Clasificación de Reseñas por Análisis de Sentimientos

1rd Palacios Flores Gonzalo Emilio
Ingeniería en Computación
Facultad de Ingeniería, UNAM
gonzooooo10@gmail.com

2nd Rosales Cañedo Daniela Susana
Ingeniería en Computación
Facultad de Ingeniería, UNAM
surosal26@gmail.com

3th Suxo Pérez Luis Axel
Ingeniería en Computación
Facultad de Ingeniería, UNAM
axelsuxo@hotmail.com

4th Valenzuela Marin Juan
Ingeniería en Computación
Facultad de Ingeniería, UNAM
juan.valenzuela08@gmail.com

***Abstract*– El siguiente artículo presenta la clasificación de reseñas así como su implementación con la herramienta Vader, de tal forma que se haga un análisis de sentimientos y una interpretación de los resultados.**

I. INTRODUCCIÓN

A. Antecedentes

Mediante el análisis de sentimientos podemos determinar el modo en el cual los seres humanos visualizamos los elementos y la manera en la cual perciben en este caso las películas.

Tomando en cuenta la implementación de las palabras utilizadas por los críticos, se logra encontrar el tipo de reseña existente sobre el tema dado, por lo cual, se permite la clasificación de las reseñas mediante este análisis de sentimientos planteado.

B. Hipótesis

Principalmente, se debe descifrar qué es lo que motiva a las personas a expresar su opinión con respecto a las carteleras del momento, dependiendo sus preferencias o incluso la edad; posteriormente, realizar el análisis de sentimientos respecto al lenguaje empleado.

Tomando esto en consideración, suponemos que puede tratarse de una reseña sobre una película animada, pensada para un público en general, y algunas críticas son de quienes no se sienten satisfechos con la película, debido a sus gustos y la edad actual que tienen al momento de la reseña.

Otra polémica que puede darse, es un género que a todos los amantes del cine les agrade, que puede ser el suspenso o terror, en el cual la mayor parte de las críticas sean positivas o negativas, tomando en cuenta el grado de suspenso o terror que se tuvo en la cinta.

C. Impacto

Para el caso planteado, el impacto será social, ya que determina y clasifica a las películas analizadas dependiendo de los resultados que se obtengan, con base a las críticas recibidas y la clasificación de edad que pueden tener.

D. Descripción del Trabajo

La finalidad que tiene este trabajo, permite realizar un análisis de sentimientos, tomando como base las reseñas dadas a las películas del momento; realizando una clasificación y la probabilidad existente de que una película futura, pueda ser tomada como buena o pésima, dependiendo su género, la clasificación del público que puede verla, y los gustos de las personas.

Teniendo la tendencia y realizando, mediante redes sociales, las recomendaciones de la cinta que puede ser del agrado del usuario, y las mejores opciones para el fin de semana.

Tomando como herramientas base, VADER y NLTK para realizar los análisis pertinentes, y concluir con los objetivos que aquí se plantean, buscando la óptima respuesta de recomendaciones válidas.

II. TRABAJO PREVIO

Se han realizado diversos estudios respecto al algoritmo de Vader y NLTK para el análisis de sentimientos, en está sección se revisarán algunos de los estudios realizados por diferentes instituciones y así mostrar los casos de uso.

- *Análisis de sentimientos para Twitter con Vader.*

  El análisis de sentimientos en apps de redes

sociales como twitter, es capaz de realizar una evaluación para cada tiut, observando si es positivo, negativo o neutro y si el entrenamiento del algoritmo puede generar discrepancias en las métricas de Precisión, Exactitud, sensibilidad, especificidad y la matriz de confusión.
El algoritmo de Vader para el análisis de sentimientos en las redes sociales utiliza los puntos

1. LIWC
2. ANEW
3. General Inquirer
4. SentiWordNet
5. Machine learning con Naive Bayes
6. Máxima entropía
7. Máquinas de soporte vectorial

- *Análisis de sentimientos de opiniones en redes sociales usando técnicas de procesamiento del lenguaje natural.*

  El análisis de sentimientos en la API de Twitter y Reddit y clasificarlas a partir de post de reddit de tipo /r/politics y en twitter eligiendo los 10 primeros buscando palabras claves, clasificándolas como positiva, negativa o neutra, resolviendo problemas de lenguaje natural utilizando el algoritmo de Vader implementado en el paquete NLTK en python, es posible aplicar a datos de texto sin etiquetas.

