CATEGORIZACIÓN DE ARTÍCULOS

Agustina Sol Ortu
Agustin Mattiussi
María Mercedes Baron
Sol Victoria Anselmo

OBJETIVO

Identificar la temática de los artículos según el lenguaje utilizado.

CORPUS

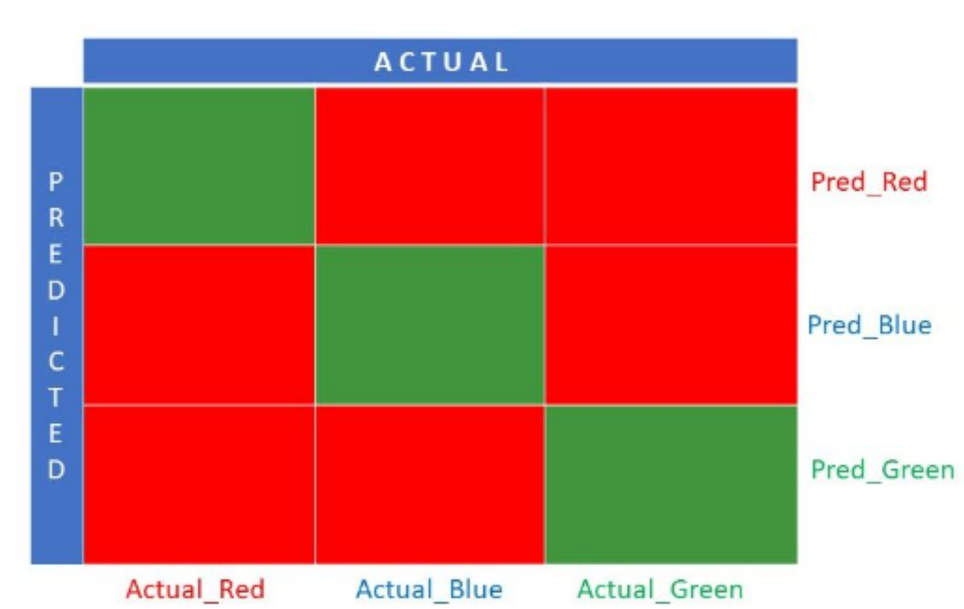
Artículos de la BBC sobre diversas temáticas.

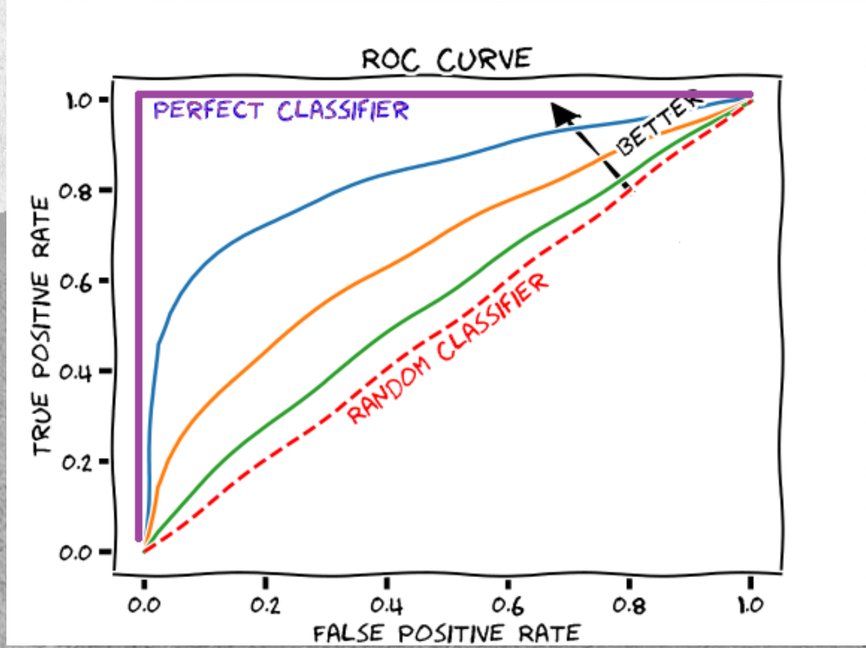
MÉTRICAS NUMÉRICAS

- **Accuracy:** Mide el porcentaje de las clasificaciones correctas.
- **Precision:** Mide el porcentaje de verdaderos positivos sobre todos los casos positivos.
- **Recall:** Mide el porcentaje de verdaderos positivos sobre la suma de verdaderos y falsos positivos.
- **F1 Score:** Combina Precision y Recall para proporcionar una única métrica que considera tanto la completitud como la exactitud.

MÉTRICAS GRÁFICAS

- **<u>Curva ROC:</u>** Representación gráfica de la sensibilidad frente a la razón de falsas alarmas.
- **<u>Matriz de confusión:</u>** Representación gráfica del rendimiento de un clasificador.

