Desarrollo de un Pipeline de Datos para Análisis de Sentimientos sobre Comentarios Relacionados a Gustavo Petro en Redes Sociales (Instagram)

Kamilt Andrés Bejarano Díaz, Jerson Osorio Cely, Daniel Vargas Peña

***Abstract* -- Este proyecto desarrolló un canal de datos para analizar los comentarios de Instagram sobre Gustavo Petro. Utilizando la API Instaloader, los datos se extrajeron, limpiaron y almacenaron en una base de datos SQLite, con múltiples cuentas creadas para manejar los límites de velocidad. El análisis de sentimiento utilizando TextBlob reveló comentarios principalmente neutrales, con períodos de sentimiento positivo y negativo que reflejan cambios en la opinión pública. El proyecto subraya el valor del análisis de redes sociales para comprender la percepción pública y demuestra la eficacia de un ETL estructurado y un proceso de análisis de sentimientos. El trabajo futuro podría ampliarse a otras redes sociales y métodos avanzados de análisis de sentimientos.**

I. Introducción

El rápido crecimiento de las redes sociales ha transformado la forma en que las figuras públicas interactúan con la sociedad, por lo que es crucial analizar el sentimiento público para comprender su impacto. Este proyecto, titulado "Desarrollo de un Pipeline de Datos para Análisis de Sentimientos sobre Comentarios Relacionados con Gustavo Petro en Redes Sociales", se enfoca en extraer, transformar y analizar comentarios relacionados con Gustavo Petro en Instagram. Este documento detalla la arquitectura del sistema, las tecnologías utilizadas, los procesos ETL implementados y la metodología de análisis de sentimientos. Además, presenta los conocimientos obtenidos del análisis de sentimiento.

II. Arquitectura del sistema

La arquitectura del sistema del proyecto está diseñada para manejar eficientemente los procesos de extracción, transformación y carga (ETL), seguidos del análisis de sentimiento. La arquitectura comprende los siguientes componentes:

A. Módulo de extracción de datos

Utiliza la API Instaloader para extraer comentarios de publicaciones de Instagram relacionadas con Gustavo Petro.

B. Módulo de transformación de datos

Procesa y limpia los datos extraídos para garantizar que estén en un formato adecuado para el análisis.

C. Módulo de carga de datos

Carga los datos transformados en una base de datos para realizar consultas y análisis eficientes.

D. Módulo de análisis de sentimiento

Aplica técnicas de procesamiento del lenguaje natural (NLP) para determinar el sentimiento de los comentarios.

E. Módulo de informes

Genera información y visualizaciones basadas en los resultados del análisis de sentimiento.

III. Tecnologías usadas

Para implementar el canal de datos, empleamos varias tecnologías y herramientas:

- API Instaloader: una biblioteca de Python utilizada para extraer datos de Instagram. Permite la extracción de comentarios en formato JSON.
- Python: el lenguaje de programación principal para crear secuencias de comandos de procesos ETL y análisis de sentimientos.
- Pandas: una biblioteca de Python utilizada para la manipulación y transformación de datos.
- NLTK y TextBlob: bibliotecas de Python para procesamiento de lenguaje natural y análisis de sentimientos.
- SQLite: una base de datos liviana que se utiliza para almacenar los datos transformados.
- Matplotlib y Seaborn: bibliotecas de Python para visualización de datos.Use SI (MKS) o CGS como unidades primarias. (Las unidades SI son las recomendadas) Las unidades inglesas pueden ser utilizadas como secundarias (entre paréntesis). Una excepción podría ser el uso de unidades inglesas como

un identificador comercial, tal como “disco de 3,5 pulgadas”.

IV. Proceso ETL (Extract, Transform, Load)

A. *Extracción de datos*

El proceso de extracción de datos implicó el uso de la API Instaloader para extraer comentarios de las publicaciones de Instagram. Inicialmente se consideró Apify, pero no proporcionó los recursos necesarios, lo que llevó a la adopción de Instaloader. Debido a los límites de tarifas de Instagram, se crearon varias cuentas de Instagram para garantizar la extracción continua de datos sin interrupciones. La API Instaloader recuperó datos en formato JSON, que incluían metadatos sobre las publicaciones y los comentarios mismos.

