



UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

# Tarea 2
# Procesamiento y clasificación de datos

**Luis Enrique López Nerio**
**1487280**

PROFESOR

**Dra. MAYRA CRISTINA BERRONES REYES**

ASIGNATURA

**Procesamiento y clasificación de datos**

26 de mayo de 2022

# Índice

# 1. Introducción

Una problemática común dentro del procesamiento del lenguaje natural es el análisis de sentimientos, esta es una técnica que tiene como principal objetivo medir y tratar de entender los datos que se presentan en forma de texto, como su nombre lo indica, el entendimiento o clasificación que se busca dar al dato en forma de texto es el sentimiento, emoción o intención detrás del mismo.

Supongamos el siguiente ejemplo, se busca analizar la calidad del servicio de una oficina del SAT que se encarga facilitar trámites y servicios relacionados con la recaudación de impuestos. Se ha creado un buzón de quejas y comentarios en donde los usuarios que han realizado algún trámite o consulta pueden calificar los servicios que les brindaron y se tienen los siguientes tres comentarios:

- Usuario 1: *Tenia cita para tramitar mi e-firma funcionaria del departamento tuvo un trato muy grosero, me citaron a las 3 y me atendieron 1 hora después, pésimo servicio.*
- Usuario 2: *Tramité mi contraseña del buzón tributario..*
- Usuario 3: *Excelente servicio, vine a aclarar dudas respecto al régimen de personas físicas con actividad empresarial, la funcionaria fue muy clara al responder mis dudas.*

Por nuestras compresión y entendimiento de lenguaje, nosotros como humanos podemos darnos cuenta de que el comentario del usuario 1 es definitivamente negativo, ya sea por el tipo de palabras que usa o en general el contenido completo del mensaje, el usuario 2 simplemente hace un comentario del servicio que recibió y el usuario 3 tiene un comentario positivo.

De acuerdo a los comentarios el SAT puede realizar conclusiones sobre cuáles son los departamentos que padecen de una mala calidad del servicio y cuáles, en cambio, van por buen camino.

Sin embargo, es importante notar que este análisis se hizo sobre solo tres comentarios, a demás, nosotros como humanos tenemos un entendimiento del mensaje con base en nuestras experiencias, comprensión lectora y conocimiento del contexto en el que se da el mensaje.

El SAT no da servicio a tres personas, sino a cientos de miles, es aquí donde entra en juego el análisis de sentimientos, la idea es que la computadora reciba un comentario y sea capaz de discernir si el mensaje tiene un sentimiento de positividad, negatividad o si es neutral. En el presente reporte propongo un análisis de sentimientos sobre una base de datos que contiene reseñas a libros de ciencia de datos vendidos en el sitio Amazon y clasificarlos utilizando 3 diferentes técnicas dentro de la literatura.

# 2. Fuente de datos

Los datos sobre los que se hará el preprocesamiento son datos provenientes de la plataforma Kaggle, específicamente la base contiene reseñas a libros que se vendieron por la plataforma de amazon y que su temática principal es la ciencia de datos, la base cuenta con 20,647 filas o entradas y 3 columnas que se describen en la tabla 1

| Nombre de la columna | Descripción |
|---|---|
| stars | En esta columna se almacena la cantidad de estrellas que el usuario le dio al libro, este valor puede tomar desde 1 a 5 estrellas |
| comment | En esta columna se almacena la reseña o comentario realizado sobre el libro comprado. |
| book url | Esta columna almacena la url que el libro tiene en el sitio de amazon, se cuentan con hasta 836 libros diferentes. |

Tabla 1: Descripción de las variables presentes en el dataset

## 3. Análisis y preprocesamiento

El primer paso para limpiar nuestros datos es remover reseñas que cuentan con emoticons como reseña, estas reseñas se ignoran por el momento y se dejaran para un futuro análisis, al realizar este primer paso nos quedamos con 20,320 filas.

Analizaremos la distribución de la longitud de las reseñas sin pre procesamiento, para esto se creara una nueva columna que mida el número de palabras que cuenta cada comentario, en la figura 1 podemos ver que la mayoría de las reseñas se concentra entre $0-100$ palabras y hay muy pocas reseñas con una gran cantidad de palabras, esto es característico de una distribución con cola derecha, además la longitud de los comentarios en promedio es de 88 palabras aproximadamente.

