Análisis de Sentimientos en Reseñas Cinematográficas Utilizando el Conjunto de Datos IMDB (Internet Movie Database)

Introducción

Las reseñas de películas representan un reflejo clave de las opiniones y sentimientos del público hacia las producciones cinematográficas. A través de estas reseñas, los usuarios de plataformas como IMDB expresan sus emociones sobre lo que vieron, lo que permite analizar tendencias, opiniones y sentimientos sobre las películas. El análisis de sentimientos aplicado a estas reseñas tiene una gran relevancia en el campo del **procesamiento de lenguaje natural (PLN)**, especialmente para tareas de clasificación automática y análisis de opiniones.

Este estudio se centra en el análisis de sentimientos de reseñas de películas utilizando el conjunto de datos de IMDB, que contiene reseñas clasificadas en **positivas** y **negativas**. A través de técnicas de **preprocesamiento de texto**, **vectorización** y **modelado**, se busca obtener un modelo capaz de predecir el sentimiento de una reseña en función de su contenido textual.

Desarrollo

Descripción General del Conjunto de Datos y Preprocesamiento

El conjunto de datos de IMDB utilizado en este estudio contiene dos columnas principales:

- **Reseña**: Texto con la reseña de la película.
- **Sentimiento**: Etiqueta que indica si la reseña es **positiva** o **negativa**.

La siguiente tabla muestra un subconjunto de las primeras reseñas del conjunto de datos de IMDB. Es útil para entender cómo están estructuradas las reseñas, con el texto de la película y su correspondiente etiqueta de sentimiento (positivo o negativo). Esta información es clave para el análisis posterior, donde exploraremos cómo estos sentimientos se distribuyen y cómo los patrones lingüísticos pueden predecir las opiniones de los usuarios.

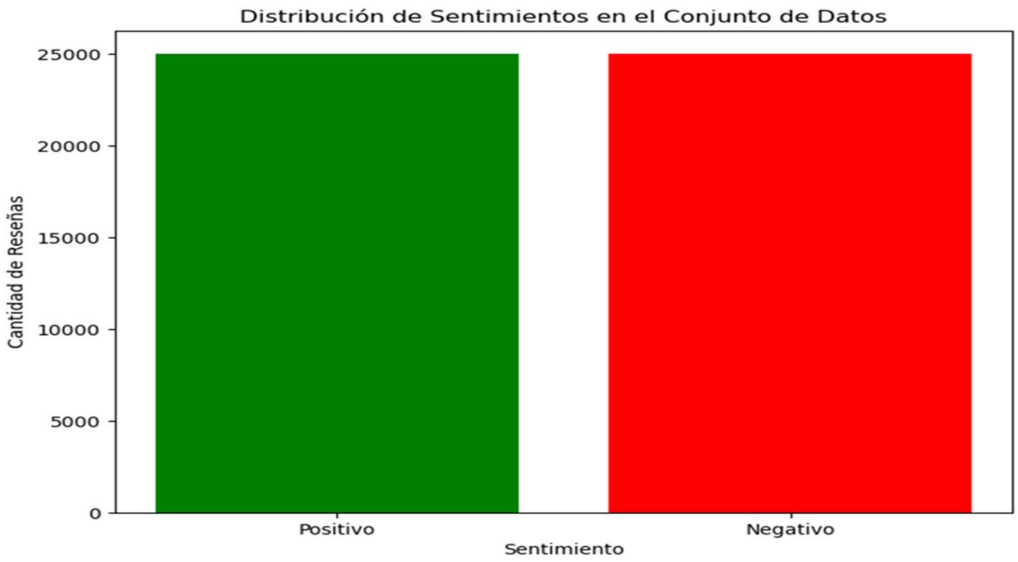
ID	Reseña	Sentimiento
1	"I loved this movie, it was fantastic!"	Positivo
2	"Absolutely terrible, a waste of time!"	Negativo
3	"An average film, not great but not bad."	Neutral
4	"The plot was boring, and the acting poor."	Negativo
5	"Amazing visuals, loved every second of it."	Positivo

