Recibido 29 de octubre 2021. Aceptado 02 de diciembre 2021. Publicado 23 de diciembre 2021.

ISSN: 2448-7775

Identificación de polaridad en Twitter usando validación cruzada

JOSÉ CARMEN MORALES CASTRO*, LUIS MANUEL LEDESMA CARRILLO, RAFAEL GUZMÁN CABRERA.
Universidad de Guanajuato, Guanajuato, México.
*Autor de correspondencia: jc.moralescastro@ugto.mx

RESUMEN En la actualidad, se observa un gran auge y crecimiento de las diferentes plataformas de redes sociales, que han provocado un gran protagonismo entre los usuarios en la generación de información día a día, una plataforma que ha evolucionado con el pasar del tiempo es Twitter, la cual nos presenta un gran reto en cuanto al procesamiento del lenguaje natural al tratar de determinar la polaridad de un tweet de manera positiva, negativa y neutra. Este trabajo presenta un estudio de emociones para analizar la polaridad de un conjunto de datos que fueron extraídos de Twitter, detallando cada uno de los recursos sobre las distintas formas que tiene un lenguaje, y poder observar cómo sentimientos como la ironía, sarcasmo, felicidad entre otros, nos pueden llegar a ayudar a clasificar la polaridad de cada uno de ellos profundamente en el corpus que se ocupa para este trabajo de investigación. Se presentan resultados de experimentos realizados utilizando distintos métodos de aprendizaje: Support Vector Machine, Naive Bayes, y Regresión Logística, con los cuales se implementó un sistema de clasificación basado en validación cruzada. Todos los experimentos fueron realizados en Python.

PALABRAS CLAVE — Polaridad sentimental, Análisis de sentimientos, Validación cruzada, Python.

I. INTRODUCCIÓN

Actualmente los sitios web de microblogging se han convertido en espacios digitales de información variada, donde los usuarios publican y diseminan información en tiempo real relacionados con una gran variedad de temas donde se pueden expresar opiniones por medio de textos que llevan implícitamente una carga emotiva. Esto quiere decir que, las opiniones llevan una carga emotiva que se convierte en una opinión positiva o negativa sobre personas, productos, o servicios que se llevan a cabo en la vida diaria.

Varias empresas organizaciones e instituciones han hecho uso de este tipo de medios para obtener retroalimentación, promocionarse, o simplemente para convertir la opinión de los usuarios en una red de mejora que ha comenzado a sondear micro blogs para tener una idea acerca del sentimiento general a sus productos y servicios [1], en este contexto se encuentra Twitter que en los años recientes ha tenido un crecimiento importante en los llamados "panoramas sociales", usado en un sistema de transmisión, así como una herramienta de conversación [2], es por eso que esta red social actualmente es muy utilizada para el desarrollo de numerosas investigaciones entre ellas el análisis de sentimientos o minería de opiniones como también se le conoce, en donde el análisis de sentimientos se llega a definir como el proceso de determinar opiniones basados en actitudes, valoraciones y emociones acerca de temas en específico [3].

Algunos trabajos como en [4] describen a la minería de opiniones como un tratamiento automático de opiniones contenidos en una frase, esto permite determinar la polaridad o sentimiento que se expresa ya sea positivo, negativo o mixto, así como permite la extracción automática de caracteres que ayuda a conocer la percepción que los usuarios tienen sobre temas y aspectos específicos.

Debido que las emociones que los usuarios expresan en los Tweets están relacionadas con los sentimientos de la persona, y la polaridad (positivo, negativo y neutral) que es la medida de las emociones expresadas en una frase. Generalmente la polaridad va de negativo (-1) a positivo (1) pasando por el neutral (0), este último valor significa que no se ha expresado ningún sentimiento u opinión [5].

II. TRABAJO RELACIONADO

Autores como [6] describen el análisis de sentimientos como una tarea que se encarga de identificar, y clasificar diferentes puntos de vista y opiniones sobre una cuestión en particular, pudiendo ser un objeto, una persona, o una actividad, entre otros; basado en el procesamiento del lenguaje natural (PNL) para identificar el estado de ánimo de las personas, recopilando comentarios, reacciones y mensajes a través de las redes sociales, donde el objetivo principal es el análisis de documentos en línea y su clasificación como sentimientos: positivo o negativo, también existe la posibilidad de que no existan y se clasificarían como neutros.

En redes sociales la investigación ha ido creciendo en el análisis de sentimientos donde su clasificación depende mucho del uso de palabras clave en los textos, un pequeño factor que puede ocasionar un problema sería la información almacenada en los gráficos, videos o imágenes ya que pueden incluir información que no se encuentra en el texto que acompaña.

