Análisis de Sentimiento con Reseñas de IMDB

Dora Alicia Guevara Villalpando
Matrícula: 1551003

Universidad Autónoma de Nuevo León)
Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

dora.guevaravll@uanl.edu.mx

***Resumen*—Este documento presenta un análisis de sentimientos en reseñas de películas utilizando el modelo VADER. Se preprocesa el texto mediante limpieza, lematización y eliminación de stopwords. Se comparan las predicciones del modelo con etiquetas originales para evaluar su desempeño. Los resultados indican que VADER es eficiente en la clasificación de sentimientos, pero presenta limitaciones en textos con lenguaje figurado o sarcasmo.**

I. Introducción

El análisis de sentimientos es una técnica clave en el procesamiento de lenguaje natural (NLP) que permite determinar la polaridad emocional de un texto, clasificándolo en positivo, negativo o neutro. Este tipo de análisis es ampliamente utilizado en la minería de opiniones para evaluar reseñas de productos, películas y servicios.

En este proyecto, se analiza un conjunto de datos de reseñas de películas de **IMDb** utilizando la herramienta VADER (Valence Aware Dictionary and sEntiment Reasoner), que se basa en un diccionario de palabras con asignación de polaridad para identificar la intensidad del sentimiento. El objetivo principal es comparar la precisión del análisis de sentimientos automático con las etiquetas de clasificación originales del conjunto de datos.

El informe presenta el procesamiento de los datos, la metodología utilizada, los resultados obtenidos y las conclusiones basadas en el rendimiento del modelo de análisis de sentimientos.

II. Descripción del Conjunto de Datos

El conjunto de datos utilizado en este estudio es el **Large Movie Review Dataset v1.0**, desarrollado por Andrew L. Maas et al. (2011) y publicado en la conferencia **ACL-HLT 2011**. Este dataset es ampliamente utilizado como referencia en tareas de clasificación de sentimientos y contiene reseñas de películas etiquetadas con polaridad positiva o negativa.

II-1. Características del dataset:

- Contiene 50,000 reseñas etiquetadas de manera balanceada en 25,000 positivas y 25,000 negativas. Se divide en:
  - 25,000 reseñas de entrenamiento (12,500 positivas y 12,500 negativas).
  - 25,000 reseñas de prueba (12,500 positivas y 12,500 negativas).
- Cada reseña proviene de una película diferente para evitar sesgos de correlación.
- Solo se incluyen reseñas con calificaciones extremas:
  - Positivas: Puntuaciones ≥ 7 en IMDb.
  - Negativas: Puntuaciones ≤ 4 en IMDb.
  - Las reseñas con calificaciones neutrales no están incluidas en los conjuntos de entrenamiento y prueba.
- Se incluyen 50,000 reseñas sin etiquetas para estudios de aprendizaje no supervisado.

El dataset está estructurado en carpetas train/ y test/, con subdirectorios pos/ y neg/ que contienen los archivos de texto con las reseñas y sus respectivas etiquetas.

III. Planteamiento del Problema

El análisis de reseñas de películas proporciona información valiosa sobre la percepción del público y su experiencia con una producción cinematográfica. Sin embargo, el proceso manual de clasificación es costoso y subjetivo, lo que resalta la necesidad de utilizar modelos automáticos de análisis de sentimientos.

Este estudio busca responder la siguiente pregunta:

¿Qué tan precisa es la herramienta VADER en la clasificación de reseñas de películas en comparación con las etiquetas originales?

Para responder estas preguntas, se implementa un pipeline de preprocesamiento de texto y se comparan los resultados del análisis de sentimientos automático con las etiquetas originales del dataset.

IV. Metodología

IV-A. Carga del Conjunto de Datos

Se descargó el dataset *Large Movie Review Dataset v1.0* desde la fuente oficial y se almacenó en un DataFrame con dos columnas: **review** (texto) y **sentiment** (etiqueta original: 1 = positivo, 0 = negativo).

