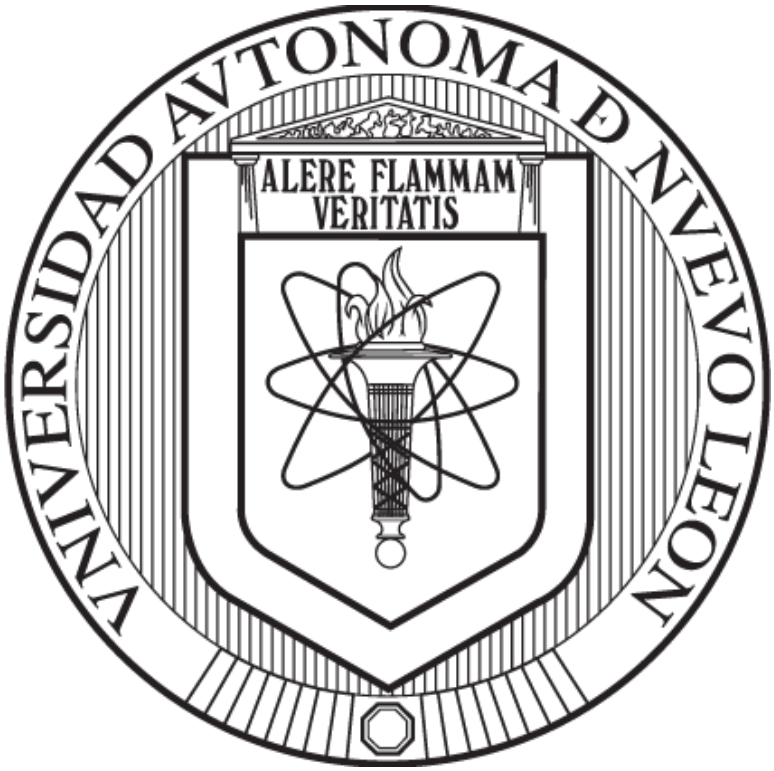
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN



Reporte de tarea

Análisis de Sentimiento

Autor: **Karla Cureño Vega** Matrícula: **2085376**

Materia: **Procesamiento y Clasificación de Datos** Profesor:**Mayra Cristina Berrones**

Actividad: **Tarea 2** Fecha: **26 de mayo de 2022**

Índice

1. Introducción **3**

2. Descripción del conjunto de datos **3**

3. Preprocesamiento de datos **3**

4. Análisis de Sentimiento **3**
4.1. Textblob 3
4.2. VADER 3
4.3. SentiWordNet 4

5. Resultados **4**
5.1. TextBlob 4
5.2. VADER 5
5.3. SentiWordNet 6
5.4. Comparación entre técnicas 7

6. Conclusiones **8**

1. Introducción

El análisis de sentimiento es el uso del procesamiento del lenguaje natural para identificar, extraer, cuantificar y estudiar connotaciones emocionales e información subjetiva sobre un texto. Este tipo de análisis tiene numerosas aplicaciones y es muy utilizado en servicio al cliente para determinar si la percecpción que el cliente tiene de un servicio es positiva, negativa o neutral.

2. Descripción del conjunto de datos

El conjunto de datos recopilado de Kaggle consiste en 50,000 reseñas de películas de la base IMDB. Estas reseñas contienen una etiqueta de sentimiento, en este caso los valores que puede tomar esa etiqueta solo son positivo o negativo. Debido a que esta etiqueta existe en el conjunto de datos original, esta se podrá utilizar para evaluar las diferentes técnicas a utilizar en el análisis de sentimiento.

El subconjunto a utilizar consta de las primeras 1,000 reseñas de este conjunto de datos. Se recortó el dataset para ahorrar costo computacional.

3. Preprocesamiento de datos

Antes de comenzar a comparar las diferentes técnicas para realizar el análisis de sentimiento se realiza un preprocesamiento de las reseñas que serán utilizadas. Este consiste básicamente en:

- Conversión a minísculas.
- Remoción de etiquetas HTML.
- Remoción de caracteres no alfábeticos.
- Remoción de "stop-words".
- Lematización de palabras.

4. Análisis de Sentimiento

4.1. Textblob

El análisis de sentimiento utilizando la librería `textblob` se apoya en el cálculo de subjetividad y polaridad para analizar los textos.

