Universidad Tecnológica Nacional
ABRIL 2020 / Año 18- Nº 37

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

Creación de un Corpus de Opiniones con Emociones usando Aprendizaje Automático

Creation of a corpus of opinions with emotions using machine learning

Presentación: 22/10/2019

Aprobación: 05/12/2019

Alejandra C. Cardoso
Facultad de Ingeniería – Universidad Católica de Salta - Argentina
acardoso@ucasal.edu.ar

María Lorena Talamé
Facultad de Ingeniería – Universidad Católica de Salta- Argentina
mltalame@ucasal.edu.ar

Matías N. Amor
Facultad de Ingeniería – Universidad Católica de Salta- Argentina
mnamor@ucasal.edu.ar

Agustina Monge
Facultad de Ingeniería – Universidad Católica de Salta
agum_96@hotmail.com

Resumen

La identificación de los sentimientos expresados en opiniones textuales puede entenderse como la categorización de los mismos según sus características, y resulta de gran interés en la actualidad. El aprendizaje supervisado es uno de los métodos más populares para la clasificación textual, pero se necesitan muchos datos etiquetados para el entrenamiento. El aprendizaje semi supervisado supera esta limitación, ya que implica trabajar con un pequeño conjunto de datos etiquetados y otro mayor sin etiquetar. Se desarrolló un método de clasificación de textos que combina ambos tipos de aprendizajes. Se recopilaron textos breves u opiniones de la red social Twitter, a los que se aplicaron una serie de acciones de limpieza y preparación, para luego clasificarlos en cuatro sentimientos: ira, asco, tristeza y felicidad.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

La precisión y recall obtenidos con el método fueron satisfactorios y como consecuencia, se logró obtener un corpus de mensajes categorizados según el sentimiento expresado.

Palabras claves: clasificación de textos, aprendizaje semi supervisado, Twitter

Abstract

The identification of feelings expressed in textual opinions can be understood as the categorization of them according to their characteristics, and is of great interest today. Supervised learning is one of the most popular methods for textual classification, but a lot of labeled data is needed for training. Semi-supervised learning overcomes this limitation, as it involves working with a small set of labeled data and a larger unlabeled data set. A text classification method was developed that combines both types of learning. Short texts or opinions from the social network Twitter were compiled, to which a series of cleaning and preparation actions were applied, and then classified into four feelings: anger, disgust, sadness and happiness. The precision and recall obtained with the method were satisfactory and as a result, a corpus of messages categorized according to the expressed feeling was obtained.

Keywords: text classification, semi-supervised learning, Twitter

INTRODUCCION

El análisis de sentimientos, también llamado minería de opiniones, consiste en el estudio de textos subjetivos, es decir, textos donde el emisor expresa una opinión o sentimiento. Un caso particular son las opiniones que generan usuarios en redes sociales, blogs, portales de noticias, foros, etc. Las mismas ayudan a revelar información importante sobre un tema específico apoyando, por ejemplo, a la inteligencia de negocios, y jugando un papel importante en la toma de decisiones. Una forma de monitorear la opinión de los usuarios sobre determinado producto o tema es proponer encuestas o resaltar comentarios de otros usuarios, por ejemplo, otorgando puntaje. Si bien esto puede ser válido, la opinión textual de los usuarios sobre un tema en cuestión se la puede clasificar por su polaridad, es decir, si es positiva o negativa, pero también se puede realizar una clasificación más avanzada identificando emociones.

La identificación del sentimiento de un usuario, expresado en un mensaje textual, puede entenderse como clasificar o categorizar el mensaje según las características del mismo. El aprendizaje supervisado se encuentra entre los métodos más populares para la categorización de textos. Los algoritmos utilizados para esta tarea pueden ser entrenados con un conjunto de datos suficientemente grande y previamente etiquetado con las categorías correspondientes. El etiquetado por parte de una o más personas expertas en el dominio de los datos es un proceso engorroso, lento y propenso a errores. En este punto, los algoritmos semi supervisados resultan un aporte al aprendizaje, ya que son capaces de aprender a partir de dos conjuntos de datos: uno etiquetado y otro sin etiquetar (Witten et al., 2017).

