Proyecto de Ingeniería de Datos

IMEDIA Project



Esteban Javier Berumen Nieto

Isabel Valladolid Dillanes

Mayo 2025

### Introducción

Actualmente, las redes sociales se han convertido en uno de los principales espacios donde las personas expresan sus opiniones, emociones y reacciones respecto a eventos sociales, políticos y culturales. Gracias al gran volumen y velocidad de publicación, es complicado extraer información significativa de manera manual. La necesidad de interpretar el sentimiento colectivo y detectar patrones emocionales es cada vez más importante para empresas, medios de comunicación, etc.

“IMEDIA”, busca contribuir a la demanda de herramientas prácticas que permitan analizar la opinión pública en plataformas o redes sociales. IMEDIA proporciona una solución práctica para extraer, clasificar y presentar insights emocionales a partir de textos publicados por los usuarios.

### Objetivo

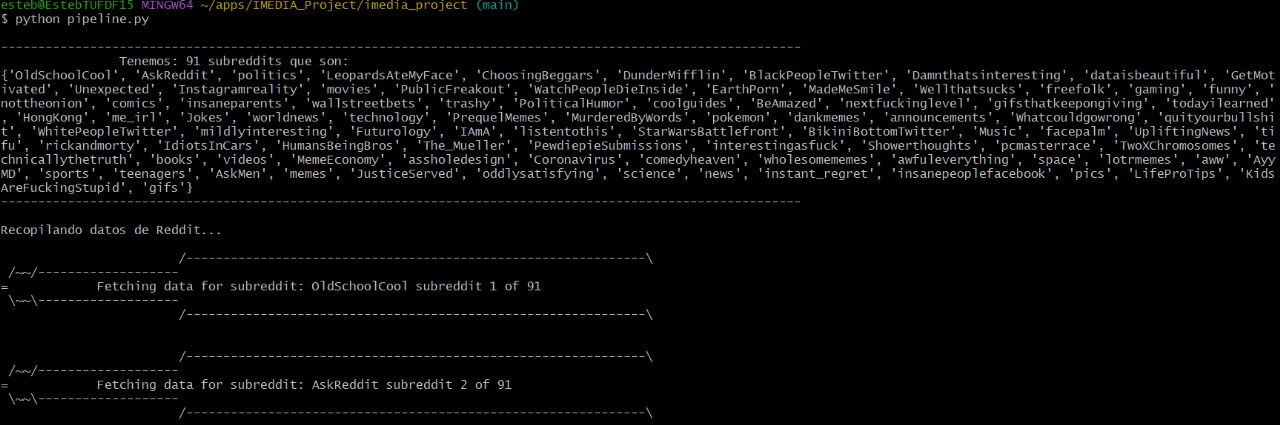
El objetivo de IMEDIA es ser una plataforma de análisis de sentimientos en redes sociales, que integre extracción de datos, procesamiento NLP y visualización interactiva de insights emocionales, que además, permita explorar patrones de opinión en tiempo real a través de un dashboard dinámico y una API de clasificación.