Teniendo el resultado de la métrica de polaridad, esta se utiliza para denfinir el sentimiento de cada texto. Si la polaridad es negativa el sentimiento es negativo, si es positiva es un sentimiento positiva, y si la polaridad es igual a 0 se define que el sentimiento del texto es neutral.

4.2. VADER

El análisis de sentimiento utilizando la librería `vaderSentiment` se apoya en el cálculo de polaridades y el compuesto de las mismas para analizar los textos.

Esta técnica realiza un cálculo de polaridad negativa, neutral y positiva para cada texto. Después de esto se realiza un compuesto de esas 3 polaridades y con este compuesto se define el sentimiento del texto. Si el cálculo compuesto es mayor o igual a 0.5 se refiere a un sentimiento positivo, menor o igual a -0.5 se refiere a uno negativo y el resto de los valores se asignan a un sentimiento neutral.

4.3. SentiWordNet

Utilizar `sentiwordnet` de la librería `nltk` consiste en un recurso léxico muy útil para el análisis de sentimiento de textos.

Esta técnica compara las palabras de un texto con el corpus de `sentiwordnet` y de acuerdo a la puntuación que tiene cada palabra en este corpus se realiza el cálculo del sentimiento del texto original. Si el cálcula resulta en un valor positivo el sentimiento del texto será positivo, si es valor negativo será un sentimiento negativo y si es igual a 0 se define un sentmiento neutral.

5. Resultados

5.1. TextBlob

De los 1,000 registros que se analizaron la técnica `textblob` definió que 714 eran textos con sentimiento positivo, 286 con negativo y no hubo ninguno que cayera en la categoría neutral. Esto es muy interesante debido a que en el conjunto de datos original ya existe una etiqueta de sentimiento y efectivamente, ninguno de los textos es neutral. Esto me hace sospechar de que esta técnica fue precisa para este conjunto de datos.

En la figura 1 se puede observar un diagrama de pastel para los sentimientos resultantes de este análisis, en la que se observa que el 71.4 % de los textos tienen una etiqueta de sentimiento positiva y el 28.6 % restante son negativos.

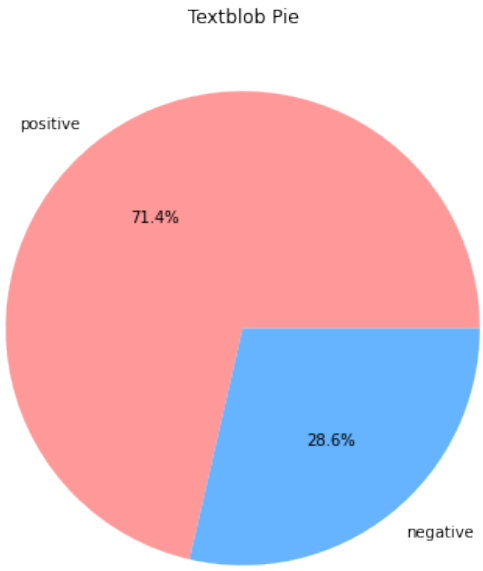


Figura 1: Resultados de Análisis de Sentimiento con TextBlob.

Para comparar con la etiqueta original se puede observar en la figura 2 la matriz de confusión resultante de esta técnica. Esta matriz muestra sobre el eje y la etiqueta original y sobre el eje x la nueva etiqueta de sentimiento, por lo que sobre la diagonal principal se posicionan los registros etiquetados de manera correcta al compararlos con la etiqueta original.

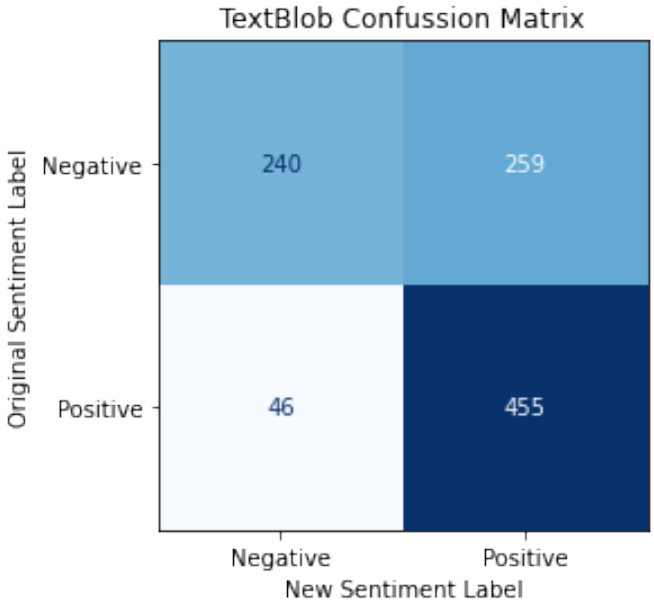


Figura 2: Matriz de Confusión para TextBlob.

