07-003

BIG DATA: APPLICATION TO THE ANALYSIS OF FEELINGS IN AIRLINES

Izquierdo García, Mariángeles[(1)]

[(1)]Universidad de Granada

Discovering new forms of data collection has been possible thanks to the rise of social networks and technology 2.0.

In this research was carried out a bibliographic review on the concept of Big Data and the techniques used for its analysis. Moreover, was analyzed and predicted consumer's sentiments in airlines using the Knime data mining platform. To do this, a model based on decision trees has been chosen, checking their reliability through confusion matrices. The used data comes from the Quandl database in which, through algorithms, more than 6000 websites are scanned every 5 minutes. The data set includes the period from January 2013 to June 2017.

The obtained results show scoring intervals that reflect the weighted average sentiment of the consumer with respect to the initial sentiment and the volume of existing news. In this line, we observe how polarized high and low feelings carry extreme intraday weighted average feelings. It should be noted that the model used in the study predicts feelings with 91% reliability in the case of the airline Ryanair and 77% in SouthWest Airlines.

Keywords: *"Big Data"; "Analysis of feelings"; "Data mining" ;"Decision trees"; "Knime"*

BIG DATA: APLICACIÓN AL ANÁLISIS DE SENTIMIENTOS EN AEROLÍNEAS

El auge de las redes sociales y la tecnología 2.0 ha permitido obtener gran cantidad de información mediante nuevas formas de recolección de datos.

En el presente trabajo se ha realizado una revisión bibliográfica sobre el concepto de Big Data y las diferentes técnicas utilizadas para su análisis. Asimismo, se han analizado y predicho los sentimientos de los consumidores en aerolíneas mediante la plataforma de minería de datos Knime. Para ello, se ha escogido un modelo basado en árboles de decisión comprobando su fiabilidad a través de matrices de confusión. Los datos utilizados en el estudio proceden de la base Quandl en la que, mediante algoritmos, se exploran más de 6000 sitios web cada 5 minutos. El conjunto de datos comprende el periodo enero 2013 a junio 2017.

Los resultados obtenidos muestran intervalos de puntuación que reflejan el sentimiento promedio ponderado del consumidor respecto al sentimiento inicial y al volumen de noticias existentes. En esta línea, se observa cómo sentimientos altos y bajos polarizados conllevan sentimientos promedios ponderados intradía extremos. Cabe destacar que el modelo utilizado en el estudio predice los sentimientos con un 91% de fiabilidad en el caso de la aerolínea Ryanair y de un 77% en SouthWest Airlines.

Palabras clave: *"Bid Data"; "Análisis de sentimientos"; "Minería de datos"; "Árboles de decisión"; "Knime"*

Correspondencia: Mariangeles Izquierdo García; m.angeles_izquierdo@hotmail.com

©2018 by the authors. Licensee AEIPRO, Spain. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

1.Introducción

El desarrollo de las tecnologías y el cambio en las necesidades empresariales, está provocando un auge en la búsqueda de datos para obtener información. Diariamente se publican en la red millones de datos acerca de un producto, un servicio o una compañía.

Ante esta situación las empresas intentan analizar los datos existentes con el objetivo de adaptarse, conocer al consumidor y fidelizarlo. El análisis de sentimientos es una de las técnicas más comunes en este campo. Una de las principales razones es el rendimiento que es capaz de obtener en el social media.

La comunicación que se presenta a continuación pretende realizar una revisión bibliográfica sobre el Big Data y conocer su aplicación en el análisis de sentimientos. En esta línea, se define, en primer lugar, en el capítulo 2, el concepto de Big Data y se explican las diversas fuentes de extracción de información. En el mismo capítulo se explican las principales técnicas de análisis que se utilizan en la actualidad. Una vez realizada la introducción general sobre los nuevos conjuntos de datos, se realiza, en el capítulo 3, un estudio sobre el análisis de sentimientos y las principales metodologías utilizadas. Del mismo modo, se muestra un análisis descriptivo de cinco importantes aerolíneas. Finalmente, en los capítulos 4 y 5, se muestra el análisis empírico realizado sobre las compañías aéreas SouthWest Airlines y Ryanair. En éstos, se describe el modelo utilizado y se exponen los resultados obtenidos del estudio. Para concluir, en el capítulo 6 se muestran las principales conclusiones del trabajo.

2.Conceptos, metodología y retos del Big Data

2.1 Definición

Si bien es cierto que no existe una definición unánimemente aceptada sobre el término Big Data, una de las más utilizada es la propuesta por la empresa Gartner[1] en 2012 que define el Big Data como: “Gran volumen, alta velocidad, y gran variedad de activos de información que requieren nuevas formas de procesamiento para permitir una mejor toma de decisiones, descubrimiento de conocimientos y optimización de procedimientos”.

Partiendo de este concepto propuesto por Gartner, IBM, SAS Insights y Oracle amplían la definición analizando diversos componentes: volumen (Jagadish et al., 2014), velocidad (Cukier, 2010), variedad (Gandomi & Haider, 2015), valor (Lambrecht & Tucker, 2015), variabilidad (Gandomi & Haider, 2015) y veracidad (Ward & Barker, 2013).

