

**UNIVERSIDAD EUROPEA MIGUEL
DE CERVANTES**

ESCUELA POLITECNICA SUPERIOR

**TITULACIÓN:
MÁSTER EN GESTIÓN Y ANÁLISIS DE
GRANDES VOLÚMENES DE DATOS: BIG
DATA**



TRABAJO FIN DE MÁSTER

**ANÁLISIS DE LOS
COMENTARIOS EN
YOUTUBE CON RELACIÓN A
LA GUERRA EN UCRANIA
EN EL AÑO 2024**

AUTOR

VLADYSLAV DODONOV

TUTOR

Fernando Alonso Pastor

VALLADOLID, 16 DE OCTUBRE DE 2025

Contenido

| | | |
|-----|--|----|
| 1. | Objetivos del trabajo | 3 |
| 2. | Análisis de la situación | 4 |
| 2.1 | Descripción del conflicto armado 2022-2025 | 4 |
| 2.2 | Historia general del conflicto (2022-2025) | 4 |
| 2.3 | Eventos clave de 2024: batallas, acuerdos y cambios territoriales..... | 8 |
| 2.4 | Cobertura mediática, desinformación y opinión pública..... | 10 |
| 2.5 | Identificación de los Canales relevantes en España. | 12 |
| 3. | Obtención, procesado y almacenamiento de los datos | 15 |
| 3.1 | Diseño de Proyecto | 15 |
| 3.2 | Obtención de información desde la API de YouTube | 19 |
| 3.3 | Limitaciones y observaciones | 21 |
| 3.4 | Almacenamiento de los datos | 22 |
| 3.5 | GitHub - versionamiento y documentación | 24 |
| 4. | Análisis exploratorio | 24 |
| 4.1 | Limpieza de datos y exploración | 24 |
| 4.2 | Visualizaciones por canal y tipo | 25 |
| 4.3 | Visualizaciones de series históricas | 27 |
| 4.4 | Nubes de Palabras | 30 |
| 4.5 | Identificación de insultos - Análisis de polaridad | 32 |
| 5. | Diseño e implementación de los modelos..... | 33 |
| 5.1 | Guía de criterios para la clasificación de comentarios | 33 |
| 5.2 | Modelo base de clasificación automática | 35 |
| 5.3 | Implementación del modelo basado en redes neuronales | 36 |
| 5.4 | Integración híbrida de la clasificación | 42 |
| 5.5 | Enriquecimiento del dataset final (bots, argumentos, país...)..... | 43 |
| 6. | Análisis de los resultados obtenidos | 45 |
| 7. | Conclusiones y planes de mejora | 48 |
| 7.1 | Conclusiones generales | 48 |
| 7.2 | Posibles mejoras y consideraciones | 49 |
| 8. | Bibliografía | 51 |
| 9. | Anexo con el código fuente desarrollado. | 55 |

1. Objetivos del trabajo

El objetivo de este trabajo consiste en analizar, aplicando métodos propios de ciencia de datos, los comentarios extraídos de los videos de Youtube y que tienen relación con la Guerra de Ucrania - Rusia. En particular se toman los videos publicados en el año 2024 y en español. En particular se hace énfasis en *youtubers* españoles y que cuentan con una base de seguidores en el país. Lo que se quiere analizar es una muestra de los comentarios que aparecen en los canales preseleccionados, aplicando metodologías NLP (Análisis de Lenguaje Natural). El objetivo es poder clasificar su postura, su polaridad, describir los patrones discursivos que allí aparecen y argumentos que se dejan entrever. Para complementar el análisis, también se suman las cuentas de los noticieros, que, aunque no representan la neutralidad, sí son un campo de lucha discursiva en común. Parte del objetivo es entender que posturas y discursos predominan en dichos espacios, y abrir una línea de investigación futura para compararla con otros países. También, se intentarán identificar la concentración de comentarios y cuentas que asimilan comportamientos de bots o en “núcleo duro”. Se intentará dar respuesta a las preguntas tales como: ¿Qué actores (personas, agrupaciones políticas y civiles, ONGs, etc.) son los que más aparecen nombradas en los discursos enfrentados? ¿En el año 2024, qué noticia fue la que disparó el mayor interés y debate relacionados a la guerra? ¿Qué tan radicalizados se encuentran los usuarios al comentar? Finalmente, al adentrarnos en los discursos que se encuentran frente a una realidad empírica de la guerra, nos preguntaremos por sus motivaciones que dejan entrever en sus argumentos. ¿Qué peso ocupan al considerar su posición la correcta?

2. Análisis de la situación

2.1 Descripción del conflicto armado 2022-2025

La guerra en Ucrania (2022-2025) representa el conflicto armado más importante en Europa desde la Segunda Guerra Mundial, con profundas implicaciones geopolíticas, estratégicas y sociales. Iniciada con la invasión rusa a gran escala en febrero de 2022, la contienda ha dejado decenas de miles de víctimas y millones de desplazados, al tiempo que ha transformado el panorama de la seguridad europea y polarizado la opinión pública mundial. En el ámbito de la comunicación, esta guerra ha estado marcada por una intensa cobertura mediática internacional, una avalancha de información (y desinformación) en redes sociales, y una lucha narrativa entre los actores involucrados. Este trabajo ofrece un análisis de la comunicación a través de la red social YouTube con énfasis en los hechos clave de 2024, los principales actores en juego, la cobertura mediática y el ecosistema informativo en torno a la guerra. Asimismo, se justifica por qué este conflicto se erige como un campo fértil para el análisis comunicacional, dado que los medios de comunicación y las redes sociales se han convertido en escenarios estratégicos de disputa discursiva. A continuación, se presenta una descripción general de la guerra de Ucrania, seguida de secciones dedicadas a los eventos de 2024, al rol de los actores principales, al tratamiento mediático del conflicto (incluyendo fenómenos como la desinformación y las tendencias de búsqueda en el mundo hispanohablante), y finalmente a la relevancia comunicacional del caso ucraniano.

2.2 Historia general del conflicto (2022-2025)

El conflicto armado entre Rusia y Ucrania escaló dramáticamente el 24 de febrero de 2022, cuando Rusia lanzó una invasión militar a gran escala del territorio ucraniano. En los primeros días, las fuerzas rusas atacaron múltiples ciudades con misiles y

avanzaron desde el norte (Bielorrusia), el este (Donbás) y el sur (Crimea), con el aparente objetivo de tomar Kiev rápidamente, derrocar al gobierno prooccidental de y devolver a Ucrania a la esfera de influencia rusa (Kirby, 2025). La ofensiva inicial causó graves daños a infraestructura civil como los hospitales, escuelas y viviendas fueron alcanzados por bombardeos indiscriminados y numerosas bajas entre la población civil. Según Human Rights Watch (2022), en la primera semana de hostilidades más de un millón de personas huyeron de sus hogares, muchas refugiándose en países vecinos, mientras en Rusia se intensificaba la censura y se reprimían protestas internas contra la guerra. La invasión violó abiertamente el derecho internacional y sentó las bases de lo que sería un conflicto prolongado y devastador. A pesar de los avances iniciales, el plan ruso de una victoria rápida fracasó. Para marzo de 2022, la resistencia ucraniana -respaldada por suministros de armas de Occidente- logró detener el asalto sobre Kiev, obligando a las tropas rusas a retirarse del norte de Ucrania. Rusia entonces concentró su esfuerzo en el Donbás (este de Ucrania) y el sur donde sus avances fueron significativos. Durante 2022, las fuerzas rusas tomaron la ciudad portuaria de Mariúpol tras un sitio de tres meses, y llegaron a ocupar alrededor del 22% del territorio ucraniano, incluyendo la península de Crimea (anexada en 2014) y partes de las regiones de Donetsk, Lugansk, Zaporiyia y Jersón. Sin embargo, el ejército ucraniano llevó a cabo contraofensivas efectivas en la segunda mitad de 2022: en septiembre reconquistó amplias zonas de la provincia de Járkov, y en noviembre recuperó la ciudad de Jersón, la única capital regional que Rusia había logrado capturar desde la invasión de 2022. Estas victorias ucranianas demostraron las debilidades logísticas y de moral de los rusos, a la vez que dieron impulso a Ucrania en el frente diplomático, consolidando el apoyo occidental. En 2023, la guerra entró en una fase de desgaste. Rusia nombró nuevos comandantes y lanzó ofensivas concentradas en el Donbás, logrando avances muy costosos en localidades como Soledar y Bajmut tras meses de combates encarnizados.