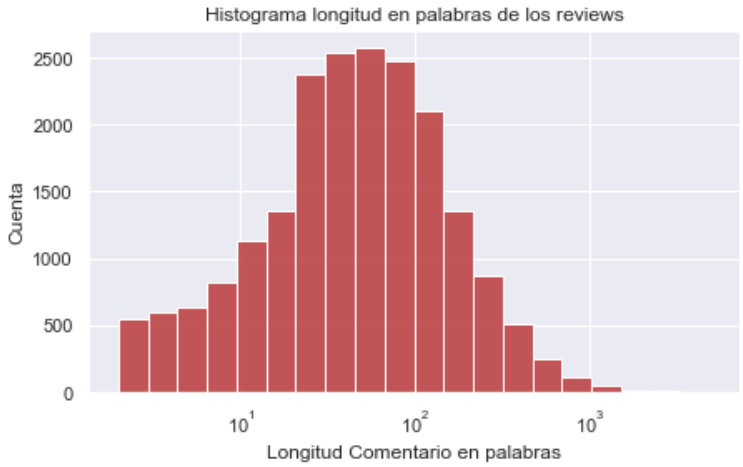



Figura 1: Histograma de la cantidad de palabras por reseña

Otra variable de interés es la columna **stars** la cual cuenta con la calificación dada al libro, esta columna toma valores de 1 a 5 y además estos valores son discretos, esto es, un libro no puede recibir una reseña de 3.5 estrellas sino solo valores enteros, observando la gráfica de barras de la figura 2 observamos que la mayoría de los libros reciben una calificación de 5 estrellas, la distribución de la calificación tiene una cola izquierda, el promedio de las reseñas es de 4.5.

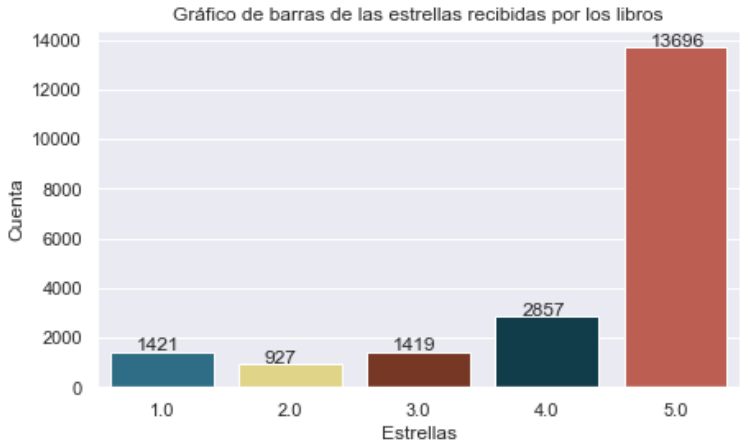



Figura 2: Cantidad de reseñas de n-estrellas

De acuerdo al número de estrellas recibidas se clasificará a la reseña como **positiva**, **negativa** o **neutral**, de igual manera, podemos observar la distribución del número de reseñas de cada tipo en la figura **??**, las reseñas positivas son las que tienen una mayor frecuencia con $16,553$.

