Revista de Turismo e Identidad
a.2 n.1 – jun 2021 – nov 2021 Mendoza, Argentina
ISSN 2718 – 8205
http://revistas.uncu.edu.ar/ojs3/index.php/turismoeindentidad
pp. 55-71

Análisis de sentimientos usando la red social Twitter ¿qué sintieron los turistas que volaron en 2020 con seleccionadas aerolíneas sudamericanas?

Sentiment analysis using the social network Twitter
What did tourists feel flying in 2020 with selected southamerican airlines?

Cristian von Matuschka

Universidad Nacional de Cuyo - Facultad de Filosofía y Letras
Instituto de Investigaciones en Turismo e Identidad. Mendoza. Argentina
cvonmatuschka@ffyl.uncu.edu.ar

RESUMEN

La activación del turismo es uno de los asuntos claves más importantes para la industria aeronáutica. Internet contiene mucha información sobre el turista. Analizar dicha información es una tarea significativa y desafiante. En este trabajo, se propone analizar la opinión de los turistas que viajaron en determinadas aerolíneas sudamericanas, mediante la técnica de análisis de sentimientos, a través del estudio de los mensajes de sus clientes. El recurso a utilizar para el análisis es la información en twitter, creada por clientes de determinadas dichas aerolíneas. Primero, se presenta un método para extraer las frases publicadas relacionadas con las ubicaciones de destino y los “hashtags”. Luego, se analizó la

polaridad de los tweets extraídos; creando opiniones positivas, negativas y eventualmente neutras. Para el proceso, se empleó un enfoque de aprendizaje automático sin supervisión que utiliza palabras semilla. El resultado experimental sobre la clasificación muestra la eficacia del método aplicado. Se adjuntan los resultados preliminares (descriptivos) así como la propuesta base para un modelo predictivo

PALABRAS CLAVE: aprendizaje automático; turismo 2020; análisis de sentimientos; industria aeronáutica

ABSTRACT

Activation of tourism is one of the key subjects for the airline industry. Internet contains a lot of information about tourists. This paper aims at analyzing the opinion of the tourists who traveled by certain South America airlines, using the sentiment analysis technique, employed in the study of their messages. The resource used for analysis is the information in twitter, provided by these airlines customers. First, a method for extracting published phrases related to target locations and "hashtags" was presented. Then, it was analyzed the polarity of the tweets extracted; creating positive, negative and eventually neutral opinions. In this process, there was utilized an unsupervised learning technique using seed words. The experimental result on the classification shows the efficacy of the applied method. Preliminary (descriptive) results as well as the basic proposal for a predictive model are herein attached.

KEYWORDS: machine learning; tourism 2020; sentiment analysis; airline industry; twitter

Fecha recepción: 8 de febrero de 2021
Fecha aprobación: 1 de junio de 2021

Introducción

Los vuelos comerciales en el turismo son de suma importancia ya que reactivar la industria aeroportuaria conduce al fomento de industrias y comunidades locales. En esta situación, la World Wide Web desempeña un papel importante (Saito, 2011). Cada vez es más significativo encontrar información relacionada con las necesidades de los destinatarios. La propuesta es extraer información sobre los turistas que viajaron en avión y

 CC BY-NC 2.5 AR

usaron Twitter para expresar sus experiencias, analizar la información desde las perspectivas vistas en trabajos similares (Pang et al., 2008; Go et al., 2009) y visualizar el resultado del análisis mediante tablas y gráficos. Dichas perspectivas hacen referencia al tipo de aprendizaje que en este trabajo es el aprendizaje automático mediante algoritmos. Siendo Twitter uno de los servicios de microblogging más usados, con mensajes de texto de hasta 280 caracteres, sus frases publicadas se describen como "tweets". En ese tipo de servicios de “microblogging”, los usuarios tienden a publicar tweets en tiempo real. Esto denota que los tweets a menudo contienen información significativa para el turismo como datos sobre lugares, eventos, opiniones, etc. (Tokuhisa et al., 2011). Los estudios anteriores más recientes (Go et al., 2009; Balabantaray et al., 2012; Mohammad et al., 2013) abordaron el problema del análisis de los sentimientos en tweets por ejemplo, mediante el uso de textos o características específicas de Twitters como emoticones, hashtags, URLs, @, símbolos y el uso de mayúsculas. El enfoque basado en embeber palabras específicas de sentimientos (Tang et al., 2014) considera al corpus. Esto significa la inclusión de pistas más afectivas que los vectores regulares de palabras y por ende la producción de un mejor resultado.

