BIG DATA PARA LA INVESTIGACIÓN LINGÜÍSTICA Y LA EDUCACIÓN BILINGÜE

ADELA GONZÁLEZ FERNÁNDEZ & ÁLVARO MAROTO CONDE[1]
UNIVERSIDAD DE CÓRDOBA

RESUMEN

La utilización de big data en el mundo de la ciencia y de la empresa está dando como resultado grandes beneficios no solo económicos, sino también en lo relacionado con la producción del conocimiento y en la calidad de este.

Sin embargo, a pesar de las reconocidas virtudes de esta metodología, pocos intentos ha habido de aplicarla al ámbito educativo y, menos aún, en el campo de las lenguas y de la enseñanza bilingüe.
En esta investigación nos proponemos estudiar cuál es el grado de interés de los usuarios de *Twitter* en distintas áreas temáticas para conocer así los gustos y las inclinaciones de los potenciales alumnos de la clase bilingüe español-inglés y conseguir un mayor acercamiento a ellos.

Para ello, hemos utilizado una herramienta de autor que, mediante la introducción de distintos parámetros, realiza una búsqueda sobre *Twitter* y nos devuelve la información procesada y lista para su posterior análisis.

De esta manera, conoceremos qué interés tiene cada país en las distintas áreas estudiadas para poder diseñar planes de estudio, materiales curriculares y proyectos docentes más personalizados, con la intención de que las clases de idiomas les resulten más interesantes y útiles.

Palabras clave: big data, *Twitter*, áreas temáticas, enseñanza bilingüe, educación.

ABSTRACT

Using big data in the scientific field and also in the business sector is undoubtedly resulting in many benefits, not only economic, but also related with knowledge and science.

However, in spite of the acknowledged advantages of this new methodology, it is not still being used in education nor in bilingual lessons.

In this project, we aim to know the degree of interest of *Twitter* users in some different topics to understand their tastes and interests as potential students of a bilingual Spanish-English class.

In order to do this, we have used authoring system that makes a search through *Twitter* by adding some preestablished parameters, such as geographic area, language, topic and date.

This way, we can learn the attitude of users towards the studied topics in order to design more effective and personalized curricula, with the intention of achieving more success in language lessons.

Keywords: big data, *twitter*, topics, bilingual education, education.

[1]Emails: adela.gonzalez@uco.es; amaroto@gmail.com

1.Introducción

La utilización de big data en el mundo de la ciencia y de la empresa está dando como resultado grandes beneficios no solo económicos, sino también en lo relacionado con la producción del conocimiento y en la calidad de este.

Esta mejora en las investigaciones científicas se está viendo reflejada a nivel cualitativo y cuantitativo, debido a las evidentes ventajas que se derivan de su utilización. Sin embargo, el trabajo con esta nueva fuente de información no puede plantearse ni acometerse sin la adopción de una nueva metodología ni de las herramientas necesarias para la gestión de datos.

Como su propio nombre indica, big data se refiere a los grandes conjuntos de información que por sus características no pueden ser obtenidos, gestionados ni procesados por herramientas tradicionales en un período de tiempo razonable (González, 2016). Estas características, según señalan algunos autores (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh y Byers, 2011; Dumbill, 2013; Provost y Fawcett, 2013; Chen, Mao y Liu, 2014), dan lugar a la teoría conocida como V^3, título que se adaptó del modelo previo de Laney (Chen*et al*. 2014): volumen, variedad y velocidad son, en líneas generales, los rasgos definitorios de este concepto que hoy en día conocemos como big data.

El inmenso volumen que caracteriza a big data crece a pasos agigantados, por lo que resulta difícil estimar el tamaño exacto del universo digital. Gantz y Reinsel (2012) estiman que su tasa de crecimiento es de un 40% anual y que no solo crece el número de datos que conforman la información, sino también el número de usuarios de internet y de aparatos electrónicos conectados.

Una de las primeras consecuencias de las enormes dimensiones de los conjuntos de datos es, sin duda alguna, su heterogeneidad. A diferencia de la metodología convencional, cuando se trata de una investigación basada en big data, la tipología de información a la que accedemos es enormemente variada. Esta es, por tanto, la segunda característica fundamental de big data y una de las responsables del cambio metodológico que se hace completamente necesario, puesto que la información aparece frecuentemente –en un porcentaje del 90%, según Gantz y Reinsel (2012)- en su forma no estructurada. Es decir, no podemos utilizar la información tal y como la encontramos, puesto que necesita ser seleccionada, ordenada y procesada previamente para su posterior aprovechamiento.