IV-B. Preprocesamiento del Texto

Los pasos de limpieza incluyen:

1. Eliminación de etiquetas HTML con *BeautifulSoup*.
2. Eliminación de caracteres no alfanuméricos.
3. Tokenización y conversión a minúsculas.
4. Lematización con WordNet.
5. Eliminación de *stopwords* en inglés.

IV-C. Análisis de Frecuencia de Palabras

Se realizó un análisis para obtener las palabras más frecuentes utilizadas en los reviews después del preprocesamiento del texto.

IV-D. Análisis de Sentimientos

Se utilizó el modelo VADER para clasificar cada reseña:

- **Positivo**: $compound \geq 0$
- **Negativo**: $compound < 0$

Se almacenó el resultado en la columna `predicted_sentiment`

IV-E. Evaluación del Modelo

Para comparar las etiquetas originales con las predicciones del modelo, se usaron:

- **Matriz de Confusión**: Muestra aciertos y errores de clasificación.
- **Precisión Global**: Porcentaje de predicciones correctas.
- **Reporte de Clasificación**: Análisis de precisión, recall y F1-score.

V. Análisis de Resultados

V-A. Frecuencia de palabras

Se obtuvo el top 20 de las palabras con mayor frecuencia entre todos los reviews. En este se observó que las palabras más usadas en los reviews son: `movie`, `film`, `one`, `like`, `time`, `good`, entre otras.

En la Figura **??** se observa el listado completo del top 20 de palabras más usadas,

V-B. Precisión del Modelo

- **Precisión Global**: 66.96 % de aciertos.
- **Negativo**:
  - Precisión: 79 %
  - Recall: 46 %
- **Positivo**:
  - Precisión: 62 %
  - Recall: 88 %

V-C. Matriz de Confusión

Acorde a los resultados mostrados en la Figura **??** podemos decir lo siguiente:

- El modelo cometió más errores al clasificar reseñas negativas como positivas.
- La ausencia de una categoría "Neutro" en el dataset original pudo afectar la precisión.

V-D. Análisis de Discrepancias

Se encontraron discrepancias en reseñas con:

- **Sarcasmo**: "I love wasting my time watching this."
- **Doble sentido**: "This movie was **too good** to be true."
- **Expresiones ambiguas**: "Not bad at all" fue clasificado erróneamente como negativo.

VI. Conclusiones

- **VADER es eficiente en la clasificación de sentimientos**, pero presenta errores en textos con subjetividad o sarcasmo.
- **El modelo tiene un sesgo hacia lo positivo**, lo que genera falsos positivos en reseñas negativas.
- **El preprocesamiento es clave**, pero la eliminación de stopwords puede afectar frases con negaciones sutiles.
- **Se recomienda el uso de modelos más avanzados**, como BERT o transformers, para mejorar la precisión en textos con lenguaje complejo.

Referencias

[1] Maas, A. et al. (2011). *Learning Word Vectors for Sentiment Analysis*. ACL-HLT 2011. URL: http://www.aclweb.org/anthology/P11-1015.

[2] Potts, C. (2011). *On the negativity of negation*. Semantics and Linguistic Theory 20, 636-659.

[3] IMDB Movie Reviews Dataset. URL: https://ai.stanford.edu/~amaas/data/sentiment/.

[4] Scikit-learn Library (TF-IDF, Logistic Regression, SVM, Random Forest). URL: https://scikit-learn.org/.

[5] Análisis de texto (text mining) con Python. URL: https://cienciadedatos.net/documentos/py25-text-mining-python.