La ciudad de Bajmut cayó en manos rusas en mayo de 2023, en lo que Moscú proclamó como una victoria significativa, aunque estratégica y simbólicamente limitada. Ucrania, por su parte, recibió sistemas de armamento más avanzados (artillería de largo alcance, defensas antiaéreas, tanques occidentales) y en junio de 2023 inició una contraofensiva en los frentes sur y este. No obstante, esta contraofensiva ucraniana avanzó lentamente frente a las densas líneas defensivas rusas (campos minados, trincheras y “dientes de dragón”) y la superioridad aérea rusa. Hacia finales de 2023, el frente se había estabilizado con solo cambios territoriales marginales: ninguna de las dos partes consiguió una ruptura decisiva, y el conflicto se estancó en una guerra de atrición, con combates intensos, pero líneas relativamente fijas. Para 2024, el equilibrio militar seguía siendo frágil pero relativamente estable en términos territoriales. Rusia aún ocupaba aproximadamente una quinta parte del territorio ucraniano, incluyendo la mayor parte de las regiones orientales de Donetsk y Lugansk y amplias zonas del sur (Zaporiyia y Jersón al este del río Dniéper), además de Crimea. Ucrania conservaba el control de las ciudades principales (Kiev, Járkov, Odesa, Leópolis, etc.) y había logrado proteger con éxito su capital y la mayor parte del centro-occidente del país. Las hostilidades continuaban en extensos frentes de batalla de más de 1.000 km de longitud, especialmente en Donbás. Como resume un informe de BBC News, en el este “la maquinaria de guerra de Moscú avanza lentamente, apenas unos metros a costa de grandes pérdidas, mientras la línea del frente ha cambiado muy poco en dos años”. Ambas partes parecían desgastadas por la contienda prolongada, pero ninguna daba señales de ceder en sus objetivos fundamentales.



Figura 1. Mapa del este de Ucrania que muestra en rojo las zonas bajo ocupación militar rusa (aprox. 20% del país) hacia mayo de 2025. Las regiones sombreadas indican los límites del avance ucraniano antes del estancamiento del frente. A esa fecha, la guerra había dejado al menos 12.654 civiles muertos y 29.392 heridos confirmados por la ONU desde 2022 (incluyendo 673 niños) (Oficina de la ONU para los Derechos Humanos, 2025). Además, cerca de 7 millones de ucranianos se convirtieron en refugiados en Europa y otras regiones, y unos 3,7 millones eran desplazados internos, configurando la mayor crisis humanitaria en Europa en décadas. Grandes extensiones de Ucrania (alrededor de 139.000 km²) han quedado contaminadas con minas y municiones sin detonar. La dimensión destructiva y humana del conflicto ilustra su escala: Ucrania estima pérdidas militares propias en decenas de miles de soldados, y atribuye a Rusia pérdidas aún mayores, si bien las cifras exactas son objeto de propaganda y difícil verificación (NYT, 2023; Le Grand Continent, 2025). En cuanto a los esfuerzos diplomáticos, hasta 2025 no se había logrado una negociación de paz efectiva. La ONU condenó la invasión en resoluciones de la Asamblea General con amplio respaldo, pero Rusia -miembro permanente del Consejo de Seguridad- bloqueó cualquier acción en ese órgano. Se sucedieron

intentos de mediación o planes de paz propuestos por diversos actores (ONU, Turquía, Francia, Santa Sede, China, entre otros), sin consenso. A mediados de 2023, se alcanzó y luego colapsó un acuerdo para la exportación de granos ucranianos vía el Mar Negro bajo mediación de Turquía y la ONU, reflejando la fragilidad de pactos limitados en medio de la desconfianza. Hacia finales de 2024 e inicios de 2025, surgieron informes sobre contactos exploratorios -incluyendo la posibilidad de negociaciones indirectas entre Estados Unidos y Rusia-, pero Ucrania insistió en su exigencia de retirada rusa como condición previa (CFR, 2025). En suma, al cumplirse tres años de guerra, el conflicto seguía abierto y sin un final claro, con las partes atrincheradas en sus posiciones militares y políticas.

2.3 Eventos clave de 2024: batallas, acuerdos y cambios territoriales

El año 2024 consolidó la guerra de Ucrania como un conflicto prolongado y de desgaste, sin avances decisivos, pero con episodios de alta repercusión militar, política y simbólica. Durante este periodo, se registraron hechos que definieron tanto la dinámica bélica en el terreno como la percepción internacional del conflicto. El primer hecho significativo del año fue el bombardeo sobre un mercado en la ciudad de Donetsk el 21 de enero, que dejó decenas de víctimas civiles. El ataque, atribuido por las autoridades locales a fuego ucraniano, generó un intercambio de acusaciones entre Kiev y Moscú y se convirtió en un símbolo de la escalada en zonas urbanas habitadas (Reuters, 2024a).

Pocas semanas después, el 16 de febrero, se confirmó la muerte del líder opositor ruso Alexéi Navalni en una colonia penitenciaria en el Ártico. Su fallecimiento provocó condenas internacionales y nuevas sanciones contra Rusia, y reactivó la narrativa occidental sobre la represión política interna del Kremlin (BBC News, 2024).

El 18 de febrero, la caída de Avdiivka marcó el acontecimiento militar más relevante de la primera mitad del año. Tras meses de asedio y bombardeos, las fuerzas ucranianas se retiraron de la localidad para evitar el cerco, lo que permitió a Rusia proclamar la captura de la ciudad y presentar el hecho como una victoria táctica y propagandística. Analistas del Institute for the Study of War (ISW, 2024) describieron la batalla como “la mayor ganancia territorial rusa desde Bajmut en 2023”, aunque lograda a un alto costo en vidas.

En abril, el Congreso de los Estados Unidos aprobó un nuevo paquete de ayuda militar y económica a Ucrania por 61 mil millones de dólares, tras meses de bloqueo legislativo (U.S. Congressional Research Service, 2024). Esta decisión aseguró la continuidad del apoyo occidental y reafirmó el liderazgo estadounidense en la coalición aliada.

El 10 de mayo, Rusia lanzó una ofensiva en el óblast de Járkov, intentando abrir un nuevo frente en el noreste. Aunque logró ocupar algunas aldeas fronterizas, las fuerzas ucranianas estabilizaron la línea defensiva, evitando avances profundos (ISW, 2024).

El 8 de julio, un ataque con misiles rusos impactó el hospital infantil Okhmatdyt en Kiev, causando víctimas civiles y conmoción internacional. Naciones Unidas y la Unión Europea condenaron el ataque como una posible violación del derecho internacional humanitario (Council on Foreign Relations, 2025).

En agosto, unidades ucranianas y grupos de voluntarios rusos opositores al Kremlin realizaron una incursión en la región rusa de Kursk, atacando infraestructura y posiciones fronterizas. Este episodio extendió el conflicto más allá de las fronteras de Ucrania y evidenció vulnerabilidades en la defensa rusa (House of Commons Library, 2024).

Durante octubre se intensificaron las operaciones en el frente oriental. El 2 de octubre, Rusia anunció la toma de Vuhledar, localidad minera en Donetsk que había resistido durante más de un año. Posteriormente, en la segunda quincena de octubre, los analistas registraron el mes con mayores avances rusos del año, estimados en unos 200 km² adicionales, concentrados en los oblasts de Donetsk y Lugansk (ISW, 2024).

El 10 de noviembre, Ucrania ejecutó el mayor ataque con drones sobre Moscú hasta la fecha, alcanzando instalaciones industriales y militares en la periferia de la capital. Estos ataques demostraron la creciente capacidad tecnológica ucraniana para proyectar fuerza a larga distancia y tuvieron un efecto simbólico considerable (Council on Foreign Relations, 2025).

Finalmente, el 8 de diciembre, se produjo un giro geopolítico inesperado con la caída del régimen sirio de Bashar al-Ásad y su huida a Moscú, lo que reconfiguró las alianzas de Rusia en Medio Oriente y sumó presión internacional al Kremlin (Reuters, 2024b).

En conjunto, los eventos de 2024 confirmaron la persistencia de un conflicto de alta intensidad y escaso movimiento estratégico. Rusia consolidó modestos avances territoriales, mientras que Ucrania mantuvo su capacidad de resistencia y expandió operaciones asimétricas. En el plano internacional, la adhesión de Suecia a la OTAN y el fortalecimiento del apoyo occidental reforzaron la percepción de que la guerra había entrado en una fase prolongada de confrontación estructural entre bloques geopolíticos (House of Commons Library, 2024).

2.4 Cobertura mediática, desinformación y opinión pública

Desde el inicio de la guerra en 2022, el conflicto en Ucrania fue ampliamente cubierto por medios internacionales, dominando titulares y generando gran interés público. Durante los primeros meses, cadenas como BBC, CNN, *El País* o *The New*

York Times ofrecían actualizaciones constantes, reflejando tanto la gravedad del conflicto como la demanda informativa. Sin embargo, con el paso del tiempo, la cobertura disminuyó y se centró solo en eventos impactantes. Un estudio de Diez-Gracia (2024) mostró que la atención mediática cayó notablemente después del primer semestre de 2022, y muchas noticias sobre la guerra tuvieron bajo nivel de lectura y difusión, evidenciando fatiga informativa.

En España, la guerra fue intensamente cubierta al inicio, con gran presencia en medios y alto interés ciudadano. No obstante, el conflicto fue perdiendo protagonismo frente a otras crisis, como la guerra entre Israel y Hamás en 2023, que desplazó el foco informativo en países como España e Italia (Le Grand Continent, 2023). Aun así, la cobertura repuntó en ciertos momentos políticos clave, como la presidencia española del Consejo de la UE.

El conflicto también ha estado marcado por una intensa guerra informativa. Rusia desplegó una estrategia sistemática de desinformación, utilizando medios estatales como RT y Sputnik, así como redes sociales, para justificar la invasión y dividir la opinión pública. Pese a estar bloqueados en Europa, estos medios han prosperado en español, especialmente en América Latina, aprovechando el déficit de medios alternativos y el escepticismo hacia EE. UU. (Reuters Institute, 2023). En España, aunque más restringidos, estos contenidos han circulado por canales de Telegram o medios marginales.

