Laboratorio 5
Minería de Textos y Análisis de Sentimiento

Diederich Solis (22952) Sara Guzmán (22097)
https://github.com/DiederichSolis/Lab5_data.git

Agosto 2025

Índice

1. Introducción

El objetivo es aplicar técnicas de **minería de textos** y **análisis de sentimiento** para procesar un conjunto de tweets y determinar si corresponden o no a desastres reales.

El dataset utilizado fue el de Kaggle *"Natural Language Processing with Disaster Tweets"*, que contiene 7,613 registros y 5 columnas: `id`, `keyword`, `location`, `text` y `target`. Este laboratorio se divide en cinco avances principales: descripción del dataset, preprocesamiento, análisis de unigramas, análisis de n-gramas (bigramas y trigramas) y construcción de un modelo preliminar.

2. Parte 1: Descripción del Dataset

El dataset presenta la siguiente estructura:

- Dimensiones: 7,613 filas y 5 columnas.
- Valores nulos: 33.27 % en la columna `location`, 0.8 % en `keyword`, y 0 % en `text` y `target`.
- Distribución de clases: 57 % de los tweets no relacionados con desastres (clase 0) y 43 % relacionados con desastres (clase 1).

Esta distribución muestra un dataset balanceado, aunque con ligera mayor representación de tweets de la clase 0. La columna `location` presenta un alto porcentaje de valores faltantes, por lo que no puede considerarse confiable como variable predictora. En cambio, las columnas `keyword` y especialmente `text` resultan fundamentales para el análisis.

3. Parte 2: Preprocesamiento

El preprocesamiento del texto incluyó:

1. Conversión a minúsculas.
2. Eliminación de URLs, menciones, símbolos y emojis.
3. Conservación de palabras en hashtags (eliminando solo el símbolo #).
4. Tratamiento especial de números relevantes como "911", que fueron preservados como tokens.
5. Eliminación de stopwords en inglés.
6. Lematización para unificar variantes gramaticales.

Ejemplo antes y después:

- **RAW**: *"Need to work in an office I can bang all my fav Future jams out loud"*
- **CLEAN**: *"need work office bang fav future jam loud"*

El texto procesado pierde su estructura gramatical original y ya no es legible como una frase natural. Sin embargo, esto es intencional: **en minería de textos no buscamos generar frases legibles, sino extraer tokens útiles para el análisis estadístico y la clasificación**. Al reducir ruido (stopwords, URLs, símbolos) y unificar formas (lematización), el modelo puede concentrarse en patrones verdaderamente informativos.

4. Parte 3: Unigramas y Visualizaciones

Se calcularon las frecuencias de palabras individuales en cada clase.

- En tweets de desastres (`target = 1`) destacan palabras como: *fire, disaster, california, suicide, police*. Estas palabras reflejan situaciones críticas y asociadas a eventos reales.
- En tweets no relacionados a desastres (`target = 0`) las palabras más frecuentes son: *like, get, new, video, day*. Estas palabras son más genéricas y típicas del lenguaje informal de Twitter.

Gráficas y su interpretación

Wordclouds: son visualizaciones cualitativas que permiten identificar rápidamente los términos más frecuentes. Útiles para *exploración*, pero no sustituyen un análisis cuantitativo.

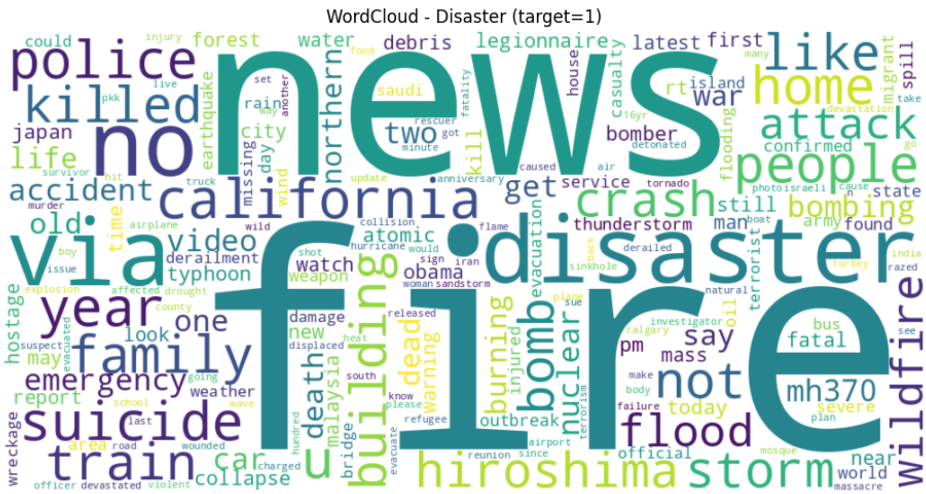

Figura 1: Wordcloud de tweets de desastres.