La tercera *v* a la que se suele hacer referencia en la definición de big data está relacionada con la velocidad que adquiere la información. Utilizamos aquí el término *velocidad* de forma general para abarcar tanto la creación de información como su transmisión. La información está en continuo movimiento, lo que se presenta como una de las principales virtudes de este nuevo concepto, aunque también un gran reto al que deben enfrentarse los científicos de todos los ámbitos. Hacia este sentido es hacia el que enfocan la cuestión de la velocidad autores como Zikopoulos, Eaton, deRoos, Deutsch y Lapis, (2012), quienes insisten en esta idea para alejarse de las definiciones convencionales.

Sin olvidarnos de otros rasgos característicos que se suelen asociar (aunque con menos frecuencia) a este concepto y que podrían elevar el número de la potencia que adorna la mencionada *v* a cinco, como son la veracidad y el valor, los tres conceptos clave que acabamos de explicar –volumen, variedad y velocidad- conforman la base sobre la que se sustentan la mayoría de las aproximaciones al concepto de big data y, paradójicamente, los principales obstáculos a los que nos enfrentamos y que provocan que, como rezaba la definición que hemos aportado unos párrafos más arriba, este tipo de información no pueda ser gestionada con las técnicas tradicionales.

Así las cosas, resulta imposible plantearse ningún tipo de investigación basada en las técnicas de big data sin el soporte informático adecuado.

En este estudio, por cuestiones prácticas, nos hemos centrado, dentro de big data, en la plataforma *Twitter*, debido a la gran cantidad de información que nos aporta y la variedad de datos útiles para la investigación lingüística. La información, por lo tanto, la extraemos de los tuits, que nos aportan datos de lo que pasa en el mundo en tiempo real y en grandes cantidades. En palabras de Yoon, Elhadad yBakken (2013: 122), el contenido de los tuits no depende de un estímulo intermitente específico, sino que representa una información más naturalista y tiene la ventaja adicional de estar disponibles en grandes cantidades.

Además, autores como Asur y Huberman (2010) demuestran que el análisis de estos medios, si es lo suficientemente amplio y está adecuadamente diseñado, suele ser más exacto que otras técnicas para la extracción de información, como los sondeos o las encuestas de opinión.

Sin embargo, a pesar de las reconocidas virtudes de esta metodología, de los numerosos ejemplos de su utilización y del enorme beneficio obtenido no solo con el trabajo con big data, sino con las redes sociales (Granovetter, 1973; Rogers, 2003; Wu, Huberman, Adamic y Tyler, 2004; Adamic y Adar, 2005), pocos intentos ha habido de aplicarla al ámbito educativo y, menos aún, en el campo de las lenguas. Aunque algunos autores sí se han planteado su utilidad como herramienta para el fomento del aprendizaje activo de idiomas (Borau, Ullrich, Feng &Shen, 2009) y como plataforma para el aprendizaje en educación superior (Ebner, Lienhardt, Rohs& Meyer, 2010).

2. Objetivos

En esta investigación nos proponemos extraer la información textual disponible en *Twitter*, a través de la recopilación, análisis y estudio de los tuits, así como la información extralingüística que se obtiene de este medio de comunicación y que contribuye al estudio de diversos factores, siendo de gran utilidad para la formulación de conclusiones de diversa índole.

Nuestro objetivo fundamental consiste en determinar cuál es el grado de interés que los usuarios de *Twitter* tienen sobre algunas áreas temáticas, para apoyarnos en ellas a la hora de la elaboración de materiales docentes y curriculares en la enseñanza bilingüe. Para ello, estudiamos cuatro países cuyo idioma oficial es el español y contrastamos el uso que se hace de esta lengua comparado con el del inglés en las áreas temáticas seleccionadas. Los países objeto de estudio son cuatro: España, México, Argentina y Chile.

3. Metodología

Como ya hemos explicado, el trabajo con una información de esta naturaleza, cuyo tamaño crece incesantemente a una velocidad vertiginosa, es inconcebible sin las tecnologías oportunas. La técnica y la informática, por tanto, entran en juego en este tipo de estudios y ocupan un lugar central en cualquier metodología que se aplique en este campo. Para este estudio hemos utilizado una herramienta de autor, creada a partir de las últimas técnicas de gestión de información y de minería de datos, que nos permite extraer la información que deseada y analizarla en un período de tiempo razonable, teniendo en cuenta el volumen de datos con los que trabajamos.

