.UBAfiuba
FACULTAD DE INGENIERÍA

Trabajo Práctico Final

Modelos grandes de lenguaje (LLM)

Alumno: Albachiaro Leandro

INTRODUCCIÓN

El objetivo de este proyecto es implementar un modelo capaz de detectar mensajes de texto SPAM para proteger a los usuarios de comunicaciones no deseadas y potencialmente peligrosas.

Para esto se utilizará DistilBERT, una versión compacta y eficiente del modelo BERT, pre-entrenado para tareas de procesamiento del lenguaje natural (NLP).

DESARROLLO

- **Preparación de datos:** se realiza un análisis exploratorio del dataset, el cual contiene una colección de 5572 mensajes de texto etiquetados como SPAM o legítimos.

  SMS Spam Collection Dataset | Kaggle

- **Carga del modelo pre-entrenado de DistilBERT:** se utiliza la biblioteca transformers de Hugging Face para cargar el modelo pre-entrenado de DistilBERT y su tokenizador correspondiente.
- **Tokenización de mensajes:** se tokenizan los mensajes de texto utilizando el tokenizador de DistilBERT.

DESARROLLO

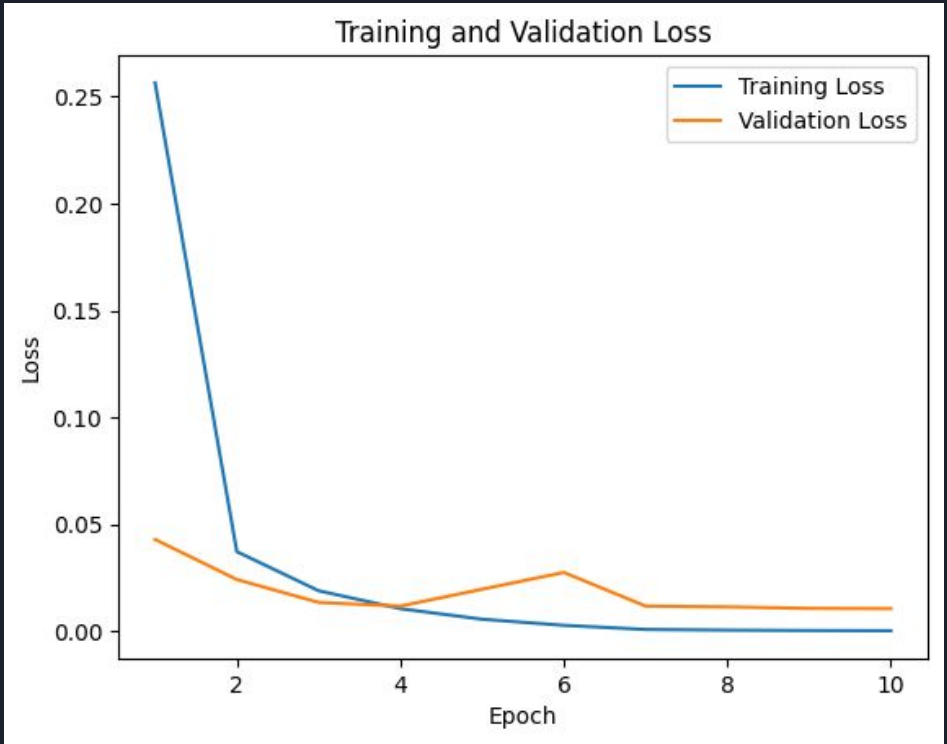
- **Entrenamiento del modelo:** Se define una arquitectura de clasificación de secuencias utilizando DistilBERT como base y se ajustan los pesos del modelo utilizando los datos de entrenamiento.

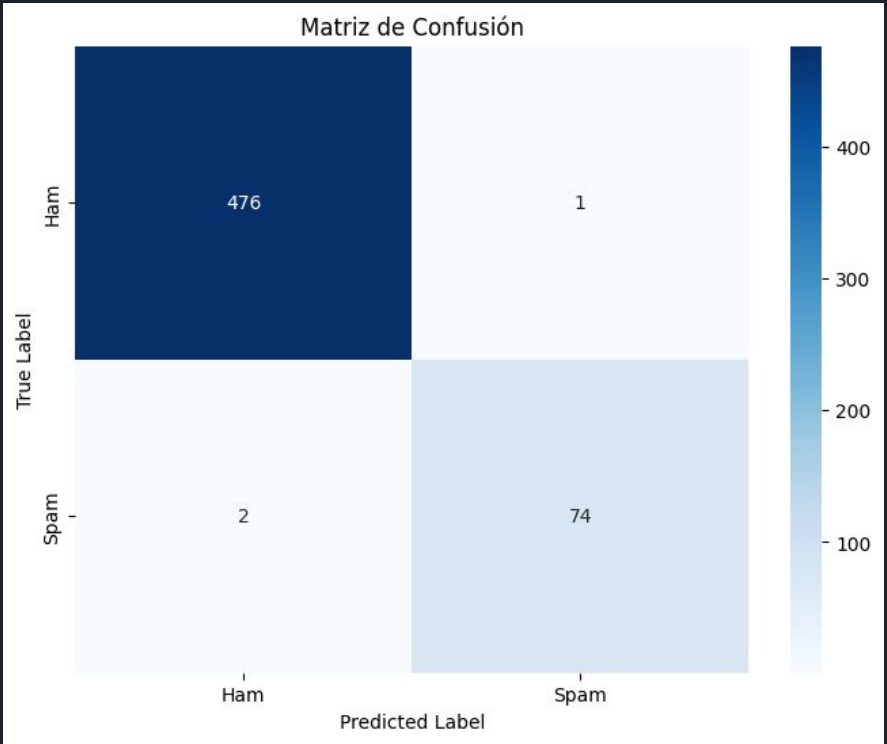
  Durante el entrenamiento, se ajustan los parámetros del modelo para minimizar la función de pérdida de entropía cruzada (Cross-Entropy Loss), que mide la discrepancia entre las predicciones del modelo y las etiquetas reales.

- **Evaluación del modelo:** Una vez entrenado el modelo, se evalúa su rendimiento utilizando datos de prueba previamente no vistos.

  Se calculan métricas como precisión, recall para medir la capacidad del modelo para distinguir entre mensajes de texto SPAM y legítimos.

RESULTADOS


Training and Validation Loss
Training Loss
Validation Loss
Loss
Epoch
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6
8
10


Matriz de Confusión
True Label
Predicted Label
Ham
Spam
476
1
2
74
400
300
200
100

RESULTADOS

```
[ ]  # Ejemplo de mensaje spam
     message_to_classify = "Hey! You've won a free vacation. Click here to claim your prize!"

     # Clasificar el mensaje
     predicted_label = classify_message(message_to_classify)
     print("Predicted label:", predicted_label)
```

```
Predicted label: spam
```

```
[ ]  # Ejemplo de mensaje legítimo
     message_to_classify = "I love u!"

     # Clasificar el mensaje
     predicted_label = classify_message(message_to_classify)
     print("Predicted label:", predicted_label)
```

```
Predicted label: ham
```

CONCLUSIONES

- Los resultados muestran la capacidad del modelo para identificar mensajes de spam con una precisión y recall significativos.
- A pesar de ser un modelo más pequeño y compacto, se demostró con éxito la viabilidad y eficacia de utilizar DistilBERT para la detección de spam en mensajes de texto.