Análisis de sentimiento y minería de opiniones en Twitter

José Medrano
joseignacio18@gmail.com

Agustina Monge
agum_96@hotmail.com

Ingeniería en Informática, Facultad de Ingeniería, UCASAL

Resumen

En los últimos años las redes sociales se han vuelto un lugar en el cual volcar nuestras más variadas opiniones. Actualmente se encuentra a Twitter como uno de los espacios más utilizados a la hora de expresar nuestros sentimientos sobre distintas temáticas debido a que dicha herramienta es una plataforma de uso gratuito y de fácil acceso. A partir de esta situación, se pueden utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para identificar comportamientos y opiniones colectivas e inferir su polaridad, definiendo qué palabras se consideran de carácter positivo y cuáles de carácter negativo. Este tipo de análisis tiene un amplio campo de aplicación ya que el mismo puede ser aplicado tanto por empresas, quienes buscan saber la opinión de los clientes respecto a un producto determinado, por organizaciones políticas que buscan determinar cuál es la postura de los ciudadanos con respecto a un candidato y a su propuesta, como por agencias de turismo para determinar qué lugar es el más popular entre los turistas y a partir de eso ofrecer paquetes, etc. Con este artículo se desea describir, mediante una investigación, qué es análisis de sentimiento, cómo se realiza actualmente su práctica y qué podemos esperar a partir de ella.

Palabras Clave

NLP, Procesamiento del lenguaje natural, Análisis de sentimientos, Big Data, Machine Learning, Aprendizaje supervisado, Diccionario

Abstract

In recent years social networks have become a place to give our most varied opinions. Currently, Twitter is one of the most used spaces to express our feelings on different topics because this tool is a platform for free use and easy to access. From this situation, Natural Language Processing (NLP) techniques can be used to identify collective behaviors and opinions and infer their polarity, defining which words are considered positive and negative. This type of analysis has a broad field of application since it can be applied both by companies, who seek to know the opinion of customers about a particular product, by political organizations that seek to determine what is the position of citizens regarding a candidate and their proposal, as for tourism agencies to determine which place is the most popular among tourists and from there offer packages, etc. With this article we want to describe, through a research, what is sentiment analysis, how it is practiced nowadays and what can we expect from it.

Keywords:

NLP, Natural Language Processing, Sentiment Analysis, Big Data, Machine Learning, Supervised Learning.

Introducción

Es muy difícil saber con certeza lo que piensan las personas en su totalidad, o al menos un gran porcentaje, con una precisión considerable, sobre los temas por los que se ven afectados en su vida cotidiana.

En general, las encuestas realizadas por organizaciones dedicadas a estas tareas se ven sesgadas por la pequeña cantidad de gente entrevistada, o por no ser esta selección de personas una muestra representativa de toda la sociedad en conjunto.

En la actualidad gran cantidad de adultos y la mayor parte de los jóvenes utilizan al menos una red social ya sea para compartir su actividad cotidiana, informarse o formular opiniones.

Según diario Clarín, en abril del 2018 Twitter contaba con 336 millones de usuarios activos mensuales alrededor del mundo, siendo la red más popular en cuanto respecta a compartir opiniones [1].

Por estas razones es que se elige como foco de este artículo la red social Twitter y las opiniones que los usuarios vierten en ella, que proveen gran cantidad de información, con representación de casi todas las edades y estratos sociales.

Entonces, la información provista por las opiniones y comentarios en esta plataforma se puede utilizar para percibir o rescatar tendencias en las inclinaciones del pensamiento social.

Esta necesidad de conocer las magnitudes reales de los sentimientos y reacciones de las personas ante los sucesos que influyen en ellos no se puede satisfacer sin encontrar la manera de ser capaces de captar, conservar y analizar grandes cantidades de comentarios y opiniones de forma eficaz, extrayendo de ellos la mayor cantidad de información posible.

Se define el término Big Data como: "*Conjunto de técnicas que permiten analizar, procesar y gestionar conjuntos de datos extremadamente grandes que pueden ser analizados informáticamente para revelar patrones, tendencias y asociaciones, especialmente en relación con la conducta humana y las interacciones de los usuarios*" [2].

Si queremos identificar grandes cantidades de personas con pensamientos u opiniones con inclinaciones similares debemos enfocar nuestro esfuerzo en investigar, asimilar y utilizar nuevas herramientas que cuenten con la capacidad de manejo de este volumen masivo de datos.

