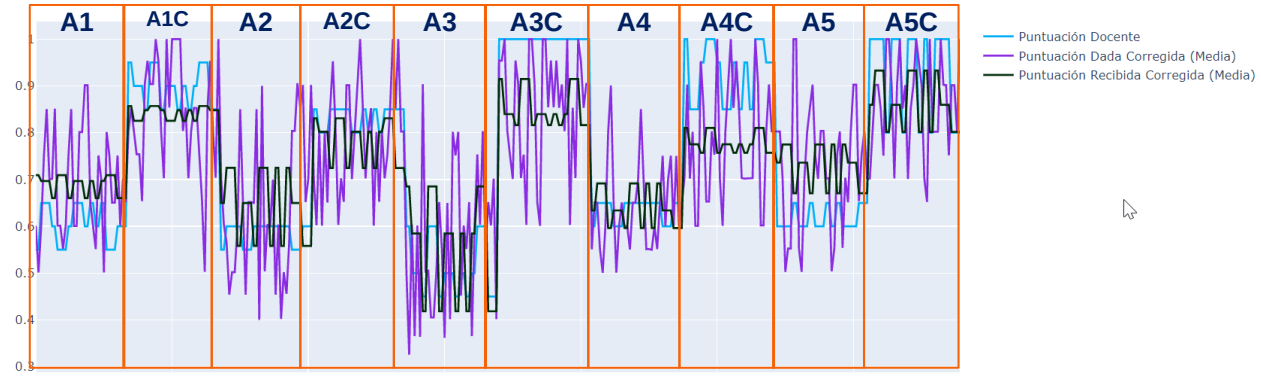
En la Figura **107** se presenta un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, por cada actividad de la asignatura de ingeniería de software del periodo académico mayo-septiembre 2022.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 107. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la asignatura de ingeniería de software del periodo académico mayo-septiembre 2022 en escenario de educación presencial

Análisis por cada actividad

En la relación entre puntuación dada y puntuación docente se obtuvo: 1 actividad (r=0.039) directa escasa, 1 actividad (r-0.002) inversa escasa, 2 actividad (r=0.107 a 0.158) directa débil, 4 actividades (r=-0.129 a -0.262) inversa débil, 1 actividad (r=0.455) directa moderada, y 1 actividad (r=-0.388) inversa moderada (Tabla **82**).

En la relación entre puntuación recibida del colectivo y puntuación docente se obtuvo: aplicando la media, 1 actividad (r=-0.064) inversa escasa, 2 actividades (r=0.103 a 0.183) directa débil, 2 actividades (r=-0.183-0.263) inversa débil, 1 actividad (r=0.408) directa moderada, y 1 actividad (r=-0.304) inversa moderada, 2 actividades (r=0.846 a 0.943) directa fuerte y 1 actividad (r=-0.939) inversa fuerte. Aplicando mediana, 2 actividades (r=0.000) directa escasa, 2 actividades (r=-0.054 a -0.253) inversa escasa, 2 actividades (r=-0.545) inversa sustancial, 3 actividades (r=0.780 a 0.951) directa fuerte y 1 actividad (r=-0.852) inversa fuerte (Tabla **82**).

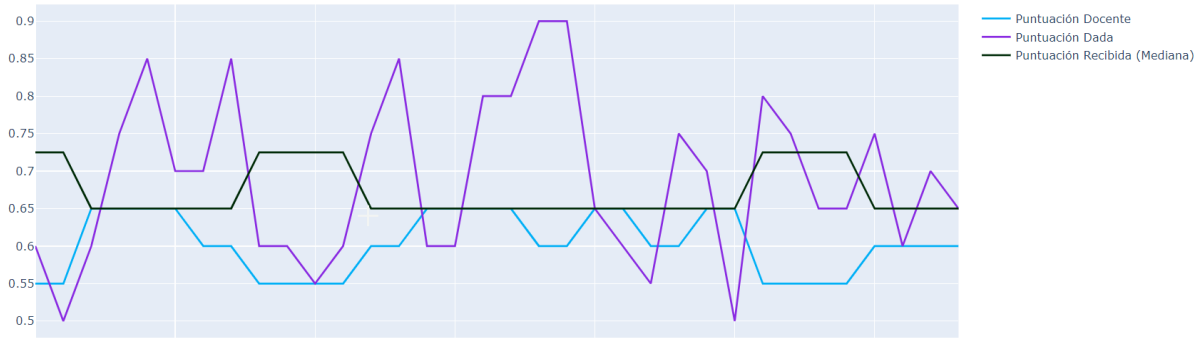
En la calibración aplicando la media, 4 actividades tendieron a bajar, 1 se mantuvo con el mismo valor y 5 tendieron a subir en la relación entre puntuación dada y puntuación docente; y, 4 actividades tendieron a bajar, y 6 actividades tendieron a subir en la relación entre puntuación recibida del colectivo y puntuación docente. Aplicando la mediana, 4 actividades tendieron a bajar, 2 se mantuvieron con el mismo valor y 4 tendieron a subir en la relación entre puntuación dada y puntuación docente; y 2 actividades tendieron a bajar, 2 se mantuvieron con el mismo valor y 6 tendieron a subir en la relación entre puntuación recibida del colectivo y puntuación docente (Tabla **82**).

Tabla 82. Relación de las puntuaciones de los pares con la del docente por periodo académico y actividad de cada asignatura en escenario de educación virtual presencial

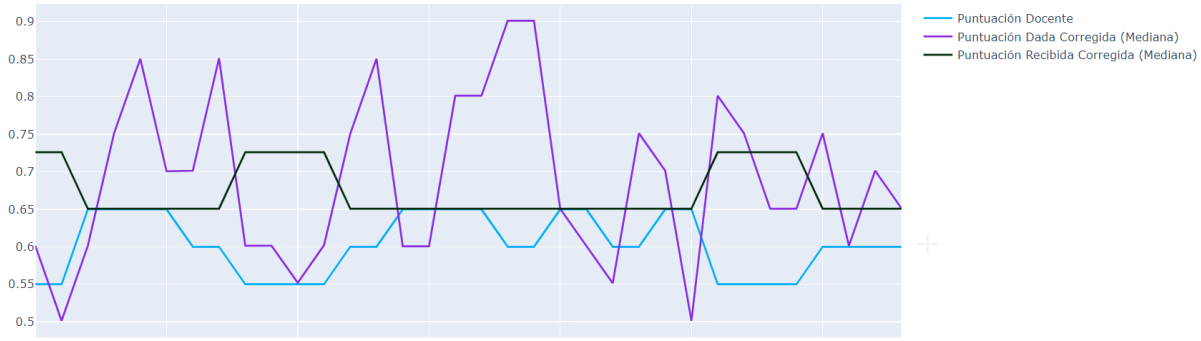
Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
Periodo académico: mayo-septiembre 2022							
Asignatura: Ingeniería de software							
A1	Puntuación Docente	0.603	0.041				
	Puntuación Dada	0.688	0.112	0.158			
	Puntuación Recibida (Media)	0.687	0.021			-0.183	
	Puntuación Recibida (Mediana)	0.672	0.035			-0.852	
	Puntuación Dada Calibrada	0.689	0.111		0.157		
	Puntuación Recibida Calibrada (Media)	0.688	0.021				-0.192
	Puntuación Dada Calibrada	0.689	0.111		0.157		
	Puntuación Recibida Calibrada (Mediana)	0.673	0.035				-0.852
A1C	Puntuación Docente	0.897	0.041				
	Puntuación Dada	0.841	0.124	-0.140			
	Puntuación Recibida (Media)	0.840	0.014			0.183	
	Puntuación Recibida (Mediana)	0.832	0.024			-0.054	
	Puntuación Dada Calibrada	0.844	0.123		-0.140		
	Puntuación Recibida Calibrada (Media)	0.843	0.013				0.224
	Puntuación Dada Calibrada	0.844	0.123		-0.140		
	Puntuación Recibida Calibrada (Mediana)	0.835	0.024				-0.062
A2	Puntuación Docente	0.585	0.023				
	Puntuación Dada	0.642	0.158	-0.262			
	Puntuación Recibida (Media)	0.641	0.072			-0.064	
	Puntuación Recibida (Mediana)	0.600	0.085			0.000	
	Puntuación Dada Calibrada	0.645	0.158		-0.267		
	Puntuación Recibida Calibrada (Media)	0.644	0.071				-0.054
	Puntuación Dada Calibrada	0.644	0.158		-0.262		
	Puntuación Recibida Calibrada (Mediana)	0.603	0.085				-0.011
A2C	Puntuación Docente	0.832	0.024				
	Puntuación Dada	0.785	0.118	0.039			
	Puntuación Recibida (Media)	0.781	0.046			-0.304	
	Puntuación Recibida (Mediana)	0.765	0.049			-0.545	
	Puntuación Dada Calibrada	0.787	0.117		0.042		
	Puntuación Recibida Calibrada (Media)	0.783	0.045				-0.311
	Puntuación Dada Calibrada	0.787	0.118		0.040		
	Puntuación Recibida Calibrada (Mediana)	0.767	0.049				-0.545
A3	Puntuación Docente	0.512	0.062				
	Puntuación Dada	0.564	0.152	-0.253			
	Puntuación Recibida (Media)	0.553	0.111			0.943	
	Puntuación Recibida (Mediana)	0.526	0.103			0.951	
	Puntuación Dada Calibrada	0.566	0.152		-0.248		
	Puntuación Recibida Calibrada (Media)	0.556	0.111				0.942
	Puntuación Dada Calibrada	0.566	0.152		-0.244		
	Puntuación Recibida Calibrada (Mediana)	0.527	0.102				0.952

Actividad	Variables	$\bar{X}$	Σ	r Puntuación Dada-Docente	r Puntuación Dada Calibrada-Docente	r Puntuación Recibida-Docente	r Puntuación Recibida Calibrada-Docente
A3C	Puntuación Docente	1.000	0.002				
	Puntuación Dada	0.856	0.129	-0.129			
	Puntuación Recibida (Media)	0.851	0.041			-0.263	
	Puntuación Recibida (Mediana)	0.862	0.062			-0.253	
	Puntuación Dada Calibrada	0.859	0.129		-0.131		
	Puntuación Recibida Calibrada (Media)	0.853	0.041				-0.262
	Puntuación Dada Calibrada	0.858	0.129		-0.132		
	Puntuación Recibida Calibrada (Mediana)	0.865	0.061				-0.254
A4	Puntuación Docente	0.635	0.023				
	Puntuación Dada	0.640	0.094	-0.002			
	Puntuación Recibida (Media)	0.640	0.040			0.103	
	Puntuación Recibida (Mediana)	0.625	0.021			0.000	
	Puntuación Dada Calibrada	0.640	0.094		-0.003		
	Puntuación Recibida Calibrada (Media)	0.641	0.041				0.112
	Puntuación Dada Calibrada	0.640	0.094		-0.003		
	Puntuación Recibida Calibrada (Mediana)	0.626	0.021				0.009
A4C	Puntuación Docente	0.929	0.063				
	Puntuación Dada	0.779	0.126	0.107			
	Puntuación Recibida (Media)	0.776	0.022			0.455	
	Puntuación Recibida (Mediana)	0.782	0.024			0.946	
	Puntuación Dada Calibrada	0.782	0.126		0.109		
	Puntuación Recibida Calibrada (Media)	0.779	0.022				0.447
	Puntuación Dada Calibrada	0.781	0.126		0.107		
	Puntuación Recibida Calibrada (Mediana)	0.784	0.024				0.946
A5	Puntuación Docente	0.618	0.024				
	Puntuación Dada	0.725	0.138	0.408			
	Puntuación Recibida (Media)	0.724	0.045			-0.939	
	Puntuación Recibida (Mediana)	0.718	0.024			-0.545	
	Puntuación Dada Calibrada	0.728	0.137		0.409		
	Puntuación Recibida Calibrada (Media)	0.727	0.045				-0.936
	Puntuación Dada Calibrada	0.728	0.137		0.409		
	Puntuación Recibida Calibrada (Mediana)	0.720	0.025				-0.572
A5C	Puntuación Docente	0.929	0.097				
	Puntuación Dada	0.863	0.103	-0.388			
	Puntuación Recibida (Media)	0.864	0.056			0.846	
	Puntuación Recibida (Mediana)	0.868	0.065			0.780	
	Puntuación Dada Calibrada	0.865	0.103		-0.387		
	Puntuación Recibida Calibrada (Media)	0.865	0.056				0.847
	Puntuación Dada Calibrada	0.865	0.103		-0.387		
	Puntuación Recibida Calibrada (Mediana)	0.869	0.065				0.777

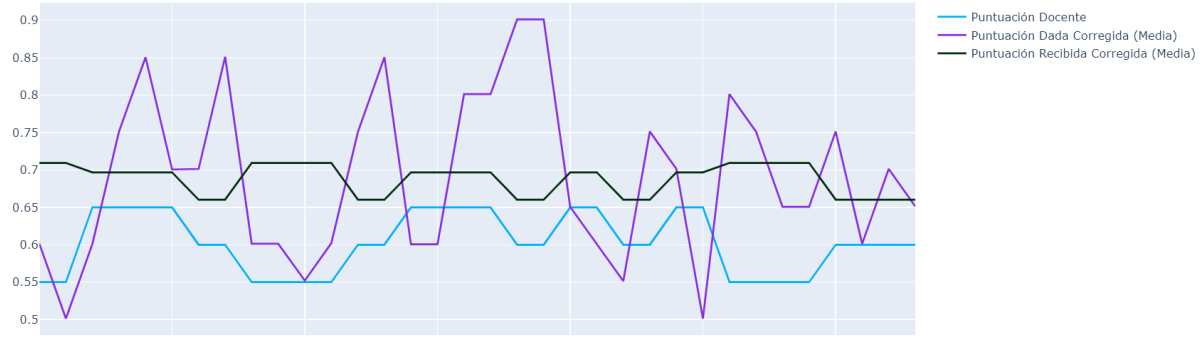
En la Figura **108** se muestra un ejemplo de comparación entre puntuación dada/recibida y puntuación del docente, y, puntuación calibrada (media/mediana) y puntuación del docente, de la actividad-1 de la asignatura de ingeniería de software del periodo académico mayo-septiembre 2022.



a) Correlación entre Puntuación Dada/Recibida y Puntuación Docente



b) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (mediana) y Puntuación Docente



c) Correlación entre Puntuación Dada Calibrada/Recibida Calibrada (media) y Puntuación Docente

Figura 108. Comparación entre: (a) Puntuación Dada/Recibida y Puntuación Docente, (b) Puntuación Calibrada (mediana) y Puntuación Docente, (c) Puntuación Calibrada (media) y Puntuación Docente, de la actividad-1 de la asignatura de ingeniería de software del periodo académico mayo-septiembre 2022 en escenario de educación presencial

En la Tabla **83** se presenta un resumen de las tablas (Tabla **78**, Tabla **80** y Tabla **82**) de la relación entre la puntación que recibe el estudiante del colectivo y la puntuación que proporciona el docente con media y mediana por escenario de educación.

Tabla 83. Resumen de relación entre puntuación recibida del colectivo y puntuación docente por escenario de educación

	Escenario de educación											
	Virtual asincrónico				**Virtual sincrónico**				**Presencial**			
Relación	Media		Mediana		Media		Mediana		Media		Mediana	
Directa escasa	1	4%	1	4%	1	6%	3	19%		0%	2	20%
Inversa escasa		0%		0%		0%	1	6%	1	10%	2	20%
Directa débil	5	21%	7	29%	5	31%	3	19%	2	20%		0%
Inversa débil	2	8%	3	13%	1	6%	1	6%	2	20%		0%
Directa moderada	3	13%	1	4%	2	13%	1	6%	1	10%		0%
Inversa moderada	1	4%	1	4%		0%		0%	1	10%		0%
Directa sustancial	**10**	42%	**9**	38%	**3**	19%	**3**	19%		0%	**2**	20%
Directa fuerte	**2**	8%	**2**	8%	**4**	25%	**4**	25%	**2**	20%	**3**	30%
Inversa fuerte		0%		0%		0%		0%	**1**	10%	**1**	10%
Total	24	100%	24	100%	16	100%	16	100%	10	100%	10	100%

En la Tabla **84** se muestra un resumen de las tablas (Tabla **78**, Tabla **80** y Tabla **82**) sobre las tendencias que se presentaron al calibrar la puntuación de evaluación de tarea con media y mediana por escenario de educación.

Tabla 84. Tendencias de aplicación de calibración por escenario de educación

	Escenario de educación											
	Virtual asincrónico				**Virtual sincrónico**				**Presencial**			
Calibración	Media		Mediana		Media		Mediana		Media		Mediana	
Tiende a subir	11	46%	11	46%	10	63%	11	69%	5	50%	6	60%
Se mantiene	2	8%	2	8%	2	13%	1	6%		0%	3	30%
Tiende a bajar	11	46%	11	46%	4	25%	4	25%	5	50%	1	10%
Total	24	100%	24	100%	16	100%	16	100%	10	100%	10	100%

Una vez realizado el análisis estadístico descriptivo, se concluye que:

- En todas las asignaturas impartidas ya sea en escenario de educación virtual asincrónico, sincrónico, o presencial, la puntuación media que proporciona el docente llega a ser menor a la de los estudiantes, exceptuando en la asignatura de ingeniería de software impartida en escenario de educación presencial, donde la puntuación del docente se encuentra muy cercana a la de los estudiantes, que resulta de la compensación que existe entre los valores altos que se obtienen después de haber realizado la corrección del trabajo.
- Los cálculos con media y mediana denotaron (Tabla **83**):
  - En escenarios de educación virtual asincrónico, no existe diferencia significativa de cálculos de media/mediana de la puntuación del colectivo con la del docente: directa sustancial (media (42%), mediana (38%)), y directa fuerte (media (8%), mediana (8%)).
  - En escenarios de educación virtual sincrónico, tanto la media/mediana de la puntuación del colectivo alcanza igual relación con la del docente: directa sustancial (media (19%), mediana (19%)), y directa fuerte (media (25%), mediana (25%)).
  - En escenarios de educación presencial, la mediana de la puntuación del colectivo alcanza mayor relación con la del docente: directa sustancial (media (0%), mediana (20%)), directa fuerte (media (20%), mediana (30%)), e inversa fuerte (media (10%), mediana (10%)).
  - Finiquitando que se puede aplicar tanto la media como la mediana para realizar los cálculos de puntuación del colectivo, ya que no existe diferencia significativa, la aplicación depende del escenario. No obstante, se destaca, que con la mediana se obtuvo bueno resultados en escenario presencial.
- De todas las asignaturas impartidas, la relación entre la puntación que da el evaluador con la puntuación que otorga el docente, obtuvo:
  - Similaridad directa escasa y débil en las asignaturas impartidas en escenario de educación virtual asincrónico.
  - Similaridad débil en las asignaturas impartidas en escenario de educación virtual sincrónico.
  - Similaridad directa sustancial en la asignatura impartida en escenario de educación presencial.
  - Coligiendo que las puntuaciones individuales no alcanzaron relación fuerte con la del docente en ningún escenario de educación.

- De todas las asignaturas impartidas, la relación entre la puntación que recibe el estudiante del colectivo con la puntuación que proporciona el docente, obtuvo:
  - Similaridad directa moderada y sustancial en las asignaturas impartidas en escenario de educación virtual asincrónico.
  - Similaridad directa sustancial en las asignaturas impartidas en escenario de educación virtual sincrónico.
  - Similaridad directa fuerte en la asignatura impartida en escenario de educación presencial.
  - Deduciendo que las puntuaciones del colectivo alcanzan mejor correlación con la del docente en escenario de educación virtual sincrónico y presencial.
- Las relaciones alcanzadas en el nivel más alto entre puntación que recibe el estudiante del colectivo y puntuación que proporciona el docente por cada actividad denotaron variabilidad en cada escenario:
  - En escenario virtual asincrónico las actividades ejecutadas alcanzaron el 42% con media y 38% con mediana en similaridad sustancial (r=0.517 a 0.688), y el 8% aplicando media y mediana en similaridad fuerte (r=0.718-0.790) (Tabla **78**).
  - En escenario virtual sincrónico se adquirió el 19% con media y mediana en similaridad sustancial (r=0.507 a 0.694), y el 25% con media y mediana en similaridad fuerte (r=0.741 a 0.971) (Tabla **80**).
  - En escenario presencial se obtuvo el 20% con mediana en similaridad sustancial (r=-0.545), el 30% con media y mediana en similaridad directa fuerte (r=0.780 a 0.951), y el 10% con media y mediana en similaridad inversa fuerte (r=-0.852) (Tabla **82**).
  - Concluyendo que el modelo se puede aplicar en todos los escenarios de educación analizados, con mayor efectividad en el presencial.
- En la calibración a la puntuación de evaluación de tarea, los resultados mostraron que: aplicando la media, la correlación entre puntuación recibida del colectivo y puntuación docente tendió a bajar el 46% de las actividades en escenario virtual asincrónico, 25% en virtual sincrónico y 50% en presencial. Se mantuvieron con el mismo valor, el 8% de las actividades en escenario virtual asincrónico, y 13% en virtual sincrónico. Tendió a subir el 46% de las actividades en escenario virtual asincrónico, 63% virtual sincrónico y 50% en presencial. Aplicando la mediana, la correlación entre puntuación recibida del colectivo y puntuación docente tendió a bajar el 46% de las actividades en escenario virtual asincrónico, 25% en

virtual sincrónico y 10% en presencial. Se mantuvieron con el mismo valor, el 8% de las actividades en escenario virtual asincrónico, 6% en virtual sincrónico y 30% en presencial. Tendió a subir el 46% de las actividades en escenario virtual asincrónico, 69% virtual sincrónico y 60% en presencial (Tabla **84**). Se colige que no existe diferencia significativa entre la aplicación de media y mediana en la calibración, sin embargo, con la mediana se obtuvieron mejores resultados en el escenario virtual sincrónico y presencial.

- La desviación estándar de las puntuaciones recibidas calibradas del colectivo (0.080-0.162) es menor en relación a la puntuación del docente (0.142-0.177) en todas las asignaturas impartidas ya sea en escenario de educación virtual asincrónico, sincrónico o presencial. Coligiendo que entre evaluadores se manejan un mismo criterio para evaluar la tarea, y que el docente proporciona un rango más amplio de puntuaciones.

5.2.2. Validación de hipótesis y utilidad del modelo propuesto

Para responder la pregunta: **¿De qué manera el modelo de evaluación entre pares basado en análisis de sentimiento podría contribuir en el proceso de enseñanza-aprendizaje?,** se evaluó el rendimiento estudiantil en dos rondas en escenario de educación virtual sincrónica/asincrónica y presencial, y la percepción de los estudiantes en el proceso de evaluación entre pares. A continuación, se detalla:

5.2.2.1. Validación de hipótesis

Se evaluó si existe mejora del rendimiento estudiantil en la segunda ronda aplicando análisis de sentimiento en la evaluación entre pares, mediante las siguientes hipótesis:

- H0: En la evaluación entre pares aplicando análisis de sentimiento no existen diferencias estadísticamente significativas en el rendimiento estudiantil.
- H1: En la evaluación entre pares aplicando análisis de sentimiento existen diferencias estadísticamente significativas en el rendimiento estudiantil.

Siendo las variables:
Dependiente: evaluación entre pares.
Independiente: análisis de sentimiento.

Para validar la hipótesis, se aplicó la prueba t de Student por cada actividad de cada asignatura en escenario virtual asincrónico, sincrónico y presencial (Tabla **85**). Los resultados mostraron:

- En escenario virtual asincrónico, las 12 actividades obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, y un incremento en el rendimiento del estudiante de (3%-12%).
- En escenario virtual sincrónico, las 8 actividades obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, y un incremento en el rendimiento del estudiante de (7%-22%).
- En escenario presencial, las 7 actividades obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, y un incremento en el rendimiento del estudiante de (15%-34%).

- Se colige que, si existe diferencia significativa en todas las actividades entre la puntuación media de la primera y segunda ronda aplicando análisis de sentimiento en la evaluación entre pares, por tanto, se rechaza la hipótesis nula.

Tabla 85. Resultados de significancia de la aplicación de análisis de sentimiento en evaluación entre pares

Escenario de educación	Actividad	$\bar{X}$ Ronda-1	$\bar{X}$ Ronda-2	p value	Grados de libertad	t	Incremento
Virtual asincrónico	**Periodo académico: mayo-septiembre 2021**						
	Asignatura: Fundamentos de ingeniería de software						
	A1	0.6657	0.7390	0.000	424	-8.2932	7%
	A2	0.6648	0.7653	0.000	424	-10.519	10%
	A3	0.6632	0.7699	0.000	424	-8.9200	11%
	A4	0.6876	0.7760	0.000	424	-11.0617	9%
	Periodo académico: octubre 2021 - febrero 2022						
	Asignatura: Fundamentos de ingeniería de software						
	A1	0.7412	0.8506	0.000	394	-12.1697	11%
	A2	0.7596	0.8166	0.000	388	-6.1209	6%
	A3	0.7795	0.8105	0.032	382	-2.1574	3%
	A4	0.7770	0.8223	0.000	382	-6.4207	5%
	Asignatura: Ingeniería de software						
	A1	0.6958	0.8189	0.000	220	-8.8063	12%
	A2	0.7542	0.8492	0.000	217	-9.4831	10%
	A3	0.7454	0.8505	0.000	214	-9.9298	11%
	A4	0.7266	0.7574	0.021	214	-2.3203	3%

Escenario de educación	Actividad	$\bar{X}$ Ronda-1	$\bar{X}$ Ronda-2	p value	Grados de libertad	t	Incremento
Virtual sincrónico	**Periodo académico: mayo-septiembre 2022**						
	Asignatura: Fundamentos de ingeniería de software						
	A1	0.763	0.854	0.000	292	-9.914	9%
	A2	0.740	0.808	0.000	882	-5.4267	7%
	A3	0.634	0.856	0.000	274	-13.5959	22%
	A4	0.668	0.802	0.000	266	-11.3415	13%
	A5	0.654	0.845	0.000	262	-14.4242	19%
	Asignatura: Fundamentos de ofimática						
	A6	0.673	0.813	0.000	130	-4.9344	14%
	A7	0.702	0.909	0.000	134	-8.3195	21%
	A8	0.747	0.896	0.000	138	-10.0398	15%
Presencial	**Periodo académico: mayo-septiembre 2022**						
	Asignatura: Ingeniería de software						
	A1	0.673	0.835	0.000	66	-22.4951	16%
	A2	0.603	0.767	0.000	66	-9.7093	16%
	A3	0.527	0.865	0.000	66	-16.5351	34%
	A4	0.626	0.784	0.000	66	-28.646	16%
	A5	0.720	0.869	0.000	66	-12.4966	15%
	Asignatura: Gestión de procesos de negocios y sistemas empresariales						
	A9	0.727	0.897	0.000	52	-5.973	17%
	Asignatura: Fundamentos de programación						
	A10	0.765	0.912	0.000	48	-5.5745	15%

A continuación, se detalla por escenario de educación:

Escenario de educación virtual asincrónico

En el escenario de educación virtual asincrónico se realizaron pruebas en 12 actividades en dos rondas. En la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2021, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, exceptuando el grupo 9 que decrece el rendimiento en la segunda ronda de la actividad 2 y 3, además también tiene rendimiento bajo en la primera ronda de la actividad 3 y 4. El historial muestra que no existe retiro de integrantes en ese grupo (Figura **109**). La misma tendencia se identifica en el periodo académico octubre 2021-febrero 2022, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, observando que el grupo 7 en todas las actividades, obtiene la menor puntuación en la primera ronda, y que logra mejorar con la mínima puntuación en la segunda ronda (Figura **110**).

