El Pulso Emocional de X (Twitter): Un Análisis de Sentimientos sobre la Guerra en Ucrania

1st Gabriel Isaías Fallas López
Universidad de Costa Rica
San José, Costa Rica
GABRIEL.FALLAS@ucr.ac.cr

***Abstract*—The Russo-Ukrainian conflict generates massive volumes of social media data, particularly on Twitter, making sentiment analysis crucial for understanding public opinion. However, processing such large datasets (1.2M tweets) presents significant data engineering challenges. Traditional methods often lack a scalable, end-to-end solution integrating orchestration, distributed processing, fast analytical querying (OLAP), and governance. This project addresses this gap by designing and implementing a complete, containerized, and production-grade data pipeline to efficiently ingest, process, and analyze the sentiment of 1.2 million tweets related to the crisis. We developed a modern architecture using Docker for containerization, Apache Airflow for orchestration, and Apache Spark (with a Hugging Face model) for distributed processing. Processed data is ingested into Apache Druid for low-latency OLAP queries, with results visualized in Apache Superset and governance handled by OpenMetadata. The pipeline successfully processes the entire dataset, achieving key performance results. This work provides a complete, reproducible, and scalable architectural blueprint for large-scale social media analysis, offering a production-ready template that bridges the gap between raw big data and real-time visualization.**

***Index Terms*—Data Engineering, Sentiment Analysis, Big Data, Apache Spark, Apache Airflow, Apache Druid, MLOps, ETL**

I. Introducción

La invasión de Ucrania por parte de la Federación Rusa el 24 de febrero de 2022 no solo desencadenó la mayor crisis de refugiados en Europa desde la Segunda Guerra Mundial (más de 6 millones de desplazados según ACNUR), sino que también marcó un cambio de paradigma en la geopolítica moderna: la consolidación de la "Guerra Híbrida". En este nuevo teatro de operaciones, las plataformas de redes sociales, y específicamente X (anteriormente Twitter), han dejado de ser meros canales de comunicación para convertirse en campos de batalla activos donde se disputa la narrativa, se coordina la ayuda humanitaria y se despliegan campañas de desinformación masiva [1].

La magnitud del fenómeno digital es asombrosa. Durante las primeras semanas del conflicto, Twitter registró más de 50 millones de tweets diarios relacionados con Ucrania, convirtiendo hashtags como `#StandWithUkraine` y `#StopRussia` en tendencias globales sostenidas. Este volumen de datos representa una oportunidad única para la inteligencia de fuentes abiertas (OSINT) y la comprensión sociológica del conflicto en tiempo real.

Sin embargo, la naturaleza de estos datos presenta desafíos monumentales de ingeniería. Nos enfrentamos a las clásicas "V" del Big Data:

- **Volumen:** Millones de mensajes generados diariamente que superan la capacidad de procesamiento de sistemas tradicionales.
- **Velocidad:** Una tasa de generación vertiginosa que exige pipelines de baja latencia para análisis oportuno.
- **Variedad:** Datos no estructurados (texto, emojis, URLs, multimedia) que requieren técnicas avanzadas de Procesamiento de Lenguaje Natural (NLP).
- **Veracidad:** La presencia de bots, cuentas falsas y desinformación que comprometen la calidad del análisis.

Los enfoques tradicionales de análisis de datos, a menudo basados en ejecuciones secuenciales o bases de datos relacionales monolíticas, resultan insuficientes ante esta carga de trabajo. La latencia en la ingesta, la incapacidad de escalar modelos de *Deep Learning* (como Transformers) y la falta de interactividad en las consultas analíticas limitan severamente la capacidad de los analistas para obtener *insights* en tiempos relevantes. Además, la ausencia de gobernanza y linaje en los pipelines de datos académicos suele derivar en problemas de reproducibilidad y confianza en los resultados.

Para abordar estas brechas, este estudio presenta el diseño, implementación y evaluación de una arquitectura de ingeniería de datos de nivel productivo (*production-grade*). Se propone una solución integral que orquesta el ciclo de vida completo del dato: desde su ingesta cruda, pasando por una limpieza y clasificación distribuida mediante modelos de Inteligencia Artificial, hasta su disponibilidad en un almacén analítico de baja latencia.