- *Enriquecimiento del modelo basado en reglas Vader a través de lexicones*

  A partir de las redes sociales que cuentan con un gran impacto a nivel mundial, es posible analizar una gran cantidad de emociones a partir de que compartan opiniones de eventos sociales, utilizando lexicón que es un listado de palabras utilizado para clasificar las emociones expresadas en las opiniones, sin embargo, dependerá de qué el estado del arte enriquezca las reglas del algoritmo de Vader.

- *Análisis de sentimientos: Herramienta para estudiar datos cualitativos en la investigación jurídica.*

  El enfoque del algoritmo Vader en la investigación jurídica se basa en el procesamiento del lenguaje natural y los números neutrosóficos de valor único utilizando la herramienta Orange, debido a que en la investigación jurídica existe una subjetividad por los prejuicios inconscientes entre autores e investigadores en las entrevistas individuales, obteniendo datos cualitativos y conclusiones amplias.

- *Análisis de la reacción del consumidor en Youtube*

  El enfoque del algoritmo Vader para el análisis de la reacción del consumidor en la plataforma de youtube se basa en el análisis de comentarios, likes, número de visualizaciones, titulo del video o canal, fecha de publicación y la descripción del video, se realiza una clasificación entre positivo, negativo o neutro asignado los valores de 1, -1, 0 respectivamente, a partir de clasificación es posible relacionar los el análisis de sentimientos con el contenido.

- *Desarrollo de una metodología y del código correspondiente para NLP*

  Es posible realizar el un análisis de sentimientos a las valoraciones, realizando una implementación de un módulo de web scraping, modelos de regresión logística y Naive Bayes, para poder clasificar los sentimientos en positivo, negativo y neutro, realizando comparaciones con otros modelos y así verificar que los resultados sean correctos.

Podemos identificar que el algoritmo de Vader incluido en la librería de NLTK de python que es posible realizar una clasificación de los sentimientos en positivo, negativo o neutro a partir de las técnicas de lenguaje natural y el aprendizaje automático, siendo posible aplicarlos en diferentes campos de estudio, como medicina, derecho, informática, filosofía, etc., observando así métricas obtenidas de acuerdo al análisis previsto, como se verá en las secciones siguientes.

III. MARCO TEÓRICO

A. Kit de herramientas de lenguaje natural (NLTK)

Es una plataforma para trabajar con el lenguaje humano mediante la creación de programas de Python, capaz de utilizar más de 50 corpus entre otros recursos léxicos, trabajando en conjunto con clasificaciones de texto, tokens, etiquetado, análisis y razonamiento semántico.

El NLTK es utilizado por lingüistas, ingenieros, estudiantes, profesores, investigadores y por empresas por igual, también por las ventajas que tiene, dado que es multiplataforma, de código abierto y tiene soporte por la comunidad.

Fig. 1 NLTK disponible en Windows, Mac OS X y Linux

Es mayormente utilizada para la enseñanza para trabajar con textos utilizando python, como objetivo se tiene un acceso fácil e interactivo donde los usuarios nuevos pueden jugar y aprender con el lenguaje natural.

Esta cuenta con una variedad de subpaquetes y submódulos.

- nltk.app package
- nltk.ccg package
- nltk.chat package
- nltk.chunk package
- nltk.classify package
- nltk.cluster package
- nltk.corpus package
- nltk.draw package
- nltk.inference package
- nltk.lm package
- nltk.metrics package
- nltk.misc package
- nltk.parse package
- nltk.sem package
- nltk.sentiment package
- nltk.stem package
- nltk.tag package
- nltk.tbl package
- nltk.test package
- nltk.tokenize package
- nltk.translate package
- nltk.tree package
- nltk.twitter package

Fig. 2 Sub Paquetes NLTK

- nltk.book module
- nltk.cli module
- nltk.collections module
- nltk.collocations module
- nltk.compat module
- nltk.data module
- nltk.decorators module
- nltk.downloader module
- nltk.featstruct module
- nltk.grammar module
- nltk.help module
- nltk.internals module
- nltk.jsontags module
- nltk.lazyimport module
- nltk.probability module
- nltk.text module
- nltk.tgrep module
- nltk.toolbox module
- nltk.treeprettyprinter module
- nltk.treetransforms module
- nltk.util module
- nltk.wsd module

Fig. 3 Submódulos NLTK

Tiene muchos ejemplos en su página oficial https://www.nltk.org/howto.html , en donde enseñan de forma básica a utilizar los paquetes, y anexan la descripción de los subprocesos y submódulos, estos también cuenta con descripción de sus funciones.

La comunidad cuenta con el apoyo de una Wiki que se encuentra en Github https://github.com/nltk/nltk/wiki , en donde se puede encontrar información de su uso, documentación y en mejor detalle lo más requerido por los usuarios.