Ucrania, por su parte, ha usado la comunicación estratégica principalmente para ganar apoyo internacional, promoviendo contenidos que destacan la resistencia civil y militar, con iniciativas de verificación como StopFake. Si bien también ha empleado elementos propagandísticos, su enfoque ha sido más defensivo y centrado en la visibilidad global.

Las redes sociales han jugado un rol central: Twitter, Facebook, Telegram y TikTok han sido espacios de difusión de información veraz y falsa. Plataformas como TikTok

dieron lugar a lo que algunos denominaron la “primera guerra TikTok”, por la abundancia de vídeos desde el frente, muchos de ellos virales pero difíciles de verificar (Hasan, 2024). En YouTube, los comentarios en vídeos sobre la guerra reflejan una clara polarización ideológica: desde apoyos fervientes a Ucrania hasta discursos prorrusos con fuerte carga propagandística. El interés público también se reflejó en las búsquedas en Google. En 2022, “Ucrania” fue uno de los términos más buscados en España. Sin embargo, en 2023 desapareció del Top 10, cediendo espacio a otros temas como elecciones o conflictos más recientes. Aun así, encuestas han mostrado que la mayoría de los ciudadanos españoles siguen apoyando la ayuda a Ucrania (Real Instituto Elcano, 2023), mientras en América Latina las posturas están más divididas.

En resumen, la guerra de Ucrania ha sido una batalla no solo militar, sino informativa. La evolución de su cobertura, el impacto de la desinformación, y la disputa narrativa en redes la convierten en un caso ejemplar para el análisis comunicacional contemporáneo.

2.5 Identificación de los Canales relevantes en España.

El análisis de canales de YouTube en el contexto español se enmarca en la transformación de la esfera pública digital, donde los medios tradicionales y los creadores de contenido compiten por la atención, el sentido y la legitimidad del discurso político (Cardón, 2010; Castells, 2012). En este ecosistema híbrido, los canales institucionales o de medios masivos representan, como señalan McCombs y Shaw (1972), un espacio de disputa simbólica donde se configura la agenda pública y se negocia el significado de los acontecimientos internacionales. Aunque estos medios pueden reflejar la opinión dominante en el país, condicionada por marcos ideológicos y políticos, también constituyen el espacio de lucha hegemónica donde

distintos actores sociales disputan la interpretación legítima de los hechos (Gramsci, 1971; Hall, 1980).

En este sentido, la clasificación de los canales en pro-ucranianos, noticieros y pro-rusos responde a una lógica analítica que busca captar la pluralidad de narrativas presentes en el ecosistema informativo español. Los medios tradicionales, como RTVE, El País, El Mundo, La Vanguardia o laSexta Noticias, tienden a alinearse con el marco europeo y atlántico de condena a la invasión rusa, reproduciendo marcos institucionales vinculados a los derechos humanos y al derecho internacional. No obstante, su papel no puede reducirse a una homogeneidad ideológica, ya que en ellos coexisten diferentes voces y tensiones internas que expresan los límites del consenso público (Sampedro, 2016; Calvo & Aruguete, 2020).

Por otro lado, canales de orientación prorrusa como Intereconomía, Negocios TV o Miguel Ruiz Calvo representan la aparición de espacios alternativos de contrainformación que cuestionan los discursos dominantes occidentales. Estas producciones articulan narrativas de desconfianza hacia la OTAN y la Unión Europea, y suelen apoyarse en argumentaciones geopolíticas y económicas que reconfiguran el sentido del conflicto. En términos de la teoría de la polarización afectiva negativa (Sarsfield & Abuchanab, 2024; Iyengar et al., 2019), tales canales refuerzan identidades políticas divergentes y movilizan emociones que consolidan comunidades digitales cohesionadas en torno a posiciones ideológicas específicas (Wilches-Tinjacá et al., 2024).

El caso español comparte con otros contextos, como el argentino analizado por Calvo (2015) y Aruguete (2020), la coexistencia de burbujas informativas o cámaras de eco (Pariser, 2011), donde la exposición selectiva al contenido mediático refuerza creencias preexistentes. Sin embargo, YouTube también funciona como un espacio de intersección discursiva (Verón, 1984), en el que los canales de medios masivos actúan como lugares de encuentro entre públicos diversos. Por eso, su inclusión en esta

investigación no implica asumir neutralidad, sino reconocerlos como arenas de interacción plural donde la disputa por la hegemonía discursiva y el humor social se hace visible en los comentarios de los usuarios.

Memorias de Pez ofrece contenidos educativos semanales sobre historia y actualidad. Durante 2024-2025 ha cubierto la guerra en su serie “La Pecera de Memorias”, mostrando simpatía hacia Ucrania y criticando las acciones rusas, especialmente en el trato a soldados y civiles.

RTVE Noticias ofrece cobertura constante y objetiva desde el servicio público. La guerra de Ucrania ha sido uno de sus temas más vistos. Aunque mantiene neutralidad institucional, adopta términos como “invasión rusa” y refleja el marco europeo de condena.

El País y El Mundo son diarios de referencia en España. Ambos han producido videos y coberturas especiales sobre la guerra. El País refuerza la narrativa europea contra la invasión, mientras El Mundo se enfoca en análisis legales y diplomáticos sin sesgo aparente.

Miguel Ruiz Calvo es un abogado y streamer con emisiones diarias sobre la guerra. Su discurso es claramente prorruso, amplificando supuestos logros militares de Rusia y criticando duramente a Ucrania y la OTAN.

Rubén Gisbert ha viajado al Donbás y publicado contenidos que cuestionan la versión occidental del conflicto. Ha aparecido en RT y promovido teorías prorrusas, como negar crímenes atribuidos a Rusia.

laSexta Noticias, como medio televisivo privado, ha ofrecido cobertura continua del conflicto con un enfoque informativo y de denuncia humanitaria. Suele utilizar un marco narrativo favorable a Ucrania y crítico con las acciones del Kremlin, especialmente en su línea editorial sobre derechos humanos.

Intereconomía ha difundido durante 2024-2025 contenidos de corte conservador y euroescéptico, incluyendo análisis y tertulias donde se relativiza la invasión rusa o se

atribuye la escalada del conflicto a las políticas de la OTAN. Su narrativa tiende a alinearse con posiciones prorrusas.

La Vanguardia mantiene un enfoque periodístico clásico, con cobertura internacional amplia y lenguaje institucional. Si bien reproduce los marcos comunicativos de la Unión Europea, su tono es moderado y evita posicionamientos ideológicos explícitos. Negocios TV, canal centrado en información económica, ha ganado relevancia en el debate sobre la guerra por sus entrevistas y programas con expertos que presentan perspectivas críticas hacia Occidente y favorables al gobierno ruso. Su narrativa combina temas geopolíticos con un discurso económico alternativo que suele coincidir con líneas prorrusas.

En conjunto, esta selección de canales permite observar cómo el debate digital sobre la guerra de Ucrania en España reproduce patrones de polarización mediática similares a los encontrados en América Latina (Bruns & Highfield, 2018; Benkler et al., 2018). Los medios tradicionales mantienen un papel central en la formación de la opinión pública, ahora coexistiendo con actores híbridos —periodistas independientes, analistas digitales y comunidades de usuarios— que disputan la definición de los hechos y las emociones colectivas en el espacio público de YouTube.

3. Obtención, procesado y almacenamiento de los datos

3.1 Diseño de Proyecto

El diseño del proyecto se estructuró bajo un enfoque modular orientado a la reproducibilidad y a la trazabilidad de todas las etapas del proceso, desde la obtención de los datos hasta la generación del producto final de análisis y visualización. El trabajo se desarrolló en Python, utilizando Visual Studio Code (VSC) como entorno principal de programación y Jupyter Notebooks (.ipynb) como herramienta de desarrollo exploratorio. Esta combinación permitió mantener una

lógica de trabajo iterativa, integrando la documentación del proceso, los resultados parciales y la validación de hipótesis dentro del mismo flujo de ejecución.

De esa forma se siguen las buenas prácticas de la industria.

El entorno de desarrollo se ejecutó en un espacio virtual independiente (venv) que concentró las dependencias necesarias del proyecto. Esta configuración garantiza la portabilidad del entorno y facilita la replicabilidad del análisis en otros equipos o plataformas, siguiendo las recomendaciones de Wilson et al. (2017) sobre buenas prácticas en ciencia de datos. En el archivo *requirements.txt* se documentaron las versiones de las librerías empleadas, asegurando que la instalación y ejecución del proyecto pueda reproducirse de manera controlada en el futuro.

El flujo general de trabajo se diseñó como un pipeline de datos secuencial y controlado, donde cada etapa genera un producto intermedio que sirve de insumo para la siguiente. De esta forma, el proceso combina elementos propios de la ingeniería de datos con un enfoque de investigación social aplicada al análisis de narrativas digitales. La estructura lógica del pipeline se organizó en cuatro fases principales:

- Obtención y captura de datos: se centró en la recolección de información desde la API de YouTube Data v3, extrayendo comentarios, metadatos de videos y características de los canales. Esta fase inicial sentó la base del corpus analítico y permitió establecer la relación entre los contenidos audiovisuales y los eventos políticos y militares relevantes del año 2024.
- Procesamiento y depuración: incluyó las tareas de limpieza, normalización de campos, manejo de duplicados y enriquecimiento con variables adicionales, tales como el número de suscriptores o la fecha de creación de las cuentas que realizan los comentarios.
- Análisis exploratorio y modelado: comprendió la aplicación de técnicas de data mining, análisis lingüístico y clasificación híbrida: utilizando

herramientas manuales y automáticas como las redes neuronales. Esto se aplicó principalmente sobre los comentarios extraídos, con el objetivo de identificar la segmentación clave en su postura frente al conflicto: pro-ruso, pro-ucraniano o neutro. Este análisis se enriqueció al trabajar con los patrones discursivos, insultos y polarización.