En [7] los autores nos presenta algunas técnicas utilizadas para la revisión del análisis sentimental, que ayudan a determinar automáticamente la polaridad en un texto, siendo los más comunes los basados en aprendizaje automático el cual forma parte importante de la Inteligencia Artificial, ya que desarrolla programas mediante algoritmos de aprendizaje y generación de conocimiento capaces de aprender a resolver problemas, dentro de las posibles aplicaciones que pueden llegar a ser tan útiles como diferentes, por lo que en la actualidad sigue siendo un tema de investigación abierto en el cual se siguen haciendo aportes muy atractivos e interesantes en el área de análisis de sentimientos, cabe mencionar que el análisis de sentimientos no solo se enfoca en la parte de identificar polaridad en opiniones expresadas por medio de textos subjetivos, ya que esta tarea puede ir mucho más allá permitiendo, incluso llevar acabo la identificación de sentimientos en particular como puede ser la clasificación de sentimientos primarios como alegría, tristeza, enojo, miedo, entre otros.

Otra técnica utilizada para la revisión de análisis de sentimientos es la orientación semántica que se encarga de extraer opiniones, en [8] nos explican que la orientación semántica de una palabra puede llegar a ser positiva cuando se muestre a través de un elogio o bien una negativa, cuando se presenta una crítica. Utiliza una técnica de aprendizaje que no forzosamente debe ser supervisada ya que no requiere una capacitación inicial, es decir no requiere de instancias manualmente etiquetadas para llevar a cabo el proceso de aprendizaje. Autores como [9] nos hablan de la forma de adaptación de este sistema de orientación semántica, como ejemplo, para poder realizar el análisis de sentimientos en un nuevo idioma, construyendo clasificadores de máquinas de vectores de soporte (SVM, por sus siglas en ingles), teniendo en cuenta el enfoque que usaron mediante el aprendizaje automático de este clasificador de texto, esto basado en que los clasificadores se pueden entrenar en cualquier idioma, para esto ellos realizaron pruebas con validación cruzada usando un clasificador basado en el método de aprendizaje SVM que fue construido con algoritmos secuenciales de mínima optimización que está incluido en el paquete de software WEKA[1] (Plataforma de software libre bajo la licencia GNU-GPL para el aprendizaje automático y la minería de datos escrito en JAVA); teniendo en cuenta que este tipo de aprendizaje no supervisado utiliza diferentes reglas léxicas en la clasificación de sentimientos.

III. PLANTEAMIENTO DEL PROBLEMA

Llevar a cabo la identificación automática de sentimientos en sistemas de información, con las mejores prestaciones, en tweets utilizando una arquitectura que combine clasificadores base y recursos léxicos.

Tratar de definir herramientas automáticas capaces de extraer información subjetiva de textos en lenguaje natural, como opiniones o sentimientos, con el fin de crear conocimiento estructurado y procesable para ser utilizado por un sistema de toma de decisiones.

IV. METODOLOGÍA

En el presente trabajo se realizó la clasificación de tweets, como conjunto de evaluación se utilizó un conjunto de datos correspondientes a opiniones emitidas en Twitter que tiene alrededor de 163mil tweets, los cuales se encuentran etiquetados en cuanto a la polaridad de la opinión como: positivo, negativo y neutro. Estos tweets y comentarios fueron hechos sobre Narendra Modi y otros líderes, así como la opinión en la sociedad hacia el próximo primer ministro de la nación (en el contexto de las elecciones generales celebradas en India en el año 2019), y que ayudara a clasificarlos de manera automática. Para este trabajo tomamos esta base de datos como objeto de estudio ya que resulta muy interesante ver como la percepción mediática de un personaje puede ser medida a través de opiniones emitidas en redes sociales y que sin duda esto puede ayudar al personaje en cuestión a corregir o moderar su discurso en relación con alguna temática en particular.

El seleccionar la base de datos fue nuestro primer paso para construir el clasificador, los textos se encuentran etiquetados con valores de -1 a 1, donde:

- 0 indica que es un Tweet/ comentario neutro
- 1 indica un sentimiento positivo
- -1 indicando un tuit/comentario negativo.

Cabe mencionar que esta es una base de datos estándar que se encuentra disponible en la red.[2]

En la Fig. 1, se muestra el diagrama que ilustra la metodología implementada en el presente trabajo. En nuestro caso utilizamos el escenario de clasificación basado en validación cruzada, el cual es uno de los métodos de remuestreo más utilizados para evaluar la capacidad de generalización de modelos predictivos y así estimar el verdadero error de predicción y ajuste de parámetros [10].