B. Diccionario consciente de valencia y razonador de sentimientos (VADER)

Es una herramienta de análisis de sentimientos que utiliza reglas y léxico que está dedicada exclusivamente a los sentimientos, VADER suele ser utilizada para fines políticos y para empresas, estudiando el ambiente de sentimientos en redes sociales o en textos digitales.

VADER puede integrarse con NLTK, permitiendo una mejor compatibilidad y manejo de datos utilizando Python.

Los datos que se obtienen son de tipo de score (puntuación), de esta forma se puede obtener con una mejor clasificación de sentimientos [1], las reglas son las siguientes:

- El Score más cercano a 1 significa que es positivo.
- El Score más cercano a -1 significa que es negativo.
- El Score más cercano a 0 significa que es neutro.

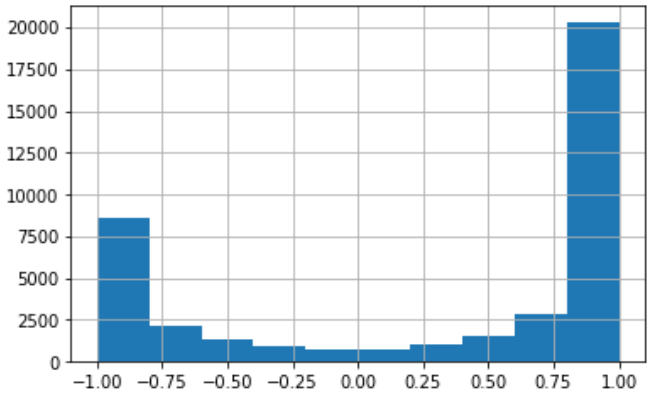

Fig. 4 Representación del score en una gráfica

En un sentido más estricto, se puede tomar el score generalmente de la siguiente manera:

- Sentimiento positivo: Score >= 0.5
- Sentimiento negativo: Score <= -0.5
- Sentimiento neutral: Score > -0.5 y Score < 0.5

C. Análisis de sentimientos

El análisis de sentimiento está basado en lógica computacional basada en opiniones, sentimientos y emociones, estas dependiendo de su enfoque pueden estar interpretadas sólo por las palabras sin entender el contexto o entendiendo el contexto del texto.

Es utilizada para personas (influencer), fines políticos y productos, con la finalidad de saber las emociones que transmiten a los receptores y el grado de las emociones.

D. Naive Bayes

Es un algoritmo de clasificación de machine learning, que utiliza el “Teorema de Bayes”. En donde las variables son independientes entre sí. Además

proporciona una manera fácil de construir modelos de comportamiento debido a su simplicidad.

Para calcularlo es necesario obtener la probabilidad a posteriori de que ocurra un evento A, dado la probabilidad a priori de qué ocurra R. [2]

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

Fig. 5 Clasificador de Naive Bayes

IV. EXPERIMENTAL

A. Corpus

Para el desarrollo de este trabajo se utilizó como datos de entrada un conjunto de 40,000 críticas de diversas películas, extraídas del portal Rotten Tomatoes, disponible en la plataforma de Kaggle (https://www.kaggle.com/datasets/yasserh/imdb-movie-ratings-sentiment-analysis)[3].

De acuerdo con la fuente, cada oración ha sido analizada en muchas frases por un analizador de Stanford. Cada oración tiene un SentenceId. Las frases que se repiten (como palabras cortas o comunes) solo se incluyen una vez en los datos. Además, se proporciona un SentenceId para que pueda rastrear qué frases pertenecen a una sola oración.

Estas oraciones ya fueron clasificadas anteriormente y se les asignó una etiqueta que representa el sentimiento que se determinó. Las etiquetas de sentimiento son:

- 0 - negativo
- 1 - positivo

Estas etiquetas obtenidas previamente ayudarán para compararlas con los resultados que nos arroje nuestro programa.

El conjunto original de datos se puede descargar aquí: https://archive.ics.uci.edu/ml/datasets/spambase

B. Metodología/Proceso

El código realizado puede verse en: https://colab.research.google.com/drive/1ncfEbxtV6Cnw6ch_MD3t43mI0W0DDN89?usp=sharing

El objetivo de este trabajo es realizar el análisis de sentimientos en Python utilizando la biblioteca NLTK, una herramienta básica para el análisis y procesamiento de textos.

Como ambiente de desarrollo utilizaremos la plataforma de Google Collaboratory.

Primero importamos todas la bibliotecas que requerimos:

- NLTK
- Pandas
- Matplotlib

Para realizar el análisis de sentimientos en Python con NLTK vamos a usar dos componentes que tenemos que descargar: vader y punkt.