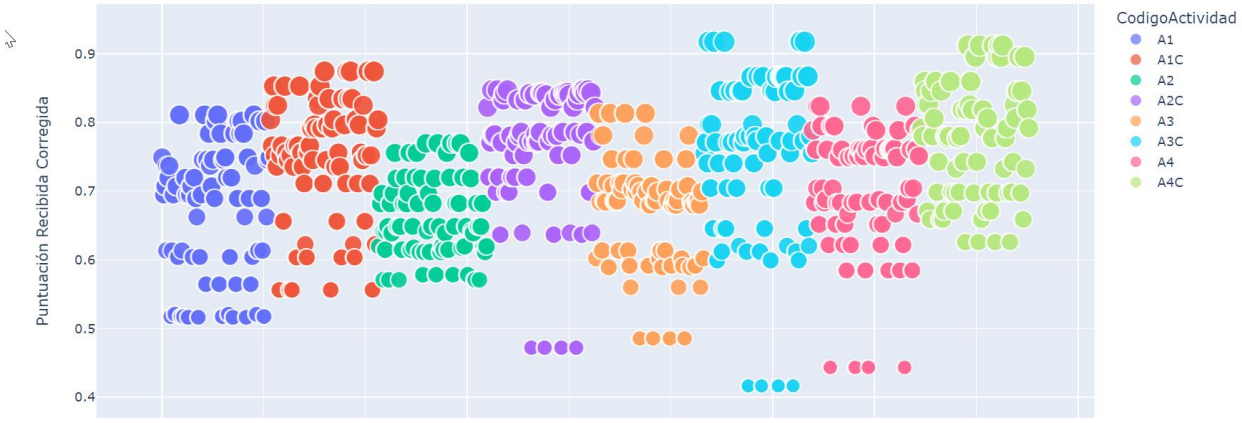


Figura 109. Comparación de primera y segunda ronda por cada actividad de la asignatura de fundamentos de ingeniería en software impartida en escenario de educación virtual asincrónico del periodo académico mayo-septiembre 2021

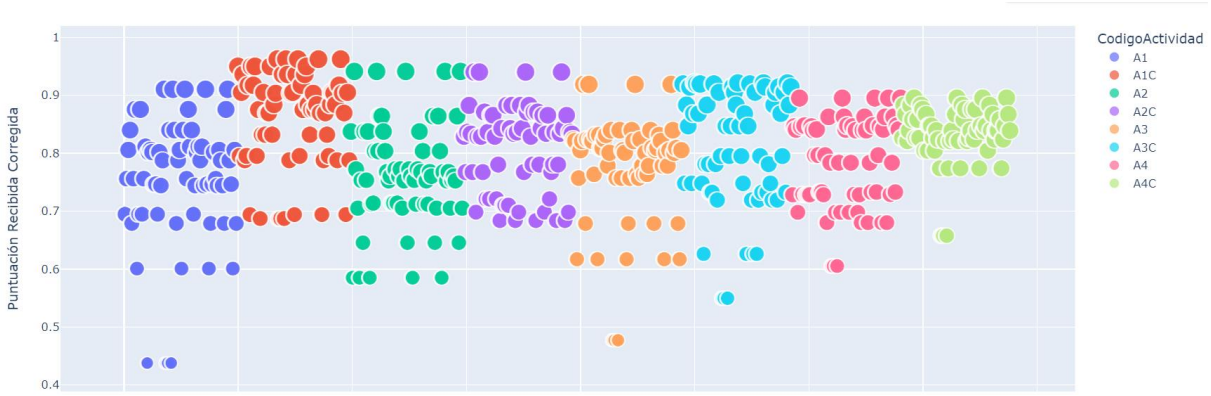


Figura 110. Comparación de primera y segunda ronda por cada actividad de la asignatura de fundamentos de ingeniería en software impartida en escenario de educación virtual asincrónico del periodo académico octubre 2021-febrero 2022

En la asignatura de ingeniería de software del periodo académico octubre 2021-febrero 2022, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, percibiendo que el grupo 8 en la actividad 3 y el grupo 5 en la actividad 4, tienden a bajar el rendimiento en la primera ronda y logran mejorar en la segunda ronda (Figura **111**).

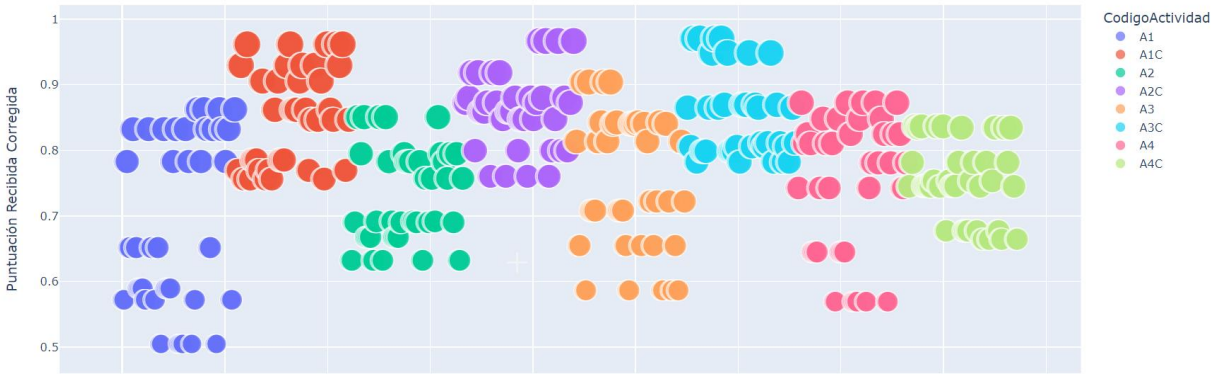


Figura 111. Comparación de primera y segunda ronda por cada actividad de la asignatura de ingeniería en software impartida en escenario de educación virtual asincrónico del periodo académico octubre 2021-febrero 2022

Escenario de educación virtual sincrónico

En el escenario de educación virtual sincrónico se realizaron pruebas en 8 actividades. En la asignatura de fundamentos de ingeniería de software del periodo académico mayo-septiembre 2022, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, exceptuando el grupo 1, que decrece el rendimiento en la segunda ronda de la actividad 3 y 5, además en la primera ronda de la actividad 4 obtienen la menor puntuación, manteniéndola en la segunda ronda, sin mejoras. El historial muestra que a partir de la actividad 3, se retiraron integrantes del grupo, quedando dos, y en la segunda ronda de la actividad 4 quedó un estudiante (Figura **112**).

Figura 112. Comparación de primera y segunda ronda por cada actividad de la asignatura de fundamentos de ingeniería en software impartida en escenario de educación virtual sincrónico del periodo académico mayo-septiembre 2022

En la asignatura de fundamentos de ofimática, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, observando que el grupo 7 en la actividad 1, y el grupo 5 en la actividad 2 y 3, han obtenido la menor puntuación en la primera y logran mejorar en la segunda ronda (Figura **113**).

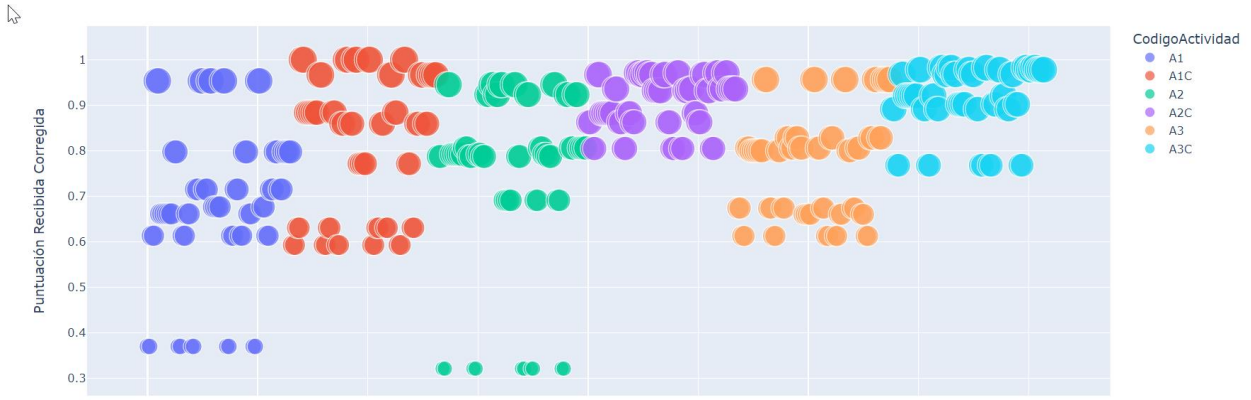


Figura 113. Comparación de primera y segunda ronda por cada actividad de la asignatura de fundamentos de ofimática impartida en escenario de educación virtual sincrónico del periodo académico mayo-septiembre 2022

Educación presencial

En el escenario de educación presencial se realizaron pruebas en 7 actividades. En la asignatura de ingeniería de software del periodo académico mayo-septiembre 2022, los estudiantes en todas las actividades tienden a mejorar el rendimiento en la segunda ronda, notando que el grupo 3 en la actividad 1, 2, 3, 5, y el grupo 2 en la actividad 1, han obtenido la menor puntuación en la primera ronda y logran mejorar en la segunda ronda (Figura **114**).

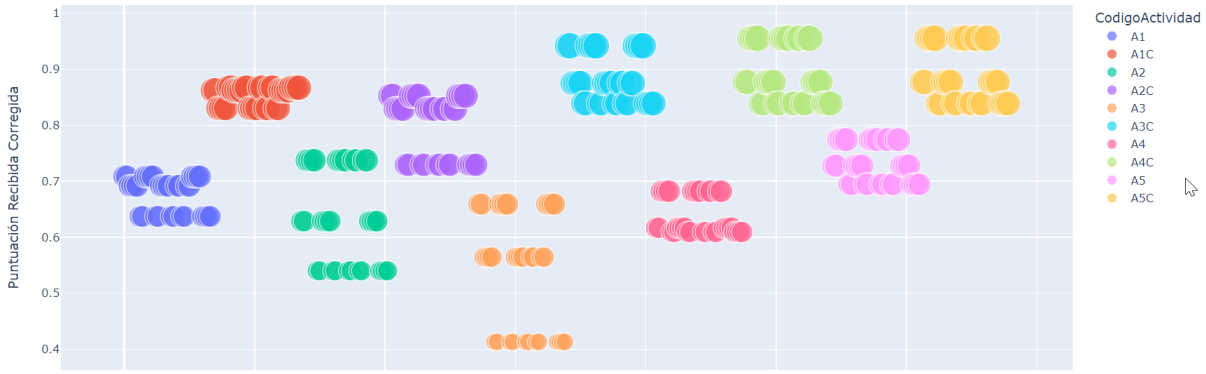


Figura 114. Comparación de primera y segunda ronda por cada actividad de la asignatura de ingeniería de software impartida en escenario de educación presencial del periodo académico mayo-septiembre 2022

En la asignatura de fundamentos de programación, los estudiantes tienden a mejorar el rendimiento en la segunda ronda, notando que el grupo 4 obtuvo la menor puntuación en la primera ronda (media (0.600), mediana (0.708)) y logró mejorar substancialmente en la segunda ronda (media (0.908), mediana (0.950)) (Figura **115**).

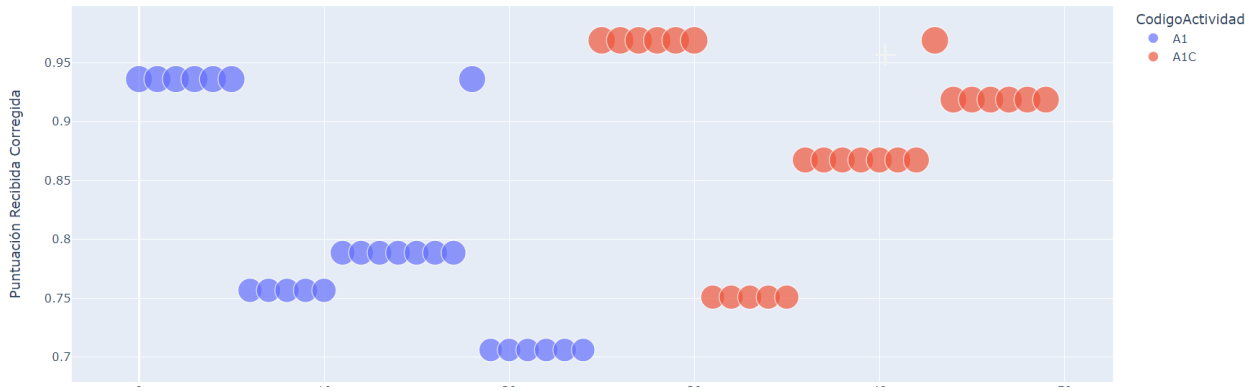


Figura 115. Comparación de primera y segunda ronda por cada actividad de la asignatura de fundamentos de programación impartida en escenario de educación presencial del periodo académico mayo-septiembre 2022

Así mismo continua la tendencia en la asignatura de gestión de procesos de negocios y sistemas empresariales, los estudiantes tienden a mejorar el rendimiento en la segunda ronda, notando que el grupo 5 obtuvo la menor puntuación (media (0.810), mediana (0.900)) en la primera ronda y logró mejorar en la segunda ronda (media (0.880), mediana (0.950)) (Figura **116**).

Figura 116. Comparación de primera y segunda ronda por cada actividad de la asignatura de gestión de procesos de negocios y sistemas empresariales impartida en escenario de educación presencial del periodo académico mayo-septiembre 2022

Una vez realizado el análisis estadístico descriptivo, se concluye que:

- En todas las asignaturas ya sea impartidas en escenario de educación virtual asincrónico, sincrónico y presencial, la tendencia de los estudiantes es mejorar el rendimiento en la segunda ronda. Existe diferencia significativa en las 27 actividades realizadas, entre la puntuación media de la primera y segunda ronda en procesos de evaluación entre pares.

5.2.2.2. Utilidad del modelo de evaluación entre pares basado en análisis de sentimiento

Se consideró establecer cuál es la percepción de los estudiantes respecto al proceso de evaluación entre pares mediante la aplicación de una encuesta adaptada de [222] que se enfoca en tres ejes:

- La percepción general de los estudiantes respecto a la evaluación entre pares.
- Evaluación sobre la intención de la retroalimentación dada.
- Evaluación de la retroalimentación recibida.

La encuesta se estructura en base a los ejes o categorías mencionadas donde cada argumento planteado en la pregunta se cuantifica de acuerdo con una escala de conformidad (Nada, Poco, Medianamente Suficiente, Suficiente y Mucho). Adicionalmente, se incluyeron preguntar de criterio abiertas, que ayudaron a determinar el impacto del modelo relacionado a los ejes planteados.

Las respuestas recibidas serán claves para determinar si el modelo de evaluación entre pares ha sido efectivo y además está generando un impacto en el proceso de enseñanza-aprendizaje.

Participaron un total de 255 estudiantes, 81 estudiantes en la asignatura de fundamentos de ingeniería de software en el periodo académico noviembre 2020-marzo 2021, 76 estudiantes en la asignatura de fundamentos de ingeniería de software en el periodo académico mayo-septiembre 2021, 62 estudiantes en la asignatura de fundamentos de ingeniería de software y 36 estudiantes en la asignatura de ingeniería de software en el periodo académico octubre 2021-febrero 2022 en escenario de educación virtual asincrónica ante la pandemia COVID-19.

El análisis de resultados se detalla a continuación:

Eje 1: Percepción general respecto a la evaluación entre pares

Dentro de esta categoría se plantearon los siguientes cuestionamientos:

- ¿El andamiaje recibido de la actividad ha sido adecuado?
- ¿La retroalimentación ha aumentado su participación en las actividades?
- ¿La frecuencia de la retroalimentación ha sido adecuada?
- ¿La carga de trabajo ha sido aceptable?

Con la información general obtenida se puede ver que los cuatro criterios que conforman la categoría inherente a la percepción de la evaluación entre pares indican una conformidad con la actividad donde se puede observar que los estudiantes están evidenciando que la forma en la que se ha implementado la actividad está siendo la correcta.

Andamiaje

En la Figura **117**, se muestra los resultados del primer argumento que es el andamiaje, donde 149 estudiantes de 255 (59%), manifiestan que su aprehensión respecto al andamiaje durante el proceso de evaluación ha sido el adecuado, y un grupo de 80 (31%), indican que es suficiente, por lo que está siendo aceptado por este sector de estudiantes. En contraste con la apreciación mayoritaria mostrada, existe un grupo de 26 (10%) que reconocen que el andamiaje ha sido medianamente o poco adecuado, por lo que se podría diseñar actividades para mejorarlo.

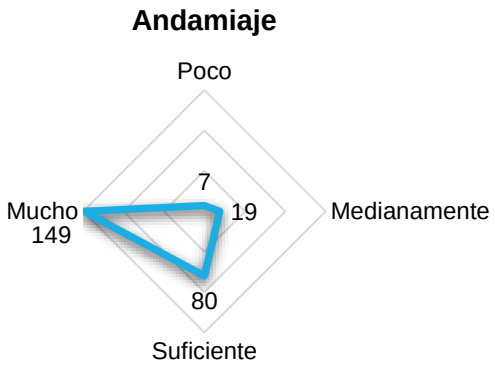


Figura 117. Percepción general de los estudiantes respecto al andamiaje bajo el cual se ejecuta la evaluación entre pares

Participación en actividades

En la Figura **118**, se muestran los resultados si la retroalimentación ha fomentado la participación del estudiante en las actividades, donde 172 estudiantes de 255 (67%) perciben que gracias a la retroalimentación que dan o reciben, han podido ser más activos en lo que corresponde a las actividades propuestas por el docente, y un grupo de 59 (23%), indican que está siendo suficiente, por lo que está siendo aceptada por este grupo de estudiantes. Sin embargo, dos grupos de 15 y 9 (10%) muestra una mediana y poca conformidad respectivamente, por lo que se podría plantear acciones para mejorar en este aspecto.

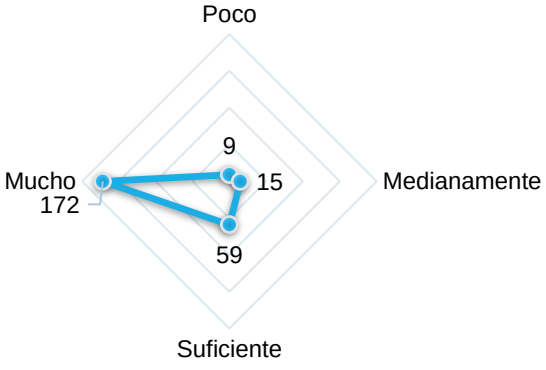


Figura 118. Percepción general de los estudiantes respecto a su participación en retroalimentación entre pares

Frecuencia de retroalimentación

En la Figura **119**, se muestra los resultados obtenidos relacionados a la impresión que tienen los estudiantes respecto a la periodicidad de la retroalimentación en el proceso de evaluación entre pares. Donde 131 estudiantes de 255 (51%) están de acuerdo en que la periodicidad de la retroalimentación es bastante adecuada, otro grupo de 91 estudiantes (36%), también muestran que se encuentran conformes. Sin embargo, también se puede observar dos grupos de 22 y 11 (13%) que reconocen que la frecuencia que se ha tenido es medianamente y poco aceptable respectivamente. Es así que se considera lo indicado por ese pequeño grupo para determinar que oportunidades de mejora podrían ayudar a improvisar el proceso de evaluación entre pares.

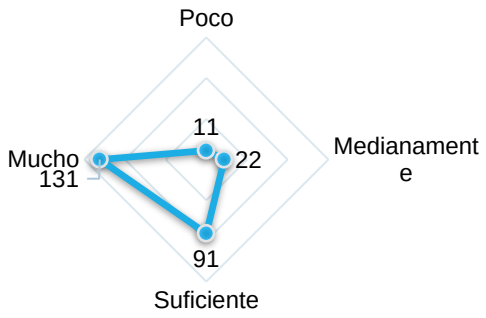


Figura 119. Percepción general de los estudiantes respecto a la frecuencia de retroalimentación en la evaluación entre pares

Carga de trabajo

En la Figura **120**, se observa que en lo que respecta a la carga de trabajo propuesta, 111 estudiantes de 255 (44%) indican que la exigencia de trabajo es muy aceptable, otro grupo de 98 (38%) ha mostrado su conformidad con la carga de trabajo. Mientras que un grupo de 30 (12%) está medianamente conforme con la cantidad de trabajo asignada, y 16 estudiantes (6%) revelan que están poco o nada de acuerdo con los trabajos asignados. Este último grupo, aunque sea pequeño ayuda a vislumbrar que se debe considerar en el plan de acción redefinir la carga de revisión de trabajo óptima para todos los estudiantes, donde no se afecte el modelo de evaluación y el proceso de aprendizaje sea consistente.

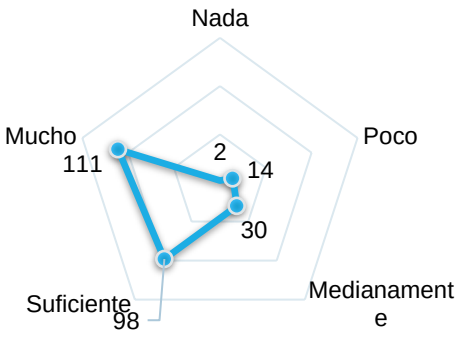


Figura 120. Percepción general de los estudiantes respecto a la carga de trabajo en la evaluación entre pares

Eje 2: Evaluación sobre la intención de la retroalimentación dada

Las preguntas realizadas para esta categoría son las siguientes:

- ¿La retroalimentación que usted proporcionó le ha ayudado a mejorar sus habilidades (críticas, reflexivas……)?
- ¿La retroalimentación que usted proporcionó le ha ayudado a mejorar el proceso de aprendizaje?
- ¿La retroalimentación que usted proporcionó le ha sido útil para mejorar sus trabajos?
- ¿La retroalimentación que usted proporcionó le ha ayudado a incrementar el nivel de implicación en su aprendizaje?
- ¿La retroalimentación que usted proporcionó le ha ayudado a desarrollar el conocimiento del contenido específico de la asignatura?

El objetivo de esta categoría es determinar si la retroalimentación que el estudiante proporciona está siendo clave para su aprendizaje. Es así como, se han seleccionado 5 aspectos relacionados al proceso de aprendizaje contemplados en cada una de las preguntas correspondientes a esta clase que son los siguientes:

- Mejora de habilidades.
- Mejora del proceso de aprendizaje.
- Mejora de los trabajos.
- Mejora de la implicación del aprendizaje.
- Mejora en la adquisición del contenido.

El parámetro utilizado para la comparación entre periodos fue la media, ya que se estableció una escala de conformidad del 1 al 5. Aunque no represente el valor propio de la variable que se está cuantificando, nos puede brindar una noción de lo que puede ocurrir respecto a los argumentos de cada pregunta realizada.

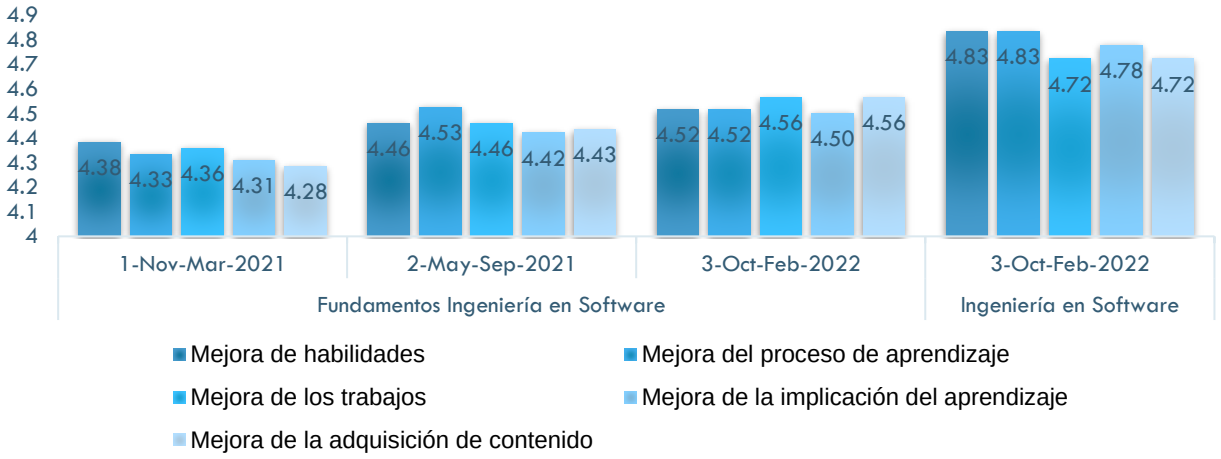


Figura 121. Percepción general de los estudiantes respecto a la retroalimentación dada en la evaluación entre pares

En la Figura **121**, se muestra las medias de los criterios evaluados en las preguntas planteadas por periodo y asignatura. En la asignatura de fundamentos de ingeniería de software de la carrera de sistema de información impartida en escenarios de educación virtual asincrónico, los resultados de la encuesta realizada a los estudiantes muestran una tendencia incremental en la percepción de la mejora de los criterios seleccionados desde el primer hasta el último periodo de evaluación. Mientras que, en la asignatura de ingeniería en software de la carrera de tecnologías de la información impartida en escenario virtual asincrónico, en el periodo octubre 2021-febrero 2022, los estudiantes denotan una mejora sustancial en cada uno de los criterios relacionados al proceso de aprendizaje gracias a la retroalimentación que han proporcionado, tal como se detalla en la Tabla **86**. Es evidente que las medias de los puntajes otorgados por el grupo de estudiantes a cada uno de los criterios de mejora son los más altos en comparación con los periodos previos a la asignatura de fundamentos de ingeniería en software.

Tabla 86. Valores medio de los criterios de mejora en las asignaturas y periodos evaluados de la retroalimentación dada

Periodo \ Criterio	Mejora de habilidades	Mejora del proceso de aprendizaje	Mejora de los trabajos	Mejora de la implicación del aprendizaje	Mejora de la adquisición de contenido
Fundamentos Ingeniería en Software					
Nov-Mar-2021	4.38	4.33	4.36	4.31	4.28
May-Sep-2021	4.46	4.53	4.46	4.42	4.43
Oct-Feb-2022	4.52	4.52	4.56	4.50	4.56
Ingeniería en Software					
Oct-Feb-2022	4.83	4.83	4.72	4.78	4.72

También se realizó un análisis holístico, considerando las opiniones de los 255 estudiantes independientemente del periodo y carrera, lo que permite tener una visión global de la percepción de estos relacionada a los criterios de mejora considerados en cada pregunta realizada.

Mejora de habilidades

En la Figura **122**, se muestra que el 61% de los estudiantes pertenecientes a las dos carreras bajo estudio indican que han mejorado sus competencias notablemente, y el 31% han podido experimentar una suficiente mejora en sus habilidades, de tal forma que se ha satisfecho lo esperado por estos estudiantes después de proporcionar su retroalimentación. Sin embargo, el 5% indica que sus expectativas respecto a la mejora de competencias han sido medianamente suficiente, y el 3% indica que han experimentado una leve mejora, por lo que se podría establecer acciones para afinar las habilidades al proporcionar retroalimentación.

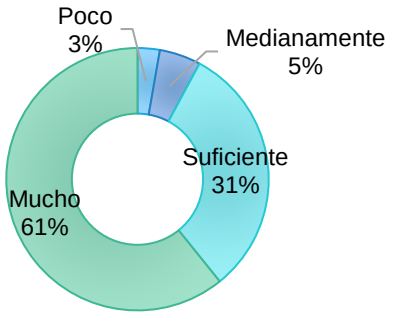


Figura 122. Percepción de mejora de habilidades al proporcionar retroalimentación

Mejora del proceso de aprendizaje

En la Figura **123**, se muestra que el 61% de los estudiantes están de acuerdo en que la retroalimentación que proporcionaron ha sido clave para mejorar el proceso de aprendizaje considerablemente, y el 32% indica que han mejorado sustancialmente. En contraste, el 5% manifiesta que ha mejorado medianamente, y el 2% indica que el aporte de su retroalimentación no ha sido consistente con la mejora del proceso de aprendizaje, por lo que se podría instituir acciones para que los estudiantes trabajen en equipo.

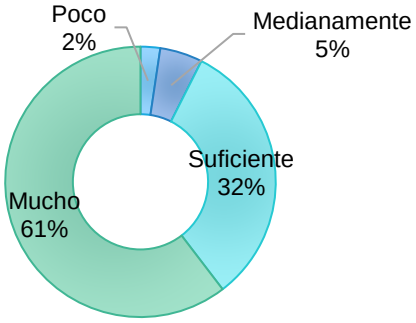


Figura 123. Percepción de mejora del proceso de aprendizaje al proporcionar retroalimentación

Mejora de los trabajos

En la Figura **124**, se muestra que el 61% de los estudiantes indica de forma favorable que los trabajos han mejorado después de aportar la retroalimentación, por lo que se considera como un aspecto positivo, y el 30% indica que los trabajos han mejorado sustancialmente, por lo que están conformes con el accionar de la metodología de evaluación. Sin embargo, al 6% indica que la retroalimentación contribuida ha ayudado a la mejora de los trabajos, pero no en la medida esperada, y el 3% señalan que la mejora en los trabajos no ha sido considerable, por lo que se podría instaurar acciones para que los estudiantes proporcionen retroalimentación basada en el contenido de la actividad.

