

Genre detection by Lyrics

SUSTENTACIÓN

FINAL

Juan Felipe Rojas De La Hoz 2220070

Diego Alejandro Arevalo Quintero 2220066

PROBLEMA

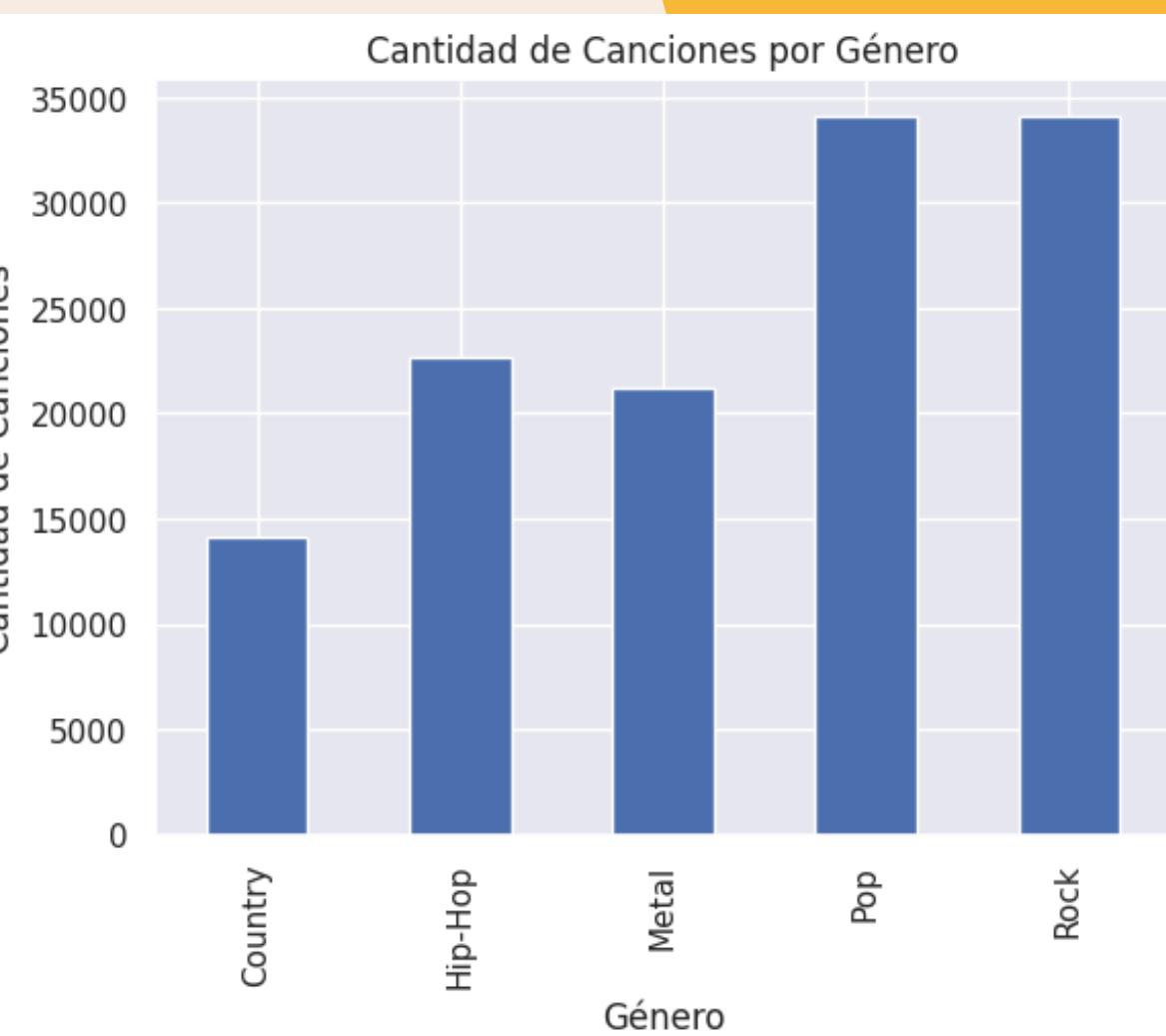
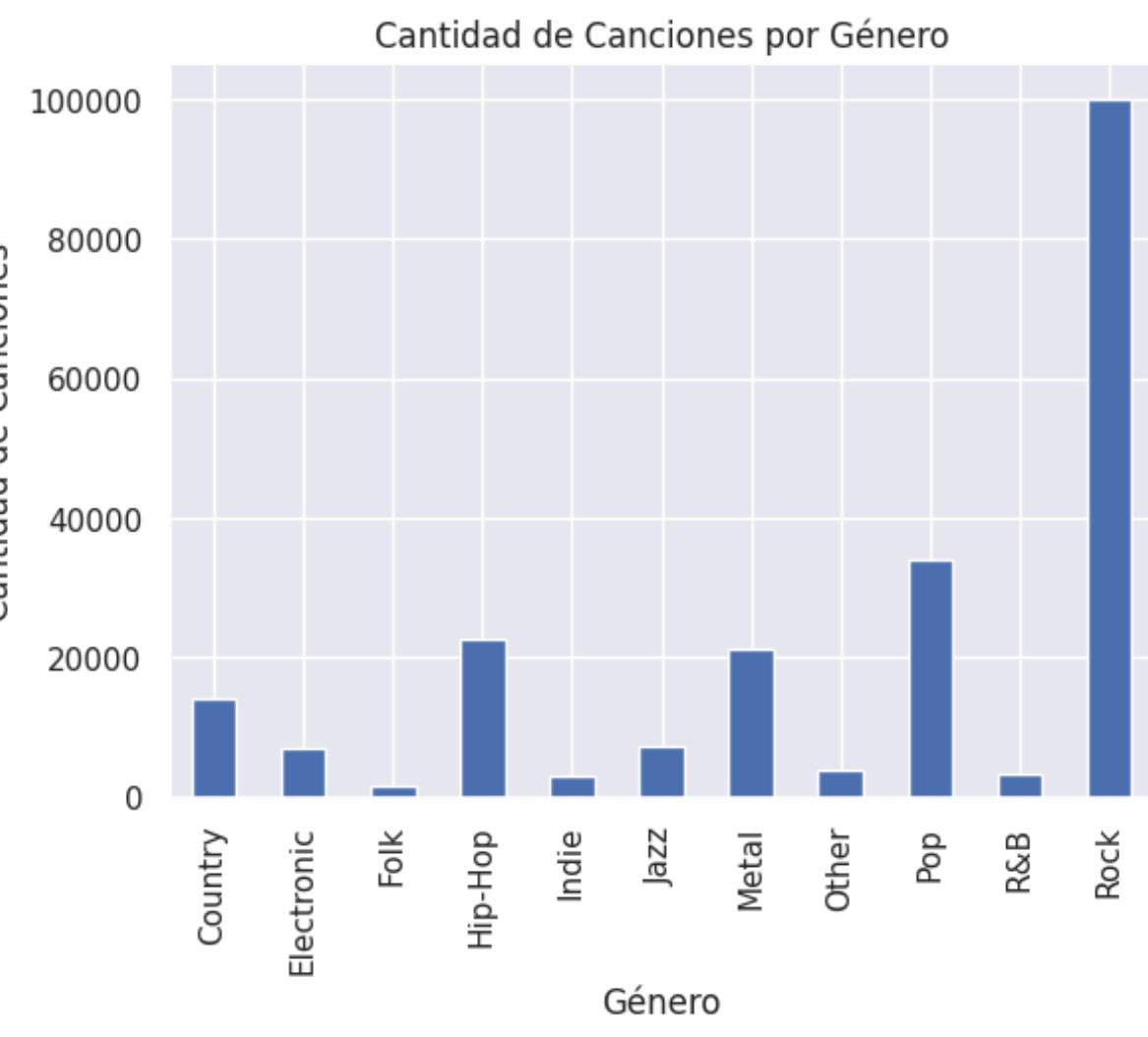
Este proyecto se enfoca en utilizar técnicas de análisis de texto para clasificar canciones por género, tomando como base un dataset de letras musicales. Mediante el uso de algoritmos de machine learning, buscaremos patrones que puedan asociarse a géneros musicales específicos.