El trabajo descrito en este artículo forma parte de un proyecto de mayor alcance que tiene

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

como objetivo la detección de emociones en textos. Ante la imposibilidad de contar con un corpus suficientemente grande y clasificado según las emociones, fue necesario obtener un corpus etiquetado altamente confiable, cuyo método se describe en este artículo. Los textos fueron obtenidos de la red social Twitter; se clasificaron en cuatro categorías o sentimientos: ira, felicidad, asco y tristeza. Debido a las características de este tipo de textos -sin formato, informales y subjetivos- fue necesario un paso previo de limpieza y preparación, a fin de mantener la mayor cantidad de información posible y descartar aquello que pudiere producir ruido en el análisis posterior.

ANÁLISIS DE SENTIMIENTOS

Muchas investigaciones relacionadas al análisis de opiniones o sentimientos se centraron principalmente en la detección de la polaridad de la opinión, es decir, determinar si la misma es positiva o negativa (algunos ejemplos en (Selva Castelló, 2015) (Amores Fernández, 2016)). Con frecuencia, también se considera la opinión neutra para diferenciar aquellos textos objetivos. Son escasos los estudios enfocados en detectar algún sentimiento particular en los textos. En algunos trabajos se crearon diccionarios de palabras asociadas a un sentimiento, tal es el caso de (Diaz Rangel et al., 2014), en el cual se catalogaron una serie de palabras según seis sentimientos: alegría, enojo, miedo, tristeza, sorpresa y repulsión. Otro recurso lingüístico se encuentra en (Mohammad, 2018), que define un diccionario de palabras con una puntuación en función de las dimensiones emocionales: valencia, dominancia y activación. En (Aguado de Cea et al., 2012) se clasificaron textos de redes sociales en ocho categorías: satisfacción, insatisfacción, confianza, temor, amor, odio, felicidad y tristeza, utilizando una base de datos de sustantivos y verbos combinadas con reglas gramaticales. En estos casos se plantearon diversas formas de utilizar las puntuaciones de las palabras para clasificar textos. Aunque este enfoque resulta válido, en este trabajo, se prefirió usar una colección de mensajes categorizados teniendo en cuenta su completo sentido y no por la valoración de cada una de sus palabras. Si bien se encontraron colecciones de tweets clasificados, como por ejemplo en TASS[1] (Taller de Análisis Semántico de la SEPLN), estos están anotados con diferentes grados de polaridad y no con emociones, otra colección disponible surge de (Navas Loro et al., 2018) pero está relacionada al marketing, por lo cual no responden a los requerimientos que se plantean en esta investigación.

Algunas aplicaciones para el análisis de texto disponibles en la web suelen detectar polaridad, entre ellas, Text Analytics de Microsoft[2] . En cambio, Tone Analyzer[3] de IBM identifica siete emociones, pero sólo está disponible para inglés y francés. Paralelldots[4] dispone de detección de seis emociones en varios idiomas (incluidos el español): happy, angry, excited, sad, fear, bored. Sin embargo, se verificó que deja de lado algunas particularidades de los textos de opinión en redes sociales, tales como el uso de emojis o hashtags.

APRENDIZAJE COMPUTACIONAL

El aprendizaje supervisado consiste en construir automáticamente un clasificador, a partir de un conjunto de textos clasificados previamente, y de manera inductiva aprender las características de las categorías. Aunque esta forma de aprendizaje obtiene una alta precisión necesita una gran cantidad de documentos etiquetados (Han et al., 2012).

El aprendizaje semi supervisado es una técnica del aprendizaje automático que mejora esta limitación ya que aprende a clasificar a partir de un número escaso de ejemplos etiquetados y otros no etiquetados. Los ejemplos etiquetados se usan para aprender modelos que

1 http://www.sepln.org/workshops/tass/2017/
2 https://azure.microsoft.com/es-es/services/cognitive-services/text-analytics/
3 https://tone-analyzer-demo.ng.bluemix.net/
4 https://www.paralleldots.com/text-analysis-apis#emotion

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

caractericen cada clase o categoría, y los ejemplos sin etiqueta se usan para refinar los límites entre las clases (Han et al., 2012). Self-training es una de las formas más simples de clasificación semi supervisada: primero se construye un clasificador usando los datos etiquetados, luego el clasificador categoriza a los datos no etiquetados. Sólo los nuevos ejemplos etiquetados con una confianza que supere cierto umbral se anexan al conjunto etiquetado y el proceso de aprendizaje se repite (Han et al., 2012). Un esquema básico de este proceso se muestra en la Figura 1. El hecho que estos algoritmos semi supervisados utilicen pocas instancias etiquetadas para entrenar hace que que sean atractivos de utilizar en aquellos casos donde el etiquetado sea muy costoso, requiera tiempo o intervención humana. En (Abudalfa and Ahmed, 2019) se evaluaron una serie de algoritmos semi supervisados aplicados a textos de micro-blogs y en idioma inglés, concluyendo que este enfoque es relevante para la clasificación textual.