Posterior a ello, debemos importar nuestro conjunto de datos con el que vamos a trabajar, para ello utilizamos la herramienta de carga de datos propia de Google Collaboratory, solo debemos seleccionar el archivo con extensión .csv desde nuestra computadora. Otra opción sería importarlos desde Google Drive.

Una vez que tenemos nuestro conjunto de datos, debemos asignarlo a un dataframe de Pandas para poder realizar el procesamiento.

Ahora, realizamos el algoritmos, para ello, utilizando el SentimentIntensityAnalyzer() generamos los indicadores de sentimiento para cada uno de nuestros datos, asignándoles a un arreglo y posteriormente agregándoles como una columna más en el dataframe.

Por último, para poder visualizar mejor los resultados graficamos cada una de las columnas que obtuvimos, así como la columna de las etiquetas que ya estaban agregadas.

C. Resultados

Podemos observar los resultados obtenidos mediante histogramas.

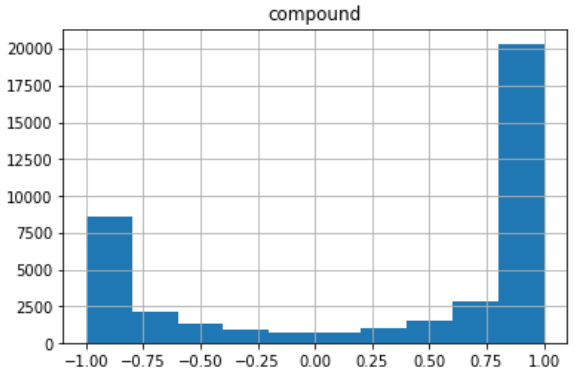

Fig. 6 Histograma del índice compuesto

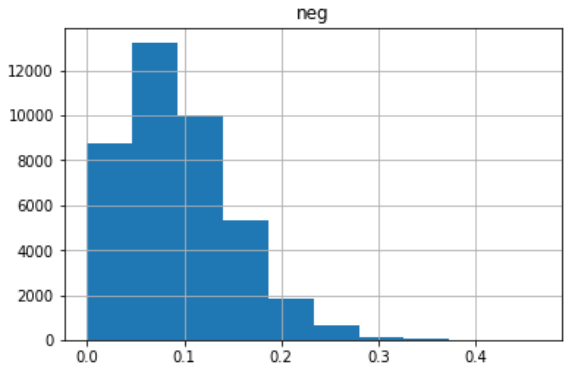

Fig. 7 Histograma del índice negativo

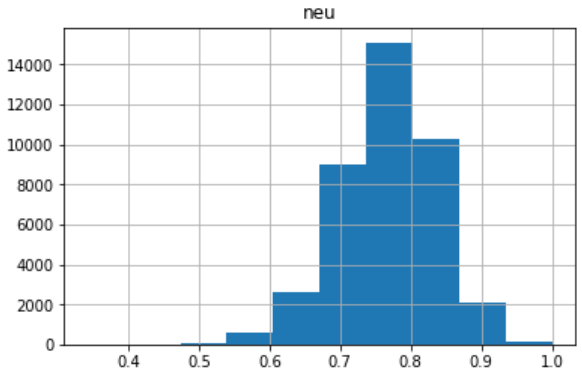

Fig. 8 Histograma del índice neutro

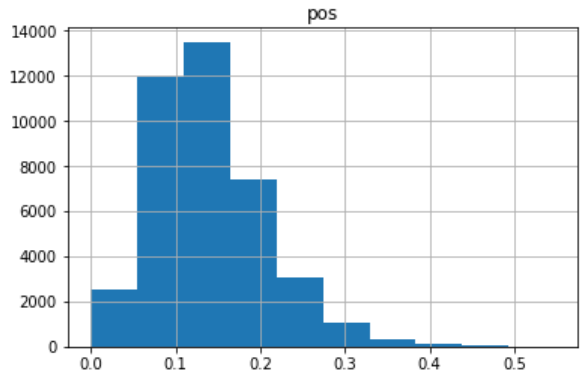

Fig. 9 Histograma del índice positivo

Para poder comparar nuestros resultados, utilizamos el histograma generado con las etiquetas que venían en el conjunto de datos

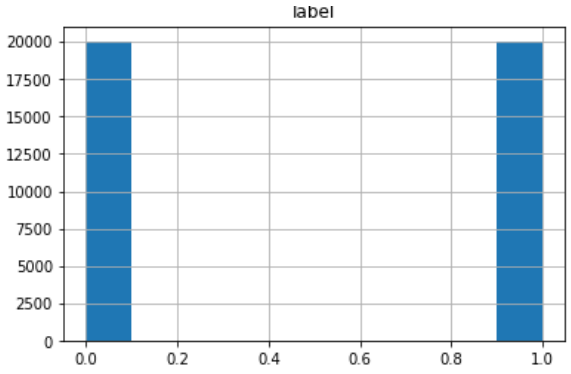

Fig. 10 Histograma de etiquetas

Podemos notar que previamente las etiquetas mostraban que los datos estaban bien distribuidos, aproximadamente la mitad de ellos son positivos y la otra mitad negativos.

