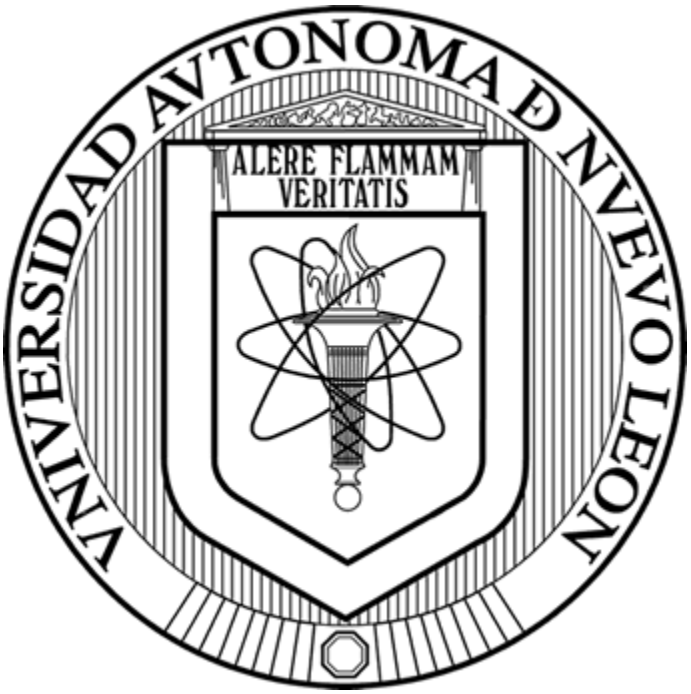
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS



REPORTE II

ANÁLISIS DE SENTIMIENTOS

Profesora: Mayra Cristina Berrones Reyes

Autor: Juan de Jesús Aguilar Solano Matrícula: 1576327

Jueves 26 de mayo de 2022

Contents

TABLA DE FIGURAS

INTRODUCCIÓN

El análisis de sentimientos es una herramienta que se ha vuelto muy popular en los últimos años, pues esto permite analizar una gran cantidad de datos en formato texto y extraer conocimientos que de otro modo se perderían. Alguna de las utilizades de este análisis es conocer la opinión general que se tiene de un objeto, persona o empresa y así poder actuar a tiempo en caso de que sea necesario.

En este proyecto se realizará un análisis de sentimientos a un conjunto de reseñas dadas a 1000 hoteles para extraer conocimiento acerca de la opinión general de sus inquilinos.

EXPERIMENTACIÓN

En este proyecto se utilizará la base de datos obtenida del sitio web kaggle [1] y que fueron extraídos mediante la API de twitter, esta base de datos contiene 515,738 reseñas de clientes hospedados en 1492 hoteles diferentes.

Diccionario de datos

- Hotel_Address: Dirección del hotel.
- Review_Date: Fecha en la que el usuario realizó la reseña.
- Average_Score: Calificación media del hotel calculada en base en los últimos comentarios del año.
- Hotel_Name: Nombre del Hotel.
- Reviewer_Nationality: Nacionalidad del usuario que realizó la reseña.
- Negative_Review: La reseña negativa que el usuario ha brindado al hotel, si no brindó reseña negativa, el campo contendrá la cadena “No Negative”.
- Review Total/NegativeWordCounts: El número de palabras negativas en la reseña.
- Positive_Review: La reseña positiva que el usuario ha brindado al hotel, si no brindó reseña positiva, el campo contendrá la cadena “No positive”.
- ReviewTotal/PositiveWordCounts: Número total de palabras positivas en la reseña.
- Reviewer_Score: La calificación que el usuario ha dado al hotel basado en su experiencia.
- TotalNumberofReviewsReviewerHasGiven: Número de reseñas que los usuarios han dado en el pasado.
- TotalNumberofReviews: El número total de reseñas válidas que el hotel tiene.
- Tags: Tags de las reseñas dadas al hotel.
- AdditionalNumber_Scoring: Indica cuantas calificaciones sin reseña ha brindado.
- lat: Latitud del hotel.
- lng: Longitud del hotel.