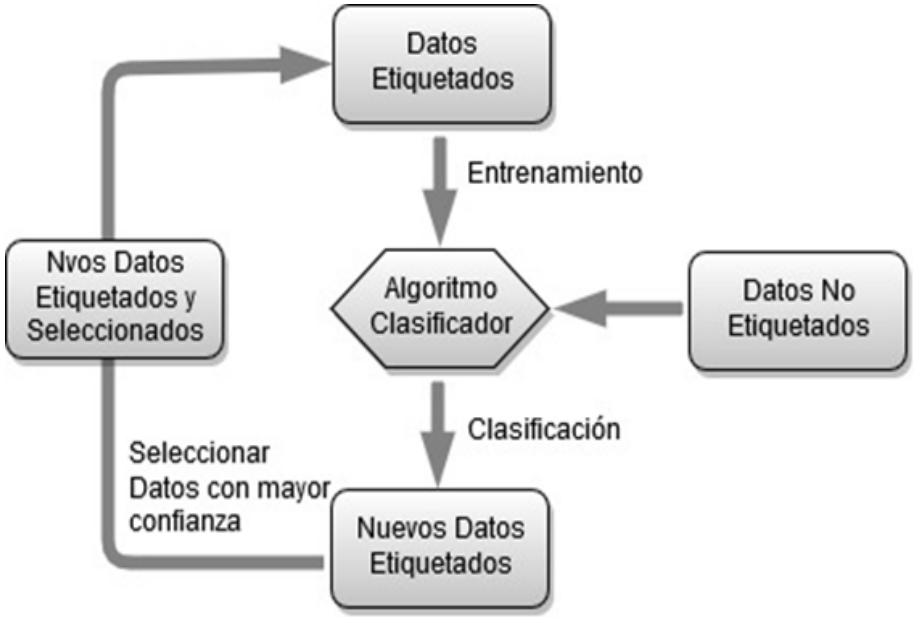

Figura 1. Esquema de método semi supervisado

CORPUS

Si bien existen colecciones de documentos categorizados con su respectiva polaridad, no existen corpus de opiniones textuales en español, disponibles para investigar y etiquetados según la emoción que expresan.

Aunque el español se hable en distintos países, su uso coloquial es diferente en muchos de ellos. Por ejemplo, “estar sin un mango” o “estar en el horno” son modismos argentinos para indicar que se tiene poco dinero, en el primer caso, y que se tiene un problema, en el segundo. Estas expresiones pueden ser frecuentes en opiniones originadas en nuestro país, pero difícilmente en opiniones generadas en otros países de habla hispana. Por esta razón, para este trabajo, se optó por crear un corpus de textos cortos obtenidos de la red social Twitter (llamados tweets), en lenguaje español y generados en Argentina (según identificador de geolocalización). Es decir, los textos que serán objeto de estudio son subjetivos

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

o coloquiales. Entre Octubre de 2018 y Febrero de 2019, utilizando la API[5] de Twitter, se recopilaron tweets públicos relacionados a los temas más populares o *trending tropic*[6] de esos meses. Para formar el corpus de estudio se descartaron los tweets que contenían solo imágenes, solo *emojis*[7] , aquellos con menos de 5 palabras o re-tweets[8] de otros.

Para el etiquetado se propusieron las 6 categorías o sentimientos considerados básicos, según la clasificación de Paul Ekman (Ekman, 1972): asco, felicidad, ira, tristeza, miedo y sorpresa. Además, se agregó otra categoría para indicar que el tweet no expresaba sentimiento (podría tratarse de una publicidad o noticia, es decir, el texto no era subjetivo) o el sentimiento era confuso. Cuatro personas realizaron el etiquetado de los tweets individualmente y se consideró como clasificación final la categoría en la que coincidieron por lo menos tres de ellas, es decir, se asignó la categoría emocional elegida por la mayoría. Es importante destacar, que este etiquetado manual, en muchos casos, resultó difícil por la falta de orden en la escritura, sentimientos confusos o irregularidades lingüísticas, lo que muchas veces generó debate en el equipo.