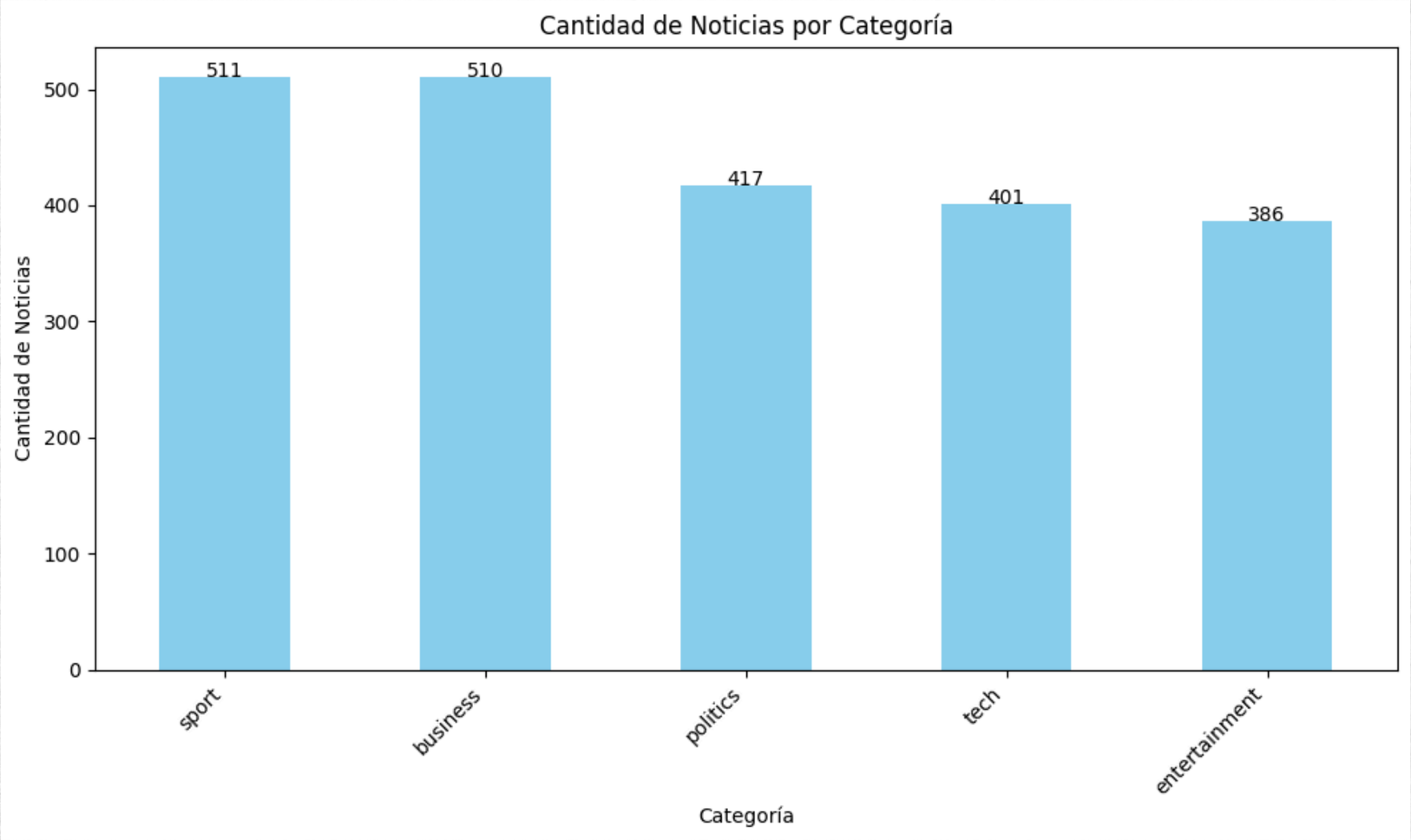




ANALISIS EXPLORATORIO DEL DATASET

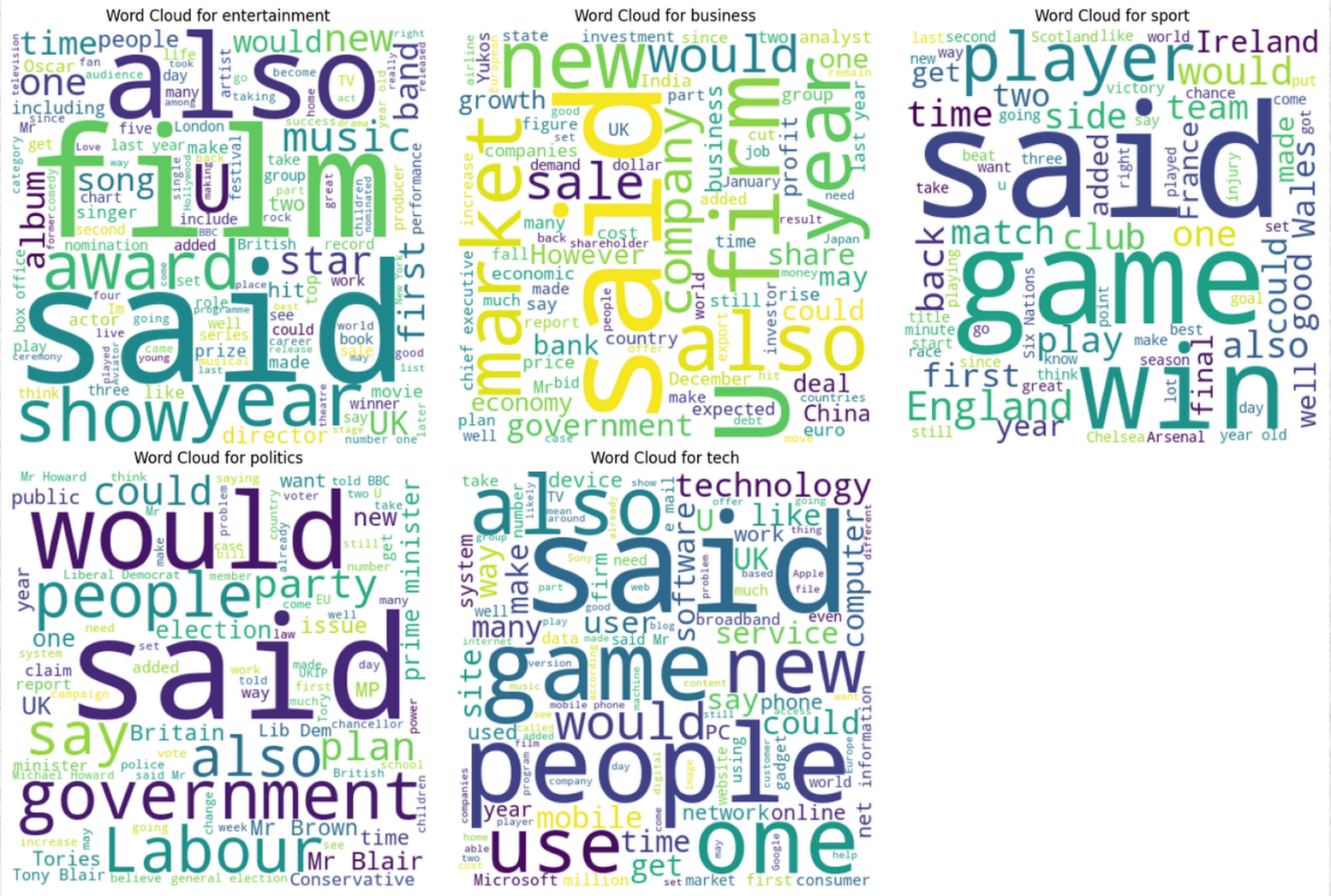
ETIQUETAS:

