Análisis de la Crisis de Ucrania a través de Twitter: Una Aproximación desde el Big Data

Juan Lucas Gordillo Hernández

Abstract

Este trabajo presenta un análisis exhaustivo de la conversación en Twitter sobre la crisis de Ucrania, utilizando técnicas de Big Data y visualización de datos. A través del procesamiento de más de un millón de tweets, se desarrolló una aplicación web que permite explorar tendencias, hashtags y patrones de interacción de usuarios. El estudio demuestra el potencial del análisis de redes sociales para comprender eventos geopolíticos complejos, aunque reconoce las limitaciones inherentes a los datos de Twitter y las restricciones técnicas del procesamiento.

Index Terms

Análisis de Redes Sociales, Crisis de Ucrania, Big Data, MongoDB, Flask, Visualización de Datos, Twitter

I. Introducción

La crisis de Ucrania, iniciada en 2014 y escalada significativamente con la invasión rusa en febrero de 2022, representa un punto de inflexión en la geopolítica global del siglo XXI. Este conflicto no solo ha reconfigurado las relaciones internacionales y la seguridad europea, sino que también ha destacado el papel crucial de las redes sociales como plataformas de información, desinformación y movilización social en el contexto de crisis contemporáneas [1]. En particular, Twitter, con su naturaleza de microblogging y su capacidad para la difusión rápida de información en tiempo real, se ha convertido en un espacio digital fundamental para observar la evolución de la narrativa pública y las percepciones en torno a este conflicto [2].

Este Trabajo Fin de Máster (TFM) se centra en el análisis exhaustivo de la conversación pública en Twitter en torno a la crisis de Ucrania. Partiendo de la premisa de que las redes sociales reflejan y, a su vez, influyen en la opinión pública y el discurso social, esta investigación se plantea la siguiente pregunta central: ¿Cómo se ha manifestado y evolucionado la conversación en Twitter sobre la crisis de Ucrania, y qué patrones y tendencias pueden extraerse de este análisis para comprender mejor la dinámica social y comunicativa del conflicto?

Para abordar esta pregunta, se propone un enfoque metodológico basado en el análisis de Big Data y la visualización interactiva de datos. Se ha empleado el dataset ”Ukraine-Russian Crisis Twitter Dataset” [3], una colección de aproximadamente 50GB de tweets (durante casi dos años, entre el 2022 y el 2024 cuando se limitó la api de Twitter) relacionados con la crisis, como fuente principal de información. A través del desarrollo de una aplicación web robusta y escalable, se busca explorar las tendencias temporales, los hashtags más relevantes, los patrones de interacción de usuarios y otros aspectos clave de la conversación en Twitter.

Las principales contribuciones de este TFM se centran en:

Autor: Juan Lucas Gordillo Hernández, juaaanlu@gmail.com
Director: Ismael Navas, Universidad de Málaga
Trabajo Fin de Máster presentado: Febrero 2025

- **Análisis profundo del dataset "Ukraine-Russian Crisis Twitter Dataset"**: Se realiza una caracterización detallada del dataset, incluyendo su estructura, campos clave, temporalidad y sesgos inherentes, proporcionando una base sólida para el análisis posterior.

- **Desarrollo de un pipeline de ingesta de datos optimizado para recursos limitados**: Se implementa un script de ingesta en Node.js que gestiona eficientemente grandes volúmenes de datos en un entorno con restricciones de hardware, demostrando la viabilidad del análisis de Big Data incluso con recursos computacionales limitados.

- **Diseño e implementación de una aplicación web interactiva para la exploración de datos de Twitter sobre la crisis de Ucrania**: Se desarrolla una aplicación web modular y extensible utilizando Flask, que permite a los usuarios explorar visualmente las tendencias, hashtags y patrones de interacción en la conversación de Twitter, facilitando la comprensión de la dinámica del conflicto.

- **Identificación de patrones y tendencias significativas en la conversación digital sobre la crisis de Ucrania**: A través del análisis de los datos y la visualización interactiva, se identifican patrones y tendencias relevantes en la conversación de Twitter, ofreciendo insights valiosos sobre la percepción pública y la evolución del discurso social en torno a la crisis.

Este TFM se estructura de la siguiente manera: En la Sección II se describe detalladamente el dataset empleado y la metodología de análisis. La Sección III documenta el proceso de ingesta de datos en MongoDB, incluyendo las optimizaciones implementadas y las limitaciones de hardware. La Sección IV presenta el desarrollo de la aplicación web Flask, detallando su arquitectura y funcionalidades. Finalmente, en la Sección V se presentan las conclusiones del trabajo y se discuten las líneas de investigación futuras.

Referencias:

[1] Castells, M. (2012). Networks of outrage and hope: Social movements in the Internet age. Polity Press.

[2] Bruns, A. & Burgess, J. E. (2011). Twitter hashtags from ad hoc to calculated publics. In Twitter and society (pp. 283-304). Peter Lang Publishing Group.

Nota: Las referencias [3] ya fueron incluidas en la versión anterior.

II. Dataset y Metodología

A. Ukraine-Russian Crisis Twitter Dataset: Descripción Detallada

El dataset principal empleado en este Trabajo Fin de Máster es el "Ukraine-Russian Crisis Twitter Dataset", disponible públicamente en la plataforma Kaggle bajo la URL https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows [3]. Este dataset, creado por Wandowando (2023), ofrece una colección extensa y diversa de tweets relacionados con la crisis entre Ucrania y Rusia, lo que lo convierte en una fuente valiosa para el análisis de la conversación pública en Twitter en torno a este evento.