Del corpus etiquetado se observó una marcada diferencia en las cantidades de tweets de las clases miedo, sorpresa y felicidad, por lo cual fue necesario aplicar algunas acciones para asegurar un conjunto de datos balanceado. Los algoritmos de aprendizaje supervisado suelen clasificar mejor los conjuntos balanceados; de otra manera, podrían catalogar erróneamente instancias de la clase minoritaria. Se observó que los tweets etiquetados con las emociones miedo y sorpresa representaban menos del 10% entre ambos, motivo por el cual, estos textos pasaron a formar parte de la categoría "SS/Otro". En esta categoría también se agruparon aquellos que no expresaban sentimientos. La clase mayoritaria fue felicidad. En este caso, se realizó un submuestreo aleatorio, a riesgo de perder información relevante, pero asegurando el equilibrio entre las categorías consideradas (Longadge and Dongre, 2013). Por lo tanto, los tweets utilizados como base del aprendizaje fueron 498, categorizados según se observa en la Tabla 1. Por otro lado, se seleccionaron otros 10000 tweets recopilados de la misma manera, en el mismo período y preparados de igual forma que el conjunto de entrenamiento. Con éstos se formó el conjunto de datos no etiquetados, utilizados en la etapa semi supervisada.

Sentimiento	Cantidad	Porcentaje
Asco	106	21,3%
Felicidad	108	21,7%
Ira	82	16,5%
Tristeza	101	20,3%
SS/Otro	101	20,3%
Total	498	

Tabla 1. Datos clasificados

ARQUITECTURA

El modelo de clasificación de textos se realizó en dos etapas: la preparación de los textos para obtener una representación adecuada de los mismos y la construcción del clasificador.

Se utilizó el lenguaje Python y las librerías: sklearn de aprendizaje automático, pandas para el manejo y análisis de datos, numpy para el manejo de vectores y matrices, y emojis para la codificación de emojis. Para la tokenización y lematización se utilizó Freeling,

5 https://developer.twitter.com/
6 Palabras o frases más repetidas en un momento concreto en los mensajes de una red social.
7 Pequeñas imágenes que representan emociones, comidas, lugares, etc.
8 Tweet replicado por otro usuario.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

una librería de código abierto para el análisis de textos y variedad de idiomas (Padró and Stanilovsky, 2012). Las ventajas de usar Freeling, frente a otras herramientas informáticas para el procesamiento de lenguaje natural, fueron analizadas por el equipo, como asimismo mencionadas en multiples trabajos, entre ellos (Aguado de Cea et al., 2012) y (Sidorov et al., 2016) .

Preparación de los textos

Para la clasificación automática de textos, utilizando técnicas de aprendizaje computacional, es necesario realizar una tarea previa. La misma consiste en preprocesarlos, descartando aquellas partes que resulten inútiles para el análisis, y transformarlos a una representación adecuada para ser utilizados por los algoritmos.

Muchos comentarios en las redes sociales no suelen tener en cuenta las reglas ortográficas, utilizan palabras del lunfardo o abreviaturas y algunas veces acentúan sus emociones con el uso de emojis, repetición de letras, palabras en mayúsculas o signos de exclamación. Por esta razón, se realizó una serie de acciones a modo de limpieza de los textos.

En primer lugar, los mensajes se convirtieron a minúsculas. Luego se realizaron las tareas típicas de preprocesamiento de textos y otras consideradas necesarias para este dominio:

-*Eliminación de stopwords*: se eliminan las palabras que carecen de un significado por si solas (artículos, pronombres, conjunciones, etc.) ya que por su alta frecuencia de aparición generan ruido en el análisis.

-*Tokenización:* consiste en la segmentación de palabras.

-*Lematización:* reduce una palabra a su lema, es decir, a la palabra que represente a todas las formas flexionadas (conjugadas, en plural, etc.) de la misma palabra.

-*Eliminación de nombres de usuarios*: los nombres de usuarios comienzan con @ y para este análisis no se consideraron relevantes. Se utilizó una expresión regular para identificarlos.

-*Eliminación de URL:* se consideró que las referencias a URL no infieren en la clasificación, por lo tanto, se eliminaron. Por lo general, son referencias a imágenes o noticias y no son utilizadas ni visualizadas para analizar el tweet.

-*Eliminación del símbolo* "#": algunos usuarios utilizan los hashtags[9] como parte del mensaje. Por ejemplo: *"Hola 2019! Por un año lleno de #alegría #felicidad #salud #amor #dinero #amigos #fútbol"*. En este ejemplo, la eliminación de todos los hashtags implicaba perder gran parte del contenido del texto. Por lo tanto, se decidió eliminar solo el símbolo "#", quedando, para este caso: *"Hola 2019! Por un año lleno de alegría felicidad salud amor dinero amigos fútbol"*.

