Análisis de Sentimientos en Reseñas de IMDb usando LSTM

Andrés Castellano y Joshua Sancho

I. Introducción

Las reseñas de películas en línea contienen información valiosa sobre las opiniones de los usuarios, pero analizarlas computacionalmente es costoso y difícil debido a la ambigüedad y complejidad del lenguaje. Este proyecto aborda el problema de clasificar automáticamente el sentimiento (positivo o negativo) en reseñas de películas utilizando técnicas de procesamiento de lenguaje natural y aprendizaje profundo [1].

Para ello, se implementó una red neuronal recurrente (RNN) de tipo Long Short-Term Memory (LSTM), arquitectura adecuada para datos secuenciales por su capacidad de retener un contexto a lo largo del tiempo. Además, se compararon dos métodos de representación del texto antes del entrenamiento: embeddings estáticos (SpaCy) y embeddings contextuales (BERT) [2]. Esta comparación permitió evaluar la influencia de la complejidad de la representación de los textos en el desempeño del modelo.

El proyecto demuestra que una LSTM junto a embeddings adecuados pueden clasificar sentimientos con una alta precisión, contribuyendo a aplicaciones como sistemas de recomendación y análisis automático de opiniones.

II. Metodología

El estudio se enfocó en comparar dos enfoques de generación de embeddings para reseñas de películas: SpaCy (vectores estáticos) y BERT (vectores contextuales), con el objetivo de analizar cómo la representación del texto impacta la clasificación de sentimientos. Este diseño permite evaluar de manera controlada la ventaja de los embeddings contextuales frente a los estáticos dentro de la misma infraestructura experimental.

Se empleó un dataset de reseñas de IMDb en formato CSV, donde cada fila contenía un enunciado y su categoría (positiva o negativa). Antes del análisis, los datos fueron depurados eliminando registros incompletos o duplicados, y luego divididos en conjuntos balanceados: 70% para entrenamiento, 20% para validación y 10% para prueba.

Los textos se normalizaron, filtraron, limpiaron y tokenizaron. Posteriormente se generaron los embeddings según cada enfoque: SpaCy generó embeddings de 300 dimensiones, ignorando stopwords y signos de puntuación; BERT generó vectores de 768 dimensiones limitando el número de textos a 19000 por restricciones de memoria. Ambos modelos procesaron un máximo de 256 tokens por texto.

El clasificador consiste en una LSTM bidireccional de dos capas con 64 unidades por dirección (128 capas en total). Los últimos 2 estados ocultos se concatenan formando un vector de 128 dimensiones, que pasa por una capa fully conected `Linear(128 → 1)` para generar un logit de salida. La función de pérdida utilizada fue `BCEWithLogitsLoss` y se optimizó con SGD durante 100 épocas con una tasa de aprendizaje de 1e-3. Esta arquitectura permite capturar dependencias contextuales en las secuencias, ofreciendo una comparación justa entre los dos tipos de embeddings.

El procedimiento experimental incluyó preprocesamiento de textos, generación de embeddings, entrenamiento de la LSTM sobre el conjunto de entrenamiento, evaluación periódica en el conjunto de validación para ajuste de hiperparámetros y medición del desempeño final en el conjunto de prueba mediante accuracy, precision, recall y F1-score. Este diseño asegura que las diferencias en desempeño se deban únicamente al tipo de embeddings.

A diferencia de enfoques previos que comparan embeddings de forma aislada o en arquitecturas distintas, este estudio mantiene la misma red neuronal, permitiendo una comparación directa y precisa. La elección de una arquitectura de tipo LSTM bidireccional permite capturar relaciones contextuales hacia atrás y hacia adelante a lo largo de secuencias largas de texto.

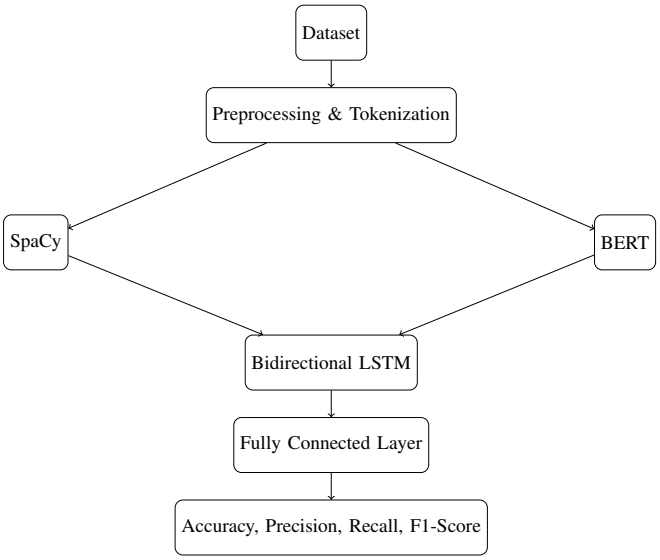

Fig. 1. Diagrama de flujo del procedimiento experimental para la clasificación de sentimientos.