A continuación, se describe brevemente cada uno de los elementos que componen la metodología propuesta. Para la entrada de datos lo primero que se realizo fue cambiar el valor en la base de datos poniendo la polaridad como positiva, negativa y neutra respectivamente, de acuerdo con la etiqueta numérica existente.

[1] https://www.cs.waikato.ac.nz/ml/weka/

[2] https://www.kaggle.com/cosmos98/twitter-and-reddit-sentimental-analysis-dataset

2448-7775 © 2021 Identidad Energética

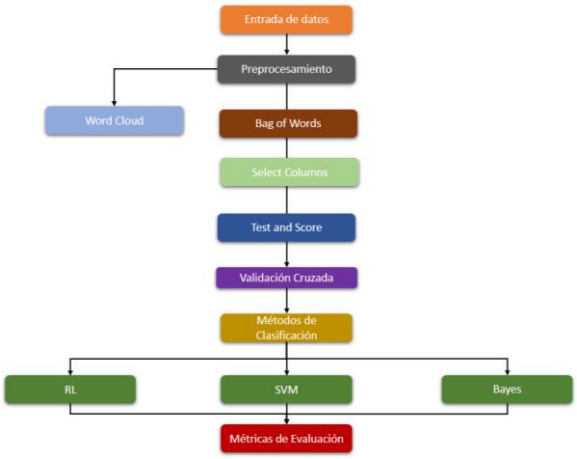


Fig. 1. Metodología implementada en los experimentos realizados en Python para el presente trabajo.

Los experimentos fueron realizados en Python, añadimos el corpus, en donde se usó como características de aprendizaje, el texto de los tweets, es decir los comentarios u opiniones que los usuarios realizaron.

En la parte del preprocesamiento se procedió a eliminar las palabras de paro, también llamadas stop words, las cuales son las palabras vacías de contenido pero que nos sirven para estructuras las oraciones y poder expresarnos correctamente. Sin embargo y dado que para el sistema de clasificación se convierte en un problema matricial, el tener menos elementos reduce la dimensionalidad de la matriz.

Una vez realizado este proceso; para corroborar, se crea nuevamente una Wordcloud, como se muestra en la Fig. 2, para asegurar que las menciones se encontraran como palabras principales, una vez eliminadas las stopwords.

Ya con el ajuste al documento se pasa a una bolsa de palabras, también conocida como bag of words en donde se selecciona como frecuencia del documento el IDF (Inverse document frequency por sus siglas en ingles).

Como siguiente paso de la metodología propuesta, se coloca la variable de etiquetas que se va a analizar, en este caso es la fila en donde se encuentran las etiquetas de positivo, negativo y neutro.

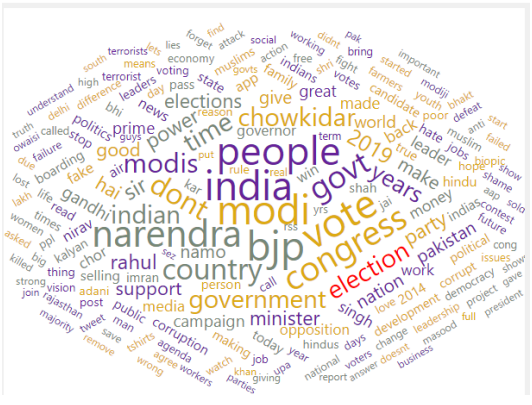


Fig. 2. Nube de palabras con lista de palabras de paro eliminadas.

Se realiza un análisis para observar que clasificador arroja mejores resultados de precisión, para ello se utilizaron los siguientes métodos de aprendizaje ampliamente conocidos en el estado del arte:

Support Vector Machine (SVM) El cual es un método que se basa en el aprendizaje y nos brinda apoyo en la resolución de problemas mediante clasificación y regresión, el cual se basa en fases de entrenamiento y resolución, este método propone una respuesta (salida) a un problema establecido [11].

Regresión Logística (RL), el cual definen en [12] como un algoritmo de aprendizaje automático de clasificación utilizado para predecir la probabilidad y datos mediante rectas, que requieren que la variable dependiente sea binaria.

Naive Bayes (NB). El cual es un clasificador que nos ayuda a calcular la probabilidad de un suceso teniendo información de este basado en el teorema y en hipótesis adicionales [13].