-*Reemplazo de emojis por su correspondiente traducción textual* (en inglés): los emojis frecuentemente son utilizados como parte del mensaje, es decir, reemplazan palabras del mensaje, por ejemplo, *"Todo listo ✈, por fin vacaciones! necesito ☀🍺y mucha ♪♪"*. Eliminarlos implicaría pérdida de información relevante, por esta razón, se utilizó la librería emoji que detecta símbolos Unicode y devuelve la correspondiente traducción textual.

-*Reemplazo de abreviaciones:* es común que se utilicen abreviaturas en los tweets. Por ejemplo, se reemplazó el *"x"* por la palabra *"por"* y *"xq"* por *"porque"*, entre otras.

-*Repetición de letras:* para intensificar la emoción en un texto, es usual que algunas palabras contengan varias letras consecutivas, por ejemplo, *"vamooooosssss Riveeeer!!"*. En este caso, la palabra *"vamooooosssss"* demuestra el énfasis que quiso poner el usuario en su tweet. Se reemplazaron las repeticiones de letras por solo dos ocurrencias, con la idea de diferenciar la palabra "normal" de aquella que manifestaba énfasis.

9 Palabra o frase precedida por el símbolo numeral (#) que se utiliza en redes sociales para destacar o agrupar mensajes

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

A modo de ejemplo, se presenta un tweet y el resultado luego del preprocesamiento (Figura 2).

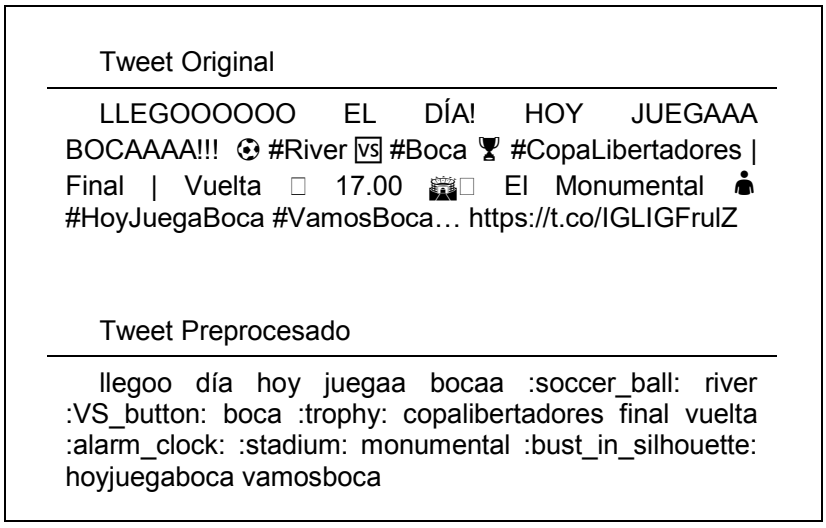

Figura 2. Ejemplo preprocesamiento de tweet

Luego de la limpieza de los textos se descartaron aquellos que contuvieran menos de cinco palabras. El paso siguiente fue obtener una estructura adecuada que los represente. La representación de textos, frecuentemente utilizada, es el modelo vectorial, que consiste en caracterizar cada documento por las palabras que aparecen en él y sus ocurrencias. En este trabajo se usó otro enfoque: consiste en asignar pesos a las palabras de acuerdo a la cantidad de apariciones en el corpus. Este esquema de ponderación se llama *Term frequency – Inverse document frequency* (Tf-Idf) (Manning et al., 2008). Si bien estas técnicas resultan efectivas, dejan de lado algunas combinaciones de palabras que puedan aparecer frecuentemente en el corpus, por ejemplo *"por favor"* o *"buen día"*. Muchas veces, estas expresiones tienen mayor significancia que las mismas palabras por separado. Por esta razón, también se consideró estructurar los textos con *n-gramas*. Los n-gramas posibilitan obtener no solo palabras, sino también secuencia de n palabras consecutivas que permiten mantener y respetar el orden en el que aparecen escritas. Como consecuencia del uso de n-gramas la cantidad de combinaciones posibles se vuelve excesivo para un *n* grande, por lo que se experimentó con n=1, n=2 y n=3. De esta forma cada mensaje (tweet) se convierte en una representación adecuada para los algoritmos de aprendizaje.