Atendiendo a las fuentes de extracción de este tipo de datos, las más utilizadas son: Internet de las coss (IoT) (Coches inteligentes, Smart Cities, teléfonos inteligentes…), datos de auto-cuantificación, dato multimedia, datos de redes sociales o datos de gobiernos y administraciones públicas

Por otra parte, al igual que ocurre con las plataformas y las bases de datos, el Big Data necesita de técnicas específicas para lograr un correcto análisis. Entre el conjunto de las existentes cabe destacar el Machine Learning y la Minería de Datos. La primera de ellas, tiene como objetivo estudiar, diseñar y desarrollar algoritmos que permitan a los ordenadores tener la capacidad de aprender sin ser programados de forma explícita. Durante dicho proceso se pueden utilizar algoritmos como redes neuronales, árboles de decisión o clasificadores bayesianos. Sin embargo, la segunda se entiende como el proceso de extracción de conocimientos y/o patrones interesantes desconocidos a partir de datos aparentemente no estructurados. Algunas de las técnicas más utilizadas son Clustering, análisis de componentes principales, análisis de correspondencias o web mining entre otros.

En esta línea, el Big Data presenta el reto de conseguir una utilización y análisis eficaz por parte de las empresas de grandes volúmenes de datos.

[1] http://www.gartner.com/it-glossary/big-data

3.Análisis de sentimientos

3.1 Concepto de análisis de sentimientos

Existen diversas definiciones acerca del análisis de sentimientos. Nasukawa & Yi (2015) lo relacionan con la identificación de cualquier expresión que muestre un sentimiento, fuerza o polaridad y la relación que tienen con el usuario. He et al., (2015) lo definen como una aplicación especial dentro de la minería de textos que se encarga del estudio computacional de las opiniones, sentimientos o emociones que aparezcan en cualquier texto de las redes sociales y que intenta diferenciarlos según la fuerza que tengan. En la misma línea se encuentra la definición de Xuan & Xu (2016) que se centran en el uso del procesamiento del lenguaje natural y en lo métodos de aprendizaje automáticos para extraer e identificar los sentimientos dentro de un texto.

Atendiendo al proceso de agrupación de expresiones en categorías, existen fundamentalmente dos tendencias: el análisis usando el aprendizaje supervisado y el no supervisado. El primero necesita tener ejemplos clasificados con anterioridad para ser capaz de predecir resultados futuros. Sin embargo, en el segundo método no es necesario obtener palabras o frases como etiquetas o ejemplos para realizar el análisis. En este caso las técnicas más comunes son el Clustering, análisis de componentes principales o el análisis de frecuencia monetario.

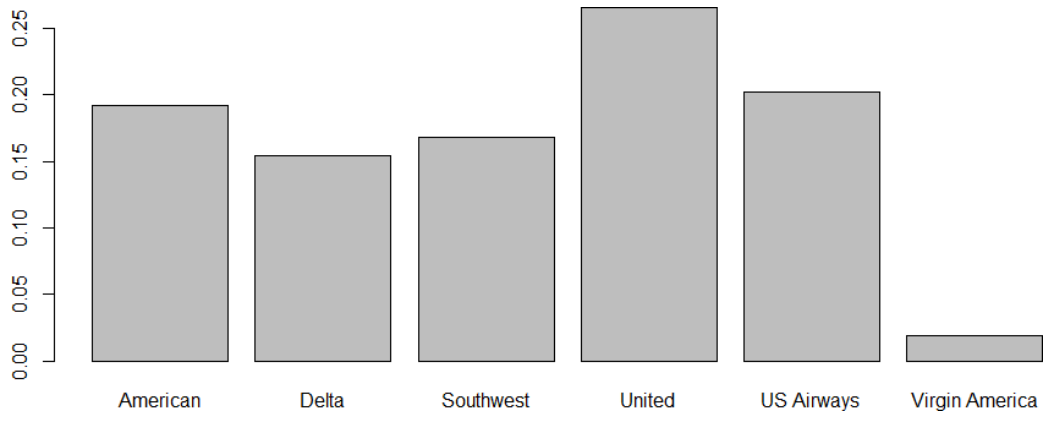
3.2 Análisis de sentimientos en las aerolíneas

Una de los sectores más conocidos por su aplicación del análisis de sentimientos es el estudio del comportamiento del consumidor en las líneas aéreas.

Con el objetivo de analizar este sector, se ha utilizado la base de datos “US Airline Sentiment analysis using twitter data” realizada por Vivek Yadav, que analiza 14640 tweets de 7700 usuarios en Twitter a partir de febrero de 2015. Las compañías aéreas que aparecen son: American Airlines, Delta Airlines, SouthWest Airlines, United Airlines, US Airways y Virgin America. Por su parte, las variables que componen la base son, los sentimientos medidos como positivos, negativos o neutros y la razón por la que ese sentimiento es negativo (vuelo cancelado, servicio al cliente, retraso del vuelo, quejas de asistencias de vuelo, problemas en la reserva, NC, mal vuelo, largas colas, equipaje perdido y equipaje dañado). En este estudio se ha utilizado el software R para analizar los sentimientos, tal y como hicieron autores como Yaqoob et al., (2016).