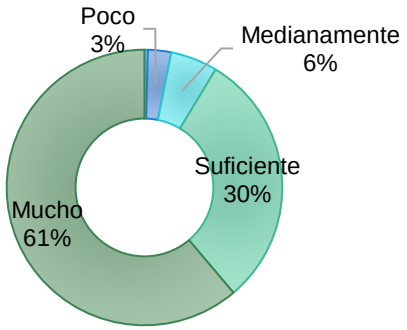


Figura 124. Percepción de mejora de los trabajos al proporcionar la retroalimentación

Mejora de la implicación del aprendizaje

En la Figura **125**, muestra que el 57% de estudiantes indican que se han podido involucrar más a lo que es el aprendizaje gracias a la retroalimentación proporcionada, y el 33% señala que la implicación hacia el aprendizaje ha mejorado. Sin embargo, el 7% revela que la mejora en este aspecto está siendo medianamente suficiente por lo que puede haber un aspecto que no esté motivando a los estudiantes a involucrarse totalmente a lo que es el aprendizaje, y el 3% advirtieron poca mejora en la implicación al aprendizaje, por lo que se podría instruir acciones para que los estudiantes proporcionen retroalimentación basada la temática.

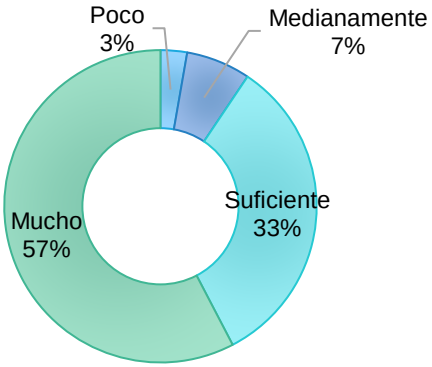


Figura 125. Percepción de mejora de la implicación del aprendizaje al proporcionar retroalimentación

Mejora en la adquisición del contenido

En la Figura **126**, se muestra que el 66% de los estudiantes ha experimentado una mejora en el provecho del contenido asociado a la retroalimentación que proporcionan y que ha sido en cierta forma superlativa, el 27% señalan que la mejora en la adquisición del contenido ha sido pasadera y que se encuentra dentro de sus expectativas como estudiantes. Mientras que el 6% revela que la mejora en este aspecto ha sido medianamente suficiente y el 1% sugiere que la mejora ha sido poca, por lo que se podría implantar acciones para que los estudiantes revisen el contenido de la actividad.

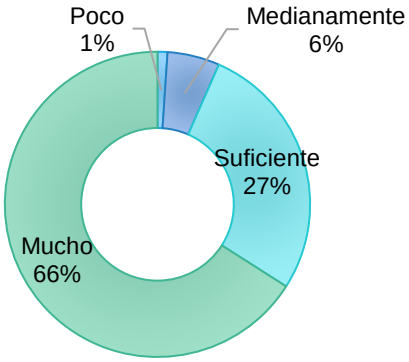


Figura 126. Percepción de mejora de la adquisición de contenido al proporcionar retroalimentación

Eje 3: Evaluación sobre la intención de la retroalimentación recibida

Las preguntas realizadas para esta categoría son las siguientes:

- ¿La retroalimentación recibida le ha ayudado a mejorar sus competencias (críticas, reflexivas……)?
- ¿La retroalimentación recibida le ha ayudado a mejorar el proceso de aprendizaje?
- ¿La retroalimentación recibida le ha sido útil para mejorar sus trabajos?
- ¿La retroalimentación recibida le ha ayudado a incrementar el nivel de implicación en su aprendizaje?
- ¿La retroalimentación recibida le ha ayudado a desarrollar el conocimiento del contenido específico de la asignatura?

El objetivo de esta categoría es determinar si la retroalimentación que el estudiante recibe está siendo clave para su desarrollo integral. Es así como, se han seleccionado 5 aspectos relacionados al proceso de aprendizaje contemplados en cada una de las preguntas correspondientes a esta clase que son los siguientes:

- Mejora de habilidades.
- Mejora del proceso de aprendizaje.
- Mejora de los trabajos.
- Mejora de la implicación del aprendizaje.
- Mejora en la adquisición del contenido.

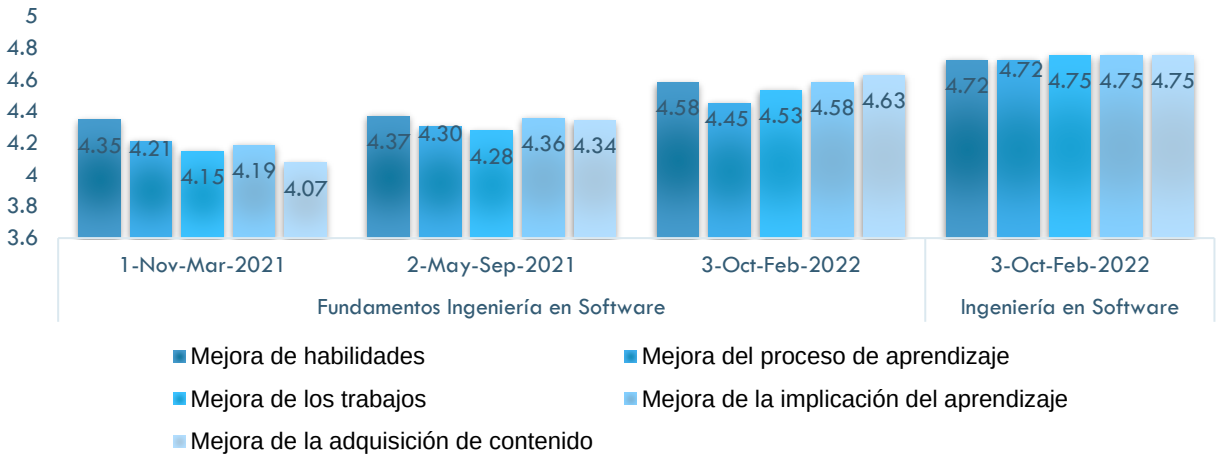


Figura 127. Percepción general de los estudiantes respecto a la retroalimentación recibida en la evaluación entre pares

En la Figura **127**, se muestra en la asignatura de fundamentos de ingeniería en software de la carrera de sistemas de información impartida en escenario virtual asincrónico, una tendencia a la mejora en los aspectos del aprendizaje relacionados a la retroalimentación recibida, la cual se incrementa cronológicamente desde el primer periodo de implementación de la evaluación entre pares hasta el último periodo evaluado. Una posible teoría de la evidente mejora es que los estudiantes se han ido familiarizando más con la metodología de evaluación, de tal forma que la fracción de estudiantes que va a evidenciar mejoras absolutas tiende a ir incrementando. Mientras que, en el periodo octubre 2021-febrero 2022 en la asignatura de ingeniería en software de la carrera de tecnología de la información impartida en escenario virtual asincrónico, los estudiantes

revelan una mejora sustancial en los criterios relacionados al proceso de aprendizaje respecto a los periodos previos, pero que corresponden a otra asignatura (Tabla **87**).

Tabla 87. Valores medio de los criterios de mejora en las asignaturas y periodos evaluados de la retroalimentación recibida

Periodo \ Criterio	**Mejora de habilidades**	**Mejora del proceso de aprendizaje**	**Mejora de los trabajos**	**Mejora de la implicación del aprendizaje**	**Mejora de la adquisición de contenido**
Fundamentos Ingeniería en Software					
Nov-Mar-2021	4.35	4.21	4.15	4.19	4.07
May-Sep-2021	4.37	4.30	4.28	4.36	4.34
Oct-Feb-2022	4.58	4.45	4.53	4.58	4.63
Ingeniería en Software					
Oct-Feb-2022	4.72	4.72	4.75	4.75	4.75

Para continuar con el análisis, se realizó un estudio holístico, considerando las opiniones de los 255 estudiantes independientemente del periodo y carrera, lo que nos permite tener una visión global de la percepción de estos relacionada a los criterios de mejora considerados en cada pregunta, todo asociado a la retroalimentación recibida.

Mejora de habilidades

En la Figura **128**, se puede distinguir que el 61% de los estudiantes pertenecientes a las dos carreras bajo estudio indican que la retroalimentación recibida les ha ayudado a mejorar sus competencias notablemente concluyendo que la metodología está permitiendo un desarrollo integral. Por otra parte, el 29% han podido experimentar una mejora en sus habilidades indicando que el aporte de la retroalimentación recibida ha sido suficiente para este grupo de estudiantes. Se puede observar también que el 7% indica que sus expectativas respecto a la mejora de competencias han sido medianamente suficientes lo que implica que ciertas competencias no se están desarrollando plenamente. Existe también la percepción del 2% que indican que han experimentado una leve mejora en sus competencias, finalmente, un porcentaje del 1% revela que no han tenido mejoras en sus competencias. Es un grupo minúsculo, pero se podría considerar para determinar posibles ajustes a la metodología de evaluación que beneficie a todos los estudiantes.

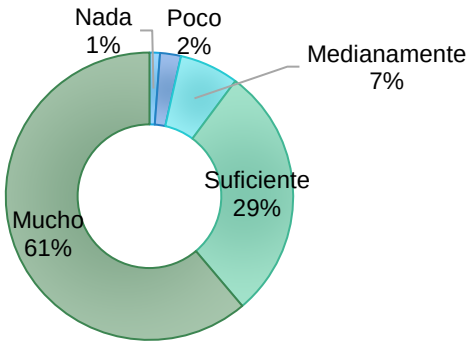


Figura 128. Percepción de la mejora de habilidades al recibir retroalimentación

Mejora del proceso de aprendizaje

En la Figura **129**, se puede distinguir que el 55% de los estudiantes están de acuerdo en que la retroalimentación que reciben ha sido clave para mejorar el proceso de aprendizaje considerablemente, fortaleciendo el conocimiento adquirido. En seguida, se nota que el 33% dejan ver que el proceso de aprendizaje ha mejorado sustancialmente gracias a la retroalimentación que se recibe y que se están cumpliendo expectativas respecto al modelo de evaluación, otro punto es el grupo que conforma el 7% el cual manifiesta que la retroalimentación recibida ha mejorado medianamente sus habilidades, por lo que sus expectativas no están siendo cubiertas totalmente. Sin embargo, el 5% de los estudiantes indica que la retroalimentación recibida no ha sido sólida en la mejora del proceso de aprendizaje y el 1% que no ha evidenciado mejora alguna, por lo que se podría diseñar acciones para que los estudiantes trabajen colaborativamente.

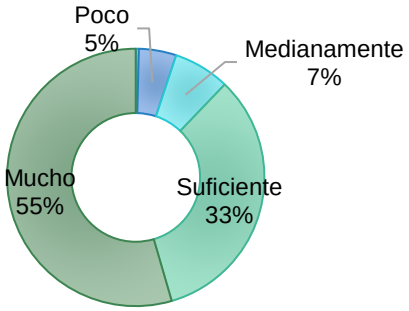


Figura 129. Percepción de la mejora del proceso de aprendizaje al recibir retroalimentación

Mejora de los trabajos

En la Figura **130**, se muestra que el 57% de los estudiantes indica que los trabajos han mejorado después de recibir la retroalimentación propiciamente, y el 29% indica que la retroalimentación recibida ha impactado satisfactoriamente en que los trabajos mejoren por lo que están conformes con el maniobrar de la metodología de evaluación. No obstante, el 9% indica que la retroalimentación recibida de sus pares ha ayudado a la mejora de los trabajos, pero no en la medida esperada, el 4% señalan que la mejora en los trabajos no ha sido considerable, y el 1% indica que no hay mejora, por lo que se podría instaurar acciones para que los estudiantes reciban retroalimentación basada en el contenido de la actividad.

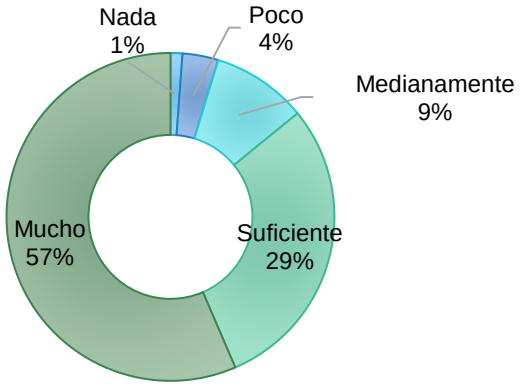


Figura 130. Percepción de la mejora de los trabajos al recibir retroalimentación

Mejora de la implicación del aprendizaje

En la Figura **131**, se muestra que el 57% de estudiantes indican que se han podido involucrar más en lo correspondiente al aprendizaje gracias a la retroalimentación recibida, y el 32% destacan que la implicación hacia el aprendizaje ha mejorado sustancialmente. Sin embargo, el 7% manifestaron que han mejorado medianamente, el 3% señalaron poca mejora en la implicación al aprendizaje, y un grupo minúsculo de equivalente al 1% indican que no se han podido involucrar de ninguna manera al proceso de aprendizaje, por lo que se podría instruir acciones para que los estudiantes reciban retroalimentación basada la temática.

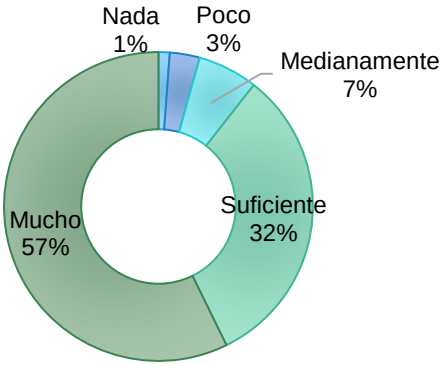


Figura 131. Percepción de la mejora de la implicación del aprendizaje al recibir retroalimentación

Mejora en la adquisición del contenido

En la Figura **132**, se muestra que el 56% de los estudiantes ha experimentado un beneficio respecto a la adquisición del contenido, deduciendo en que la retroalimentación recibida ha sido relevante, el 31% señalan que la mejora en la adquisición del contenido ha sido suficiente y que cumple sus expectativas. Mientras que el 8% indican que la mejora en este aspecto ha sido medianamente suficiente, el 3% y 2% sugiere que la mejora ha sido poca o nula respecto a la retroacción recibida respectivamente, por lo que se podría implantar acciones para que los estudiantes revisen el contenido de la actividad.

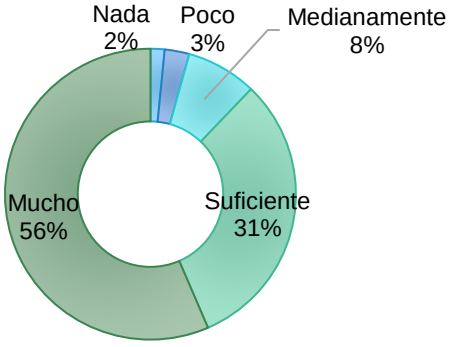


Figura 132. Percepción de la mejora de la adquisición de contenido al recibir retroalimentación

En la Tabla **88** y Figura **133**, se presenta un resumen de las (Tabla **86** y Tabla **87**) de la percepción general de los estudiantes sobre la retroalimentación dada y recibida.

Tabla 88. Percepción general de los estudiantes sobre la retroalimentación dada y recibida

Criterio	Dada	Recibida	Promedio
Mejora de habilidades	4.50	4.46	4.48
Mejora del proceso de aprendizaje	4.51	4.37	4.44
Mejora de los trabajos	4.49	4.36	4.43
Mejora de la implicación del aprendizaje	4.45	4.41	4.43
Mejora de la adquisición de contenido	4.46	4.38	4.42

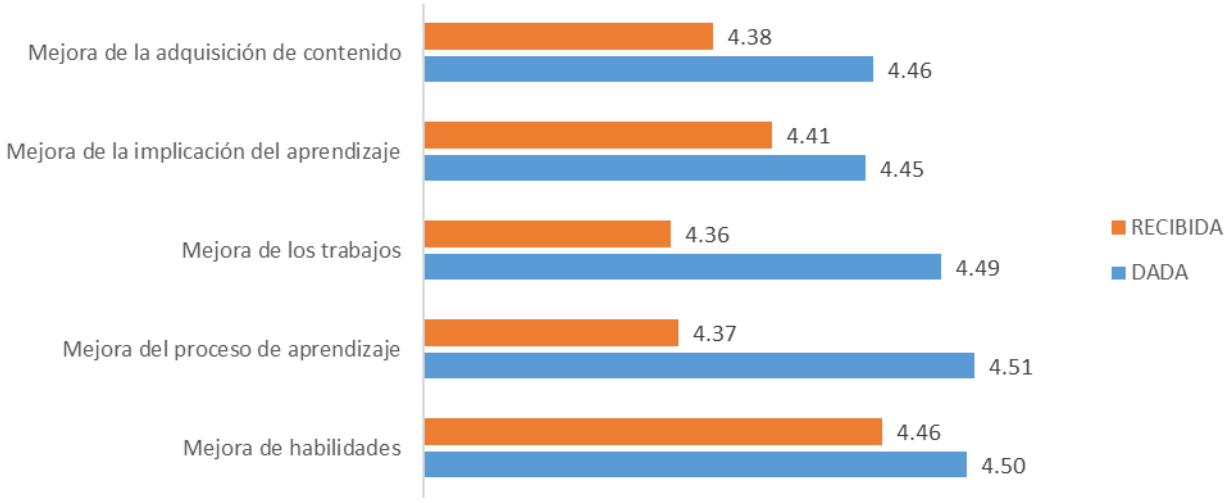


Figura 133. Puntajes promedios sobre la retroalimentación dada y recibida

Evaluación del impacto de la evaluación entre pares

Las preguntas abiertas realizadas para esta categoría son las siguientes:

- ¿La evaluación entre pares ha sido útil?
- ¿La retroalimentación que proporcioné ha sido útil?
- ¿La retroalimentación recibida ha sido útil?

Los argumentos planteados en cada pregunta corresponden a la utilidad que está teniendo la metodología de evaluación entre pares y si la retroalimentación es beneficiosa. De esta manera, los comentarios que los estudiantes proporcionen se generará la puntuación mediante el modelo de análisis de sentimiento que puede tomar 3 valores; +1 (Positivo) si el comentario reafirma la pregunta, 0 (Neutral) si el comentario es neutral y -1 (Negativo) si el comentario refuta el argumento de la pregunta.

Utilidad de la evaluación entre pares

Se seleccionó respuestas aleatorias de los estudiantes donde se ha de evidenciar la categoría asignada en base a su comentario, con el fin de ilustrar mediante un ejemplo real obtenido durante la recolección de los datos. Los ejemplos se presentan a continuación:

Argumento Positivo (+1)

La evaluación entre pares nos ayudó mucho a darnos cuentas de nuestros errores, y a saber cómo corregirlos de manera adecuada, la mejor parte es que se puede aprender mucho comentando y viendo el diagrama de tus compañeros, y para realizar un buen comentario antes se tuvo que empaparse del tema.

Neutro (0)

Sí, fue útil, pero como se ha dicho anteriormente, el tiempo en el que se hizo todo fue largo y se fue volviendo cansado realizar el proceso.

Negativo (-1)

La evaluación entre pares no ha sido de tanta utilidad ya que no es lo mismo tener la evaluación de alguien experto que nos pueda ayudar a mejorar y saber en que realmente estamos correctos y mejorar de esta manera, a diferencia de que, aunque somos estudiantes no poseemos el conocimiento total del tema y calificar algo a nuestro parecer que creemos que es correcto es complicado y confuso para quienes evaluamos.

En la Figura **134**, se muestra que el 88% de los estudiantes está evidenciando un impacto positivo de la evaluación, lo que nos permite inferir que para este grupo la evaluación entre pares está siendo bastante útil. Por otra parte, se puede distinguir que el 6% indican que el impacto ha sido neutral en algunos aspectos. No obstante, el 6% restante menciona que la actividad no está siendo provechosa, y que, a su criterio está teniendo un impacto negativo.

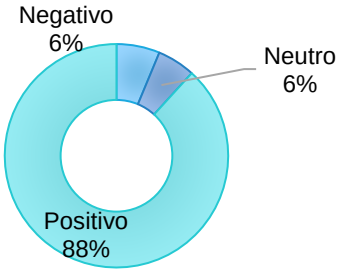


Figura 134. Percepción de la utilidad de evaluación entre pares

Utilidad de la retroalimentación proporcionada

Se seleccionó aleatoriamente varias de las respuestas proporcionadas por los estudiantes, relacionadas con la pregunta abierta donde se evalúa la calidad de la retroalimentación dada. De la misma forma se indica la categoría en base al argumento expuesto. Los ejemplos se presentan a continuación:

Positivo (+1)

La retroalimentación que proporcione me ayudó a mejorar mi reflexión y aprender de los errores de los demás para aplicar correcciones a mi trabajo.

Neutro (0)

Ayuda a enfocarse en la materia y sus distintos contenidos y comprensión de los temas, pero era cansado

Negativo (-1)

Me resultaba difícil dar retroalimentación a un grupo porque no se entendía el diagrama y me absorbía mucho tiempo.

En la Figura **135**, se muestra de forma general que la utilidad ha sido positiva, donde el 93% de los estudiantes están de acuerdo que han evidenciado mejoras en los aspectos relacionados al aprendizaje al proporcionar retroalimentación. Sin embargo, se puede divisar que el 6% difícilmente puede descifrar la utilidad de la retroalimentación que están proporcionando, y que el 1% conciertan que al dar retroalimentación no está teniendo un impacto positivo, por lo que no está siendo útil completamente para este grupo de estudiantes.

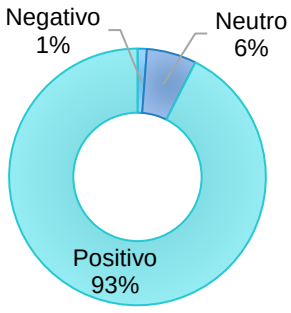


Figura 135. Percepción de la utilidad de la retroalimentación dada

Utilidad de la retroalimentación recibida

Se tomaron ejemplos aleatoriamente de las respuestas que se han categorizado en base al criterio expuesto en las mismas y que responden al argumento planteado en la pregunta abierta. Los ejemplos se presentan a continuación:

Positivo (+1)

La retroalimentación recibida ha sido útil para aceptar críticas constructivas y conocer desde el punto de vista de mis compañeros, en lo que podría mejorar en el trabajo.

Neutro (0)

La retroalimentación recibida ha sido parcialmente útil, pudiendo mejorar y aprender de los errores cometidos en el trabajo, reforzar los conocimientos adquiridos durante el semestre. Pero, por otra parte, ciertos comentarios no eran críticos, ni tenían argumentos relacionado con los criterios de evaluación.

Negativo (-1)

Sinceramente no, ya que muchos solo generalizaban comentarios y no especificaron, esto puedo entenderlo ya que muchos parámetros la gente no conocía dichos términos o tenían ausencia de contenidos, pero esto más que alentarlos a tratar de comprender el trabajo, hacían repetir las mismas frases y hasta veces escribir incoherencias que no ayudaban a corregir los errores

En la Figura **136**, se muestra que el 86% de los estudiantes han podido notar que el impacto la retroalimentación recibida está siendo beneficioso. Mientras que el 12% argumentan aspectos positivos/negativos respecto a la utilidad de la retroalimentación, y el 2% ha podido advertir un impacto negativo de la retroalimentación recibida, por lo que se puede establecer que no ha sido en concreto útil para este grupo.

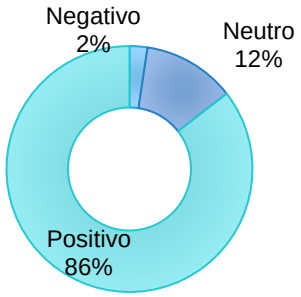


Figura 136. Percepción de la utilidad de la retroalimentación recibida

Una vez realizado el análisis estadístico descriptivo, se diseñó el plan de acción:

Plan de acción propuesto

Como se ha podido evidenciar a lo largo del análisis, se han identificado diferentes inconvenientes con el fin de establecer un plan de acción que permita mitigar los efectos causados e instituir que el beneficio de la metodología de evaluación entre pares sea indiscutible para todos los estudiantes en el futuro.

En Tabla **89**, se presentan las actividades que se proponen para mejorar el proceso de evaluación entre pares donde intervienen el docente y los estudiantes.

Tabla 89. Actividades propuestas en el plan de acción con el objetivo de mejorar la metodología de evaluación entre pares

Actividad	Metas	Responsable	Tiempo	Factores
Presentación de material audiovisual y capacitación respecto a la metodología de evaluación.	Proveer un andamiaje más adecuado.	Docente / Estudiantes.	Primer mes del periodo académico mayo-septiembre 2022.	El objetivo debe ser reducir la reincidencia de los errores cometidos durante la evaluación para que los recursos empleados para realizarla sean claramente entendidos por los estudiantes.
Mostrar los beneficios de una evaluación entre pares realizada correctamente.	Fomentar participación estudiantil.	Docente / Estudiantes.	Por cada ronda.	Sin duda, un grupo de estudiantes no es aún consciente de los beneficios que les trae la evaluación por lo que una iniciativa aplicable sería mostrar los beneficios de una evaluación correctamente ejecutada para asegurar una mejor participación.
Se asignarán actividades no consecutivas con 2 rondas.	Periodicidad de la retroalimentación.	Docente / Estudiantes.	Previo a cada actividad.	Al recibir comentarios de que la periodicidad era tediosa, en muchos casos incomoda se ha decidido ajustar la periodicidad en no ejecutar al mismo tiempo varias actividades de evaluación entre pares en un mismo curso.
Se asignarán uno o dos trabajos para la revisión.	Carga de trabajo adecuada.	Estudiantes.	Por cada actividad.	Se plantea disminuir la carga de trabajo debido a los comentarios que indican que el proceso se hace más tedioso por la cantidad de actividades y la evaluación como tal.
Comprometer a los estudiantes a evaluar con criterio reflexivo, critico basándose en lo aprendido y de forma imparcial, para evitar conflictos entre estudiantes.	Mejora de habilidades de reflexión, criticidad.	Docente / Estudiantes.	Por cada actividad.	La mejora de habilidades es una función de la manera correcta en la que se proporciona la retroalimentación. En muchos casos, la calificación proporcionada no refleja la realidad de la actividad.

Actividad	Metas	Responsable	Tiempo	Factores
Fomentar el trabajo colaborativo y cooperativo en cada actividad.	Mejora del proceso de aprendizaje.	Estudiantes.	Por cada actividad.	En cierto punto la falla en la evaluación es porque no todos los integrantes del grupo han revisado completamente el tema. Por lo que, si se trabaja en equipo se incrementa las probabilidades de hacer un mejor trabajo.
Fomentar en los estudiantes que proporcionen una retroalimentación útil detallando el error sustentando con los contenidos de la actividad.	Mejora de los trabajos.	Docente / Estudiantes.	Por cada actividad.	Se plantea mejorar la comprensión de los temas para evitar que los estudiantes proporcionen argumentos muy simples, ambiguos y nada asociados al tema en revisión y así corrijan los trabajos en base a retroalimentaciones fundamentadas con el contenido.
Explicar claramente a los estudiantes los criterios de la metodología de evaluación para que puedan entregar una retroalimentación basada en temática.	Mejora de la implicación del aprendizaje.	Estudiantes.	Por cada ronda.	Una retroalimentación carente de contenido es lo que se pretende evitar, por lo que se fomentará que el estudiante tenga el interés en revisar el error y comprender de mejor forma el tema que les ayudará a la corrección.
Establecer clases de forma virtual sincrónico o presencial que permitan a los estudiantes hacer preguntas y puedan clarificar dudas respecto a los contenidos de la asignatura.	Mejora en la adquisición del contenido.	Docente / Estudiantes.	Por cada actividad.	De acuerdo con los indicado por un número de estudiantes no se han podido solventar dudas, debido a que han atendido la clase de forma virtual asincrónica.

Una vez ejecutado el plan de acción de concluye que:

- La capacitación en el andamiaje, la disminución de asignación de trabajos, ayudo a mejorar la validez del modelo, puesto que la correlación entre la puntación que recibe el estudiante del colectivo con la puntuación que proporciona el docente en el periodo académico mayo-septiembre 2022 obtuvo similaridad directa sustancial en las asignaturas impartidas en escenario de educación virtual sincrónico (Tabla **79**), y directa fuerte en la asignatura impartida en escenario de educación presencial (Tabla **81**). Mientras que, en periodos académicos anteriores en asignaturas impartidas en escenario de educación virtual asincrónico, la similaridad fue directa débil y moderada (Tabla **77**).
- El hecho de que las puntuaciones de los pares a partir de las pruebas con escenarios virtual sincrónico y presencial, arrojaron mejores coeficientes de validez puede explicarse porque se dotó a los estudiantes del conocimiento fundamental para comprender los criterios de evaluación y las habilidades críticas de evaluación en tiempo real para que

brinden retroalimentación acertada, y se disminuyó a 1 o 2 asignaciones de revisión de tarea.

- Por lo tanto, se deduce que el modelo de evaluación entre pares se optimizó con el plan de acción, y podría ser una herramienta útil a la labor del docente, en contribuir a mejorar el proceso de enseñanza-aprendizaje, puesto que la retroalimentación dada y recibida, ayudo a mejorar: habilidades de reflexión, criticidad, proceso de aprendizaje, trabajos, implicación del aprendizaje y adquisición del contenido.

