UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

Tarea 3
Procesamiento y clasificación de datos

Luis Enrique López Nerio
1487280

PROFESOR

Dra. MAYRA CRISTINA BERRONES REYES

ASIGNATURA

Procesamiento y clasificación de datos

31 de mayo de 2022

Índice

1. Introducción

En la actualidad la capacidad que se tiene para recopilar datos e información ha crecido exponencialmente, los datos que se generan a partir de una gran variedad de fuentes, por esta razon se pueden clasificar de muchas maneras, una forma de clasificar estos datos es como:

- Datos estructurados
  Este tipo de datos por lo general es almacenado en bases de datos relacionales,los campos de información estan bien definidos y delimitados como por ejemplo, números de telefono, números de seguro social, etc. Este tipo de datos es de facil acceso y explotación.

- Datos no estructurados
  Los datos no estructurados son esencialmente todo lo demas que no se puede catalogar como dato estructurado como por ejemplo:

  - Archivos de texto
  - Emails
  - Imagenes
  - Audio

Los datos no estructurados forman parte de nuestra vida diaria, dentro de esta categoria caen los archivos de texto, este tipo de archivos puede ser una fuente de información importante y de la cual se puede extraer valor, sin emabrgo, su naturaleza no estructurada puede hacer dificil esta labor, sin embargo, en la actualidad existen muchas maneras de extraer valor de los datos en forma de texto que van desde análisis descriptivo hasta modelos de de aprendizaje de maquina.

Dentro del campo del aprendizaje de maquina podemos identificar dos grandes areas, el aprendizaje supervisado y el no supervisado, en este reporte se propone el uso de modelos de aprendizaje de maquina supervisado para la clasificación de texto.

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 20,636 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.

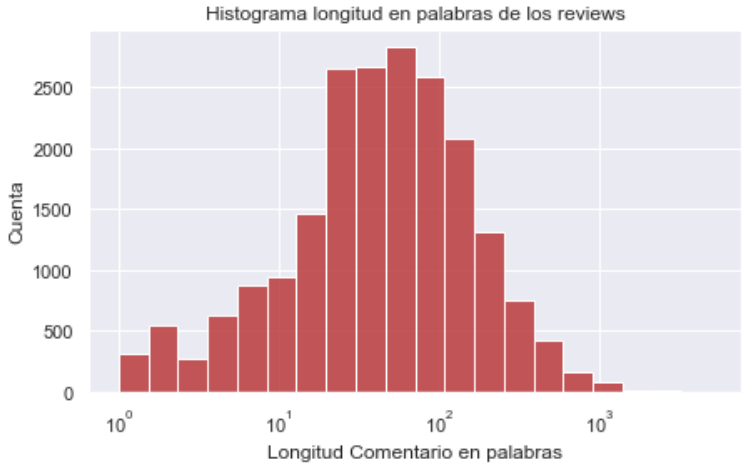


Figura 1: Histograma de la cantidad de palabras por reseña

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 4.3.

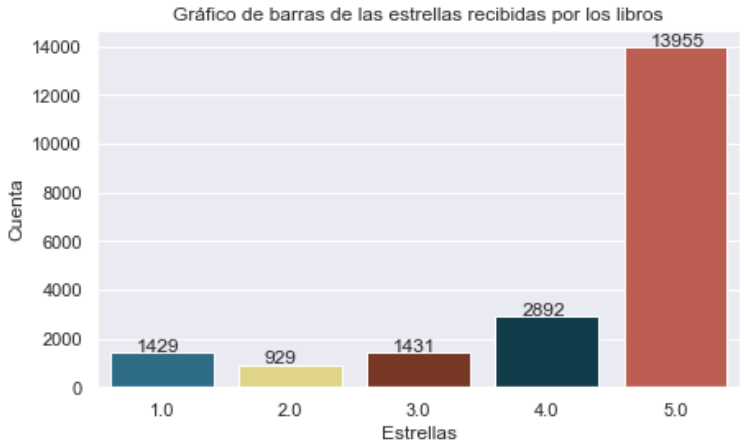


Figura 2: Cantidad de reseñas de n-estrellas

De acuerdo al número de estrellas recibidas se clasificará a la reseña como **positiva** o **negativa**, si una reseña tiene una calificación de 4-5 estrellas se calificara como una reseña positiva y de otra manera se clasificara como una reseña negativa, de igual manera, podemos observar la distribución del número de reseñas de cada tipo en la figura 3, las reseñas positivas son las que tienen una mayor frecuencia con $16,847$.