3.1. Parámetros

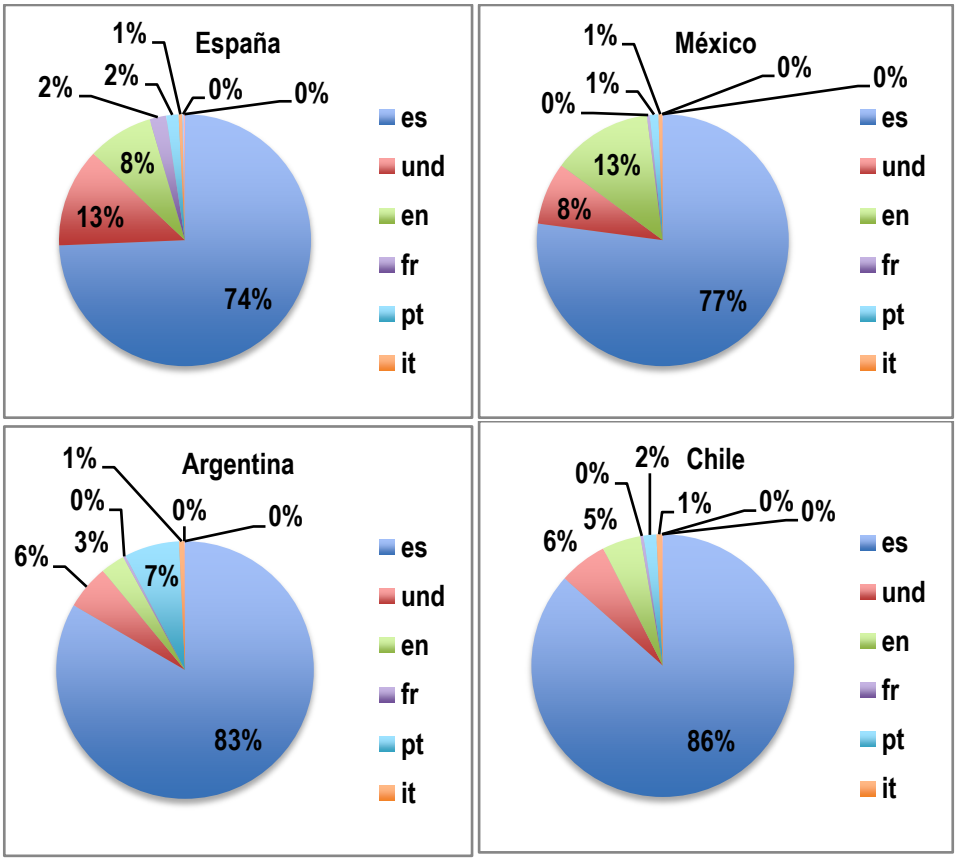
Para la realización de este estudio en concreto se introdujeron en la herramienta una serie de parámetros previos que determinaron la búsqueda de tuits que posteriormente se lanzaría sobre todo *Twitter*. Estos parámetros atienden a cuatro aspectos distintos: zona de aplicación, fecha, idiomas y áreas temáticas.

El primero de ellos, como se ha apuntado más arriba, se delimitó a cuatro países hispanohablantes para poder comparar el uso que se hace de los idiomas español e inglés, aunque previamente se realizó un estudio de los ocho idiomas más utilizados en cada país. Los países estudiados han sido, por tanto, España, México, Argentina y Chile. La delimitación de la zona geográfica se llevó a cabo mediante la introducción en la herramienta de las coordenadas de cada país.

Una vez seleccionados los países, se determinaron las fechas que comprenderían el estudio, que quedaron fijadas entre el 1 de enero de 2016 y el 28 de febrero del mismo año, lo que significa que la herramienta recogería los tuits publicados en los países mencionados dentro de este rango de fechas de dos meses.

Por otro lado, dado que el objetivo fundamental del estudio es la determinación del grado de interés de los usuarios de *Twitter* en determinadas áreas temáticas y comparar en qué medida se utilizan tanto el español como el inglés, se llevó a cabo una distinción de los idiomas de publicación de los tuits. En este caso, por cuestiones prácticas y de optimización de recursos, se realizó la búsqueda solo con seis idiomas para obtener una visión general del mapa lingüístico de cada país y poder comparar posteriormente el español y el inglés. Los idiomas son español, inglés, francés, italiano, portugués y alemán. Como vemos en los gráficos siguientes (Gráfico 1, Gráfico 2, Gráfico 3 y Gráfico 4), aparece

también un idioma indeterminado, que se corresponde con tuits cuyo idioma *Twitter* es incapaz de distinguir por la presencia de caracteres extraños o de contenido multimedia.