B. *Transformación de datos*

Una vez extraídos los datos, se sometieron a una serie de pasos de transformación para limpiarlos y estandarizarlos. Esto incluyó eliminar duplicados, manejar valores faltantes y normalizar datos de texto (por ejemplo, convertir a minúsculas, eliminar caracteres especiales). Luego, los datos transformados se estructuraron en un formato tabular adecuado para cargarlos en una base de datos.

C. *Carga de datos*

Los datos limpios y estructurados se cargaron en una base de datos SQLite. Se eligió SQLite por su simplicidad y eficiencia en el manejo de conjuntos de datos de tamaño moderado. Esta base de datos facilitó la consulta y recuperación eficiente de datos para el posterior análisis de sentimientos.

V. Metodología de análisis de sentimiento

El módulo de análisis de sentimientos aplicó técnicas de procesamiento del lenguaje natural para analizar los sentimientos expresados en los comentarios. El proceso implicó los siguientes pasos:

- Tokenización: dividir el texto en palabras o tokens individuales.
- Eliminación de palabras irrelevantes: eliminación de palabras comunes que no contribuyen al sentimiento (por ejemplo, "y", "el").
- Puntuación de sentimiento: uso de TextBlob para asignar una puntuación de polaridad de sentimiento a cada comentario. TextBlob proporciona una puntuación de polaridad que oscila entre -1 (negativo) y 1 (positivo).
- Agregación: calcular el sentimiento general de cada publicación y la tendencia del sentimiento general en todos los comentarios.

VI. Resultados y discusiones

El análisis de sentimiento de los comentarios relacionados con Gustavo Petro en Instagram reveló varios conocimientos sobre la percepción pública. Se encontró que la mayoría de los comentarios eran neutrales, y una proporción menor mostraba fuertes sentimientos positivos o negativos. Las tendencias del sentimiento a lo largo del tiempo permitieron comprender cómo fluctuó la opinión pública con los diferentes eventos y declaraciones de Gustavo Petro.

Las visualizaciones creadas con Matplotlib y Seaborn ilustraron estas tendencias de sentimiento, mostrando picos de sentimiento positivo correspondientes a publicaciones o anuncios específicos, mientras que los picos de sentimiento negativo indicaron períodos de controversia o crítica.

VII. Conclusiones

Este proyecto desarrolló con éxito un canal integral de análisis de sentimientos y ETL para comentarios relacionados con Gustavo Petro en Instagram. A pesar de los desafíos iniciales con los límites de tarifas, el uso de múltiples cuentas de Instagram y la API Instaloader permitió una extracción de datos eficiente. El análisis de sentimiento proporcionó información valiosa sobre la opinión pública, demostrando la utilidad del análisis de redes sociales para comprender el impacto de las figuras públicas.

Si bien el proyecto desarrolló con éxito un proceso eficaz de análisis de sentimientos y ETL, enfrentó varias limitaciones y restricciones. La dependencia de múltiples cuentas de Instagram debido a los límites de velocidad de la API planteó un desafío importante que podría afectar la coherencia de los datos. Además, el análisis de sentimientos se limitó a las capacidades de TextBlob, que puede no capturar sentimientos matizados con tanta eficacia como los modelos más avanzados. Las mejoras futuras podrían abordar estas limitaciones explorando métodos de extracción de datos más sólidos e integrando técnicas sofisticadas de análisis de sentimientos.

VIII. Trabajo a futuro

Las mejoras futuras podrían incluir la integración de datos de redes sociales adicionales para proporcionar una visión más holística del sentimiento público. También podrían explorarse técnicas avanzadas de análisis de sentimientos, como los modelos de aprendizaje profundo, para mejorar la precisión y profundidad de los conocimientos.