Entonces, al concepto de Big Data debemos sumar el concepto de Data Mining (DM) o minería de datos. Sas Institute Inc. define a la minería de datos como:

"*Proceso de encontrar anomalías, patrones y correlaciones dentro de grandes conjuntos de datos para predecir resultados. Usando una amplia gama de técnicas, se puede utilizar esta información para aumentar los ingresos, reducir costos, mejorar las relaciones con clientes, reducir riesgos y más*" [3].

Uno de los derivados de la minería de datos tradicional, utilizada sobre bases de datos relacionales con datos estructurados, es la minería de texto, que busca extraer información sobre datos no estructurados, en distintos formatos de documentos.

A su vez, de la minería de texto surge otra rama conocida como análisis de sentimiento o minería de opinión, que es la parte de la minería en la cual se enfoca este artículo.

Entonces, el uso minería de opiniones para determinar la existencia o el cambio de comportamientos sociales se basa en herramientas de procesamiento de lenguaje natural (PLN), capaces de dar una connotación positiva o negativa del lenguaje utilizado en el texto escrito.

Clasificar la polaridad de un comentario u opinión se encuentra como la tarea primordial en análisis de sentimientos: se determina si la opinión es positiva, negativa o neutra según las palabras utilizadas para expresarse y su asociación con otros casos ya clasificados.

Con clasificaciones más avanzadas podemos determinar estados sentimentales, tanto positivos (felicidad, alegría), como negativos (ira, enojo, tristeza).

2. Twitter: el auge del microblogging

Desde su creación en 2006, esta plataforma fue adquiriendo más relevancia hasta llegar a ser lo que es en la actualidad, en donde millones de usuarios de todas partes del mundo comparten

sus experiencias, opiniones y visiones con respecto a una variedad notable de temas o sucesos.

A medida que fue evolucionando paulatinamente con el paso de los años se fueron incorporando distintas formas de enriquecer las publicaciones realizadas en esta red social, pudiendo ahora encontrar dentro de un tweet información en distintos formatos como vídeos, imágenes, GIFs, emojis, hashtags y texto.

Debido a estas características nombradas se puede considerar a Twitter como la fuente más importante de opiniones a la hora de llevar a cabo un proyecto de análisis de sentimiento sobre una determinada temática, gracias al enorme abanico de posibilidades que posee la gente a la hora de expresarse.

En la Figura 1 se pueden observar algunas de las relaciones de Twitter [4].

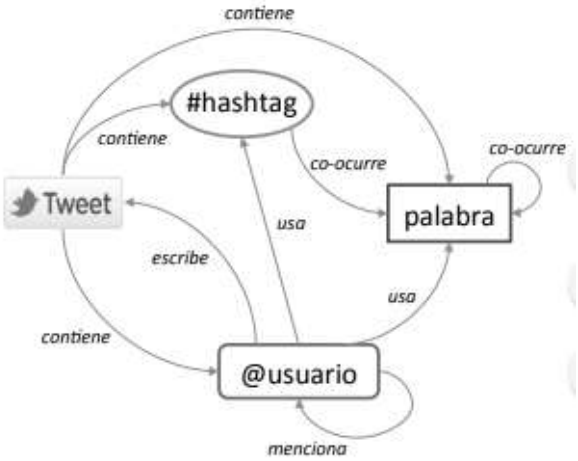

Figura 1: Relaciones en Twitter

publicado en redes sociales, principalmente Twitter, por su formato de microblogging.

Las aplicaciones en la vida real son sólo parte del porqué el análisis de sentimiento se ha vuelto un tópico popular de investigación.

Adicionalmente, es visto como un gran desafío en lo que respecta al PNL, creciendo exponencialmente el interés en esta problemática desde su aparición [6].

Luego de definir el nivel de análisis se debe determinar la técnica de análisis de sentimiento a utilizar.

Los dos enfoques más populares se tratan de técnicas de Machine Learning y el uso de un Diccionario léxico-emocional, como muestra la Figura 2 [7].