Para finalizar, se establecieron cinco áreas temáticas mediante la utilización de campos semánticos que delimitaran los tuits pertenecientes a cada una de ellas. De esta manera, se obtuvieron los tuits relacionados con cada temática seleccionada. Dichas áreas fueron la política, la música, los deportes, la moda y la tecnología. Los tuits recogidos de cada una mediante la búsqueda por palabras clave dentro de los campos semánticos específicos fueron ya recogidos solo en español y en inglés.

4. Resultados

Tras la búsqueda ininterrumpida de tuits durante los dos meses de recogida de datos, registramos, en primer lugar, un recuento del total de publicaciones obtenidas en cada país. De los cuatro estudiados, como podemos ver a continuación en el Gráfico 5, es en España donde más producción hubo, mientras que Chile es el país con menos participación:

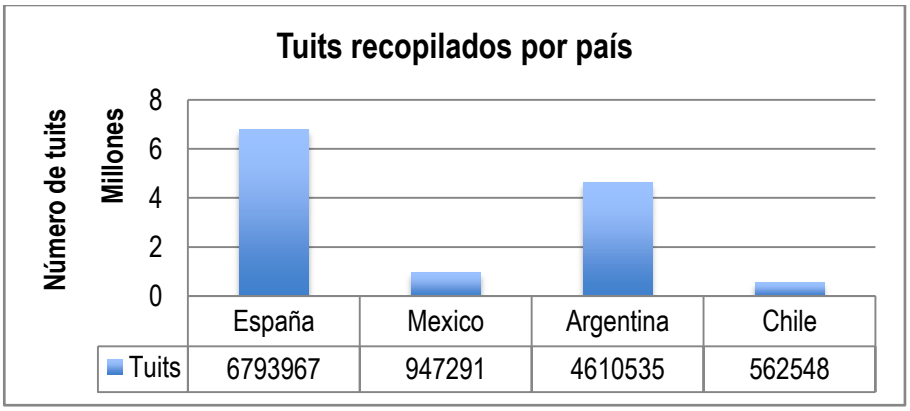
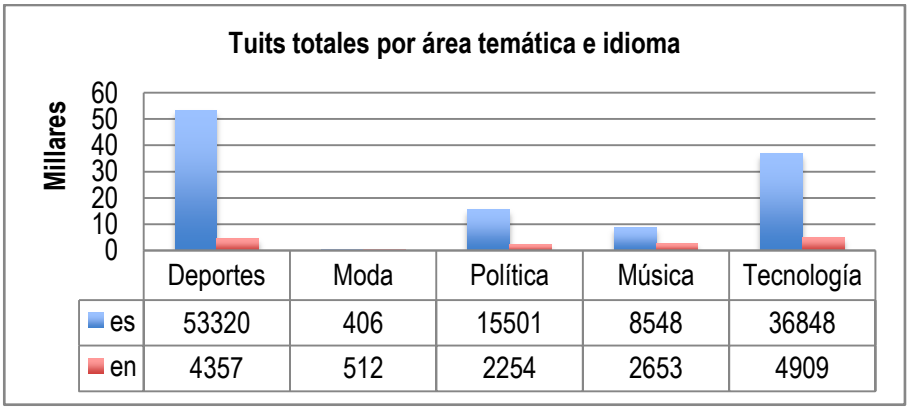
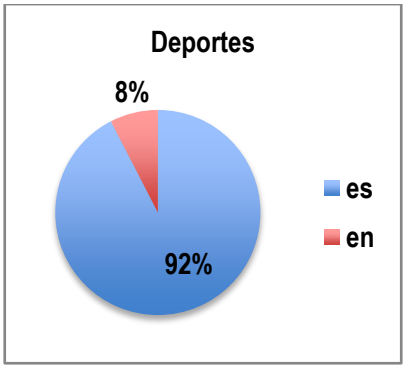


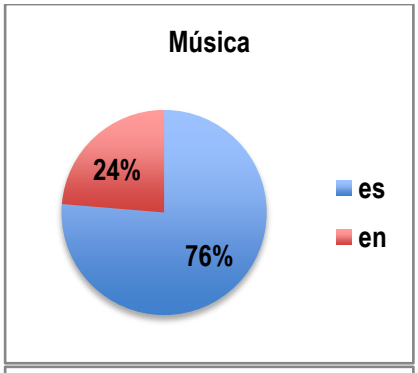
Gráfico 5: Número total de tuits recopilados en cada país