PREPRoCESAMiENTO

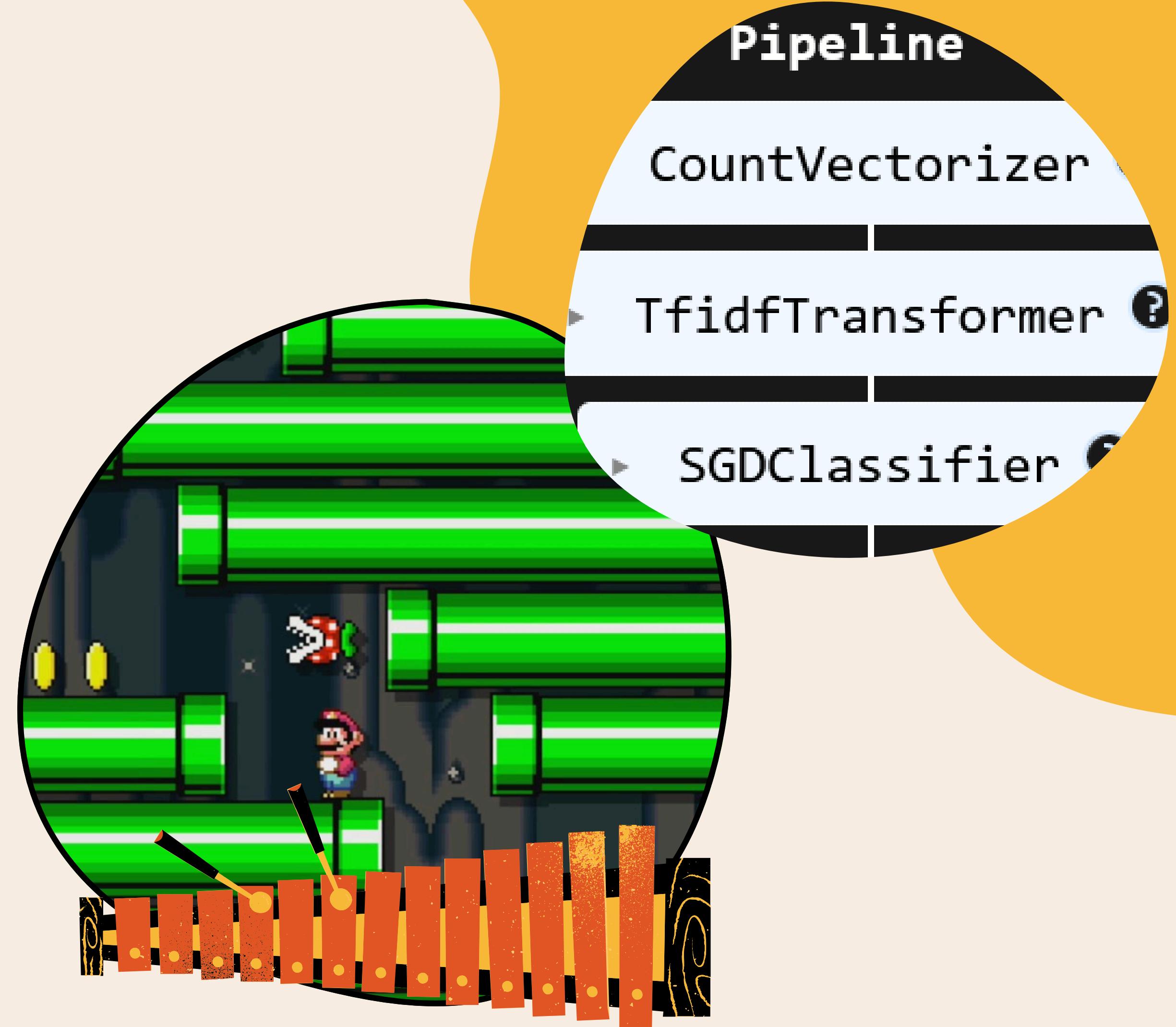
Al inicio el dataset contenía una gran variedad de géneros, claramente algunos con más predominancia que otros gracias a su popularidad en el tiempo en donde se recogieron los datos.

Se escogieron solo los géneros con mayor cantidad de canciones para entrenar a nuestro modelo.



PIPELINE

Un pipeline es una forma de encadenar pasos para realizar un proceso de análisis o modelado de datos.