VII. Anexo: Gráficos.

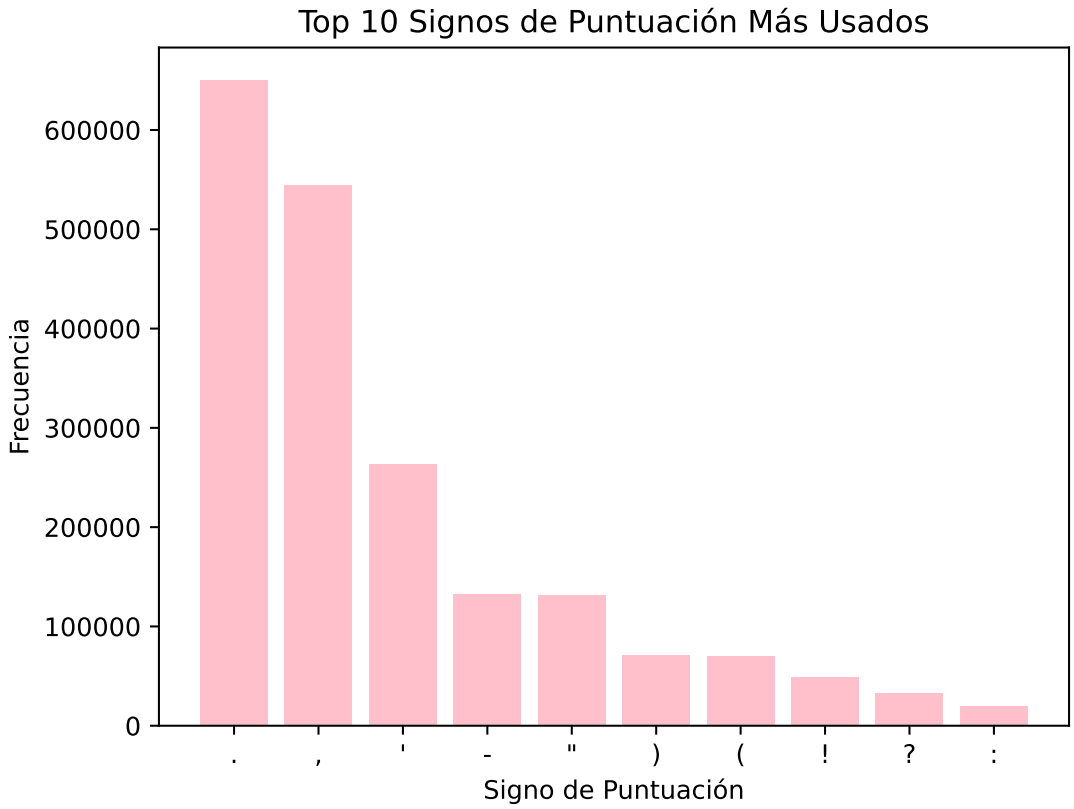

Figura 1. Gráfico del top 10 de los signos de puntuación mas usados.