El objetivo es determinar si es posible detectar sentimientos de turistas usando los tweets para luego construir un modelo predictivo. El objeto de estudio son los clientes de las aerolíneas comerciales que volaron durante los meses completos de enero a octubre de 2020, usando las siguientes aerolíneas: Azul Linhas Aéreas (Brasil), Avianca (Colombia), LATAM (Chile), Sky Airline (Chile), Aerolíneas Argentinas (Argentina) y TAME (Ecuador). La Tabla 1 muestra la cantidad de tweets analizados por aerolínea (y luego de la depuración).

Tabla 1. Cantidad total de tweets por aerolínea sudamericana

Nombre de la aerolínea	Total
Azul Linhas Aéreas (Brasil)	4774
Avianca (Colombia)	3734
Sky Airline (Chile)	3424
LATAM (Chile)	3044
Aerolíneas Argentinas (Argentina)	2943
TAME (Ecuador)	573

Fuente: elaboración propia

Metodología

La investigación es de índole no experimental ya que no se manipula la variable independiente, esto significa que se observan los fenómenos tal y como se dan en su contexto natural, para después analizarlos. El estudio es de corte longitudinal por analizar distintos intervalos de tiempo. Y es descriptiva ya que se busca interpretar los sentimientos reflejados por los turistas. La hipótesis podría escribirse de la siguiente manera:

> H0: La aplicación de técnicas de análisis de sentimientos permitiría identificar la polaridad de los mensajes de clientes en redes sociales (en este caso específico, a través de twitter), y por consiguiente clasificar las múltiples opiniones sobre los servicios ofrecidos por las aerolíneas analizadas de manera desatendida.

Para la minería de comentarios de turistas se tomó de base a la página de Twitter (https://www.twitter.com). El trabajo se realizó con la Interfaz de programación de aplicaciones (API) oficial (https://developer.twitter.com/en/docs) y un cuaderno de trabajo vinculado al software de analítica de datos KNIME (https://www.knime.com/software-overview), el mismo consistió en tres partes:

 CC BY-NC 2.5 AR

1) En la etapa de extracción de datos, se construye una base de datos sin manipulación. En esta parte, se aplicó la propuesta metodológica de minería de datos basada en Shafer et al. (2000) que sirve para extraer las palabras de los comentarios del cliente y definir las palabras más utilizadas. Luego crear un conjunto de características para construir el modelo de análisis de sentimientos.

2) Luego la etapa de análisis consiste en la descripción y clasificación del texto. Se calcula para los datos de prueba un coeficiente de correlación. El mismo se obtiene usando la columna semilla y la columna de cada palabra más utilizada. Esto significa una medida de la correlación de las dos variables que según Lai et al. (2019) que nos habilita para el uso de los datos de prueba en la etapa de clasificación.

3) Para la clasificación se determina la mejor herramienta clasificadora. En este estudio se compara un árbol de decisión con el atributo objetivo nominal y un clasificador de Bayes. En el caso del árbol se aplica un método de poda para reducir el tamaño del árbol y aumentar la precisión de la predicción. El método de poda se basa en el principio de la "longitud mínima" de descripción (Rokatch et al., 2008).