- Sport
- Politics
- Tech
- Business
- Entertainment

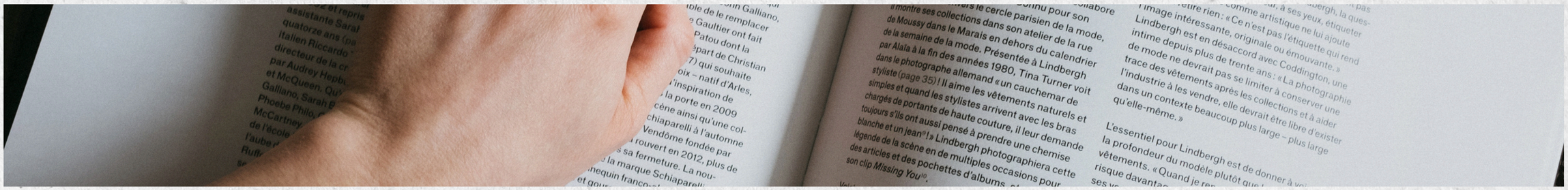


ANALISIS EXPLORATORIO DEL DATASET

Word Cloud de las palabras más comunes para cada etiqueta.



BERT



Usamos BERT ya que el pre-entrenamiento es bidireccional y es pre-entrenado en largos corpus para luego analizar los resultados con un fine-tuning del dataset.

PREPROCESAMIENTO DEL DATASET

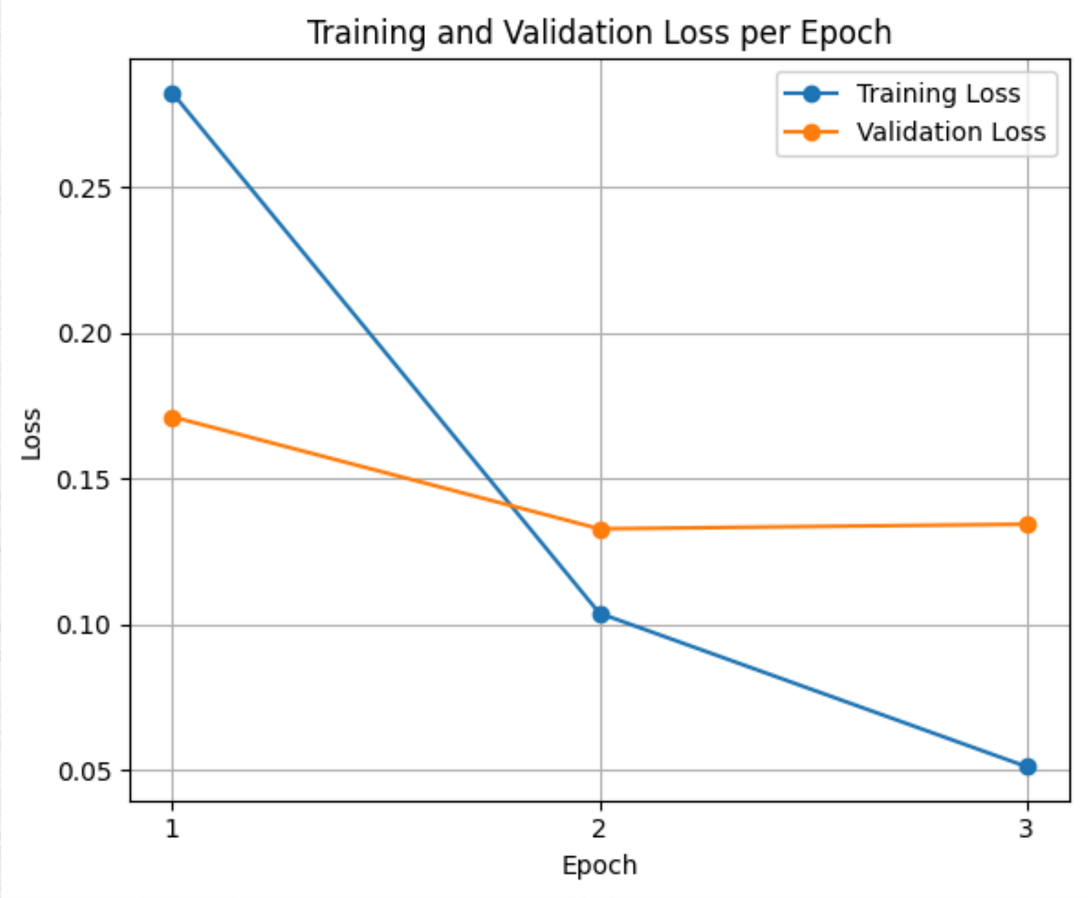
Este modelo tiene contextos limitados por lo que aquellos artículos que superaban el límite de caracteres los dividimos en chunks y les asignamos la categoría del artículo al que pertenecían.