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.

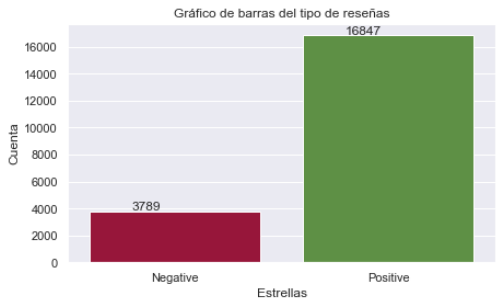


Figura 3: Cantidad de reseñas de n-estrellas

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

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 con modelos de aprendizaje de maquina

Se procede a realizar el análisis de sentimientos utilizando tres modelos diferentes:

- Regresión logistica
- Naive-Bayes
- Random Forest

Nuestro siguiente paso es desarrollar un modelo dado una reseña en forma de texto, sea capaz de clasificar la reseña como positiva o negativa, de nuestra base solo se utilizara la etiqueta, (i.e. reseña positiva o negativa), y el texto de la reseña.

Antes de entrenar nuestros modelos es necesario convertir el mensaje de texto en un vector,para la computadora es dificil trabajar con datos en forma de texto , como los modelos utilizan información númerica es necesario convertir el texto en un vector número, un metodo de vectorización de texto es el de bolsa de palabras, este metodo basicamente toma todos los mensajes de texto, indentifica todas las palabras diferentes unicas(bolsa de palabras) y transforma a cada texto en una representación de la frecuencia(u otra metrica) de está bolsa de palabras.

El segundo paso es separar nuestros datos en entrenamiento y test, los datos de entrenamiento serviran para entrenar a nuestro modelo y los datos de test serviran para evaluar el desempeño de nuestro modelo sobre datos a los que no ha tenido acceso y asi evitar un sobre ajuste de los parametros del modelo.

A continuación se muestran los resultados obtenidos en cada uno de los modelos, se utilizaran las siguientes metricas-visualizaciones para evaluar el desempeño de los modelos.

- Matriz de confusión.
  La matriz de confusión es una forma de presentar los resultados del desempeño de un modelo de clasificación en forma de tabla o de matriz, las filas representan instancias de la clase real mientras que las columnas representan instancias de la clase predecida.

- Accurary
  Es una metrica que sirve para evaluar el desempeño de nuestro clasificador, es la fracción de clasificaciones que nuestro modelo predijo de manera correcta.

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

- Precission
  Precission es una metrica que al igual que las otras evalua el desempeño de un clasificador en una clase en particular y busca contestar la siguiente pregunta: ¿Cuántas clasficaciones de la clase "x"fueron realmente correctas?Se calcula con la formula:

$$Precission = \frac{\#\text{ de predicciones "x"que fueron correctas}}{\#\text{ de predicciones de "x"totales}}$$

- Recall
  El recall es una metrica de desempeño que busca contestar la siguiente pregunta: ¿ Cuál es la proporcion de datos que eran de la clase "xz que fueron clasificados correctamente ?, la formula para calcular esta metrica es:

$$Recall = \frac{\#\text{ de predicciones de "x"que fueron correctas}}{\#\text{ de datos que son en realidad de la clase "x".}}$$

- f1-score
  El f1-score es una metrica que combina el precission y recall y evalua de manera más general el desempeño del modelo. Se calcula de la siguiente manera:

$$F1 - score = 2\frac{precission * recall}{precission + recall}$$

- ROC curve y AUC score
  La curva ROC(receiver operating characteristic curve) es una gráfica que muestra el desempeño de un modelo de clasificación a diferentes umbrales de clasificación. Esta curva grafica dos parametros:

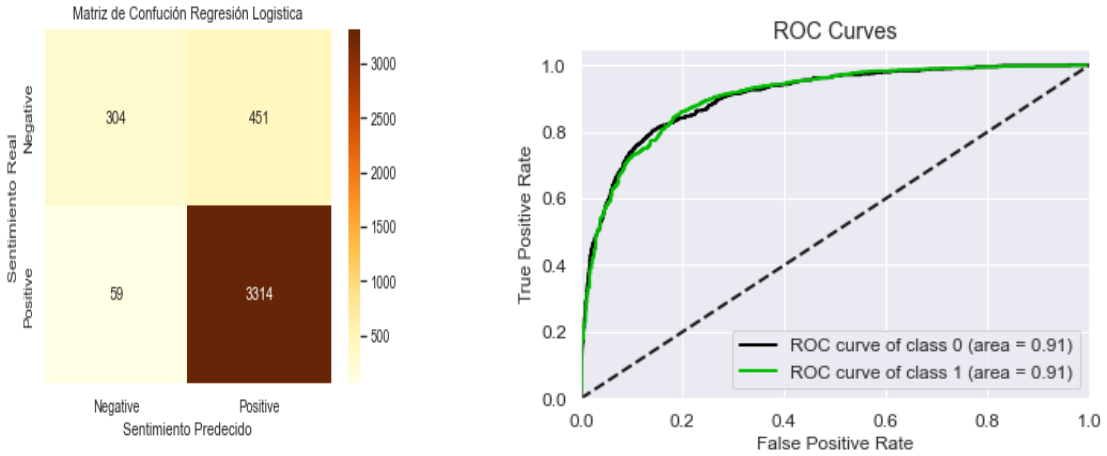
  - Tasa de verdaderos positivos. $TPR = \frac{VP}{VP+FN}$
  - Tasa de falsos positivos $TFP = \frac{FP}{FP+VN}$

  Una curva roc grafica la tasa de verdaderos positivos contra la tasa de falsos positivos a diferentes umbrales de decisión, bajar el umbral de clasificación significa clasificar mas observaciones como "positivas.[a] cambio de incrementar la tasa de falsos positivos, incrementando los falsos positivos y los verdaderos positivos en el modelo.

  La métrica AUC, por sus siglas en inglés "Area under the ROC curve"(Area bajo la curva ROC) mide el area que se forma bajo la curva de la curva ROC, tomando valores entre 0 y 1, nos provee de una medida de desempeño para todos los posibles umbrales de decisión, un modelo con un valor de $AUC = 0$ significaria que hace todas sus clasificaciones de manera incorrecta, un $AUC = 1$ seria un modelo que es perfecto en sus predicciones y un $AUC = .5$ seria un modelo que practicamente es igual a lanzar una moneda y decidir en base a esto.

4.1. Regresión logistica

El modelo de regresión logistica es un modelo de clasificación clasico, es ampliamente utilizado y aunque no es un modelo tan sofisticado como otros modelos aun es ampliamente utilizado y se puede tomar como benchmark para comparar el desempeño de un modelo nuevo, a continuación se presentan los resultados obtenidos en el modelo.



(a) Matriz de confusión modelo regresión logistica

(b) Curva ROC modelo regresión logistica

Figura 4: Modelo Regresión Logistica

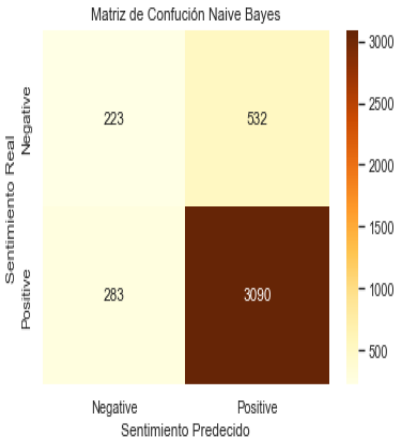
En la figura 4a podemos ver que se obtuvo una buena clasificación de la clase positiva y hay algunos problemas con la clase negativa, en la curva roc 4b observamos que se tiene un buen desempeño para distinctos umbrales de clasificación.

$$\begin{aligned} \textbf{Accuracy} &= 87.6\,\% \\ \textbf{F1-Score promedio ponderado} &= 86\,\% \\ \textbf{AUC score} &= 69.3 \end{aligned} \tag{1}$$