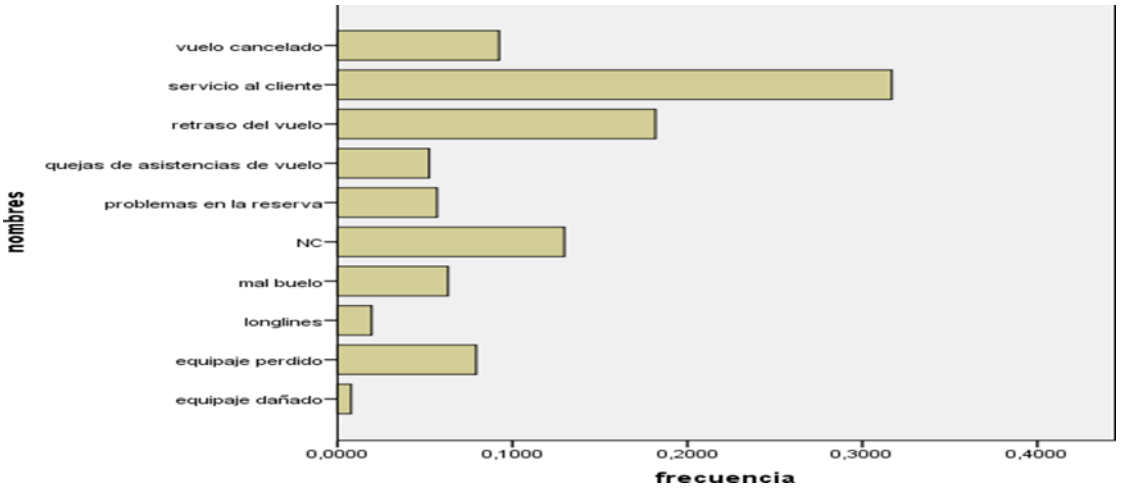
Atendiendo a los resultados obtenidos, en la figura 1, se muestra qué aerolínea es la más usada, siendo esta United Airlines seguida de US Airways.

Figura 1: Diagrama de barras de uso de aerolíneas



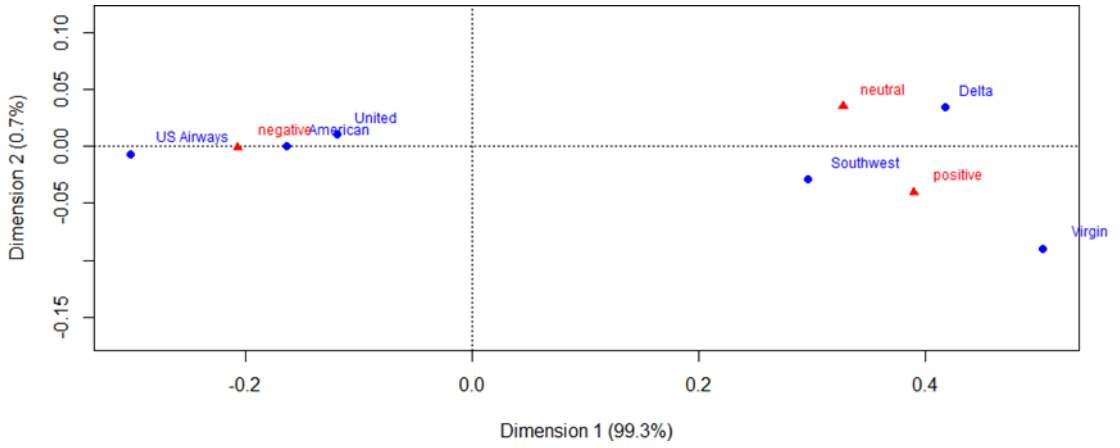
Sabiendo que una gran parte de los usuarios tienen sentimientos negativos hacia las compañías, es necesario para éstas conocer cuál ha sido la razón de ello[2].

Figura 2: Diagrama de frecuencia de razones negativas



Por su parte, a través del análisis multidimensional, para conocer qué compañías son las mejor y peor valoradas se realizó un análisis de correspondencia:

Figura 3: Análisis de correspondencia de las aerolíneas



Para ello se ha realizado una tabla de contingencia con los individuos con sentimientos negativos y las razones de cada uno.

[2] Todas las figuras y tablas presentadas se han realizado mediante elaboración propia

Tabla 1: Tabla de contingencia de razones negativas

	American	Delta	SouthWest	United	US Airways	Virgin America
Bad flight	87	64	90	216	104	15
Can´t tell	198	186	159	379	246	12
Cancelled flight	246	51	162	181	189	12
Customer service issue	768	199	391	680	811	34
Damaged luggage	12	11	14	22	11	1
Flight attendant complaints	87	60	38	168	123	1
Flight booking problems	130	44	61	144	122	18
Late flight	249	269	152	525	453	6
Longliness	34	14	29	48	50	3
Lost luggage	149	57	90	268	154	4

4.Estudio empírico

4.1 Modelo aplicado

El modelo propuesto tiene el objetivo de predecir los sentimientos futuros de los consumidores respecto a dos aerolíneas. En esta línea, los sentimientos actuales de los consumidores hacia la compañía, junto con otros aspectos como el volumen y contenido de las noticias que aparezcan, influirán en las intenciones de compras o utilización de servicios de los mismos. Para ello se han utilizado árboles de decisión que, a través de las diferentes opciones mostradas en sus nodos, ayudarán a predecir el sentimiento futuro que tendrá cada consumidor. Según Han & Kamber (2006), un árbol de decisión es un diagrama de flujo, donde cada nodo interno denota una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo de hoja tiene un rótulo de clase, siendo el nodo superior el conocido como nodo raíz. Asimismo, para obtener el árbol, se realiza una partición de los datos en dos partes y en función del porcentaje y los datos de la partición se construye el árbol. Dichos valores aparecen en lo que es conocido como matriz de training.

Del mismo modo, es necesario medir la bondad del modelo. Para ello se utiliza la matriz de confusión de test también denominada como matriz de error o contingencia, la cual contiene información acerca de cada clasificación actual y las predicciones realizadas.