6. DISCUSIÓN DE RESULTADOS, CONCLUSIONES, Y LÍNEAS FUTURAS

A lo largo de los diferentes capítulos que forman este trabajo se ha ido elaborando y modelando el proceso de evaluación entre pares. El punto inicial, sin duda, ha sido el estudio y compresión de conceptos tan importantes como son el contexto de evaluación entre pares basada en retroalimentación, análisis de sentimiento, y lógica difusa entre otros. A partir de la definición de estos conceptos, se puede observar que en los diferentes capítulos se ha ido produciendo como resultado un modelo de evaluación entre pares con técnicas computacionales.

En sucesivos apartados, se va a concretar la discusión de los resultados de las diferentes preguntas de investigación planteadas, las conclusiones de los diferentes objetivos propuestos y finalmente, tras comentar las limitaciones y las futuras líneas de investigación, se van a enumerar las publicaciones que se han realizado acerca de este desarrollo.

6.1. Discusión de resultados

En este apartado se analizan los resultados obtenidos y como estos son concordantes con otros estudios, así como los aportes que se han generado desde esta investigación, en tal sentido se presentan los cuestionamientos que se estudiaron.

¿Como aplicar evaluación entre pares cuantitativa, cualitativa, inversa, en dos rondas y calibrada en escenarios de educación superior?

En esta investigación se aplicó evaluación entre pares armonizando varias evaluaciones: cualitativa, cuantitativa, inversa, en dos rondas y calibrada en la Universidad Técnica de Manabí (Ecuador), con el propósito de que el estudiante desarrolle habilidades cognitivas al dar y recibir retroalimentación, y en base a las retroalimentaciones de la primera ronda mejore los trabajos, y el rendimiento en la segunda ronda en una temática específica, y que el docente se libere de la carga de evaluación de trabajos abiertos y de revisión de la calidad de evaluación de los pares evaluadores. En contraste con otros estudios que solo aplicaron uno o dos tipos de evaluaciones, como por ejemplo, [78], [304]–[307] aplicaron evaluación formativa para mantener a los estudiantes enfocados principalmente en el contenido de la actividad, y en el beneficio de dar y recibir retroalimentación. Y otros autores [18], [19] emplearon evaluación sumativa y formativa.

La evaluación entre pares en los últimos años ha crecido en modelos de aprendizaje automático sobre características de retroalimentación [16]–[21], [280] y en modelos difusos sobre

la imprecisión inherente a la evaluación [22], [23], [26], [27]. Sin embargo, la aplicación conjunta de ambas técnicas de computación blanda ha sido poco explorada.

La literatura no reveló investigaciones sobre evaluación entre pares utilizando técnicas de computación blanda en idioma español; por tanto, se diseñó un modelo híbrido (ver Sección 3) con la técnica de aprendizaje automático para obtener la predicción de sentimiento de la retroalimentación textual (evaluación cualitativa), y la técnica de lógica difusa para correlacionar la puntuación numérica (evaluación cuantitativa) y retroalimentación textual (evaluación cualitativa), en dos rondas y finalmente calibrar la puntuación de evaluación de tarea considerando el rendimiento y índice (rating) de confianza de los evaluadores (evaluación inversa) (Figura **17**).

El entorno del prototipo (ver Subsección 5.1) que se implementó para la recolección de datos de evaluación entre pares permitió incorporar rúbricas divididas en criterios con aspectos específicos de cualquier temática para que el estudiante realice una evaluación constructiva de la tarea, y de la calidad de la evaluación de los pares con puntuación numérica y retroalimentación textual, apoyándose de un texto instructivo con ejemplos y videos de cómo se debe formular y justificar la retroalimentación mediante argumentos razonadas considerando el contenido de la temática a evaluar. Similar a otros estudios que exteriorizaron que el diseño de la rúbrica es posiblemente el aspecto más importante de una actividad de revisión por pares, ya que el docente decide qué comentarios serán útiles para los autores, así como la naturaleza de las habilidades críticas que se fomentarán en el revisor [308]. El uso de preguntas en la rúbrica hizo que los estudiantes se familiarizaran con criterios y estándares de evaluación claros que les permitieron participar en procesos de aprendizaje entre pares, y que sugieren apoyar a los estudiantes usando tutoriales, listas de verificación, indicaciones de preguntas, guiones o plantillas [309]. Para explicar las respuestas a las preguntas de la rúbrica, los estudiantes deben revisar cuidadosamente los contenidos de la temática para ver si cumplen con los criterios específicos o no, y luego, mediante un profundo proceso cognitivo, articular el porqué de sus respuestas [99]. Las sesiones de andamiaje son esenciales para fomentar la colaboración, crear un entorno de aprendizaje positivo y fomentando la internalización del proceso. Se debe hacer explícito a los estudiantes que el esfuerzo en el rol de revisión conduce a una mejor comprensión y también a una retroalimentación de mejor calidad para todos [310].

La integración de la retroalimentación textual por cada criterio en el proceso de evaluación ayudó a promover habilidades de pensamiento crítico, mismas que fueron analizadas mediante

técnicas computacionales. Varios autores [221], [254], [264], [274] coinciden en que algunos enfoques de NLP funcionan para el análisis de sentimiento aplicado a oraciones cortas con excelentes resultados.

El conjunto de datos se creó en idioma español debido a que no existe un corpus de educación sobre evaluación entre pares en este idioma. La mayoría de los métodos propuestos de evaluación entre pares se centran principalmente en el idioma inglés [16]–[21]. Del mismo modo, otras propuestas de evaluación de contextos educativos también se han centrado en analizar sentimiento en idioma inglés [2], [126], [144], [247], [258], y en otros idiomas como: vietnamita [219], [253], [263], chino [273], turco [248], tailandés [107], [143], serbio [125], birmano [258], y español [249], [255].

Por lo tanto, se obtuvo el diseño de un modelo de evaluación entre pares con evaluación cualitativa, cuantitativa, inversa, en dos rondas y calibrada, un prototipo para la recolección de datos, un corpus para entrenamiento con 21280 instancias de evaluación de tarea y 13740 instancias de evaluación de la calidad de evaluación, etiquetadas en positivo, negativo y neutral (Tabla **49**), y un corpus para validación con 20322 instancias de evaluación de tarea y 19997 instancias de evaluación de la calidad de evaluación (Tabla **50**).

¿Cómo agilizar la generación de puntuación de sentimiento de retroalimentación textual (evaluación cualitativa) en procesos de evaluación entre pares?

Para generar la puntuación de sentimiento de retroalimentación textual, se realizó modelos predictivos mediante aprendizaje automático y profundo. Con una muestra de 2134 instancias se evaluó algoritmos de aprendizaje automático clásico (NB, K-NN (IBk) y LibSVM). LibSVM (F-Measure de 0.870) con representación de Bag of Words+TF-IDF+Stop-Words, obtuvo mejor rendimiento (Tabla **19**) (ver Modelo). Con una muestra de 3968 instancias, se evaluó modelos predictivos con aprendizaje automático clásico (MNB, SVM, LR, RF y DT), el modelo SVM (F-Measure de 0.879) fue el mejor con función lineal. Con aprendizaje automático moderno, el modelo VE (F-Measure de 0.882) superó a SVM, pero con costos computacionales más altos; ambos modelos demostraron su eficacia cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF; sin embargo, con el modelo de aprendizaje profundo, el rendimiento de predicción con Bi-LSTM (F-Measure F de 0.837) fue más débil, por lo que el tamaño de la muestra fue pequeño (Tabla **39**). Se seleccionó como modelo baseline a SVM para predecir el valor de puntuación (1-positivo/-1-negativo) porque obtuvo un

buen desempeño con bajo costo computacional (ver Modelo). Los resultados previos de esta investigación, han sido publicados en [297], [300].

Finalmente, con una muestra de 21280 instancias de evaluación de tarea y 13740 instancias de evaluación de calidad de evaluación, recopiladas en escenario de educación presencial y virtual asincrónico, se entrenó los algoritmos LSTM/Bi-LSTM con enriquecimiento semántico Word2Vec/Glove preentrenado. Los resultados revelaron que el modelo que predice la puntuación de sentimiento (1-positivo/0-neutral/-1-negativo) de evaluación de tarea con mejor rendimiento fue Bi-LSTM con representación de Glove (F-Measure de 0.963) (Tabla **57**), y que el modelo que predice la puntuación de sentimiento (1-positivo/0-neutral/-1-negativo) de evaluación de calidad de evaluación con mejor rendimiento fue LSTM con representación de Glove (F-Measure de 0.980) (Tabla **63**) (ver Modelo). Por lo tanto, las representaciones de Glove superaron al modelo Baseline de esta investigación, que ha sido publicado en [300].

Estos resultados con algoritmos de aprendizaje automático y profundo son similares a otros trabajos destacados de esta área de investigación. Por ejemplo, en investigaciones de evaluación por pares con corpus en inglés, [19] utilizaron SVM para evaluar calificaciones numéricas junto con texto de forma libre y obtuvieron un promedio de 0.780 de 2943 evaluaciones analizadas; [18] aplicó LSTM para detectar automáticamente las inconsistencias entre las puntuaciones numéricas y comentarios textuales, y obtuvo un error absoluto medio de 0.22 de 1925 revisiones; [252] aplicó NB (Precisión de 0.861) de 300 comentarios de varios foros de aprendizaje. En otras investigaciones del contexto educativo sobre la evaluación de la enseñanza sobre la retroalimentación de los estudiantes, [253] en idioma vietnamita obtuvo 0.876 en LSTM y 0.920 en Bi-LSTM, en términos de F-Measure de 16000 comentarios; [107] en tailandés utilizaron VE (Precisión de 0.872) de 400 instancias etiquetadas; [248] en idioma turco utilizaron SVM (F-Measure of 0.990) de 27586 tweets; [125] en serbio usó SVM (F-Measure of 0.760) de 3403 anotaciones y (F-Measure of 0.670) de 8101 anotaciones; y [258] en el idioma birmano utilizó SVM (F-Measure of 0.972) de 3000 comentarios.

En tal sentido, se obtuvo dos modelos predictivos de retroalimentación textual con mejor rendimiento, un modelo (Bi-LSTM con representación de Glove) que predice la puntuación de sentimiento de retroalimentación textual de evaluación de tarea, y otro modelo (LSTM con representación de Glove) que predice la puntuación de sentimiento de retroalimentación textual de evaluación de la calidad de evaluación.

¿Cómo correlacionar puntuación numérica (evaluación cuantitativa) con retroalimentación textual (evaluación cualitativa), y generar una puntuación equilibrada entre estas dos evaluaciones?

Para detectar precisión/imprecisión entre puntuación numérica y de sentimiento generada por el modelo predictivo, y calcular la puntuación de evaluación por cada criterio, con todos los criterios por evaluador o grupo, se aplicó lógica difusa utilizando el enfoque de Mamdani y los métodos de desfuzzificación centroide, bisector, SOM, MOM y LOM (ver Modelo). Los resultados previo de esta investigación han sido publicados en [300], mismos que fueron refinados, cuando existe impresión entre puntuación numérica y de sentimiento por cada criterio no se aplica ninguna penalización (ver Modelo). Los métodos SOM, MOM y LOM fueron los métodos más apropiados para este estudio porque calcularon el resultado más plausible seleccionando el valor típico del término lingüístico de salida más válido.

Estos resultados son similares a otras investigaciones que consideraron la vaguedad y la inexactitud; [311] mostró que los métodos de máximos, como la media de los máximos, se utilizan a menudo en los sistemas de razonamiento difuso para calcular el resultado más plausible, mientras que los métodos de distribución y los métodos de área, como el centro de gravedad, son cada vez más populares en los controladores difusos debido a la propiedad de continuidad; [212] demostraron que, con datos sin intervalos, la elección del método de defuzzificación no influye en la salida.

En tal sentido, se obtuvo un modelo con lógica difusa que genera una puntuación equilibrada entre puntuación numérica y puntuación de sentimiento por cada criterio tanto para evaluación de tarea y evaluación de calidad de la evaluación.

¿Cómo determinar un índice (rating) de confidencia de los pares evaluadores (evaluación inversa)?

Los evaluados valoraron la calidad de la evaluación de sus pares evaluadores por cada criterio de la rúbrica con puntuación numérica, y respondieron si la retroalimentación recibida es correcta o parcialmente correcta o incorrecta, justificando el por qué, en base al contenido de la temática. Posteriormente mediante el modelo predictivo se generó la puntuación de sentimiento de la retroalimentación textual y se la correlacionó con la puntuación numérica mediante lógica difusa por cada criterio. Seguidamente se calculó la puntuación de evaluación que dio cada evaluado y

finalmente se obtuvo la puntuación del colectivo, que sería el índice (rating) de confianza de cada evaluador por cada actividad, ya que se consideró que tanto el evaluador y el evaluado producen un trabajo similar en la misma temática (ver Subsección 4.1.3).

En contraste con otros estudios que determinaron solo la utilidad de las retroalimentaciones de manera general, [305] evaluaron la calidad de la retroalimentación de los pares categorizándolas en ninguna, general, especifica, constructiva, los tutores dieron mayor peso a los comentarios que consideraron constructivos y menos peso a los comentarios que son muy generales; [280] evaluaron la utilidad de cada comentario usando la votación me gusta o disgusto, luego permitieron que los evaluados acentúen si están o no de acuerdo con el resultado y finalmente ofrezcan de forma anónima comentarios adicionales para compartir su experiencia y posibles preocupaciones sobre el proceso de evaluación por pares, no lo hacen por cada criterio ni obtienen el rating de confianza de cada evaluador.

Varios autores [85], [280], [281], [312] destacan que para que los procesos de retroalimentación sean efectivos, deben involucrar a los estudiantes activamente en la generación, el procesamiento y la respuesta a la información de la retroalimentación. La respuesta o refutación por escrito requiere que los estudiantes participen críticamente con los comentarios y justifiquen las decisiones tomadas [85]. Este aspecto alienta a los estudiantes a procesar la información de la retroalimentación, proporcionar justificaciones para los comentarios que se han utilizado o no, y este tipo de interacción y co-construcciones de comprensiones se alinean bien con las teorías de aprendizaje constructivista social [312].

En tal sentido, se obtuvo el índice (rating) de confianza de cada evaluador, acrecentando la confiabilidad del proceso de evaluación entre pares, ya que el estudiante participó en evaluar y brindar retroalimentación sobre la evaluación de sus pares evaluadores por cada criterio.

¿Qué método de tendencia central (media/mediana) tiene el mejor ajuste para generar una puntuación del colectivo confiable en procesos de evaluación entre pares?

El trabajo abierto, fue evaluado en dos rondas por los evaluadores con puntuación numérica y retroalimentación textual justificando el porqué de tal puntaje en base al contenido de la temática a evaluar por cada criterio de la rúbrica. En la segunda ronda los evaluadores evaluaron el trabajo corregido, pudiendo revisar la evaluación de la primera ronda y la respuesta del evaluado para verificar si corrigieron o no el trabajo, o si evaluado tenía razón o no en su refutación. Tanto en la primera y segunda ronda se aplicó el modelo predictivo para generar la puntuación de sentimiento

de la retroalimentación textual, y se la correlacionó con la puntuación numérica mediante lógica difusa, obteniéndose una puntuación equilibrada entre estas dos evaluaciones por cada criterio. Seguidamente se obtuvo la puntuación individual por cada evaluador calculando la media de todos los criterios, ya que se consideró que todos los criterios son importantes (ver Modelo).

Los resultados (ver Subsección 5.2.1) mostraron que la relación entre la puntación individual que da cada evaluador con la puntuación que otorga el docente alcanzaron similaridad directa escasa y débil en las asignaturas impartidas en escenario de educación virtual asincrónico (Tabla **77**), similaridad débil en las asignaturas impartidas en escenario de educación virtual sincrónico (Tabla **79**) y similaridad directa sustancial en la asignatura impartida en escenario de educación presencial (Tabla **81**). Coligiendo que las puntuaciones individuales no alcanzaron relación fuerte con la del docente en ningún escenario de educación. Similar a otras investigaciones que denotaron que el resultado de múltiples evaluadores es más confiable que un resultado individual [313]. El problema de las valoraciones incorrectas para producir la nota final puede reducirse considerando varias calificaciones en una misma actividad [18]. Recibir múltiples revisiones por pares es beneficioso porque expone a los estudiantes a una diversidad de perspectivas [314]. Se recomiendan múltiples revisiones por pares para que los estudiantes estén expuestos a una variedad de trabajos de diferente calidad y ellos mismos reciban más de una revisión por pares [52].

Finalmente, para lograr una evaluación confiable, se calcula la puntuación del colectivo aplicando la media/mediana del conjunto de puntajes individuales. En la aplicación de la media se consideró que todas las puntuaciones individuales son igualmente importantes, y en la aplicación de la mediana se consideró que algunas puntuaciones pueden estar sesgadas por alguna razón tanto en términos de beneficio como de perjuicio. Los resultados muestran en los histogramas que las distribuciones no siguen una tendencia Gaussiana, además presentan una gran cantidad de valores extremos que hacen que la media cambie su magnitud, en contraste la mediana no se ve afectada por valores muy grandes o muy pequeños, por lo tanto, la mediana es robusta y filtra los valores extremos (Tabla **71**, Figura **61** y Figura **62**).

En cuanto al comportamiento de la aplicación de media/mediana no hubo diferencias significativas. En escenario de educación virtual asincrónico, la puntuación del colectivo alcanza mayor relación con la del docente empleando media, en directa moderada (13%), y directa sustancial (42%); utilizando mediana en directa débil (29%), e inversa débil (13%); y aplicando media/mediana, en directa escasa (4%), inversa moderada (4%) y directa fuerte (8%). En

escenario de educación virtual sincrónico, la puntuación del colectivo alcanza mayor relación con la del docente usando media, en directa débil (31%), y directa moderada (13%); utilizando mediana en directa escasa (19%), e inversa escasa (6%); y aplicando media/mediana, en inversa débil (6%), directa sustancial (19%) y directa fuerte (25%). En escenario de educación presencial, la puntuación del colectivo alcanza mayor relación con la del docente usando media, en directa débil (20%), inversa débil (20%), directa moderada (10%), e inversa moderada (10%); utilizando mediana en directa escasa (20%), inversa escasa (20%), directa sustancial (20%), y directa fuerte (30%); y aplicando media/mediana en inversa fuerte (10%) (Tabla **83**). Se concluye que media/mediana son métodos de cálculos válidos en procesos de evaluaciones entre pares, su aplicabilidad depende del escenario de educación. La media tiene mejor aplicabilidad en escenario virtual asincrónico y sincrónico, y la mediana en virtual sincrónico y presencial.

Del mismo modo, [315] argumentaron que la mediana o la media aritmética son métodos de agregación válidos, y cada uno tiene puntos fuertes y débiles, la media utiliza todos los datos disponibles, mientras que la mediana es robusta y filtra los valores extremos [316]. Calcularon la calificación final de la tarea mediante la mediana de las calificaciones recibidas de los revisores de pares [317]. Calcularon el valor central de cada ítem de la matriz de revisión, aplicando la mediana de las diferentes valoraciones dadas a una misma actividad, y destacaron que si bien se podrían considerar otros descriptores estadísticos como el máximo/mínimo o la media aritmética, recurrieron al descriptor de la mediana debido a su robustez frente a los valores atípicos [96].

Por lo tanto, se determinó que la mediana tiene el mejor ajuste para generar una puntuación del colectivo confiable, ya que estadísticamente es más efectiva para lidiar con puntajes extremos, que ocurren con frecuencia en el proceso de evaluación entre pares, debido a que los estudiantes varían en sus capacidades y/o motivación para calificar.

¿Cómo calibrar la puntuación de evaluación de tarea, que establezca fiabilidad en el proceso de evaluación entre pares?

Se calibró la puntuación de evaluación de tarea por cada actividad, ya que cada actividad es sobre una temática diferente y en escenario diverso, adicionando o restando la proporción que se obtiene de la varianza (tendencia central) multiplicada por desviación estándar (dispersión) de todas puntuaciones dadas por los evaluadores, en función del rendimiento y índice (rating) de

confianza que obtuvo el evaluador. Posteriormente se recálculo las puntuaciones del colectivo aplicando media/mediana (ver Subsección 4.1.4).

Los resultados (Tabla **84**) mostraron que la relación entre la puntación del colectivo y puntuación que proporciona el docente con la calibración tiende a: a) subir, el 46% de las actividades aplicando media/mediana en escenario de educación virtual asincrónico; el 63% de las actividades empleando media, y el 69% usando mediana en escenario de educación virtual sincrónico; y, el 50% de las actividades aplicando media, y el 60% utilizando mediana en escenario de educación presencial; b) mantener el mismo valor, el 8% de las actividades aplicando media/mediana en escenario de educación virtual asincrónico; el 13% de las actividades empleando media, y el 6% usando mediana en escenario de educación virtual sincrónico; y, el 30% de las actividades aplicando mediana en escenario de educación presencial; c) bajar, el 46% de las actividades aplicando media/mediana en escenario de educación virtual asincrónico; el 25% de las actividades empleando media/mediana en escenario de educación virtual sincrónico; y, el 50% de las actividades aplicando media, y el 10% utilizando mediana en escenario de educación presencial.

Este diseño experimental no ha sido comparable con lo de otros estudios, porque los autores han aplicado otros factores o medidas en la calibración, por ejemplo: [318] introduce una fase de calibración que ayuda a los estudiantes a aprender a calificar, practicando primero con envíos de ejemplo; [319] realizaron la calibración eliminando todas las evaluaciones poco confiables, si un estudiante califica alto un trabajo sin terminar, o con cero un trabajo terminado, la calificación para ese trabajo será irrelevante; [320] aplicó un enfoque mixto en que los estudiantes calibraron contra el trabajo real de sus compañeros que también había sido corregido por un tutor para marcar a los revisores malos o buenos; [321] los estudiantes tenían que demostrar alineación con un puntaje predefinido, si un estudiante no cumplía con las expectativas, podría repetir los ejercicios de calibración hasta lograr una alineación aceptable, la variación aceptable del 10% en los puntajes de calibración entre las evaluaciones de los estudiantes y el predeterminado; [322] proponen un método calibrado colaborativo para seleccionar a los evaluadores según el perfil para promover la productividad entre los alumnos; [96] proponen un método basado en el análisis estadístico en el que el docente sólo interviene en el proceso de corrección cuando existen discrepancias severas entre los diferentes pares revisores, para esto calcularon el valor central de cada ítem de la matriz de revisión, luego la distancia de cada evaluador de la central; [313] realizaron un modelo de motivación, recompensando a los

estudiantes que dan puntajes justos, así como penalizar a aquellos que no dan puntajes justos, dependiendo de qué tan cerca esté la puntuación del valor medio de todas las puntuaciones dadas por un grupo de revisión; y otros autores [99], [323], [324] utilizaron el software calibrated peer review, en la etapa de calibración, los estudiantes aplican criterios de evaluación a tres muestras, que han sido seleccionados y calificados por el instructor; las calificaciones de cada muestra deben estar dentro de una desviación estándar establecida por el instructor antes de que los estudiantes puedan ingresar a la etapa de revisión.

Por lo tanto, se obtuvo un modelo de calibración que contribuyó a mejorar la fiabilidad en el proceso de evaluación entre pares, ya que mediante el ajuste de la puntuación individual que dio cada evaluador en base a sus perfiles, se logró que la relación entre la puntación del colectivo y puntuación que proporciona el docente tendiera a subir en más del 50% de las actividades ejecutadas en escenarios de educación virtual sincrónico y presencial.

¿Cuál es la incidencia de la modalidad de educación en procesos de evaluación entre pares?

Para determinar la incidencia de la modalidad de educación en procesos de evaluación entre pares, se probó la validez del modelo propuesto en 3 escenarios de educación: virtual asincrónico, virtual sincrónico y presencial (ver Subsección 5.2.1).

Los resultados (Tabla **83**) mostraron que las mejores correlaciones que se obtuvieron en términos de acuerdo entre la puntación que recibe el estudiante del colectivo y la puntuación que proporciona el docente fueron: a) similaridad directa sustancial, el 42% de las actividades aplicando media, y el 38% aplicando mediana obtuvieron (r=0.517 a 0.688) en escenario de educación virtual asincrónico; y, el 19% de las actividades aplicando media/mediana obtuvieron (r=0.507 a 0.694) en escenario de educación virtual sincrónico; b) similaridad inversa sustancial, el 20% de las actividades aplicando mediana obtuvieron (r=-0.545) en escenario de educación presencial; c) similaridad directa fuerte, el 8% de las actividades aplicando media/mediana obtuvieron (r=0.718-0.790) en escenario de educación virtual asincrónico; el 25% de las actividades aplicando media/mediana obtuvieron (r=0.741 a 0.971) en escenario de educación virtual sincrónico; y, el 30% de las actividades aplicando media/mediana obtuvieron (r=0.780 a 0.951) en escenario de educación presencial; d) similaridad inversa fuerte, el 10% de las actividades aplicando mediana obtuvieron (r=-0.852) en escenario de educación presencial.

La literatura muestra puntos de vista contrastantes sobre la correlación entre la puntuación de los evaluadores y el docente. En estudios realizados en escenarios de educación MOOC, [317] obtuvieron correlación moderada entre los instructores y evaluadores pares (r=0.49); [96] presentaron que las notas obtenidas de las actividades tras aplicar la corrección de valores atípicos son muy similares a las que ofrece la evaluación manual del docente (r=0.98) en la primera actividad, y (r=0.95) en la segunda actividad; [325] alcanzaron una correlación general positiva moderada (r=0.39) entre los evaluadores pares y el docente; [326] lograron una correlación (r=0.71) entre las calificaciones de los instructores y los pares; [315] obtuvieron mejores puntuaciones, considerando como referencia las del instructor, al utilizar pesos basados en el compromiso de tres evaluadores y en mayor medida al utilizar pesos basados en el desempeño, la correlación mejoró de (r=0.45) correlación moderada a (r=0.76) correlación fuerte; [327] encontraron una correlación moderada (r=0.57) entre el instructor y la puntuación media de tres evaluadores; y [316] encontraron una correlación moderada (r =0.62) entre la calificación del instructor y la de los evaluadores basada en la mediana cuando se asignaba tres evaluadores; aumentaron ligeramente (r =0.66) cuando se utilizó la media, en lugar de la mediana, para calcular las calificaciones finales. En escenarios de educación presencial, [328] obtuvo correlación moderada (de 0.44 a 0.55) entre calificaciones de grupos de expertos y cinco evaluadores pares utilizando una interfaz de evaluación basada en la web.

Por lo tanto, se determinó que la modalidad de educación incide de forma proporcional en la evaluación entre pares, es así que el escenario de educación presencial es el más adecuado para poder aplicar el modelo propuesto alcanzando (r=0.780 a 0.951) en similaridad directa fuerte y (r=-0.852) en similaridad inversa fuerte.

¿De qué manera el modelo de evaluación entre pares basado en análisis de sentimiento podría contribuir en el proceso de enseñanza-aprendizaje?

Se evaluó si existe mejora del rendimiento estudiantil en la segunda ronda aplicando análisis de sentimiento en evaluación entre pares, mediante la prueba de t de Student (ver Subsección 5.2.2.1). Los resultados (Tabla **85**) mostraron que existe diferencia significativa en las 27 actividades ejecutadas. En el escenario virtual asincrónico, las 12 actividades obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, y un aumento del (3% a 12%) en el rendimiento del estudiante. En escenario virtual sincrónico, las 8 actividades adquirieron la puntuación media en la segunda ronda mayor que la

primera ronda con un valor de significancia menor a 0.05, y un incremento del (7% a 22%) en el rendimiento del estudiante. En escenario presencial, las 7 actividades tuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05, y un acrecentamiento del (15%-34%) en el rendimiento del estudiante. Se colige que, tanto en el escenario de educación virtual asincrónico/sincrónico y presencial, los estudiantes mejoraron el rendimiento en la segunda ronda, y con mayor efectividad en el escenario presencial.

Estos hallazgos pueden justificarse utilizando la taxonomía de dominios de aprendizaje de Bloom, ya que la evaluación entre pares es una forma de evaluación, que se considera una actividad cognitiva de nivel superior en la jerarquía taxonómica que promueve el aprendizaje significativo [329]. El aprendizaje que resulta de la revisión por pares se debe principalmente a la retroalimentación interna que los estudiantes generan sobre su propio trabajo durante la revisión [42]. La evaluación formativa es un proceso que implica el intercambio de retroalimentaciones, con el objetivo de mejorar la calidad del proceso de enseñanza- aprendizaje, lo que resulta en una mejora de la autonomía del estudiante y la maximización del resultado del aprendizaje [330]. Los procesos de aprendizaje entre pares de alta calidad no solo se reflejaron en la redacción de la retroalimentación argumentativa de los estudiantes, sino también en el aprendizaje específico de un tema [234]. Al recibir retroalimentación de los compañeros con argumentos justificados ayudó a los estudiantes a adquirir perspectivas divergentes y diversas sobre el tema en cuestión durante el proceso de evaluación entre pares [309].