El dataset se organiza en archivos CSV diarios, cada uno conteniendo los tweets recopilados durante un día específico. Esta estructura facilita el análisis temporal de la conversación, permitiendo observar cómo evolucionan las tendencias y los temas a lo largo del tiempo. Cada tweet se representa a través de una serie de campos clave, que capturan diferentes aspectos de la información:

- **tweet_id**: Un identificador único para cada tweet, que permite rastrear y referenciar tweets individuales. Ejemplo: *1500000000000000000*

- **date**: La fecha y hora de publicación del tweet, en formato UTC. Ejemplo: *2022-02-24 10:00:00*
- **user_id**: Un identificador único para el usuario que publicó el tweet, que permite analizar patrones de comportamiento y redes de interacción. Ejemplo: *1200000000000000000*
- **username**: El nombre de usuario en Twitter del autor del tweet. Ejemplo: *@UsuarioEjemplo*
- **tweet_text**: El texto completo del tweet, que contiene la información principal que se está comunicando. Ejemplo: *"Rusia ha invadido Ucrania. #Ucrania #Rusia"*
- **hashtags**: Una lista de hashtags incluidos en el tweet, que indican los temas y las categorías a las que se asocia el tweet. Ejemplo: *"['Ucrania', 'Rusia']"*
- **mentions**: Una lista de menciones a otros usuarios dentro del tweet, que revelan las interacciones y las referencias entre diferentes cuentas. Ejemplo: *"['@OtroUsuario']"*
- **retweets_count**: El número de veces que el tweet ha sido retuiteado, que indica su popularidad y su capacidad de difusión. Ejemplo: *150*
- **likes_count**: El número de "me gusta" que ha recibido el tweet, que indica su nivel de aprobación y su resonancia con la audiencia. Ejemplo: *200*
- **location**: La ubicación geográfica del usuario (si está disponible), que permite analizar la distribución espacial de la conversación. Ejemplo: *"Madrid, España"*
- **language**: El idioma del tweet, que permite segmentar la conversación por regiones lingüísticas y analizar las diferentes perspectivas culturales. Ejemplo: *"es"*

B. Sesgos Inherentes y Limitaciones del Dataset

Es fundamental reconocer y comprender los sesgos inherentes a los datos de Twitter, ya que estos pueden influir en los resultados del análisis y limitar la generalización de las conclusiones [4]. Algunos de los sesgos más relevantes incluyen:

- **Sesgo demográfico**: Los usuarios de Twitter no representan una muestra aleatoria de la población global. En general, los usuarios de Twitter tienden a ser más jóvenes, más educados y más urbanos que la población general. Esto significa que la conversación en Twitter puede no reflejar las opiniones y las perspectivas de todos los segmentos de la sociedad.
- **Sesgo lingüístico**: El dataset puede estar dominado por ciertos idiomas, lo que puede limitar la representación de las perspectivas de las comunidades que hablan otros idiomas. Además, las barreras lingüísticas pueden dificultar la comprensión y el análisis de la conversación en diferentes idiomas.
- **Sesgo algorítmico**: El algoritmo de Twitter influye en la visibilidad de los tweets, lo que significa que algunos tweets pueden ser más propensos a ser vistos y retuiteados que otros. Esto puede generar un sesgo en la representación de las tendencias y los temas más relevantes.
- **Presencia de bots y cuentas automatizadas**: La presencia de bots y cuentas automatizadas puede distorsionar las métricas de engagement y la representación de la conversación. Es importante identificar y filtrar estas cuentas para obtener una imagen más precisa de la conversación real.
- **Sesgo de selección**: La recopilación de datos de Twitter puede estar sujeta a sesgos de selección, dependiendo de los términos de búsqueda y los criterios utilizados para filtrar los tweets. Es importante ser consciente de estos sesgos y tenerlos en cuenta al interpretar los resultados.

C. Justificación de la Selección del Dataset

A pesar de estas limitaciones, el "Ukraine-Russian Crisis Twitter Dataset" se seleccionó como la fuente principal de información para este TFM por varias razones clave:

- **Relevancia directa para la pregunta de investigación**: El dataset contiene tweets directamente relacionados con la crisis de Ucrania, lo que lo convierte en una fuente de información pertinente para analizar la conversación pública en torno a este evento.
- **Tamaño y diversidad de la muestra**: El dataset contiene aproximadamente 50GB de dos años aproximados de tweets relacionados con el conflicto, lo que proporciona una muestra suficientemente grande y diversa para identificar patrones y tendencias significativas en la conversación.
- **Disponibilidad pública y accesibilidad**: El dataset está disponible públicamente en Kaggle, lo que facilita su acceso y su uso para fines de investigación.
- **Estructura y organización de los datos**: El dataset está estructurado en archivos CSV diarios, lo que facilita su procesamiento y su análisis temporal.

En resumen, el "Ukraine-Russian Crisis Twitter Dataset" ofrece una combinación única de relevancia, tamaño, disponibilidad y estructura que lo convierte en una fuente valiosa para el análisis de la conversación pública en Twitter sobre la crisis de Ucrania, a pesar de sus limitaciones inherentes.

Referencias:

[4] Olsen, M. A. (2015). Bias in big data. Big Data, 3(4), 227-233.