Para esta técnica, se observa que hubo 240 reseñas etiquetadas de manera correcta como reseñas negativas y 455 correctamente etiquetadas como positivas. Hubo 46 reseñas positivas que la técnica etiquetó incorrectamente como negativas y 259 reseñas negativas que fueran incorrectamente etiquetadas como positivas.

Conociendo los datos provenientes de la matriz de confusión se puede hacer el cálculo de la exactitud (accuracy) que tuvo esta técnica. La exactitud se calcula diviendo la cantidad de datos que obtuvieron una etiqueta correcta sobre los datos totales. Para la técnica de `textblob` la exactitud resultó en un **69.5 %**. Este número no es malo, sin embargo tampoco es muy confiable por lo que hay que seguir evaluando técnicas para decidir cuál es la adecuada para este conjunto de datos.

5.2. VADER

De los 1,000 registros que se analizaron la técnica `vaderSentiment` definió que 606 eran textos con sentimiento positivo, 264 con negativo y 130 con un sentimiento neutral.

En la figura 3 se puede observar un diagrama de pastel para los sentimientos resultantes de este análisis, en la que se observa que el 60.6 % de los textos tienen una etiqueta de sentimiento positiva, el 26.4 % son negativos y el 13 % restante tienen un sentimiento neutral.

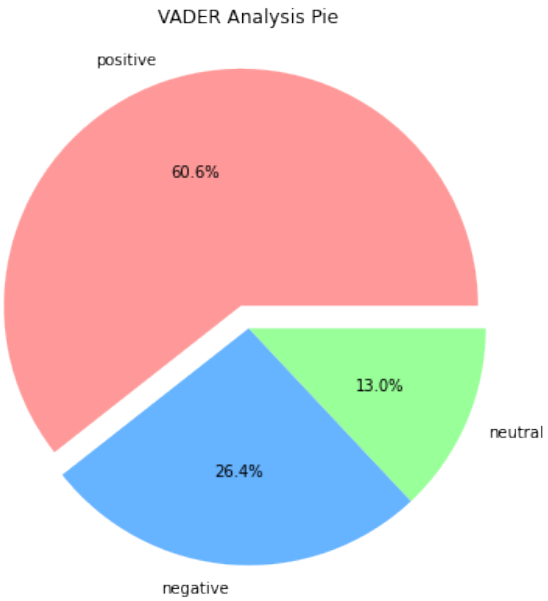


Figura 3: Resultados de Análisis de Sentimiento con VADER.

Comparando con la etiqueta original se puede observar en la figura 4 la matriz de confusión resultante de esta técnica. Conociendo los datos provenientes de esta matriz se puede hacer el cálculo de la exactitud que resultó en un **61.9 %** para esta técnica. Esta disminución en exactitud se puede deber a que en el conjunto de datos original no existe la etiqueta neutral y en esta técnica hay algunos textos que son etiquetados de esta manera.

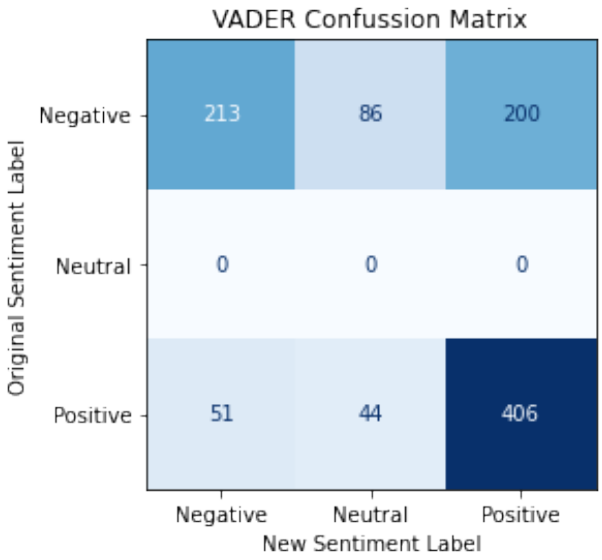


Figura 4: Matriz de Confusión para VADER.