Los resultados de la encuesta (ver Subsección 5.2.2.2) también respaldaron la utilidad del proceso de evaluación entre pares, ya que todos los aspectos considerados a la valoración de la retroalimentación proporcionada y recibida tienen una puntuación media muy positiva (Tabla **88** y Figura **133**). Los estudiantes han considerado que la retroalimentación les ha ayudado a mejorar sus habilidades (dada (4.50) recibida (4.46)), a mejorar el proceso de aprendizaje (dada (4.51) recibida (4.37)), a mejorar sus trabajos (dada (4.49) recibida (4.36)), a mejorar la implicación en el aprendizaje (dada (4.45) recibida (4.41)), y a mejorar la adquisición del contenido (dada (4.46) recibida (4.38)). Se colige que la tendencia de los estudiantes fue valorar más altas las retroalimentaciones que dieron a sus compañeros en todos los indicadores planteados.

Estos resultados están alineados a otros estudios [42], [99], [304], [305] que destacan que los estudiantes aprenden más al dar retroalimentación sobre el trabajo de sus compañeros que al recibir retroalimentación de sus compañeros. No obstante otros autores indicaron que tanto al dar como recibir retroalimentación condujeron a mejorar el rendimiento [222], [307]. El revisar

un trabajo y dar retroalimentación involucra procesos de orden superior, como la aplicación de criterios, la emisión de juicios y la elaboración de sugerencias para mejorar [42]. Activa un proceso reflexivo mediante el cual el estudiante se involucra en la resolución de problemas en relación con el trabajo de sus compañeros, detecta debilidades, fortalezas y propone soluciones [78]. Esto hace que los estudiantes durante la revisión examinen su propio entendimiento sobre la temática, comparen su propio trabajo con el de sus compañeros, y construyan nuevos conocimientos, que luego aplican a su propio trabajo [305]. Al recibir retroalimentación, el estudiante compara el trabajo que ha producido con una descripción textual de lo que es bueno o deficiente en su trabajo o sobre cómo se podría mejorar ese trabajo, y a partir de esa comparación generan nuevos conocimientos [306], por lo tanto el recibir retroalimentación también tiene impacto en el aprendizaje, coligiendo que los estudiantes adquieren nuevos conocimientos al dar y recibir retroalimentación.

Hay varias teorías de aprendizaje que sustentan la investigación y la práctica de la retroalimentación. El enfoque cognitivista enfatiza el procesamiento de la información en el sentido de que las personas reciben retroalimentaciones, los procesan y toman medidas para cerrar las brechas entre los niveles de rendimiento actuales y deseados [331]. El enfoque constructivista social adopta la perspectiva de que el conocimiento se forma a través de la participación, el diálogo y la co-construcción entre individuos [312]. El enfoque sociocultural destaca el papel de la participación y la construcción de significado mediada por la actividad dentro de contextos sociales y culturales [332].

En tal sentido, el modelo de evaluación entre pares podría implementarse como una herramienta pedagógica para apoyar al docente en enriquecer el proceso de enseñanza-aprendizaje, ya que los estudiantes dieron y recibieron retroalimentaciones detalladas sobre lo que deben mejorar en una tarea específica, y pudieron refutar sobre las retroalimentaciones dadas, lo que a su vez indujo que mejoraran el trabajo y el rendimiento en la segunda ronda.

6.2. Conclusiones finales y contribución

A lo largo de este estudio se ha podido dar respuesta a las preguntas de investigación, y estas respuestas se sintetizan en los siguientes objetivos cumplidos, resultados y conclusiones:

Revisión bibliográfica

Tras realizar una revisión sistemática de la literatura (ver Subsección 2.2). Se obtuvieron las siguientes conclusiones:

- Se puede concluir que, de los artículos revisados, existe un 58% que han realizado análisis de sentimiento en retroalimentación de procesos de enseñanza-aprendizaje, y un 4% que han aplicado análisis de sentimiento en retroalimentación entres pares; sin embargo, un 37% de los estudios revelan que no han aplicado técnicas computacionales. Reflejándose que el uso de análisis de sentimiento ha sido escasamente explorado en retroalimentación entre pares. Por lo tanto, el diseño de herramientas tecnológicas basadas en enfoque de análisis de sentimiento y NLP, se considera una contribución importante a la comunidad científica para apoyar la mejora del proceso de evaluación entre pares.
- La investigación realizada, evidencia que los estudios en evaluación entre pares basada en retroalimentación, y en análisis de sentimiento de texto educativo se han dado en mayor medida en el dominio de computación (57 estudios); sin embargo, cabe señalar que también se destacan estudios relacionados con el dominio de educación (32 estudios), por cuanto la evaluación entre pares basada en retroalimentación ayuda a mejorar el proceso de enseñanza-aprendizaje. Para lo cual han utilizado rúbricas de evaluación, donde el estudiante participa, reflexiona y adquiere aprendizaje, también han utilizado sistemas o herramientas para procesar grandes cantidades de datos educativos, mismas que han sido extraídas y exploradas por análisis de sentimiento utilizando algoritmos, recursos léxicos o herramientas.
- La revisión de la literatura revela que los investigadores para extraer y/o seleccionar características han utilizado con mayor frecuencia la técnica de POS (17 estudios), seguida N-Grams (13 estudios), TF-IDF (9 estudios), LDA (7 estudios), Word2Vec (7 estudios), GloVe (4 estudios), FastText (3 estudios), y BoW (2 estudios), y con menor frecuencia (1 estudio) utilizaron: SVD, LSA, E-LDA, BTM, WF, TP, TF, TextBlob3,

Doc2Vec, DEP y Cove; y para clasificar el sentimiento han aplicado con mayor frecuencia el método de aprendizaje automático (35 estudios) con los algoritmos: SVM (24 estudios), seguido NB (21 estudios), K-NN y DT (7 estudios), RF (6 estudios), ME, LR y CNB (4 estudios), y NN (3 estudios), y con menor frecuencia (1 estudio): ANN, Attribute Selected Classifier, Bagging, BFTree, GBT, Hoeffding Tree (VFDT), K (IBK), K-Star, Logistic, MaxEnt, Multi-Class Classifier, One R, PART, Random Subspace, REPT, RL, SPPM, Zero R, AdaBoost, Bayes Net, ID3, J48, SimpleLogistic, Stacking y Voting; además han aplicado el método de aprendizaje profundo (9 estudios), con mayor énfasis los algoritmos: LSTM (9 estudios), seguido CNN (4 estudios) y MLP (3 estudios), y con menor énfasis (1 estudio): Bi-LSTM, DT-LSTM, GRU, RNN-AM y RNN. Sin embargo, cabe señalar que también han aplicado léxicos (7 estudios) y herramientas de análisis de sentimiento (6 estudios). Estos resultados abren un abanico de opciones para realizar análisis de sentimiento de retroalimentación entre pares con diferentes métodos, técnicas NLP, algoritmos de clasificación, léxicos o herramientas.

Diseño de artefactos

Se diseñó diferentes artefactos que colaboran para que tanto la interacción de un escenario de evaluación entre pares y técnicas computacionales produzcan resultados de una calificación fiable entre puntuación cuantitativa, cualitativa, inversa, en dos rondas y calibrada (ver Sección 3).

Construcción de un modelo de evaluación entre pares

Durante la experimentación de esta investigación, se desarrolló modelos, con técnicas de: minería de texto, NLP, análisis de sentimiento y lógica difusa, y algoritmos de aprendizaje automático y profundo (ver Sección 4). A continuación, se detalla:

- **Evaluación del efecto de distintas técnicas de minería de texto, NLP sobre la tarea de clasificación de sentimiento**
  - Para determinar si el uso de determinadas técnicas de minería de texto y NLP podría aumentar o disminuir la eficacia de la tarea de clasificación de sentimiento de retroalimentación textual en español, se analizó (Bag of Words+Stemmer+TF-IDF+Stop-Words), combinaciones (N-Grams+TF-IDF+Stop-Words), Word2Vec/Glove preentrenados para formar los distintos vocabularios de los dos conjuntos de datos utilizados en esta investigación. Las principales conclusiones obtenidas al respecto

han sido; a) aplicando Bag of Words se obtuvo bajo rendimiento (Tabla **19**) en la tarea de clasificación, no se logró capturar el contexto de las palabras; b) la combinación de tokenización basada en 1-g y 2-g, cuya relevancia se ponderó con TF-IDF (Tabla **38**) mejoró la tarea de clasificación de sentimiento con algoritmos de aprendizaje automático; c) mediante la incorporación de fuentes externas Word2Vec/Glove preentrenados, se consiguió mejorar el rendimiento de algoritmos de aprendizaje profundo (Tabla **57** y Tabla **63**), ya que se generó características con enriquecimiento semántico, que hicieron aumentar la calidad del vocabulario que forman los distintos conjuntos de datos utilizados en la tarea de clasificación de sentimiento. Estos algoritmos preentrenados permitieron resolver ambigüedades terminológicas, expandir el vocabulario y mejorar la detección de negaciones, entre otras funcionalidades.

- **Evaluación del rendimiento de distintos algoritmos de clasificación**
  - Para determinar qué algoritmos arrojan un mejor rendimiento en la tarea de clasificación de sentimiento de retroalimentación textual en español, se evaluó algoritmos de aprendizaje automático clásico (NB, K-NN (IBk), MNB, SVM, LR, RF, DT), de aprendizaje automático moderno (VE), y de aprendizaje profundo (LSTM, Bi-LSTM). Las principales conclusiones obtenidas al respecto han sido; a) LibSVM (F-Measure de 0.870) con representación de Bag of Words+TF-IDF+Stop-Words, obtuvo mejor rendimiento con una muestra de 2134 instancias (Tabla **19**); b) SVM (F-Measure de 0.879) con una muestra de 3968 instancias superó el rendimiento con bajo costo computacional, sin embargo VE (F-Measure de 0.882) superó a SVM, pero con costos computacionales más altos; ambos modelos demostraron su eficacia cuando la retroalimentación se representó como un vector de 1-g y 2-g, cuya relevancia se ponderó con TF-IDF (Tabla **38**); c) Bi-LSTM (F-Measure de 0.963) con representación empleando Glove, supero el rendimiento de todos los algoritmos con una muestra de 21280 instancias de evaluación de tarea; e) se deduce que la representación de características y el tamaño de la muestra influyen en el rendimiento de los algoritmos.
- **Evaluación del modelo predictivo**
  - Los modelos predictivos seleccionados son confiables, ya que tuvo una concordancia igual o superior al 80% entre puntuación de sentimiento generada por el modelo, y polaridad de sentimiento etiquetada por el anotador (Tabla **40**).

  - El modelo predictivo es portable, ya que se lo aplicó en diferentes asignaturas, carreras y escenarios de educación superior presencial, virtual asincrónico y virtual sincrónico ante la pandemia COVID 19 (Tabla **41** y Tabla **69**).
- **Evaluación del modelo de cálculo con lógica difusa**
  - Para generar la puntuación de la evaluación de la correlación entre puntuación de sentimiento y numérica, por cada criterio o evaluador o grupo, se evaluó los métodos de defuzzificación en lógica difusa SOM, MOM, LOM, Centroid, y Bisector. Las principales conclusiones obtenidas al respecto han sido; a) SOM, MOM y LOM fueron los métodos más precisos para generar la puntuación de la evaluación (Tabla **45** y Tabla **46**); b) se determinó que la correlación por cada criterio es la más apropiada, puesto que la correspondencia es la predicción de puntuación de sentimiento de cada retroalimentación con su respectiva puntuación numérica.
- **Evaluación del modelo de calibración**
  - EL modelo de calibración acrecentó la fiabilidad del modelo ajustando la puntuación de evaluación de tarea en base a los perfiles del evaluador, puesto que la correlación entre la puntación que recibe el estudiante del colectivo con la puntuación que proporciona el docente tendió a subir: 46% de las actividades en escenario de educación virtual asincrónico, 69% en escenario virtual sincrónico y 60% en escenario presencial (Tabla **84**).

Evaluación del modelo de evaluación entre pares implementado

El modelo de evaluación entre pares se lo llevó a la práctica creando un prototipo capaz de interpretar la retroalimentación textual de evaluación de tarea y de evaluación de calidad de la evaluación, prediciendo la puntuación de sentimiento en positivo/negativo/neutral, generar una puntuación equilibrada de la correlación de puntuación de sentimiento y numérica mediane lógica difusa por cada criterio de la rúbrica, y generar la puntuación por cada evaluador y del colectivo (ver Sección 5).

- **Validez del modelo**
  - Se probó la validez del modelo propuesto en 3 escenarios de educación: virtual asincrónico, virtual sincrónico y presencial. Se correlacionó mediante Pearson la puntación que recibe el estudiante del colectivo con la puntuación que proporciona el docente (ver Subsección 5.2.1). Las principales conclusiones obtenidas al respecto

han sido; a) se obtuvo similaridad fuerte en el 8% de las actividades en virtual asincrónico (r=0.718-0.790), en el 25% de las actividades en virtual sincrónico (r=0.741 a 0.971) y en el 40% de las actividades en presencial (r=0.780 a 0.951) (Tabla **83**); b) se determinó que la modalidad de educación incide directamente en el proceso de aplicación de este modelo, y que el modelo si contribuye a la mejora sostenida del criterio del estudiante hacia el docente. *Sin embargo, se requiere realizar más pruebas para ir validando en diferentes contextos el mantener esta tendencia en escenario presencial;* c) por lo tanto, se deduce que el modelo se lo podría aplicar en un sistema en producción para que los estudiantes evalúen a sus compañeros con calificación cuantitativa (sumativa) y calificación cualitativa (formativa), sin que el docente evalúe.

- **Validación de hipótesis**
  - Se valido la hipótesis: En la evaluación entre pares utilizando análisis de sentimiento existen diferencias estadísticamente significativas en el rendimiento estudiantil, mediante la prueba t de Student por cada actividad de cada asignatura. La principal conclusión obtenida al respecto ha sido; a) el 100% de las actividades evaluadas obtuvieron la puntuación media en la segunda ronda mayor que la primera ronda con un valor de significancia menor a 0.05 (Tabla **85**); b) el incremento en la segunda ronda del rendimiento del estudiante en virtual asincrónico fue de 3%-12%, en virtual sincrónico de 7%-22%, y en presencial de 15%-34% (ver Subsección 5.2.2.1).
- **Utilidad del modelo**
  - Para establecer cuál es la percepción de los estudiantes respecto al proceso de evaluación entre pares, se realizó una encuesta (ver Subsección 5.2.2.2). La principal conclusión obtenida al respecto ha sido: a) todos los aspectos considerados han recibido una puntuación media muy positiva, puesto que los estudiantes consideran que la retroalimentación les ha ayudado a: mejorar sus habilidades (4.48), mejorar el proceso de aprendizaje (4.44), mejorar sus trabajos (4.43), mejorar la implicación en el aprendizaje (4.43), y mejorar la adquisición de contenidos (4.42) (Tabla **88**).
  - En tal sentido, el modelo propuesto se postula como una herramienta útil para ayudar a los estudiantes a mejorar sus trabajos y rendimiento en base a las retroalimentaciones dadas por sus compañeros en el proceso de evaluación de pares, a obtener resultados inmediatos de la evaluación, y ganancias cognitivas de reflexión y de análisis crítico; a los docentes a reducir la carga de calificaciones y mejorar los

procesos de enseñanza-aprendizaje, y a las autoridades a mejorar el entorno educativo, y satisfacer las necesidades educativas actuales que todos enfrentan con la tecnología y la pandemia. S*e precisa que este resultado está de acuerdo con el contexto del experimento; sin embargo, se requiere ampliar los espacios de experimentación para que pueda ser generalizado.*

La contribución de este estudio se centró en:

- Diseñar un modelo de evaluación entre pares con evaluación cuantitativa (sumativa), cualitativa (formativa), inversa, en dos rondas y calibrada.
- Crear un corpus en idioma español para entrenamiento con alrededor de 21280 instancias de evaluación de tarea, y 13740 instancias de evaluación de la calidad de evaluación, etiquetadas en positivo, negativo y neutral; y otro corpus para validación con 20322 instancias de evaluación de tarea y 19997 instancias de evaluación de la calidad de evaluación.
- Obtener un modelo que predice la puntuación de sentimiento de retroalimentación textual de evaluación de tarea, y otro modelo que predice la puntuación de sentimiento de retroalimentación textual de evaluación de la calidad de evaluación en positivo, negativo y neutral.
- Obtener un modelo de cálculo, con lógica difusa para generar una puntuación equilibrada entre puntuación numérica y puntuación de sentimiento por cada criterio, con la métrica de media para generar la puntuación individual de todos criterios y con la métrica media/mediana para generar la puntuación del colectivo.
- Obtener un modelo de calibración de la puntuación de evaluación de tarea en base al rendimiento y índice (rating) de confianza (evaluación inversa) del evaluador.
- Desarrollar un prototipo de evaluación entre pares e implementar los modelos.

6.3. Líneas de trabajo futuro

Se evidenció el gran potencial que pueden aportar las técnicas computacionales en el ámbito de la educación a lo largo del desarrollo de esta investigación. Se pudo desarrollar un modelo de evaluación entre pares que permite predecir la puntuación de sentimiento de retroalimentación textual mediante aprendizaje automático, correlacionar la puntuación numérica con puntuación de sentimiento mediante lógica difusa, y se logró llevarlo a la práctica, a través de la

implementación del prototipo, con el propósito de apoyar en el proceso de enseñanza-aprendizaje. Logrados estos aspectos, se observa que todavía queda un largo camino por recorrer y que son varias las líneas de investigación que podrían abrirse para mejorar el modelo de evaluación entre pares y el prototipo desarrollado. Algunas de estas líneas futuras se recogen a continuación:

- El principal inconveniente encontrado desde el inicio de esta investigación fue no poder encontrar un corpus de evaluación entre pares en español, de uso público para investigadores, y proveniente de un entorno real. Ante la dificultad de obtener este corpus se optó por realizar un procedimiento para la recopilación de datos y crear los conjuntos de datos para la experimentación. Por lo tanto, un aspecto pendiente es aumentar los conjuntos de datos mediante el prototipo de evaluación entre pares, corregir las predicciones incorrectas y reentrenar los algoritmos seleccionados.
- Los conjuntos de datos están desbalanceados. El conjunto de datos de evaluación de tarea contiene 8647 instancias etiquetadas con clase positiva, 6303 etiquetadas con clase neutral y 6330 etiquetadas con clase negativa. El conjunto de datos de evaluación de calidad de la evaluación contiene 6327 instancias etiquetadas con clase positiva, 3700 etiquetadas con clase neutral y 3713 etiquetadas con clase negativa. La aplicación de técnicas de rebalanceo adaptadas a los conjuntos de datos podría hacer mejorar el rendimiento del clasificador.
- El modelo de predicción de sentimiento desarrollado en esta investigación se ha basado a nivel de oración. Una futura línea de investigación podría ser predecir el sentimiento a nivel de aspecto por cada criterio de la rúbrica. El apoyo de esta ardua tarea podría simplificarse con aplicación de técnicas de minería de texto, NLP, y utilización de algoritmos de aprendizaje profundo.
- Se evidenció que los estudiantes si mejoran el rendimiento en la segunda ronda, sin embargo para acrecentar la fiabilidad del modelo, se podría adaptar un punto de control, que a partir de una puntuación base recibida en la primera ronda podrán corregir la tarea, para que el estudiante no se limite a mejorar solo en la segunda ronda, sino que se esfuerce en realizar correctamente la tarea desde el inicio, y que no envíe la tarea solo con un mínimo, ya que si al menos desde el inicio demuestra un conocimiento general, con las retroalimentaciones podrían mejorar en lo particular y lograr un mejor aprendizaje, y así se podrá obtener mejores resultados con el modelo propuesto.

- Con la calibración se logró que la correlación entre la puntuación del colectivo y la del docente tendiera a subir. No obstante, para mejorar la fiabilidad del modelo, se podría analizar otro tipo de medidas en la detección de puntuaciones sesgadas por cada criterio entre todos los evaluadores, y así obtener mejores resultados.
- Se demostró que el modelo es portable. Sin embargo, para mejorar se podría aplicar el modelo en otras facultades y escenarios de educación superior, que permita ir generando una cultura de evaluación entre pares en los estudiantes de toda la universidad.

6.4. Publicaciones asociadas a la tesis

6.4.1. Revistas internacionales

Título: Peer assessment using soft computing techniques [300]
Autores: Maricela Pinargote-Ortega, Jaime Meza y Sebastián Ventura

Journal of Computing in Higher Education, volume 33, pages 684–726 (2021)
DOI. https://doi.org/10.1007/s12528-021-09296-w

Ranking:
Factor de impacto (JCR 2021): 4.045
Área de conocimiento: educación e investigación educativa
Cuartil: Q1

Título: Peer Feedback Sentiment Analysis Prototype
Autores: Maricela Pinargote-Ortega, Jaime Meza y Sebastián Ventura
Aceptado para publicación en RISTI (Revista Ibérica de Sistemas y Tecnologías de la Información). 2023. (ver Apéndice C)

6.4.2. Conferencias internacionales

Título: Accuracy' Measures of Sentiment Analysis Algorithms for Spanish Corpus generated in Peer Assessment [297]
Autores: Maricela Pinargote Ortega, Jaime Meza Hormaza y Sebastián Ventura Soto
Proceedings of the 6th International Conference on Engineering & MIS 2020 (ICEMIS'20). Association for Computing Machinery, New York, NY, USA, Article 102, 1–7. https://dl.acm.org/doi/10.1145/3410352.3410838

Título: Sentiment Analysis Techniques for Peer Feedback: A Review [279]
Autores: Maricela Pinargote-Ortega, Jaime Meza y Sebastián Ventura
2023 Ninth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 2023, pp. 1-8. https://ieeexplore.ieee.org/document/10122085

Título: Sentiment Analysis of Peer Feedback in Higher Education
Autores: Maricela Pinargote-Ortega, Jaime Meza y Sebastián Ventura
Aceptado para publicación en AIP (American Institute of Physics) Conference Proceedings. 2023 (ver Apéndice B)

Título: Tendencias de investigación en la evaluación por pares basada en comentarios. Revisión literaria 2014-2018
Autores: Maricela Pinargote Ortega, Sebastián Ventura Soto
Congreso internacional y multidisciplinar de investigadores en formación-Universidad de Córdoba. 2019. Página 103. https://cidecuador.org/congreso/i-congreso-internacional-y-multidisciplinario-de-investigadores-en-formacion-en-ecuador/ (ver Apéndice D)

Título: Especificación de requerimientos para un sistema de evaluación entre pares en la Universidad Técnica de Manabí
Autor: Maricela Pinargote Ortega
XII encuentro internacional de investigación multidisciplinaria. Universidad Nacional Autónoma de México. 2019 (ver Apéndice E)

Título: Estudio de técnicas de minería de datos educativas para realizar predicciones en la educación superior
Autor: Maricela Pinargote Ortega
II convención científica internacional de la Universidad Técnica de Manabí. 2018 (ver Apéndice F)

Título: Revisión de sistemas de evaluación por pares en instituciones educativas
Autor: Maricela Pinargote Ortega
II convención científica internacional de la Universidad Técnica de Manabí. 2018 (ver Apéndice G)

Título: Análisis de sentimiento de la retroalimentación entre pares como alternativa para la evaluación formativa y sumativa
Autores: Maricela Pinargote Ortega, Jaime Meza Hormaza
V congreso internacional de ciencias informáticas-VI convención científica internacional de la Universidad Técnica de Manabí. 2022 (ver Apéndice H)

Título: Técnicas de análisis de sentimiento para la retroalimentación entre pares: una revisión
Autores: Maricela Pinargote Ortega, Jaime Meza Hormaza
V congreso internacional de ciencias informáticas-VI convención científica internacional de la Universidad Técnica de Manabí. 2022 (ver Apéndice I)

Título: Prototipo de análisis de sentimiento de retroalimentación entre pares
Autores: Maricela Pinargote Ortega, Jaime Meza Hormaza, Sebastián Ventura
VI Congreso Internacional de Tecnologías de la Información y Computación – CITIC. 2023 (ver Apéndice J)

REFERENCIAS

[1] Keppell, Au, Ma, and Chan, “Peer learning and learning-oriented assessment in technology-enhanced environments,” *Assessment \& Evaluation in Higher Education*, vol. 31, no. 4, pp. 453–464, 2006, doi: 10.1080/02602930600679159.

[2] Obeleagu, Abass, and Adeshina, “Sentiment analysis in student learning experience,” *2019 15th International Conference on Electronics, Computer and Computation, ICECCO 2019*, no. Icecco, pp. 0–4, 2019, doi: 10.1109/ICECCO48375.2019.9043293.

[3] Zott, Amit, and Massa, “The business model: Recent developments and future research,” *Journal of Management*, vol. 37, no. 4, pp. 1019–1042, 2011, doi: 10.1177/0149206311406265.

[4] Taylor, Ryan, and Pearce, “Enhanced student learning in accounting utilising web-based technology, peer-review feedback and reflective practices: a learning community approach to assessment,” *Higher Education Research & Development*, vol. 34, no. 6, pp. 1251–1269, Nov. 2015, doi: 10.1080/07294360.2015.1024625.

[5] Fang, Cassim, Hsu, and Chen, “Effects of reciprocal peer feedback on EFL learners’ communication strategy use and oral communication performance,” *Smart Learning Environments*, vol. 5, no. 1, p. 11, Dec. 2018, doi: 10.1186/s40561-018-0061-2.

[6] Cardoso, Hurst, and Nespoli, “Reflective Inquiry in Design Reviews: The Role of Question-Asking During Exchanges of Peer Feedback,” *International Journal of Engineering Education*, vol. 36, no. 2, pp. 614–622, 2020.

[7] Vu and Dall’Alba, “Students’ experience of peer assessment in a professional course,” *Assessment and Evaluation in Higher Education*, vol. 32, no. 5, pp. 541–556, 2007, doi: 10.1080/02602930601116896.

[8] Deiglmayr, “Instructional scaffolds for learning from formative peer assessment: effects of core task, peer feedback, and dialogue,” *European Journal of Psychology of Education*, vol. 33, no. 1, pp. 185–198, 2018, doi: 10.1007/s10212-017-0355-8.

[9] Daou, Sabra, and Zgheib, “Factors That Determine the Perceived Effectiveness of Peer Feedback in Collaborative Learning: a Mixed Methods Design,” *Medical Science Educator*, vol. 30, no. 3, pp. 1145–1156, 2020, doi: 10.1007/s40670-020-00980-7.

[10] Liaqat, Munteanu, and Demmans Epp, “Collaborating with Mature English Language Learners to Combine Peer and Automated Feedback: a User-Centered Approach to Designing Writing Support,” *International Journal of Artificial Intelligence in Education*, Jul. 2020, doi: 10.1007/s40593-020-00204-4.

[11] Ibarra-Sáiz, Rodríguez-Gómez, and Boud, “Developing student competence through peer assessment: the role of feedback, self-regulation and evaluative judgement,” *Higher Education*, vol. 80, no. 1, pp. 137–156, Jul. 2020, doi: 10.1007/s10734-019-00469-2.

[12] Teräs, Suoranta, Teräs, and Curcher, “Post-Covid-19 Education and Education Technology ‘Solutionism’: a Seller’s Market,” *Postdigital Science and Education*, vol. 2, no. 3, pp. 863–878, 2020, doi: 10.1007/s42438-020-00164-x.

[13] Chakraborty, Mittal, Gupta, Yadav, and Arora, “Opinion of students on online education during the COVID-19 pandemic,” *Human Behavior and Emerging Technologies*, no. November, pp. 1–9, 2020, doi: 10.1002/hbe2.240.

[14] VanSchenkhof, Houseworth, McCord, and Lannin, “Peer evaluations within experiential pedagogy: Fairness, objectivity, retaliation safeguarding, constructive feedback, and experiential learning as part of peer assessment,” *International Journal of Management Education*, vol. 16, no. 1, pp. 92–104, 2018, doi: 10.1016/j.ijme.2017.12.003.

[15] Goldin, Narciss, Foltz, and Bauer, “New Directions in Formative Feedback in Interactive

Learning Environments," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 3, pp. 385–392, 2017, doi: 10.1007/s40593-016-0135-7.

[16] Ramachandran, Gehringer, and Yadav, "Automated Assessment of the Quality of Peer Reviews using Natural Language Processing Techniques," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 3, pp. 534–581, Sep. 2017, doi: 10.1007/s40593-016-0132-x.