Etapa de extracción de datos

En esta sección, se describe en detalle el método usado para preparar la base de datos. Se comienza con las palabras relacionadas seleccionadas. Se definieron entonces los nombres de las instalaciones, destinos o sucesos como "consultas básicas". Por ejemplo, "aerolínea", "vuelo" y "aeropuerto" son consultas básicas. El número de consultas básicas en esta investigación es de aproximadamente 300 palabras debido sobre todo al nombre de las aerolíneas involucradas y los aeropuertos. El lenguaje utilizado es exclusivamente el castellano.

Se ha detectado que la gente no siempre menciona en las consultas básicas, los nombres de las instalaciones o eventos. Para resolver este problema, se generó manualmente abreviaturas de consultas. Las cuestiones semánticas también fueron consideradas, por ejemplo, agregando al listado de consultas palabras como: "vuelos2020", "vuelosen2020" y "volaren2020". Por lo tanto, se realizó una expansión de la consulta significando una división de cada frase en palabras. Para el proceso se utilizó segtok (https://pypi.org/project/segtok/), un segmentador de oraciones basado en reglas (splitter) y un llamado “tokenizer” de palabras usando características ortográficas.

A continuación, se aplicó el post-procesamiento al proceso de extracción; es decir, filtrado. Los tweets extraídos no siempre se relacionan con el objeto de estudio, aunque contengan consultas reales y válidas. Por ejemplo, la palabra "pasajeros" se utiliza no sólo para el turismo, sino también en un contexto temporal. Por lo tanto, hay que eliminar los datos de ruido. Para el proceso, se utilizó la salida de segtok y un enfoque basado en reglas. Los resultados en segtok incluyen algunos indicadores como la etiqueta del habla; por ejemplo, [ProperWord-verbo] y [ProperWord-sustantivo]. Al usar estas etiquetas, se juzgó si cada tweet es adecuado o no. Además, se prepararon algunas reglas de sufijo manualmente y se utilizaron en este proceso. También se removieron datos incompletos, redundantes o anómalos bajo un criterio manual.

La nube de palabras es una gran herramienta para visualizar datos de PNL (como planteado en Higashiyama et al., 2008). Cuanto más grandes son las palabras en la imagen, mayor es la frecuencia de esa palabra en la base de datos textual. En la Figura 1 se pueden observar palabras como “vuelo cancelado”, “ayuda”, “por favor”, totalmente relacionadas a quejas u opiniones negativas. Mientras que otras como “avión”, “reserva”, “personal” requieren un análisis más profundo.

 CC BY-NC 2.5 AR

Figura 1. Nube de palabras para los sentimientos negativos en tweets

Fuente: elaboración propia

En la Figura 2 se observa lo mismo que en la Figura 1, pero ahora para aspectos positivos. Algunos ejemplos son: “gracias”, “mejor”, “amor”, “mejor aerolínea.

Figura 2. Nube de palabras para los sentimientos positivos en tweets

Fuente: elaboración propia

Este tipo de análisis del texto que usa diccionarios de palabras sirve para otorgar una idea de primera instancia. En la investigación se busca mayor precisión combinándolo con el método P/N.

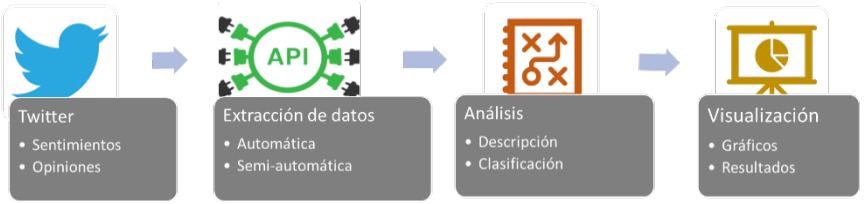
Etapa de análisis

La hipótesis de trabajo está relacionada al análisis de sentimientos. El análisis de sentimientos es una técnica que utiliza el procesamiento de lenguaje, análisis de texto y herramientas computacionales para clasificar comentarios subjetivos de diferentes usuarios. Esta técnica moderna es muy demandada y pertenece al Procesamiento del Lenguaje Natural (PNL) (Pang et al., 2008). Una manera de realizar dicho análisis es a través de clasificadores. Una de las clasificaciones más usadas es la clasificación P/N, que consiste en clasificar un sentimiento de entrada en opiniones positivas (P) o negativas (N) (Hermanto et al., 2018). En esta investigación, se aplicó un enfoque de aprendizaje automático sin supervisión basado en un clasificador de Bayes teniendo en cuenta también la tercera posibilidad (opiniones neutras).