- Visualización y presentación: etapa final del pipeline, donde se integraron los resultados procesados en una capa analítica de visualización mediante PowerBI, permitiendo representar gráficamente los ejes argumentativos, los eventos de mayor impacto y la distribución ideológica de los comentarios.

El proyecto se diseñó con el siguiente árbol de carpetas que se explicita en el archivo README.md.

La recolección de datos se realizó mediante la API de YouTube Data v3, a través de un proceso de extracción incremental que permitió descargar de forma controlada los comentarios y metadatos asociados a videos relevantes sobre la guerra en Ucrania. Este procedimiento se ejecutó de manera automatizada en lotes (batches), respetando las cuotas de uso de la API y asegurando la continuidad del flujo de datos.

La primera etapa del pipeline se implementó en el notebook

01_YouTube_2024_Ukraine_data_extracting.ipynb, donde se define la conexión con la API, los parámetros de búsqueda y las rutinas de paginación. Los resultados se almacenaron en formato CSV y JSON (para mapeo si se interrumpe la recolección de datos por el límite de la cuota) dentro de la carpeta data/raw/, generando dos archivos fundamentales:

- *0_comments_raw.csv*: contiene los comentarios y sus metadatos básicos.
- *1_comments_youtube_clean.csv*: metadatos enriquecidos desde diferentes llamadas de las APIs (Users, Channels).
- *video_list_full.json*: guarda la información estructurada de los videos incluidos en el análisis.

A partir de estos datos, se aplicaron limpiezas y mejoras en el notebook *02_EDA_comments.ipynb*, donde se eliminaron duplicados y registros inconsistentes, se analizaron las métricas básicas y se agregaron variables auxiliares. El resultado se consolidó en el archivo *2_comments_youtube_refined.csv*, ubicado en data/processed/ que constituye la base limpia y normalizada para el análisis posterior.

La siguiente fase, ejecutada en *03_sample_comments_classifications.ipynb*, generó varios outputs ya que ahí se prepararon subconjuntos del dataset para el etiquetado manual y la detección de lenguaje ofensivo. Los archivos

4_9000_comments_to_label.xlsx y *5_9000_comments_with_insults.xlsx* representan los puntos de control (checkpoints) de esta etapa, utilizados para el proceso de clasificación híbrida (manual + automática). Finalmente el output general de esta notebook es el que se encuentra en: *6_9000_comments_hibrid_class.xlsx* ya que presenta la clasificación final de la muestra tomada.

La integración de las etiquetas resultantes y su validación se realizaron en el notebook *04_final_classification_core.ipynb*, que genera el archivo *7_final_label.csv*. Posteriormente, en *05_master_dataset_enrichment.ipynb*, se combinan los resultados y se crean las versiones consolidadas del corpus:

- *8_final_master_enriched.csv*: dataset Enriquecido y preparado para análisis de visualización y modelado.
- *pbi_unigrams.csv* y *pbi_bigrams.csv*: salidas derivadas para análisis léxico y visualización en Power BI.
- Todas las tablas de hechos y dimensionales dentro de la carpeta *bi_layer*.

3.2 Obtención de información desde la API de YouTube

La información utilizada en este trabajo se obtuvo mediante la API oficial de YouTube Data v3, que permite acceder a datos públicos como comentarios, metadatos de videos, información de canales y estadísticas de interacción. La implementación técnica se realizó en Python, utilizando la biblioteca *googleapiclient* para gestionar solicitudes autenticadas a través de una clave privada almacenada en un archivo *.env*, garantizando la seguridad y confidencialidad de las credenciales (Google Developers, 2024).

El proceso de extracción fue diseñado para ejecutarse en lotes (batches), dada la limitación de cuotas de la API (10.000 quota units diarias por cuenta). Cada lote recuperó un conjunto de videos asociados a canales previamente identificados y,

para cada video, los comentarios disponibles en formato jerárquico. Esta metodología permitió construir un corpus progresivo y verificable, evitando la pérdida de información y facilitando la recuperación en caso de interrupciones. La primera etapa del pipeline se centró en obtener los comentarios y atributos básicos de cada interacción. Los campos recolectados y su descripción se detallan a continuación:

| Campo | Descripción breve |
|---------------------|---|
| comment_id | Identificador único del comentario. |
| comment | Texto completo del comentario. |
| comment_text_length | Longitud del comentario en caracteres. |
| user_id | ID del usuario o canal que realiza el comentario. |
| user_name | Nombre público del autor del comentario. |
| comment_time | Fecha y hora en que se publicó el comentario. |
| comment_likes | Número de “me gusta” recibidos. |
| total_reply_count | Número de respuestas al comentario. |
| video_title | Título del video asociado. |
| channel_title | Nombre del canal propietario del video. |
| video_published_at | Fecha de publicación del video. |
| video_views | Total de visualizaciones. |
| video_likes | Total de “me gusta” del video. |
| video_duration | Duración del video en formato ISO 8601. |
| video_tags | Lista de etiquetas asignadas al video. |
| video_category_id | Identificador de la categoría del video. |

| | |
|-----------------|---|
| relacion_evento | Indica la vinculación del comentario con un evento político o militar concreto. |
| evento | Nombre del evento relevante (por ejemplo, “Caída de Avdiivka” o “Ataque a hospital infantil”). |
| tipo_evento | Clasificación del evento según su naturaleza (militar, político, geopolítico, simbólico, etc.). |

Además de la información proveniente de la API, se integraron variables personalizadas diseñadas para contextualizar los comentarios dentro del marco analítico de la investigación. Estas variables complementarias fueron creadas manualmente y se incorporaron en las etapas posteriores del procesamiento.

3.3 Limitaciones y observaciones

El proceso de extracción presentó ciertas limitaciones técnicas inherentes a la API de YouTube, como la cuota diaria de peticiones y el límite máximo de comentarios recuperables por video, lo que impacta directamente en la completitud de los datos obtenidos. Se optó, por lo tanto, traer una muestra de los videos (hasta 50 por cuenta de interés) para poder observar los resultados sobre datos controlados. Finalmente, si se pudo mejorar la extracción de los datos (utilizando varios proyectos y api keys para esquivar los límites de cuota diaria) y traer más de 2000 videos con sus comentarios, pero ese dataset quedó guardado para futuras investigaciones. En esta oportunidad se trabajó con la muestra de 499 videos que más vistas y/o interacciones tuvieron en el año 2024. Otra limitación relevante fue la restricción o

bloqueo de comentarios por parte de ciertos canales en videos específicos, algo común en temas controvertidos (Google Developers Policy, 2024). En ese contexto por la API de la llamada pública no pude acceder con la búsqueda tradicional, a los videos de los canales VisualPolitik. Nuevamente como lo comenté, si se pudo realizar con otros métodos, pero quedando fuera del alcance de dicho trabajo. También hubo problemas para ordenar y filtrar efectivamente los videos de interés ya que el nombre por el que se buscaban dichos videos no era altamente restrictivo y trajo videos que además de la temática principal contenía otras no relacionadas. Es el caso que se da fuertemente en el canal Miguel Ruiz Calvo que agrupa sus videos por varias temáticas a la vez. Se decidió conservarlo como parte del objeto de estudio por su relevancia y alta interacción (*engagement*) con los usuarios. También se consideró sumar a Liu Sivaya (liusivaya), pero tenía menos suscriptores e interacciones que Miguel Ruiz Calvo que ya tenía el doble de comentarios que el *youtuber* ucraniano Memorias de Pez.

3.4 Almacenamiento de los datos

Los datos fueron almacenados en archivos CSV, lo que facilitó la manipulación inicial y la realización de análisis exploratorios rápidos. Considerando la escalabilidad del proyecto y la posibilidad de incorporar análisis más avanzados, se prevé una futura migración hacia un sistema de almacenamiento más robusto, como DuckDB, que ofrece una alternativa económica y eficiente para el análisis local con recursos computacionales limitados (Raasveldt & Mühleisen, 2022). Esta solución permitiría escalar, sin costos adicionales, el *pipeline* de datos y acelerar el procesamiento de estos. Realmente por el volumen de lo que se está trabajando no hubo problemas en las lecturas y guardado de los archivos en formatos csv directamente en el repositorio de archivos local. Sin embargo, para el desarrollo de las futuras

investigaciones sí se plantea realizar una migración a un almacén de datos (Data Warehouse) e, inclusive a BigQuery y/o Google Cloud Storage (GCS) que podría suplir todas las necesidades de escalamiento y aprovechar la integración con las APIs de Youtube de forma nativa. Las tablas se podrían importar y exportar y se realizarían las consultas en SQL para la exploración de datos.

