

Clasificación de reseñas de películas IMBD - BERT

Tabla de Accuracy

	Bag-of-Words (BOW)		TD-IDF		BERT	
	Logistic Regression	Random Forest	Logistic Regression	Random Forest	Red Neuronal Entrenamiento	Red Neuronal Validación
Nsamp = 1000, maxtokens = 50, maxtokenlen = 20	66%	67.16%	69.5%	66.83%	77.14%	68.5%
Nsamp = 1000, maxtokens = 100, maxtokenlen = 100	75.5%	75.5%	72.16%	68%	81.07%	76.33%
Nsamp = 1000, maxtokens = 200, maxtokenlen = 100	80%	80%	84%	78.16%	81.57%	79.67%
Nsamp = 1000, maxtokens = 200, maxtokenlen = 200	-	-	-	-	81.14%	77.33%
Nsamp = 1000, maxtokens = 230, maxtokenlen = 100	-	-	-	-	82.93%	81.67%

Coclusión

En la tabla se puede observar un análisis comparativo entre diferentes representaciones de texto (Bag-of-Words, TF-IDF y embeddings contextuales con BERT) utilizando Logistic Regression (LR), Random Forest y Red Neuronal como modelos de clasificación. En los primeros tres experimentos, Bag-of-Words y TF-IDF logran desempeños buenos. Sin embargo, en el panorama general el desempeño de BERT supera significativamente a BOW y TG-IDF alcanzando precisiones superiores de 81% mientras las demás se quedan en 80% o por debajo de esta, solo LR en TD-IDF con Nsamp = 1000, maxtokens = 200, maxtokenlen = 100 alcanza un 84% superando a BERT, sin embargo considerando los desempeños generales observados BERT tiene un desempeño superior.

La ventaja principal del uso de los embeddings contextuales generados por BERT radica en su capacidad para capturar relaciones semánticas profundas en los textos. A diferencia de BOW y TF-IDF, que se basan en representaciones estáticas y no

consideran el contexto de las palabras, BERT emplea una arquitectura de Transformer que permite entender cómo el significado de una palabra varía dependiendo de su posición en la oración. Esto resulta en una mejora notable en tareas complejas de clasificación, especialmente con datos que requieren entender el contexto lingüístico. En esta actividad logramos observar como BERT mostró un desempeño consistentemente superior, evidenciando su eficacia como herramienta de modelado en aplicaciones de procesamiento del lenguaje natural.