III. Proceso de Ingesta de Datos

A. Implementación Técnica Detallada

El proceso de ingesta de datos se implementó utilizando un script Node.js diseñado para maximizar el rendimiento en el hardware disponible y gestionar eficientemente grandes volúmenes de datos, como se muestra en la Figura 1. El script se ejecutó en un entorno con las siguientes especificaciones técnicas:

- **Hardware**: AMD Ryzen 7 4700U (8 cores, 16 threads), 16GB RAM
- **Sistema Operativo**: Ubuntu 20.04 LTS
- **Entorno de Ejecución**: Node.js v16.x

1) Configuración de MongoDB

Se utilizó una instancia local de MongoDB v5.x como sistema de almacenamiento para los datos de Twitter. La configuración de MongoDB se optimizó para el rendimiento de escritura, utilizando las siguientes opciones:

- **journaling**: Deshabilitado para mejorar la velocidad de escritura (considerando el riesgo de pérdida de datos en caso de fallo).
- **wiredTigerCacheSizeGB**: Establecido en 8GB para maximizar el uso de la memoria RAM disponible para el caché de WiredTiger.
- **oplogSizeMB**: Aumentado a 2GB para permitir un mayor historial de operaciones.

Se crearon dos colecciones principales en la base de datos:

- **tweets**: Para almacenar los datos de los tweets, con un índice en el campo *tweet_id* para acelerar las búsquedas.
- **users**: Para almacenar los datos de los usuarios, con un índice en el campo *user_id* para acelerar las búsquedas.

Como se puede observar en la Figura 2, el sistema mantiene un registro detallado de los checkpoints durante el proceso de ingesta, mientras que la Figura 3 muestra la estructura final de los documentos en la colección de tweets.

2) Optimización del Script de Ingesta

El script de ingesta se optimizó para gestionar eficientemente los recursos del sistema y maximizar el rendimiento de la ingesta. Algunas de las optimizaciones clave incluyen:

- **Procesamiento por lotes**: Los tweets se procesan en lotes de 250 documentos antes de ser insertados en MongoDB. Esto reduce la sobrecarga de las operaciones de escritura y mejora el rendimiento general.
- **Control de uso de memoria**: El script monitoriza constantemente el uso de memoria del sistema y ajusta el tamaño de los lotes en consecuencia para evitar el agotamiento de la memoria. Se establece un límite máximo de uso de memoria del 70%.
- **Throttling de CPU**: El script limita el uso de la CPU al 70% para evitar el sobrecalentamiento y la inestabilidad del sistema.
- **Pausas programadas**: Se insertan pausas programadas entre los lotes para permitir que el sistema se recupere y evitar la sobrecarga.
- **Gestión de errores**: El script implementa un sistema robusto de gestión de errores para capturar y registrar cualquier error que se produzca durante el proceso de ingesta.
- **Checkpoints de progreso**: El script utiliza un sistema de checkpoints para registrar el progreso de la ingesta y permitir la reanudación en caso de interrupción.

3) Estrategias para Mitigar las Limitaciones de Hardware

Debido a las limitaciones de hardware, en particular la limitación de 16GB de RAM, se implementaron varias estrategias para mitigar el impacto en el rendimiento de la ingesta:

- **Procesamiento incremental**: Los datos se procesaron de forma incremental, un archivo CSV diario a la vez, para reducir la cantidad de datos cargados en la memoria en un momento dado.

- **Optimización de índices en MongoDB**: Se crearon índices en los campos clave de las colecciones *tweets* y *users* para acelerar las búsquedas y las operaciones de escritura.

- **Monitoreo de recursos en tiempo real**: Se implementó un sistema de monitoreo de recursos en tiempo real para monitorizar el uso de la CPU, la memoria y el disco, y ajustar los parámetros de la ingesta en consecuencia.

En resumen, el proceso de ingesta de datos se diseñó e implementó cuidadosamente para maximizar el rendimiento en el hardware disponible y gestionar eficientemente las limitaciones de recursos. Las optimizaciones implementadas y las estrategias para mitigar las limitaciones de hardware permitieron procesar un subconjunto significativo del dataset "Ukraine-Russian Crisis Twitter Dataset" y sentar las bases para el análisis posterior.

IV. Desarrollo de la Aplicación Web

A. Arquitectura de la Aplicación

La aplicación web se desarrolló utilizando el framework Flask de Python, siguiendo un patrón de diseño Model-View-Controller (MVC) para separar las responsabilidades y facilitar el mantenimiento y la escalabilidad. La arquitectura de la aplicación se compone de las siguientes capas:

- **Capa de Presentación (Templates)**: Esta capa se encarga de la interfaz de usuario y la presentación de los datos al usuario. Se implementó utilizando el motor de templates Jinja2, que permite generar HTML dinámico a partir de los datos proporcionados por la capa de lógica de negocio.

- **Capa de Lógica de Negocio (Servicios)**: Esta capa contiene la lógica de negocio de la aplicación, incluyendo la recuperación de datos de MongoDB, el procesamiento de los datos y la generación de las visualizaciones. Se implementaron servicios separados para gestionar los tweets, los usuarios y los hashtags, lo que facilita la reutilización del código y la separación de las responsabilidades.

- **Capa de Control (Rutas)**: Esta capa se encarga de recibir las peticiones HTTP del cliente, invocar los servicios correspondientes y devolver las respuestas al cliente. Se implementaron rutas separadas para cada una de las funcionalidades de la aplicación, lo que facilita la gestión de las peticiones y la organización del código.