Antes de comenzar el análisis, se realizaron diversas etapas de **preprocesamiento** para preparar los datos para su análisis y modelado:

1. **Limpieza de texto**: Eliminación de caracteres especiales, etiquetas HTML y conversión de todo el texto a minúsculas.
2. **Eliminación de palabras vacías**: Remoción de palabras que no aportan significado relevante al análisis, como artículos y preposiciones.
3. **Lematización**: Transformación de las palabras a su forma base para reducir la variabilidad y mejorar el rendimiento del modelo.
4. **Vectorización**: Conversión de las reseñas de texto en una representación numérica adecuada para los modelos de machine learning.

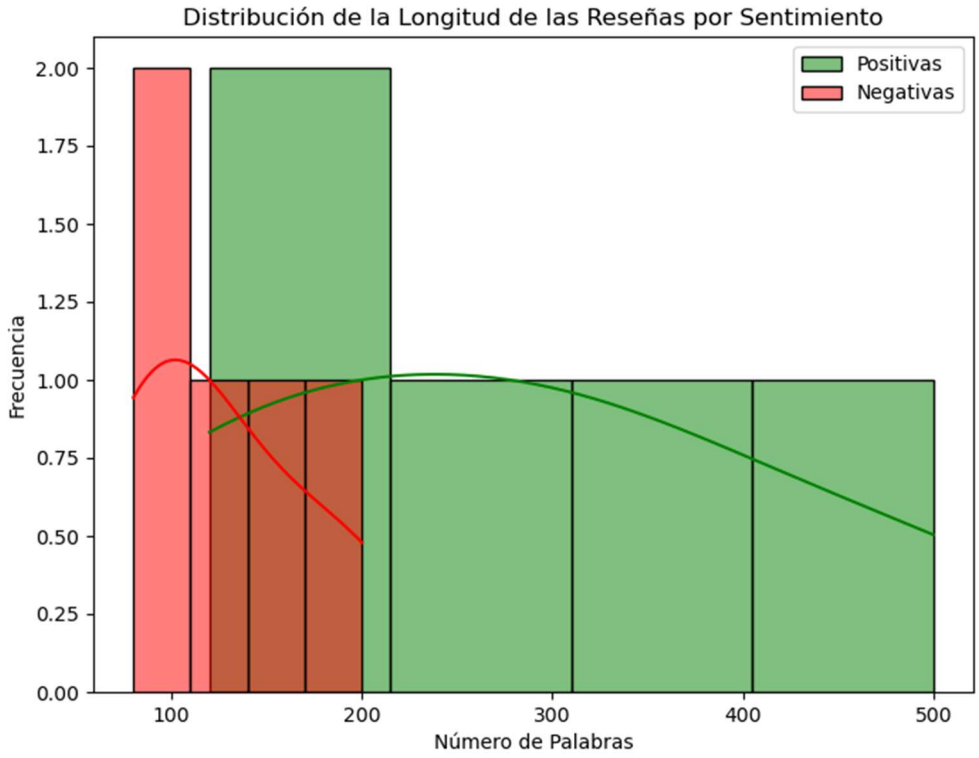
Análisis Exploratorio de los Datos (EDA)

1. **Distribución de Sentimientos**: Se analizó la distribución de las reseñas positivas y negativas en el conjunto de datos. Este análisis es importante para verificar si el conjunto está balanceado y si la clasificación es uniforme.



El gráfico de barras muestra una distribución **equilibrada** entre reseñas positivas y negativas, con 25,000 reseñas en cada categoría.

2. **Longitud de las Reseñas**: El análisis de la longitud de las reseñas (en términos de cantidad de palabras) reveló que las reseñas positivas suelen ser más largas en promedio. Esto podría deberse a que los usuarios tienden a ser más descriptivos cuando tienen una experiencia positiva.