Las dos compañías analizadas son Ryanair y Southwest Airlines para el período comprendido entre enero de 2013 y junio de 2017. La predicción de sentimientos se utiliza para conocer la cotización tal y como realiza la empresa de big data Quandl. El dataset ha sido comprado de Quandl que analiza los sentimientos según el léxico cada 5 minutos en millones de blogs, en twitter y otras redes sociales. Ésta ofrece indicadores diarios sobre sentimientos de los usuarios a través de las publicaciones en Twitter. En esta línea, la obtención de los datos se realiza mediante algoritmos que exploran cada cinco minutos 6000 sitios web. Para la empresa comentada, la información obtenida es crucial en la toma de decisiones y posicionamiento en el mercado.

Por su parte, las variables utilizadas son las siguientes: en primer lugar, el promedio ponderado intradía del sentimiento que es una medida del lenguaje utilizado en los medios para un determinado valor en un día concreto, cuyo rango fluctúa entre -5 y +5. En segundo lugar, el máximo sentimiento intradía definido como el valor máximo intradía (compra o venta de un valor en una misma sesión bursátil) de la puntuación del sentimiento, sondeado a intervalos de cinco minutos. En tercer lugar, el mínimo sentimiento intradía expresado como el valor mínimo intradía de la puntuación del sentimiento, sondeado a intervalos de cinco minutos. En cuarto lugar, el volumen de noticias que hace referencia al número total de artículos sobre un valor, que han sido publicados y analizados en un determinado día. Por último, el buzz de noticias definido como una

medida de la tasa de cambio en la cobertura de noticias de una acción dada en un día determinado, normalizado en una escala de uno a diez. Por su parte, el software utilizado se trata de una plataforma de minería de datos conocida como Knime (Konstanz Information Miner), que permite el desarrollo de modelos mediante destrezas visuales, siendo una de sus principales características es que está basado en la plataforma Eclipse y programado en Java.

5.Resultados

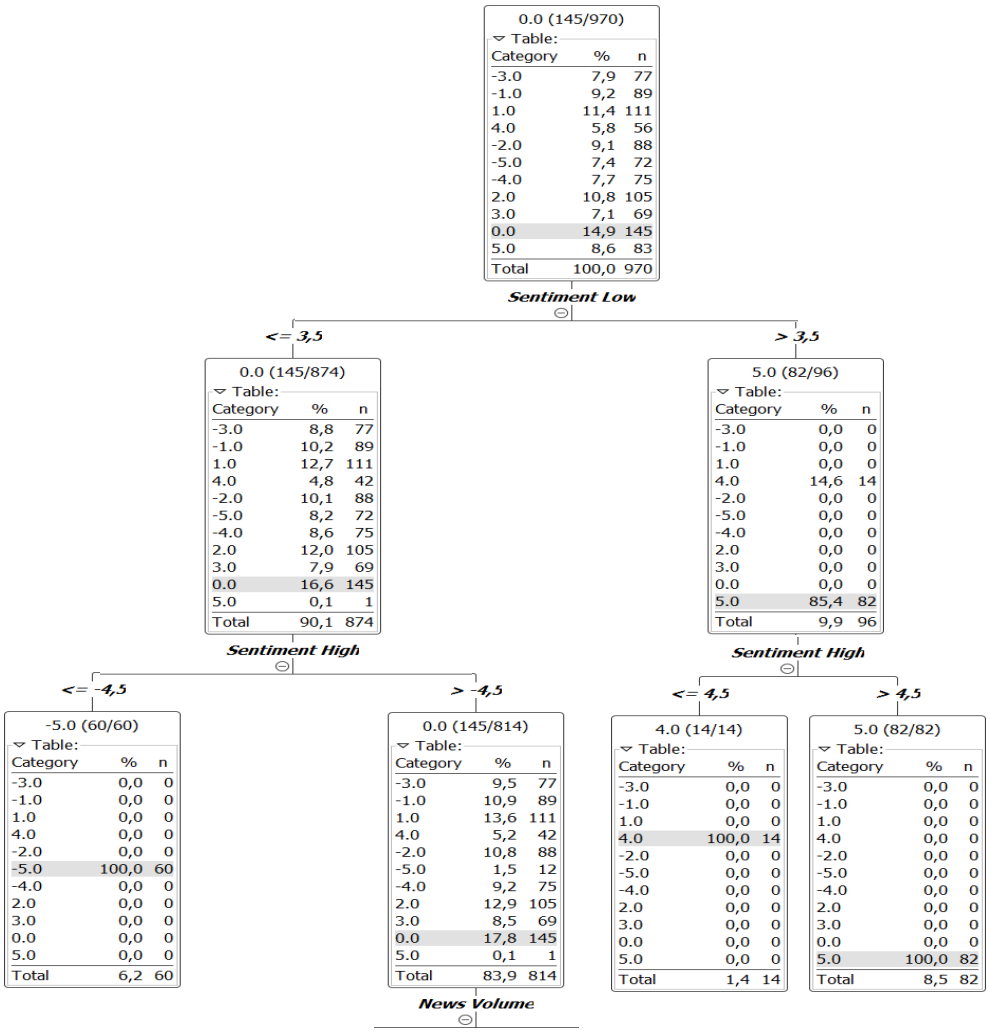
En primer lugar, la matriz de training de la compañía Southwest Airlines es la que se muestra en la tabla 2:

Tabla 2: Matriz de confusión de training de la compañía SouthWest

Sentimiento	-2.0	-3.0	-1.0	1.0	4.0	0.0	-1.0	-5.0	-4.0	2.0	5.0
-2.0	45	18	23	2	0	0	0	0	0	0	0
-3.0	9	62	0	1	0	0	0	0	0	0	0
-1.0	10	0	60	19	0	0	0	0	0	0	0
1.0	0	0	7	82	0	0	0	0	22	0	0
4.0	0	0	0	0	54	0	0	0	0	2	0
0.0	0	0	0	0	0	145	0	0	0	0	0
-5.0	0	0	0	0	0	0	60	12	0	0	0
-4.0	0	11	0	0	0	0	0	64	0	0	0
2.0	0	0	0	15	0	0	0	0	74	16	0
3.0	0	0	0	3	7	0	0	0	11	48	0
5.0	0	0	0	0	1	0	0	0	0	0	82

El conjunto de valores utilizados en el training es de 776. A raíz de la matriz, se calcula el árbol. A continuación, en la figura 4 se presentan una de las ramas cortadas y el modo de análisis:

Figura 4: Nodo del árbol de decisión de la compañía SouthWest



En la figura 4. se muestra un primer corte del árbol. Se parte de un sentimiento neutro, que dependiendo de cómo varíe dadas las experiencias del consumidor, de las noticas que aparezcan o de otros factores irá variando hasta lograr predecir el sentimiento futuro del consumidor. Las puntuaciones de los mismos, indican la polaridad del sentimiento.

Esta serie de clasificaciones nos ayudan a observar que un sentimiento bajo y alto elevado conllevan a que exista un sentimiento de 4 o 5, mientras que un sentimiento bajo moderado pero un sentimiento alto muy negativo, provocan un sentimiento final de -4 o -5. En los casos en los que el sentimiento final es 0 (neutro), suelen estar marcados a lo largo del tiempo,por sentimientos altos y bajos moderados y un volumen de noticias neutral.

Por otra parte, la evaluación de la fiabilidad del árbol obtenido mediante la matriz de test es la que se presenta en la tabla 3:

Tabla 3: Matriz de confusión de test de la compaía SouthWest

Sentimiento	-2.0	-3.0	-1.0	1.0	4.0	0.0	-5.0	-4.0	2.0	3.0	5.0
-2.0	29	6	19	3	0	0	0	0	0	0	0
-3.0	9	39	0	1	0	0	0	3	0	0	0
-1.0	6	0	35	19	0	0	0	0	0	0	0
1.0	0	0	4	52	0	0	0	0	18	0	0
4.0	0	0	0	1	35	0	0	0	0	1	1
0.0	0	0	0	0	0	97	0	0	0	0	0
-5.0	0	0	0	0	0	0	42	6	0	0	0
-4.0	0	13	0	1	0	0	0	36	0	0	0
2.0	0	0	0	7	0	0	0	0	49	14	0
3.0	0	0	0	1	4	0	0	0	11	30	0
5.0	0	0	0	0	1	0	0	0	0	0	55

Como se puede observar, el número de valores clasificados correctamente es 499, y la medida Accuracy es del 77,006%, lo que implica que se ha realizado una buena exactitud en la clasificación. El indicador Cohen´s Kappa por su parte, es 0,745 que se refiere a que se están evitando un 74,5% de los errores que se cometería en una clasificación al azar, lo que implica que tenemos un algoritmo que sabe predecir los sentimientos de la compañía.

Si analizamos la matriz por dentro, se puede ver como la suma de la diagonal principal es 499. Cada una de las casillas de dicha diagonal son verdaderos positivos para el valor, por ejemplo, para el valor -2 de sentimiento, el modelo predijo el verdadero valor en 32 casos.

Una vez analizada la compañía SouthWest, se realizó el mismo estudio para Ryanair. En este caso, el árbol completo predice unos resultados similares. No obstante, en primer lugar, se va a exponer la matriz de training mediante la cual, se ha creado el árbol.

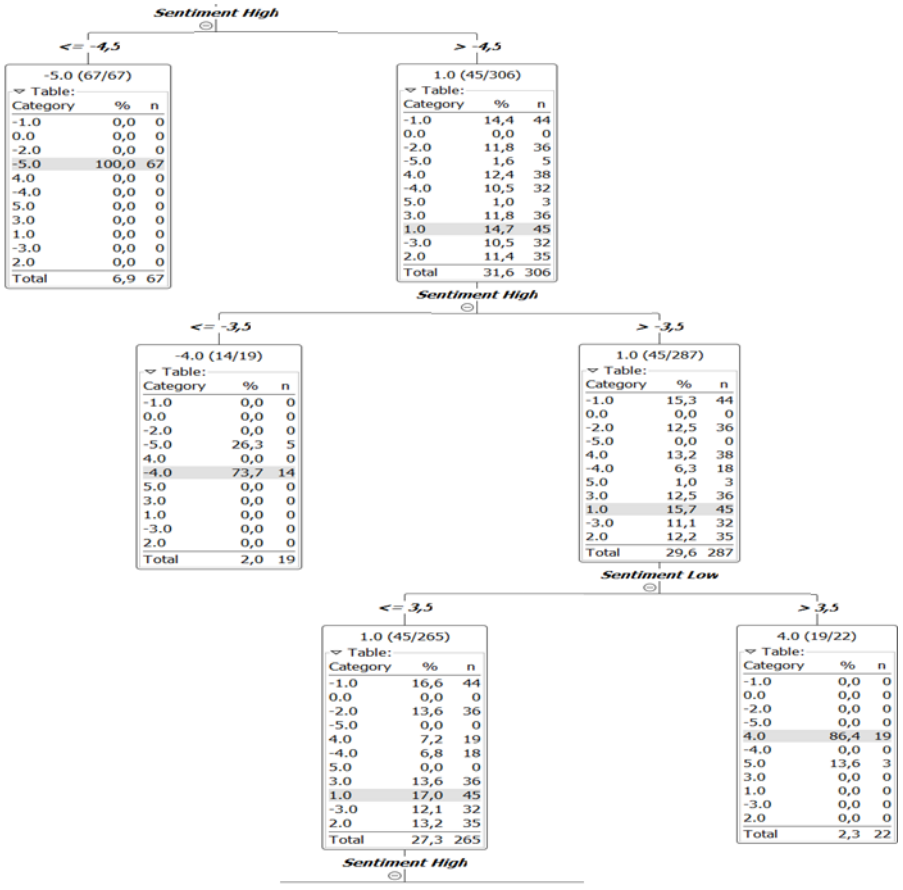
Tabla 4: Matriz de confusión de training de la compaía Ryanair