Figura 2: Wordcloud de tweets no relacionados a desastres.

Barras Top-10: permiten comparar *cantidades* de los términos más frecuentes por clase y son más fieles para análisis cuantitativo. El patrón observado es consistente con distribuciones tipo Zipf, con pocas palabras muy frecuentes y una larga cola de términos raros.

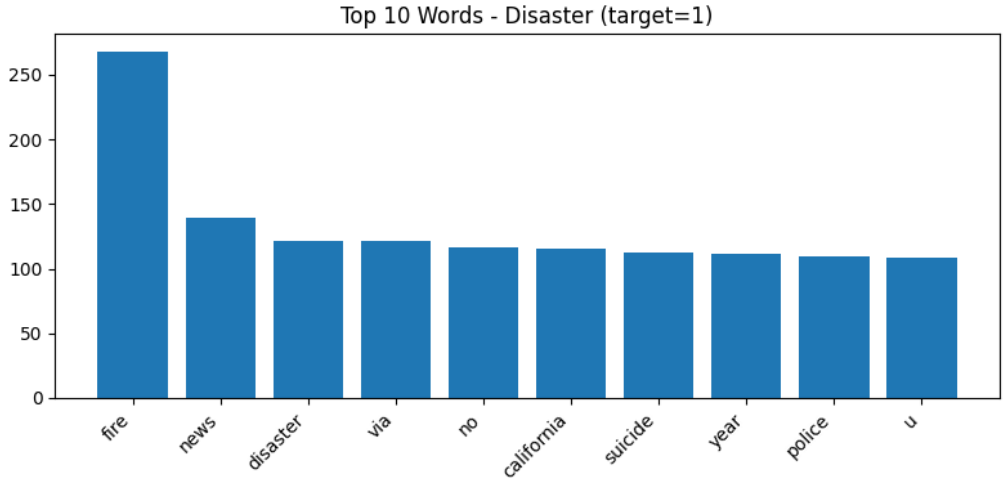

Figura 3: Top-10 unigramas en tweets de desastres (frecuencia absoluta).

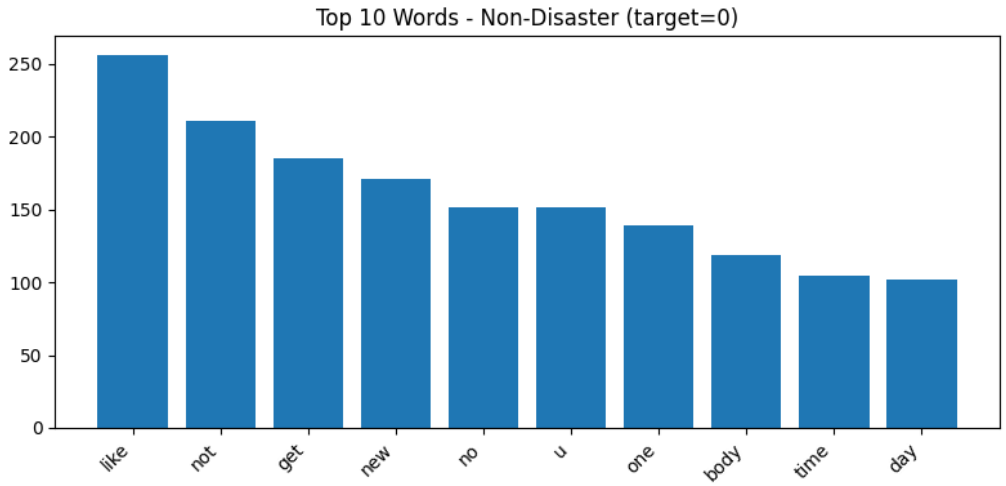

Figura 4: Top-10 unigramas en tweets no relacionados a desastres (frecuencia absoluta).

El contraste entre los unigramas más frecuentes en cada clase es evidencia de que las palabras clave capturan bien la diferencia semántica entre los contextos de desastre y no-desastre. Esto valida que los unigramas son una base sólida para construir clasificadores.