Esta investigación, se centra en un enfoque de aprendizaje automático no supervisado, que no necesita datos de entrenamiento anotados manualmente. Se basa en palabras semilla y datos de entrenamiento pseudo extraídos de un corpus no agregado. Turney (2002) ha propuesto un método para clasificar las revisiones como recomendadas o no, mediante el uso de algunas palabras semilla. Usó las palabras "excelente" y "pobre" como palabras semilla, y calculó la orientación semántica de una frase usando la Información Mutua de Pointwise (PMI) entre la palabra semilla y la frase. Esto ayuda a evitar el uso exclusivo de expresiones lingüísticas, como "bueno" y "malo", ya que los tweets en Twitter son informales y a menudo contienen muchos emoticones y símbolos. Para ello se usa el test no paramétrico de Pearson (Lai et al., 2019). Los valores que faltan en una columna se ignoran de tal manera que para el cálculo de la correlación entre dos columnas sólo se tienen en cuenta los registros completos. Aquí el valor de esta medida oscila entre -1 (fuerte correlación negativa) y 1 (fuerte

correlación positiva). Un valor de 0 no representa correlación lineal (aunque las columnas podrían ser muy dependentes entre sí). La Figura 3 muestra el esquema propuesto de esta investigación.

Figura 3. El esquema del procedimiento para analizar sentimientos



Fuente: elaboración propia basado en Kobayashi et al. 2004

Etapa de clasificación

La clasificación P/N, significa clasificar una entrada en opiniones positivas o negativas. Muchos investigadores han estudiado diversos enfoques para la identificación P/N de una entrada. Por ejemplo, Pang et al. (2002) han reportado una tarea de clasificación para documentos relacionados a la revisión de películas. Compararon varias técnicas de aprendizaje automático y mostraron la eficacia de los datos en la revisión. En general, las técnicas de aprendizaje automático generan un clasificador de alta precisión mediante el uso de una enorme cantidad de datos de entrenamiento. Sin embargo, construir dichos datos es costoso.

Se utilizó entonces el "clasificador bayesiano ingenuo" como un medio para clasificar los tweets en positivos o negativos. El método usa un modelo probabilístico basado en el teorema de Bayes.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

Aquí se puede usar sólo el numerador de la fracción ya que el denominador P (d) no depende de c.

$$\hat{c} = argmax_c P(c) \prod_{i=1}^{n} P(x_i|c)$$

Donde c es la clase (P/N) y xi es una palabra en una oración.

Luego se desea evaluar el método con un conjunto de datos de prueba. El método necesita un corpus no marcado para la adquisición de datos de entrenamiento. El corpus depurado consistió en 18.492 mil tweets. La base se divide en la relación estándar de Pareto 80/20 (Shimada et al., 2010). Los datos de prueba consistieron entonces en 3.698 tweets. Las Tablas 2 y 3 muestran los resultados de cada clasificador. Dichas figuras representan la matriz de confusión para la predicción de los sentimientos versus los sentimientos reales (negativo, neutral y positivo).

Tabla 2. Clasificador de árboles de decisión (accuracy=0,62)

		Predictivo		
		Negativos	Neutral	Positivos
Real	Negativos	1.583	306	232
	Neutral	300	361	172
	Positivos	206	174	364

Fuente: elaboración propia.