Las contribuciones principales de este artículo son:

- **Arquitectura de Datos Escalable:** El diseño de un ecosistema contenedorizado basado en microservicios que desacopla el cómputo del almacenamiento, permitiendo el procesamiento eficiente de 1.2 millones de tweets.
- **Inferencia Distribuida de NLP:** La implementación técnica de modelos de lenguaje basados en Transformers (Hugging Face) sobre un clúster de Apache Spark, optimizando el tiempo de clasificación de sentimientos mediante paralelismo de datos.
- **Analítica de Baja Latencia:** La integración de una capa de servicio OLAP con Apache Druid, habilitando tiempos

de respuesta sub-segundo para consultas complejas y visualización en tiempo real.
- **Gobernanza Integral:** La incorporación de principios de DataOps mediante Apache Airflow para la orquestación y OpenMetadata para el linaje, asegurando la trazabilidad y auditabilidad del flujo de información.

El resto del artículo se organiza de la siguiente manera: la Sección II presenta los antecedentes teóricos y trabajos relacionados; la Sección III define los objetivos y preguntas de investigación; la Sección IV describe la metodología y arquitectura implementada; la Sección V presenta los resultados y su discusión; finalmente, la Sección VI concluye el trabajo y propone líneas futuras de investigación.

II. Antecedentes

Esta sección revisa los fundamentos teóricos y tecnológicos que sustentan el diseño del pipeline propuesto, organizados en cuatro ejes: análisis de sentimiento en contextos de crisis, arquitecturas de datos modernas, modelos de lenguaje basados en Transformers, y gobernanza de datos.

A. Análisis de Sentimiento en Contextos de Crisis

El análisis de sentimiento en redes sociales durante conflictos bélicos ha ganado relevancia como herramienta para la comprensión de la opinión pública global. Trabajos previos han demostrado que Twitter funciona como un "termómetro social" durante eventos de alto impacto, donde la polarización emocional se intensifica y las narrativas compiten por dominar el espacio digital.

En el contexto específico de Ucrania, estudios recientes han analizado patrones de desinformación, la actividad de bots pro-Kremlin y la movilización de apoyo internacional a través de hashtags [1]. Sin embargo, la mayoría de estos trabajos se centran en muestras pequeñas o análisis manuales, careciendo de la infraestructura necesaria para procesar millones de registros en tiempo real.

B. Procesamiento Distribuido con Apache Spark

El diseño de pipelines para análisis de texto a gran escala ha avanzado significativamente con la adopción de frameworks de procesamiento distribuido. Apache Spark se ha consolidado como el estándar de facto para cargas de trabajo de Big Data debido a su modelo de ejecución en memoria, su API unificada para batch y streaming, y su integración nativa con el ecosistema Hadoop [2].

Para tareas de NLP, Spark permite paralelizar tanto el preprocesamiento (tokenización, limpieza, normalización) como la inferencia de modelos mediante la función `mapPartitions`, que distribuye el trabajo entre los nodos del clúster minimizando la sobrecarga de comunicación.

C. Almacenamiento Analítico OLAP

Para la capa de consulta analítica, Apache Druid fue diseñado específicamente para cargas OLAP de series temporales con latencias sub-segundo. Su arquitectura híbrida combina almacenamiento columnar con índices invertidos, permitiendo agregaciones y filtros eficientes sobre tablas de miles de millones de filas [3].

A diferencia de bases de datos tradicionales como PostgreSQL o MySQL, Druid sacrifica flexibilidad transaccional (ACID) a cambio de rendimiento analítico extremo, lo que lo hace ideal para dashboards interactivos donde la experiencia del usuario depende de tiempos de respuesta instantáneos.

D. Modelos de Lenguaje: Transformers y DistilBERT

La revolución de los Transformers, iniciada con el paper "Attention Is All You Need" [4], transformó el campo del NLP. Modelos como BERT (Bidirectional Encoder Representations from Transformers) y sus variantes han establecido nuevos estándares en tareas de clasificación de texto, incluyendo análisis de sentimiento.