Sentimiento	0.0	3.0	-1.0	-2.0	-5.0	4.0	-4.0	5.0	-3.0	1.0	2.0
0.0	333	0	0	0	0	0	0	0	0	0	0
3.0	2	20	0	0	0	1	0	0	0	0	4
-1.0	0	0	19	3	0	0	0	0	0	6	1
-2.0	1	0	2	20	0	0	0	0	2	0	0
-5.0	2	0	0	0	42	0	4	0	0	0	0
4.0	0	5	0	0	0	20	0	0	0	0	0
-4.0	0	0	0	1	0	0	18	0	3	0	0
5.0	0	0	0	0	0	5	0	55	0	0	0
-3.0	0	0	0	4	0	0	0	0	17	0	0
1.0	0	1	3	0	0	0	0	0	0	21	6
2.0	1	0	0	0	0	0	0	0	0	3	21

El conjunto de valores utilizados en el training es de 586. A raíz de dicha matriz se ha construido el árbol de decisión de la compañía.

Por su parte, en la figura 5 se muestra un nodo de árbol de decisión de la compañía:

Figura 5: Nodo del árbol de decisión de la compañía Ryanair



En esta compañía las conclusiones son bastante parecidas a la de la anterior aerolínea, sin embargo, los sentimientos finales suelen estar bastante polarizados. Si el consumidor actualmente tiene sentimientos altos y bajos muy negativos, en un futuro presentará un sentimiento extremadamente negativo, mientras que con sentimientos altos se obtiene justo lo contrario. Todo esto, influirá en las futuras decisiones de compras de vuelos respecto a la aerolínea. Atendiendo a la matriz de confusión que se presenta en la tabla 5 se puede observar como la exactitud en la clasificación es aún mayor:

Tabla 5: Matriz de confusión de test de la compañía Ryanair

Sentimiento	0.0	3.0	-1.0	-2.0	-5.0	4.0	-4.0	5.0	-3.0	1.0	2.0
0.0	500	0	0	0	0	0	0	0	0	0	0
3.0	4	28	0	0	0	1	0	0	0	0	7
-1.0	0	0	28	5	0	0	0	0	10	10	1
-2.0	1	0	7	25	0	0	0	4	0	0	0
-5.0	1	0	0	0	67	0	5	0	0	0	0
4.0	0	7	0	0	0	31	0	0	0	0	0
-4.0	1	0	0	5	0	0	24	3	0	0	0
5.0	0	0	0	0	0	3	0	0	0	0	0
-3.0	0	0	0	2	0	0	0	30	0	0	0
1.0	1	0	3	0	0	0	0	0	36	36	6
2.0	1	6	0	0	0	0	0	0	0	3	26

El número de valores clasificados en este caso es 882 y la fiabilidad es del 91,022%, lo que nos deja ver que el modelo es muy bueno para predecir la clasificación. El indicador Cohen´s Kappa por su parte es 0,872.

Sin embargo, dado que la fiabilidad de la matriz de Ryanair es más elevada, se entiende que el algoritmo predice mejor los sentimientos en Ryanair y que por lo tanto el modelo es más exacto. Este hecho puede deberse a que usualmente existen más noticias acerca de la compañía que de SouthWest, debido a los múltiples incidentes ocurridos en los últimos años como los aterrizajes forzosos, las despresurizaciones de cabina o falta de combustible, lo que provocó que los pasajeros realizasen millones de comentarios en redes sociales.

6.Conclusiones

Los resultados analizados muestran que hemos obtenido un buen modelo de clasificación para predecir los sentimientos en ambas compañías. Si bien es cierto, que el modelo es más preciso en Ryanair ya que, la matriz de confusión presenta un 91% de fiabilidad en el conjunto de test mientras que en SouthWest es del 77%. Este hecho se ha podido deber a que, aunque ambas aerolíneas tienen un perfil similar de bajo coste, Ryanair ha tenido a lo largo de los últimos años numerosos incidentes que ha provocado una gran cantidad de noticias y comentarios en la web y que, por ello, el modelo pueda captar mejor los sentimientos respecto a ésta.

Finalmente, una vez analizadas las características y ámbitos de aplicación del Big Data este trabajo ha intentado resaltar que en numerosas ocasiones no se trata de la cantidad de información de la que se disponga, sino de cómo se materializan esos datos para dar valor a las personas. Según Philip &Zhang (2014), los volúmenes masivos de datos podrían producir 250 mil millones de euros para la Administración Pública Europea y generar un superávit potencial de 600000 millones de

dólares anuales a nivel mundial. Asimismo, el informe del Instituto Mckinsey (Manyika et al., 2011) afirma que el uso efectivo de Big Data posibilitaría la transformación de las economías a través de un crecimiento renovado y el estudio de Accentura and General Electric en 2014 expone que el 87% de las compañías creen que este nuevo tipo de análisis cambiará totalmente el panorama de las industrias en los próximos años (Fondevila Gascón et al., 2017).En definitiva, tal y como expone Pat Gelsinger [3](2017): “los datos son la nueva ciencia. El Big Data son las respuestas”.