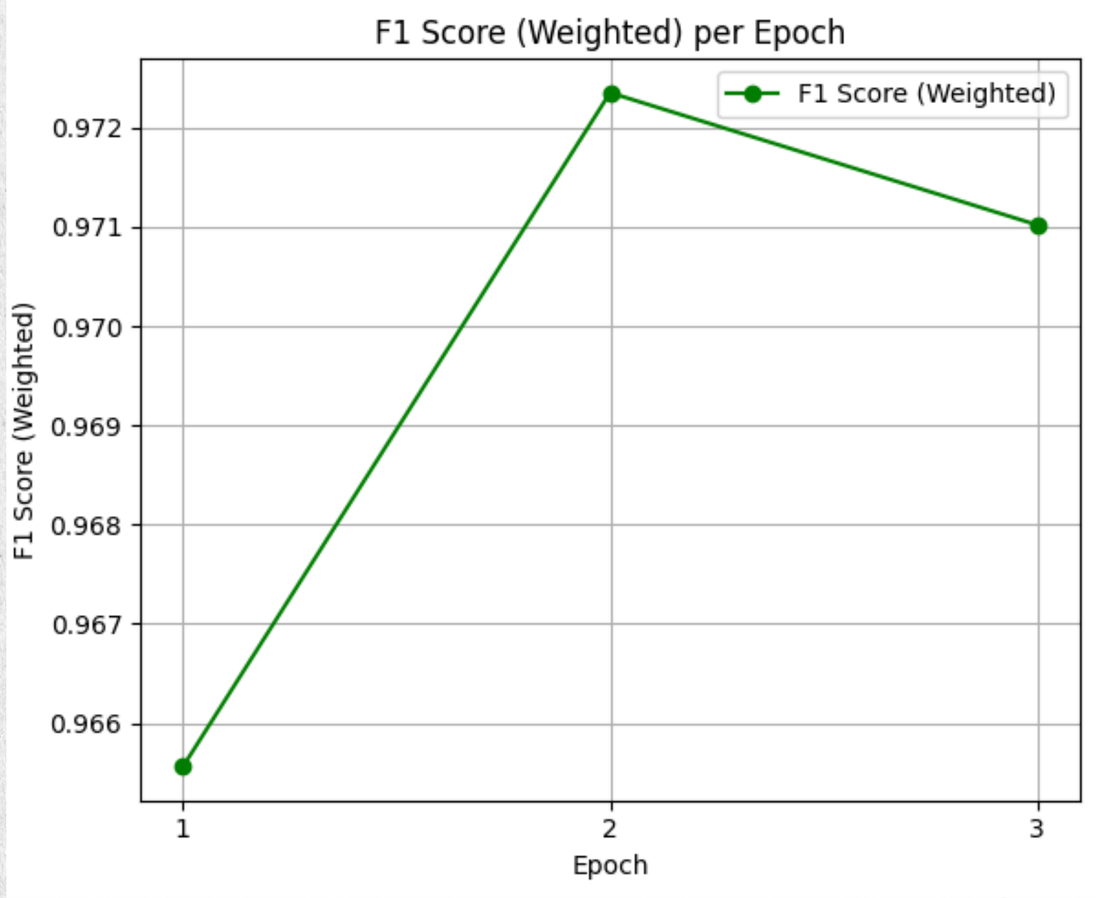
Además, se elimaron stopwords, caracteres especiales y números.