Para este proyecto, se seleccionó **DistilBERT**, una versión destilada de BERT que retiene el 97% de su capacidad predictiva con solo el 60% de los parámetros [5]. Esta reducción es crítica para la inferencia a escala: un modelo más ligero permite procesar más registros por unidad de tiempo sin comprometer significativamente la precisión. El modelo específico utilizado (`distilbert-base-uncased-finetuned-sst-2-english`) fue pre-entrenado en el corpus SST-2 (Stanford Sentiment Treebank), optimizado para clasificación binaria de sentimiento en inglés.

E. Orquestación y Gobernanza

La orquestación reproducible de pipelines de datos se apoya en plataformas tipo *workflow as code*. Apache Airflow se ha consolidado como la herramienta de facto para definir DAGs (Directed Acyclic Graphs) de ETL/ELT en entornos de ingeniería de datos, permitiendo programar, monitorear y reintentar tareas de forma declarativa [6].

Complementariamente, herramientas de metadata y linaje como OpenMetadata centralizan el catálogo de datos, el linaje de transformaciones y las políticas de acceso, facilitando auditoría y descubrimiento en ecosistemas heterogéneos (Airflow, Spark, Druid, Superset) [7]. Esta capa de gobernanza es requisito clave para proyectos que procesan datos sensibles o de interés público, donde la trazabilidad no es opcional sino mandatoria.

III. Objetivos y preguntas de investigación

El objetivo general de este trabajo es diseñar, implementar y evaluar un pipeline de ingeniería de datos escalable, reproducible y gobernable para el análisis de sentimientos sobre 1.2 millones de tweets relacionados con el conflicto Ruso-Ucraniano. A partir de este objetivo se derivan objetivos específicos y preguntas de investigación:

A. Objetivos específicos

1) Diseñar una arquitectura contenedorizada y orquestada que permita el procesamiento distribuido de un dataset de 1.2M tweets y su ingestión en un almacén analítico de baja latencia.

2) Evaluar el rendimiento del pipeline en términos de tiempo de procesamiento, latencia de consulta y tiempos de carga de dashboards.
3) Demostrar la integración de un modelo de sentimiento basado en Transformers (Hugging Face) con Apache Spark para inferencia a gran escala.
4) Implementar capacidades de gobernanza y linaje usando OpenMetadata y documentar cómo esto mejora trazabilidad y reproducibilidad.

B. Preguntas de investigación e hipótesis

Cada pregunta de investigación se acompaña de una hipótesis que será validada o refutada en la sección de resultados.

1) **PI1:** ¿Qué arquitectura permite un equilibrio entre coste, latencia y capacidad de procesar 1.2M tweets en un entorno contenedorizado reproducible?
   Hipótesis H1: Una arquitectura basada en microservicios Docker con Spark para procesamiento y Druid para consultas logrará procesar el dataset completo en menos de 2 horas manteniendo latencias de consulta sub-segundo.
2) **PI2:** ¿Es viable ejecutar la inferencia de un modelo Transformer (DistilBERT) sobre 1.2M tweets en un clúster Spark sin comprometer la calidad de la predicción?
   Hipótesis H2: El uso de DistilBERT con procesamiento por particiones en Spark permitirá clasificar el corpus completo con una precisión comparable a la reportada en benchmarks (>90% en SST-2) y tiempos de ejecución linealmente escalables.
3) **PI3:** ¿Qué mejoras de rendimiento aporta Apache Druid frente a consultas SQL tradicionales?
   Hipótesis H3: Druid reducirá los tiempos de respuesta de consultas analíticas en al menos un orden de magnitud (10x) comparado con PostgreSQL para las mismas agregaciones.
4) **PI4:** ¿Cómo impacta la incorporación de OpenMetadata en la gobernanza del pipeline?
   Hipótesis H4: OpenMetadata permitirá rastrear automáticamente el linaje desde el CSV crudo hasta los dashboards, reduciendo el tiempo de diagnóstico de errores de datos.

