ITESO, Universidad Jesuita de Guadalajara

Proyecto No. 1 - N-Grams & Sentiment Analysis

Rafael Takata García - 750625

Roi Jared Garza Stone - 752696

Esteban Gómez Valerio -

Ingeniería y Ciencia de datos

Minería de textos

Ing. Juan Antonio Vega Fernández, M. Sc., M. T. Ed

O2025_MAF3654H

28/09/2025

Reporte - Análisis de sentimientos con n-gramas

1. Introducción:

El análisis de sentimientos es una rama del procesamiento de lenguaje natural (NLP, por sus siglas en inglés) que busca identificar y clasificar las opiniones expresadas en textos, como positivas, negativas o neutras. En este proyecto, se llevó a cabo un análisis de sentimientos sobre el conjunto de datos de reseñas de películas de IMDB, el cual contiene miles de opiniones de usuarios que reflejan sus valoraciones subjetivas sobre distintas producciones cinematográficas.

2. Metodología

El proyecto aplicó un enfoque de **aprendizaje automático tradicional** (Machine Learning) a datos textuales:

- **Preprocesamiento**: Se tokenizaron las reseñas, se convirtieron a minúsculas y se eliminaron la puntuación y las *stop words*.
- **Extracción de Características**: Se utilizó **CountVectorizer** para transformar el texto en un vector numérico, explorando tres rangos de N-gramas:
    - Unigrama (1, 1)
    - Unigrama + **Bigrama (1, 2)**
    - Unigrama + Bigrama + Trigram (1, 3)
- **Modelos**: Se entrenaron y evaluaron tres clasificadores: **Regresión Logística**, **Naive Bayes Multinomial** y **SVM Lineal**.

3. Análisis de Rendimiento (Resultados Clave)

Rango N-grama	Modelo	Accuracy	Precision	Recall	F1-score
Unigrama (1, 1)	Logistic Regression	0.8686	0.8745	0.8606	0.8675
	Naive Bayes	0.8280	0.8689	0.7726	0.8180
	SVM	0.8472	0.8552	0.8361	0.8455
Unigrama+Bigrama (1, 2)	**Logistic Regression**	**0.8896**	0.8911	0.8878	0.8894
	Naive Bayes	0.8540	0.8908	0.8070	0.8468
	SVM	0.8854	0.8874	0.8828	0.8851
Unigrama+Bigrama+Trigram (1, 3)	Logistic Regression	0.8892	0.8883	**0.8902**	**0.8893**
	Naive Bayes	0.8599	**0.8918**	0.8191	0.8539
	SVM	0.8876	0.8874	0.8879	0.8876

El mejor rendimiento se logró con **Regresión Logística,** dominando 3 de las 4 categorías (Accuracy, Recall y F1-Score). Funcionó relativamente mejor con Trigrama y con Bigrama, el otro que tuvo el mayor score fue naive bayes en combinación con Trigrama.

Impacto de los N-gramas:

- La inclusión de **Bigramas** (palabras de dos en dos, ej., "no bueno" o "gran película") resultó en una **mejora sustancial** de la precisión para todos los modelos en comparación con el uso exclusivo de Unigramas.
- Agregar Trigramas no generó una mejora significativa, lo que indica que los **Bigramas fueron el factor clave** para la capacidad predictiva del modelo.

4. Análisis de Errores

El modelo de Regresión Logística (Unigrama+Bigrama) demostró un buen equilibrio en sus predicciones, pero los errores residuales se dividieron principalmente en:

	Predicho 0 (Negativo)	Predicho 1 (Positivo)
Actual 0 (Negativo)	11,109 (TN)	1,391 (FP)
Actual 1 (Positivo)	1,555 (FN)	10,945 (TP)

Estos errores suelen deberse a la incapacidad de los modelos de bolsa de N-gramas para capturar **ironía**, **sentimiento sutil** o **estructuras gramaticales complejas** donde una frase positiva es seguida por una conclusión negativa, y viceversa.

5. Conclusión

La **Regresión Logística con características Unigrama y Bigrama** es la configuración más eficiente para la clasificación de sentimiento en este conjunto de datos.

Se concluye que el modelo es **sólido** y que el enfoque de N-gramas es efectivo para la tarea que se está buscando resolver. Sin embargo, sería bueno explorar métodos más avanzados para encontrar un mejor rendimiento en el problema.