Para la solución actual vamos a optar por los archivos locales en formatos .csv y .xlsx (para la parte de clasificación por humanos). La estructura adoptada sigue los principios de la arquitectura *Medallion*, ampliamente utilizada en la ingeniería de datos moderna (Databricks, 2022; O’Leary, 2023). Este enfoque organiza los datos en tres capas:

- Bronze, que almacena los datos en bruto tal como se extraen
- Silver, donde se limpian, transforman y enriquecen
- Gold (bi_layer en nuestro caso), optimizada para el consumo analítico, visualizaciones y modelos predictivos.

Como se observa en la siguiente imagen la distribución de los archivos internos representa una lógica de movimientos intermedios, propio de un ETL y más seguro ante los errores y fallas en los procesos.

```
analysis_guerra_ucrania_youtube/
├── data/
│   ├── raw/                                # Datos originales descargados desde la API de YouTube
│   ├── processed/                           # Datos limpios y enriquecidos
│   └── bi_layer/                            # Tablas finales para visualización en Power BI
```

Aunque en este caso la implementación se realiza sobre archivos locales, la estructura de carpetas y el flujo de transformación reproducen la lógica de un Data Lakehouse modular, garantizando la trazabilidad de los datos y una mayor resiliencia ante errores en los procesos ETL.

Finalmente, las tablas procesadas se integraron en Power BI para la elaboración del dashboard final, complementado con la documentación académica disponible en formato PDF.

3.5 GitHub - versionamiento y documentación

Se utiliza GitHub como sistema de control de versiones para el código fuente del proyecto, garantizando reproducibilidad y trazabilidad de los cambios realizados. El código desarrollado es documentado mediante comentarios claros y detallados que explican las funciones, variables y metodologías aplicadas, siguiendo prácticas recomendadas para proyectos de ciencia de datos. Consecuentemente, además de la presentación formal en los anexos de este trabajo y el Tablero final en PowerBI los resultados se podrán consultar también en ese enlace de GitHub que será compartido en los anexos.

4. Análisis exploratorio

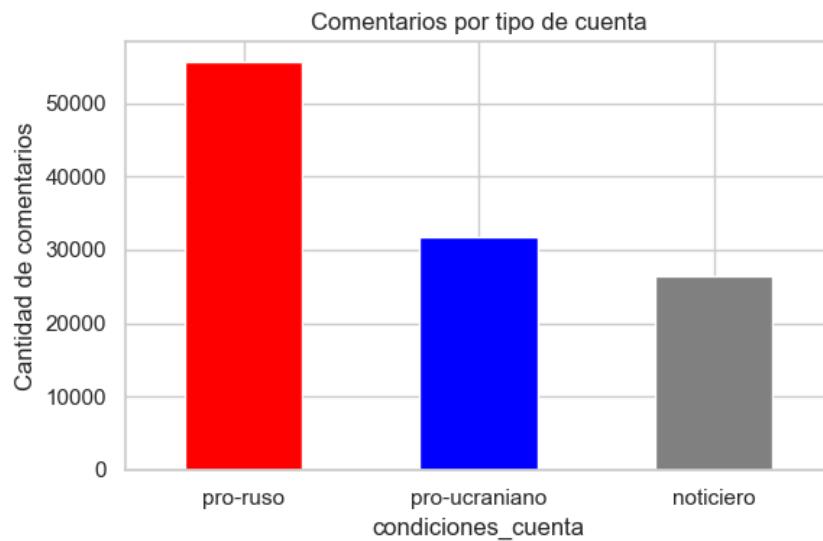
4.1 Limpieza de datos y exploración

El análisis exploratorio se inició con la utilización de la biblioteca Pandas (McKinney, 2022). Inicialmente se hizo la limpieza de los nulos que ya se mencionó para calcular el total de comentarios recuperados por canal y otras distribuciones claves. Por otro lado, se limpiaron los casos de comentarios nulos (vacíos) o casos sin el nombre de usuario. Con métodos `describe()` e `info()` se pudo ver cómo están los datos en cuanto a su calidad y distribución. Para poder visualizar mejor las métricas históricas, se convirtieron en `datetime` las columnas de fechas que venían con otro tipo de dato. Se sumó la columna de la antigüedad de la cuenta: `days_since_account_creation` para

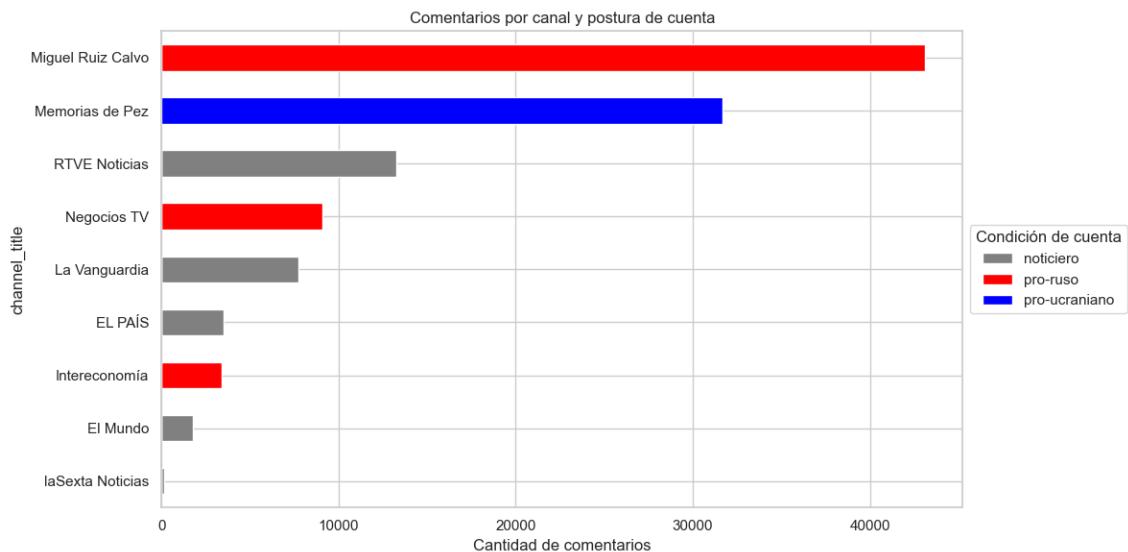
los casos donde teníamos la fecha de la creación de esta.

4.2 Visualizaciones por canal y tipo

Todo esto nos permitió empezar a trabajar con las visualizaciones y obtener una clara visualización inicial del volumen relativo de participación según la orientación editorial del canal.

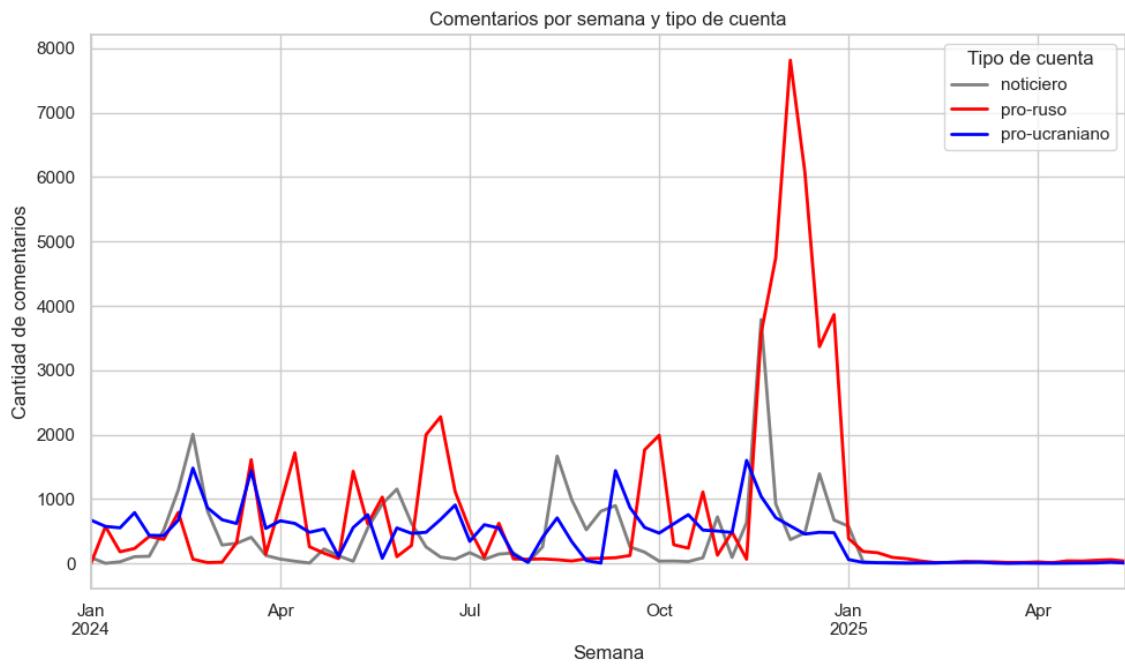


Observamos que, así como están los datos, tenemos mayor presencia de comentarios en los canales prorrusos. Al representar dichos resultados a mayor detalle vemos que el canal prorruso de Miguel Ruiz Calvo está arriba del ranking. Es sorprendente ya que sobrepasa con holgura a Memorias de Pez que tiene hoy cinco veces más de suscriptores.