Pérdidas de entrenamiento y validación

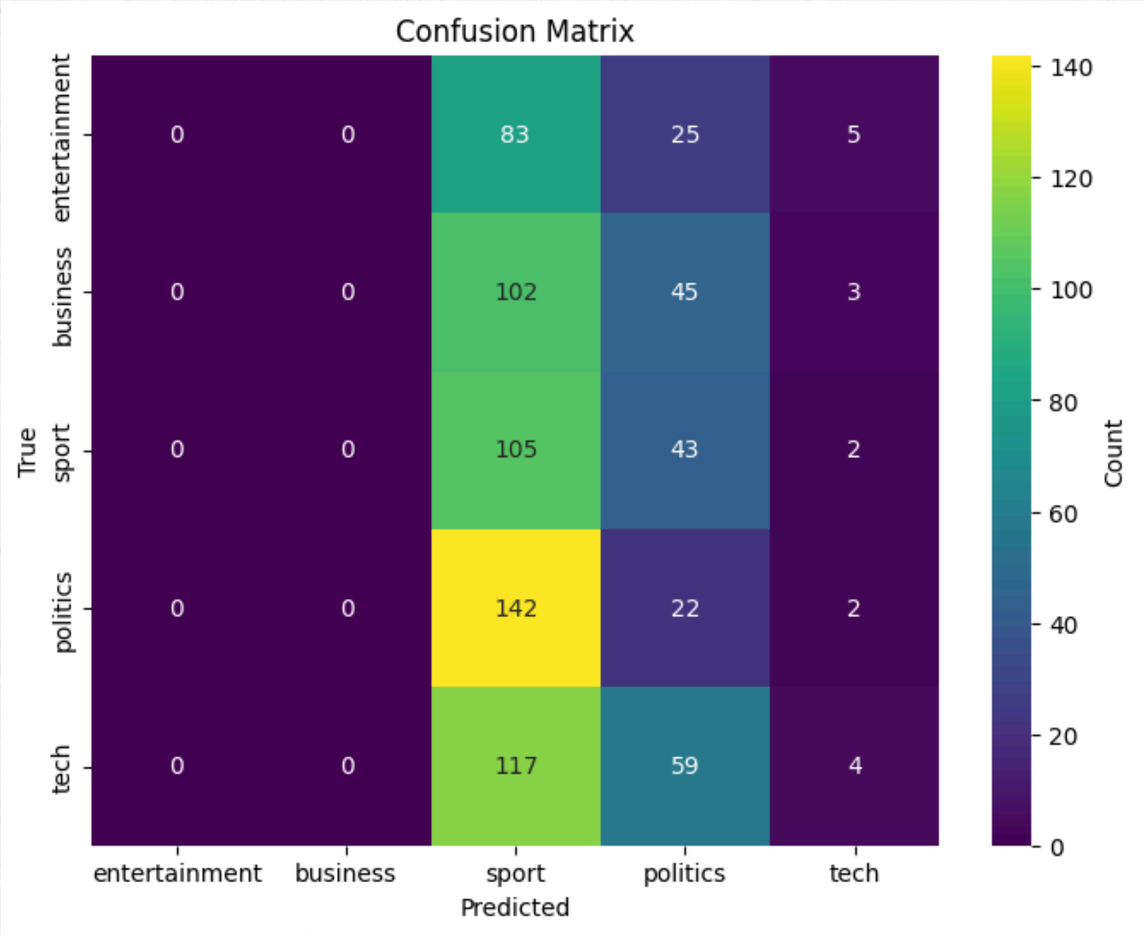
Fine-Tuned

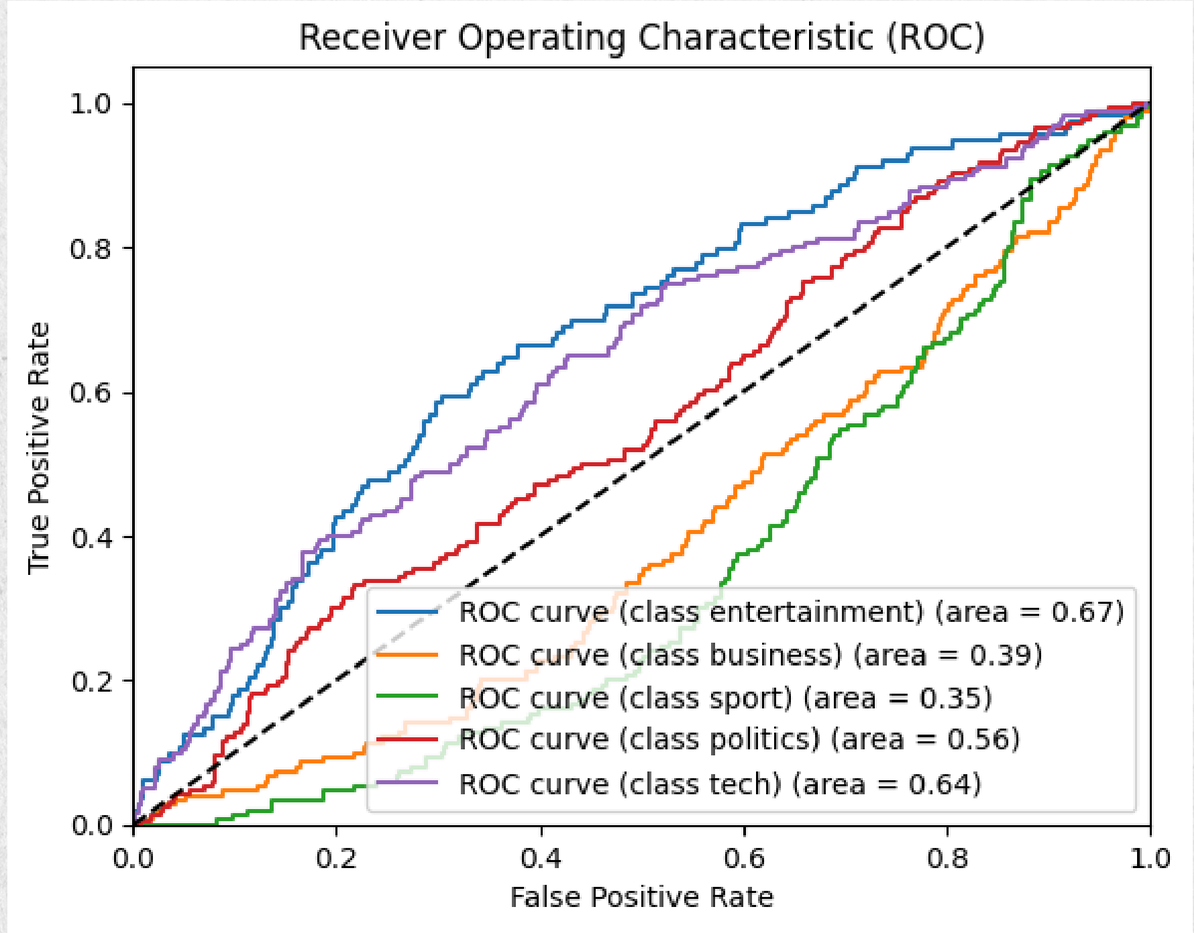
- 1 época = buenos resultados.
- 3 épocas = sobreajuste