Construcción del clasificador

Las principales estrategias para realizar la validación de clasificadores son dos. La primera consiste en separar los datos en un conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento se utiliza para construir un clasificador, que luego asigna una categoría a cada instancia del conjunto de prueba. El desempeño del modelo construido se mide en función de las instancias correctamente clasificadas. La segunda estrategia consiste en usar validación cruzada con k iteraciones o pliegues. En este caso, las instancias son divididas en k partes iguales y se realizan k iteraciones. En cada iteración se toman k-1 partes como conjunto de entrenamiento y la parte restante como conjunto de prueba (Witten et al., 2017).

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

Accuracy, precisión y recall son métricas de evaluación más frecuentes para medir el desempeño de los clasificadores de textos. Accuracy representa la proporción de ejemplos correctamente clasificados. La medida accuracy se considera adecuada cuando las clases del dataset se distribuyen uniformemente. Otras medidas como recall, precisión y F1 son más adecuadas para problemas con conjuntos desbalanceados (Han et al., 2012). Precisión representa la habilidad del clasificador para identificar solo las instancias correctas de cada clase. Recall es la habilidad de un clasificador para encontrar todas las instancias correctas por clase. F1 es la media armónica de precisión y recall, y actúa como una medida que combina ambas.

EXPERIMENTOS

El método propuesto consiste en obtener un clasificador semi supervisado con el enfoque self-training y un modelo inicial fiable, de tal forma que los pasos iterativos del aprendizaje semi supervisado propaguen la menor cantidad de errores posible y los resultados sean más confiables.

Los experimentos se llevaron a cabo en dos etapas. En la primera, con los datos etiquetados, se realizaron una serie de pruebas para obtener los dos mejores modelos clasificadores. En la etapa siguiente, estos modelos se utilizaron para categorizar el conjunto de datos no etiquetado. A continuación se explican las etapas del proceso.

Aprendizaje supervisado

A partir de los textos vectorizados, se realizó una serie de pruebas con algoritmos usualmente utilizados para clasificación: Naive Bayes, Support Vector Machine (SVM), Random Forest, Arbol de decisión, K-Nearest Neighbors, Ridge. Se empleó la técnica de validación cruzada con 10 pliegues y distintas cantidades de atributos, con el objetivo de encontrar los mejores modelos para la etapa siguiente. A los atributos se les otorgó una puntuación según la relevancia en el corpus, calculada con el test estadístico chi-cuadrado, y se seleccionaron los n primeros de este ranking[10] . Las pruebas consistieron en verificar el comportamiento de cada algoritmo con distintas cantidades de n atributos. La Tabla 2 muestra la performance de los algoritmos frente a las distintas cantidades de atributos considerados. Los dos algoritmos con mejor performance fueron: Naive Bayes entrenado con 500 atributos que logró 0,76 de recall y 0,79 de precision, y SVM que logró 0,79 de recall y 0,82 de precision. Para ambos algoritmos, la métrica F1 resultó la de mayor valor entre todas las pruebas.

Atributos	100			200			300		
	rec	prec	F1	rec	prec	F1	rec	prec	F1
DecTree	0,65	0,75	0,7	0,64	0,74	0,69	0,65	0,73	0,69
Ridge	0,71	0,79	0,75	0,72	0,77	0,74	0,74	0,8	0,77
KNN	0,62	0,66	0,64	0,58	0,64	0,61	0,57	0,63	0,6
Random	0,69	0,72	0,71	0,71	0,77	0,74	0,71	0,78	0,74
SVM	0,73	0,79	0,76	0,74	0,78	0,76	0,76	0,8	0,78
N.Bayes	0,66	0,74	0,70	0,73	0,79	0,76	0,76	0,79	0,77

10 https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

Atributos	400			500			600		
	rec	prec	F1	rec	prec	F1	rec	prec	F1
DecTree	0,65	0,72	0,68	0,63	0,71	0,67	0,62	0,67	0,64
Ridge	0,76	0,81	0,78	0,76	0,81	0,79	0,76	0,81	0,78
KNN	0,56	0,63	0,59	0,50	0,55	0,52	0,45	0,49	0,47
Random	0,69	0,78	0,73	0,70	0,77	0,73	0,69	0,76	0,72
SVM	0,77	0,82	0,79	0,79	0,82	0,81	0,79	0,82	0,81
N.Bayes	0,76	0,78	0,77	0,76	0,79	0,78	0,76	0,78	0,77

Tabla 2. Performance de algoritmos

Aprendizaje Semi Supervisado

En esta etapa se aplicó el método self-training pero con la variante que participaron dos clasificadores, en lugar de uno, seleccionados en el paso previo. Cada clasificador asigna las categorías a los datos no etiquetados otorgando, en cada caso, un nivel de confianza a esta asignación. En la etapa iterativa, aquellos ejemplos, ahora etiquetados, para los cuales ambos algoritmos asignen la misma categoría y que superen el 80% de confianza en la categorización, se agregan al conjunto de entrenamiento y se eliminan del conjunto de datos no etiquetados. Este proceso (Figura 3) continúa hasta que no queden instancias sin etiquetar o el nivel de confianza en la predicción de cada instancia no supere el estipulado.