Lectura de datos

Para la lectura de datos se utilizó la biblioteca pandas, haciendo uso del método read_csv() yse aplicó un strip a la columna “Reviewer_Nationality” debido a que presentaba espacios al inicio y al final de la palabra. Para el propósito de este proyecto, la base de datos ha sido recortada y limitada a 100,000 observaciones de esta base de datos.

Cada observación contiene una reseña positiva y una negativa (al usuario se le solicito hacer esto), se realizará la operación melt, de moto que se tenga una observación por reseña positiva y negativo, por lo que ahora la base de datos se duplica a 200,000 observaciones y se ha añadido una columna llamada “Sentiment” que contiene justamente la etiqueta de la reseña (positiva/negativa).

Preprocesamiento

En esta fase del procesamiento de texto se procedió a realizar procesos básicos, los cuales son los siguientes:

- Convertir todos los caracteres a minúsculas.
- Remover signos de puntuación y caracteres especiales, los cuales son los siguientes:

  [¿, ?, ¡, ‘, “, #, :, “, ”, _, ., ,), (, \, /, *, –, $,%,&,+,;,<,=,>,@,[,],^_,`,{,|,},~]
- Normalizar los espacios, es decir, el espaciado doble, triple, cuádruple, etc. Es sustituido por un espaciado simple.
- Separar la cadena de texto en palabras y esta es guardada en una lista, tomando como referencia de separación el espaciado simple.
- Removemos las “stopwords” de la lista de palabras, las “stopwords” se obtuvieron de la biblioteca nltk [3].
- Añade contexto a las palabras mediante el uso de la función pos_tag() de la biblioteca nltk.
- Lemmatización de las palabras. Este proceso se lleva a cabo mediante el lematizador WordNetLemmatizer, también contenido en la biblioteca nltk tomando en cuenta el contexto de la palabra.

Función de preprocesamiento

Todos los pasos descritos durante este apartado se introdujeron en una sola función llamada preprocesamiento_words() y lemmatize(), se muestra a continuación el antes y después de la función:

```
Texto de entrada:
https://sitioweb.com Nombre: Juan, Correo: juan.agsolano@gmail.com

Lista de salida
['nombre', 'juan', 'correo', 'juan', 'agsolano', 'gmail', 'com']
```

El resultado de la función es una lista de palabras ya procesadas contenidas en jupyter notebook adjunto en este proyecto.

Una vez procesados los textos de las reseñas se procedió a filtrar las observaciones tomando como criterio la longitud de las palabras retornadas por la función antes descrita, es decir, si la reseña solo contiene 2 palabras, se eliminaron de nuestro análisis debido su contenido muy limitado para el análisis, además así quitamos las reseñas nulas marcadas como “no positive” y “no negative”.

Por último, se añadió un identificador único que nos permitirá identificar correctamente una reseña al realizar una unión o buscar una reseña en particular.

Conteo de palabras

En esta parte del contenido se obtiene una serie de instrumentos descriptivos que nos brindan información sobre el comportamiento de los datos, entre ellos se encuentra tabla de frecuencias para todas las palabras incluidas en el listado descrito anteriormente, gráfico de barras y un gráfico de nube de palabras. Para llevar a cabo este conteo utilizamos la función Counter() de la biblioteca collections, el resultado es un diccionario con la palabra y su frecuencia, misma que es procesada mediante un dataframe de la biblioteca pandas para ser ordenado por la frecuencia de la palabra. Las herramientas descritas en esta sección se encuentran en la parte de resultados.