Tabla 3. Clasificador bayesiano (accuracy=0,69)

		Predictivo		
		Negativos	Neutral	Positivos
Real	Negativos	1.765	268	201
	Neutral	228	409	149
	Positivos	146	153	379

Fuente: elaboración propia

Por lo tanto, si comparamos ambas figuras vemos que el clasificador bayesiano otorga la mejor puntuación de precisión, puntuaciones de

 CC BY-NC 2.5 AR

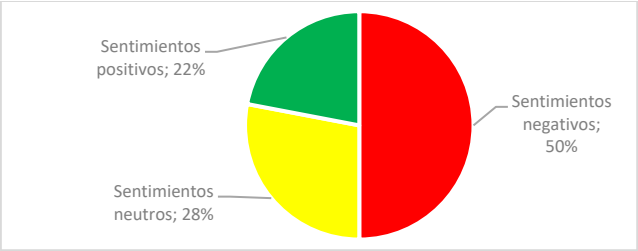
precisión de acuerdo con el informe de clasificación (mayor intensidad de verde representa mayor precisión).

Se optó por usar entonces el clasificador de Bayes a diferencia de los árboles de decisión debido a su mayor precisión clasificatoria (árboles de decisión 0,62 versus bayesiano con 0,69). Esto tiene concordancia con estudios similares (Kazutaka et al., 2009).

Resultados

En la Figura 4 se ven los resultados de la clasificación usando el método Bayesiano, siendo aquellos tweets que no son ni positivos ni negativos, los llamados "neutros".

Figura 4. Distribución total de la clasificación

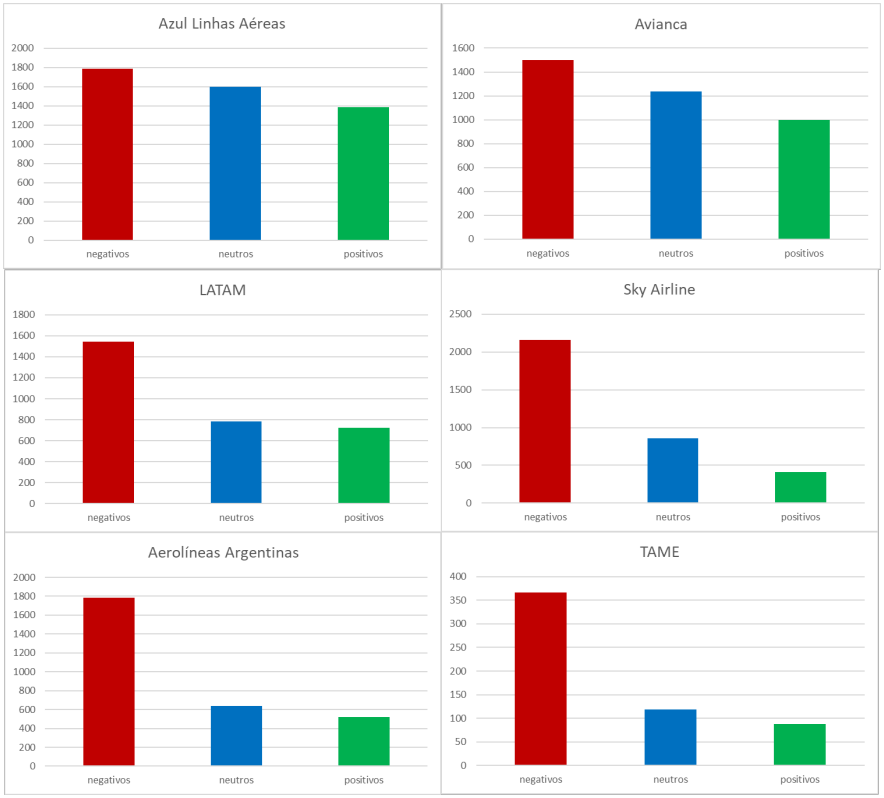


Fuente: elaboración propia

Una vez consolidada la base de datos, se procede a transformarlos en información. Para ello se usa la visualización a través de gráficos. Se comienza por la clasificación por aerolínea. La figura 5 muestra que en general se puede decir que hay más tweets negativos que positivos y neutros. Esto se debe a varias razones. La primera razón es que los turistas suelen comunicar más sus problemas usando Twitter. Este comportamiento es generalizado en el rubro de Atención al Cliente. Estudios muestran que los clientes tienden a comunicar más cuestiones negativas que positivas (Simon et al., 2014). Ahora bien, en la Figura 5 se puede observar también que TAME, Aerolíneas Argentinas y Sky Airline tienen sustancialmente reacciones negativas. Estas