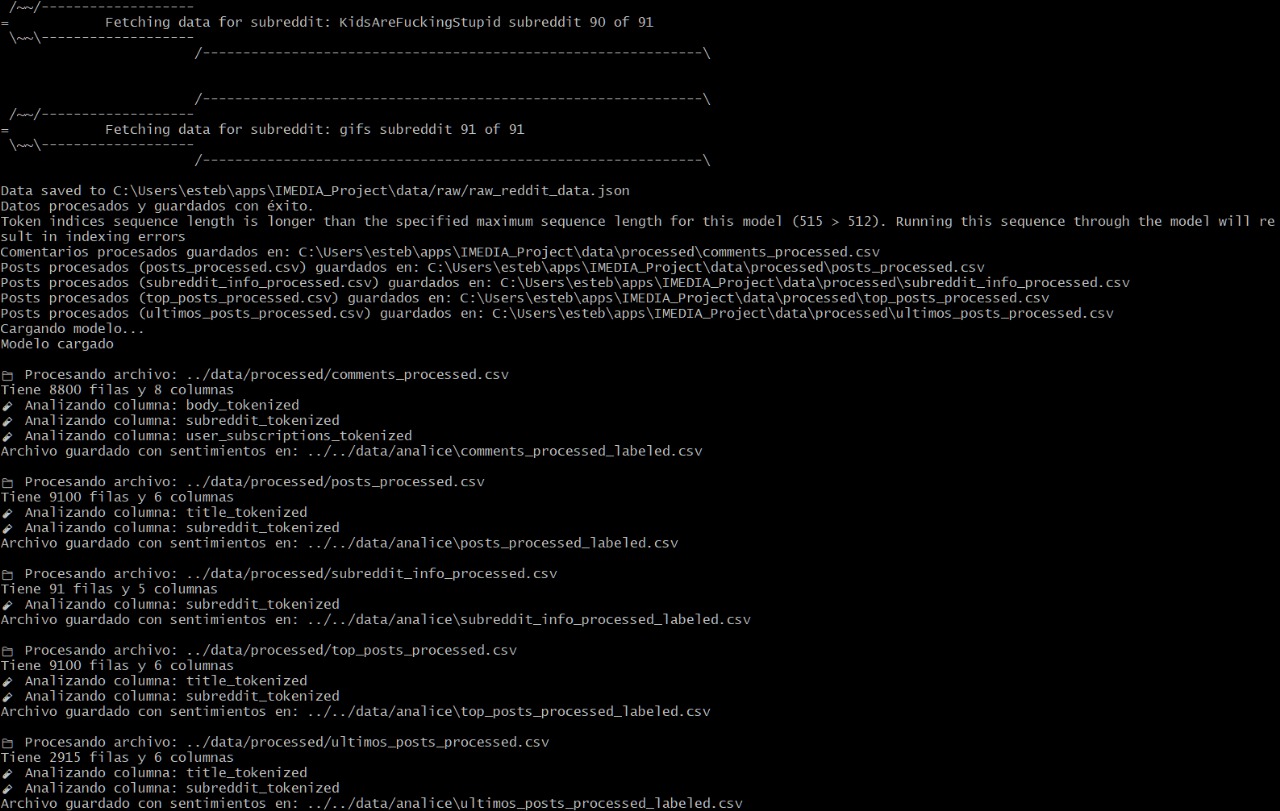
## Metodos

El proyecto realiza la extracción de datos desde subreddits populares, incluyendo publicaciones, comentarios y metadatos, enriquecidos con información de los autores. Los datos en formato JSON se transforman en archivos CSV organizados. Luego, se aplica limpieza textual y tokenización con *bert-base-uncased*. Los textos se clasifican por sentimiento o categoría usando un modelo ajustado previamente, y los resultados se exportan. Finalmente, se visualizan mediante dashboards en Power BI que muestran análisis por subreddit, tipo de publicación, nube de palabras y tendencias temporales.

### Resultados obtenidos

El flujo de datos sigue una estructura organizada: los archivos JSON se almacenan en data/raw/, se transforman en CSVs en data/interim/, luego se limpian y procesan en data/processed/, y finalmente se clasifican en data/analice/. Los resultados se visualizan mediante dashboards interactivos en Power BI, facilitando el análisis y la interpretación de la información extraída.





# Conclusiones

Este proyecto desarrolló un sistema automatizado para extraer y procesar datos de Reddit usando técnicas de NLP. Mediante la API de Reddit, librerías de Python y modelos pre entrenados, se recolectaron y limpiaron publicaciones, clasificándolas por sentimiento y almacenados localmente. La estructura modular del código y el uso de archivos CSV facilitan su reutilización y escalabilidad. Este trabajo demuestra cómo el NLP puede transformar datos textuales en información útil para analizar opiniones y comportamientos en redes sociales.

Bibliografía

*PRAW 7.7.1 documentation*. (s. f.). <https://praw.readthedocs.io/en/stable/>

*Tokenizer*. (s. f.). <https://huggingface.co/docs/transformers/en/main_classes/tokenizer>

Repositorio de github: <https://github.com/XTEP63/IMEDIA_Project.git>