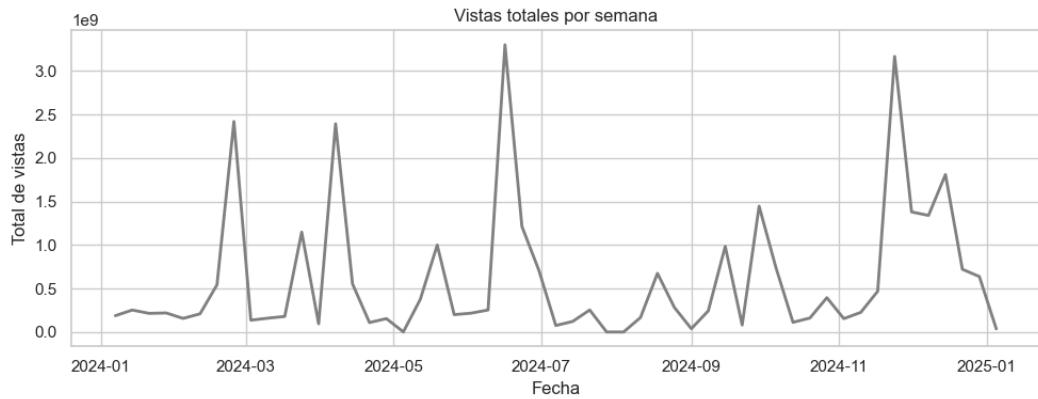


En cuanto a la distribución temporal de los comentarios, justamente podemos observar una incidencia alta en noviembre-diciembre de 2024 marcando la relevancia del canal prorruso mencionado en ese periodo en particular. Veremos más adelante a que evento podría corresponder este hecho. También se menciona que en ningún momento se eliminaron los comentarios realizados en 2025 ya que son relativos a los videos publicados en 2024 (nuestro recorte temporal aceptado), sin embargo, sí se limitan por default en las visualizaciones del Panel de Control en PowerBI.

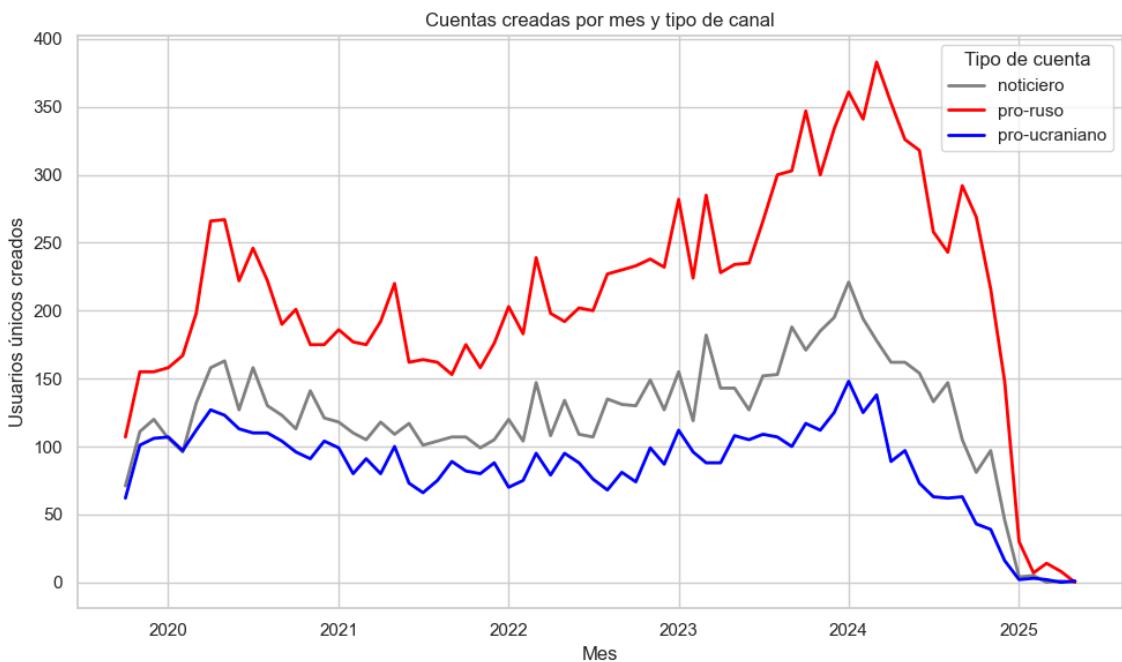
4.3 Visualizaciones de series históricas



En cuanto a las vistas totales, se observa que hay pequeñas diferencias ya que los picos de marzo, abril y Julio están subrepresentados en cuanto a los comentarios.



Adentrándonos en el análisis exploratorio de los usuarios, podemos observar que la distribución de su antigüedad al momento del recorte muestra pequeñas diferencias. Podemos ver que realmente no hay diferencias sustanciales entre los comentarios segmentados por el tipo de canal. Las variaciones son similares con picos en 2020 y el mismo año 2024.



Posteriormente hemos realizado algunas exploraciones adicionales para representar los rankings por usuarios y canales que nos ayudaron a desarrollar nuestro modelo para identificar a los bots y el núcleo duro más adelante. También, se refuerza claramente la superioridad que hemos observado en cuanto a los comentarios por parte del youtuber Miguel Ruiz Calvo ocupando gran parte del top 10 de los videos por cantidad de comentarios:

| video_title | channel_title | cantidad_comentarios | |
|-------------|---|----------------------|------|
| 0 | RUSIA ACORRALA A LA OTAN EN UCRANIA: ÚLTIMA HORA | Miguel Ruiz Calvo | 2746 |
| 1 | ÚLTIMA HORA: RUSIA ATACÓ EL AERÓDROMO FAVORITO... | Miguel Ruiz Calvo | 2116 |
| 2 | ÚLTIMA HORA: SIRIA HA CAÍDO ISRAEL EMPIEZA A... | Miguel Ruiz Calvo | 2050 |

| video_title | channel_title | cantidad_comentarios | |
|-------------|---|----------------------|------|
| 3 | ÚLTIMA HORA: ISRAEL ENLOQUECE! CASTIGO MASIVO ... | Miguel Ruiz Calvo | 1874 |
| 4 | <input checked="" type="checkbox"/> ¿Cuántas bajas llevan RUSIA y UCRANIA en la ... | Memorias de Pez | 1730 |
| 5 | ÚLTIMA HORA: RUSIA NO PUEDE HACER NADA: SIRIA ... | Miguel Ruiz Calvo | 1644 |
| 6 | DIRECTO URGENTE: DESTRUCTORES ATACADOS! ORESHN... | Miguel Ruiz Calvo | 1628 |
| 7 | <input checked="" type="checkbox"/> RESUMEN de los DOS AÑOS de guerra entre RUSI... | Memorias de Pez | 1579 |
| 8 | GUERRA UCRANIA OTAN: "Rusia no tiene las tro... | EL PAÍS | 1540 |
| 9 | ÚLTIMA HORA Alemania dice que la OTAN no def... | Negocios TV | 1457 |

Otra parte del análisis exploratorio que hemos realizado en las *Jupyter notebooks*, pero que posteriormente se migraron al Panel de Control de PowerBi son las nubes de palabras. Adjunto una imagen de la muestra general desde la notebook donde claramente se visualizan las palabras representativas del conflicto: países en guerra, líderes, actores geopolíticos relevantes, entre otros.

4.4 Nubes de Palabras



En el caso del Panel de Control de PowerBi se realizaron trabajos de refinamiento para limpiar, a través de los stop-words, los términos irrelevantes y se segmentaron (y se cuantificaron) los términos a nivel de unigramas (1-gramas) y bigramas. Esto se realizó en la notebook 05_master_dataset_enrichment.ipynb que veremos más adelante. Sin embargo, podemos adelantar los resultados:



En los *bigramas* observamos que vuelven a aparecer los actores relevantes, pero también eslógans (viva rusia), agradecimientos y burlas a los youtubers (“gracias miguel” en referencia a Miguel Ruiz Calvo y “memorias otan” en referencia a la posición occidentalista de Memorias de Pez).

También aparece un nuevo término frente al miedo de la escalada del conflicto: guerra mundial.

Por otro lado, en el caso de unigramas (1-gramas) la perspectiva es similar a lo que veíamos en la nube de palabras inicial.



Aunque la gran parte de la analítica descriptiva aparece en el panel del control de PowerBi, quiero mencionar solamente estos dos puntos antes de avanzar. En primer lugar, esta exploración e iteración (no visible en el código) me permitió descubrir algunos patrones y plantearme preguntas que no tenía al inicio del trabajo. Por ejemplo, de sumar el estudio de insultos y quejas hacia el creador (es algo que pude observar al revisar los comentarios del canal Memorias de Pez). También sumé las categorías de palabras reservadas (lenguaje propio del conflicto con lugares, personas y términos que se interpretan de la forma específica en ese contexto) y ejes argumentativos que

pude ir descubriendo al explorar los datos. Pude observar que no es inusual mandar saludos o mencionar el país desde dónde las personas están escribiendo y eso me permitió considerar la categoría para segmentar los países de los usuarios. Claramente España siendo subrepresentada (no es usual mandar saludos desde el mismo país de dónde es el canal/youtuber), si se vieron varios comentarios desde Latinoamérica con mayor representación de Argentina.

4.5 Identificación de insultos - Análisis de polaridad

En este segmento del código nos orientamos a identificar patrones de comportamiento lingüístico y, en particularidad, la detección de la presencia de insultos. Eso es relevante ya que posteriormente forma parte del input para los modelos y también para en análisis de la polaridad.

