ANÁLISIS DE SENTIMIENTOS

AMAZON DATASET

Gabriel Almeida y Diana Almeida

INDICE

CODER

amazon

Objetivo

Crear un modelo de Machine Learning capaz de predecir la emoción del usuario de compra de productos musicales en Amazon, por medio del comentario (review) que escribe en la plataforma.

Preguntas

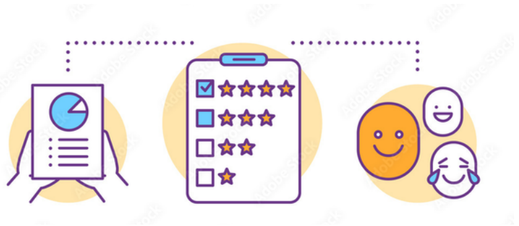
1. ¿Cuáles son las variables que determinan la categoría emocional en el modelo?
2. ¿Cómo se pueden distinguir emociones a partir de un texto utilizando un modelo de machine learning?

Data Acquisition

El dataset cuenta con 1.584.082 de registros

Data Acquisition

El dataset cuenta con 1.584.082 de registros, los campos seleccionados para realizar el análisis y por tanto el modelo fueron los siguientes:



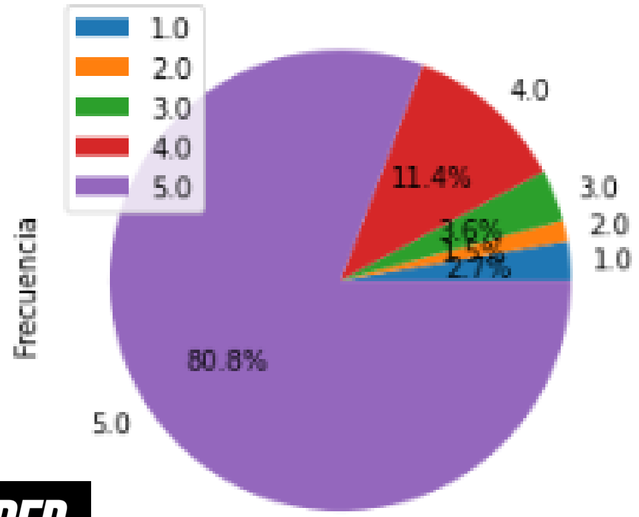
1. **Overall: Puntuación de 1 a 5 que brinda el usuario según su satisfacción con el producto, siendo 1 el más bajo y 5 el puntaje más alto.**
2. **Verified: Campo true/false que indica si la cuenta que realiza el comentario es o no verificada**
3. **review_time: Fecha en la que se generó el comentario y puntaje del producto.**
4. **reviewer_id: Documento hasheado del cliente.**
5. **review_text: Comentario completo sobre el producto adquirido.**
6. **summary: Comentario resumido sobre el producto adquirido.**
7. **style.format: Formato del contenido musical comprado.**
8. **vote: Cantidad de votos sobre el comentario realizado.**
9. **sentiment: Campo creado según el overall, donde se determina 1(sentimiento positivo) si es mayor a 3 el puntaje , y 0 (sentimiento negativo) si es menor o igual a 3.**
10. **review_length: Campo creado en donde se cuenta el número de caracteres del comentario a fin de determinar si existe alguna relación desde la longitud del mismo con el sentimiento que genera en el usuario.**

Enlace dataset Digital Music Amazon modificado:
https://drive.google.com/file/d/1tNfteg9zZvlcwmOQFDPkxMrBmWtnwN5A/view?usp=sharing