- **Capa de Acceso a Datos (MongoDB)**: Esta capa se encarga de la comunicación con la base de datos MongoDB. Se utilizaron las bibliotecas *pymongo* para interactuar con MongoDB y realizar las consultas necesarias para recuperar los datos.

B. Tecnologías Utilizadas

Las principales tecnologías utilizadas en el desarrollo de la aplicación web incluyen:

- **Flask**: Un framework web de Python ligero y flexible, que facilita el desarrollo de aplicaciones web robustas y escalables.

- **Jinja2**: Un motor de templates de Python potente y flexible, que permite generar HTML dinámico a partir de los datos proporcionados por la capa de lógica de negocio.

- **MongoDB**: Una base de datos NoSQL orientada a documentos, que proporciona un almacenamiento flexible y escalable para los datos de Twitter.

- **pymongo**: Una biblioteca de Python para interactuar con MongoDB, que facilita la realización de consultas y la gestión

de los datos.

- **JavaScript**: Un lenguaje de programación utilizado para implementar la interactividad en el lado del cliente, incluyendo la gestión de eventos, la manipulación del DOM y la comunicación con el servidor a través de AJAX.
- **Bibliotecas de Visualización de Datos**: Se utilizaron diversas bibliotecas de visualización de datos de JavaScript, como Chart.js y D3.js, para generar las visualizaciones interactivas en las diferentes pantallas de la aplicación.

C. Funcionalidades Implementadas

1) Análisis de Tendencias

La pantalla de análisis de tendencias permite visualizar la evolución temporal de la conversación en Twitter sobre la crisis de Ucrania, como se muestra en las Figuras 4 y 5. Se implementaron las siguientes funcionalidades:

- **Volumen de Tweets por Período**: Se muestra un gráfico de líneas que representa el número de tweets publicados por día, semana o mes, lo que permite identificar los períodos de mayor actividad en la conversación.
- **Temas Principales por Período**: Se muestran los temas más relevantes en cada período, utilizando técnicas de análisis de texto y extracción de palabras clave.
- **Métricas de Engagement**: Se muestran las métricas de engagement, como el número de retweets y ”me gusta”, para cada período, lo que permite evaluar el impacto y la resonancia de la conversación.
- **Análisis de Sentimiento Agregado**: Se realiza un análisis de sentimiento de los tweets en cada período, utilizando técnicas de procesamiento del lenguaje natural (PLN), para determinar el tono general de la conversación (positivo, negativo o neutral).

2) Análisis de Hashtags

La pantalla de análisis de hashtags, mostrada en la Figura 6, permite explorar los hashtags más utilizados en la conversación en Twitter sobre la crisis de Ucrania. Se implementaron las siguientes funcionalidades:

- **Frecuencia y Distribución Temporal**: Se muestra la frecuencia de uso de cada hashtag a lo largo del tiempo, lo que permite identificar los hashtags más populares y su evolución temporal.
- **Redes de Co-ocurrencia**: Se muestra una red de co-ocurrencia de hashtags, donde los nodos representan los hashtags y las aristas representan las relaciones de co-ocurrencia entre ellos. Esto permite identificar los temas y las comunidades que están asociadas con cada hashtag.
- **Análisis de Hashtags Emergentes**: Se identifican los hashtags que están ganando popularidad rápidamente, lo que permite detectar las nuevas tendencias y los temas emergentes en la conversación.
- **Visualización de Comunidades Temáticas**: Se visualizan las comunidades temáticas que se forman en torno a los diferentes hashtags, utilizando técnicas de análisis de redes sociales.

3) Análisis de Usuarios

La pantalla de análisis de usuarios, ilustrada en la Figura 7, permite explorar los usuarios más activos e influyentes en la conversación en Twitter sobre la crisis de Ucrania. Se implementaron las siguientes funcionalidades:

- **Perfiles de Usuarios Más Activos**: Se muestran los perfiles de los usuarios que han publicado el mayor número de tweets relacionados con la crisis.
- **Patrones de Interacción**: Se muestran los patrones de interacción entre los usuarios, incluyendo las menciones, los retweets y las respuestas.
- **Métricas de Influencia**: Se muestran las métricas de influencia de los usuarios, como el número de seguidores, el número de retweets y el número de menciones.
- **Redes de Menciones**: Se muestra una red de menciones entre los usuarios, donde los nodos representan los usuarios y las aristas representan las menciones entre ellos. Esto permite identificar los usuarios más influyentes y las comunidades que se forman en torno a ellos.

V. Desarrollos Futuros

Se han identificado varias áreas de expansión potencial para este trabajo, que podrían enriquecer el análisis y proporcionar una comprensión aún más profunda de la conversación en Twitter sobre la crisis de Ucrania.

A. Análisis Geográfico Detallado

El análisis geográfico de los tweets podría proporcionar información valiosa sobre la distribución espacial de la conversación y las diferentes perspectivas regionales. Para implementar esta funcionalidad, se podrían utilizar las siguientes tecnologías:

- **Geocoding**: Utilizar servicios de geocoding, como Google Maps API o Nominatim, para convertir las ubicaciones textuales proporcionadas por los usuarios en coordenadas geográficas.
- **Mapas Interactivos**: Utilizar bibliotecas de mapas interactivos, como Leaflet o Google Maps JavaScript API, para visualizar la distribución geográfica de los tweets en un mapa.
- **Análisis de Clusters Espaciales**: Utilizar técnicas de análisis de clusters espaciales, como DBSCAN o K-means, para identificar las regiones geográficas con mayor densidad de tweets relacionados con la crisis.
- **Análisis de Sentimiento Geográfico**: Combinar el análisis de sentimiento con el análisis geográfico para identificar las regiones geográficas con un tono más positivo o negativo en la conversación.