5.3. SentiWordNet

De los 1,000 registros que se analizaron la técnica `SentiWordNet` definió que 679 eran textos con sentimiento positivo, 310 con negativo y 11 con un sentimiento neutral.

En la figura 5 se puede observar un diagrama de pastel para los sentimientos resultantes de este análisis, en la que se observa que el 67.9 % de los textos tienen una etiqueta de sentimiento positiva, el 31 % son negativos y el 1.1 % restante tienen un sentimiento neutral.

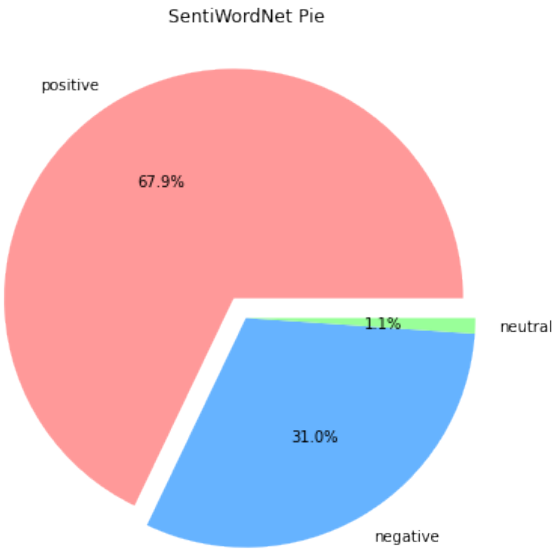


Figura 5: Resultados de Análisis de Sentimiento con SentiWordNet.

Comparando con la etiqueta original se puede observar en la figura 6 la matriz de confusión resultante de esta técnica. Conociendo los datos provenientes de esta matriz se puede hacer el cálculo de la exactitud que resultó en un **64.2 %** para esta técnica. Esta técnica también se ve afectada por la clasificación de textos neutrales, sin embargo como son muy pocos los textos que caen en esta categoría no se ve tan negativamente afectada como la técnica anterior.

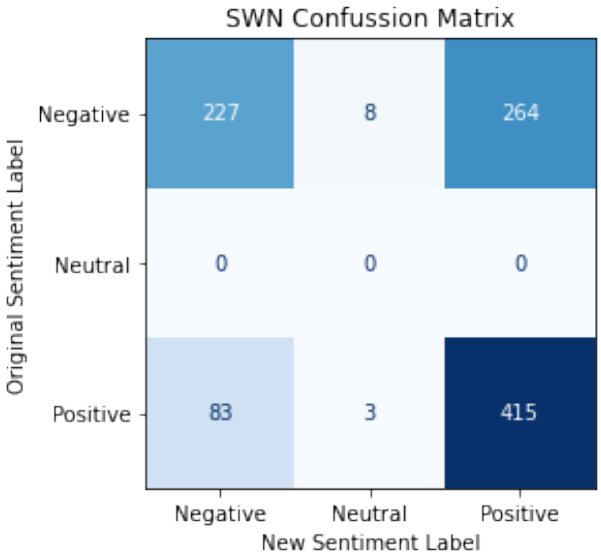


Figura 6: Matriz de Confusión para SentiWordNet.

5.4. Comparación entre técnicas

Es importante realizar una comparación entre las 3 técnicas utilizadas sobre este conjunto de datos. Para comenzar esta comparativa, se observan en 7 los resultados del análisis de las 3 técnicas. Es interesante notar que solo en la técnica de `textblob` no se encontró ningún texto con sentimiento neutral. Tal vez, esto en otro contexto causaría ruido en el análisis, sin embargo en esta caso se sabe de antemano que no existe una reseña neutral en el conjunto de datos por lo que desde aquí se puede intuir que la técnica más precisa fue esta.

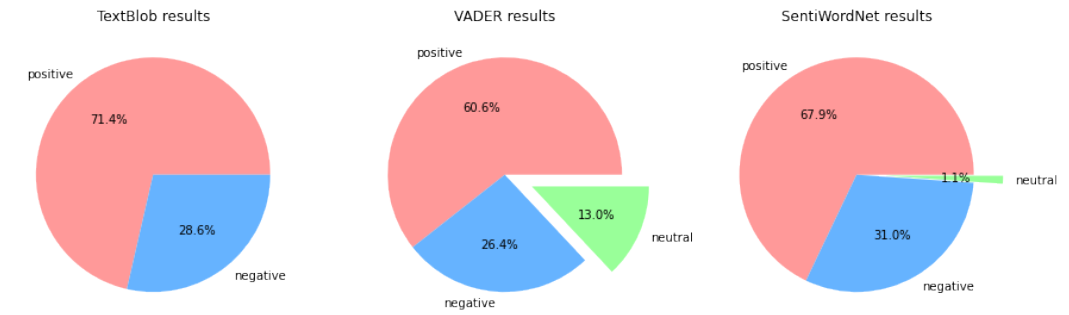


Figura 7: Resultados obtenidos por las 3 técnicas de análisis de sentimiento.