Análisis Univariado

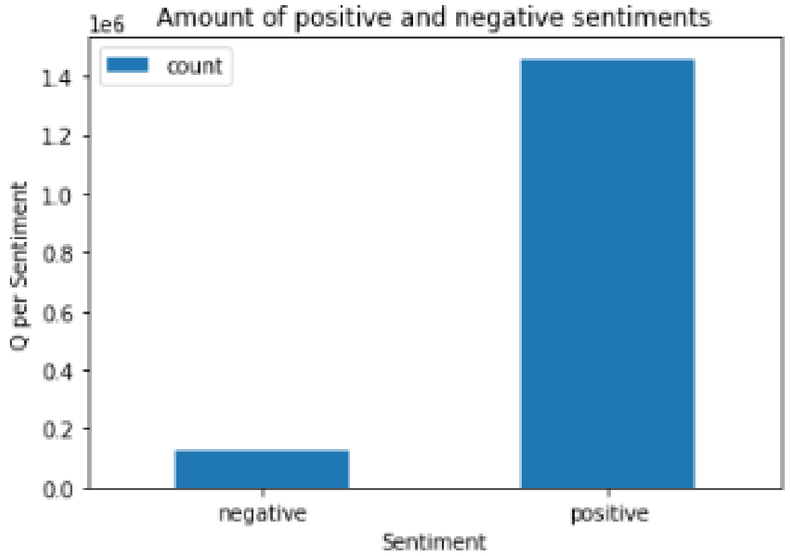
Pie Chart porcentajes overall

80% de los productos de música digital en Amazon tienen puntuaciones de 5.0 estrellas. El 20% restante está dividido en las puntuaciones inferiores; de manera descendente a medida que baja la puntuación: 4.0 (11.4%). 3.0 (3.6%), 2.0 (1.5%), 1.0 (2.7%).



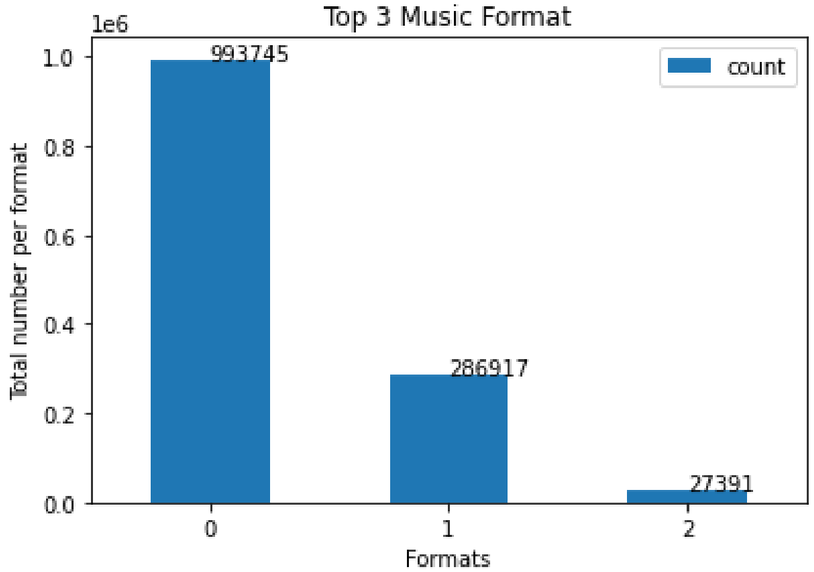
Cantidad de sentimientos negativos y positivos

En cuanto a la distribución de la valencia (positiva o negativa) de los comentarios, se muestra que la mayoría de los comentarios son positivos, solo 124.006 de ellos son negativos.



Ranking formatos musicales más vendidos

Finalmente, el formato de música con más reseñas es Audio CD, seguido del MP3 y finalmente Vinilo. Los otros 6 formatos tienden a no ser representativos ni comparables con la cantidad de comentarios que presentan los 3 primeros.