7.Bibliografía

Cukier, K. (2010). Data, data everywhere: A special report of non managineting information. The economist newspaper. Recuperado de: https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data- everywhere.pdf.

Fernández, V., García, X. (Eds), Media and Metamedia Management. Advance in Intelligent Systems and Computing, Vol 503. Switzerland, Spring.

Fondevila Gascón, J.M., Mir, P., Santana, E., E., Rom, J. (2017). New Digital Metrics in Marketing: A comparative study on Social Media Use. En: Freire, F., Rúas Araújo, X., Martinez

Gandomi, A., Haider, M. (2015). Beyond thehype: Big Data concepts, methods, and analytics. International Journal of Information Management, Vol. 35, pp. 137-144.

Han, J. Kamber, M. (2006). Data Mining: Concepts and Techniques. Waltham. Elsevier.

He, W., Wu, H., Yan, G., Akula, V., Shen, J. (2015). A novel social media competitive analytics framerowkframe work with sentiment benchmarks. Information & Management. Vol. 52, No. 7, pp. 801-812.

Jagadish, H.V., Gehrke, J., Labrindis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishman, R., Shahabi, C. (2014). Big Data and its technical challenges. ACM, Vol. 57, No. 7, pp. 86-94.

Lambrecht, A., Tucker, C. (2015). Can Big Data Protect a Firm from Competition? Recuperado de: http://ec.europa.eu/information_society/newsroom/image/document/2016-6/computer_and_communications_industry_association_-_can_big_data_protect_a_firm_from_competition_13846.pdf.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., HungByers, A. (2011). Big Data: The Next Frontier for innovation, competition and productivity. Mckinsey Global Institute.

Nasukawa, T., Yi, J. (2015). Sentyment analysis: Capturing favorability using natural language processing. Paper presented at the 2nd International Conference on Knowledge Capture: Sanbel Island, USA.Recuperado de https://www.researchgate.net/publication/220916772.

Philip Chen, C.L., Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciencies, Vol. 275, pp. 314-347.

Ward, J.S., Barker, A. (2013). Undefined By Data: A Survey of Big Data definitions. School of Compute Science. University of StAndrews, Uk.

Xuan, J., Xu, Z. (2016). Sentiment computing for the News Event based on the Big Social Media Data. Disponible en: https://www.researchgate.net /publication /308204666.

Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B., Vasilakos, A.V. (2016). Big data: From beginning to future. International Journal of Information Management, Vol. 36, pp. 1231 - 1247.

Wan, X. (2009). Co-training for cross- lingual sentiment classification. Paper presented at the 47th Annual Meeting. Singapore.

Agradecimientos

Agradecimientos a mis tutoras Encarnación Álvarez Verdejo y Mari Carmen Pegalajar Jimenez por su calidad profesional y personal