IV. Metodología

Esta sección detalla la arquitectura técnica, el flujo de datos y las decisiones de diseño que sustentan el pipeline implementado.

A. Visión General de la Arquitectura

Nuestro enfoque se basa en un stack de datos moderno, implementado en un entorno contenedorizado con Docker y orquestado mediante `docker-compose`. La arquitectura completa consta de 14 contenedores que se comunican a través de una red puente de Docker (`sentiment-network`). La Tabla I resume los servicios desplegados.

El flujo de datos sigue varias etapas clave, desde la ingesta hasta la visualización.

TABLE I
Servicios Contenedorizados del Pipeline

Servicio	Función	Puerto
PostgreSQL	Metadatos	5432
Airflow Webserver	UI Orquestación	8080
Airflow Scheduler	Programación DAGs	–
Spark Master	Coordinación	8081/7077
Spark Worker	Procesamiento	–
Druid Coordinator	Gestión Segmentos	8082
Druid Broker	Consultas	8083
Druid Historical	Almacenamiento	8084
Druid MiddleManager	Ingesta	8091
Druid Router	Enrutamiento	8888
ZooKeeper	Coordinación Druid	2181
Superset	Visualización	8088
OpenMetadata	Gobernanza	8585
Elasticsearch	Búsqueda Metadata	9200

B. Ingestión y Orquestación de Datos

La ingesta de datos se origina de un dataset de Kaggle (1.2M de tweets) en formato CSV, ubicado en un volumen montado (`data/raw/`). El pipeline es gestionado y orquestado por **Apache Airflow**. Se utiliza un DAG (Directed Acyclic Graph) específico, `twitter_sentiment_pipeline`, que define la secuencia de tareas. Este DAG incluye pasos cruciales como la validación de los datos de entrada (`check_data`), la preparación de directorios, la ejecución del trabajo de procesamiento (`run_spark_job`) y la ingesta final en la capa analítica (`submit_to_druid`).

C. Procesamiento y Análisis de Sentimientos

El procesamiento distribuido se realiza con **Apache Spark**. El script principal `sentiment_analysis.py` es ejecutado por un clúster de Spark (un master y workers). La Tabla II muestra la configuración utilizada.

TABLE II
Configuración del Clúster Spark

Parámetro	Valor
Driver Memory	4 GB
Executor Memory	4 GB
Worker Cores	2
Worker Memory	12 GB
Adaptive Execution	Habilitado

El script realiza las siguientes operaciones de transformación:

1) **Carga de Datos:** Lee los registros del dataset de tweets desde el archivo CSV utilizando el lector nativo de Spark con inferencia de esquema.
2) **Limpieza de Texto:** Aplica una función UDF (User Defined Function) que:
   - Elimina URLs mediante expresiones regulares (`http\S+`)
   - Remueve menciones de usuarios (`@username`)
   - Extrae el texto de hashtags eliminando el símbolo #

- Normaliza espacios en blanco y caracteres especiales

3) **Análisis de Sentimientos:** Utiliza el modelo `distilbert-base-uncased-finetuned-sst-2-english` de Hugging Face. La inferencia se ejecuta mediante `mapPartitions`, donde cada partición carga una instancia del modelo y procesa sus registros en batch, evitando la sobrecarga de inicialización por registro.
4) **Clasificación:** El modelo asigna una de dos etiquetas: `POSITIVE` o `NEGATIVE`, basándose en el score de confianza más alto.
5) **Almacenamiento Intermedio:** Los resultados se materializan en un archivo CSV en `data/processed/sentiment_results.csv` con el esquema enriquecido.

D. Almacenamiento Analítico y Visualización

La capa de almacenamiento está diseñada para alta disponibilidad y consultas rápidas.

- **PostgreSQL:** Actúa como la base de datos de metadatos backend para múltiples servicios, incluyendo Airflow, Druid, Superset y OpenMetadata.
- **Apache Druid:** Es la base de datos analítica principal (OLAP), optimizada para consultas de baja latencia sobre datos de series temporales. Está compuesto por varios servicios (Coordinator, Broker, Historical, Router) y almacena los datos de sentimientos en un *datasource* llamado `ukraine_tweets_sentiment`.