Análisis Bivariado

`overall` is highly correlated with `sentiment`

`sentiment` is highly correlated with `overall`

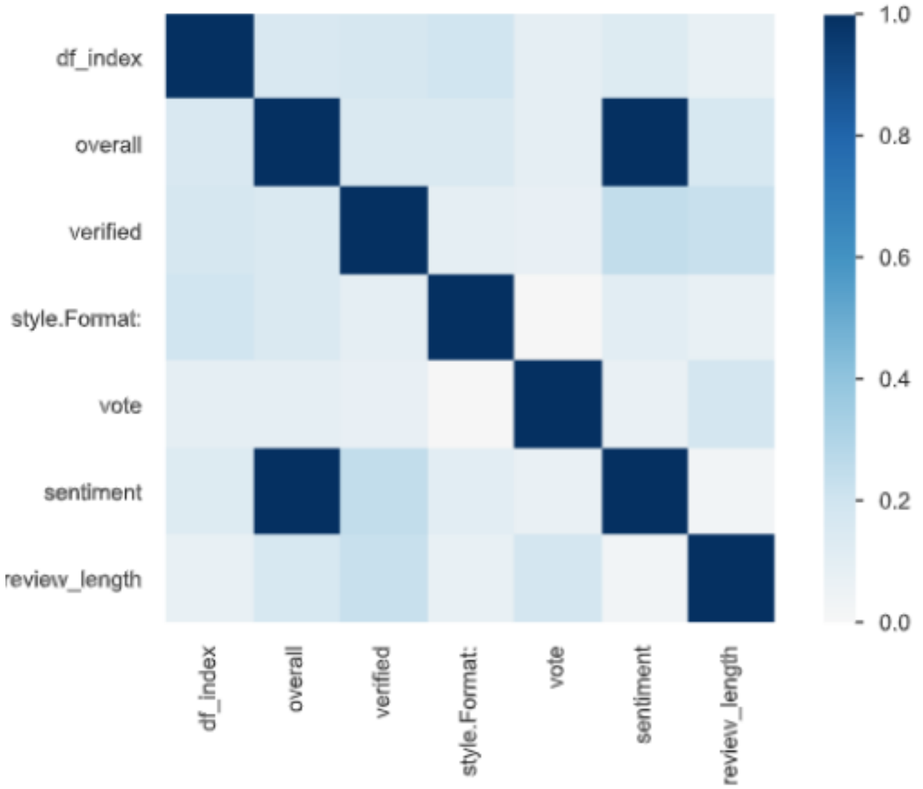
`overall` is highly correlated with `sentiment`

`sentiment` is highly correlated with `overall`

`reviewerID` is uniformly distributed

`reviewText` is uniformly distributed

`df_index` has unique values




```
sentiment
0    552.773364
1    562.277436
Name: review_length, dtype: float64
```