aerolíneas podrían considerarse como aquellas con menos satisfacción comparativa del cliente; durante su servicio en pandemia. En comparación con otras como Azul Linhas Aéreas y Avianca que tienen los tweets más balanceados.

Figura 5. Clasificación de sentimientos por aerolínea

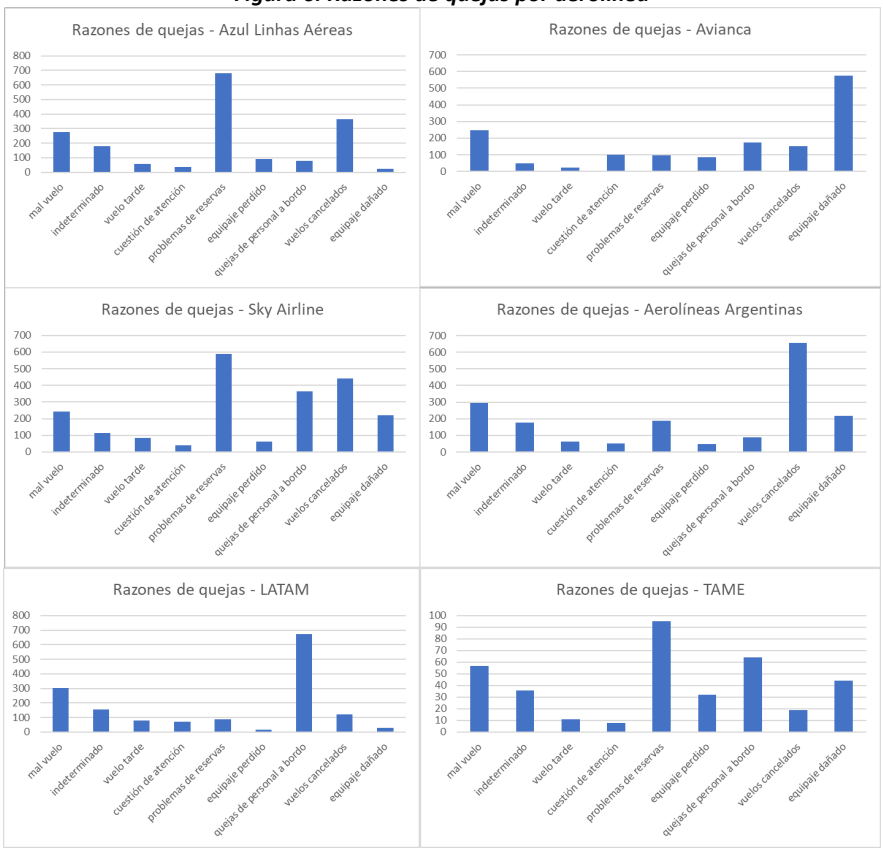


Fuente: elaboración propia

De interés para esta investigación es detectar la razón de por ejemplo cada tweet calificado como negativo. Para ello se explora entonces la columna de razones negativas de la base de datos para extraer conclusiones sobre los

sentimientos negativos en los tweets de los clientes. Dicha columna se construye usando la salida de segtok y el enfoque basado en reglas (ver extracción de datos). La Figura 6 es un resumen de las razones más frecuentes según cada aerolínea.