Análisis de sentimientos

Para el apartado del análisis de sentimientos se utilizaron tres herramientas:

- Análisis de sentimientos basado en un lexicón: Este método utiliza un diccionario de palabras a la cual se le asigna un peso o valor, de modo que si un valor es negativo, está asociado a un sentimiento negativo y así mismo, si el resultado es positivo, es que la palabra está asociada a un sentimiento positivo, de modo que al final, se suman los valores para cada una de las palabras contenidas en la reseña y así obtener una conclusión final, es decir, estos resultados tienen tres posibilidades, “Positivo” si el resultado es mayor a cero, “Negativo” si el resultado es menor a cero y “Neutral” si es igual a cero.
- Análisis de sentimientos mediante TextBlob: El análisis mediante este método corresponde al uso de la biblioteca TextBlob, la cual utiliza diferentes métodos con la finalidad de obtener correctamente el sentimiento ligado a dichas palabras.
- Análisis de sentimientos utilizando VADER: Este análisis sigue el mismo principio que TextBlob, sin embargo, los umbrales utilizados son un poco diferentes, aquí tomamos 0.5 como umbral.

Comparación de modelos

Al final de este proceso se realizó una comparación de dichos modelos con diferentes métricas, las cuales son F1, Accuracy, Recall y Presicion.

RESULTADOS

Como primer resultado podemos ver un listado de las 30 palabras más frecuentes en todas las reseñas, puede verse que estas palabras corresponden justamente con el contexto de la base de datos.

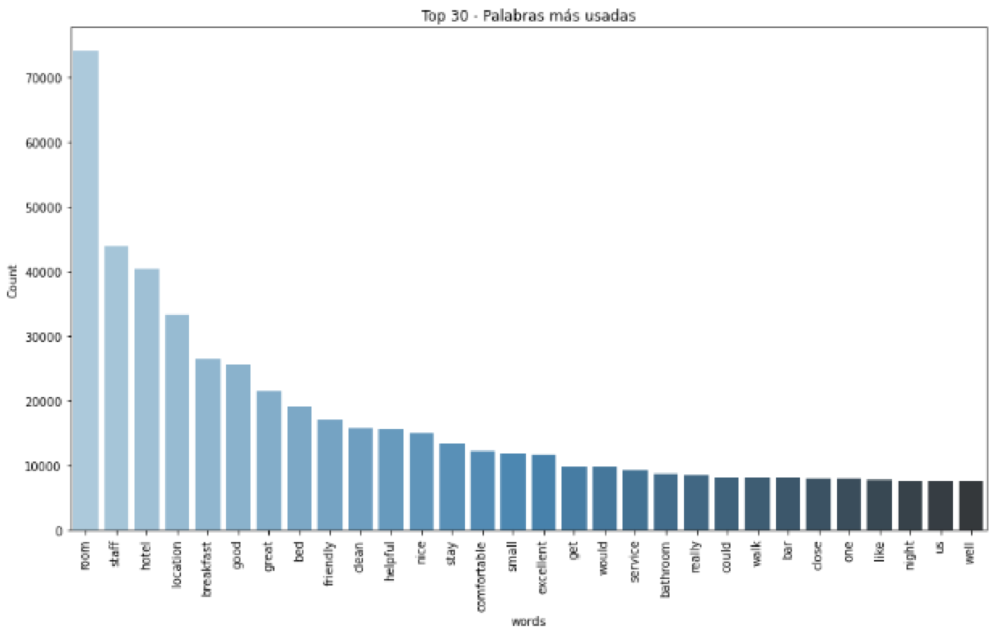


Ilustración 1.- Gráfico con las 30 palabras más utilizadas en todas las reseñas.

Nuble de palabras

Para mostrar la relevancia de resultados se muestra una nuble de palabras que contiene las 200 palabras principales.