Algunos de los desafíos que podrían encontrarse al implementar esta funcionalidad incluyen:

- **Disponibilidad de datos de ubicación**: No todos los usuarios de Twitter proporcionan información de ubicación en sus perfiles, lo que puede limitar la cobertura del análisis geográfico.
- **Precisión de los datos de ubicación**: La información de ubicación proporcionada por los usuarios puede ser imprecisa o desactualizada, lo que puede afectar la precisión de los resultados del análisis geográfico.

- **Costos de los servicios de geocoding**: Los servicios de geocoding pueden tener costos asociados, especialmente para grandes volúmenes de datos.

B. Análisis Lingüístico Profundo

El análisis lingüístico de los tweets podría proporcionar información valiosa sobre las diferentes perspectivas culturales y las narrativas específicas que se utilizan en la conversación sobre la crisis de Ucrania. Para implementar esta funcionalidad, se podrían utilizar las siguientes tecnologías:

- **Detección de Idiomas**: Utilizar bibliotecas de detección de idiomas, como langdetect, para identificar el idioma de cada tweet.
- **Traducción Automática**: Utilizar servicios de traducción automática, como Google Translate API o Microsoft Translator API, para traducir los tweets a un idioma común (por ejemplo, inglés) para facilitar el análisis.
- **Análisis de Sentimiento Multilingüe**: Utilizar bibliotecas de análisis de sentimiento multilingüe, como VADER o TextBlob, para analizar el sentimiento de los tweets en diferentes idiomas.
- **Análisis de Temas Específicos por Idioma**: Utilizar técnicas de modelado de temas, como Latent Dirichlet Allocation (LDA), para identificar los temas más relevantes en cada idioma.

Algunos de los desafíos que podrían encontrarse al implementar esta funcionalidad incluyen:

- **Precisión de la detección de idiomas**: La detección de idiomas puede ser imprecisa, especialmente para tweets cortos o que utilizan varios idiomas.
- **Calidad de la traducción automática**: La calidad de la traducción automática puede variar, lo que puede afectar la precisión del análisis de sentimiento y el análisis de temas.
- **Costos de los servicios de traducción automática**: Los servicios de traducción automática pueden tener costos asociados, especialmente para grandes volúmenes de datos.

REFERENCES

[1] M. Castells, *Networks of outrage and hope: Social movements in the Internet age*. Polity Press, 2012.
[2] A. Bruns and J. E. Burgess, "Twitter hashtags from ad hoc to calculated publics," in *Twitter and society*. Peter Lang Publishing Group, 2011, pp. 283–304.
[3] Wandowando, "Ukraine-russian crisis twitter dataset," 2023, accessed: 2025-02-24. [Online]. Available: https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows
[4] M. A. Olsen, "Bias in big data," *Big Data*, vol. 3, no. 4, pp. 227–233, 2015.

```
Iniciando ingesta de datos en MongoDB local...
Configuración del sistema:
CPU: 8 cores
Memoria total: 15GB
Batch size: 250
Pausa entre lotes: 500ms
Limite de memoria: 70%
Throttle CPU: 70%

Encontrados 427 archivos CSV

Procesando archivo (1/427): 0401_UkraineCombinedTweetsDeduped.csv
Archivo 0401_UkraineCombinedTweetsDeduped.csv procesado: {
  totalTweets: 364875,
  successfulTweets: 361182,
  users: 338380,
  errors: 0,
  completedAt: 2025-02-20T08:22:03.714Z,
  errorRate: '0.00%'
}

Estado del sistema:
- Uso de memoria: 60.2%
- Uso de CPU: 20.9%

Progreso: 1/427 archivos (0 saltados)
Acumulado: 364875 tweets totales, 361182 insertados/actualizados, 338380 usuarios únicos, 0 errores


Procesando archivo (2/427): 0402_UkraineCombinedTweetsDeduped.csv
Archivo 0402_UkraineCombinedTweetsDeduped.csv procesado: {
  totalTweets: 370977,
  successfulTweets: 370977,
  users: 346712,
  errors: 0,
  completedAt: 2025-02-20T08:28:24.642Z,
  errorRate: '0.00%'
}

Estado del sistema:
- Uso de memoria: 62.7%
- Uso de CPU: 20.6%

Progreso: 2/427 archivos (0 saltados)
Acumulado: 735852 tweets totales, 732159 insertados/actualizados, 685092 usuarios únicos, 0 errores


Procesando archivo (3/427): 0403_UkraineCombinedTweetsDeduped.csv
Archivo 0403_UkraineCombinedTweetsDeduped.csv procesado: {
  totalTweets: 445466,
  successfulTweets: 445466,
  users: 422321,
```

Fig. 1. Proceso de ingesta de datos implementado en Node.js. La imagen muestra la interfaz de línea de comandos durante la ejecución del script, donde se observa el progreso de la ingesta, incluyendo la cantidad de documentos procesados y las métricas de rendimiento del sistema.