Figura 6. Razones de quejas por aerolínea



Fuente: elaboración propia

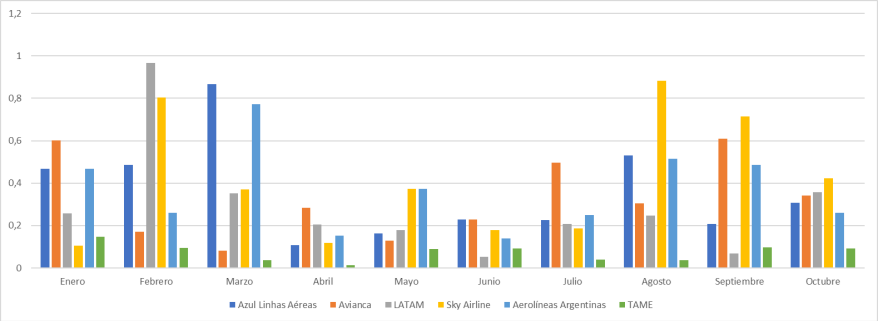
Ahí se observa que hay ciertas temáticas muy recurrentes. Un ejemplo es “problemas de reservas”, razón con más frecuencia en Azul Linhas Aéreas, Sky Airline y TAME. Otra situación es la de LATAM con “quejas de personal a

bordo". En este caso LATAM muestra la mayor frecuencia de todas las aerolíneas. Mientras que los "vuelos cancelados" parecen ser el tema central de las quejas en Aerolíneas Argentinas.

Interesante es analizar el tema "mal vuelo". Ahí se nota que en cada aerolínea existe en una relativa fuerza. Bajo esa categoría se agrupa por lo general lo relacionado al trayecto: turbulencia, servicios básicos dentro del avión, etc.

Otra visualización de interés en este trabajo es la temporal. Como la base contiene datos del 01-01-2020 al 31-10-2020, la idea aquí es mostrar los sentimientos de los tweets para cada fecha y cada aerolínea, ver Figura 7.

Figura 7. Sentimientos negativos por mes en relación al total de viajes



Fuente: elaboración propia

Se observan dos focos grandes de sentimientos negativos. El primero al principio de la pandemia probablemente debido a la explosión de casos (febrero y marzo). Luego hacia el final del año (agosto, septiembre y octubre) donde se retoma la circulación con los diversos protocolos de cada país.

Cabe señalar que el número total de tweets para TAME fue significativamente menor en comparación con el resto de aerolíneas, y por lo tanto posee menos tweets negativos.

Los tweets negativos de Aerolíneas Argentinas y Sky Airlines en el mes de mayo donde no hubo vuelos regulares comerciales tiene su origen en que las quejas de los clientes se relacionan justamente a los problemas con las reservas y comunicación de las políticas de cancelación.

Conclusiones

Para determinar la validez de la hipótesis se recurrió a comparar el ajuste del clasificador por árboles de decisión con el clasificador bayesiano (ajuste estadístico de 0,62 y 0,69 correspondiéndose con el mismo resultado para Tabla 2 y 3). Esto indica que podría ser aceptada la hipótesis H0. No obstante a ello, se necesita manejar la información contextual para ganar aún mayor precisión en el análisis. Una de estas informaciones contextuales es la relacionada al lugar de residencia del usuario de cada tweet. Otra es la información en los tweets antes y después del tweet de destino (el tweet analizado en este estudio). Es por esta razón que el método utilizado de Análisis de Sentimientos es apropiado para el llamado nivel de entendimiento circunscripto al "mensaje y frase" (Pozzi et al., 2017). Del análisis se aprecia la cuantía de sentimientos negativos para con los positivos o neutros y sus posibles causas.

Para el nivel de entendimiento circunscripto a la "entidad y aspecto" se recomienda recurrir al Análisis de Redes Sociales, lo que permite considerar los datos contextualizados, aumentando su valor explicativo y predictivo (West et al., 2014; Tan et al. 2011; Pozzi et al. 2013). Integrar este tipo de datos es un trabajo de perfeccionamiento a futuro y por ende se procede a aceptar solo de manera tentativa la hipótesis planteada.