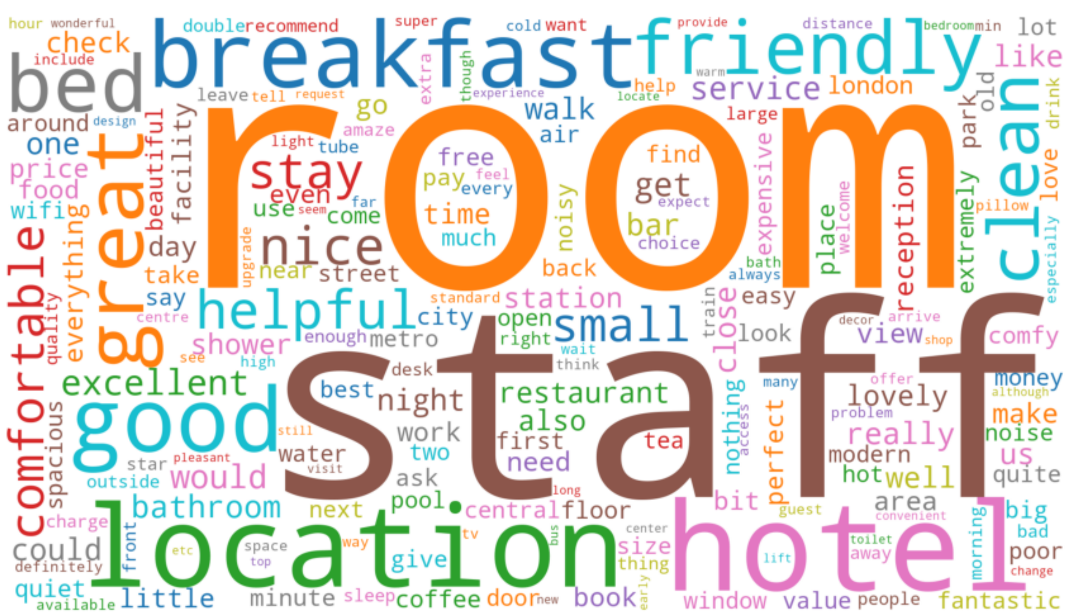


Ilustración 2.- Gráfico de nube que ilustra las 200 palabras más utilizadas.

Análisis de sentimientos - Modelos

La evaluación de los 4 modelos se llevó a cabo bajo el mismo preprocesamiento de texto y se decidió tomar las reseñas neutras como positivas para que ésta pueda ser comparada con las etiquetas originales y así evaluar el desempeño de éstos.

A continuación, se muestra una comparación de proporciones entre las reseñas positivas, negativas y neutras con respecto a las etiquetas originales.

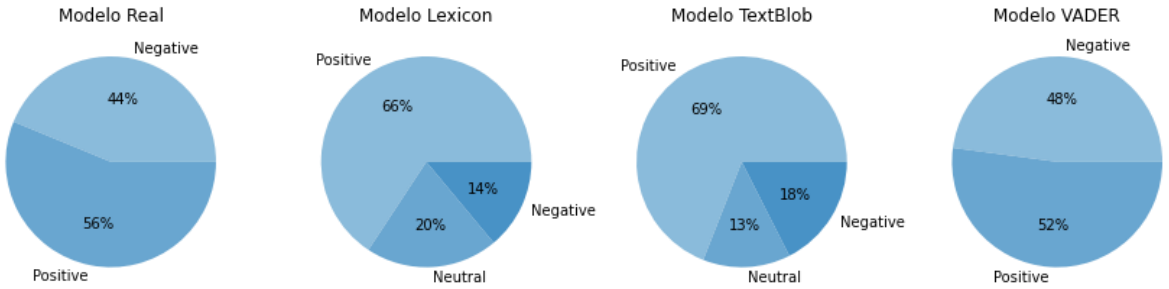


Ilustración 3.- Proporción de etiquetados para diferentes modelos.

El modelo VADER solo logró clasificar reseñas positivas y negativas, el resultado concuerda bastante con las etiquetas originales.

Métricas utilizadas

En la evaluación de desempeño de los modelos se decidió probar con cuatro métricas F1 score, Accuracy, Recall y Precision, esto con la finalidad de medir el rendimiento de cada uno de los modelos bajo diferentes propósitos. El resultado se resume en el siguiente gráfico de barras.

