Ejercicio 2

Aplicaciones de la Lingüística Computacional – Curso 2023/2024

Iñaki Diez Lambies

# Ejercicio 4

En el ejercicio propuesto, se realizó el ajuste fino del modelo DistilBERT, específicamente diseñado para la clasificación de tokens, con el propósito de identificar elementos narrativos en textos clasificados según su inclinación hacia narrativas conspirativas o críticas. Este proceso de ajuste fino es crucial para adaptar modelos preentrenados a tareas específicas, permitiendo que el modelo reconozca patrones y características únicas de los conjuntos de datos en cuestión.

Para llevar a cabo este ajuste fino, se utilizó el modelo base `distilbert-base-uncased`, seleccionado por su eficiencia y eficacia comprobadas en tareas de procesamiento de lenguaje natural. Los hiperparámetros empleados durante el entrenamiento incluyen una tasa de aprendizaje de 2e-5, un tamaño de *batch* de 8 para el entrenamiento y la evaluación, y un total de 3 *epochs*. Este conjunto de hiperparámetros fue escogido para equilibrar entre la precisión del modelo y la viabilidad computacional, considerando las limitaciones de tiempo y recursos. Además, se aplicó un peso de decaimiento de 0.01 para regularizar el modelo y evitar el sobreajuste.

Los resultados obtenidos tras el proceso de ajuste fino y evaluación en el conjunto de prueba revelaron un puntaje F1 de 0.5949, una precisión de 0.5803 y un recall de 0.6104. Estos resultados indican una capacidad moderada del modelo para identificar correctamente las etiquetas relevantes dentro del texto, demostrando un equilibrio entre la precisión y la capacidad de recuperación de información relevante. La precisión, aunque no es excepcionalmente alta, es indicativa de que cuando el modelo predice una etiqueta, es relativamente confiable. Por otro lado, el recall más elevado sugiere que el modelo es capaz de identificar una porción significativa de las etiquetas relevantes, aunque esto también pueda conllevar a una mayor cantidad de falsos positivos. La combinación de estos factores resulta en el puntaje F1 mencionado, que sirve como una medida de balance entre precisión y recall.

A pesar de los resultados prometedores, existe un margen considerable para mejorar el rendimiento del modelo. Una estrategia que se considera podría tener un impacto positivo significativo es la ampliación del conjunto de datos de entrenamiento. Incorporar una mayor variedad de ejemplos y contextos puede ayudar al modelo a generalizar mejor en situaciones no vistas anteriormente, lo cual es esencial para mejorar tanto la precisión como el recall. Este enfoque se basa en la premisa de que modelos de aprendizaje automático, especialmente aquellos basados en redes neuronales, suelen beneficiarse de grandes volúmenes de datos diversificados.

Por otro lado, un enfoque más experimental que podría explorarse es la modificación de la arquitectura del modelo, incorporando capas de atención adicional o ajustando la estructura de las capas existentes para adaptarse mejor a las características específicas de la tarea. Aunque este método implica un riesgo mayor debido a la complejidad y la posibilidad de sobreajuste, la personalización de la arquitectura del modelo ofrece una vía intrigante para mejorar el rendimiento específico en tareas de clasificación de tokens complejas.

En general, los resultados obtenidos del ajuste fino del modelo DistilBERT para la clasificación de tokens en el contexto de análisis de narrativas oposicionistas ofrecen un punto de partida sólido para futuras investigaciones y mejoras. Las estrategias sugeridas para la mejora del rendimiento, tanto la ampliación del conjunto de datos como la experimentación con la arquitectura del modelo, representan direcciones prometedoras que podrían llevar a un avance significativo en la precisión, recall y, por ende, en el puntaje F1 del modelo en tareas futuras.