Input:
E: conjunto de datos etiquetados
NE: conjunto de datos no etiquetados

Sean cl_1 y cl_2 los mejores clasificadores obtenidos de la etapa previa
Repetir hasta que E no tenga cambios
Entrenar cl_1 con E
Entrenar cl_2 con E
Para cada *x* de NE hacer:
Sea x_1.clase la categoría asignada a *x* usando cl_1
Sea x_2.clase la categoría asignada a *x* usando cl_2
Si (x_1.clase=x_2.clase) y (x_1.confianza>80%) y (x_2.confianza>80%)
Agregar *x* al conjunto E
Eliminar *x* del conjunto NE

Figura 3. Algoritmo semi supervisado

En cada paso iterativo, los dos modelos se entrenaron de la misma manera que en la etapa anterior, es decir, con validación cruzada y 10 pliegues. La Tabla 3 muestra cómo se fueron modificando los conjuntos de datos etiquetados y no etiquetados en cada paso. En la primera iteración se entrenó ambos algoritmos con los 498 ejemplos etiquetados, luego ambos algoritmos coincidieron en la clasificación de 1267 ejemplos, con una confianza mayor a 80%. Estos nuevos ejemplos etiquetados se suman a los etiquetados manualmente, con lo cual en la iteración 2 los algoritmos se entrenaron con 1765 ejemplos.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

Iter	Etiq.	No Etiq	Nvas	N.Bayes	SVM
1	498	10000	1267	0.76	0.79
2	1765	8733	2577	0.90	0.89
3	4342	6156	2711	0.92	0.92
4	7053	3445	1078	0.91	0.92
5	8131	2367	385	0.91	0.92
6	8516	1982	161	0.92	0.91
7	8677	1821	89	0.91	0.91
8	8766	1732	51	0.91	0.92
9	8817	1681	15	0.91	0.92
10	8832	1666	6	0.91	0.91
11	8838	1660	6	0.91	0.92
12	8844	1654	26	0.91	0.91
13	8870	1628	22	0.91	0.92
14	8892	1606	4	0.92	0.92
15	8896	1602	2	0.91	0.91
16	8898	1600	3	0.91	0.92
17	8901	1597			

Etiq: instancias etiquetadas. No Etiq.: instancias no etiquetadas. Nvas: instancias etiquetadas por ambos clasificadores. N.Bayes recall del algoritmo Naïve Bayes. SVM: recall del algoritmo Support Vector Machine

Tabla 3. Resultados método semi supervisado

Como se observa en la tabla, en las primeras iteraciones se clasificaron la mayor parte de los nuevos ejemplos. El valor recall evaluado en los dos algoritmos fue incrementándose a medida que aumentaba el conjunto de entrenamiento, finalizando ambos con más del 90% de efectividad. Como resultado final, en 16 iteraciones, se clasificaron 8403 ejemplos inicialmente no etiquetados. La distribución de las categorías asignadas por los modelos se observa en la Tabla 4. Como se mencionó anteriormente, las aplicaciones disponibles para la clasificación en emociones no contemplan las mismas categorías de este trabajo, por lo tanto, no fue posible usarlas para comparar la asignación realizada por el método. Por esta razón, se constataron los ejemplos manualmente, siguiendo el mismo criterio del etiquetado inicial. Se verificaron que en alrededor del 5% la etiqueta era errónea, por lo cual el método se consideró aceptable.