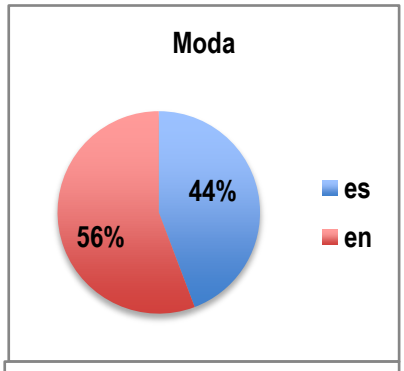
4.1. España

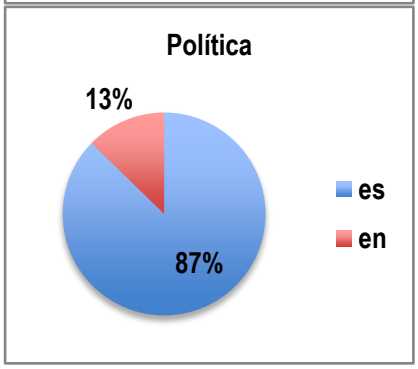


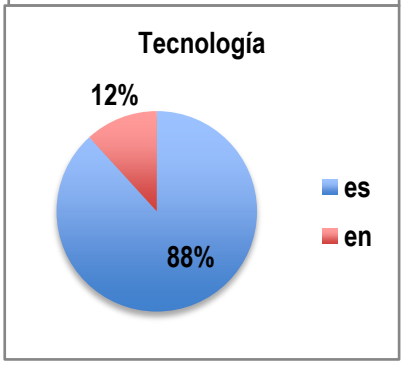
	Deportes	Moda	Política	Música	Tecnología
es	53320	406	15501	8548	36848
en	4357	512	2254	2653	4909