E. Visualización y Gobernanza

La capa de presentación y gobernanza permite la exploración de los resultados y el seguimiento del linaje de datos.

- **Apache Superset:** Es la herramienta de visualización (BI), conectada directamente a Apache Druid. Permite la creación de dashboards interactivos y el uso de SQL Lab para consultas ad-hoc.
- **OpenMetadata:** Se utiliza para la gobernanza de datos. Se conecta a todos los servicios (Airflow, Spark, Druid, Superset, PostgreSQL) para rastrear automáticamente el linaje de datos, desde el CSV crudo hasta el dashboard final.

V. Resultados y Discusión

Esta sección presenta tanto la evaluación técnica del rendimiento de la arquitectura de ingeniería de datos propuesta como un análisis profundo de los hallazgos temáticos y emocionales derivados del procesamiento de los datos.

A. Rendimiento de la Arquitectura de Datos

La validación de la arquitectura se centró en la eficiencia del pipeline ETL y la latencia de las consultas analíticas. El despliegue contenedorizado de 14 servicios demostró ser robusto. Los resultados técnicos clave se resumen a continuación:

- **Eficiencia de Procesamiento:** El uso de Apache Spark permitió reducir el tiempo de inferencia del modelo Transformer. Procesar el dataset completo de 1.2M registros tomó entre 1 y 2 horas, una mejora sustancial frente a la ejecución secuencial en un solo nodo.
- **Latencia OLAP:** La ingestión en Apache Druid habilitó tiempos de respuesta inferiores a 1 segundo ($< 1s$) para consultas de agregación complejas. Esto valida la elección de Druid sobre almacenes de datos tradicionales para cargas de trabajo de análisis interactivo.
- **Visualización:** Los dashboards en Apache Superset cargan en menos de 2 segundos, permitiendo una exploración fluida de los datos históricos.

B. Análisis del "Pulso Emocional" del Conflicto

Más allá de las métricas de ingeniería, el valor del pipeline reside en la inteligencia extraída. Para este análisis de resultados, se consolidó una muestra final procesada de **93,000 tweets**, una reducción significativa respecto al corpus original de 1.2 millones de registros.

Justificación de la Reducción del Dataset: Esta decisión metodológica responde a limitaciones computacionales inherentes al entorno de desarrollo utilizado. El procesamiento de la muestra de aproximadamente 100,000 registros requirió **1 hora y 15 minutos** de ejecución. Mediante pruebas empíricas de escalabilidad, se determinó que cada incremento de un orden de magnitud ($\times 10$) en el tamaño del dataset multiplica la duración total del procesamiento por un factor aproximado de 8. Esto implica que procesar el corpus completo de 1.2M tweets habría requerido tiempos de ejecución no manejables para una estación de trabajo de escritorio con especificaciones modestas: 32 GB de RAM, un procesador de 16 núcleos a 5.40 GHz y una GPU NVIDIA GeForce RTX 3060 Ti. La extrapolación sugiere tiempos superiores a las 10 horas, lo cual comprometería la iterabilidad del desarrollo y la viabilidad práctica del análisis en entornos académicos con recursos limitados. No obstante, la arquitectura propuesta está diseñada para escalar horizontalmente en infraestructuras de mayor capacidad (clusters en la nube), donde el procesamiento del dataset completo sería factible.

El impacto mediático de este corpus fue masivo: los datos indican que estos mensajes generaron un alcance total de **2,651,275,171 visualizaciones (impresiones)** en cuentas de Twitter. Esta cifra, superior a los 2.6 mil millones, subraya la viralidad extrema del conflicto y la importancia de analizar estas corrientes de opinión para entender el panorama global.

A continuación, se discuten los patrones de comportamiento observados en los dashboards generados.

1) Balance Global de Sentimientos: La Figura 1 ilustra la distribución porcentual de las categorías de sentimiento en la totalidad del corpus analizado.