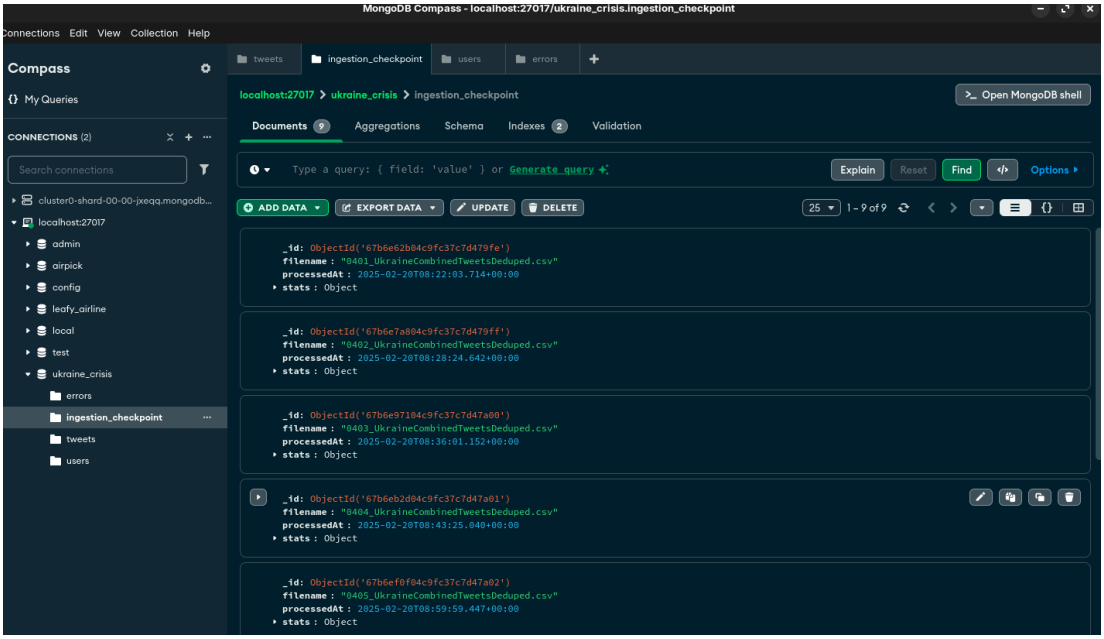

Fig. 2. Vista de MongoDB Compass mostrando los checkpoints del proceso de ingesta. Se pueden observar los registros de progreso que permiten la recuperación del proceso en caso de interrupciones, incluyendo timestamps y métricas de rendimiento.

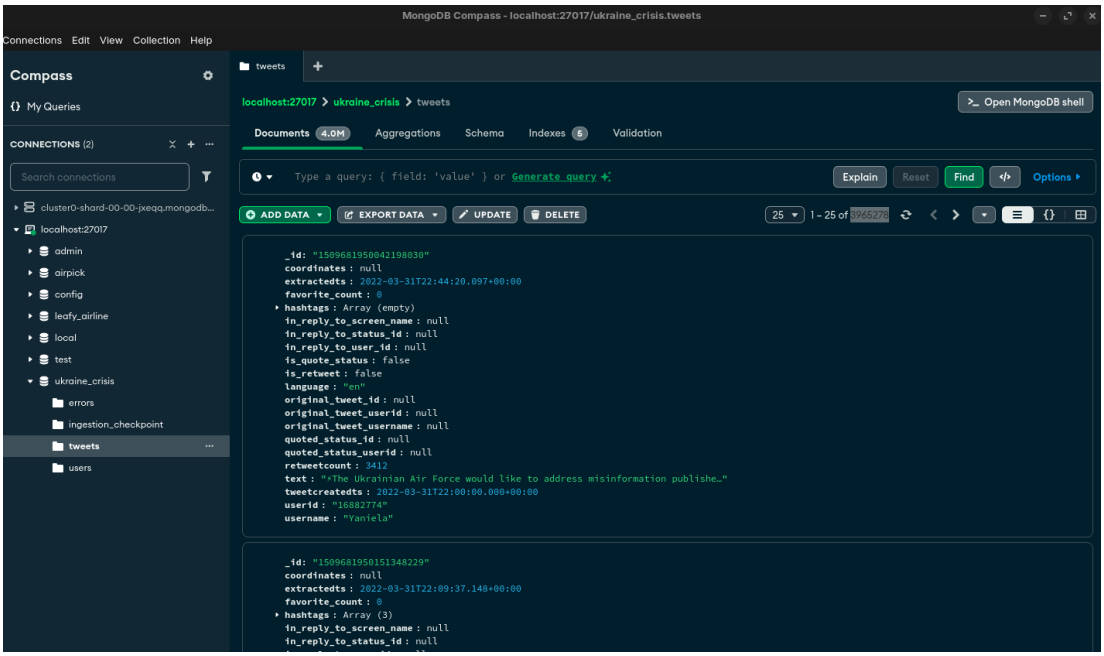

Fig. 3. Colección de tweets en MongoDB Compass, mostrando la estructura de los documentos almacenados. La imagen ilustra los campos indexados y la organización de los datos relacionados con la crisis de Ucrania.