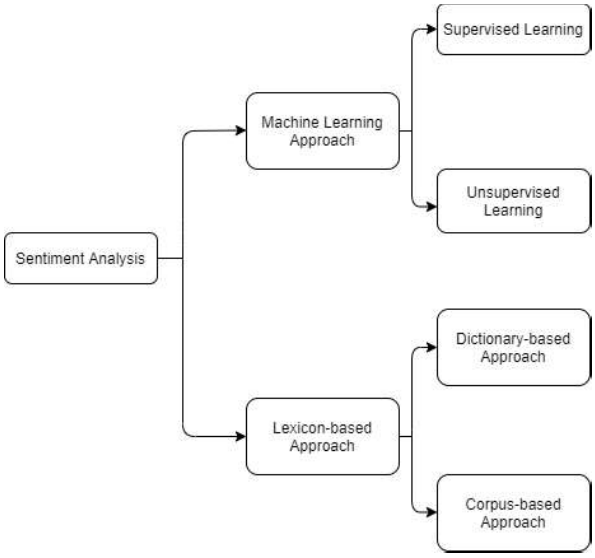

Figura 2: Tipología de las técnicas de Análisis de Sentimiento en Twitter

3. Análisis de sentimiento en la comunicación

Análisis de sentimiento (también conocido como minería de opinión) se refiere al uso de procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de los recursos.

"*Desde el punto de vista de la minería de textos, el análisis de sentimientos es una tarea de clasificación masiva de documentos de manera automática, en función de la connotación positiva o negativa del lenguaje ocupado en el documento.*" [5]

Actualmente, esta práctica se realiza mayoritariamente para la interpretación de lo

3.1. Aprendizaje Automático (Machine Learning)

El aprendizaje automático o también conocido como Machine Learning se divide en dos grandes tipos según haya o no retroalimentación, por un lado, está el aprendizaje automático no supervisado y por el otro el aprendizaje automático supervisado.

Los algoritmos utilizados por el aprendizaje automático no supervisado realizan el procesamiento en base únicamente a las entradas, es decir que el programa se va configurando a medida que procesa las observaciones que va recibiendo, en cambio, los algoritmos utilizados por el aprendizaje

automático supervisado cuentan con un corpus manualmente clasificado, sobre el cual se llevan a cabo dos procesos: encontrar los mejores parámetros para el algoritmo, y evaluar el nivel de fiabilidad con esos parámetros.
A esta fase se le llama de aprendizaje o entrenamiento.
A la hora de analizar tweets el algoritmo más adecuado de utilizar es del aprendizaje automático supervisado, a su vez este se divide en algoritmos de regresión o de clasificación.
Es conveniente utilizar este último ya que la tarea del mismo es asignar una categoría al texto de entrada, la cual puede ser dividida en tres (positivo/negativo/neutro) o en dos (positivo/negativo) [7].
Una vez definido cada uno de los puntos citados anteriormente, se debe proceder a extraer los atributos de cada uno de los tweets recolectados. Es por ello que se debe realizar un preprocesamiento de los mismos, en donde el objetivo principal es reducir la dimensionalidad y así, poder captar los conceptos más importantes a través de la eliminación de las *stopwords* o palabras vacías y la lematización. Las *stopwords* son aquellas palabras que no aportan información, mientras que la lematización consiste en reducir las palabras a su lema [8].
Luego, se debe definir si un atributo estará compuesto por una palabra o por varias (en general dos o tres) y a partir de ello seleccionar el algoritmo más adecuado, un ejemplo de ellos es Support Vector Machine (SVM) el cual es uno de los más utilizado a la hora de realizar este tipo de análisis ya que al utilizarlo como clasificador da buenos resultados. Cabe destacar que estos algoritmos serán aplicados al conjunto de entrenamiento y luego al conjunto de prueba que se quiera clasificar.