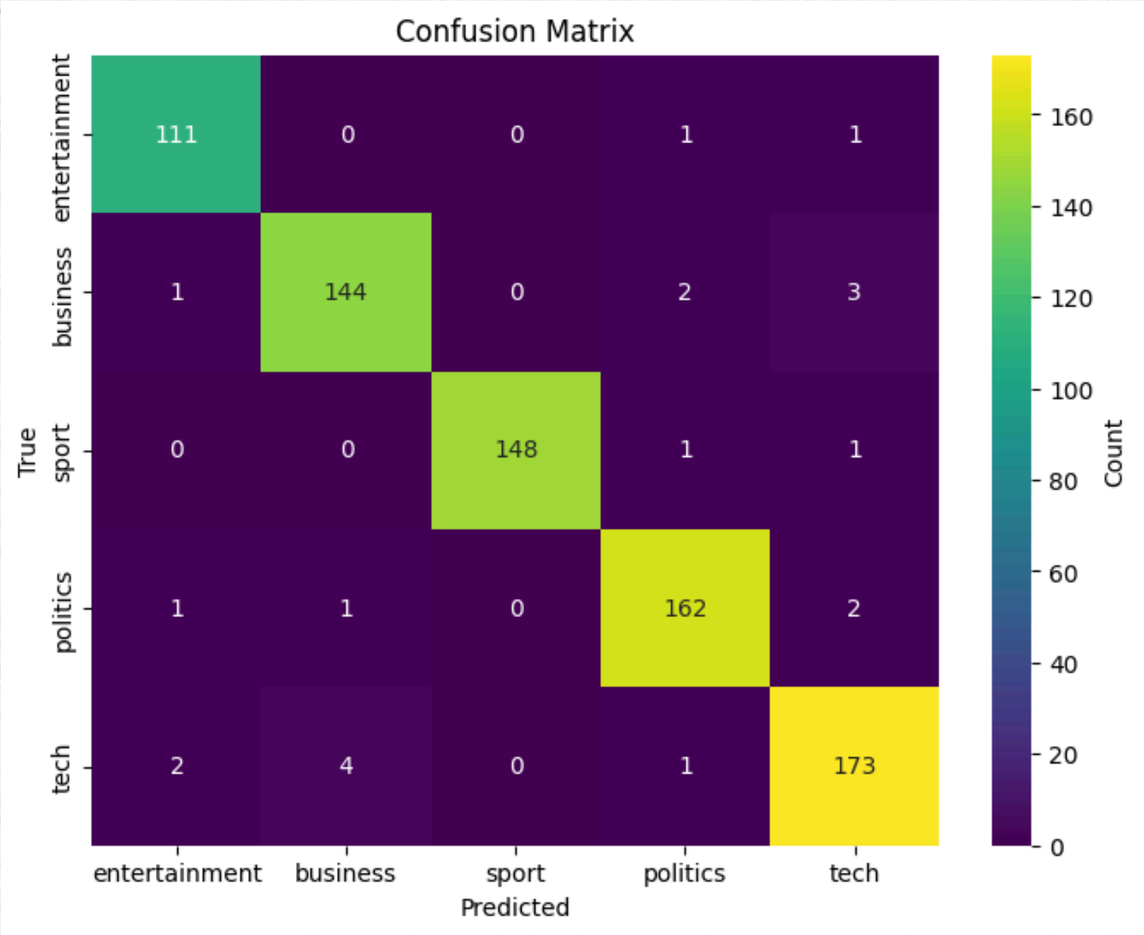


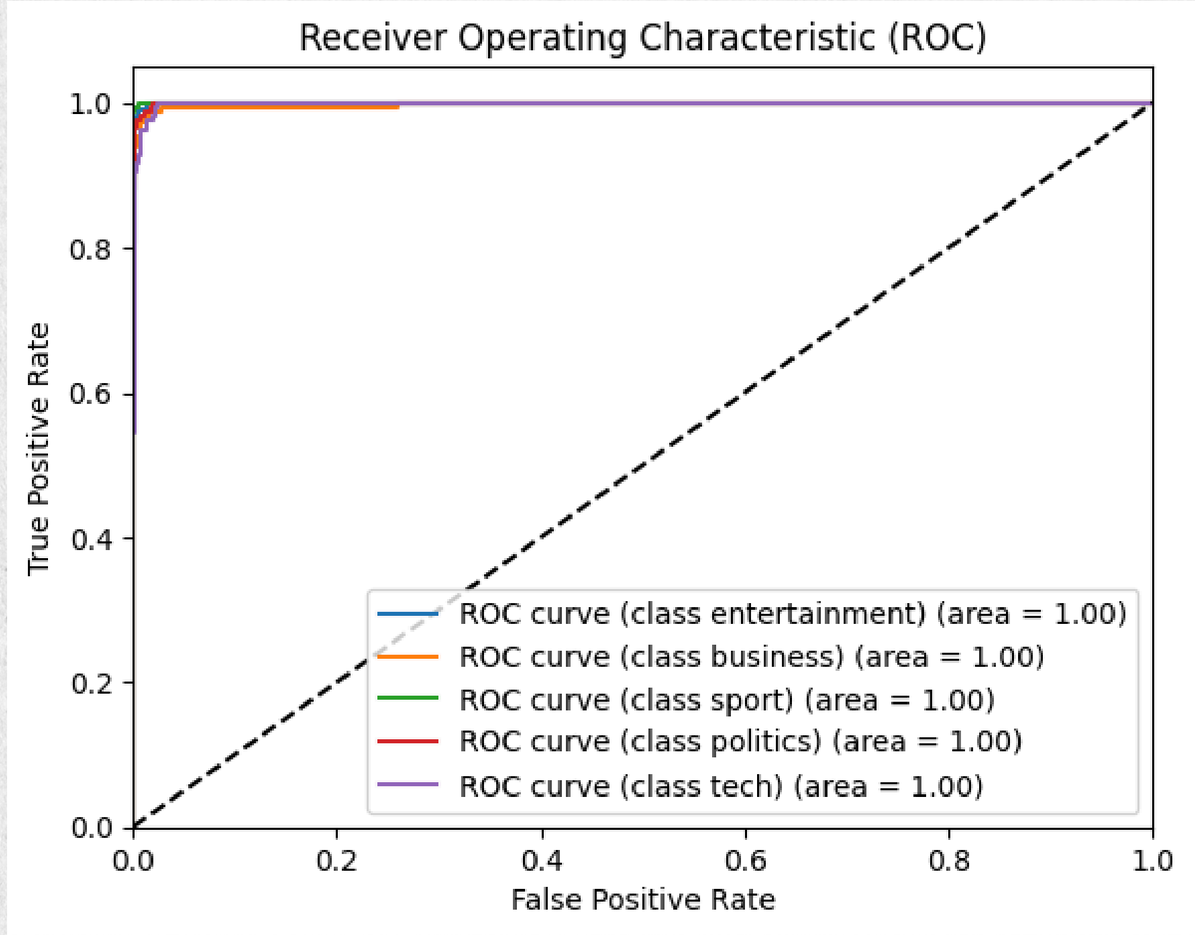
BERT - NO FINE TUNED





BERT - FINE TUNED



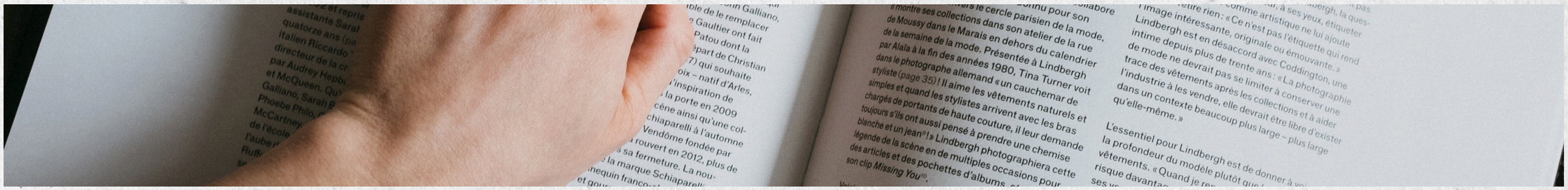


BERT - NO FINE TUNED VS FINE TUNED

Model Metric	Finetuned	No Finetuned
Accuracy	0.972332	0.172596
F1 Score	0.972351	0.095784
Precision	0.972419	0.121888
Recall	0.972332	0.172596