1. No existe evidencia suficiente para afirmar que hay independencia entre el formato y el sentimiento

2. Si existe una dependencia entre el formato y el sentimiento, siendo el top 3 : CD.MP3 y Vinyl.

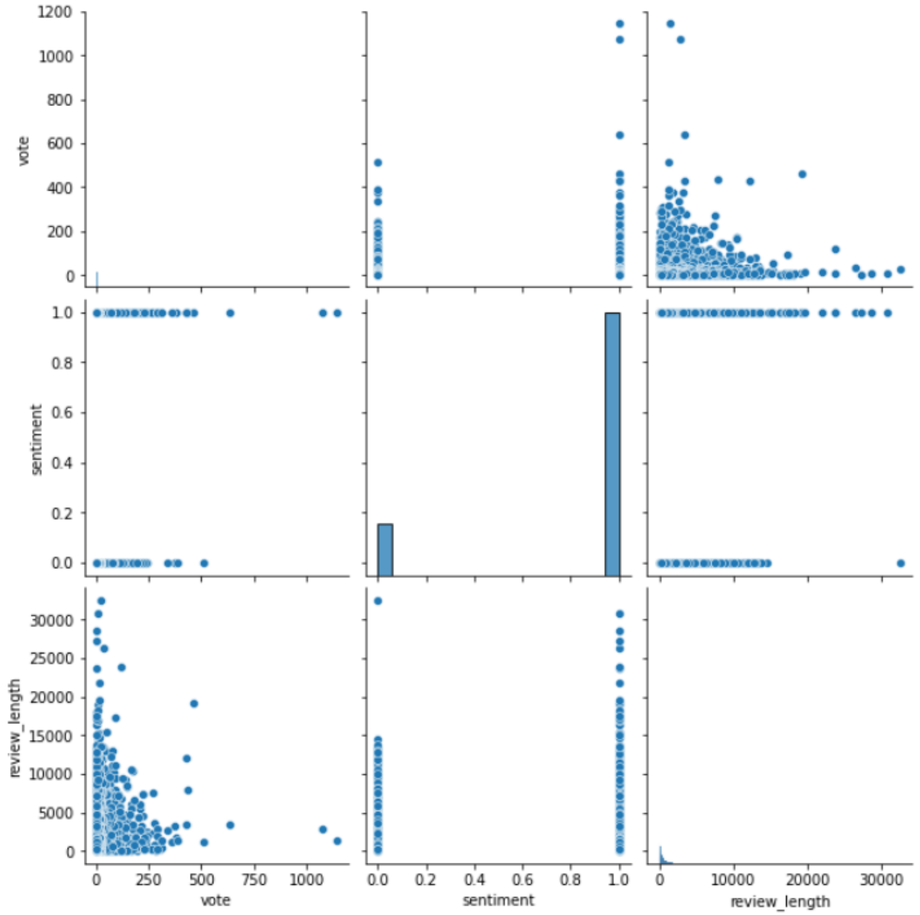
3. Si existe una dependencia entre el comentario verificado y el sentimiento

Análisis Multivariado

```
from sklearn import metrics
#Accuracy
print("Accuracy:",metrics.accuracy_score(y_test, y_pred)) # Se tiene un accuracy del 83%
```

