Análisis y clasificación de sentimientos de reseñas de peliculas

Cynthia Selene Martínez Espinoza
Matrícula: 1011238
cynthia.martineze@uanl.edu.mx

30 de enero de 2025

Introducción

El cine es un arte, donde se muestran sentimientos, mensajes profundos, etc. Lo cual es atractivo para la sociedad ya que despierta en ella el interés para despertar emociones y generar debates. Constantemente vemos estrenos de películas, de los cuales se realizan comentarios o reseñas sobre ellas, que infiere en el público si lo ve o no lo ve. Por lo que se vuelven relevantes estos comentarios en la industria del cine.

Además de poder conocer acerca de lo que trata una película, también se puede analizar calidad del guión, música, dirección, actuación, duración de la película, etc.

Actualmente con el tema digital está reseñas cobran importancia, existen plataformas donde cualquier persona puede dejar tu reseña y compartir opiniones acerca de la película. *(IMDb, Rotten Tomatoes y Letterboxd)*

Planteamiento del Problema

Este informe documenta el análisis y clasificación de sentimientos aplicado a un conjunto de 50,000 reseñas de películas, el dataset fue tomado de la página de *kaggle*. El análisis incluye tareas de procesamiento de texto, visualización de resultados y evaluación y comparación del desempeño de los modelos de clasificación de texto basado en algoritmos de aprendizaje automático.

Metodología

Recopilación de Datos

Las reseñas se fueron obtuvieron de un conjunto de datos de *kaggle.*

Identificar y clasificar los sentimientos sobre las reseñas de películas, comprender las tendencias de opinión y evaluar el desempeño del modelo Random Forest, Regresión Logística.

Preparación de los Datos

- **Descarte de Palabras**: Se eliminaron palabras comunes irrelevantes utilizando listas de stopwords en español.
- **Tokenización**: Las reseñas fueron divididos en palabras individuales.
- **Stemming**: Se redujeron las palabras a su forma base para unificar variantes.
- **Vectorización**: Se generó una representación numérica de los textos, con ello generando bigramas.
- **Análisis de Sentimientos**: Se agregaron datos como:

  Polaridad Indica el sentimiento general del texto, variando de -1 (negativo) a 1 (positivo).
  Subjetividad: Mide cuán subjetivo es el texto, de 0 (objetivo) a 1 (muy subjetivo).

- **Comparación de modelos**: Se entrenaron los modelos de regresión logística y Random forest, los cuales se utilizan para clasificacion de texto.

Visualizaciones Generadas

- **Nube de Palabras**: Destaca términos como *historia*, *final*, *personaje*, y *hombre*,
- **Gráfico de Dispersión**: Muestra la relación entre polaridad y subjetividad en las reseñas analizadas.
- **Gráfico de Barras**: Representa la distribución de reseñas sobre las películas de positivos, negativos y neutrales.

Resultados

- Destaca los términos más frecuentes en las reseñas. **F**igura 1:
- Para realizar un Análisis de Opiniones usamos la Subjetividad vs Polaridad

  Identificar si los comentarios o reseñas son objetivos o subjetivos.

  Subjetividad:

  Mide qué tan subjetivo o basado en opiniones es el texto.

  Valor cercano a 0: El texto es más objetivo (por ejemplo, .[El] cielo es azul").
  Valor cercano a 1: El texto es más subjetivo (por ejemplo, "Me encanta este producto")

  Clasificación de Texto: Utilizar la subjetividad como una característica para clasificar textos en análisis de sentimiento o temas.

  Detección de Sesgo: Analizar textos de noticias para evaluar el nivel de subjetividad y detectar sesgos.

  Polaridad Determinar si el texto expresa un sentimiento positivo, negativo o neutral. Ejemplo: Calificaciones de productos, opiniones de clientes, de las reseñas.

  Mide si el sentimiento del texto es negativo, neutro o positivo:
  Valor cercano a -1: Sentimiento negativo (por ejemplo, .[Es]to es horrible").
  Valor cercano a 0: Sentimiento neutral (por ejemplo, .[Es] un día").
  Valor cercano a 1: Sentimiento positivo (por ejemplo, "Me encanta este lugar")

  Clasificación de Textos: Utilizar la polaridad como una característica clave en algoritmos de machine learning.

  Tendencias de Opinión: Analizar grandes volúmenes de texto (como tweets o reseñas) para obtener una visión general de las opiniones de los usuarios.
  Recibe un valor de polaridad (score), que representa el análisis previo del sentimiento del texto.