5. Parte 4: Bigramas y Trigramas

El análisis de n-gramas más largos reveló secuencias muy informativas:

- En desastres: bigramas como *"suicide bomber"*, *"oil spill"*, y *"california wildfire"*; trigramas como *"suicide bomber detonated"* y *"northern california wildfire"*.

- En no-desastres: expresiones más coloquiales como *"look like"*, *"feel like"*, o secuencias de uso cotidiano en redes sociales como *"reddit new content"*.

Mientras los unigramas muestran palabras aisladas, los bigramas y trigramas revelan **contexto lingüístico**, lo cual es crucial para mejorar modelos de clasificación. Frases como *suicide bomber* no podrían identificarse solo con palabras individuales.

6. Parte 5: Modelo Preliminar

Se implementó un modelo de clasificación como baseline con las siguientes características:

- División: 80 % entrenamiento, 20 % validación (estratificada).
- Vectorización: TF-IDF con unigramas y bigramas, hasta 20,000 tokens.
- Clasificador: Regresión Logística.

Resultados:

- Accuracy: $\approx 82\,\%$
- F1 macro: $\approx 0{,}81$
- Recall clase 0: 0.90, Recall clase 1: 0.71

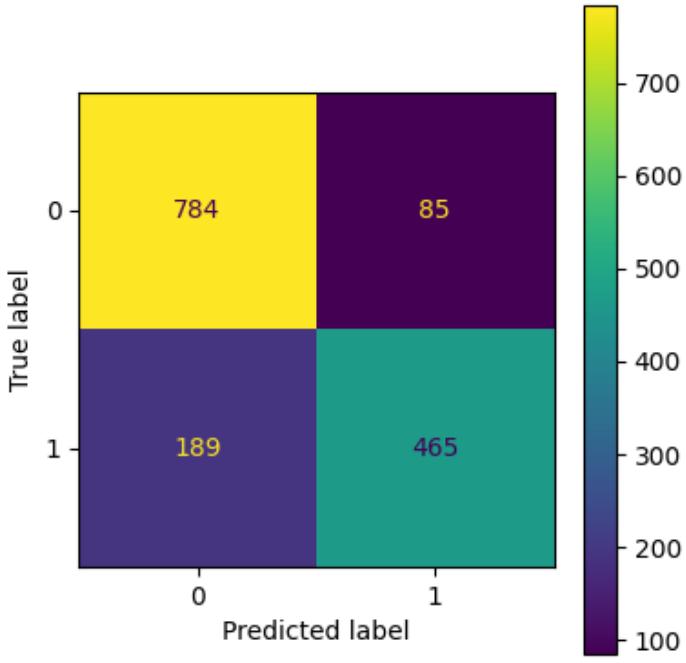

Figura 5: Matriz de confusión del modelo preliminar.

Interpretación de la matriz de confusión: el modelo acierta con mayor frecuencia en la clase 0 (no-desastre) y falla más al identificar algunos tweets de clase 1 (desastre).

Esto se refleja en el *recall* menor para la clase 1 y es típico cuando existen mensajes ambiguos o lenguaje metafórico.

El modelo tiene un buen desempeño inicial, especialmente en la clase 0. No obstante, muestra dificultad para identificar correctamente todos los tweets de desastres (clase 1). Los errores de clasificación revelan que algunos tweets de desastre usan un lenguaje similar al de tweets coloquiales (p. ej., referencias a películas o metáforas).

El modelo baseline es un punto de partida sólido, pero se podrían considerar mejoras como:

- Ajuste de hiperparámetros de la regresión logística.
- Uso de embeddings preentrenados (Word2Vec, GloVe, FastText).
- Modelos más complejos como SVM, Random Forests o redes neuronales.

7. Conclusiones

- El preprocesamiento permitió eliminar ruido y concentrar la información en tokens significativos, aunque las frases resultantes pierden sentido gramatical. Esto es intencional: el objetivo no es la legibilidad, sino la utilidad para análisis y clasificación.
- El análisis de unigramas y n-gramas mostró diferencias claras entre tweets de desastre y no-desastre, confirmando que estas características capturan bien el contexto.
- Las **gráficas** (wordclouds y barras Top-10) facilitaron interpretar visualmente la distribución de términos: los wordclouds ayudaron a la exploración cualitativa y las barras permitieron comparar frecuencias de forma cuantitativa.
- El modelo baseline alcanzó un 82 % de exactitud, evidenciando un desempeño competitivo. Sin embargo, el bajo recall en la clase 1 muestra la necesidad de modelos más sofisticados.