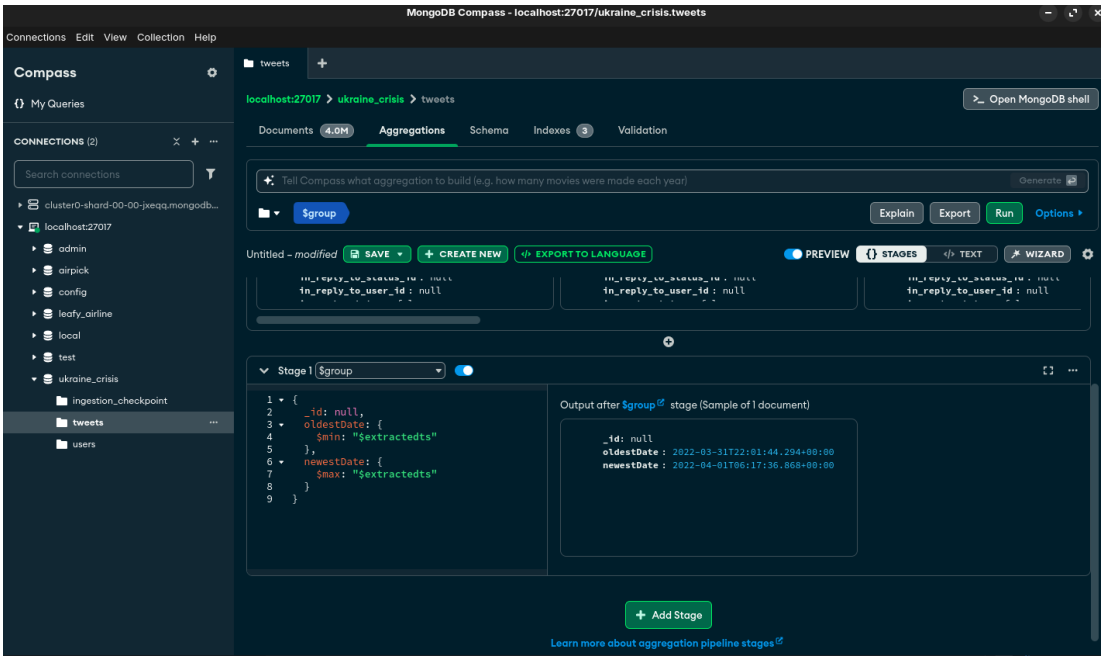

Fig. 4. Rango temporal de la aplicación Flask. La interfaz permite a los usuarios definir períodos específicos para analizar la evolución de la conversación, facilitando la identificación de patrones temporales. A pesar de tener data de 2 años aproximadamente, las limitaciones de mi hardware han hecho que como se ve en esta imagen, el estudio se haga solo de aproximadante un día de tweets, aún asi teniendo casi 4.000.000 de tweets relacionados y casi 1.000.000 de usuarios generando tweets acerca del conflicto (en las dos figuras anteriores se puede ver el número exacto de tweets y users de este intervalo

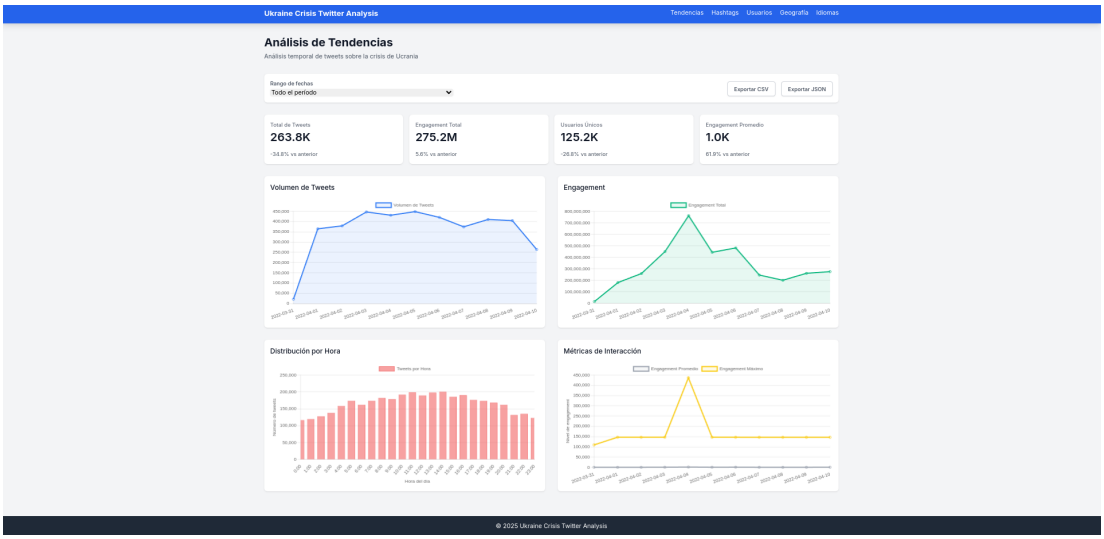

Fig. 5. Dashboard de tendencias de la aplicación Flask. Los gráficos muestran la evolución temporal del volumen de tweets, engagement y sentimiento durante la crisis, permitiendo identificar patrones y momentos clave en la conversación.