Figura 1: Frecuencia de palabras

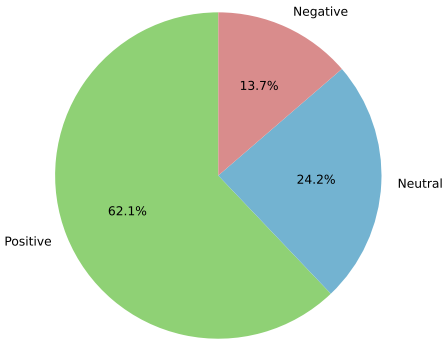

Figura 2: Distribución de Sentimientos

Distribución de Subjetividad

(¿es una opinión o un hecho?).

- **Muy Objetivos (menor a 0.3)**: 14195 reseñas).
- **Parcialmente Objetivos (entre 0.3 y 0.6)**: 20652 reseñas).
- **Muy Subjetivos (mayor 0.6)**: 15153 reseñas).

Distribución de Sentimientos

(¿qué tan positivo o negativo es el sentimiento?).

- **Positivos**: 31055 reseñas (62 %).
- **Negativos**: 6837 reseñas (14 %).
- **Neutrales**: 12108 reseñas (24 %). **F**igura 2: Distribución de sentimientos.

Cada punto en el gráfico representa un texto analizado, y su posición depende de los valores de polaridad y subjetividad asociados. **F**igura 3: Dispersion Polaridad Subjetividad

Mide el sentimiento: Eje X (Polaridad)
Izquierda (-1): Negativo.
Centro (0): Neutro.
Derecha (1): Positivo.

Mide la subjetividad: Eje Y (Subjetividad):
Inferior (0): Objetivo (hechos).
Superior (1): Subjetivo (opiniones).

La mayoría de los textos analizados son positios en términos de polaridad, pero con un grado moderado de subjetividad son parcialmente objetivos.

Existe una cantidad considerable de textos positivos y neutros, pero su proporción es menor en comparación con los negativos. La distribución de subjetividad sugiere que los textos contienen tanto opiniones como referencias a hechos, las opiniones y hechos estan equilibrados.

Evaluación del Modelo

Regresión Logística en Clasificación de Texto

- Precisión más alta (0.82)
- Mejor cuando las características del texto están bien representadas numéricamente (TF-IDF, embeddings, etc.).
- Menos propenso al sobreajuste en conjuntos de datos pequeños.

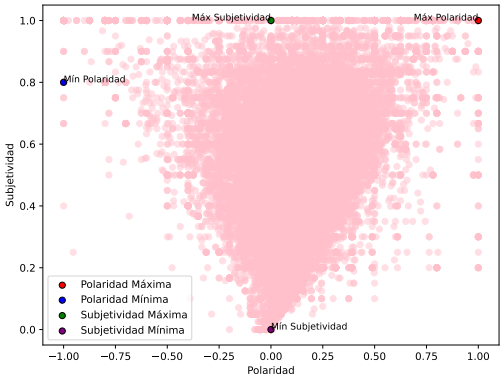

Figura 3: Dispersion Sentimiento / (hecho /opinion)Subjetividad (hecho /opinion)

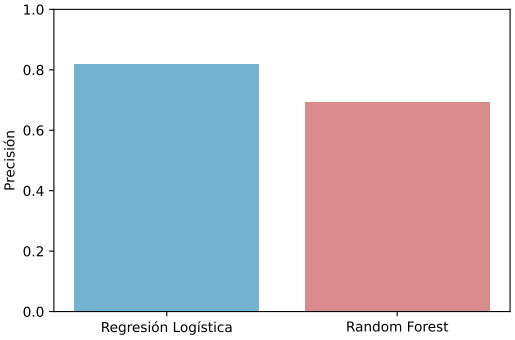

Figura 4: Comparacion de modelos

-

Random Forest en Clasificación de Texto

- Precisión menor (0.70)
- Modelo basado en árboles de decisión: Construye múltiples árboles y promedia sus predicciones.
- Mejor cuando las relaciones entre palabras son no lineales y se necesitan interacciones entre características.
- Puede verse afectado por datos de alta dimensión como los generados por TfidfVectorizer.
- Más robusto contra valores atípicos pero propenso a sobreajustarse en conjuntos pequeños.

Conclusiones

- Regresión Logística funciona bien en problemas de texto donde la relación entre palabras y la clase objetivo es lineal.

- Random Forest es útil para datos con interacciones complejas, pero en clasificación de texto, donde TF-IDF ya captura relaciones clave, puede no ser la mejor opción.

  En análisis de texto, Regresión Logística suele ser mejor que Random Forest, especialmente cuando usamos TF-IDF, porque el problema es lineal y no requiere la complejidad de modelos de árboles.