Proyecto M7 - Clasificación de Sentimientos

Camila Acosta

Introducción

Problema:

- La empresa necesita entender la recepción de las aplicaciones e identificar fallos.
- Hacerlo manualmente es poco práctico.

Objetivo:

- Crear un modelo capaz de clasificar los reviews según el tipo de sentimiento (positivo/negativo/neutral) automáticamente.

Presentación

Datos:

- Kaggle: Google Play Store Apps for analysing the Android market.
- 50.000+ reviews clasificadas como positiva, negativa o neutra.

Limpieza y preparación de los datos

Nuestro dataset contiene un total de 5 columnas. De estas, eliminamos 3 columnas dejando solamente Translated_Review y Sentiment, para poder enfocarnos completamente en el objetivo de este proyecto.

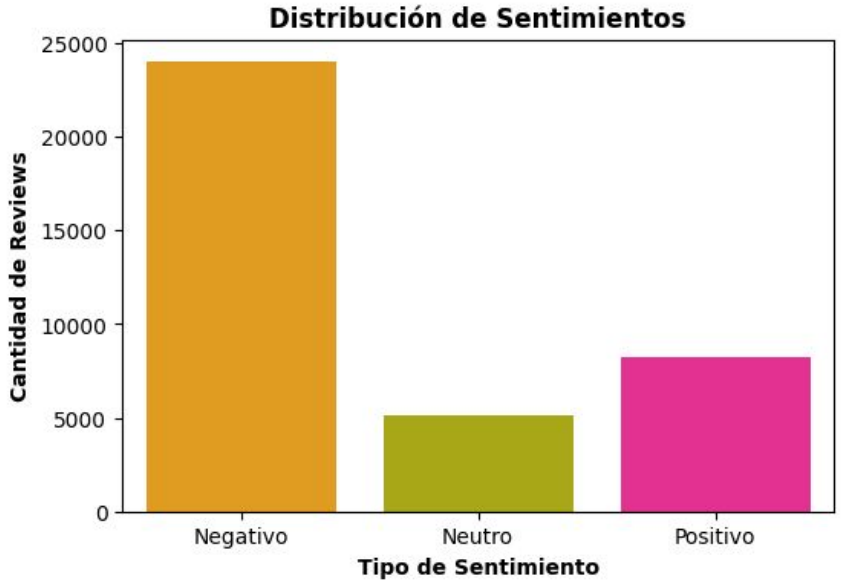
Las columnas eliminadas no nos ofrecen información relevante.

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000

Limpieza y preparación de los datos

- Se eliminaron todas las filas con valores nulos, pasando de tener 64.295 filas a 37.427

- Se creó un gráfico de barras, mostrando la distribución de sentimientos según el tipo de sentimiento y la cantidad de reviews correspondientes.



Limpieza y preparación de los datos

Para la limpieza del texto:

- Eliminación de stopwords (una lista específica de palabras que no contemplan relevancia para que el modelo aprenda).
- Se convirtió todo el texto a minúsculas.
- Eliminación de URLs, números y caracteres especiales (como signos de puntuación.

- Se codificó la columna de sentimiento para convertir estos en números.
- Negativo = 0, Neutro = 1, Positivo = 2.

	tokens	Sentiment_encoded
like eat delicious food cooking food case best...		2
help eating healthy exercise regular basis		2
works great especially going grocery store		2
best idea us		2
best way		2

Limpieza y preparación de los datos

En este punto, hacemos un split de los datos para tener nuestro conjunto de entrenamiento (70% de los datos) y de prueba (20%), lo que permitirá que el algoritmo (modelo de clasificación) pueda entrenarse y luego hacer una comparación con los datos de prueba.

Por último, vectorizamos el texto ya limpio, para así asignar un número a cada palabra y el modelo sea capaz de interpretar y aprender de los datos que le estamos proporcionando.

Vectorizador utilizado:

- TF-IDF

Modelo de Clasificación

Algoritmos:

- Logistic Regression
- Naive Bayes

Estos algoritmos fueron elegidos ya que tienen un buen desempeño y personalmente han dado los mejores resultados.

Luego de probar cada uno, elegimos definitivamente a Logistic Regression.