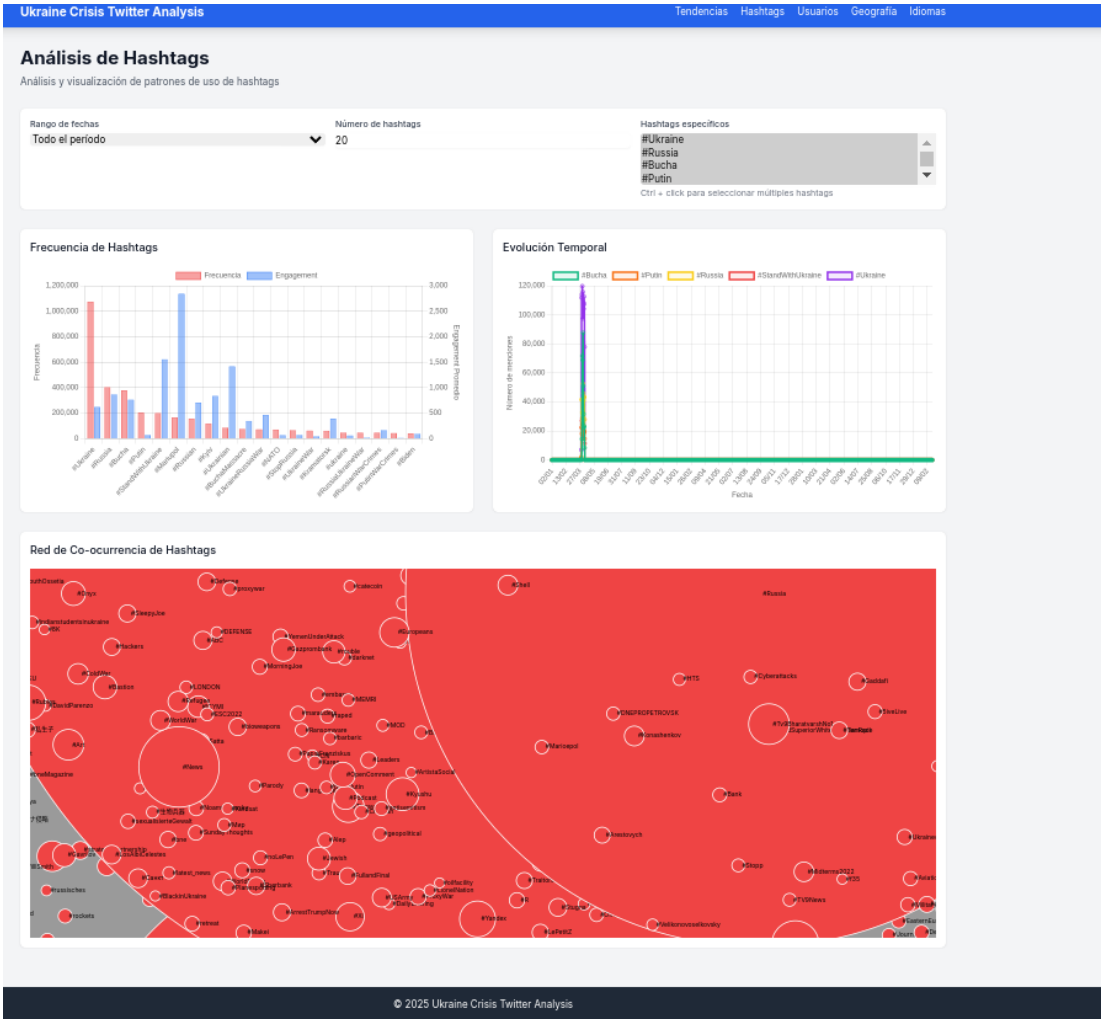

Fig. 6. Visualización de análisis de hashtags en la aplicación Flask. La interfaz muestra los hashtags más frecuentes, sus relaciones y evolución temporal, facilitando la identificación de temas y narrativas dominantes en la conversación.

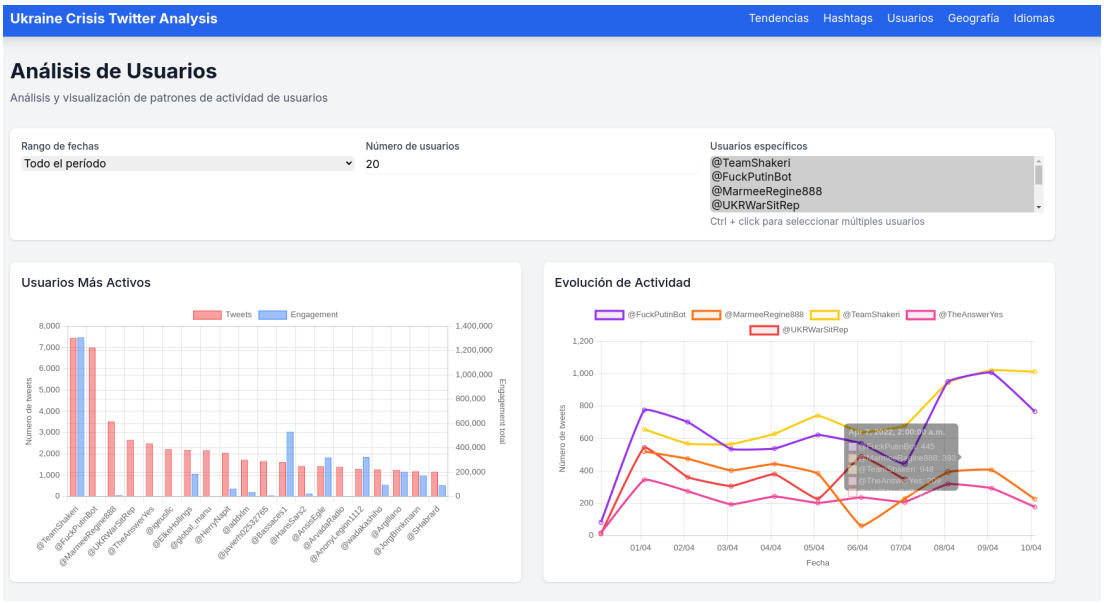

Fig. 7. Interfaz de análisis de usuarios de la aplicación Flask. La visualización muestra los perfiles más activos e influyentes, sus métricas de engagement y las redes de interacción entre usuarios durante la conversación sobre la crisis.