Pero al observar nuestros resultados, podemos notar que esta distribución no es tan simétrica. Con el índice compuesto podemos ver que la gran mayoría de los datos muestran un sentimiento positivo, mientras que el resto muestran un sentimiento negativo o neutro.

Las otras tres gráficas nos muestran que todos los datos tienen componentes negativos, neutros y positivos, pero nunca llegan al límite superior de la gráfica. En el caso de los índices positivo y negativo muestran que, en la mayoría de los casos, estos componentes no superan el 0.3, mientras que el componente neutro se encuentra mejor distribuido.

V. CONCLUSIONES

Mediante el análisis de sentimientos podemos determinar el sentir de las personas con respecto a un tema en específico, en este caso, a través de críticas determinamos si una película les gustó o no. Esta técnica de análisis del texto es de gran ayuda para analizar estrategias durante la toma de decisiones, no solo en película, si no en cualquier ámbito, ya que podemos observar el sentir de las personas con respecto a un producto o servicio, y con ello determinar qué acciones llevar a cabo y cuales evitar para mejorar.

Podemos usar esta técnica para realizar múltiples tareas, como la recomendación automática de materiales a usuarios o la clasificación de estos. En este caso, podríamos clasificar las películas que recibieron críticas con un índice de positividad alto para recomendarlas a otros usuarios, o utilizar las críticas realizadas por un usuario para recomendar películas que puedan causar una impresión positiva

VI. REFERENCIAS

[1] cjhutto. (1 de abril de 2022). VaderSentiment. GitHub. Recuperado el 31 de mayo de 2022 de:

https://github.com/cjhutto/vaderSentiment#installation

[2] Roman V. (25 de abril de 2019). Algoritmos Naive Bayes: Fundamentos e Implementación. Medium. Recuperado el 31 de mayo de 2022 de: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementación-4bcb24b307f

[3] Yasser H. (febrero de 2022). IMDB Movie Ratings Sentiment Analysis. Kaggle. Recuperado el 31 de mayo de 2022 de: https://www.kaggle.com/datasets/yasserh/imdb-movie-ratings-sentiment-analysis

[4] Hopkins M., Reeber E., Forman G., Suermondt J. Hewlett-Packard Labs. (1 de julio de 1999). Spambase Data Set. UCI. Recuperado el 31 de mayo de 2022 de: https://archive.ics.uci.edu/ml/datasets/spambase

[5] Álvarez D. (6 de junio de 2019). Análisis de la reacción del consumidor en Youtube. Universidad de la Laguna. Recuperado el 31 de mayo de 2022 de: https://riull.ull.es/xmlui/bitstream/handle/915/14520/Analisis%20de%20la%20reaccion%20del%20consumidor%20en%20Youtube.pdf?sequence=1

[6] Valle G. (2021). Desarrollo de una metodología y del código correspondiente para NLP. Universidad Pontificia Comillas. Recuperado el 31 de mayo de 2022 de: https://repositorio.comillas.edu/xmlui/handle/11531/55160

[7] Alemán S. (2021). Análisis de sentimientos para Twitter con Vader y TextBlob. Revista Odigos. Recuperado el 31 de mayo de 2022 de: https://revista.uisrael.edu.ec/index.php/ro/article/view/494

[8] Huto C. & Gilberto E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM. Recuperado el 31 de mayo de 2022 de: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[9] Scotto J. (2021). Análisis de sentimientos de opiniones en redes sociales usando técnicas de procesamiento del lenguaje natural. Universidad de Belgrano. Recuperado el 31 de mayo de 2022 de: http://repositorio.ub.edu.ar/bitstream/handle/123456789/9534/Scotto.pdf?sequence=1&isAllowed=y

[10] Mejía K. (12 de julio de 2018). Enriquecimiento del modelo basado en reglas Vader a través de lexicones. UAEM. Recuperado el 31 de mayo de 2022 de: https://ri.uaemex.mx/handle/20.500.11799/99592