3.2. Diccionario léxico:

La alternativa más popular al aprendizaje supervisado es el uso de diccionario para la orientación semántica. En este caso no es necesario entrenar un modelo y supervisarlo. El uso de esta técnica se basa en algoritmos mucho más sencillos y en la utilización de un diccionario comúnmente llamado lexicón, en el cual encontramos generalmente una importante cantidad de palabras polarizadas y ponderadas, según su índole positiva o negativa y la intensidad del sentimiento que infieren las mismas.
Es posible también encontrar diccionarios en los cuales se clasifiquen a las palabras según la emoción que connotan.
El algoritmo usado por estos métodos es de características mucho más simples que los anteriormente nombrados.
La evaluación de un tweet implica solamente la detección de coincidencias en el texto con elementos del conjunto del diccionario, y realizar una ponderación final según todas las palabras encontradas, su concurrencia y la fuerza del sentimiento, para determinar entonces la orientación sentimental de la oración.
Existen varios diccionarios presentes en la web. La mayor parte de éstos (y los más refinados) están diseñados para el inglés. Existen diccionarios en español, pero no tan precisos y todavía faltos de maduración. Otra desventaja detectada a la hora de llevar a cabo esta técnica para realizar análisis de sentimientos es que la misma no considera el contexto en el que fue escrito un tweet en particular, ya que consiste solo en ponderar el mismo a partir del valor de cada una de las palabras que lo integran que se encuentran en el diccionario, es por ello que un tweet que fue escrito con una connotación negativa podría ser considerado positivo por el modelo y viceversa.
Algunos de los diccionarios más reconocidos son *Linguistic Inquirty and Word Count (LIWC), SentiWordNet, Q-WordNet, MPQA Subjectivity Lexicon, Big Liu Opinion Lexicon, AFINN y The General Inquirer.*

4. Conclusión

El análisis de sentimientos basado en la extracción de opiniones realizadas por las personas en redes sociales es una de las prácticas más utilizadas dado a su múltiple aplicación en diferentes temáticas, y por el hecho de que la misma es más económica que realizar encuestas en forma física y nos permite abarcar un gran segmento del mercado ya que las redes sociales

no discriminan las edades de sus usuarios ni los estatutos sociales.

Como se fue describiendo a lo largo del artículo, el análisis de sentimientos conlleva la aplicación de técnicas de Procesamiento del Lenguaje Natural el cual está muy relacionado al idioma y territorio en el cual se realizará el análisis. A partir del español se identifican los siguientes problemas [8]:

- Ambigüedad a nivel palabra: una palabra puede tener más de un significado.
- Ambigüedad sintáctica: Por ejemplo "María se encontró con Raquel para calmar su preocupación" no se puede definir si la preocupación es de María o de Raquel
- Resolución de la anáfora: se define anáfora al uso de una expresión cuya interpretación depende de otra expresión presente en el contexto del discurso (llamado su antecedente).
- Presuposición: "Ha dejado de fumar" implica que antes fumaba
- Ironía- Sarcasmo: "Había olvidado que tú eras el más inteligente, y que todos los demás éramos unos tontos."

Sin embargo, considerando dichos problemas y utilizando las técnicas y algoritmos adecuados se puede llegar a resultados muy beneficiosos para la persona u organización que realiza el análisis.

Un trabajo futuro a realizar en este ámbito sería el desarrollo de herramientas fiables para el análisis de sentimiento en tiempo real en base a *keywords* o palabras claves dadas utilizando cualquiera de las técnicas más conocidas nombradas en este artículo

Referencias Bibliográficas

[1] *Twitter crece en ganancias y usuarios, y supera las expectativas.* (25 de abril de 2018). Diario Clarín. Obtenido de:https://www.clarin.cim/tecnologia/twitter-crece-ganancias-usuarios-supera-expectativas_0_H1NnFZo2G.html

[2] Real Academia Española y Diccionario del Español Jurídico. (2019). *Big Data*. Obtenido de: https://dej.rae.es/lema/big-data.

[3] Delwiche, L. D., & Slaughter, S. J. (2019). *The little SAS book: a primer.* SAS Institute.

[4] Cotelo J.M, Cruz F., Ortega F.J, Troyano J.A. (2015). *Explorando Twitter mediante la integración de información estructurada y no estructurada*. Sociedad Española para el PNL (55), 75-82.

[5] Liu, B. (2007). *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Chicago, Estados Unidos: Springer.

[6] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. doi:10.2200/S00416ED1V01Y201204HLT016.

[7] Baviera, T. (2017) *Técnicas para el Análisis de Sentimientos en Twitter: Aprendizaje Automático Supervisado y SentiStrength.* Dígitos (3),33-50. Recuperado de: https://www.researchgate.net/publication/317256429_Tecnicas_para_el_Analisis_de_Sentimiento_en_Twitter_Aprendizaje_Automatico_Supervisado_y_SentiStrength .

[8] Hernández Orallo, J. (2004) *Introducción a la minería de datos.* Madrid, España:Pearson.