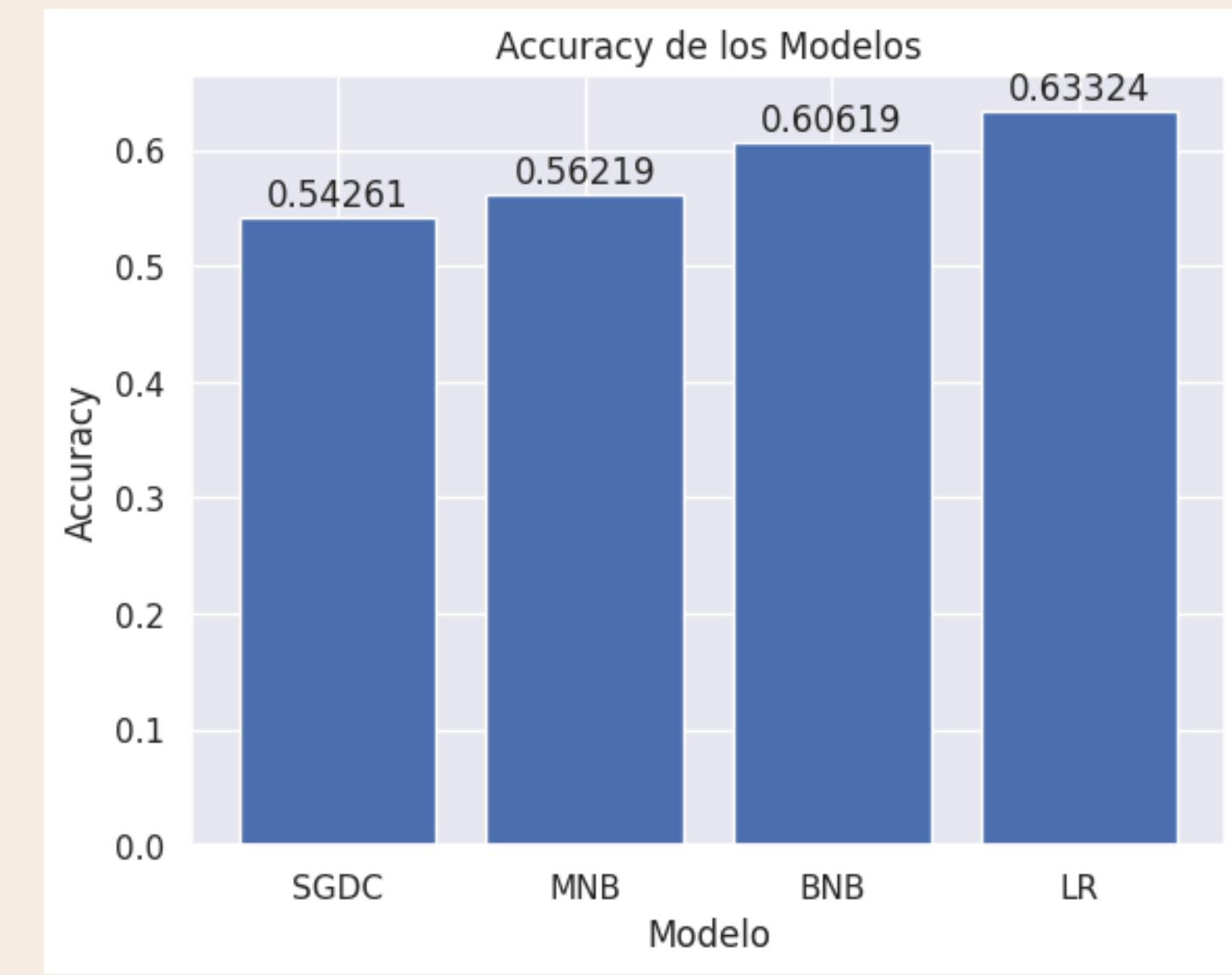


CLASIFICADOR ESCOGIDO LOGISTICREGRESSION

116000
PALABRAS

LogisticRegression

- Es una variante de la regresión lineal que utiliza la función logística que transforma predicciones en probabilidades , generalmente funciona con features binarias.