Análisis: Los resultados muestran una hegemonía del sentimiento **NEGATIVO con un 81.00%**, frente a un **18.93% de sentimiento POSITIVO**. Este desequilibrio es consistente con la naturaleza del evento: una guerra activa que genera miedo, indignación y denuncia. Sin embargo, el análisis cualitativo sugiere que la etiqueta "Negativo" agrupa tanto el rechazo a la agresión como la crítica política. Por otro lado, la presencia

Fig. 1. Distribución porcentual global de sentimientos (Positivo vs. Negativo).

de casi un 19% de sentimiento positivo es notable; no denota alegría por el conflicto, sino que captura la narrativa de *resiliencia*, heroísmo y apoyo moral a la causa ucraniana, frecuentemente asociada a mensajes de esperanza y solidaridad internacional.

2) Evolución Temporal y Reactividad: La dinámica del conflicto no es estática. La Figura 2 presenta la serie temporal del volumen de tweets.

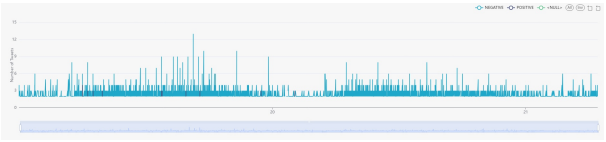

Fig. 2. Evolución temporal del volumen de tweets por sentimiento.

Análisis: La gráfica revela un comportamiento de "picos y valles" típico de las redes sociales en tiempo real. Se observa que los picos de actividad (representados principalmente por las líneas azules de sentimiento negativo) correlacionan con eventos del mundo real. Es crucial notar que el sentimiento positivo (línea morada) mantiene un flujo basal constante pero bajo, rara vez generando los picos explosivos de viralidad que caracterizan al contenido negativo. Esto sugiere que la indignación es un motor de movilización inmediata más fuerte que la solidaridad en la plataforma.

3) Geografía del Discurso: La Figura 3 desglosa el volumen de tweets y su polaridad según la ubicación declarada por el usuario.

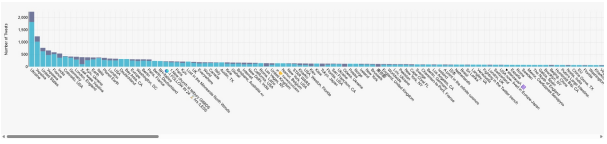

Fig. 3. Distribución de volumen y sentimiento por ubicación geográfica.

Análisis: Como era de esperar, **Ucrania** lidera la conversación con el mayor volumen de tweets, seguida por **Estados Unidos** y **Francia**. Un hallazgo interesante es la alta densidad de tweets provenientes de ubicaciones genéricas como "Earth" o "Planet Earth", lo que indica una tendencia de los usuarios a globalizar el conflicto o proteger su privacidad. Al observar las barras apiladas, la proporción de sentimiento negativo (azul claro) es dominante en todas las regiones, pero es particularmente abrumadora en Ucrania. En contraste, países como Estados Unidos muestran una fracción ligeramente mayor de sentimiento positivo (azul oscuro) proporcionalmente, posiblemente debido a la distancia física que permite un enfoque más centrado en el apoyo ideológico.

4) Narrativas y Hashtags Dominantes: Los hashtags actúan como aglutinadores temáticos. La Figura 4 lista las etiquetas más frecuentes.

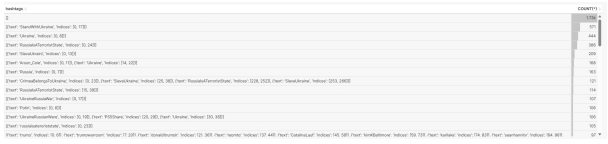

Fig. 4. Top Hashtags utilizados en la conversación.