Como métricas de evaluación se utilizaron:

Área bajo la curva (AUC por sus siglas en inglés) se calcula usando el área bajo la curva ROC y cuanto mayor es el área más precisa es formalmente el predictor, la fórmula para calcular el AUC es representada por la Ec. (1).

$$AUC = \int_0^1 f(x)dx \tag{1}$$

Donde f(x) representa la función de la curva característica de funcionamiento del receptor (ROC por sus siglas en ingles), sin embargo desde f(x) tiende a no tener una forma de integración como una parábola; varios autores sugieren utilizar métodos de aproximación para calcular las AUC[14].

Accuracy es el grado de cercanía al valor verdadero, se refiere a una medición con resultados tanto verdaderos como consistentes. La fórmula para calcular el Accuracy es representada por Ec. (2).

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \tag{2}$$

Donde tp representa un valor verdadero-positivo, tn un valor verdadero-negativo, fp un valor falso-positivo, y fn un valor falso-negativo.

$F1$ es una medida de precisión en una prueba que se calcula a partir de la precisión y el recall de la prueba que se está llevando a cabo, en pocas palabras F1 es la media armónica de la precisión y el recall, que se muestra en la Ec. (3) a continuación.

$$F1 = \frac{tp}{tp+\frac{1}{2}(fp+fn)} \tag{3}$$

Donde tp es un valor verdadero-positivo, fp un valor falso-positivo y fn un valor falso-negativo.

Precision es una métrica de rendimiento que se aplica en datos recuperados de una colección, corpus o espacio muestral; también se le conoce como valor predictivo positivo el cual es una fracción de instancias relevantes entre las instancias recuperadas tal como se muestra en Ec. (4).

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

Donde *tp* equivale a un valor verdadero-positivo y *fp* a un valor falso-positivo.

Recall también conocido como sensibilidad es una fracción de instancias relevantes que se lograron recuperar, la Ec. (5) la representa de la siguiente manera:

$$Recall = \frac{tp}{tp+fn} \quad (5)$$

Donde *tp* representa un valor verdadero-positivo y *fn* representa un valor falso-negativo[15].

V. RESULTADOS

Llevando una secuencia y trabajo como se muestra en la Fig. 3, utilizando Python, se muestra a continuación los resultados obtenidos en los experimentos realizados.

Este experimento se llevó en dos etapas, donde se utilizaron dos listas de palabras de paro también conocidas como stopwords, las cuales fueron eliminadas de los documentos bajo estudio, la primera consta de 150 palabras de paro y se puede encontrar en línea[3], la cual nos arrojó como resultado lo que se muestra en la Tabla I, y la segunda que contiene un total de 574 palabras de paro y se encuentra disponible en la red[4], se comparó con el primer análisis obteniendo un resultado que se muestra en la Tabla II.

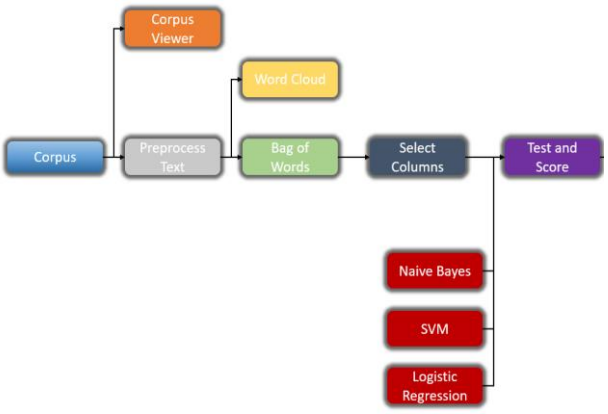


Fig. 3. Diagrama de secuencia para la realización de los experimentos en Python.

TABLA I. MÉTRICAS DE EVALUACIÓN LISTA DE STOPWORDS 1.

Model	AUC	CA	F1	Precision	Recall
SVM	0.671	0.492	0.438	0.510	0.492
Naive Bayes	0.776	0.513	0.464	0.625	0.513
Logistic Regression	0.795	0.516	0.421	***0.724***	0.516

TABLA II. MÉTRICAS DE EVALUACIÓN CON LISTA DE STOPWORDS 2

Model	AUC	CA	F1	Precision	Recall
SVM	0.618	0.452	0.390	0.464	0.452
Naive Bayes	0.746	0.500	0.450	0.595	0.500
Logistic Regression	0.743	0.496	0.400	***0.682***	0.496

Como resultado podemos observar que la primera lista de palabras de paro nos arroja un valor de precisión del 72.4% y la segunda un valor de 68.2%, teniendo una diferencia de poco más del 4% para esta métrica de evaluación.