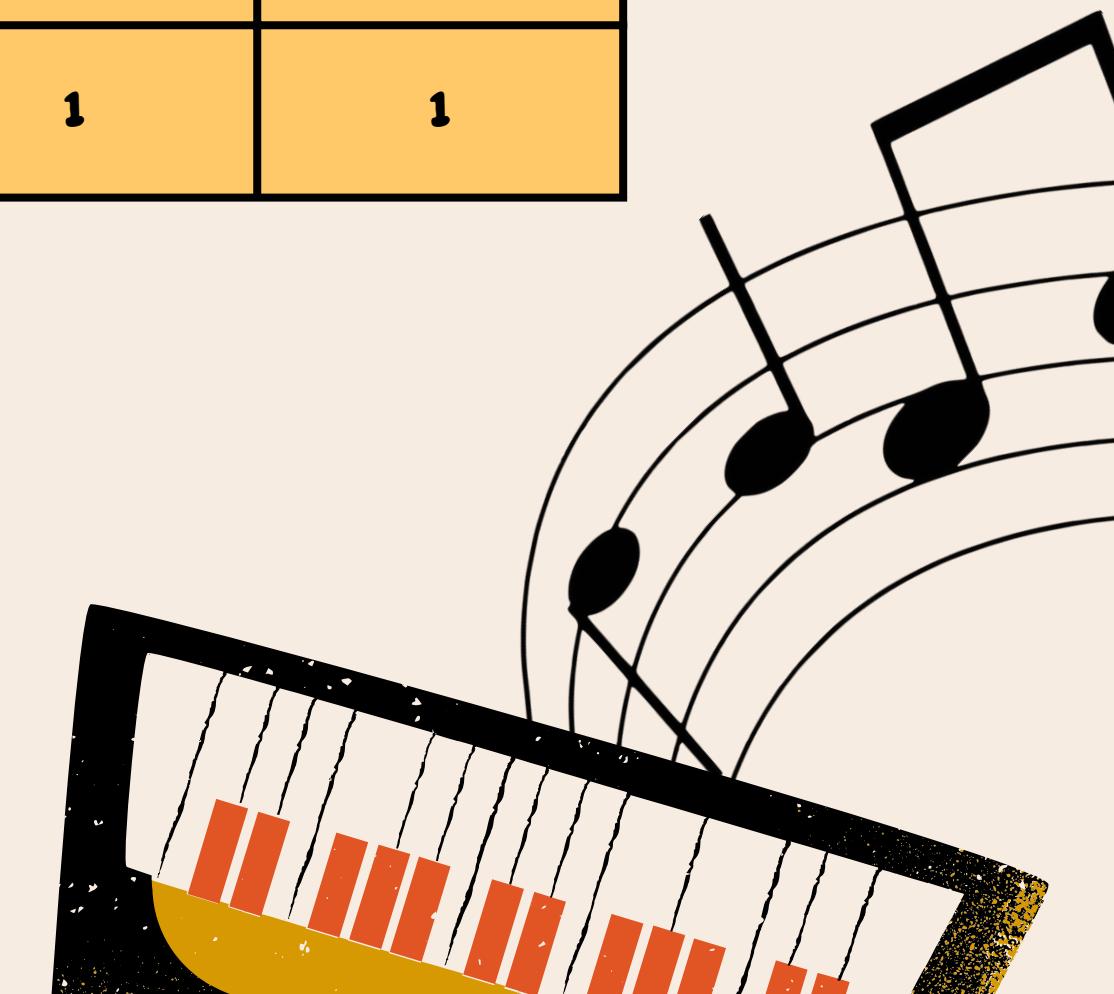
COUNT VECTORIZER

Convierte texto en una representación numérica llamada matriz de frecuencia de términos.

Básicamente, cuenta cuántas veces aparece cada palabra en los textos y genera una tabla donde cada fila representa un documento (por ejemplo, una canción) y cada columna, una palabra.