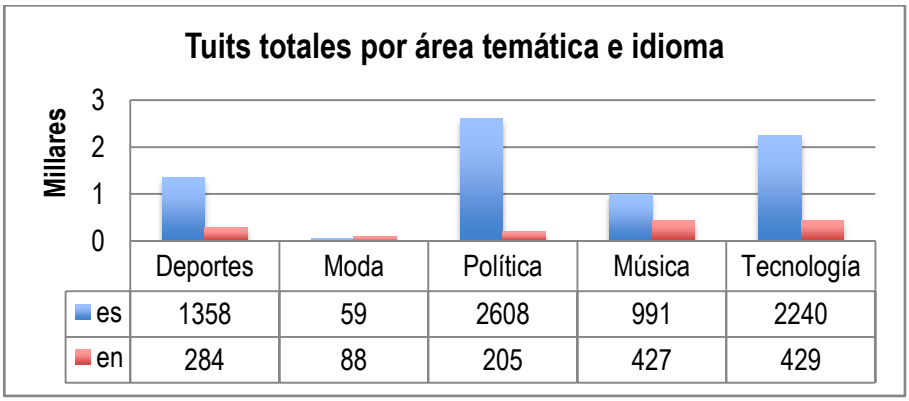


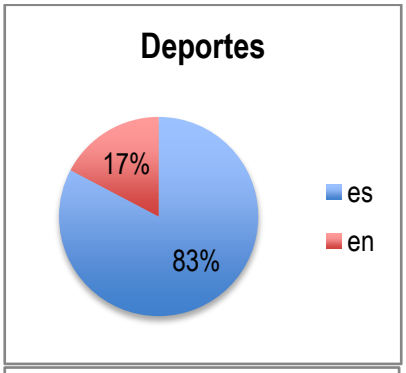


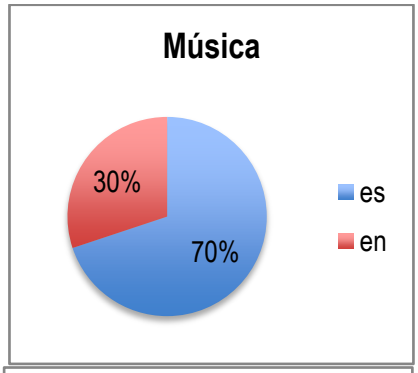


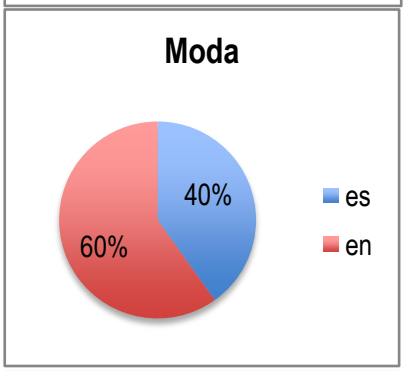


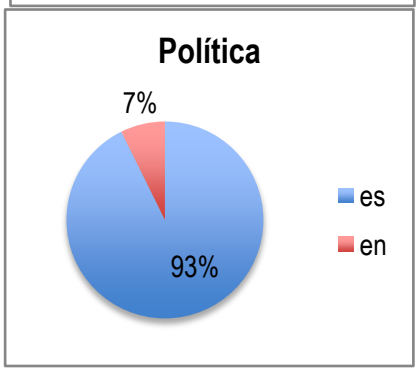
4.2. México

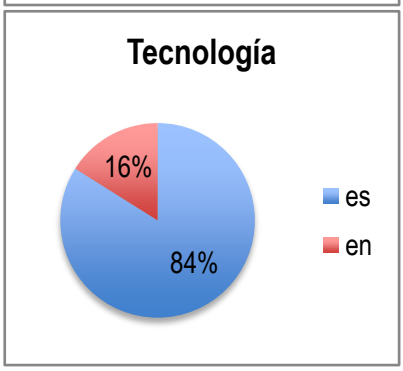

Tuits totales por área temática e idioma
Millares
3
2
1
0
Deportes
Moda
Política
Música
Tecnología
es
1358
59
2608
991
2240
en
284
88
205
427
429