[3] https://bigdata299.wordpress.com/

7.Anexos

Figura 5: Árbol de decisión de la compañía SouthWest

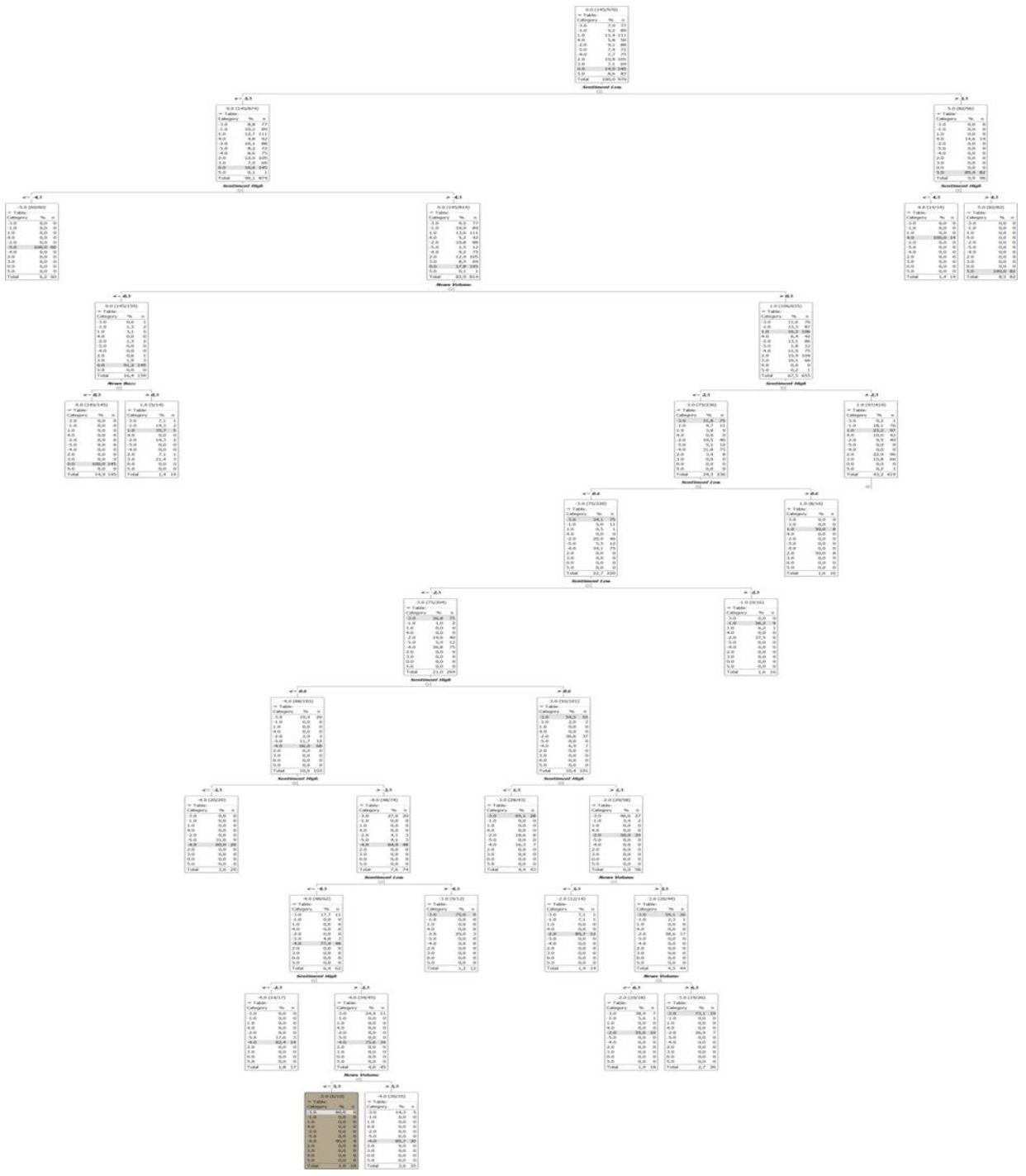
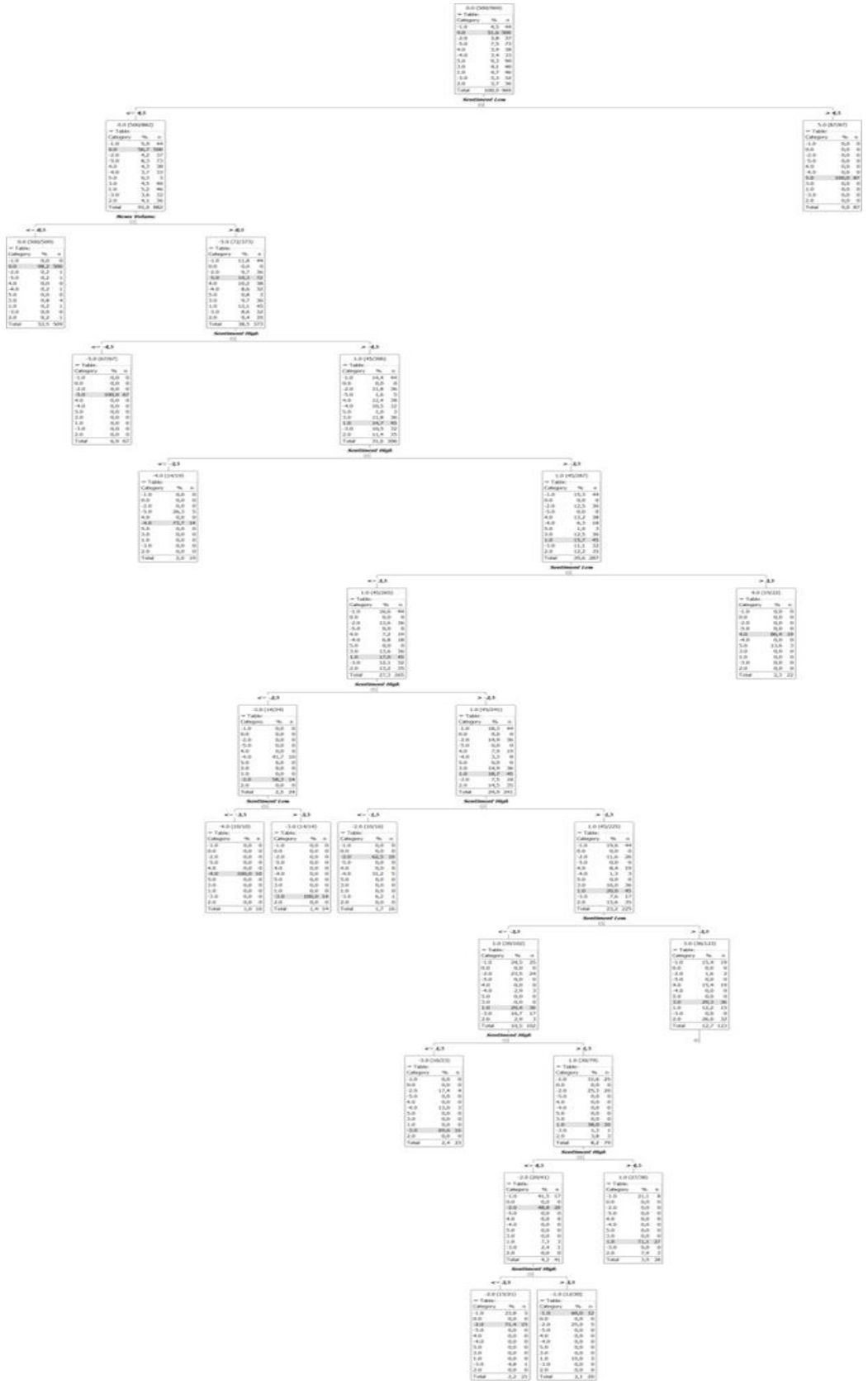



Figura 6: Árbol de decisión de la compañía Ryanair