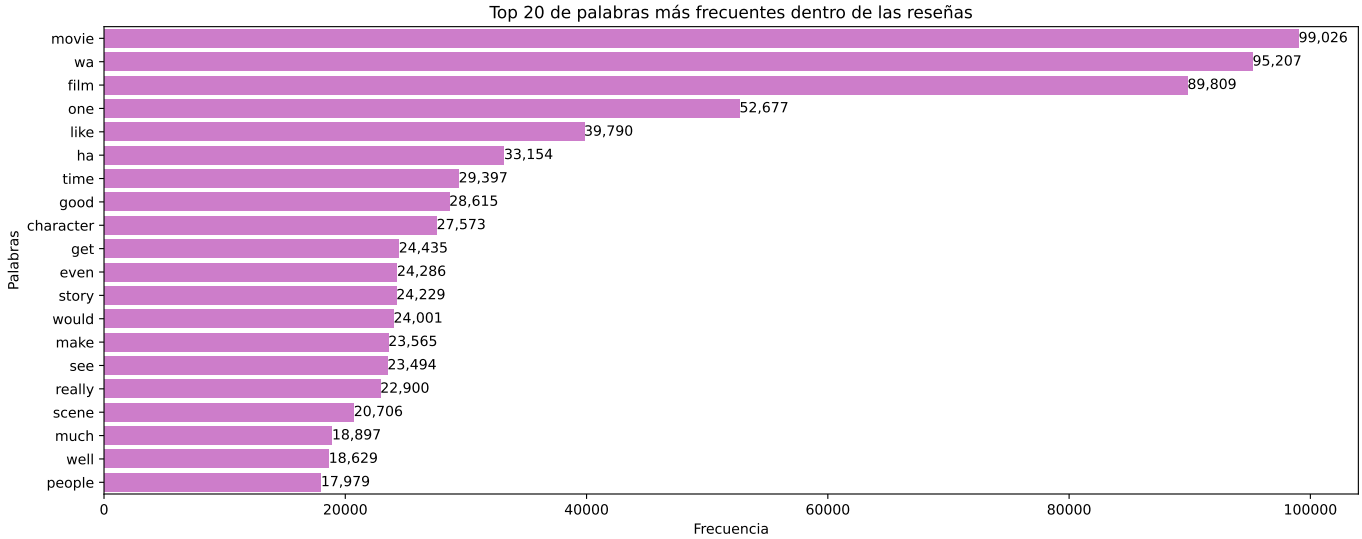

Figura 2. Gráfico del top 20 palabras mas frecuentes.

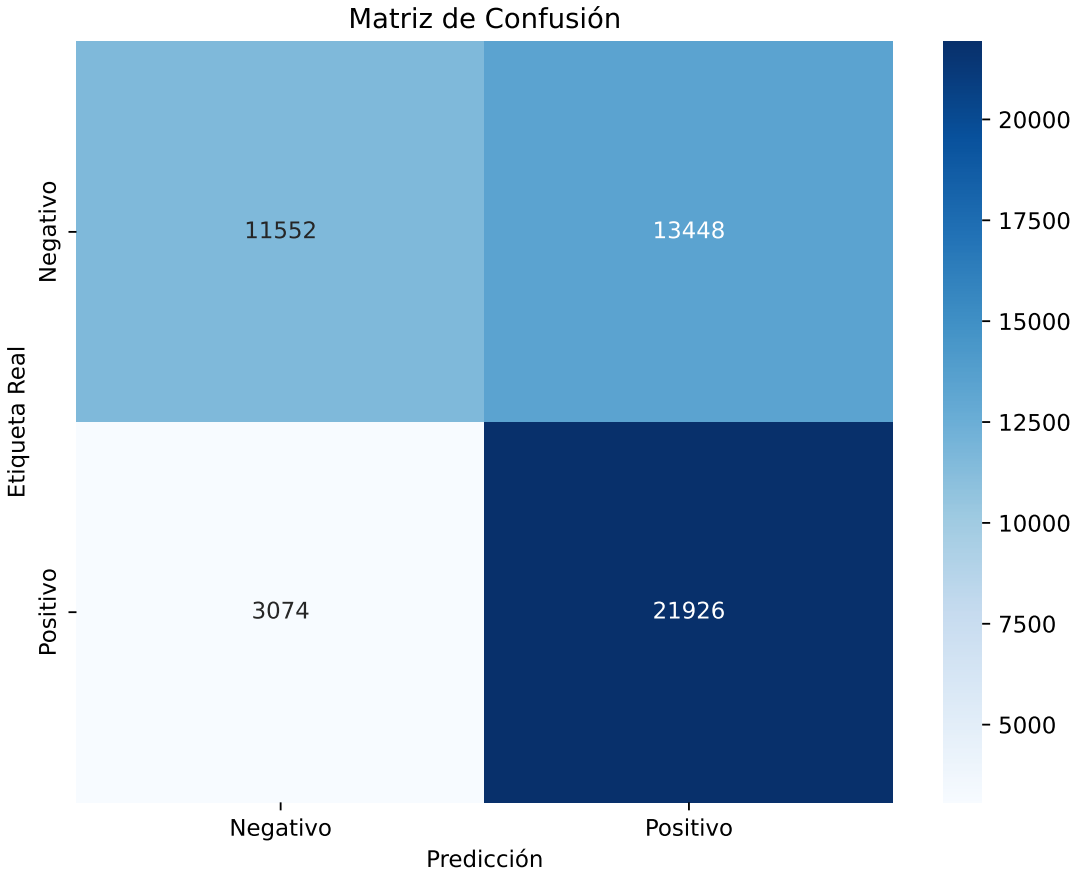

Figura 3. Matriz de Confusión.