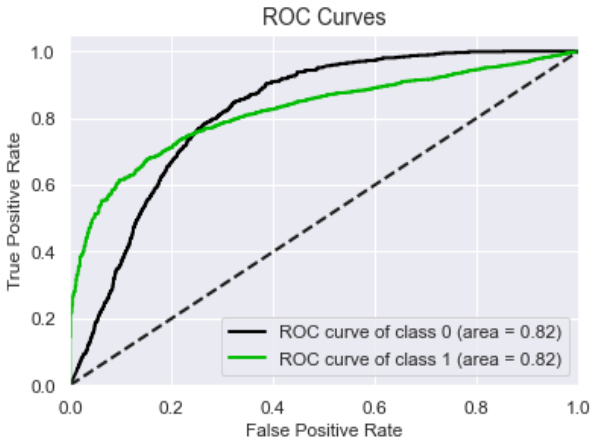
Clase	precision	recall	f1-score	support
Negativa	0.84	0.40	0.54	755
Positiva	0.88	0.98	0.93	3373

Tabla 2: Metricas modelo regresión logistica

4.2. Naive-Bayes



(a) Matriz de confusión modelo Naive-Bayes



(b) Curva ROC modelo Naive-Bayes

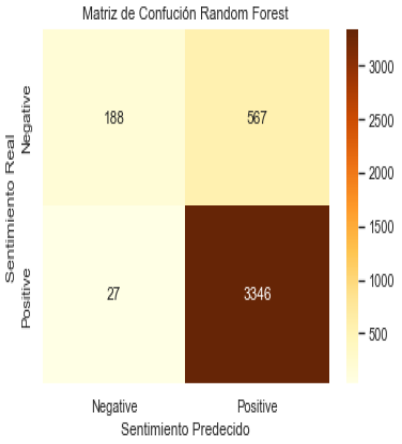
Figura 5: Modelo Naive-Bayes

$$\begin{aligned} \textbf{Accuracy} &= 80.3\,\% \\ \textbf{F1-Score promedio ponderado} &= 79\,\% \end{aligned} \tag{2}$$

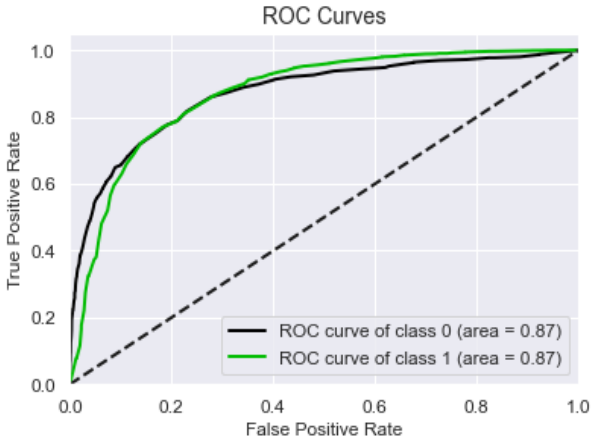
Clase	precision	recall	f1-score	support
Negativa	0.44	0.30	0.35	755
Positiva	0.85	0.92	0.88	3373

Tabla 3: Metricas modelo Naive Bayes

4.3. Random Forest



(a) Matriz de confusión modelo random forest



(b) Curva ROC modelo Random Forest

Figura 6: Modelo Ranfom Forest

$$\begin{aligned} \textbf{Accuracy} &= 85.6\,\% \\ \textbf{F1-Score promedio ponderado} &= 79\,\% \end{aligned} \tag{3}$$

Clase	precision	recall	f1-score	support
Negativa	0.87	0.25	0.39	755
Positiva	0.86	0.99	0.92	3373

Tabla 4: Metricas modelo random forest

5. Discusión de resultados

En general podemos observar que el modelo de regresión logistica obtuvo un mejor desempeño general al contar con un F1 score más alto, está metrica es la más adecuada para evaluar el desempeño debido al desbalance que se presenta en las clases, ademas de que el AUC-score tambien fue mejor que el de los otros dos modelos.

6. Conclusión

Se toma el modelo de regresión logistica como el mejor modelo para clasificar entre reseñas positivas o negativas debido al mejor desempeño en todas la metricas.

7. Referencias

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