Accuracy: 0.9118

Logistic Regression

Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.91	0.80	0.85	1653
1	0.87	0.84	0.85	1049
2	0.92	0.97	0.94	4784
accuracy			0.91	7486
macro avg	0.90	0.87	0.89	7486
weighted avg	0.91	0.91	0.91	7486

Accuracy: 0.7384

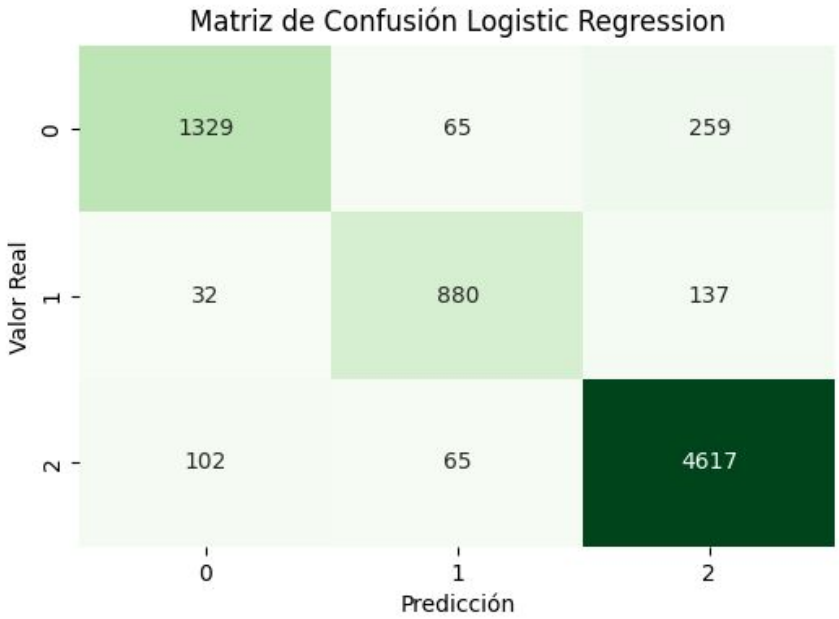
Naive Bayes

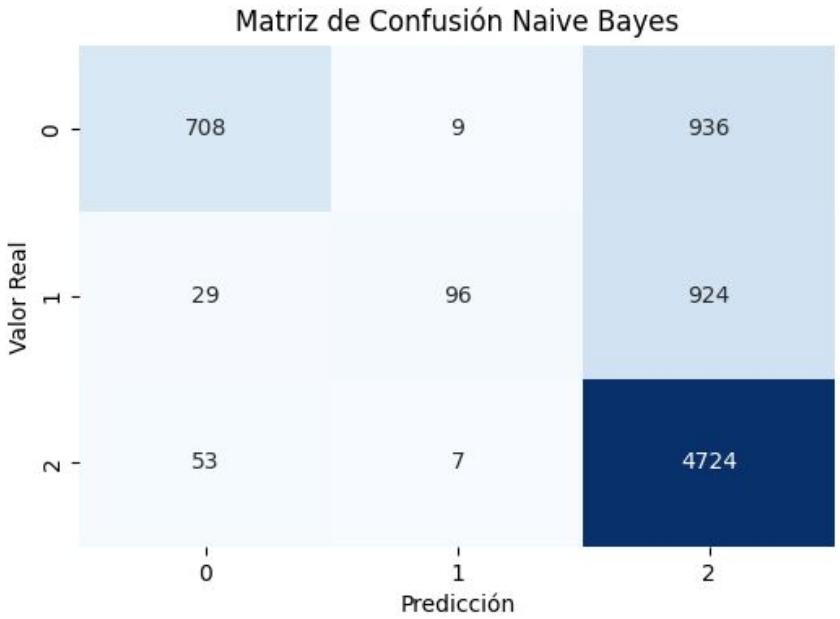
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.90	0.43	0.58	1653
1	0.86	0.09	0.17	1049
2	0.72	0.99	0.83	4784
accuracy			0.74	7486
macro avg	0.82	0.50	0.53	7486
weighted avg	0.78	0.74	0.68	7486

Logistic Regression tiene un accuracy de 0.91 sobre naive bayes que tiene un accuracy de 0.74.

Modelo de Clasificación





Modelo de Clasificación

Puntos importantes de los resultados y observaciones:

- Accuracy: 91% - en general, es un buen resultado pero podría mejorar.

- F1 - Score según sentimiento:
  Negativo: 85%
  Neutro: 85%
  Positivo: 94%
- Se podría ajustar el modelo para obtener mejor resultados en la predicción de los sentimientos negativos y neutros.

Implementación de la API

¿Qué es?
Es una URL (endpoint) que recibe datos, en este vaso texto (reviews), y devuelve una respuesta (para este proyecto, devolvería el resultado de nuestro modelo de clasificación).

El -endpoint- se creó usando **Flask**.

Para exponerlo al público, utilizamos **ngrok**, una herramienta que crea un túnel seguro HTTPS desde internet a nuestro servidor local.

Con esta API puedes usar el modelo de clasificación de forma remota y con nuevos reviews.

Uso de la API

Para probar el URL usamos Talend API Tester, donde solo ingresas el URL, colocas un texto (formato JSON) y este te devolverá una clasificación de sentimiento.