Veamos que como mejoran las metricas una vez que aplicamos finetuned.

TF-IDF



Usamos TF-IDF para identificar qué tan importante es un término en el documento a través de su frecuencia de aparición en el mismo.

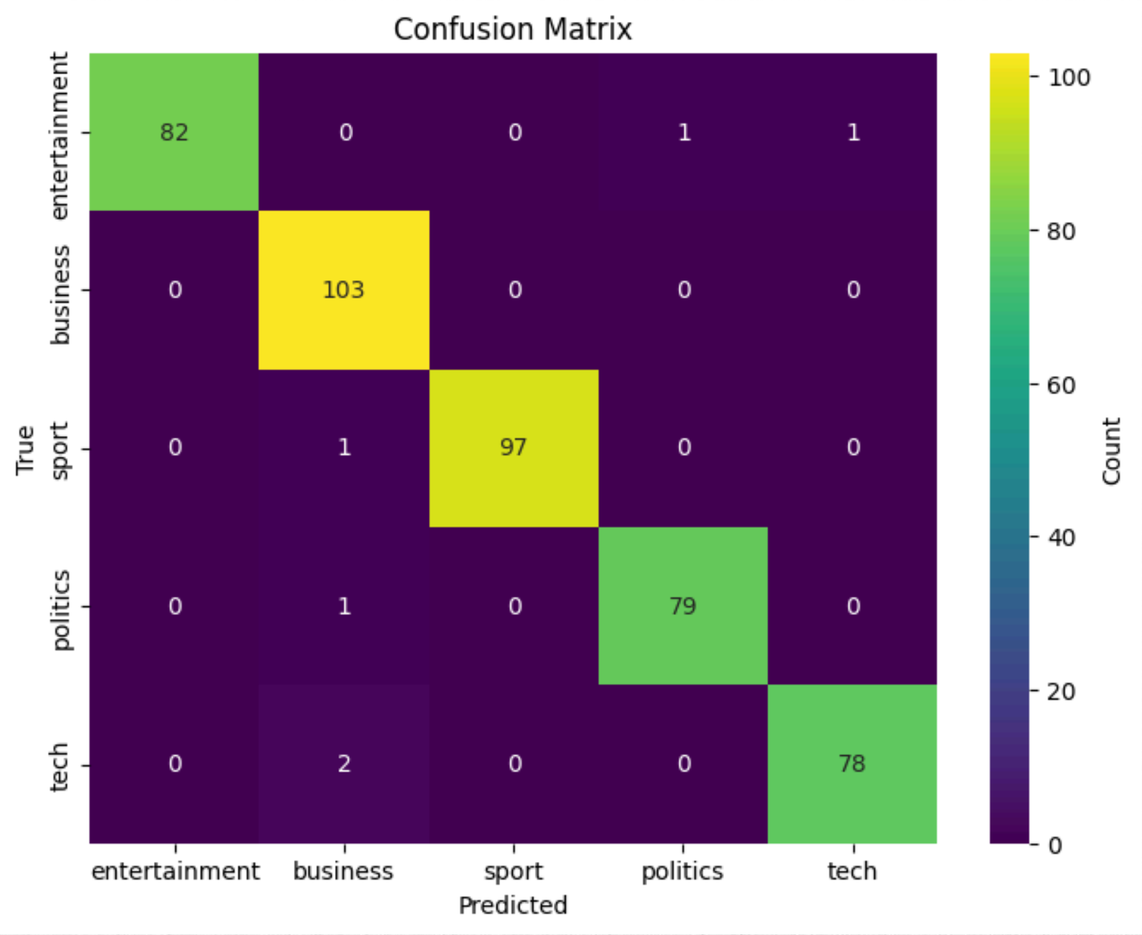
Para el preprocesamiento de los datos se eliminan las stopwords (él, la, pero, etc).