En primer lugar, se aplicó un filtro léxico sobre el texto de los comentarios con el objetivo de identificar expresiones agresivas o insultantes. Esta operación consistió en la búsqueda de coincidencias en una lista curada de términos ofensivos, considerando variaciones morfológicas, de género y capitalización, e insultos propios del conflicto (ejemplo: “ucranazi”). El proceso generó una nueva variable binaria denominada insulto, cuyo valor indica la presencia o ausencia de insultos explícitos. Los resultados mostraron que el seis por ciento de los comentarios contenían algún tipo de expresión ofensiva (6.817 sobre un total de 113.583), mientras que el resto (94 %) no presentaba agresividad textual. Si bien esta proporción es relativamente baja, el indicador resulta útil para estudiar la relación entre tono discursivo y

orientación ideológica, así como para detectar posibles patrones de comportamiento automatizado o de desinformación.

Posteriormente se exploraron los resultados en el Panel de control por el Canal junto a otras métricas para mapear los espacios “polarizados”:

| Canal | Inclinación Canal | Comentarios | Suscriptores | Índice polarización | % Coment Insulto |
|-------------------|-------------------|-------------|--------------|---------------------|------------------|
| Miguel Ruiz Calvo | pro-ruso | 42.307 | 574000 | 0,53 | 6,7% |
| Memorias de Pez | pro-ucraniano | 31.448 | 2520000 | 0,62 | 4,8% |
| RTVE Noticias | noticiero | 12.994 | 2460000 | 0,50 | 6,4% |
| Negocios TV | pro-ruso | 8.810 | 2110000 | 0,59 | 5,4% |
| La Vanguardia | noticiero | 7.648 | 2170000 | 0,46 | 6,1% |
| EL PAÍS | noticiero | 3.412 | 3030000 | 0,46 | 4,6% |
| Intereconomía | pro-ruso | 3.297 | 329000 | 0,52 | 7,5% |
| El Mundo | noticiero | 1.733 | 1450000 | 0,41 | 10,1% |
| laSexta Noticias | noticiero | 112 | 439000 | 0,65 | 9,8% |
| Total | | 111.761 | 15082000 | 0,55 | 6,0% |

5. Diseño e implementación de los modelos

5.1 Guía de criterios para la clasificación de comentarios

Con el fin de asegurar coherencia y rigor en la interpretación ideológica de los mensajes, se elaboró una guía de criterios para la clasificación de los comentarios de YouTube relacionados con la guerra en Ucrania. Este instrumento metodológico combina el enfoque del análisis del discurso político (van Dijk, 2003; Fairclough, 1995) con principios del análisis computacional de opinión (Liu, 2012), integrando tanto la dimensión semántica como el contexto de publicación de cada comentario.

La clasificación se realizó considerando tres categorías principales: pro-ucraniano, pro-ruso y neutral. Cada categoría se definió en función de los marcos discursivos predominantes, el posicionamiento explícito o implícito del comentario y, en algunos casos, la orientación editorial del canal donde fue publicado.

La categoría pro-ucraniana incluye aquellos comentarios que manifiestan apoyo al gobierno, ejército o ciudadanía ucraniana; expresan simpatía hacia figuras políticas del país (como el presidente Zelensky); condenan las acciones militares rusas o reproducen narrativas asociadas a la Unión Europea, la OTAN o Estados Unidos. En contraste, la categoría pro-rusa agrupa los mensajes que justifican la intervención militar de Rusia, respaldan a su gobierno o liderazgo político, desacreditan los medios occidentales y utilizan categorías discursivas propias del Kremlin como “desnazificación” o “provocación occidental”. Por último, la categoría neutral comprende los comentarios que expresan preocupación humanitaria sin alinearse con ningún bando, formulan críticas equilibradas hacia ambos, o simplemente agradecen el contenido informativo sin emitir juicios valorativos.

La guía también contempló casos especiales que pudieran alterar la interpretación semántica directa. En particular, se establecieron criterios para identificar ironía o sarcasmo: si el comentario irónico busca desacreditar al canal y este posee orientación pro-ucraniana, se clasifica como pro-ruso, mientras que si el sarcasmo refuerza el mensaje del canal, se considera alineado con la orientación de este. Asimismo, se previó una etiqueta adicional (*es_sarcastico*) para futuros análisis discursivos más finos. Los comentarios fuera de contexto —aquellos que no refieren al contenido del video ni al conflicto— fueron marcados como tales y excluidos del análisis ideológico, pues suelen corresponder a intervenciones automatizadas, autopromocionales o de desinformación general.

Por último, los comentarios que elogian directamente al canal (“excelente contenido”, “gran trabajo”) se interpretaron según la orientación del canal, asumiendo adhesión implícita a su línea editorial. De modo análogo, las menciones a hechos bélicos sin juicio explícito se clasificaron en función del tono discursivo: la admiración o validación de acciones militares rusas se codificó como pro-rusa, mientras que la neutralidad o la crítica implicaron clasificación neutral.

El dataset utilizado para el etiquetado manual incluyó las siguientes columnas: `label_comentario` (categoría ideológica principal), `es_sarcastico` (presencia de ironía o sarcasmo) y `fueras_de_contexto` (marcador de pertinencia temática). Esta estructura permitió registrar información contextual adicional sin alterar la variable de clasificación principal y facilitó la validación cruzada entre codificadores humanos y procedimientos automáticos.

Este protocolo de codificación constituyó el punto de partida para el entrenamiento del modelo automático de clasificación, asegurando la consistencia semántica y metodológica entre la etapa manual y la fase computacional del proceso.

5.2 Modelo base de clasificación automática

A partir del conjunto etiquetado manualmente se entrenó un modelo base de clasificación con el objetivo de extender la cobertura del corpus y establecer un punto de referencia sobre el comportamiento lingüístico de los comentarios. Dado que la tarea consiste en asignar cada comentario a una de tres categorías mutuamente excluyentes (pro-ruso, pro-ucraniano o neutro), se optó por un enfoque supervisado clásico de aprendizaje automático de texto, basado en la combinación de un vectorizador TF-IDF (*Term Frequency-Inverse Document Frequency*) con un clasificador *Logistic Regression multinomial*.

Esta técnica fue seleccionada por su solidez empírica en escenarios con volumen moderado de datos, características textuales dispersas y clases desbalanceadas, condiciones habituales en estudios de análisis político y de opinión (Manning, Raghavan & Schütze, 2008). El modelo pondera la relevancia de los términos dentro del corpus en función de su frecuencia relativa, ofreciendo además interpretabilidad al identificar las palabras con mayor peso predictivo en cada orientación discursiva. Su eficiencia computacional y su bajo riesgo de sobreajuste lo convierten en una

alternativa adecuada para esta fase exploratoria, previa a la implementación de modelos más complejos.

El preprocessamiento textual incluyó la normalización del texto, la eliminación de stopwords y la tokenización en minúsculas, generando un espacio vectorial de alta dimensión a partir de un vocabulario controlado. El modelo de regresión logística se entrenó sobre el conjunto etiquetado manualmente ($n = 1.877$), utilizando validación cruzada estratificada para estimar la variabilidad del rendimiento entre clases. Para cada comentario, el algoritmo calculó la probabilidad de pertenecer a cada categoría, asignando la clase final cuando la probabilidad superaba un umbral de confianza definido experimentalmente.

La evaluación del modelo se realizó sobre un subconjunto independiente de 376 comentarios con etiqueta humana, alcanzando una precisión global del 64 % y un F1 macro promedio de 0,54. El desempeño fue superior en la categoría pro-rusa ($F1 = 0,76$), mientras que las clases neutra y pro-ucraniana presentaron resultados más bajos ($F1 \approx 0,42$). Estas diferencias se explican tanto por el desbalance de clases como por la ambigüedad semántica de algunos comentarios, una característica frecuente en contextos de polarización digital (Calvo & Araguete, 2020).

A pesar de sus limitaciones, el modelo cumplió adecuadamente su propósito inicial: establecer una base cuantitativa para la etiquetación semiautomática y la expansión controlada del corpus, manteniendo interpretabilidad y trazabilidad sobre las decisiones del algoritmo. Además, permitió validar la correspondencia entre los patrones lingüísticos detectados y las narrativas predominantes observadas en los canales analizados, ofreciendo así un punto de partida robusto para la etapa de clasificación híbrida y el análisis posterior de polarización discursiva.

5.3 Implementación del modelo basado en redes neuronales

Con el propósito de mejorar la capacidad predictiva y superar las limitaciones del

modelo base lineal, se implementó una arquitectura neuronal basada en transformadores ligeros. El modelo seleccionado fue DistilBERT, una versión reducida y optimizada del modelo Bidirectional Encoder Representations from Transformers (BERT) propuesto por Devlin et al. (2018). Este modelo fue elegido por su equilibrio entre rendimiento y eficiencia computacional, dado que mantiene más del 95 % del rendimiento de BERT original con un 40 % menos de parámetros y un tiempo de inferencia significativamente inferior (Sanh et al., 2019).