Deportes
17%
83%
es
en

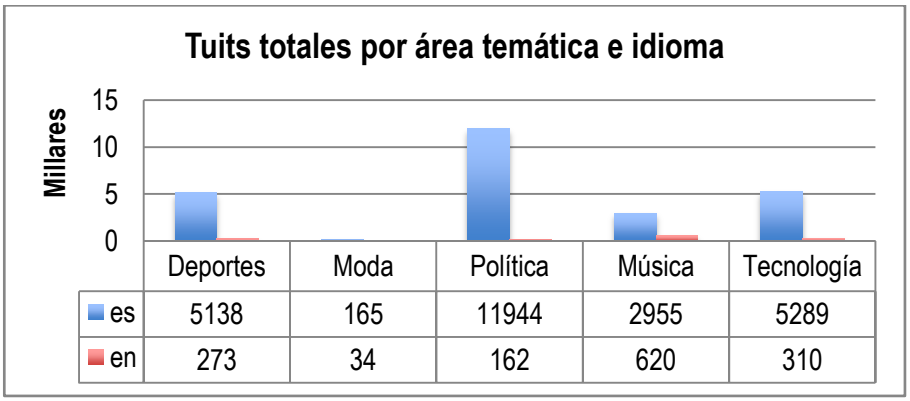

Música
30%
70%
es
en


Moda
40%
60%
es
en

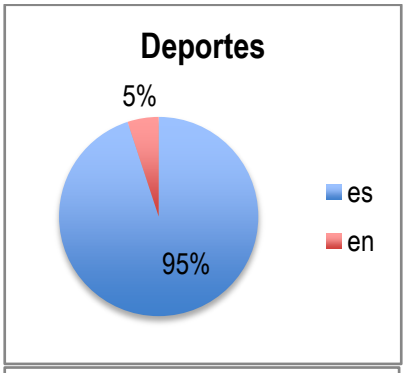

Política
7%
93%
es
en

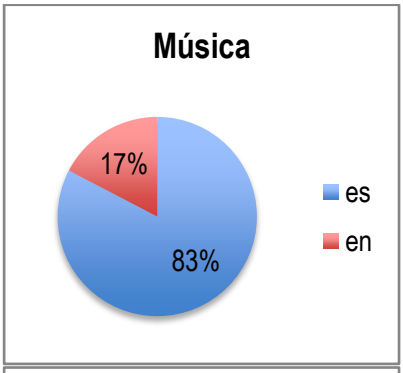

Tecnología
16%
84%
es
en

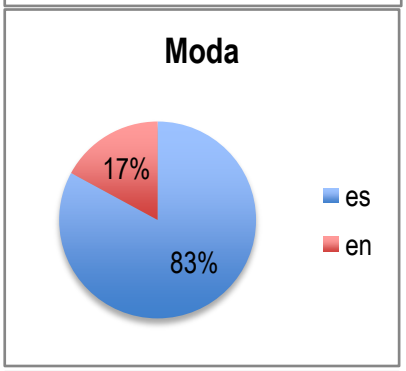
4.3. Argentina

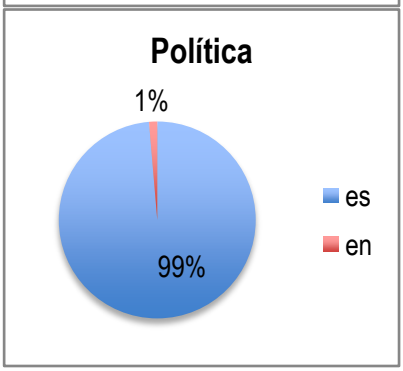


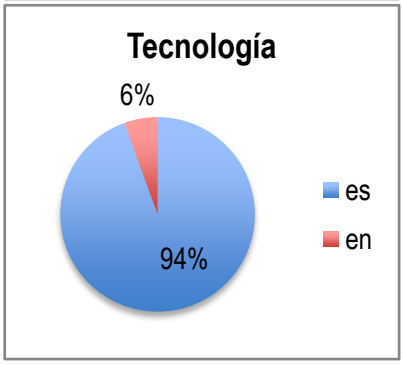
	Deportes	Moda	Política	Música	Tecnología
es	5138	165	11944	2955	5289
en	273	34	162	620	310