ENTRADA: ["HOLA MUNDO", "MUNDO FELIZ"]

Hola: 0	Mundo: 1	feliz: 2
1	1	0
0	1	1



TFIDF TRANSFORMER

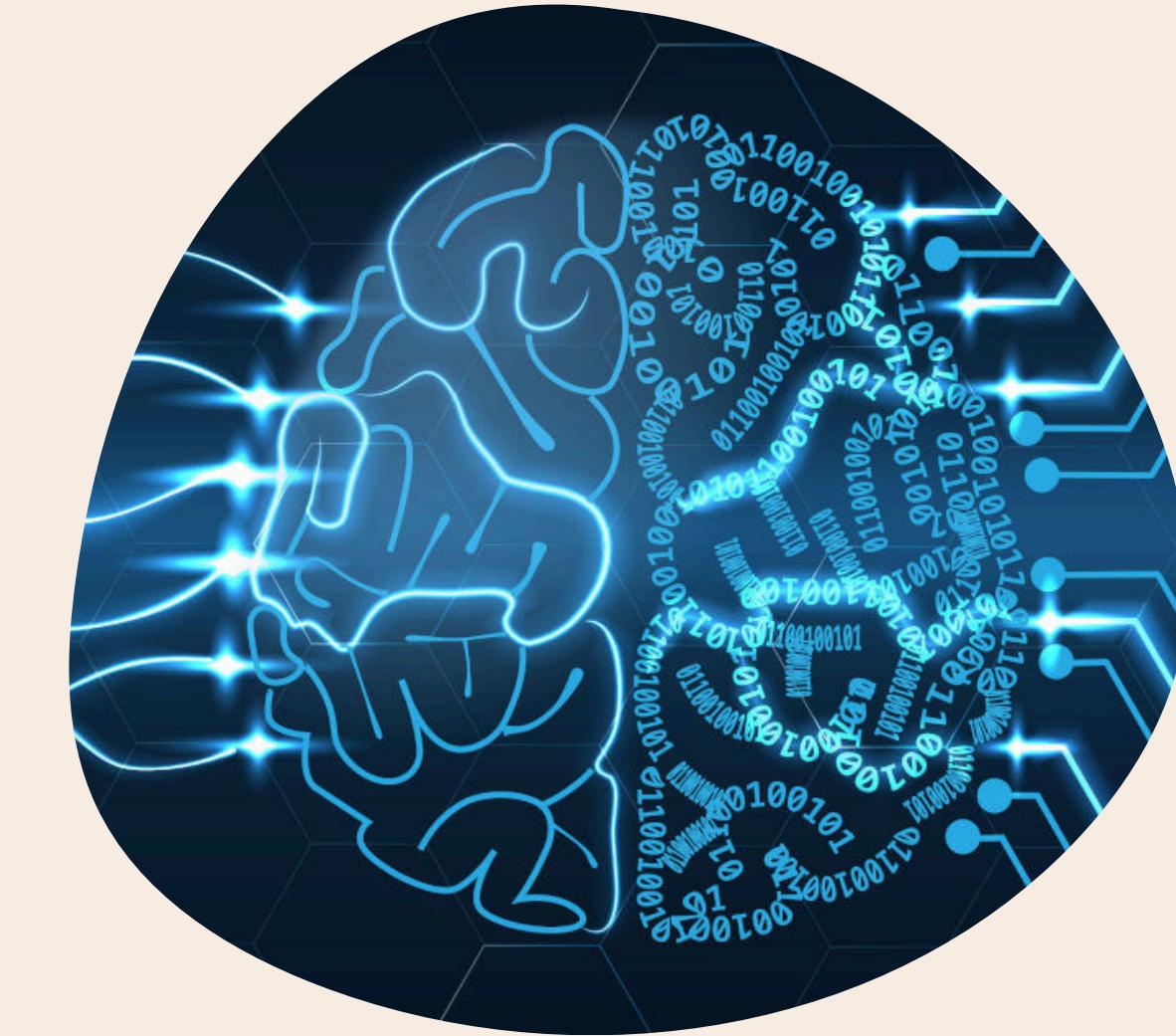
Calcula el Term Frequency–Inverse Document Frequency

Este paso ajusta las frecuencias generadas en el paso anterior para dar más peso a las palabras importantes (que aparecen en pocos documentos) y reducir la importancia de las palabras comunes (que aparecen en muchos documentos).