Bibliografía y referencias

Balabantaray, R., Mohammad, M. & Sharma, N. (2012): *Multi-class twitter emotion classification: a new approach*, Int. J. Appl. Inform. Syst. 4 (1), 48–53.

Go, A., Bhayani, R. & Huang, L. (2009): *Twitter sentiment classification using distant supervision,* in: CS224N Project Report, Stanford 1, 12.

Hermanto, D., Ziaurrahman, M., Bianto, M. & Setyanto, A. (2018): *Twitter Social Media Sentiment Analysis in Tourist - Destinations Using Algorithms Naive Bayes Classifier*, Journal of Physics, p. 2.

Higashiyama, M., Inui, K. & Matsumoto, Y. (2008): *Acquiring noun polarity knowledge using selectional preferences*. 14nd Annual Meeting of the Association for Natural Language Processing, pp. 584– 587.

Kazutaka, S., Inoue, S., Maeda, H. & Endo, T. (2009): *Analyzing Tourism Information on Twitter for a Local City*, Hiroshi Maeda and Tsutomu Endo, Department of Artificial Intelligence, Kyushu Institute of Technology, p. 5.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. & Fukushima, T. (2004): *Collecting evaluative expressions for opinion extraction*. Proceedings of the First International Joint Conference on Natural Language Processing, pp. 584–589.

Lai, C., Tao, Y., Xu, F., Wing, W., Jia, Y., Yuan, H. & Locatelli, G. (2019): *A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty*. Information Sciences. pp. 58–77.

Mohammad, S.M., Kiritchenko, S. & Zhu, X. (2013): *NRC-Canada: building the state-of-the-art in sentiment analysis of tweets*, in: Joint Conference on Lexical and Computational Semantics.

Pang B., Lee L. & Vaithyanathan S. (2002): *Thumbs up? sentiment classification using machine learning techniques*. Conference on Empirical Methods in Natural Language Processing, pp. 79–86.

Pang, B. & Lee, L. (2008): *Opinion mining and sentiment analysis.* Foundations and Trends in Information Retrieval.

Pozzi, F.A., Maccagnola, D., Fersini, E. & Messina, E. (2013*): Enhance user-level sentiment analysis on microblogs with approval relations*, in: AI*IA 2013: Advances in Artificial Intelligence, Springer, New York, pp. 133–144.

Pozzia, F.A., Fersinib, E., Messinab, E. & Liuc, S. (2017): *Challenges of sentiment analysis in social networks: an overview*, Sentiment Analysis in Social Networks, pp. 16-17.

Rokach, L. & Maimon, O. (2008): *Data mining with decision trees: theory and applications*. World Scientific Pub.

Saito, H. (2011): *Analysis of tourism informatics on we*b. Journal of the Japanese Society for Artificial Intelligence, pp. 234–240.

Shafer, J., Agrawal, R. & Mehta. M. (2000): *A Scalable Parallel Classifier for Data Mining*. Springer.

Shimada, K., Yamaumi, M., Tadano, R., Hadano, M. & Endo, T. (2010): *Interactive aspect summarization using word-aspect relations for review documents.* 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems, pp. 189–188.

Simon, D. & Gómez, M. (2014): *Customer Satisfaction, Competition, and Firm Performance*, Managerial and Decision Economics. Vol. 35, No. 6, pp. 371-386.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M. & Li, P. (2011): *User-level sentiment analysis incorporating social networks, in*: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1397–1405.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. & Qin, B. (2014): *Learning sentiment-specific word embedding for twitter sentiment classification,* in: Annual Meeting of the Association for Computational Linguistics.

Tokuhisa, M., Okumura, H. & Murata, M. (2011): *Sentiment analysis of weblog articles to support tourism development.* Journal of Society for Tourism Informatics, 7(1), pp. 85–98.

Turney, P. (2002): *Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews*. 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424.

West, R., Paskov, H.S., Leskovec, J. & Potts, C. (2014): *Exploiting social network structure for person-to-person sentiment analysis,* Trans. Assoc. Comput. Linguist. 2, pp. 297–310.