III. Resultados

Los experimentos realizados en este estudio tuvieron como objetivo evaluar el desempeño de una red neuronal LSTM bidireccional entrenada sobre reseñas de películas, utilizando dos enfoques distintos de embeddings: estáticos (SpaCy) y

conxtuales (BERT). La evaluación se centró en demostrar que el uso de embeddings contextuales puede mejorar la clasificación de sentimientos y comparar cuantitativamente el desempeño de cada enfoque mediante métricas estándar de clasificación.

TABLE I
DESEMPEÑO DE LA LSTM BIDIRECCIONAL CON EMBEDDINGS ESTÁTICOS Y CONTEXTUALES

Embeddings	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SpaCy	87.11	86.02	88.70	87.34
BERT	89.32	90.31	88.13	89.21

Los resultados del modelo entrenado con embeddings de BERT indican que el modelo captura de manera equilibrada las clases positiva y negativa, logrando una alta precisión en la predicción de reseñas positivas, con un leve margen de error. Las métricas calculadas permiten evidenciar que BERT logra representar de manera efectiva las relaciones contextuales entre palabras, lo que se refleja en un desempeño consistente y confiable en la clasificación de sentimientos, incluso si este fue entrenado con muchos menos ejemplos, lo que quiere decir que es muy probable que, aumentando la ventana de contexto y el número de ejemplos, el modelo alcance un mayor desempeño.

Por su parte, aunque no hubo una diferencia tan alta en la precisión con el modelo entrenado con embeddings de SpaCy (<5%), este fue alimentado con todos los ejemplos del dataset, lo que implica que lo único que podría mejorar el desempeño es aumentar el límite de tamaño de los textos para capturar patrones más largos.

Las comparaciones cuantitativas muestran que el enfoque con BERT supera consistentemente a SpaCy en métricas clave, demostrando la ventaja de utilizar embeddings contextuales que capturan información semántica más rica. En términos prácticos, esto significa que un sistema de recomendación o análisis de opiniones basado en BERT tendría mayor confiabilidad al interpretar el sentimiento de los usuarios, minimizando errores de clasificación que podrían afectar la personalización de contenidos.

Entre las limitaciones del enfoque, se destaca que a ninguno de los modelos se le realizó fine-tuning, por lo que aún podría mejorar mediante el ajuste de sus pesos a la tarea específica de clasificación de reseñas. Además, el límite de secuencias a 256 tokens puede generar pérdida de información en reseñas más largas. Para mejoras futuras, se sugiere explorar estrategias de aumento de datos, fine-tuning de embeddings, o arquitecturas de tipo Transformer para capturar mejor las dependencias de largo alcance sin incrementar excesivamente el costo computacional y de memoria.

En resumen, los resultados confirman que la metodología propuesta logra clasificar de manera efectiva el sentimiento en reseñas de películas, con un desempeño superior al usar embeddings contextuales de BERT frente a embeddings estáticos de SpaCy, cumpliendo el objetivo del proyecto y proporcionando información práctica sobre la efectividad de diferentes estrategias de representación textual.

IV. DISCUSIÓN

Los resultados obtenidos muestran que la elección de embeddings tiene un impacto significativo en el desempeño de la red LSTM para clasificación de sentimientos. El modelo con BERT superó levemente a SpaCy, lo que evidencia que las representaciones contextuales capturan mejor relaciones semánticas y dependencias en el texto. Este hallazgo concuerda con estudios previos que muestran la ventaja del uso de embeddings generados por modelos de tipo Transformer frente a embeddings estáticos en tareas de análisis de sentimiento.

El éxito del enfoque propuesto refleja que los embeddings generados con BERT le permiten al modelo diferenciar matices y expresiones más complejas. Esto tiene implicaciones prácticas importantes para sistemas de recomendación y análisis de opiniones, donde la precisión en la interpretación de sentimientos mejora la personalización de contenidos.

Entre las limitaciones destacan el tamaño del dataset, la complejidad y ambigüedad del lenguaje, la limitaciones de memoria y costo computacional, y la sensibilidad a hiperparámetros. Además, posibles sesgos en el contenido de las reseñas podrían afectar el desempeño del modelo. Como sugerencias futuras, se propone realizar fine-tuning a los modelos de embeddings para análisis de sentimientos, aumentar la ventana de contexto más allá de 256 tokens, el uso de arquitecturas de tipo Transformer y el entrenamiento y evaluación en dominios o idiomas distintos para mejorar la generalización y robustez del modelo.

En síntesis, la metodología propuesta demuestra que los embeddings contextuales combinados con arquitecturas de tipo LSTM permiten clasificar eficazmente el sentimiento en reseñas de películas, mostrando ventajas claras sobre embeddings estáticos y ofreciendo una base sólida para mejoras futuras en análisis de opinión y sistemas de recomendación.

REFERENCES

[1] Bo Pang and Lillian Lee, *Opinion Mining and Sentiment Analysis*, Foundations and Trends in Information Retrieval, 2008.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL-HLT, 2019.