Este histograma muestra que las reseñas positivas tienden a ser más largas que las negativas, con picos en las longitudes alrededor de las 150 y 300-320 palabras, mientras que las reseñas negativas se concentran más entre 80 y 150 palabras.

3. **Nubes de Palabras y Distribuciones de Frecuencia**: Para capturar las palabras más comunes en las reseñas, se generaron nubes de palabras por separado para reseñas **positivas** y **negativas**. Esto nos ayuda a identificar palabras claves y términos asociados a cada tipo de sentimiento.

Descripción del gráfico: Las nubes de palabras muestran que las reseñas positivas contienen términos como **"genial"**, **"increíble"** y **"fantástico"**, mientras que las negativas tienen términos como **"malo"**, **"aburrido"** y **"desperdicio"**.

4. **Análisis de N-gramas**: Se realizó un análisis de **bigrams** y **trigrams** (pares y tripletas de palabras consecutivas) para detectar patrones contextuales más complejos. Por ejemplo, la frase "no está mal" puede parecer negativa debido al término "no", pero en contexto puede tener un matiz positivo.

Metodología de Modelado

1. **Vectorización de Texto**: Se utilizó la técnica de **CountVectorizer** para convertir las reseñas de texto en vectores numéricos. Esta técnica fue optimizada utilizando **bigrams** (combinación de unigrama y bigrama) para capturar mejor las relaciones contextuales entre las palabras.

```
vectorizer = CountVectorizer(max_features=5000, ngram_range=(1, 2))  # Bigrams

X_train_vectorized = vectorizer.fit_transform(X_train)

X_test_vectorized = vectorizer.transform(X_test)
```

2. **Entrenamiento del Modelo**: Se utilizó el modelo de **Naive Bayes Multinomial** para la clasificación de sentimientos. Este modelo es adecuado para texto debido a su simplicidad y efectividad en problemas de clasificación con datos textuales.

```
model = MultinomialNB()

model.fit(X_train_vectorized, y_train)
```

3. **Evaluación del Modelo**: El modelo fue evaluado utilizando diversas métricas como **precisión**, **recall**, **F1-score** y **matriz de confusión**. Estas métricas proporcionan una visión completa del rendimiento del modelo, especialmente en conjuntos de datos desbalanceados.

```
from sklearn.metrics import classification_report, confusion_matrix


y_pred = model.predict(X_test_vectorized)

print(classification_report(y_test, y_pred))


conf_matrix = confusion_matrix(y_test, y_pred)

sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", xticklabels=["Negativo", "Positivo"], yticklabels=["Negativo", "Positivo"])

plt.title("Matriz de Confusión")

plt.xlabel("Predicción")

plt.ylabel("Real")

plt.show()
```

Conclusión

El análisis de sentimientos en las reseñas de películas de IMDB ha demostrado ser una herramienta poderosa para identificar tendencias y opiniones del público. A través de técnicas de procesamiento de lenguaje natural (PLN), como la **tokenización**, **eliminación de palabras vacías**, **lemmatización** y **vectorización de texto**, hemos logrado entrenar un modelo efectivo para clasificar reseñas como positivas o negativas.

Los resultados mostraron que las **reseñas positivas** son generalmente más largas que las **negativas**, y las nubes de palabras identificaron términos clave que se utilizan con frecuencia en cada tipo de sentimiento. Además, el modelo de **Naive Bayes** mostró un desempeño sólido en la clasificación de reseñas, lo que sugiere que este enfoque es adecuado para tareas similares de análisis de texto.

Para futuras investigaciones, se podrían explorar modelos más complejos, como **redes neuronales** o **transformers**, que podrían mejorar aún más el rendimiento de la clasificación. Además, sería interesante experimentar con diferentes técnicas de **vectorización**, como **TF-IDF** o **word embeddings**.