UNIVERSIDAD DE ANTIOQUIA
1803

Análisis de sentimiento en reseñas de películas utilizando LSTM

Daniel Brand Taborda

Jhonier Raúl Jiménez

Deep Learning 2025

Entrega 1

Contexto de aplicación

El análisis de sentimiento, también conocido como minería de opinión, es una técnica dentro del procesamiento del lenguaje natural (PNL) que identifica y clasifica el tono emocional presente en un texto. Su objetivo primordial es discernir si la opinión expresada es positiva, negativa o neutral. Esta metodología ha adquirido relevancia en la comprensión de la opinión pública, la retroalimentación de los clientes y la dinámica de las redes sociales, impulsada por el crecimiento exponencial del contenido generado por los usuarios en línea.

Dentro de la industria cinematográfica, las reseñas de películas constituyen una aplicación clásica del análisis de sentimiento. Esta técnica permite evaluar la recepción de una película o su impacto en una audiencia específica mediante la clasificación de las reseñas. El análisis de sentimiento proporciona a los realizadores una valiosa retroalimentación sobre cómo se percibe su trabajo, y puede incluso contribuir a la predicción del éxito de una película. Además, facilita la formulación de estrategias de marketing al comprender las reacciones del público.

Objetivo de machine learning

El objetivo primordial de aprendizaje automático para este proyecto es desarrollar un modelo capaz de predecir con precisión el sentimiento (positivo o negativo) expresado en las reseñas de películas del conjunto de datos de IMDB. Para alcanzar este objetivo, se empleará específicamente una red de memoria a corto plazo (LSTM, por sus siglas en inglés). Las LSTM son un tipo de red neuronal recurrente (RNN) diseñada para aprender y

predecir secuencias de datos de manera efectiva. Resultan particularmente eficaces para capturar dependencias a largo plazo en los datos, lo cual es crucial para el análisis de sentimiento donde el sentimiento de una oración puede depender del contexto establecido por palabras precedentes.

Dataset: reseñas de películas de IMDB

El conjunto de datos utilizado en este proyecto proviene de Kaggle. El enlace específico es https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. El tipo de datos que contiene son reseñas de películas en formato de texto. Cada reseña es un fragmento de texto escrito por un usuario que expresa su opinión sobre una película.

En cuanto al tamaño del conjunto de datos, este consta de 50,000 reseñas de películas. Este total se divide en 25,000 reseñas altamente polarizadas para entrenamiento y 25,000 para pruebas. El archivo del conjunto de datos, denominado IMDB Dataset.csv, tiene un tamaño de 66.21 MB.

La distribución de las clases dentro del conjunto de datos está equilibrada. Se trata de un conjunto de datos para la clasificación binaria de sentimiento, donde los sentimientos se etiquetan como positivos (1) o negativos (0). Existe una distribución equitativa de las clases, con 25,000 reseñas positivas y 25,000 negativas.

Métricas de desempeño

Para evaluar los modelos de clasificación de sentimiento en aprendizaje automático, se utilizan comúnmente varias métricas de desempeño. Estas incluyen:

- **Precisión (Accuracy):** Es el porcentaje de reseñas clasificadas correctamente. Se calcula como (Verdaderos Positivos + Verdaderos Negativos) / Total de Predicciones.
- **Precisión (Precision):** Indica, de todas las reseñas que el modelo clasificó como positivas, cuántas fueron realmente positivas. Se calcula como Verdaderos Positivos / (Verdaderos Positivos + Falsos Positivos).
- **Recuperación (Recall) o sensibilidad:** Mide la capacidad del clasificador para encontrar todos los casos positivos reales. Se calcula como Verdaderos Positivos / (Verdaderos Positivos + Falsos Negativos).
- **Puntuación F1 (F1 Score):** Es la media armónica de la precisión y la recuperación, proporcionando una medida equilibrada.
- **Curva ROC y AUC:** La curva Característica Operativa del Receptor (ROC) grafica la tasa de verdaderos positivos contra la tasa de falsos positivos en diferentes umbrales. El Área Bajo la Curva (AUC) proporciona una medida agregada del rendimiento en todos los umbrales.
- **Matriz de confusión:** Una tabla que visualiza el rendimiento de un clasificador mostrando los conteos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

A nivel de negocio, podrían ser relevantes las siguientes métricas de desempeño para un sistema de análisis de sentimiento de reseñas de películas:

- **Ahorro de tiempo:** Reducción del tiempo necesario para analizar las reseñas de películas utilizando el aprendizaje automático en comparación con los métodos manuales.
- **Eficiencia de costos:** Reducción de los costos operativos lograda mediante el uso de modelos de aprendizaje automático en el proceso de análisis de sentimiento.
- **Rendimiento/Tiempo de ejecución:** Número de reseñas que se pueden procesar dentro de un período de tiempo determinado, lo que indica la velocidad del análisis.

Referencias y resultados previos

1. IMDB Dataset of 50K Movie Reviews - Kaggle, accessed April 7, 2025, https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
2. Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory - ResearchGate, accessed April 7, 2025, https://www.researchgate.net/profile/Saeed-Qaisar/publication/346511493_Sentiment_Analysis_of_IMDb_Movie_Reviews_Using_Long_Short-Term_Memory/links/626174a8bca601538b5cd022/Sentiment-Analysis-of-IMDb-Movie-Reviews-Using-Long-Short-Term-Memory.pdf
3. LSTMs explained with sentiment analysis on IMDb reviews | by Sarmadkhattak - Medium, accessed April 7, 2025, https://medium.com/@sarmadkhattak858/lstms-explained-with-sentiment-analysis-on-imdb-reviews-a4ccd592d856