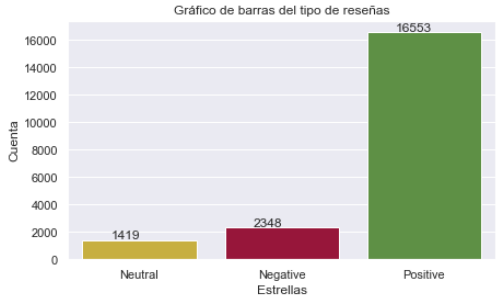


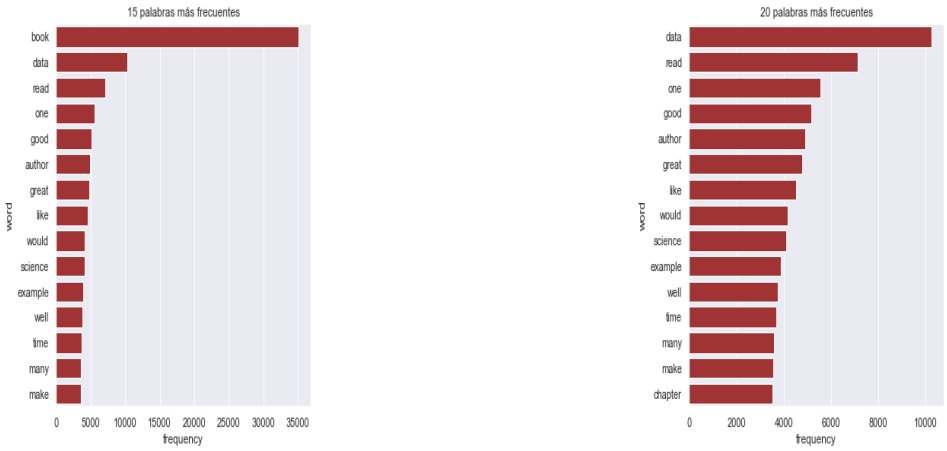

Figura 3: Cantidad de reseñas de n-estrellas

Con el análisis y las gráficas podemos darnos una idea general de nuestros datos y notar que contamos con un set de datos desbalanceados, esto es, contamos con muchas más reseñas positivas que negativas, es importante tomar esto en cuenta cuando se haga el análisis de sentimientos.

El siguiente paso es el preprocesamiento de nuestros datos, para esto utilizaremos la librería `nltk` los pasos a seguir en el preprocesamiento son:

1. Remover signos de puntuación.
2. Tokenizar las palabras.
3. Convertir letras mayúsculas a minúsculas
4. Remoción de stopwords
5. Lematización

En el preprocesamiento se encontró que la palabra *book* tenía una frecuencia mucho mayor que las otras palabras, dentro del contexto que es libros de ciencia de datos, esto es de esperar, sin embargo, está palabra no aporta mucha información al texto, es por eso que se decide agregar a los *stopwords*, en la gráfica 4 podemos observar las frecuencias de los términos una vez hecho el preprocesamiento, en la figura 4a no se eliminó la palabra *book* y en la gráfica 4b si.




(a) Frecuencia de las 15 palabras más frecuentes en todas las reseñas

(b) Frecuencia de las 15 palabras más frecuentes en todas las reseñas eliminando book

Figura 4: Frecuencias de las palabras en todas las reseñas

A continuación, se presenta la función en python que se encarga de realizar el preprocesamiento para una cadena de texto.

```
def preprocesamiento2(comentario):
    comentario =  re.sub('[^a-zA-Z]',' ',comentario)
    palabras = word_tokenize(comentario)
    palabras = [palabra.lower() for palabra in palabras ]
    palabras = [palabra for palabra in palabras if palabra.isalpha()]
    palabras = [palabra for palabra in palabras if not palabra in stop_words]
    palabras = [lemmatizer.lemmatize(palabra) for palabra in palabras]
    return palabras

```

# 4. Análisis de sentimientos

Se procede a realizar el análisis de sentimientos sobre nuestras reseñas preprocesadas, la idea es utilizar tres librerías diferentes para hacer el análisis, se utilizan la librerias:

- `textblob`
- `Vader`
- `Sentinet`

Cada una de estas librerías está especializada en el análisis de textos, se utilizaran para clasificar nuestro texto dentro de alguna de las tres categorías existentes, **Positivo, Negativo, Neutral**, el resultado de esta clasificación se comparara con el sentimiento R̈ealq̈ue se obtuvo con base en las estrellas otorgadas al libro en la reseña, por ejemplo, si en una reseña, el usuario otorga entre 4 y 5 estrellas al libro, se considerara que el sentimiento "real.$^{es}$ positivo, se espera que el análisis de sentimientos concluya que la reseña fue positiva, de otra manera, la clasificación fue errónea.