[17] Nguyen, Xiong, and Litman, "Iterative Design and Classroom Evaluation of Automated Formative Feedback for Improving Peer Feedback Localization," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 3, pp. 582–622, Sep. 2017, doi: 10.1007/s40593-016-0136-6.

[18] Rico-Juan, Gallego, and Calvo-Zaragoza, "Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning," *Computers & Education*, vol. 140, p. 103609, Oct. 2019, doi: 10.1016/j.compedu.2019.103609.

[19] Izzo and Maloy, "86 Sentiment Analysis Demonstrates Variability in Medical Student Grading," *Annals of Emergency Medicine*, vol. 70, no. 4, pp. S35–S36, Oct. 2017, doi: 10.1016/j.annemergmed.2017.07.111.

[20] Marsico, Sciarrone, Sterbini, and Temperini, "Educational Data Mining for Peer Assessment in Communities of Learners," in *The Future of Innovation and Technology in Education: Policies and Practices for Teaching and Learning Excellence*, Visvizi, Lytras, and Daniela, Eds., Emerald Publishing Limited, 2018, pp. 203–217. doi: 10.1108/978-1-78756-555-520181016.

[21] Sciarrone and Temperini, "K-OpenAnswer: a simulation environment to analyze the dynamics of massive open online courses in smart cities," *Soft Computing*, vol. 24, no. 15, pp. 11121–11134, 2020, doi: 10.1007/s00500-020-04696-z.

[22] El Alaoui, El Yassini, and Ben-Azza, "Peer Assessment Improvement Using Fuzzy Logic," Springer International Publishing, 2019, pp. 408–418. doi: 10.1007/978-3-030-11196-0_35.

[23] Capuano, Caballé, Percannella, and Ritrovato, "FOPA-MC: fuzzy multi-criteria group decision making for peer assessment," *Soft Computing*, vol. 24, no. 23, pp. 17679–17692, 2020, doi: 10.1007/s00500-020-05155-5.

[24] Ma, Zeng, Peng, Fortino, and Zhang, "Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis," *Future Generation Computer Systems*, vol. 93, pp. 304–311, 2019, doi: 10.1016/j.future.2018.10.041.

[25] Moreno-Marcos, Alario-Hoyos, Munoz-Merino, Estevez-Ayres, and Kloos, "Sentiment analysis in MOOCs: A case study," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, Apr. 2018, pp. 1489–1496. doi: 10.1109/EDUCON.2018.8363409.

[26] Chai, Tay, and Lim, "A new fuzzy peer assessment methodology for cooperative learning of students," *Applied Soft Computing*, vol. 32, no. April 2015, pp. 468–480, Jul. 2015, doi: 10.1016/j.asoc.2015.03.056.

[27] Capuano, Loia, Member, and Orciuoli, "A Fuzzy Group Decision Making Model for Ordinal Peer Assessment," vol. 10, no. 2, pp. 247–259, 2017.

[28] Barlybayev, Sharipbay, Ulyukova, Sabyrov, and Kuzenbayev, "Student's Performance Evaluation by Fuzzy Logic," *Procedia Computer Science*, vol. 102, no. August, pp. 98–105, 2016, doi: 10.1016/j.procs.2016.09.375.

[29] Voskoglou, "Fuzzy Logic as a Tool for Assessing Students' Knowledge and Skills," *Education Sciences*, vol. 3, no. 2, pp. 208–221, May 2013, doi: 10.3390/educsci3020208.

[30] Jyothi, Parvathi, Srinivas, and Althaf Rahaman, "Fuzzy Expert Model for Evaluation of Faculty Performance in Technical Educational Institutions," *Journal of Engineering*

Research and Applications www.ijera.com, vol. 4, no. 5, pp. 41–50, 2014.
[31] Wang, Subhan, Shamshirband, Zubair Asghar, Ullah, and Habib, “Fuzzy-based Sentiment Analysis System for Analyzing Student Feedback and Satisfaction,” *Computers, Materials & Continua*, vol. 62, no. 2, pp. 631–655, 2020, doi: 10.32604/cmc.2020.07920.
[32] Schneider, Oliveira, and de Souza, “Designing, building and evaluating a social news curation environment using the action design research methodology,” *Cluster Computing*, vol. 20, no. 2, pp. 1731–1748, 2017, doi: 10.1007/s10586-017-0781-z.
[33] Sein, Henfridsson, Purao, Rossi, and Lindgren, “Action design research,” *MIS Quarterly: Management Information Systems*, vol. 35, no. 1, pp. 37–56, 2011, doi: 10.2307/23043488.
[34] de Vries and Berger, “An Action Design Research Approach within Enterprise Engineering,” *Systemic Practice and Action Research*, vol. 30, no. 2, pp. 187–207, 2017, doi: 10.1007/s11213-016-9390-7.
[35] Tsai and Liang, “The development of science activities via on-line peer assessment: the role of scientific epistemological views,” *Instructional Science*, vol. 37, no. 3, pp. 293–310, 2009, doi: 10.1007/s11251-007-9047-0.
[36] Molenda, “Educational technology: an encyclopedia,” *Santa Barbara, CA: ABC-Clio*, 2003.
[37] Van Gennip, Segers, and Tillema, “Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions,” *Learning and Instruction*, vol. 20, no. 4, pp. 280–290, 2010, doi: https://doi.org/10.1016/j.learninstruc.2009.08.010.
[38] Van Zundert, Sluijsmans, and van Merriënboer, “Effective peer assessment processes: Research findings and future directions,” *Learning and Instruction*, vol. 20, no. 4, pp. 270–279, 2010, doi: https://doi.org/10.1016/j.learninstruc.2009.08.004.
[39] Topping, “Peer Assessment Between Students in Colleges and Universities,” *Review of Educational Research*, vol. 68, no. 3, pp. 249–276, 1998, doi: 10.3102/00346543068003249.
[40] Anderson and Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Longman, 2001.
[41] Rubin and Turner, “Student performance on and attitudes toward peer assessments on Advanced Pharmacy Practice Experience assignments,” *Currents in Pharmacy Teaching and Learning*, vol. 4, no. 2, pp. 113–121, 2012, doi: https://doi.org/10.1016/j.cptl.2012.01.011.
[42] Nicol, D., Thomson, A & Breslin, “Rethinking feedback practices in higher education: a peer review perspective,” *Assessment and Evaluation in Higher Education.*, vol. 39, no. 1, pp. 102–122, 2014.
[43] Cho, Schunn, and Wilson, “Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives,” *Journal of Educational Psychology*, vol. 98, no. 4, pp. 891–901, 2006, doi: 10.1037/0022-0663.98.4.891.
[44] Hsia, Huang, and Hwang, “Effects of different online peer-feedback approaches on students’ performance skills, motivation and self-efficacy in a dance course,” *Computers and Education*, vol. 96, pp. 55–71, 2016, doi: 10.1016/j.compedu.2016.02.004.
[45] Wen and Tsai, “Online peer assessment in an inservice science and mathematics teacher education course,” *Teaching in Higher Education*, vol. 13, no. 1, pp. 55–67, 2008, doi: 10.1080/13562510701794050.
[46] Bouzidi and Jaillet, “Can online peer assessment be trusted?,” *Journal of Educational Technology \& Society*, vol. 12, no. 4, pp. 257–268, 2009.
[47] Suen, “Peer assessment for massive open online courses (MOOCs),” *International Review of Research in Open and Distributed Learning*, vol. 15, no. 3, pp. 312–327, 2014.
[48] Meek, Blakemore, and Marks, “Is peer review an appropriate form of assessment in a

MOOC? Student participation and performance in formative peer review," *Assessment \& Evaluation in Higher Education*, vol. 42, no. 6, pp. 1000–1013, 2017.

[49] Gamage, Whiting, Rajapakshe, Thilakarathne, Perera, and Fernando, "Improving Assessment on MOOCs Through Peer Identification and Aligned Incentives," in *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, in L@S '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 315–318. doi: 10.1145/3051457.3054013.

[50] Garcia-Loro, Martin, Ruipérez-Valiente, Sancristobal, and Castro, "Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform," *Computers & Education*, vol. 154, p. 103894, 2020, doi: https://doi.org/10.1016/j.compedu.2020.103894.

[51] Taras, "Assessment – Summative and Formative – Some Theoretical Reflections," *British Journal of Educational Studies*, vol. 53, no. 4, pp. 466–478, 2005, doi: 10.1111/j.1467-8527.2005.00307.x.

[52] Cho and MacArthur, "Student revision with peer and expert reviewing," *Learning and Instruction*, vol. 20, no. 4, pp. 328–338, 2010, doi: https://doi.org/10.1016/j.learninstruc.2009.08.006.

[53] Sancho-Thomas, Fuentes-Fernández, and Fernández-Manjón, "Learning teamwork skills in university programming courses," *Computers & Education*, vol. 53, no. 2, pp. 517–531, 2009, doi: https://doi.org/10.1016/j.compedu.2009.03.010.

[54] Falchikov, *Learning together: Peer tutoring in higher education*. Routledge, 2003.

[55] Lin, "Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system," *Computers and Education*, vol. 116, pp. 81–92, 2018, doi: 10.1016/j.compedu.2017.08.010.

[56] Dijks, Brummer, and Kostons, "The anonymous reviewer: the relationship between perceived expertise and the perceptions of peer feedback in higher education," *Assessment & Evaluation in Higher Education*, vol. 43, no. 8, pp. 1258–1271, 2018, doi: 10.1080/02602938.2018.1447645.

[57] Kobayashi, "Does anonymity matter? Examining quality of online peer assessment and students' attitudes," *Australasian Journal of Educational Technology*, vol. 36, no. 1, pp. 98–110, 2020, doi: 10.14742/ajet.4694.

[58] Double, McGrane, and Hopfenbeck, "The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies," *Educational Psychology Review*, vol. 32, no. 2, pp. 481–509, 2020, doi: 10.1007/s10648-019-09510-3.

[59] Peters, Körndle, and Narciss, "Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance," *European Journal of Psychology of Education*, vol. 33, no. 1, pp. 117–143, 2018, doi: 10.1007/s10212-017-0344-y.

[60] Li … Suen, "Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings," *Assessment \& Evaluation in Higher Education*, vol. 41, no. 2, pp. 245–264, 2016, doi: 10.1080/02602938.2014.999746.

[61] Li, Liu, and Steckelberg, "Assessor or assessee: How student learning improves by giving and receiving peer feedback," *BRITISH JOURNAL OF EDUCATIONAL TECHNOLOGY*, vol. 41, no. 3, pp. 525–536, May 2010, doi: 10.1111/j.1467-8535.2009.00968.x.

[62] Liu and Li, "Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment," *Assessment & Evaluation in Higher Education*, vol. 39, no. 3, pp. 275–292, Apr. 2014, doi: 10.1080/02602938.2013.823540.

[63] Falchikov and Goldfinch, "Student peer assessment in higher education: A meta-analysis

comparing peer and teacher marks," *Review of Educational Research*, vol. 70, no. 3, pp. 287–322, 2000, doi: 10.3102/00346543070003287.

[64] Topping, "Peer Assessment," *Theory Into Practice*, vol. 48, no. 1, pp. 20–27, 2009, doi: 10.1080/00405840802577569.

[65] Ng, "Fostering pre-service teachers' self-regulated learning through self- and peer assessment of wiki projects," *Computers and Education*, vol. 98, pp. 180 – 191, 2016, doi: 10.1016/j.compedu.2016.03.015.

[66] Topping, "Peer assessment: Learning by judging and discussing the work of other learners," *Interdisciplinary Education and Psychology*, vol. 1, no. 1, pp. 1 – 17, 2017, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049638257&partnerID=40&md5=1c41f1346dff037331c95a00873bed67

[67] Kizilcec, Davis, and Cohen, "Towards equal opportunities in MOOCs: Affirmation reduces gender & social-class achievement gaps in China," in *L@S 2017 - Proceedings of the 4th (2017) ACM Conference on Learning at Scale*, 2017, pp. 121 – 130. doi: 10.1145/3051457.3051460.

[68] Li and Grion, "The Power of Giving Feedback and Receiving Feedback in Peer Assessment.," *The Power of Giving Feedback and Receiving Feedback in Peer Assessment.*, vol. 11, no. 2, pp. 1–17, 2019.

[69] Roscoe and Chi, "Tutor learning: the role of explaining and responding to questions," *Instructional Science*, vol. 36, no. 4, pp. 321–350, 2008, doi: 10.1007/s11251-007-9034-5.

[70] Cook … Cornes, "The impact of peer-review on undergraduate grades when students decide whether to participate," *Journal of Geography in Higher Education*, vol. 45, no. 2, pp. 238–254, 2021, doi: 10.1080/03098265.2020.1804844.

[71] Hughes, "But isn't this what you're paid for? The pros and cons of peer and self assessment," *Planet*, vol. 3, no. 1, pp. 20–23, 2001, doi: 10.11120/plan.2001.00030020.

[72] Chevalier, Dolton, and Lührmann, "Making it count': incentives, student effort and performance," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, vol. 181, no. 2, pp. 323–349, 2018, doi: 10.1111/rssa.12278.

[73] Gillanders, Karazi, and O'Riordan, "Loss aversion as a motivator for engagement with peer assessment," *Innovations in Education and Teaching International*, vol. 57, no. 4, pp. 424–433, Jul. 2020, doi: 10.1080/14703297.2020.1726203.

[74] Ng, "Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring," *Computers and Education*, vol. 73, pp. 141 – 148, 2014, doi: 10.1016/j.compedu.2013.12.016.

[75] Miller, "Formative computer-based assessment in higher education: the effectiveness of feedback in supporting student learning," *Assessment \& Evaluation in Higher Education*, vol. 34, no. 2, pp. 181–192, 2009, doi: 10.1080/02602930801956075.

[76] Kilic, "An Examination of Using Self-, Peer-, and Teacher-Assessment in Higher Education: A Case Study in Teacher Education," *Higher Education Studies*, vol. 6, no. 1, p. 136, 2016, doi: 10.5539/hes.v6n1p136.

[77] Chang, Tseng, and Lou, "A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students," *Computers \& Education*, vol. 58, no. 1, pp. 303–320, 2012.

[78] To and Panadero, "Peer assessment effects on the self-assessment process of first-year undergraduates," *Assessment and Evaluation in Higher Education*, vol. 44, no. 6, pp. 920–932, 2019, doi: 10.1080/02602938.2018.1548559.

[79] Ion, Sánchez Mart\'\i, and Agud Morell, "Giving or receiving feedback: which is more

beneficial to students' learning?," *Assessment \& Evaluation in Higher Education*, vol. 44, no. 1, pp. 124–138, 2019.
[80] Adachi, Tai, and Dawson, "A framework for designing, implementing, communicating and researching peer assessment," *Higher Education Research and Development*, vol. 37, no. 3, pp. 453 – 467, 2018, doi: 10.1080/07294360.2017.1405913.
[81] Hogg, "Empowering students through peer assessment: interrogating complexities and challenges," *Reflective Practice*, vol. 19, no. 3, pp. 308–321, 2018, doi: 10.1080/14623943.2018.1437404.
[82] Nicol, "The power of internal feedback: exploiting natural comparison processes," *Assessment \& Evaluation in Higher Education*, vol. 46, no. 5, pp. 756–778, Jul. 2021, doi: 10.1080/02602938.2020.1823314.
[83] Nicol and Selvaretnam, "Making internal feedback explicit: harnessing the comparisons students make during two-stage exams," *Assessment and Evaluation in Higher Education*, vol. 47, no. 4, pp. 507–522, 2022, doi: 10.1080/02602938.2021.1934653.
[84] Van Popta, Kral, Camp, Martens, and Simons, "Exploring the value of peer feedback in online learning for the provider," *Educational Research Review*, vol. 20, pp. 24–34, 2017, doi: https://doi.org/10.1016/j.edurev.2016.10.003.
[85] Harland, Wald, and Randhawa, "Student peer review: enhancing formative feedback with a rebuttal," *Assessment and Evaluation in Higher Education*, vol. 42, no. 5, pp. 801–811, 2017, doi: 10.1080/02602938.2016.1194368.
[86] Melville, "Crowd-Sourced Peer Feedback (CPF) for Learning Community Engagement: Results and Reflections from a Pilot Study," in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 32–41. doi: 10.1109/HICSS.2014.14.
[87] Nicol, "Reconceptualising feedback as an internal not an external process," *Italian Journal of Educational Research*, vol. 9744, 2019, doi: 10.7346/SIRD-1S2019-P71.
[88] Chien, Hwang, and Jong, "Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions," *Computers and Education*, vol. 146, no. October 2019, p. 103751, 2020, doi: 10.1016/j.compedu.2019.103751.
[89] Goldin and Ashley, "Peering inside peer review with bayesian models," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6738 LNAI, pp. 90 – 97, 2011, doi: 10.1007/978-3-642-21869-9_14.
[90] Pedrosa-de-Jesus, Guerra, and Watts, "Active Co-Constructive Written Feedback: Promoting Students' Critical Thinking in a Higher Education Context," in *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, in TEEM'19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 129–136. doi: 10.1145/3362789.3362825.
[91] Chen and Kuo, "An optimized group formation scheme to promote collaborative problem-based learning," *Computers and Education*, vol. 133, no. 64, pp. 94–115, 2019, doi: 10.1016/j.compedu.2019.01.011.
[92] Hbscher, "Assigning students to groups using general and context-specific criteria," *IEEE Transactions on Learning Technologies*, vol. 3, no. 3, pp. 178–189, 2010, doi: 10.1109/TLT.2010.17.
[93] van der Laan Smith and Spindle, "The impact of group formation in a cooperative learning environment," *Journal of Accounting Education*, vol. 25, no. 4, pp. 153–167, 2007, doi: 10.1016/j.jaccedu.2007.09.002.
[94] Herrera-Pavo, "Collaborative learning for virtual higher education," *Learning, Culture and*

Social Interaction, vol. 28, no. June 2020, 2021, doi: 10.1016/j.lcsi.2020.100437.
[95] Ravenscroft, Buckless, and Zuckerman, “Student team learning-replication and extension,” *Accounting Education*, vol. 2, no. 2, 1997.
[96] Rico-Juan, Gallego, Valero-Mas, and Calvo-Zaragoza, “Statistical semi-supervised system for grading multiple peer-reviewed open-ended works,” *Computers and Education*, vol. 126, no. December 2017, pp. 264–282, 2018, doi: 10.1016/j.compedu.2018.07.017.
[97] Stanley and Porter, *Engaging large classes: Strategies and techniques for college faculty.* Anker Publishing Company, 2002.
[98] Hicks, Pandey, Fraser, and Klemmer, “Framing feedback: Choosing review environment features that support high quality peer assessment,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 458–469.
[99] Russell, Van Horne, Ward, Bettis, and Gikonyo, “Variability in students’ evaluating processes in peer assessment with calibrated peer review,” *Journal of Computer Assisted Learning*, vol. 33, no. 2, pp. 178–190, 2017, doi: 10.1111/jcal.12176.
[100] Anglin, Anglin, Schumann, and Kaliski, “Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics,” *Decision Sciences Journal of Innovative Education*, vol. 6, no. 1, pp. 51–73, 2008, doi: 10.1111/j.1540-4609.2007.00153.x.
[101] Romero and Ventura, “Educational data science in massive open online courses,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 1, pp. 1–12, 2017, doi: 10.1002/widm.1187.
[102] Baker and Inventado, “Educational Data Mining and Learning Analytics,” in *Learning Analytics: From Research to Practice*, Larusson and White, Eds., New York, NY: Springer New York, 2014, pp. 61–75. doi: 10.1007/978-1-4614-3305-7_4.
[103] Berry and Castellanos, “Survey of text mining,” *Computing Reviews*, vol. 45, no. 9, p. 548, 2004.
[104] Ferreira-Mello, André, Pinheiro, Costa, and Romero, “Text mining in education,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 6, 2019, doi: 10.1002/widm.1332.
[105] Shi, Zhu, Li, Guo, and Zheng, “Survey on Classic and Latest Textual Sentiment Analysis Articles and Techniques,” *International Journal of Information Technology \& Decision Making*, vol. 18, no. 04, pp. 1243–1287, 2019, doi: 10.1142/S0219622019300015.
[106] Education, “What is text mining.” Retrieved from ibm. com: https://www. ibm. com/cloud/learn/text-mining, 2020.
[107] Pong-Inwong and Kaewmak, “Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration,” in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, IEEE, Oct. 2016, pp. 1222–1225. doi: 10.1109/CompComm.2016.7924899.
[108] Kovanović, Joksimović, Gašević, Hatala, and Siemens, “Content analytics: The definition, scope, and an overview of published research,” *Handbook of learning analytics and educational data mining*, pp. 77–92, 2017.
[109] Litman, “Natural language processing for enhancing teaching and learning,” *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 4170–4176, 2016, doi: 10.1609/aaai.v30i1.9879.
[110] Celestial-Valderama, Vinluan, and ..., “Mining students’ feedback in a general education course: Basis for improving blended learning implementation,” *International Journal of …*, vol. 5, no. 1, pp. 568–583, 2021, doi: 10.25147/ijcsr.2017.001.1.60.
[111] P. Miranda and T. Martin, “Topic Modeling and Sentiment Analysis of Martial Arts Learning

Textual Feedback on YouTube," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 2712–2718, 2020, doi: 10.30534/ijatcse/2020/35932020.

[112] Chowdhury, "Natural language processing," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 352–357, 2010, doi: 10.1002/wics.76.

[113] Bock and Garnsey, "Language Processing," *A Companion to Cognitive Science*, vol. 349, no. 6245, pp. 226–234, 2008, doi: 10.1002/9781405164535.ch14.

[114] Cambria and Hussain, "Tools," in *Sentic Computing: Techniques, Tools, and Applications*, Dordrecht: Springer Netherlands, 2012, pp. 69–101. doi: 10.1007/978-94-007-5070-8_4.

[115] Medhat, Hassan, and Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[116] Buche, "Opinion Mining and Analysis: A Survey," *International Journal on Natural Language Computing*, vol. 2, no. 3, pp. 39–48, 2013, doi: 10.5121/ijnlc.2013.2304.

[117] Jagtap and Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 3, pp. 164–170, 2013, [Online]. Available: http://ijset.com/ijset/publication/v2s3/paper11.pdf

[118] Wiebet, Brucet, and O'Hara, "Development and Use of a Gold-Standard Data Set for Subjectivity Classifications," p. 246, 1998.

[119] Wilson, Wiebe, and Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Computational Linguistics*, vol. 35, no. 3, pp. 399–433, 2009, doi: 10.1162/coli.08-012-R1-06-90.

[120] Fu and Wang, "Chinese sentence-level sentiment classification based on fuzzy sets," *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, vol. 2, no. August, pp. 312–319, 2010.

[121] Zhang and He, "Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification," *Journal of Information Science*, vol. 41, no. 4, pp. 531–549, 2015, doi: 10.1177/0165551515585264.

[122] Zhai, Liu, Xu, and Jia, "Clustering product features for opinion mining," *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, pp. 347–354, 2011, doi: 10.1145/1935826.1935884.

[123] Sindhu, Muhammad Daudpota, Badar, Bakhtyar, Baber, and Nurunnabi, "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation," *IEEE Access*, vol. 7, pp. 108729–108741, 2019, doi: 10.1109/ACCESS.2019.2928872.

[124] Sivakumar and Reddy, "Aspect based sentiment analysis of students opinion using machine learning techniques," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, Nov. 2017, pp. 726–731. doi: 10.1109/ICICI.2017.8365231.

[125] Nikolić, Grljević, and Kovačević, "Aspect-based sentiment analysis of reviews in the domain of higher education," *The Electronic Library*, vol. 38, no. 1, pp. 44–64, Jan. 2020, doi: 10.1108/EL-06-2019-0140.

[126] Kastrati, Imran, and Kurti, "Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs," *IEEE Access*, vol. 8, pp. 106799–106810, 2020, doi: 10.1109/ACCESS.2020.3000739.

[127] Riloff and Wiebe, "Learning extraction patterns for subjective expressions," pp. 105–112, 2003, doi: 10.3115/1119355.1119369.

[128] Kumar and Sebastian, "Sentiment Analysis: A Perspective on its Past, Present and Future," *International Journal of Intelligent Systems and Applications*, vol. 4, no. 10, pp. 1–14, 2012, doi: 10.5815/ijisa.2012.10.01.

[129] Pang, Lee, and Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” *American Journal of Orthodontics and Oral Surgery*, vol. 31, no. 9, pp. 481–482, 2002, doi: 10.1016/0096-6347(45)90048-2.
[130] Gupta and Agrawal, “Chapter 1 - Application and techniques of opinion mining,” in *Hybrid Computational Intelligence*, Bhattacharyya, Snášel, Gupta, and Khanna, Eds., in Hybrid Computational Intelligence for Pattern Analysis and Understanding. Academic Press, 2020, pp. 1–23. doi: https://doi.org/10.1016/B978-0-12-818699-2.00001-9.
[131] Pang and Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, no. June, pp. 115–124, 2005.
[132] Liu, “Sentiment Analysis and Opinion Mining,” p. 168, 2012.
[133] Zhang and Liu, “Identifying noun product features that imply opinions,” *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, no. 2009, pp. 575–580, 2011.
[134] Ullah, “Sentiment analysis of students feedback: A study towards optimal tools,” in *2016 International Workshop on Computational Intelligence (IWCI)*, IEEE, Dec. 2016, pp. 175–180. doi: 10.1109/IWCI.2016.7860361.
[135] Haddi, Liu, and Shi, “The role of text pre-processing in sentiment analysis,” *Procedia Computer Science*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[136] Dhanalakshmi, Bino, and Saravanan, “Opinion mining from student feedback data using supervised learning algorithms,” in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, IEEE, Mar. 2016, pp. 1–5. doi: 10.1109/ICBDSC.2016.7460390.
[137] Agarwal, “Data mining: Data mining concepts and techniques,” *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, pp. 203–207, 2014, doi: 10.1109/ICMIRA.2013.45.
[138] Farisi, Sibaroni, and Faraby, “Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier,” *Journal of Physics: Conference Series*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012024.
[139] Altrabsheh, Cocea, and Fallahkhair, “Sentiment Analysis: Towards a Tool for Analysing Real-Time Students Feedback,” in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, IEEE, Nov. 2014, pp. 419–423. doi: 10.1109/ICTAI.2014.70.
[140] Gupta and Lehal, “A survey of text mining techniques and applications,” *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60–76, 2009, doi: 10.4304/jetwi.1.1.60-76.
[141] Kennedy and Inkpen, “Sentiment classification of movie reviews using contextual valence shifters,” *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006, doi: 10.1111/j.1467-8640.2006.00277.x.
[142] Bringula, Ulfa, Miranda, and Atienza, “Text mining analysis on students’ expectations and anxieties towards data analytics course,” *Cogent Engineering*, vol. 9, no. 1, 2022, doi: 10.1080/23311916.2022.2127469.
[143] Pong-inwong and Songpan, “Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2177–2186, Aug. 2019, doi: 10.1007/s13042-018-0800-2.
[144] ONAN, “Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach,” *Computer Applications in Engineering Education*, vol. 29, no. 3, pp. 572–589, May 2021, doi: 10.1002/cae.22253.
[145] Kasthuriarachchy, De Zoysa, and Premaratne, “Enhanced bag-of-words model for phrase-

level sentiment analysis," *2014 14th International Conference on Advances in ICT for Emerging Regions, ICTer 2014 - Conference Proceedings*, no. December, pp. 210–214, 2014, doi: 10.1109/ICTER.2014.7083903.

[146] Hagenau, Liebmann, and Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decision Support Systems*, vol. 55, no. 3, pp. 685–697, 2013, doi: 10.1016/j.dss.2013.02.006.

[147] Khadjeh Nassirtoussi, Aghabozorgi, Ying Wah, and Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306–324, 2015, doi: 10.1016/j.eswa.2014.08.004.

[148] Brindha, Prabha, and Sukumaran, "The Comparison Of Term Based Methods Using Text Mining," *International Journal of Computer Science and Mobile Computing*, vol. 5, no. 9, pp. 112–116, 2016.

[149] Shiri, "Introduction to Modern Information Retrieval (2nd edition)," *Library Review*, vol. 53, no. 9, pp. 462–463, 2004, doi: 10.1108/00242530410565256.

[150] Nasim, Rajput, and Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, IEEE, Jul. 2017, pp. 1–6. doi: 10.1109/ICRIIS.2017.8002475.

[151] Pham, "Exploring the Effect of Word Embeddings and Bag-of-Words for Vietnamese Sentiment Analysis," in *Ubiquitous Intelligent Systems*, Karuppusamy, García Márquez, and Nguyen, Eds., Singapore: Springer Nature Singapore, 2022, pp. 595–605.