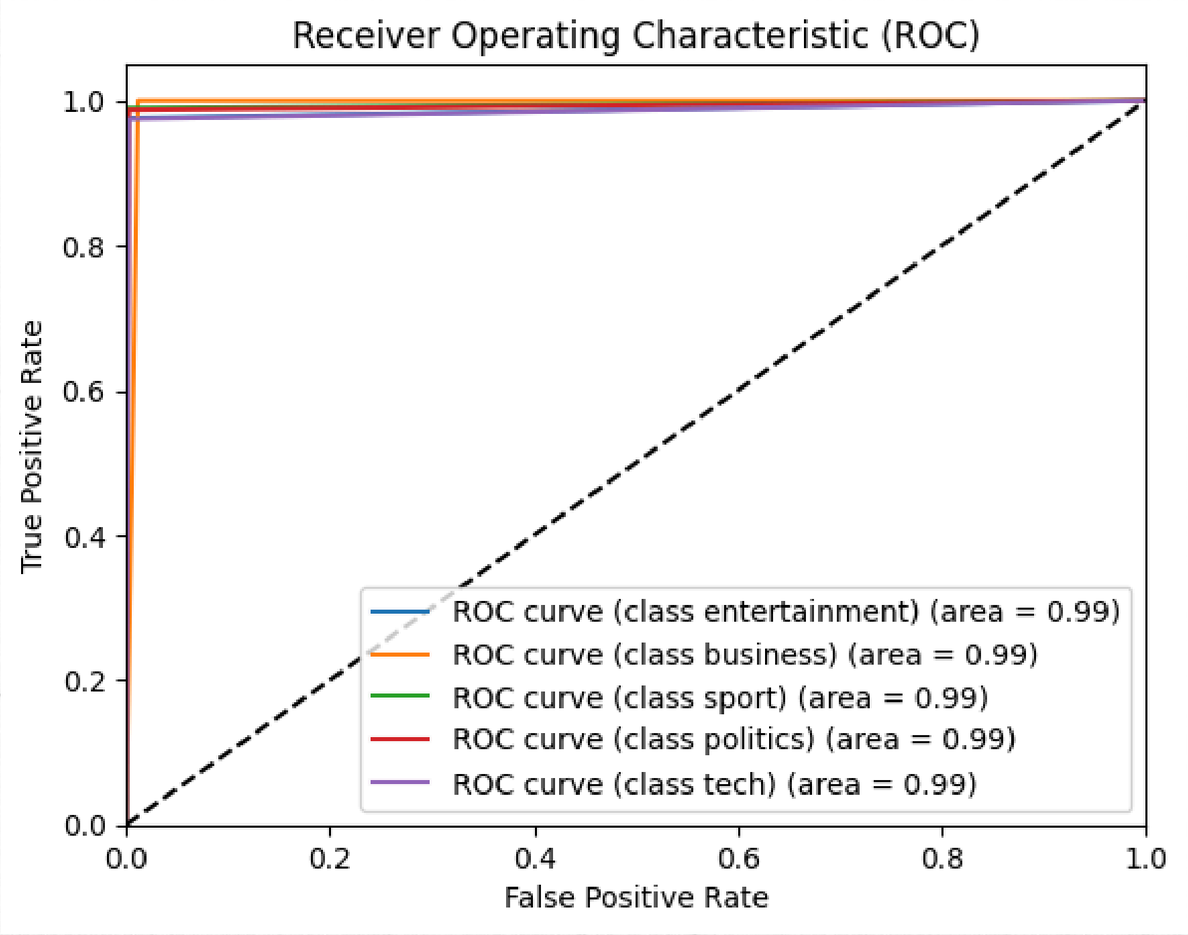
$$TF = \frac{t \in d}{T \in d}$$

$$TF - IDF = TF \times IDF$$

$$IDF = x * \log\left(\frac{D}{\{d \in D : t \in d\}}\right)$$

TF-IDF


Confusion Matrix
True
entertainment
business
sport
politics
tech
82 0 0 1 1
0 103 0 0 0
0 1 97 0 0
0 1 0 79 0
0 2 0 0 78
Predicted
Count
100
80
60
40
20
0


Receiver Operating Characteristic (ROC)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (class entertainment) (area = 0.99)
ROC curve (class business) (area = 0.99)
ROC curve (class sport) (area = 0.99)
ROC curve (class politics) (area = 0.99)
ROC curve (class tech) (area = 0.99)

TF-IDF

Model	Test	Train
Metric		
Accuracy	0.986517	0.997191
F1 Score	0.986482	0.997095
Precision	0.987492	0.997017
Recall	0.985697	0.997178

Podemos ver en las métricas que el modelo tiene un rendimiento muy bueno tanto en los datos de entrenamiento como en los datos de prueba,

CONCLUSIONES

BERT

- Para obtener mejores resultados se debe aplicar el ajuste fino (fine-tuning).
- Las métricas demuestran una diferencia de casi un 80% entre el entrenamiento realizado con fine-tuning y el no-fine-tuned.

TF-IDF

- El modelo TF-IDF junto con la regresión logística aplicada está funcionando correctamente en la tarea de clasificación de textos, mostrando un buen equilibrio entre precisión y sensibilidad, y generalizando bien a datos no entrenado.

GENERALES

- Ambos modelos cumplieron el propósito de manera efectiva, dando resultados superiores a 95% en cada una de las métricas.

LIMITACIONES Y MEJORAS

BERT

- Para el modelo Fine-Tuned, entrenamos 3 épocas y tomó entre 6 y 7 horas en terminar de correr, al realizar el ajuste fino se tarda mucho en procesar los datos.
- La cantidad de palabras que puede procesar por artículo es limitado.

TF-IDF

- Se podría usar técnicas de búsqueda como Grid Search o Random Search para optimizar mejor los parámetros del modelo de regresión logística.
- Se podría añadir una validación de datos de entrada para las funciones train y evaluate_model.

GRACIAS