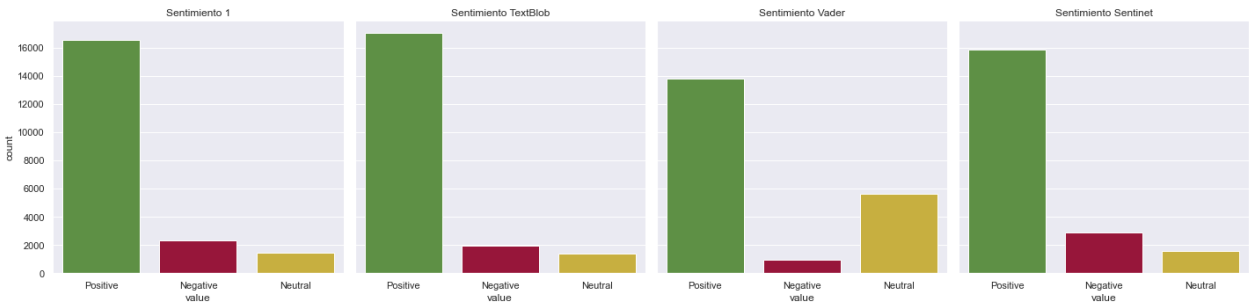



Figura 5: Cantidad de reseñas de n-estrellas por las diferentes librerías

Los resultados de clasificación de las tres librerías se presentan a continuación en la gráfica 5, observamos que el análisis de sentimiento realizado por la librería`textblob` es el que da como resultado una distribución de reseñas positivas-negativas-neutras más parecida a la del sentimiento "real", el análisis realizado por la librería vader arroja una mayor cantidad de reseñas clasificadas como neutras y la menor cantidad de positivas y por último, el análisis utilizando la librería sentinet clasifica como negativas un número mayor de reseñas que las otras librerías.

Para poder comparar de manera más puntual que tan bien clasificó una librería a un texto en específico es necesario comparar en cada texto cuál era su clasificación "realz cuál es la clasificación que le dio la librería, esto se puede realizar con una matriz de confusión, la matriz de confusión nos permite plasmar de manera gráfica de la cantidad de reseñas positivas-negativas-neutras cuantas fueron clasificadas dentro de cada clase, en la diagonal principal desearíamos ver el mayor número de reseñas ya que representan las reseñas que fueron clasificadas de manera correcta.

## 5. Discusión de resultados

Los resultados en la gráfica 6 nos muestran las matrices de confusión de las tres librerías, una vez que tenemos las librerías podemos obtener métricas para analizar el desempeño de nuestros análisis, por ejemplo, una métrica que se puede obtener de nuestras matrices de confusión es el accurary:

$$Accurary = \frac{\#\text{ de predicciones correctas}}{\#\text{ total de predicciones}}$$

Si ponemos atención en el accuracy obtenido con las librerías, nos damos cuenta de que la librería textblob obtuvo un mayor accuracy con 87 %, sin embargo, el accuracy es una métrica que se ve afectada cuando nuestros datos están desbalanceados, esto es, que una clase se presenta con mayor frecuencia que la otras, lo ideal es analizar otras métricas en conjunto con el accuracy para medir el desempeño de nuestra clasificación.

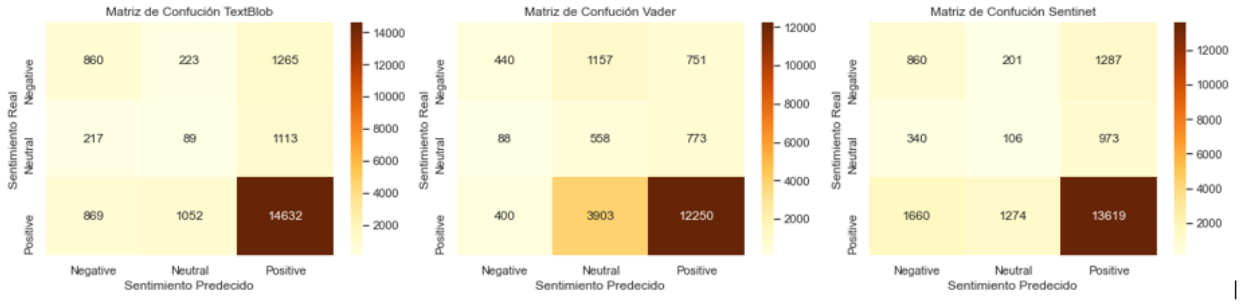



Figura 6: Matrices de confusión

## 6. Conclusión

Analizando las matrices de confusión en conjunto con diferentes métricas se concluye que la librería `textblob` obtuvo un mejor desempeño ya que presento un accuracy de 87 %, además de que, en el contexto de problema, una clase que es de suma importancia es la de reseñas negativas ya que es la que nos puede ofrecer un mayor entendimiento de cuales libros tienen un peor recibimiento entre los clientes. En vista de lo anterior, una métrica de importancia es el recall de la clase negativa, esta métrica nos indica del total de reseñas que eran negativas, el porcentaje que fueron clasificadas correctamente, la libreria textblob fue la que tuvo un mejor desempeño en esta categoría con un recall de la clase negativa de 36 %, lo cual es bajo, sin embargo, fue el mejor comparado con las otras dos librerias.

## 7. Referencias


- Repositorio de Github
- Origen de los datos en la plataforma de Kaggle