[152] Calandra, Achmad Nizar, Nur Fitriah, Abidin, and Wati, "Mining Student Feedback to Improve the Quality of Higher Education through Multi Label Classification, Sentiment Analysis, and Trend Topic," *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, vol. 6, pp. 0–5, 2019.

[153] Shitanshu Jain and Vishwakarma, "Analysis of Text Classification with various Term Weighting Schemes in Vector Space Model," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 10, pp. 390–393, 2020, doi: 10.35940/ijitee.d1938.0891020.

[154] Kompan and Bieliková, "News Article Classification Based on a Vector Representation Including Words ' Collocations News Article Classification Based on a Vector Representation Including Words ' Collocations," no. February, 2011, doi: 10.1007/978-3-642-23163-6.

[155] Martín-Valdivia, Martínez-Cámara, Perea-Ortega, and Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications*, vol. 40, no. 10, pp. 3934–3942, 2013, doi: 10.1016/j.eswa.2012.12.084.

[156] Bengio, Ducharme, and Vincent, "A Neural Probabilistic Language Model," in *Advances in Neural Information Processing Systems*, Leen, Dietterich, and Tresp, Eds., MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf

[157] Rezaeinia, Rahmani, Ghodsi, and Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, pp. 139–147, 2019, doi: 10.1016/j.eswa.2018.08.044.

[158] Mikolov, Chen, Corrado, and Dean, "Efficient Estimation of Word Representations in Vector Space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop*

Track Proceedings, pp. 1–12, Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781
[159] Alami, Meknassi, and En-nahnahi, “Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning,” *Expert Systems with Applications*, vol. 123, pp. 195–211, 2019, doi: 10.1016/j.eswa.2019.01.037.
[160] Joulin, Grave, Bojanowski, Douze, Jégou, and Mikolov, “FastText.zip: Compressing text classification models,” pp. 1–13, 2016, [Online]. Available: http://arxiv.org/abs/1612.03651
[161] Bhardwaj, Di, and Wei, *Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling*. Packt Publishing Ltd, 2018.
[162] Jeffrey Pennington, Socher, and Manning, “GloVe: Global Vectors forWord Representation,” *Proceedings ofthe 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 19, no. 5, pp. 417–425, 2014.
[163] D’Andrea, Ferri, Grifoni, and Guzzo, “Approaches, Tools and Applications for Sentiment Analysis Implementation,” *International Journal of Computer Applications*, vol. 125, no. 3, pp. 26–33, 2015, doi: 10.5120/ijca2015905866.
[164] Maimon and Rokach, *Data Mining and Knowledge*, vol. 40, no. 6. 2005.
[165] Bosch, “Sentiment Analysis : Incremental learning to build domain models,” 2014.
[166] Abbas, Ali, Jamali, Ali Memon, and Aleem Jamali, “Multinomial Naive Bayes Classification Model for Sentiment Analysis Classification for Sentiment Analysis View project Antenna Designing View project Multinomial Naive Bayes Classification Model for Sentiment Analysis,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, p. 62, 2019, [Online]. Available: https://www.researchgate.net/publication/334451164
[167] Kibriya, Frank, Pfahringer, and Holmes, “Multinomial Naive Bayes for Text Categorization Revisited.” 2004.
[168] Bahary, Sibaroni, and Mubarok, “Sentiment analysis of student responses related to information system services using Multinomial Naïve Bayes (Case study: Telkom University),” *Journal of Physics: Conference Series*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012046.
[169] Zhang, “Social Sentiment Analysis Using Classifiers and Ensemble Learning,” *Journal of Physics: Conference Series*, vol. 1237, no. 2, 2019, doi: 10.1088/1742-6596/1237/2/022193.
[170] Joachims, “Making Large-Scale SVM Learning Practical,” no. October 1999, 1999, doi: 10.17877/DE290R-5098.
[171] Na, Sui, Khoo, Chan, and Zhou, “Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews,” *Advances in Knowledge Organization*, vol. 9, pp. 49–54, 2004.
[172] Jaswanth, Muni, Kumar, Sudhan, Vijaya Kumar, and Rajagopalam, “Sentiment analysis using logistic regression algorithm,” *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 2081–2086, 2020, [Online]. Available: https://ejmcm.com/article_1947.html
[173] Xu, Davoine, and Denoeux, “Evidential logistic regression for binary SVM classifier calibration,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8764, pp. 49–57, 2014, doi: 10.1007/978-3-319-11191-9_6.
[174] Nath … Kaushik, “A Sentiment Analysis of Food Review using Logistic Regression,” vol. 2, no. 7, pp. 251–260, 2017, [Online]. Available: https://www.researchgate.net/publication/334654833
[175] Reddy, Sri, Reddy, and Shaik, “Sentimental Analysis using Logistic Regression,” vol. 11, no. July, pp. 36–40, 2021, doi: 10.9790/9622-1107023640.

[176] Al Amrani, Lazaar, and El Kadirp, “Random forest and support vector machine based hybrid approach to sentiment analysis,” *Procedia Computer Science*, vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.
[177] Karthika, Murugeswari, and Manoranjithem, “Sentiment Analysis of Social Media Network Using Random Forest Algorithm,” *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2019*, 2019, doi: 10.1109/INCOS45849.2019.8951367.
[178] Bahrawi, “Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based,” *Journal of Information Technology and Its Utilization*, vol. 2, no. 2, p. 29, 2019, doi: 10.30818/jitu.2.2.2695.
[179] Abbas, Salih, Hussein, Hussein, and Abdulwahhab, “Twitter Sentiment Analysis Using an Ensemble Majority Vote Classifier,” *Journal of Southwest Jiaotong University*, vol. 55, no. 1, 2020, doi: 10.35741/issn.0258-2724.55.1.9.
[180] Varshney, Sharma, and Yadav, “Sentiment analysis using ensemble classification technique,” *2020 IEEE Students’ Conference on Engineering and Systems, SCES 2020*, no. July, 2020, doi: 10.1109/SCES50439.2020.9236754.
[181] Yann Lecun, “Generalization and network design strategies,” *Connectionism in perspective Elsevier*, vol. 19, 1089.
[182] Ni and Cao, “Sentiment Analysis based on GloVe and LSTM-GRU,” *Chinese Control Conference, CCC*, vol. 2020-July, pp. 7492–7497, 2020, doi: 10.23919/CCC50068.2020.9188578.
[183] Pal, Ghosh, and Nag, “Sentiment Analysis in the Light of LSTM Recurrent Neural Networks,” *International Journal of Synthetic Emotions*, vol. 9, no. 1, pp. 33–39, 2018, doi: 10.4018/ijse.2018010103.
[184] Tripathi, “Sentiment Analysis of Nepali COVID19 Tweets Using NB , SVM AND LSTM,” vol. 03, no. 03, pp. 151–168, 2021.
[185] Gers, Schmidhuber, and Cummins, “Learning to Forget: Continual Prediction with LSTM,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000, doi: 10.1162/089976600300015015.
[186] Srinivas, Satyanarayana, Divakar, and Sirisha, “Sentiment Analysis using Neural Network and LSTM,” *IOP Conference Series: Materials Science and Engineering*, vol. 1074, no. 1, p. 012007, 2021, doi: 10.1088/1757-899x/1074/1/012007.
[187] Mrhar, Benhiba, Bourekkache, and Abik, “A Bayesian CNN-LSTM Model for Sentiment Analysis in Massive Open Online Courses MOOCs,” *International Journal of Emerging Technologies in Learning*, vol. 16, no. 23, pp. 216–232, 2021, doi: 10.3991/ijet.v16i23.24457.
[188] Gers and Schraudolph, “Learning Precise Timing with LSTM Recurrent Networks,” vol. 3, pp. 115–143, 2002.
[189] Long, Zhou, and Ou, “Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention,” *IEEE Access*, vol. 7, pp. 141960–141969, 2019, doi: 10.1109/ACCESS.2019.2942614.
[190] Anusha and Shashirekha, “BiLSTM-Sentiments Analysis in Code-Mixed Dravidian Languages,” 2021.
[191] Wang, Huang, and Zhou, “Sentiment analysis of ALBERT-BiLSTM model MOOC reviews via,” vol. 05008, pp. 1–8, 2021.
[192] Xu, Meng, Qiu, Yu, and Wu, “Sentiment analysis of comment texts based on BiLSTM,” *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.
[193] Alex Graves, “Hybrid Speech Recognition with Deep Bidirectional LSTM,” pp. 273–278,

2013.
[194] Huang, Zheng, Wang, and Zhu, “Sentiment Analysis of Chinese Text Based on CNN-BiLSTM Serial Hybrid Model,” *ACM International Conference Proceeding Series*, pp. 309–313, 2021, doi: 10.1145/3497623.3497673.
[195] Pawar and Ganguli, “Genetic Fuzzy System BT - Structural Health Monitoring Using Genetic Fuzzy Systems,” Pawar and Ganguli, Eds., London: Springer London, 2011, pp. 25–40. doi: 10.1007/978-0-85729-907-9_2.
[196] Hansen, “Analog forecasting of ceiling and visibility using fuzzy sets,” *Preprints of the 2nd Conference on Artificial Intelligence, American Meteorological Society*, pp. 1–7, 2000, [Online]. Available: http://chebucto.ns.ca/~bjarne/ams2000
[197] Zadeh, “The concept of a linguistic variable and its application to approximate reasoning-I,” *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975, doi: 10.1016/0020-0255(75)90036-5.
[198] Zadeh, “Fuzzy Logic,” *Computer*, vol. 21, no. 4, pp. 83–93, 1988.
[199] Chen and Tsao, “A description of the dynamic behavior of fuzzy systems,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 4, pp. 745–755, 1989, [Online]. Available: https://www.ptonline.com/articles/how-to-get-better-mfi-results
[200] Lagunes, Castillo, Valdez, Soria, and Melin, “Parameter Optimization for Membership Functions of Type-2 Fuzzy Controllers for Autonomous Mobile Robots Using the Firefly Algorithm - Fuzzy Information Processing,” in *Fuzzy Information Processing. NAFIPS 2018. Communications in Computer and Information Science*, 2018, pp. 569–579. doi: 10.1007/978-3-319-95312-0.
[201] Castillo and Melin, “Type-2 fuzzy logic systems,” *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 6, pp. 643–658, 1999, doi: 10.1007/978-3-642-28956-9_2.
[202] Lambat, Ayres, Maglaras, and Ferrag, “A mamdani type fuzzy inference system to calculate employee susceptibility to phishing attacks,” *Applied Sciences (Switzerland)*, vol. 11, no. 19, 2021, doi: 10.3390/app11199083.
[203] Singh and Thongam, “Mobile Robot Navigation Using Fuzzy Logic in Static Environments,” *Procedia Computer Science*, vol. 125, pp. 11–17, 2018, doi: 10.1016/j.procs.2017.12.004.
[204] Wang, *Computational Intelligence in Agile Manufacturing Engineering*. Elsevier Science Ltd., 2001. doi: 10.1016/b978-008043567-1/50016-4.
[205] Abdolkarimzadeh, Fazel Zarandi, and Castillo, “Interval Type II Fuzzy Rough Set Rule Based Expert System to Diagnose Chronic Kidney Disease BT - Fuzzy Information Processing,” in *Fuzzy Information Processing. NAFIPS 2018. Communications in Computer and Information Science*, Barreto and Coelho, Eds., Cham: Springer International Publishing, 2018, pp. 559–568.
[206] Arora, “A Decision-Making System for Corona Prognosis Using Fuzzy Inference System,” vol. 2, no. 4, pp. 344–354, 2021.
[207] Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965, doi: https://doi.org/10.1016/S0019-9958(65)90241-X.
[208] Shirvani, Abdollahi, Dusti, and Baghbani, “Fuzzy based stabilizer for upfc in multi machine power system,” no. September, 2015, doi: 10.7813/2075-4124.2014/6-4/A.26.
[209] Mamdani, “Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis,” *IEEE Transactions on Computers*, vol. C–26, no. 12, pp. 1182–1191, 1977, doi: 10.1109/TC.1977.1674779.
[210] Takagi and Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, 1985, doi: 10.1109/TSMC.1985.6313399.

[211] P. Jaques, Pursula, Niittymak, and Kosonen, “The impact of different approximate reasoning methods on fuzzy signal controllers,” no. June, 2002.
[212] Mogharreban and DiLalla, “Comparison of Defuzzification Techniques for Analysis of Non-interval Data,” in *NAFIPS 2006 - 2006 Annual Meeting of the North American Fuzzy Information Processing Society*, IEEE, Jun. 2006, pp. 257–260. doi: 10.1109/NAFIPS.2006.365418.
[213] Kitchenham and Charters, “Procedures for Performing Systematic Literature Reviews in Software Engineering,” *Keele University & Durham University, UK*, 2007.
[214] Higgins and Green, “Cochrane handbook for systematic reviews of interventions 4.2.6. The Cochrane collaboration,” *Text*, no. September, 2006.
[215] Tacconelli, *Systematic Reviews. CRD’s guidance for undertaking reviews in health care.* 2009.
[216] Santos, Rita, and Guerreiro, “Improving international attractiveness of higher education institutions based on text mining and sentiment analysis,” *International Journal of Educational Management*, vol. 32, no. 3, pp. 431–447, Apr. 2018, doi: 10.1108/IJEM-01-2017-0027.
[217] Canham, “Comparing Web 2.0 applications for peer feedback in language teaching: Google Docs, the Sakai VLE, and the Sakai Wiki,” *Writing & Pedagogy*, vol. 9, no. 3, pp. 429–456, Jan. 2018, doi: 10.1558/wap.32352.
[218] Cabada, “Mining of Educational Opinions with Deep Learning,” vol. 24, no. 11, pp. 1604–1626, 2018.
[219] Nguyen, Hong, Nguyen, and Nguyen, “Deep Learning versus Traditional Classifiers on Vietnamese Students’ Feedback Corpus,” in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, Nov. 2018, pp. 75–80. doi: 10.1109/NICS.2018.8606837.
[220] Rajput, Haider, and Ghani, “Lexicon-Based Sentiment Analysis of Teachers’ Evaluation,” *Applied Computational Intelligence and Soft Computing*, vol. 2016, pp. 1–12, 2016, doi: 10.1155/2016/2385429.
[221] Bird, Osheroff, Pettepher, Cutrer, and Carnahan, “Using Small Case-Based Learning Groups as a Setting for Teaching Medical Students How to Provide and Receive Peer Feedback,” *Medical Science Educator*, vol. 27, no. 4, pp. 759–765, Dec. 2017, doi: 10.1007/s40670-017-0461-x.
[222] Ion, Barrera-Corominas, and Tomàs-Folch, “Written peer-feedback to enhance students’ current and future learning,” *International Journal of Educational Technology in Higher Education*, vol. 13, no. 1, 2016, doi: 10.1186/s41239-016-0017-y.
[223] Cañabate, Nogué, Serra, and Colomer, “Supportive Peer Feedback in Tertiary Education: Analysis of Pre-Service Teachers’ Perceptions,” *Education Sciences*, vol. 9, no. 4, p. 280, Nov. 2019, doi: 10.3390/educsci9040280.
[224] Mercader, Ion, and Díaz-Vicario, “Factors influencing students’ peer feedback uptake: instructional design matters,” *Assessment & Evaluation in Higher Education*, pp. 1–12, Feb. 2020, doi: 10.1080/02602938.2020.1726283.
[225] Seroussi, Sharon, Peled, and Yaffe, “Reflections on peer feedback in disciplinary courses as a tool in pre-service teacher training,” *Cambridge Journal of Education*, vol. 49, no. 5, pp. 655–671, Sep. 2019, doi: 10.1080/0305764X.2019.1581134.
[226] Ostuzzi and Hoveskog, “Education for flourishing: an illustration of boundary object use, peer feedback and distance learning,” *International Journal of Sustainability in Higher Education*, vol. 21, no. 4, pp. 757–777, May 2020, doi: 10.1108/IJSHE-09-2019-0271.
[227] Adiguzel, “Examining a Web-Based Peer Feedback System in an Introductory Computer

Literacy Course," *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 13, no. 1, pp. 237–251, Jan. 2017, doi: 10.12973/eurasia.2017.00614a.

[228] Panadero and Alqassab, "An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading," *Assessment & Evaluation in Higher Education*, vol. 44, no. 8, pp. 1253–1278, Nov. 2019, doi: 10.1080/02602938.2019.1600186.

[229] Zhang, Yu, and Yuan, "Understanding Master's students' peer feedback practices from the academic discourse community perspective: A rethinking of postgraduate pedagogies," *Teaching in Higher Education*, vol. 25, no. 2, pp. 126–140, Feb. 2020, doi: 10.1080/13562517.2018.1543261.

[230] Han and Xu, "The development of student feedback literacy: the influences of teacher feedback on peer feedback," *Assessment & Evaluation in Higher Education*, vol. 45, no. 5, pp. 680–696, Jul. 2020, doi: 10.1080/02602938.2019.1689545.

[231] Friess and Goupee, "Using Continuous Peer Evaluation in Team-Based Engineering Capstone Projects: A Case Study," *IEEE Transactions on Education*, vol. 63, no. 2, pp. 82–87, May 2020, doi: 10.1109/TE.2020.2970549.

[232] Yu, Zhang, Zheng, Yuan, and Zhang, "Understanding student engagement with peer feedback on master's theses: a Macau study," *Assessment & Evaluation in Higher Education*, vol. 44, no. 1, pp. 50–65, Jan. 2019, doi: 10.1080/02602938.2018.1467879.

[233] Dickson, Harvey, and Blackwood, "Feedback, feedforward: evaluating the effectiveness of an oral peer review exercise amongst postgraduate students," *Assessment & Evaluation in Higher Education*, vol. 44, no. 5, pp. 692–704, Jul. 2019, doi: 10.1080/02602938.2018.1528341.

[234] Noroozi and Hatami, "The effects of online peer feedback and epistemic beliefs on students' argumentation-based learning," *Innovations in Education and Teaching International*, vol. 56, no. 5, pp. 548–557, Sep. 2019, doi: 10.1080/14703297.2018.1431143.

[235] Yu, "Learning from giving peer feedback on postgraduate theses: Voices from Master's students in the Macau EFL context," *Assessing Writing*, vol. 40, pp. 42–52, 2019, doi: 10.1016/j.asw.2019.03.004.

[236] López-Pellisa, Rotger, and Rodríguez-Gallego, "Collaborative writing at work: Peer feedback in a blended learning environment," *Education and Information Technologies*, no. 2016, 2020, doi: 10.1007/s10639-020-10312-2.

[237] Demirbilek, "Social media and peer feedback: What do students really think about using Wiki and Facebook as platforms for peer feedback?," *Active Learning in Higher Education*, vol. 16, no. 3, pp. 211–224, Nov. 2015, doi: 10.1177/1469787415589530.

[238] Potter, Englund, Charbonneau, MacLean, Newell, and Roll, "ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback," *Teaching & Learning Inquiry*, vol. 5, no. 2, p. 89, Sep. 2017, doi: 10.20343/teachlearninqu.5.2.8.

[239] Roman, Callison, Myers, and Berry, "Facilitating Authentic Learning Experiences in Distance Education: Embedding Research-Based Practices into an Online Peer Feedback Tool," *TechTrends*, vol. 64, no. 4, pp. 591–605, Jul. 2020, doi: 10.1007/s11528-020-00496-2.

[240] Hoo, Tan, and Deneen, "Negotiating self- and peer-feedback with the use of reflective journals: an analysis of undergraduates' engagement with feedback," *Assessment & Evaluation in Higher Education*, vol. 45, no. 3, pp. 431–446, Apr. 2020, doi: 10.1080/02602938.2019.1665166.

[241] Dingyloudi and Strijbos, "Just plain peers across social networks: Peer-feedback networks

nested in personal and academic networks in higher education," *Learning, Culture and Social Interaction*, vol. 18, no. April, pp. 86–112, 2018, doi: 10.1016/j.lcsi.2018.02.002.

[242] Kamp, van Berkel, Popeijus, Leppink, Schmidt, and Dolmans, "Midterm peer feedback in problem-based learning groups: the effect on individual contributions and achievement," *Advances in Health Sciences Education*, vol. 19, no. 1, pp. 53–69, Mar. 2014, doi: 10.1007/s10459-013-9460-x.

[243] O'Neill … Brennan, "Introducing a scalable peer feedback system for learning teams," *Assessment & Evaluation in Higher Education*, vol. 44, no. 6, pp. 848–862, Aug. 2019, doi: 10.1080/02602938.2018.1526256.

[244] Jiranantanagorn and Shen, "Sentiment analysis and visualisation in a backchannel system," in *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*, Parker, Ed., New York, New York, USA: ACM Press, 2016, pp. 353–357. doi: 10.1145/3010915.3010992.

[245] Abdulsalami … Ekoja, "Sentiment analysis of students' perception on the use of smartphones: A cross sectional study," in *2017 Second International Conference on Informatics and Computing (ICIC)*, IEEE, Nov. 2017, pp. 1–5. doi: 10.1109/IAC.2017.8280625.

[246] Balachandran and Kirupananda, "Online reviews evaluation system for higher education institution: An aspect based sentiment analysis tool," in *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, IEEE, Dec. 2017, pp. 1–7. doi: 10.1109/SKIMA.2017.8294118.

[247] Hew, Hu, Qiao, and Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Computers & Education*, vol. 145, p. 103724, Feb. 2020, doi: 10.1016/j.compedu.2019.103724.

[248] İskender and Batı, "Comparing Turkish Universities Entrepreneurship and Innovativeness Index's Rankings with Sentiment Analysis Results on Social Media," *Procedia - Social and Behavioral Sciences*, vol. 195, pp. 1543–1552, Jul. 2015, doi: 10.1016/j.sbspro.2015.06.457.

[249] Esparza … de Jesus Nava, "A Sentiment Analysis Model to Analyze Students Reviews of Teacher Performance Using Support Vector Machines," in *Advances in Intelligent Systems and Computing*, 2018, pp. 157–164. doi: 10.1007/978-3-319-62410-5_19.

[250] Ortigosa, Martín, and Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, no. 1, pp. 527–541, Feb. 2014, doi: 10.1016/j.chb.2013.05.024.

[251] Troisi, Grimaldi, Loia, and Maione, "Big data and sentiment analysis to highlight decision behaviours: a case study for student population," *Behaviour & Information Technology*, vol. 37, no. 10–11, pp. 1111–1128, Nov. 2018, doi: 10.1080/0144929X.2018.1502355.

[252] Gawron, Cheng, and Meinel, "Automatic Vulnerability Classification Using Machine Learning," in *9th International Conference on Risks and Security of Internet and Systems, CRiSIS 2014*, Lopez, Ray, and Crispo, Eds., in Lecture Notes in Computer Science, vol. 8924. Cham: Springer International Publishing, 2015, pp. 131–147. doi: 10.1007/978-3-319-17127-2.

[253] Kiet Van, Vu Duc, V., Phu X., Tham T. H, and Ngan Luu-Thuy, "UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, Nov. 2018, pp. 19–24. doi: 10.1109/KSE.2018.8573337.

[254] Lange, Schmitt, and Wanka, "Towards a Better Understanding of the Local Attractor in Particle Swarm Optimization: Speed and Solution Quality," in *Lecture Notes in Computer*

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, pp. 90–99. doi: 10.1007/978-3-319-11298-5_10.

[255] Cobos, Jurado, and Blazquez-Herranz, "A Content Analysis System That Supports Sentiment Analysis for Subjectivity and Polarity Detection in Online Courses," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 14, no. 4, pp. 177–187, Nov. 2019, doi: 10.1109/RITA.2019.2952298.

[256] Spatiotis, Perikos, Mporas, and Paraskevas, "Sentiment Analysis of Teachers Using Social Information in Educational Platform Environments," *International Journal on Artificial Intelligence Tools*, vol. 29, no. 02, p. 2040004, Mar. 2020, doi: 10.1142/S0218213020400047.

[257] Janda, Pawar, Du, and Mago, "Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation," *IEEE Access*, vol. 7, pp. 108486–108503, 2019, doi: 10.1109/ACCESS.2019.2933354.

[258] Lwin, Oo, Ye, Kyaw Lin, Aung, and Paing Ko, "Feedback Analysis in Outcome Base Education Using Machine Learning," in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, IEEE, Jun. 2020, pp. 767–770. doi: 10.1109/ECTI-CON49241.2020.9158328.

[259] Al Bashaireh, Sabeeh, and Zohdy, "Towards a new indicator for evaluating universities based on twitter sentiment analysis," *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, pp. 1398–1404, 2019, doi: 10.1109/CSCI49370.2019.00261.

[260] Soe and Soe, "Domain Oriented Aspect Detection for Student Feedback System," in *2019 International Conference on Advanced Information Technologies (ICAIT)*, IEEE, Nov. 2019, pp. 90–95. doi: 10.1109/AITC.2019.8921372.

[261] Katragadda, Ravi, Kumar, and Lakshmi, "Performance Analysis on Student Feedback using Machine Learning Algorithms," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020, pp. 1161–1163. doi: 10.1109/ICACCS48705.2020.9074334.

[262] Sultana, Sultana, Yadav, and AlFayez, "Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE, 2018.

[263] Nguyen, Nguyen, and Nguyen, "Variants of Long Short-Term Memory for Sentiment Analysis on Vietnamese Students' Feedback Corpus," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, Nov. 2018, pp. 306–311. doi: 10.1109/KSE.2018.8573351.

[264] Sangeetha and Prabha, "Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 4117–4126, Mar. 2021, doi: 10.1007/s12652-020-01791-9.

[265] Akhtar, "An interactive multi-agent reasoning model for sentiment analysis: a case for computational semiotics," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 3987–4004, 2020, doi: 10.1007/s10462-019-09785-6.

[266] Cunningham-Nelson, Baktashmotlagh, and Boles, "Visualizing Student Opinion through Text Analysis," *IEEE Transactions on Education*, vol. 62, no. 4, pp. 305–311, 2019, doi: 10.1109/TE.2019.2924385.

[267] Karunya, Aarthy, Karthika, and Jegatha Deborah, "Analysis of Student Feedback and Recommendation to Tutors," *2020 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1579–1583, 2020, doi: 10.1109/ICCSP48568.2020.9182270.