✓ 0.1s

```
Accuracy: 0.8337638701344449
```



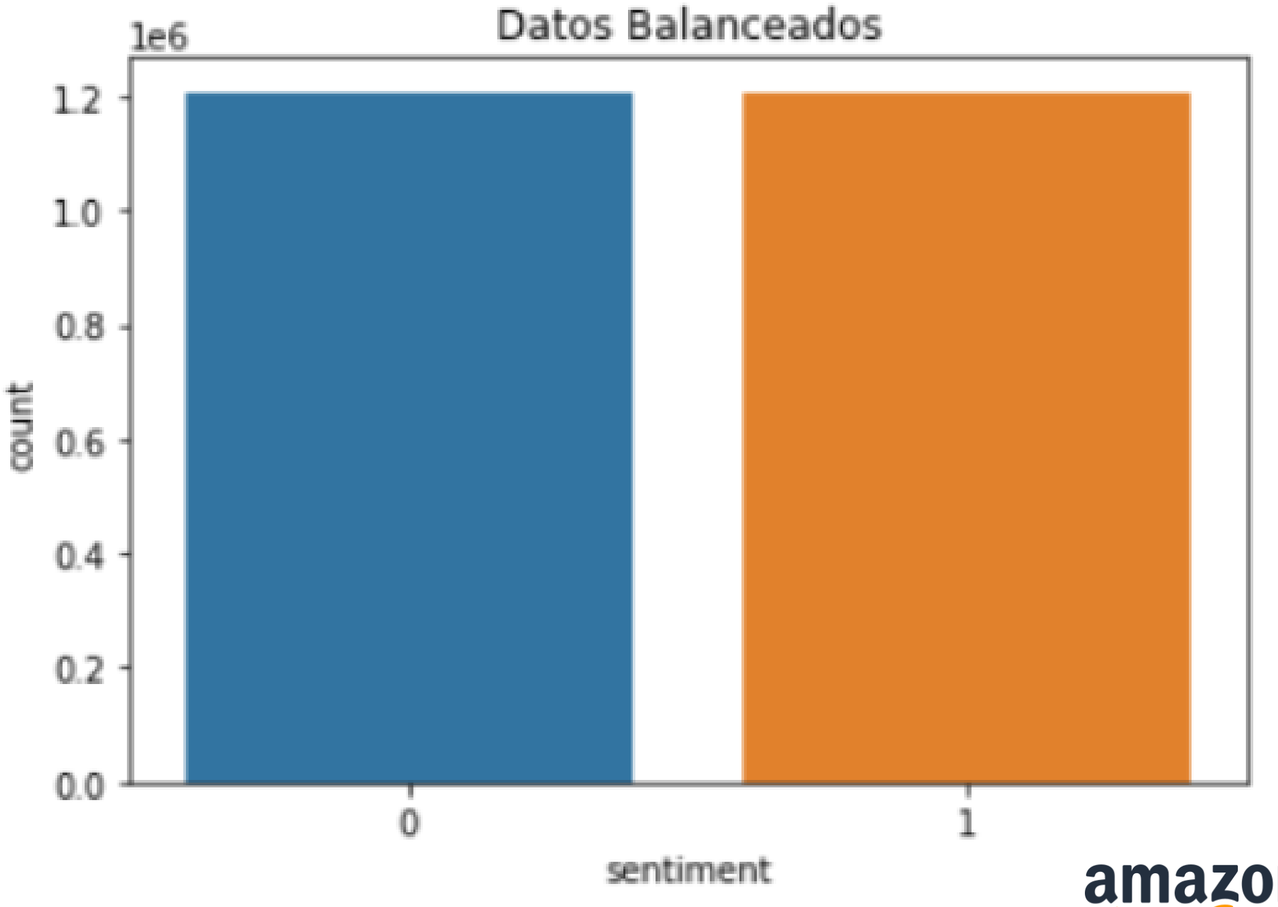
Se predice que el modelo puede identificar con un accuracy del 83% si un comentario es positivo o negativo

Si se hace un PairPlot se encuentra que entre más carácteres menos votos recibe el producto comprado. Las otras dos gráficas no muestran una correlación significativa para ser analizada.

BALANCEO: Optimizaciones del modelo SVM

```
Cambiode X antes de SMOTE: (1306426, 3)
Cambio de X despues SMOTE (2413292, 3)

Balance position and negativo de la clase (%):
0    50.0
1    50.0
Name: sentiment, dtype: float64
```