VI. CONCLUSIONES

El objetivo fundamental del análisis de sentimientos es definir herramientas automáticas capaces de extraer información subjetiva de textos en lenguaje natural, como opiniones o sentimientos, con el fin de crear conocimiento estructurado y procesable para ser utilizado por un sistema de toma de decisiones.

Como conclusión se puede decir que el llevar a cabo la identificación de sentimientos en textos no estructurados, como los que se encuentran en Twitter, es una tarea no trivial que cada vez se utiliza más tanto por empresas como por instituciones de gobierno. Con base en los resultados obtenidos podemos observar que el mejor resultado en la identificación de sentimientos se obtiene al utilizar la lista corta de palabras de paro facilitado de esta manera el procesamiento de grandes volúmenes de información y permitiendo además el poder identificar las áreas de oportunidad de mejora en el caso de las opiniones negativas.

REFERENCIAS

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30-38.
[2] J. Comm and K. J. S. P. E. G. BURGE, "O poder do Twitter," 2009.
[3] P. M. Fiorini, L. R. J. C. S. Lipsky, and Interfaces, "Search marketing traffic and performance models," vol. 34, no. 6, pp. 517-526, 2012.
[4] J. Fernández, E. Boldrini, J. M. Gómez, and P. J. P. d. l. n. Martínez-Barco, "Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog," vol. 47, pp. 179-187, 2011.
[5] A. Reyes, P. Rosso, T. J. L. r. Veale, and evaluation, "A multidimensional approach for detecting irony in twitter," vol. 47, no. 1, pp. 239-268, 2013.
[6] B. Saberi and S. J. I. J. o. A. S. E. I. T. Saad, "Sentiment analysis or opinion mining: a review," vol. 7, pp. 1660-1667, 2017.
[7] R. Hierons, "Machine learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, UK, International Student Edition, 1997.

[3] https://www.ranks.nl/stopwords

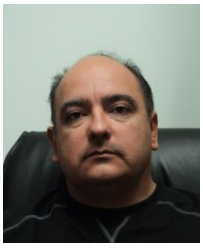
[4] https://github.com/manishkanadje/reuters21578/blob/master/stopwords.txt

ISBN: 0-07-115467-1, 414 pages. Price: UK£ 22.99, soft cover," ed: Wiley Online Library, 1999.
[8] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," in *Proceedings of the 38th annual Hawaii international conference on system sciences*, 2005, pp. 112c-112c: IEEE.
[9] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-linguistic sentiment analysis: From English to Spanish," in *Proceedings of the international conference RANLP-2009*, 2009, pp. 50-54
[10] P. Refaeilzadeh, L. Tang, and H. J. E. o. d. s. Liu, "Cross-validation," vol. 5, pp. 532-538, 2009.
[11] L. J. M. A. E. Rouhiainen, "Inteligencia artificial," 2018.
[12] R. E. Wright, "Logistic regression," 1995.
[13] W. Morales Castro and R. J. C. y. S. Guzman Cabrera, "Tuberculosis: Diagnosis by Image Processing," vol. 24, no. 2, 2020.
[14] A. J. Bowers and X. J. J. o. E. f. S. P. a. R. Zhou, "Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes," vol. 24, no. 1, pp. 20-46, 2019.
[15] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*, 1999, pp. 249-252: Herndon, VA.

BIOGRAFÍAS

JOSÉ CARMEN MORALES CASTRO Estudiante de Posgrado en el área de Maestría en Administración de Tecnologías de la Información en el Departamento de Estudios Multidisciplinarios de la Universidad de Guanajuato; Ingeniero en Mecatrónica por el Instituto Tecnológico Superior de Irapuato.

LUIS MANUEL LEDESMA CARRILLO Profesor investigador del Departamento de Estudios Multidisciplinarios (Sede Yuriria) de la Universidad de Guanajuato. Doctor en Ingeniería Eléctrica en el área de instrumentación y visión robótica por la Universidad de Guanajuato, SNI-1,



RAFAEL GUZMÁN CABRERA Profesor Titular del departamento de Ingeniería Eléctrica de la División de Ingenierías del campus Irapuato-Salamanca de la Universidad de Guanajuato desde hace 21 años, Dr. en Reconocimiento de Formas e Inteligencia Artificial por la Universidad Politécnica de Valencia, España. Miembro de la Academia Mexicana de Ciencias, SNI-1. Miembro del cuerpo académico de física aplicada y tecnologías avanzadas.