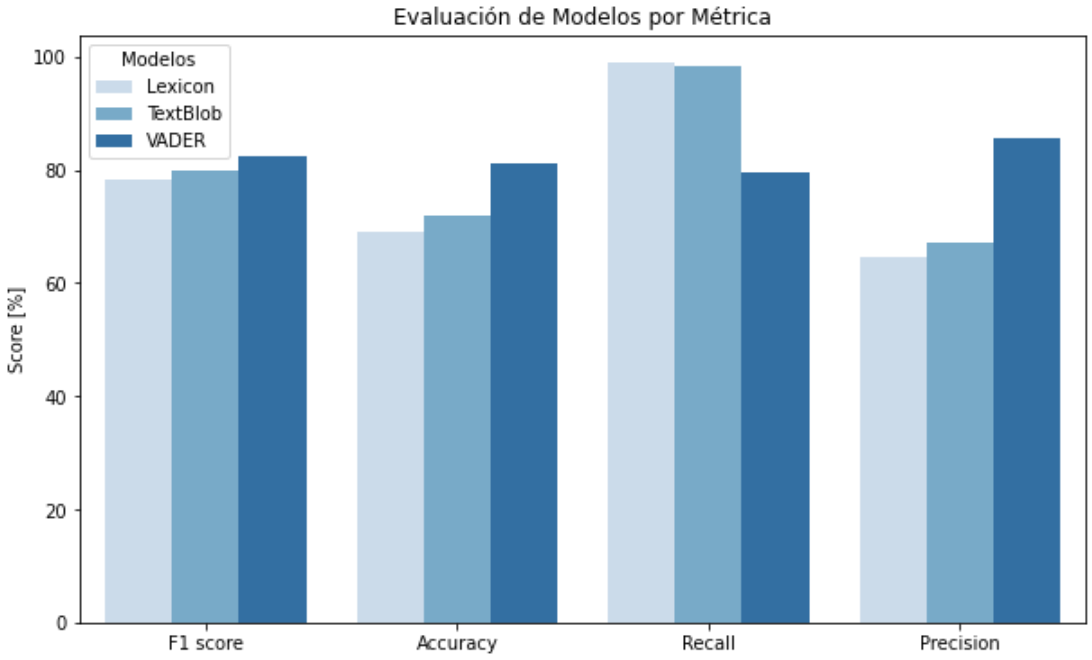


Ilustración 4.- Métricas de evaluación para cada uno de los modelos usados.

CONCLUSIÓN

El uso del análisis de sentimientos debe llevarse a cabo con cautela, debido a que hay factores que influyen en el rendimiento de éste, como lo son el procesamiento de los datos, el modelo a utilizar y sus características particulares como lo son el umbral de decisión además la métrica utilizada también es un factor importante sobre todo a la hora de evaluar el rendimiento o para la creación de un modelo de clasificación dado el contexto del problema.

En nuestro caso particular, si nos basamos en la métrica Accuracy, F1 score y Presicion, el modelo VADER es el que presenta un mejor rendimiento, mientras que con Recall el modelo Lexicon y TextBlob presentan mejor rendimiento que VADER. Para los propósitos de la base de datos, el clasificar correctamente tanto las reseñas como positivas son igual de importantes, de modo que la métrica que más se ajusta a este propósito es F1 score, de modo que el mejor modelo dado el propósito del análisis es VADER.

BIBLIOGRAFÍA

[1]. J. (2018, 18 diciembre). *Sentiment analysis with hotel reviews*. Kaggle. Recuperado 26 de mayo de 2022, de https://www.kaggle.com/code/jonathanoheix/sentiment-analysis-with-hotel-reviews/data

[2]. Go, A., Bhayani, R. and Huang, L., 2009. *Twitter sentiment classification using distant supervision*. *CS224N Project Report, Stanford, 1(2009), p.12*.

[3]. *Sentiment Analysis | Lexalytics*. (s. f.). Lexalytics. Recuperado 26 de mayo de 2022, de https://www.lexalytics.com/technology/sentiment-analysis

REPOSITORIO

Los archivos y códigos se encuentran en el repositorio cuyo enlace es:

https://github.com/juanagsolano/Procesamiento-y-clasificaci-n-de-datos