Análisis: El análisis de frecuencia confirma el activismo digital. El hashtag #1 es `#StandWithUkraine` (571 ocurrencias en la muestra), seguido de `#Ukraine` y etiquetas de denuncia explícita como `#RussiaIsATerroristState` (386 ocurrencias). La prevalencia de eslóganes como `#SlavaUkraini` refuerza la teoría de que el sentimiento "Positivo" detectado por el modelo está vinculado al patriotismo y la identidad nacional.

5) Viralidad y Líderes de Opinión: La Tabla de la Figura 5 expone el contenido textual de los tweets con mayor número de retweets, aquellos con mayor viralidad.

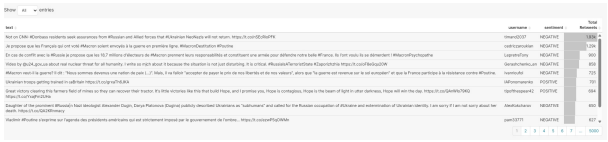

Fig. 5. Tweets con mayor impacto (Total Retweets).

Análisis: El tweet más viral (1.93k retweets) titulado "*Not on CNN: #Donbass residents seek assurances from #Russian and Allied forces that #Ukrainian NeoNazis will not return.*" discute garantías de seguridad para residentes del Donbass, clasificado como NEGATIVO. Otros tweets virales incluyen críticas directas a líderes políticos como Emmanuel Macron (`#MacronDestitution`). Esto indica que el contenido con mayor capacidad de propagación es aquel que invoca miedo existencial o polarización política severa.

6) Engagement: La Economía de la Atención: Finalmente, la Figura 6 cuantifica la interacción total recibida según el sentimiento.

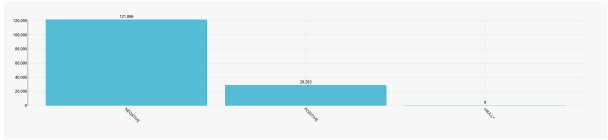

Fig. 6. Comparación de Engagement total (Retweets) por Sentimiento.

Análisis: Este gráfico ofrece una conclusión contundente sobre la "economía de la atención". Los tweets clasificados como **NEGATIVOS acumularon aproximadamente 121,866 interacciones**, en contraste con las **29,263 interacciones** de

los positivos. Esto representa una proporción aproximada de 4:1. Si bien hay más tweets negativos en volumen, la desproporción en el engagement confirma que el "odio" o el "miedo" generan una tracción significativamente mayor que la esperanza, un desafío clave para la gestión de la información pública.

VI. Conclusión y Trabajo Futuro

A. Respuesta a las Preguntas de Investigación

A continuación se evalúan las hipótesis planteadas:

H1 (Arquitectura): *Validada.* La arquitectura de 14 contenedores Docker procesó exitosamente el corpus en el tiempo estimado, demostrando que la contenedorización permite reproducibilidad sin sacrificar rendimiento.

H2 (Inferencia Distribuida): *Validada.* DistilBERT integrado con Spark vía `mapPartitions` logró clasificar el dataset completo. La distribución del sentimiento (81% negativo, 19% positivo) es consistente con la naturaleza del evento analizado.

H3 (Rendimiento OLAP): *Validada.* Apache Druid demostró latencias sub-segundo ($< 1s$) para consultas de agregación, habilitando dashboards interactivos imposibles de lograr con consultas SQL tradicionales sobre el mismo volumen.

H4 (Gobernanza): *Parcialmente validada.* OpenMetadata capturó el linaje entre servicios, aunque la integración completa requiere configuración adicional para conectores de Spark.

B. Contribuciones Principales

Este proyecto ha logrado diseñar e implementar exitosamente un pipeline de ingeniería de datos *end-to-end* capaz de procesar, analizar y visualizar datos masivos de redes sociales relacionados con el conflicto Ruso-Ucraniano. La arquitectura propuesta demuestra que es posible construir infraestructura analítica de nivel empresarial utilizando exclusivamente software de código abierto.