Sentimiento	Cantidad	Porcentaje
Asco	134	0.2%
Felicidad	6286	79.5%
Ira	670	1.8%
Tristeza	516	1,0%
SS/Otro	797	17.5%
Total	8403	

Tabla 4. Nuevos ejemplos clasificados

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

CONCLUSIONES

La identificación del sentimiento que un usuario de una red social expresa en un mensaje textual ayuda a revelar información importante. Los textos se obtuvieron de la red social Twitter de diversas temáticas según las tendencias ocurridas en Argentina en un período de tiempo determinado. En general, en las redes sociales, los usuarios suelen escribir sin respetar reglas ortográficas y lo hacen con la jerga de su país de origen. Esto implica que la preparación del corpus sea de suma importancia. Además, utilizan emojis o hashtags como parte de las oraciones en reemplazo de palabras o pareceres. Por lo tanto, eliminarlos supone pérdida de información importante, pero mantenerlos implicaría necesariamente la realización de acciones para transformarlos en datos significativos.

Una parte de los tweets recopilados, 498 ejemplos, fueron clasificados manualmente en cuatro sentimientos: ira, felicidad, asco y tristeza. Se agregó una quinta categoría para aquellos textos que no expresaban sentimientos o indicaban uno diferente a los considerados. Otra porción de tweets, 10000 ejemplos, participaron en la etapa semi supervisada del método propuesto. Este método consistió en combinar algoritmos supervisados con semi supervisados. En una primera instancia, se realizó validación cruzada para elegir los dos mejores modelos de clasificación supervisada, sus respectivos parámetros y la cantidad de n-gramas que mejor represente al corpus. Luego, en una segunda etapa, se utilizó un proceso semi supervisado para etiquetar automáticamente los textos con la variante que participan dos clasificadores, en lugar de uno. Las métricas de ambos modelos, en sus pasos iterativos, se mantuvieron por encima del 90% logrando clasificar más del 80% de los ejemplos. De ellos, se constató que alrededor del 5% fueron clasificados erróneamente según el criterio del equipo. Por lo tanto, los resultados obtenidos demuestran el poder de clasificación de este método. Ante la falta de un corpus de textos coloquiales que se encuentren etiquetados en sentimientos, el mayor logro alcanzado fue el de crear un corpus de 8901 tweets clasificados por emociones en lenguaje español.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

REFERENCIAS

Abudalfa, S. I., & Ahmed, M. A. (2019). Semi-Supervised Target-Dependent Sentiment Classification for Micro-Blogs. Journal of Computer Science & Technology, 19(01). doi:https://doi.org/10.24215/16666038.19.e06

Aguado de Cea, G., Barrios, M., Bernardos, M., Campanella, I., Montiel-Ponsoda, E., Muñoz-García, O., & Rodriguez, V. (2012). Análisis de sentimientos de un corpus de redes sociales. 31er Congreso Asociación Española de Lingüística Aplicada. "Comunicación, Cognición y Cibernética". San Cristóbal de la Laguna, Tenerife.

Amores Fernández, M. A. (2016). Detección de la polaridad de las opiniones basada en nuevos recursos léxicos. Universidad Central “Marta Abreu” de Las Villas.

Cámara, E., Valdivia, M., Ortega, J., & Ureña Lopez, A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español. Procesamiento del Lenguaje Natural(47), 163-170.

Diaz Rangel, I., Sidorov, G., & Suarez Guerra, S. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomazein. Revista semestral de lingüística, filología y traducción, 31-46.

Dubiau, L. (Octubre de 2013). Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos. Buenos Aires: Facultad de Ingenieria, Universidad de Buenos Aires.

Ekman, P. (1972). Emotion in the Human Face. Pergamon.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. USA: Morgan Kaufmann Publishers.

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. International Journal of Computer Science and Network (IJCSN). Obtenido de arXiv preprint arXiv:1305.1707

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction toinformation retrieval. Cambridge University Press.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1: Long Papers, págs. 174-184.

Montesinos García, L. (2014). Análisis de sentimientos y predicción de eventos en Twitter. Santiago de Chile: Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile.

Navas Loro, M., Rodriguez-Doncel, V., Santana-Perez, I., Fernández-Izquierdo, A., & Sanchez, A. (2018). MAS: A corpus of tweets for marketing in spanish. European Semantic Web Conference, (págs. 363-375). Springer, Cham.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul.

Selva Castelló, J. (2015). Desarrollo de un sistema de análisis de sentimiento sobre Twitter. Universitat Politècnica de València Escuela.

Sidorov, G., Haro, S. N., & Vázquez, V. A. (2016). Construcción de un corpus marcado con

Revista Tecnología y Ciencia
DOI:https://doi.org/10.33414/rtyc.37.11-23.2020 - ISSN 1666-6933

emociones para el análisis de sentimientos en Twitter en español. Revista Escritos BUAP, 1(1).

Witten, I., Hall, M., Frank, E., & Pal, C. (2017). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.