[268] Hujala, Knutas, Hynninen, and Arminen, “Improving the quality of teaching by utilising written student feedback: A streamlined process,” *Computers and Education*, vol. 157, no. June, p. 103965, 2020, doi: 10.1016/j.compedu.2020.103965.
[269] Wang and Zhang, “Topic Sentiment Analysis in Online Learning Community from College Students,” *Journal of Data and Information Science*, vol. 5, no. 2, pp. 33–61, Apr. 2020, doi: 10.2478/jdis-2020-0009.
[270] Srinivas and Rajendran, “Topic-based knowledge mining of online student reviews for strategic planning in universities,” *Computers and Industrial Engineering*, vol. 128, no. July 2018, pp. 974–984, 2019, doi: 10.1016/j.cie.2018.06.034.
[271] Rani and Kumar, “A Sentiment Analysis System to Improve Teaching and Learning,” *Computer*, vol. 50, no. 5, pp. 36–43, May 2017, doi: 10.1109/MC.2017.133.
[272] Aung and Myo, “Sentiment analysis of students’ comment using lexicon based approach,” in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, IEEE, May 2017, pp. 149–154. doi: 10.1109/ICIS.2017.7959985.
[273] Liu, Qi, Ma, and Yang, “Sentiment Analysis by Exploring Large Scale Web-based Chinese Short Text,” in *International Conference on Computer Science and Application Engineering (CSAE)*, in DEStech Transactions on Computer Science and Engineering, vol. 190. 439 DUKE STREET, LANCASTER, PA 17602-4967 USA: DESTECH PUBLICATIONS, INC, 2017, pp. 930–939.
[274] Jiménez, Casanova, Nunes, and Finamore, “CourseObservatory: Sentiment analysis of comments in course surveys,” *Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019*, pp. 176–178, 2019, doi: 10.1109/ICALT.2019.00053.
[275] Hixson, “Reactions vs. Reality: Using Sentiment Analysis to Measure University Students’ Responses to Learning ArcGIS,” *Journal of Map and Geography Libraries*, vol. 15, no. 2–3, pp. 263–276, 2019, doi: 10.1080/15420353.2020.1719266.
[276] da Silva and Bernardino, *Combining Sentiment Analysis Scores to Improve Accuracy of Polarity Classification in MOOC Posts*, vol. 1. Springer International Publishing, 2019. doi: 10.1007/978-3-030-30241-2.
[277] Lundqvist, Liyanagunawardena, and Starkey, “Evaluation of student feedback within a MOOC using sentiment analysis and target groups,” *International Review of Research in Open and Distance Learning*, vol. 21, no. 3, pp. 140–156, 2020, doi: 10.19173/irrodl.v21i3.4783.
[278] Selmi, Hage, and A Ï Meur, “Latent semantic analysis for privacy preserving peer feedback,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8924, pp. 100–115, 2015, doi: 10.1007/978-3-319-17127-2_7.
[279] Pinargote-Ortega, Bowen-Mendoza, Meza, and Ventura, “Sentiment Analysis Techniques for Peer Feedback: A Review,” in *2023 Ninth International Conference on eDemocracy & eGovernment (ICEDEG)*, IEEE, Apr. 2023, pp. 1–8. doi: 10.1109/ICEDEG58167.2023.10122085.
[280] Darvishi, Khosravi, Sadiq, and Gašević, “Incorporating <scp>AI</scp> and learning analytics to build trustworthy peer assessment systems,” *British Journal of Educational Technology*, vol. 53, no. 4, pp. 844–875, Jul. 2022, doi: 10.1111/bjet.13233.
[281] Ashenafi, “Peer-assessment in higher education – twenty-first century practices, challenges and the way forward,” *Assessment and Evaluation in Higher Education*, vol. 42, no. 2, pp. 226–251, 2017, doi: 10.1080/02602938.2015.1100711.
[282] Ravi and Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and

applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015, doi: 10.1016/j.knosys.2015.06.015.
[283] Amershi … Zimmermann, "Software Engineering for Machine Learning Applications," *Icse*, vol. 2020, pp. 1–10, 2019, doi: 10.1109/ICSE-SEIP.2019.00042.
[284] Gielen and De Wever, "Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content," *Computers in Human Behavior*, vol. 52, pp. 315–325, 2015, doi: 10.1016/j.chb.2015.06.019.
[285] Su and Markert, "Eliciting subjectivity and polarity judgements on word senses," no. August, pp. 42–50, 2008, doi: 10.3115/1611628.1611635.
[286] Maks and Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, vol. 53, no. 4, pp. 680–688, 2012, doi: 10.1016/j.dss.2012.05.025.
[287] Russell and Norvig, *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.
[288] Hossein Kazemi, Shiri, Marti, and Majnooni-Heris, "Assessing temporal data partitioning scenarios for estimating reference evapotranspiration with machine learning techniques in arid regions," *Journal of Hydrology*, vol. 590, no. June, p. 125252, 2020, doi: 10.1016/j.jhydrol.2020.125252.
[289] Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Computational Statistics and Data Analysis*, vol. 53, no. 11, pp. 3735–3745, 2009, doi: 10.1016/j.csda.2009.04.009.
[290] Gibaja and Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys*, vol. 47, no. 3, 2015, doi: 10.1145/2716262.
[291] Jeni, Cohn, and De La Torre, "Facing imbalanced data - Recommendations for the use of performance metrics," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, no. September, pp. 245–251, 2013, doi: 10.1109/ACII.2013.47.
[292] Monllaó Olivé, Huynh, Reynolds, Dougiamas, and Wiese, "A supervised learning framework: using assessment to identify students at risk of dropping out of a MOOC," *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 9–26, 2020, doi: 10.1007/s12528-019-09230-1.
[293] Alves … Guimarães, "Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs," *Computers in Biology and Medicine*, vol. 132, no. December 2020, 2021, doi: 10.1016/j.compbiomed.2021.104335.
[294] Kontogiannis, Bargiotas, and Daskalopulu, "Fuzzy control system for smart energy management in residential buildings based on environmental data," *Energies*, vol. 14, no. 3, 2021, doi: 10.3390/en14030752.
[295] Chang and Lin, "LIBSVM: A Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011, doi: 10.1145/1961189.1961199.
[296] Mohey, "Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016, doi: 10.14569/ijacsa.2016.070134.
[297] Pinargote Ortega, Mendoza Bowen, Hormaza, and Ventura Soto, "Accuracy' measures of sentiment analysis algorithms for spanish corpus generated in peer assessment," *ACM International Conference Proceeding Series*, 2020, doi: 10.1145/3410352.3410838.
[298] Martínez-Cámara, Martín-Valdivia, and Ureña-López, "Opinion Classification Techniques Applied to a Spanish Corpus," in *Natural Language Processing and Information Systems*, Muñoz, Montoyo, and Métais, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 169–176.

[299] Wilson, Wiebe, and Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2005, pp. 347–354. doi: 10.3115/1220575.1220619.
[300] Pinargote-Ortega, Bowen-Mendoza, Meza, and Ventura, “Peer assessment using soft computing techniques,” *Journal of Computing in Higher Education*, vol. 33, no. 3, pp. 684–726, 2021, doi: 10.1007/s12528-021-09296-w.
[301] Cardellino, “Spanish Billion Word Corpus and Embeddings.” 2019. [Online]. Available: https://crscardellino.github.io/SBWCE/
[302] Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[303] García and Herrera, “An extension on ‘statistical comparisons of classifiers over multiple data sets’ for all pairwise comparisons,” *Journal of Machine Learning Research*, vol. 9, no. December 2008, pp. 2677–2694, 2008.
[304] Khatoon and Jones, “Flipped small group classes and peer marking: incentives, student participation and performance in a quasi-experimental approach,” *Assessment and Evaluation in Higher Education*, vol. 47, no. 6, pp. 910–927, 2022, doi: 10.1080/02602938.2021.1981823.
[305] Gaynor, “Peer review in the classroom: student perceptions, peer feedback quality and the role of assessment,” *Assessment and Evaluation in Higher Education*, vol. 45, no. 5, pp. 758–775, 2020, doi: 10.1080/02602938.2019.1697424.
[306] Nicol and McCallum, “Making internal feedback explicit: exploiting the multiple comparisons that occur during peer review,” *Assessment and Evaluation in Higher Education*, vol. 47, no. 3, pp. 424–443, 2022, doi: 10.1080/02602938.2021.1924620.
[307] Huisman, Saab, van Driel, and van den Broek, “Peer feedback on academic writing: undergraduate students’ peer feedback role, peer feedback perceptions and essay performance,” *Assessment and Evaluation in Higher Education*, vol. 43, no. 6, pp. 955–968, 2018, doi: 10.1080/02602938.2018.1424318.
[308] Purchase and Hamer, “Peer-review in practice: eight years of Aropä,” *Assessment and Evaluation in Higher Education*, vol. 43, no. 7, pp. 1146–1165, 2018, doi: 10.1080/02602938.2018.1435776.
[309] Latifi, Noroozi, and Talaee, “Peer feedback or peer feedforward? Enhancing students’ argumentative peer learning processes and outcomes,” *British Journal of Educational Technology*, vol. 52, no. 2, pp. 768–784, 2021, doi: 10.1111/bjet.13054.
[310] Voelkel, Varga-Atkins, and Mello, *Students tell us what good written feedback looks like*, vol. 10, no. 5. 2020. doi: 10.1002/2211-5463.12841.
[311] Leekwijck and Kerre, “Defuzziÿcation : criteria and classiÿcation,” *Fuzzy Sets and Systems*, vol. 108, pp. 159–178, 1999.
[312] Carless, “From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes,” *Active Learning in Higher Education*, vol. 23, no. 2, pp. 143–153, 2022, doi: 10.1177/1469787420945845.
[313] Wang, Ai, Liang, and Liu, “Toward motivating participants to assess peers’ work more fairly: Taking programing language learning as an example,” *Journal of Educational Computing Research*, vol. 52, no. 2, pp. 180–198, 2015, doi: 10.1177/0735633115571303.
[314] Mulder, Pearce, and Baik, “Peer review in higher education: Student perceptions before and after participation,” *Active Learning in Higher Education*, vol. 15, no. 2, pp. 157–171, 2014, doi: 10.1177/1469787414527391.
[315] García-Martínez, Cerezo, Bermúdez, and Romero, “Improving essay peer grading

accuracy in massive open online courses.pdf.” 2018.
[316] Luo, Robinson, and Park, “Peer grading in a MOOC: Reliability, validity, and perceived effects,” *Journal of Asynchronous Learning Network*, vol. 18, no. 2, pp. 1–14, 2014, doi: 10.24059/olj.v18i2.429.
[317] Formanek, Wenger, Buxner, Impey, and Sonam, “Insights about large-scale online peer assessment from an analysis of an astronomy MOOC,” *Computers and Education*, vol. 113, pp. 243–262, 2017, doi: 10.1016/j.compedu.2017.05.019.
[318] Knight, Leigh, Davila, Martin, and Krix, “Calibrating assessment literacy through benchmarking tasks,” *Assessment and Evaluation in Higher Education*, vol. 44, no. 8, pp. 1121–1132, 2019, doi: 10.1080/02602938.2019.1570483.
[319] Martha, Wang, Jiang, and Varghese, *Interactive tool to find focal spots in human computer interfaces in ecommerce: Ecommerce consumer analytics tool (eCCAT)*, vol. 529. 2015. doi: 10.1007/978-3-319-21383-5_27.
[320] Song, Hu, Gehringer, Morris, Kidd, and Ringleb, “Toward better training in peer assessment: Does calibration help?,” *CEUR Workshop Proceedings*, vol. 1633, 2016.
[321] Isaacs, Miller, Hu, Johnson, and Weber, “Inter-rater reliability of web-based calibrated peer review within a pharmacy curriculum,” *American Journal of Pharmaceutical Education*, vol. 84, no. 4, pp. 427–431, 2020, doi: 10.5688/ajpe7583.
[322] Boudria, Lafifi, and Bordjiba, “Collaborative calibrated peer assessment in massive open online courses,” *International Journal of Distance Education Technologies*, vol. 16, no. 1, pp. 76–102, 2018, doi: 10.4018/IJDET.2018010105.
[323] Saterbak, Moturu, and Volz, “Using a Teaching Intervention and Calibrated Peer Review™ Diagnostics to Improve Visual Communication Skills,” *Annals of Biomedical Engineering*, vol. 46, no. 3, pp. 513–524, 2018, doi: 10.1007/s10439-017-1946-x.
[324] Culver, Bowman, Youngerman, Jang, and Just, “Promoting equitable achievement in STEM: lab report writing and online peer review,” *Journal of Experimental Education*, vol. 90, no. 1, pp. 23–45, 2022, doi: 10.1080/00220973.2020.1799315.
[325] Badea and Popescu, “Analyzing the Validity of the Peer Assessment Process in a Project-Based Learning Scenario:Preliminary Results,” *2020 24th International Conference on System Theory, Control and Computing, ICSTCC 2020 - Proceedings*, pp. 831–836, 2020, doi: 10.1109/ICSTCC50638.2020.9259671.
[326] Yang, Luo, Song, and Yin, “Enhancing Peer Assessment Validity with Engagement Behaviors: A Structural Equation Modeling Approach,” *TALE 2021 - IEEE International Conference on Engineering, Technology and Education, Proceedings*, pp. 1163–1166, 2021, doi: 10.1109/TALE52509.2021.9678860.
[327] Reilly, Stafford, Williams, and Corliss, “Evaluating the validity and applicability of automated essay scoring in two massive open online courses,” *International Review of Research in Open and Distance Learning*, vol. 15, no. 5, pp. 83–98, 2014, doi: 10.19173/irrodl.v15i5.1857.
[328] Paré and Joordens, “Peering into large lectures: Examining peer and expert mark agreement using peerScholar, an online peer assessment tool,” *Journal of Computer Assisted Learning*, vol. 24, no. 6, pp. 526–540, 2008, doi: 10.1111/j.1365-2729.2008.00290.x.
[329] Gogus, “Bloom’s Taxonomy of Learning Objectives,” in *Encyclopedia of the Sciences of Learning*, Seel, Ed., Boston, MA: Springer US, 2012, pp. 469–473. doi: 10.1007/978-1-4419-1428-6_141.
[330] Black and Wiliam, “Classroom assessment and pedagogy,” *Assessment in Education: Principles, Policy and Practice*, vol. 25, no. 6, pp. 551–575, 2018, doi:

10.1080/0969594X.2018.1441807.
[331] Thurlings, Vermeulen, Bastiaens, and Stijnen, “Understanding feedback: A learning theory perspective,” *Educational Research Review*, vol. 9, pp. 1–15, 2013, doi: 10.1016/j.edurev.2012.11.004.
[332] Esterhazy and Damşa, “Unpacking the feedback process: an analysis of undergraduate students’ interactional meaning-making of feedback comments,” *Studies in Higher Education*, vol. 44, no. 2, pp. 260–274, 2019, doi: 10.1080/03075079.2017.1359249.

APÉNDICES

Apéndice A. Percepción de los estudiantes sobre la usabilidad del prototipo de evaluación entre pares

Se realizaron pruebas de usuario final con la interacción de 79 estudiantes de la asignatura de fundamentos de ingeniería de software impartida en escenario de educación virtual asincrónica en el periodo académico mayo-septiembre 2021 ante la pandemia COVID-19. Se dio la capacitación mediante Google Meet, explicando cada una de las interfaces del usuario estudiante.

Posteriormente se aplicó el cuestionario, para conocer la percepción de los estudiantes sobre la usabilidad del prototipo, en el que participaron 47 estudiantes.

Los resultados exteriorizaron lo siguiente:

En la interfaz de envío de tarea, el 2% de los estudiantes tuvo muchas dificultades, el 15% tuvo pocas dificultades y el 83% no tuvo ninguna dificultad (Figura A1).

En la interfaz de evaluación de tarea, el 2% de los estudiantes tuvo muchas dificultades, el 30% tuvo pocas dificultades y el 68% no tuvo ninguna dificultad (Figura A2).

En la interfaz de evaluación de la calidad de la evaluación, el 2% de los estudiantes tuvo muchas dificultades, el 28% tuvo pocas dificultades y el 70% no tuvo ninguna dificultad (Figura A3).

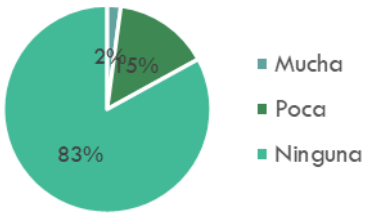


Figura A1. Dificultades en la interfaz de enviar tarea

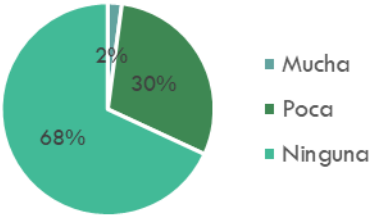


Figura A2. Dificultades en la interfaz de evaluación de tarea

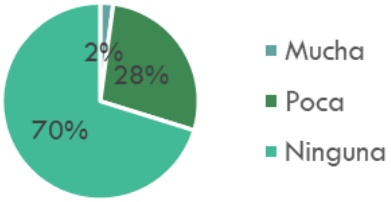


Figura A3. Dificultades en la interfaz de evaluación de revisión del evaluador

También los estudiantes manifestaron sugerencias, se destacan las siguientes:

- Que se puedan guardar la retroalimentación, aunque no estén todas completadas.
- Aumentar el tamaño de las letras.
- Que sea un poco más amigable la interfaz.
- Mejoras en la interfaz de inicio de sesión.
- Añadir un botón de retroceder, mejorar en la navegación de las páginas
- Cambiar el formato de horario.
- Que se distinga cuando se ha ejecutado la evaluación.
- Visualizar el nombre de la tarea que se está evaluando, entre otros.

Se colige que fueron pocas las dificultades encontradas en cada una de las interfaces. En base a las dificultades encontradas y sugerencias dada por los estudiantes se realizaron los ajustes necesarios al prototipo.

Apéndice B. Artículo: Sentiment Analysis of Peer Feedback in Higher Education. Aceptado para publicación en AIP Conference Proceedings

UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

INSTITUTO DE INVESTIGACIÓN

4 May 2023

Portoviejo, Ecuador

ACCEPTANCE LETTER

Article Title: Sentiment Analysis of Peer Feedback in Higher Education.

Corresponding Author: Maricela Pinargote-Ortega

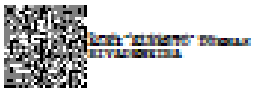
Authors: Maricela Pinargote-Ortega, Lorena Bowen-Mendoza, Jaime Meza, Sebastián Ventura

We are pleased to inform you that your manuscript referenced above has been accepted for publication in AIP Conference Proceedings (ISSN: 0094243X, 15517616) which has collaborated with VI International Scientific Convention 2022 of Universidad Técnica de Manabí in publishing the research works.

Your accepted manuscript will now be transferred to our AIP Publishing and work will begin on creation of the proof. If we need any additional information to create the proof, we will let you know. If not, you will be contacted again in the next few days with a request to approve the proof and to complete a form (License to publish agreement for conference proceedings) that is required for publication.

Many thanks for submitting your fine paper to VI International Scientific Convention 2022 of UTM and to AIP Conference Proceedings. We look forward to receiving additional papers from you in the future events.

With kind regards,

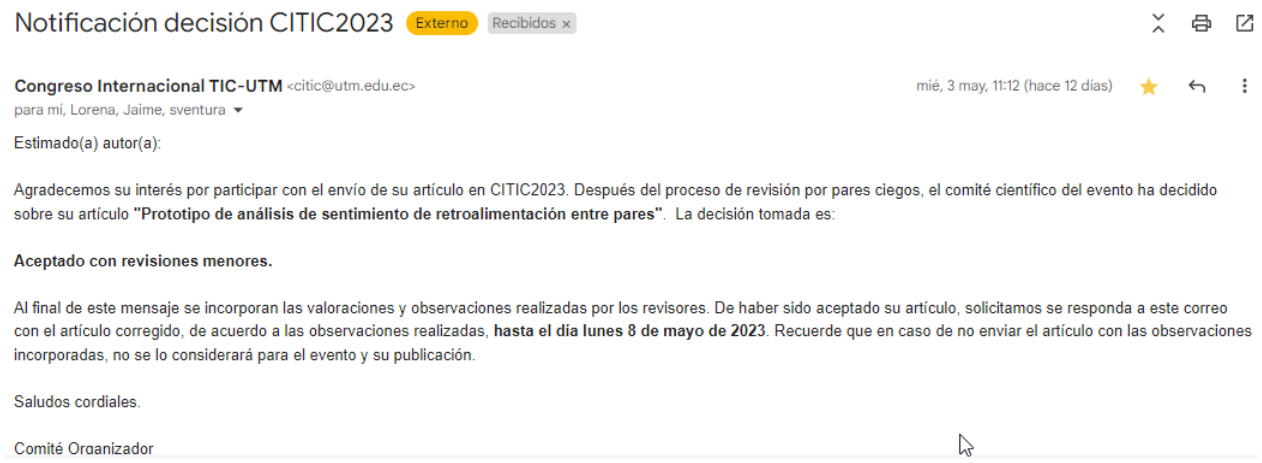


Prof. Dr. Alex Alberto Dueñas-Rivadeneira, Ph.D.,

Editor-in-Chief

AIP Conference Proceedings – VI CCI UTM 2022

Apéndice C. Artículo: Peer Feedback Sentiment Analysis Prototype



Notificación decisión CITIC2023 Externo Recibidos ×

Congreso Internacional TIC-UTM <citic@utm.edu.ec> mié, 3 may, 11:12 (hace 12 días)
para mí, Lorena, Jaime, sventura

Estimado(a) autor(a):

Agradecemos su interés por participar con el envío de su artículo en CITIC2023. Después del proceso de revisión por pares ciegos, el comité científico del evento ha decidido sobre su artículo **"Prototipo de análisis de sentimiento de retroalimentación entre pares"**. La decisión tomada es:

Aceptado con revisiones menores.

Al final de este mensaje se incorporan las valoraciones y observaciones realizadas por los revisores. De haber sido aceptado su artículo, solicitamos se responda a este correo con el artículo corregido, de acuerdo a las observaciones realizadas, **hasta el día lunes 8 de mayo de 2023**. Recuerde que en caso de no enviar el artículo con las observaciones incorporadas, no se lo considerará para el evento y su publicación.

Saludos cordiales.

Comité Organizador

Apéndice D. Ponencia: Tendencias de investigación en la evaluación por pares basada en comentarios. Revisión literaria 2014-2018

UNIVERSIDAD DE CÓRDOBA

CONGRESO INTERNACIONAL Y MULTIDISCIPLINAR DE INVESTIGADORES EN FORMACIÓN

Manta - Ecuador, 18 al 21 de febrero de 2019

Confieren el siguiente **Certificado** a:

JENMER MARICELA PINARGOTE ORTEGA

Por su participación en calidad de PONENTE con el tema:

Tendencias de investigación en la evaluación por pares basada en comentarios. Revisión literaria 2014-2018

Congreso Internacional y Multidisciplinar de Investigadores en Formación, organizado por la Universidad de Córdoba y el CIDE el 18, 19, 20 y 21 de febrero, con una duración de 40 horas académicas, realizado en la ciudad de Manta.

Manta, febrero 2019

Dr. Francisco Sánchez Tortora
Presidente del Congreso Internacional

Dra. Rocío Serrano Rodríguez,
Secretaria Académica del Congreso Internacional

CIDE

Lic. Max Oswaldo Alvarado
Director Central - CIDE

ISBN: 978-9942-802-07-1

Apéndice E. Ponencia: Especificación de requerimientos para un sistema de evaluación entre pares en la Universidad Técnica de Manabí

Apéndice F. Ponencia: Estudio de técnicas de minería de datos educativas para realizar predicciones en la educación superior

Segunda Convención Científica Internacional de la Universidad Técncica de Manabí

ESTUDIO DE TÉCNICAS DE MINERÍA DE DATOS EDUCATIVAS PARA REALIZAR PREDICCIONES EN LA EDUCACIÓN SUPERIOR.

Pinargote Ortega Jenmer Maricela[1]*, Bowen Mendoza Lorena Elizabeth[1], Cruz Felipe Marely Del Rosario[1], Demera Ureta Gabriel[1]
[1]Facultad de Ciencias Informáticas, Universidad Técnica de Manabí, Portoviejo, Ecuador
*Autor de correspondencia: jmpinargote@utm.edu.ec

Resumen

Este trabajo plantea explorar datos históricos de información útil y oculta a partir de bases de datos de estudiantes para aplicar técnicas de minería de datos en la predicción temprana del fracaso escolar con mayor precisión; transformando una gran cantidad de información en una riqueza de previsibilidad, persistencia y beneficios. El fracaso escolar universitario ha sido objeto de estudio e intervención desde algunas décadas. En Latinoamérica y el Caribe, se han reportado altas tasas de abandono, la deserción estudiantil es uno de los más grandes problemas que aborda la mayoría de las instituciones de educación superior de toda Latinoamérica. El Ecuador no está aislado de este problema, el ingreso a una carrera universitaria se establece mediante el puntaje obtenido en el Examen Nacional para la Educación Superior (ENES); lo que ha conducido a que los estudiantes ingresen en carreras no deseadas con escasa motivación escolar y bajas expectativas de éxito. La Universidad Técnica de Manabí (UTM), situada en el Ecuador, desde el 2008 hasta la actualidad almacena información de estudiantes mediante sistemas informáticos. Uno de los problemas que se presentan en la Universidad es el alto porcentaje de fracaso escolar, la directiva no ha podido hacer mucho al respecto; ya que no se realiza un diagnóstico precoz, ni se da el seguimiento adecuado; existen varios factores que han influido como personales, familiares, sociales, económicos, académicos, entre otros; hasta la fecha es desconocido el impacto o aportación que tiene cada uno de estos factores en la predicción del posible fracaso escolar de un estudiante. Por tales razones es necesario aplicar técnicas de minería de datos para predecir a tiempo a estudiantes que se encuentra en situación de riesgo de fracaso escolar e identificar los factores que lo suscita.

Palabras clave: predicción, técnicas de minería de datos, UTM

Apéndice G. Ponencia: Revisión de sistemas de evaluación por pares en instituciones educativas

UNIVERSIDAD TÉCNICA DE MANABÍ

CONFIERE EL PRESENTE

CERTIFICADO A

PINARGOTE ORTEGA JENMER MARICELA
BOWEN MENDOZA LORENA ELIZABETH

Por haber participado con el tema:

REVISIÓN DE SISTEMAS DE EVALUACIÓN POR PARES EN INSTITUCIONES EDUCATIVAS

En la **SEGUNDA CONVENCIÓN CIENTÍFICA INTERNACIONAL DE LA UTM 2018**
Realizadas en la ciudad de Portoviejo-Ecuador desde el 17 al 19 de octubre de 2018.

Vicente Véliz Briones, Ph D.
RECTOR UTM
PRESIDENTE DE HONOR CCIUTM

Apéndice H. Ponencia: Análisis de sentimiento de la retroalimentación entre pares como alternativa para la evaluación formativa y sumativa

UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

CCIUTM 2022
CONVENCIÓN CIENTÍFICA INTERNACIONAL
UNIVERSIDAD TÉCNICA DE MANABÍ

V CONGRESO INTERNACIONAL DE CIENCIAS INFORMÁTICAS

CONFIERE EL PRESENTE

CERTIFICADO

A

PINARGOTE ORTEGA JENMER MARICELA
BOWEN MENDOZA LORENA ELIZABETH
MEZA HORMAZA JAIME ALCIDES

Por su participación como PONENTE con el tema:

ANÁLISIS DE SENTIMIENTO DE LA RETROALIMENTACIÓN ENTRE PARES COMO ALTERNATIVA PARA LA EVALUACIÓN FORMATIVA Y SUMATIVA

En el marco de la SEXTA CONVENCIÓN CIENTÍFICA INTERNACIONAL DE LA UTM 2022 realizada en la ciudad de Portoviejo - Ecuador del 26 al 28 de octubre de 2022.

ING. SANTIAGO QUIROZ FERNÁNDEZ, PHD.
RECTOR UTM
PRESIDENTE CCIUTM

ING. MARA MOLINA DE LOZANO, PH.D.
VICERRECTORA ACADÉMICA UTM
VICEPRESIDENTA PRIMERA

ALEX DUEÑAS RIVADENEIRA, PHD.
COORDINADOR CIENTÍFICO

LIC. MONSERRATE RUIZ CEDEÑO, PH.D.
SECRETARIA EJECUTIVA

Apéndice I. Ponencia: Técnicas de análisis de sentimiento para la retroalimentación entre pares: una revisión

UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

CCIUTM 2022
CONVENCIÓN CIENTÍFICA INTERNACIONAL
UNIVERSIDAD TÉCNICA DE MANABÍ

V CONGRESO INTERNACIONAL DE CIENCIAS INFORMÁTICAS

CONFIERE EL PRESENTE

CERTIFICADO

A

PINARGOTE ORTEGA JENMER MARICELA
BOWEN MENDOZA LORENA ELIZABETH
MEZA HORMAZA JAIME ALCIDES

Por su participación como PONENTE con el tema:

TÉCNICAS DE ANÁLISIS DE SENTIMIENTOS PARA LA RETROALIMENTACIÓN ENTRE PARES: UNA REVISIÓN

En el marco de la SEXTA CONVENCIÓN CIENTÍFICA INTERNACIONAL DE LA UTM 2022 realizada en la ciudad de Portoviejo - Ecuador del 26 al 28 de octubre de 2022.

ING. SANTIAGO QUIROZ FERNÁNDEZ, PHD.
RECTOR UTM
PRESIDENTE CCIUTM

ING. MARA MOLINA DE LOZANO, PH.D.
VICERRECTORA ACADÉMICA UTM
VICEPRESIDENTA PRIMERA

ALEX DUEÑAS RIVADENEIRA, PHD.
COORDINADOR CIENTÍFICO

LIC. MONSERRATE RUIZ CEDEÑO, PH.D.
SECRETARIA EJECUTIVA

Apéndice J. Ponencia: Prototipo de análisis de sentimiento de retroalimentación entre pares

UNIVERSIDAD TÉCNICA DE MANABÍ
Fundada en 1952

ESPAM MFL
ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ

Uleam
UNIVERSIDAD LAICA ELOY ALFARO DE MANABÍ

Confieren el presente

CERTIFICADO A:

PINARGOTE ORTEGA MARICELA
BOWEN MENDOZA LORENA ELIZABETH
MEZA HORMAZA JAIME ALCIDES
VENTURA SEBASTIAN

Por su participación como PONENTE en el
VI Congreso Internacional de Tecnologías de la Información y Computación

CITIC 2023

Con el tema:
Prototipo de análisis de sentimiento de retroalimentación entre pares

En el marco del evento realizado en la ciudad de Portoviejo - Ecuador,
los días 18 y 19 de mayo de 2023, con una duración de 16 horas académicas.

Portoviejo, 19 de mayo 2023

ING. SANTIAGO QUIROZ FERNÁNDEZ, PHD.
RECTOR

ING. ALEX DUEÑAS RIVADENEIRA, PHD.
DIRECTOR DEL INSTITUTO DE INVESTIGACIÓN

ING. JAIME MEZA HORMAZA , PHD.
COORDINADOR DEL EVENTO