Aunado a esto, se observa en 8 las matrices de confusión obtenidas para cada una de las técnicas. Con la información de estas matrices se construye la tabla mostrada en 1 para comparar las exactitudes de las técnicas.

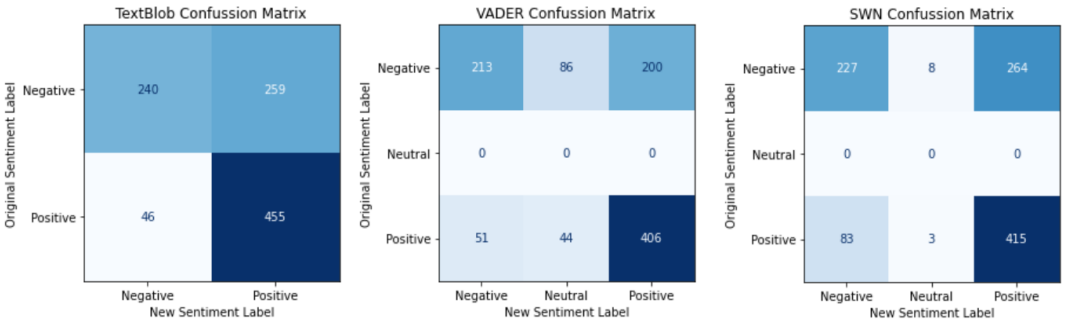


Figura 8: Matrices de confusión obtenidas por las 3 técnicas de análisis de sentimiento.

Técnica	**Exactitud (Accuracy)**
TextBlob	69.5 %
VADER	61.9 %
SentiWordNet	64.2 %

Tabla 1: Tabla comparativa de exactitud de las 3 técnicas empleadas.

Basado en esto, se concluye que la mejor técnica para este conjunto de datos fue la de `textblob`, sin embargo la diferencia de exactitud con las otras dos técnicas no es muy grande por lo que sería recomendable añadir más registros al análisis para hacerlo más robusto o cambiar por una base de datos cuyas etiquetas de sentimiento originales incluyan los 3 sentimientos posibles, esto con la finalidad de no sesgar los resultados hacia alguna técnica en particular.

6. Conclusiones

El análisis de sentimiento es una técnica analítica muy útil para traducir las emociones que pueden existir dentro de un texto. Conocer esto proporciona información relevante sobre el conjunto de datos a analizar y puede ser el inicio de un análisis mucho más complejo para la toma de decisiones.

Los resultados obtenidos con el análisis de sentimiento sobre el conjunto de datos estudiado permiten comparar las diferentes técnicas para realizar este análisis. Es importante considerar

las características del conjunto de datos para seleccionar la técnica adecuada. El uso de alguna métrica para evaluar la precisión del análisis también puede ser útil para elegir el método más preciso respecto al conjunto de datos.

Referencias

[1] Lakshmipathi, N. (2019). IMDB Dataset of 50K Movie Reviews. [Online]. Disponible en: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] Cureno, K. (2021). Análisis de Sentimiento. [Online]. Disponible en: https://github.com/karlacuv/MCD_Procesamiento/blob/main/Tarea2_AnalisisSentimiento.ipynb

[3] Berrones, M. (2021). Análisis de Sentimientoo. [Online]. Disponible en: https://github.com/mayraberrones94/FCFM/blob/master/Semana_2_Analisis_de_sentimiento.ipynb

[4] Loria, S. (2022). TextBlob: Simplified Text Processing. [Online]. Disponible en: https://textblob.readthedocs.io/en/dev/

[5] Beri, A. (2020). Sentimental Analysis Using Vader. [Online]. Disponible en: https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[6] Esuli, A. (2019). SentiWordNet. [Online]. Disponible en: https://github.com/aesuli/SentiWordNet