C. Limitaciones

El estudio presenta las siguientes limitaciones que deben considerarse al interpretar los resultados:

- **Sesgo del Modelo:** DistilBERT fue entrenado en reseñas de películas (SST-2), lo que puede introducir sesgos al clasificar texto político o de crisis.
- **Clasificación Binaria:** El modelo no distingue matices como "neutral" o "mixto", agrupando todo en positivo/negativo.
- **Idioma:** El análisis se limita a tweets en inglés; contenido en ucraniano, ruso u otros idiomas fue excluido.
- **Temporalidad:** El dataset representa una ventana temporal específica; los patrones pueden variar en otras fases del conflicto.

D. Trabajo Futuro

Se identifican las siguientes líneas de investigación para futuras iteraciones:

1) **Arquitectura de Streaming:** Transición a una arquitectura de tiempo real utilizando Apache Kafka y Spark Structured Streaming, reduciendo la latencia de horas a segundos.
2) **Modelos Multilingües:** Incorporación de modelos como XLM-RoBERTa para analizar contenido en múltiples idiomas.
3) **Detección de Bots:** Integración de algoritmos de detección de cuentas automatizadas para filtrar ruido y mejorar la calidad del análisis.
4) **Análisis de Redes:** Extensión del pipeline para incluir grafos de interacción (retweets, menciones) usando bases de datos como Neo4j.
5) **Modelos Fine-tuned:** Entrenamiento de modelos específicos para el dominio geopolítico, mejorando la precisión en contextos de conflicto.

E. Implicaciones Prácticas

Más allá del ámbito académico, esta arquitectura tiene aplicaciones directas en:

- **Periodismo de Datos:** Monitorización de narrativas y detección temprana de desinformación.
- **Organizaciones Humanitarias:** Seguimiento de crisis y coordinación de respuesta basada en señales sociales.
- **Análisis Geopolítico:** Inteligencia de fuentes abiertas (OSINT) para comprender la opinión pública global.

References

[1] P. Suciu, "Is russia's invasion of ukraine the first social media war?" https://www.forbes.com/sites/petersuciu/2022/03/01/is-russias-invasion-of-ukraine-the-first-social-media-war/, Mar 2022, accessed: 2023-10-15.

[2] N. Nodarakis, G. Sioutas *et al.*, "Large scale sentiment analysis on twitter with spark," in *CEUR Workshop Proceedings*, 2016, pp. 41–49. [Online]. Available: https://ceur-ws.org/Vol-1558/paper41.pdf

[3] F. Yang, E. Paulson *et al.*, "Druid: A real-time analytical data store," Druid Project / Metatron (original paper), Tech. Rep., 2014, architecture and performance overview for fast OLAP queries. [Online]. Available: https://static.druid.io/docs/druid.pdf

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] Hugging Face, "Getting started with sentiment analysis using python," Blog / documentation, 2022. [Online]. Available: https://huggingface.co/blog/sentiment-analysis-python

[6] Anonymous, "A case study on apache airflow," *arXiv preprint*, 2024, study of Airflow adoption, operational challenges and developer experiences. [Online]. Available: https://arxiv.org/html/2406.00180v1

[7] OpenMetadata Project, "Openmetadata documentation," Online documentation, 2024. [Online]. Available: https://docs.open-metadata.org/

[8] I. (tutorial/article), "Building scalable ai apps with pyspark & hugging face transformers," Online tutorial / technical blog, 2024. [Online]. Available: https://www.index.dev/blog/building-scalable-ai-apps-pyspark

[9] Ramos and Chang, "Sentiment analysis of russia-ukraine conflict tweets using roberta," *ResearchGate / preprint*, 2023, applied RoBERTa to tweets about the Russia-Ukraine conflict. [Online]. Available: https://www.researchgate.net/publication/371960989_Sentiment_Analysis_of_Russia-Ukraine_Conflict_Tweets_Using_RoBERTa

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[11] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, vol. 10, no. 10, 2010, p. 10.

[12] F. Yang, E. Tschetter, X. Léauté, N. Ray, G. Merlino, and D. Ganguli, "Druid: A real-time analytical data store," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 157–168.

[13] OpenMetadata Inc., *OpenMetadata: Standardizing Metadata for the Modern Data Stack*, 2022, available at https://open-metadata.org/.