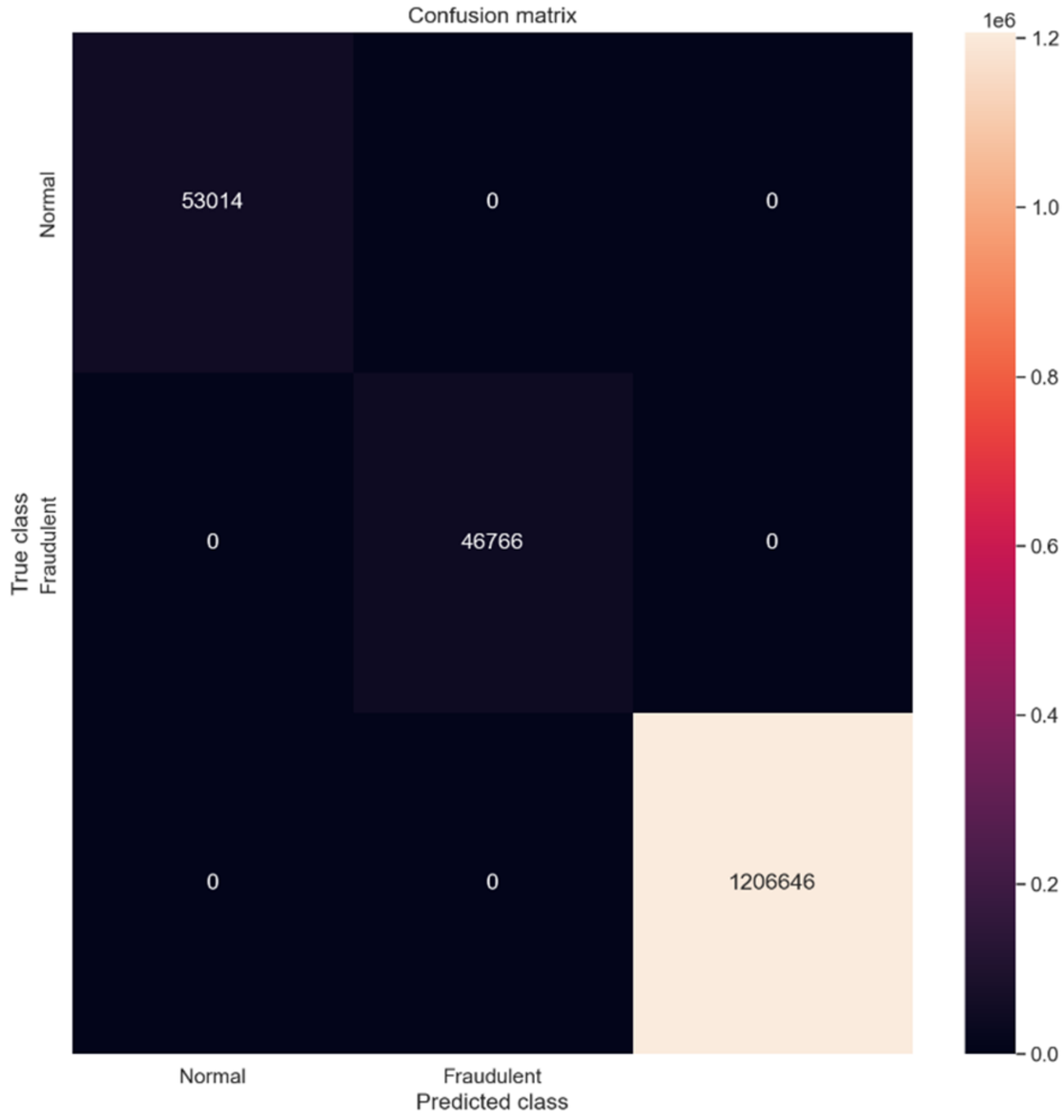


CODER

amazon

BALANCEO: Optimizaciones del modelo BERT

```
Distribution before resampling Counter({2: 1206646, 0: 53014, 1: 46766})
Distribution after resampling Counter({0: 46766, 1: 46766, 2: 46766})
```



CODER

amazon

Metricas finales del modelo optimizado

Capacidad para correr el modelo SVM

En el caso del SVM, luego de la configuración con el SMOTE, el programa no permitió correr y generar las métricas finales del modelo, python indica que está realizando el código que permite entrenar el modelo pero no termina de ejecutarse.

Modelo BERT

En el caso del modelo BERT ocurre algo similar y es que si bien todo está codificado y aparenta no tener ningún error, la capacidad de nuestros computadores no permite que el modelo corra y termine de ejecutarse correctamente, sino que se queda procesando.

CODER

amazon

Futuras líneas y Conclusiones

No existe ninguna relación significativa entre la cantidad de caracteres del comentario y el sentimiento generado.

Capacidad de procesamiento para los modelos

Ambos modelos si bien aparentan tener un buen accuracy y estar desarrollados correctamente, no pudieron terminar de procesarse debido a la gran cantidad de información que se estaba manejando.

Observaciones

1. Por la defnición que se le dio al campo sentiment, se observa que es el Overall con la que tiene mayor correlación.
2. El top tres de formatos escuchados son: CD.MP3 y Vinyl
3. Entre más caracteres tiene un revie, menos votos recibe el producto comprado.