DEEP LEARNING

Para el modelo de deep learning se aplicó una función de pérdida “sparse_categorical_crossentropy” el cual es útil si se quiere aplicar deep learning a Data Sets llenos de 0s como es nuestro caso donde la gran parte de nuestros registros tienen 0



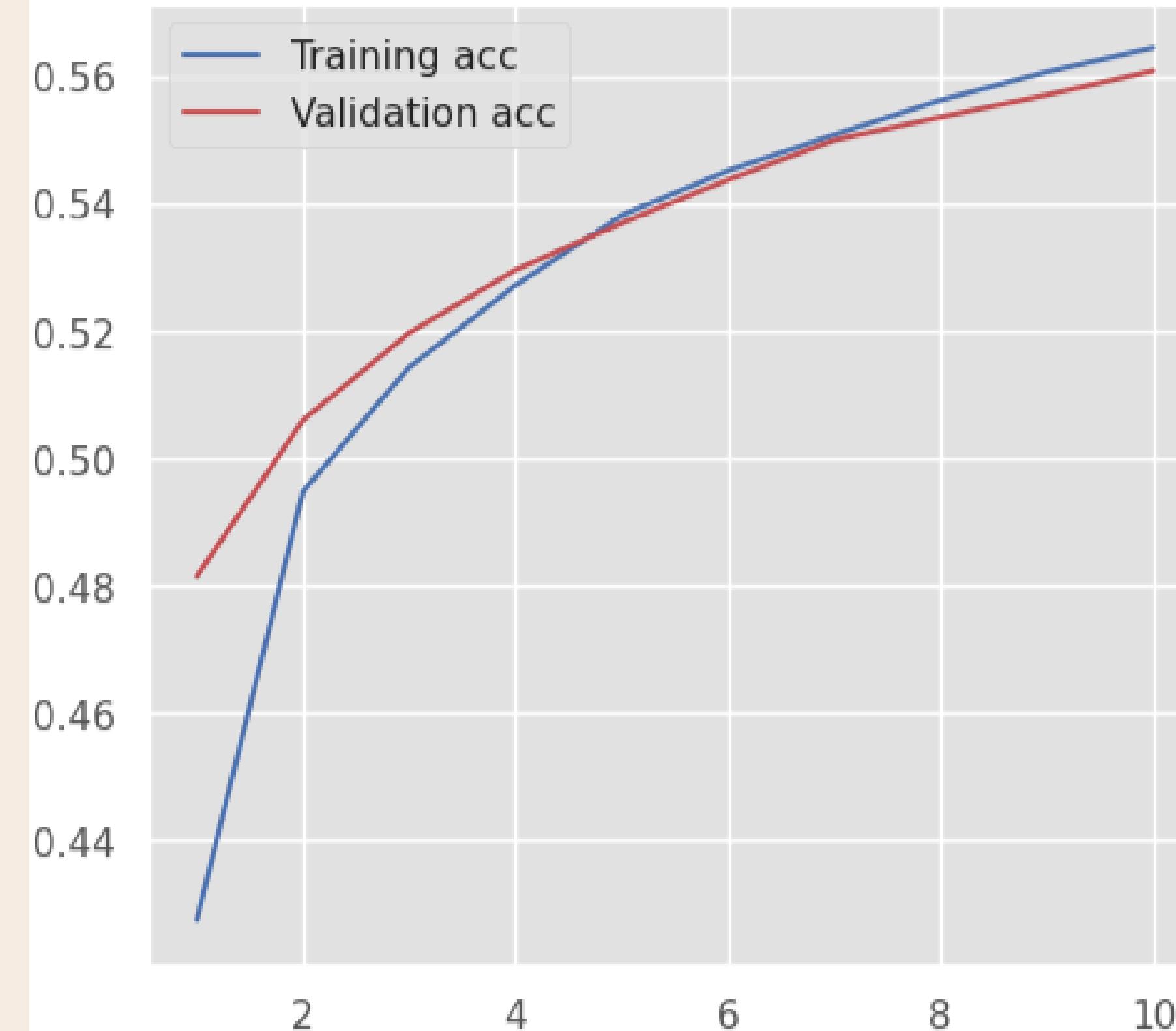
```
model.compile(loss='sparse_categorical_crossentropy',
              optimizer='adagrad',
              metrics=['accuracy'])
model.summary()
```

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 10)	1,532,830
dense_1 (Dense)	(None, 6)	66