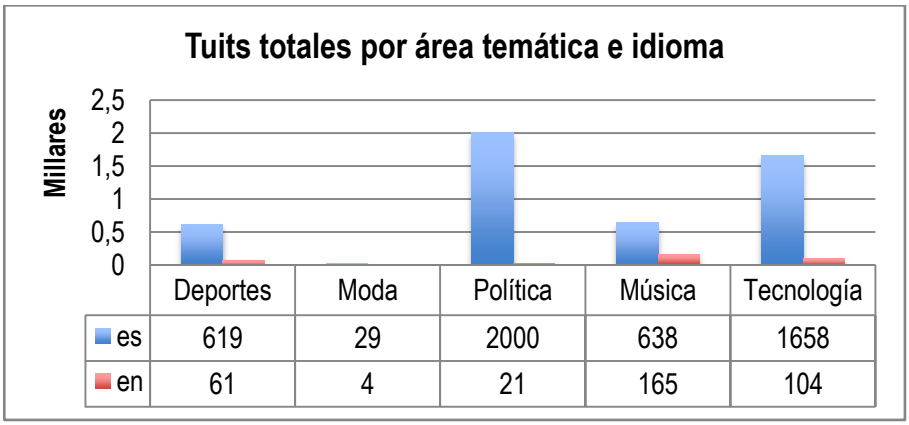


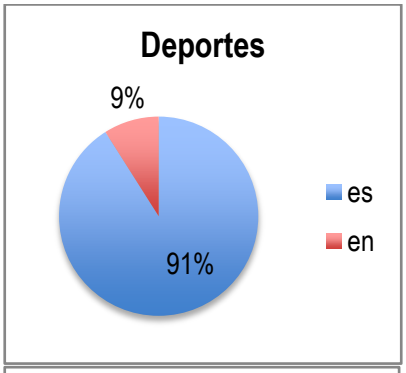


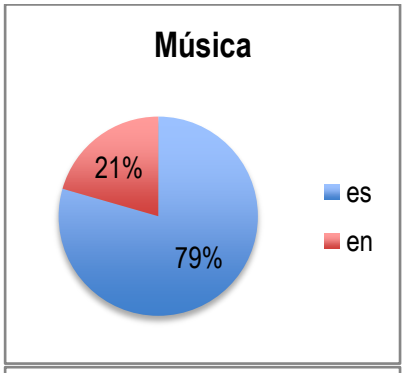


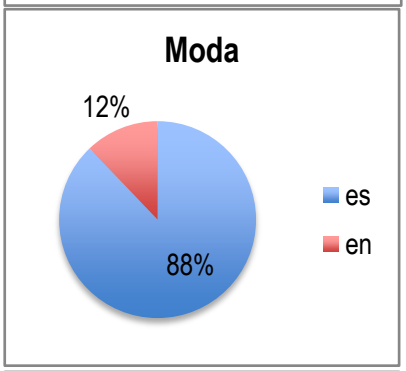


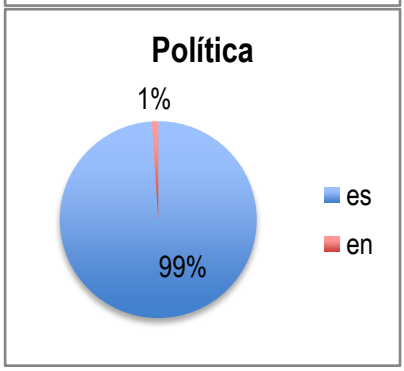
4.4. Chile

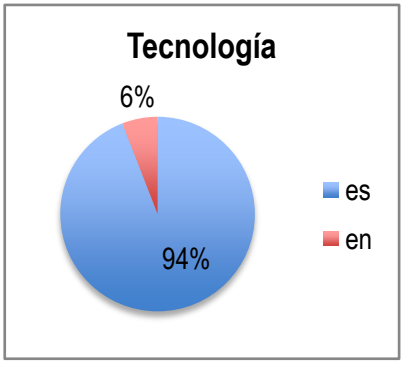












5. Conclusiones

El trabajo con big data supone un salto cualitativo y cuantitativo en los estudios científicos, en general, y en el trabajo lingüístico, en particular. Dentro del campo de la educación, y más concretamente de la educación bilingüe, debemos aprovechar las ventajas que surgen de la utilización de esta nueva metodología para ampliar los horizontes y obtener el máximo provecho posible.

Dejando a un lado los innumerables beneficios aplicables al ámbito educativo, nos centramos aquí en la posibilidad de investigar acerca de los intereses reales de los usuarios de *Twitter*, estudiantes potenciales de idiomas, para poder acercarnos un poco más a su realidad, a sus gustos y a sus preferencias.

En esta investigación, por tanto, se ha llevado a cabo una búsqueda y posterior recopilación de los tuits publicados en España, México, Argentina y Chile durante enero y febrero de 2016. Posteriormente, se ha realizado un estudio comparativo entre aquellos posts escritos en español y en inglés, separados en las distintas áreas temáticas que hemos establecido, para hacernos una idea de cuáles son las vías con mayores posibilidades de penetración del bilingüismo en cada uno de los países.

De esta manera, conociendo cuáles son sus áreas de interés, resulta más accesible la confección de planes de estudio, materiales curriculares y proyectos docentes dirigidos a un alumnado más receptivo, que se enfrentaría a clases que atenderían sus intereses de una forma mucho más personalizada.

No cabe duda de que la única forma de adaptarnos a los nuevos tiempos, en los que los estudiantes tienen un acceso cada vez mayor a cualquier tema que les suscite interés, es hacerlos partícipes de la clase, permitir que ellos sean los protagonistas del aprendizaje y el eje en torno a cual gire la planificación del profesorado. Solo así conseguiremos despertar en ellos la curiosidad por la clase y la implicación que requiere la enseñanza de idiomas.

6. Referencias bibliográficas

Adamic, L. A., & Adar, E. (2005). How to search a social network. *Social Networks,* 27(3),187–203.

Asur, S. & Huberman, B. (2010). Predicting the future with Social Media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference of Web Intelligence and Intelligent Agent Technology*, 1 (pp. 492-499). Washington, DC: IEE Computer Society.

Borau, K., Ullrich, C., Feng, J. &Shen. R. (2009). Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence. En Spaniol, M. *et al*. (Eds.), *ICWL 2009* (pp.78-87). Berlin: Springer-Verlag Berlin Heidelberg.

Ebner, M., Lienhardt, C., Rohs, M. & Meyer, I. (2010). Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning? *Computers & Education*, 55, 92-100.

González, A. (2016). Análisis de las necesidades traductológicas en Europa a través de big data. *Skopos*, 7, 45-74. ISSN: 2255-3703 [en prensa].

Granovetter, M. S. (1973). The Strengh of Weak Ties. *American Journal of Sociology*, 78(6), 1360-1380.

Manyika, J., Chui, M., Brown, J. Dobbs, R., Roxburgh, C. &Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Mckinsey Global Institute.

Dumbill, E. (2013). Making sense of Big Data. *Big Data*, 1(1), 1-3. doi: 10.1089/big.2012.1503.

Chen, M., Mao, S., Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), 171-209. doi: 10.1007/s11036-013-0489-0.

Gant, J. & Reinsel, D. (2012). The digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. *IDC*. Recuperado de https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in- 2020.pdf

Provost, F. & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59. doi: 10.1089/big.2013.1508.

Rogers, E. (2003). *Diffusion of Innovations, 5th Edition*. New York: Simon and Schuster.

Wu, F., Huberman, B., Adamic, L. A. & Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337 (1-2), 327- 335. doi:10.1016/j.physa.2004.01.030.

Yoon, S., Elhadad, N. &Bakken, S. (2013). A Practical Approachfor Content Mining of Tweets. *American Journal of Preventive Medicine*, 45(1), 122-129.

Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T. & Lapis, G. (2012). *IBM. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw Hill.

Received: 22/11/2016
Accepted: 01/09/2017