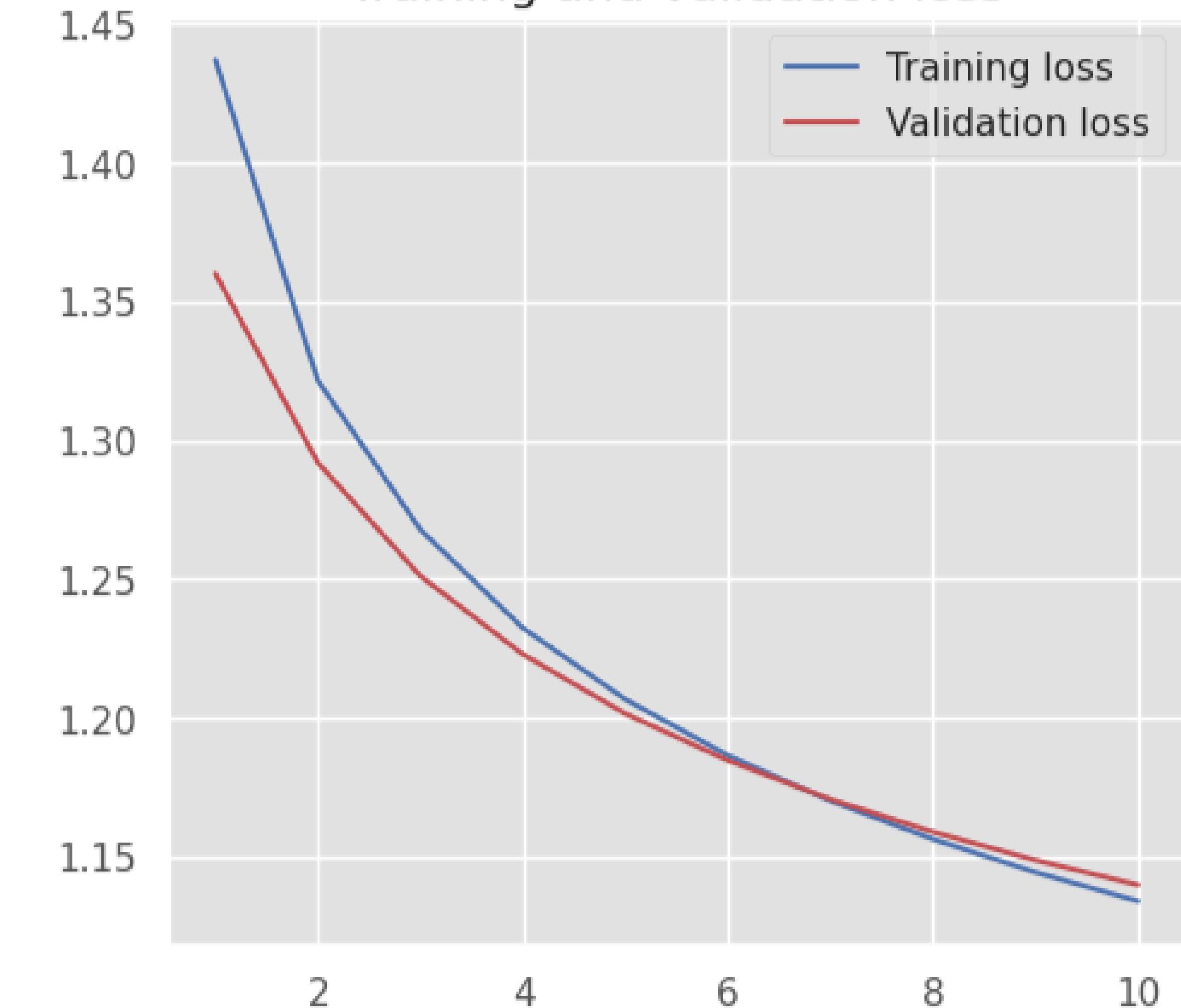
Total params: 1,532,896 (5.85 MB)
Trainable params: 1,532,896 (5.85 MB)
Non-trainable params: 0 (0.00 B)

Training and validation accuracy



TESTING ACCURACY: 0.5609

Training and validation loss



TRAINING ACCURACY: 0.5642

CONCLUSIONES

Predecir el género de una canción tan solo con la letra de esta no es viable por similitud de temas entre algunos generos, lo cual se llega a un límite de entre 60% y 65% en accuracy, con respecto tanto a nuestro proyecto como otros similares en internet.