La elección de esta arquitectura se fundamenta en su capacidad para capturar relaciones semánticas bidireccionales en el texto, un aspecto esencial para el análisis del discurso político en entornos de polarización, donde el significado de una frase depende con frecuencia del contexto completo. A diferencia de los enfoques clásicos de bag-of-words o TF-IDF, los modelos basados en transformadores generan representaciones contextuales de las palabras, permitiendo detectar matices, ironías o contradicciones implícitas, que resultan determinantes para distinguir entre mensajes pro-rusos, pro-ucranianos y neutros.

Diseño e implementación

El modelo fue implementado en TensorFlow 2.x y Keras, utilizando la variante distilbert-base-uncased con pesos preentrenados sobre grandes corpus en inglés. Dado que el dataset del proyecto contenía comentarios en español e inglés, se aplicó un proceso de pretokenización mediante el tokenizer de Hugging Face compatible con DistilBERT. Las secuencias fueron truncadas y normalizadas a una longitud fija de 128 tokens para optimizar el rendimiento durante el entrenamiento.

El entrenamiento se realizó íntegramente en entorno local, aunque se dejó en el repositorio una versión que se podría reajustarse para adaptarla para la ejecución en Azure Machine Learning, con soporte de GPU y escalamiento en clúster, para futuras iteraciones del modelo. Esta decisión respondió a la necesidad de garantizar

reproducibilidad sin depender de recursos externos, priorizando la estabilidad del entorno local frente a la disponibilidad del servicio en la nube.

El modelo fue entrenado sobre la muestra ampliada del corpus híbrido, utilizando tres épocas (epochs) y un tamaño de lote ajustado experimentalmente. La duración total del entrenamiento fue de aproximadamente 25 minutos por época, con una tasa de aprendizaje baja ($2e-5$) y optimizador AdamW, lo que permitió una convergencia estable evitando el sobreajuste. Se optó por un número reducido de épocas debido al tamaño relativamente pequeño del conjunto de entrenamiento (≈ 6.000 registros) y para prevenir una sobreespecialización en ejemplos con alta frecuencia léxica.

Entrenamiento y evaluación

Durante el entrenamiento, se observó una evolución progresiva del rendimiento tanto en el conjunto de entrenamiento como en el de validación. La precisión (accuracy) aumentó de 0,74 en la primera época a 0,92 en la tercera, mientras que la pérdida (loss) se redujo de 0,59 a 0,22. En validación, la precisión final alcanzó 0,827 con una pérdida de 0,451, indicando una buena capacidad de generalización para un modelo de estas dimensiones. La Figura 5.1 muestra la evolución del accuracy y la pérdida a lo largo de las tres épocas, donde se aprecia una mejora sostenida sin indicios de sobreajuste prematuro.

A partir del modelo entrenado, se generó la versión final *tf_distilbert_stance_v1*, cuyas métricas de evaluación sobre el conjunto de validación fueron las siguientes:

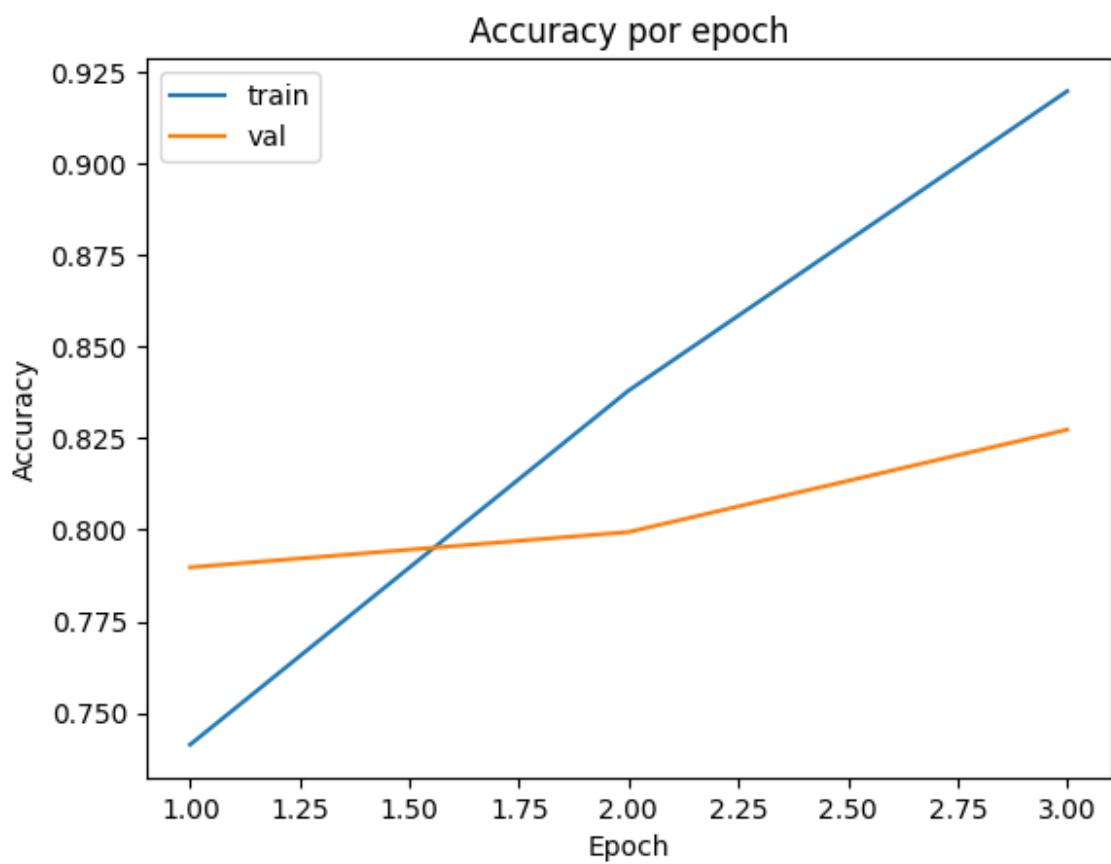
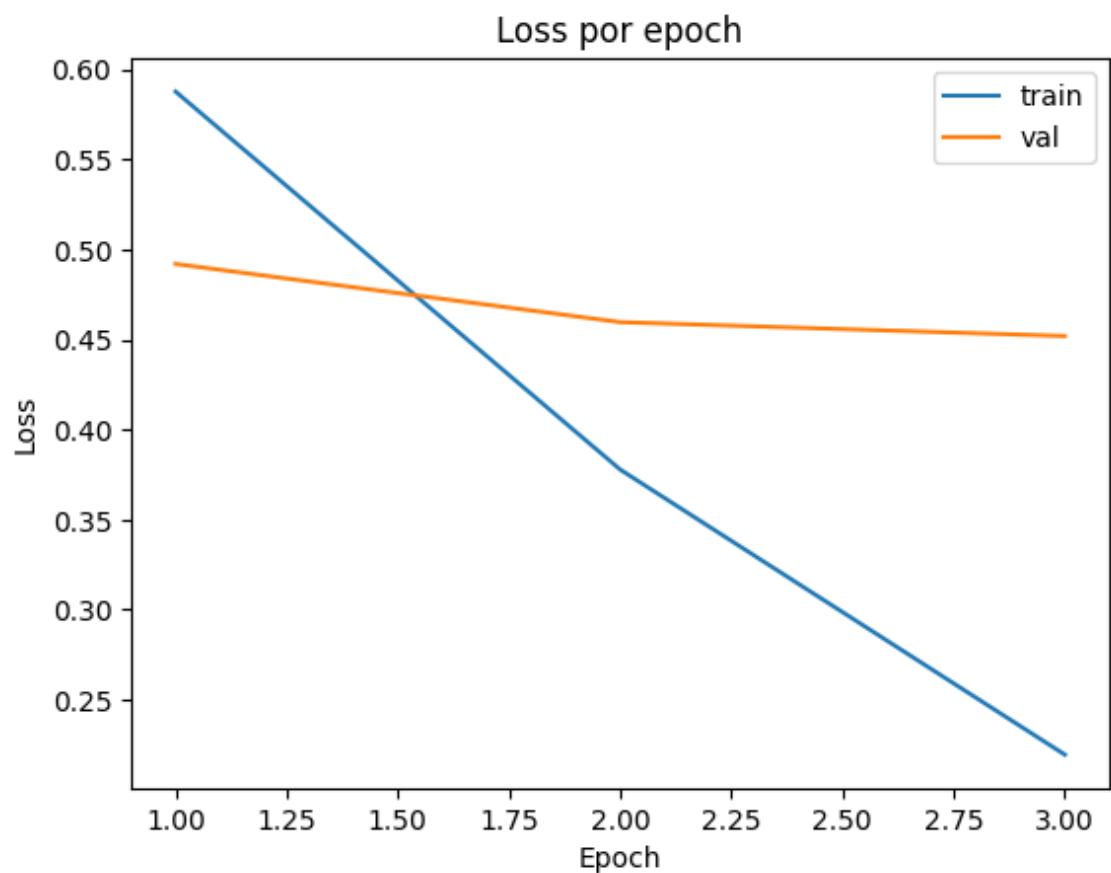
| Métrica global | Valor |
|----------------|-------|
| Accuracy | 0.813 |
| F1 macro | 0.703 |
| F1 ponderado | 0.805 |

| | precision | recall | f1-score | support |
|--------------------------------------|-----------|--------|----------|--------------------|
| ruso | 0.849 | 0.925 | 0.886 | 585.1999999999996 |
| ucraniano | 0.744 | 0.542 | 0.627 | 135.00000000000014 |
| neutro | 0.645 | 0.556 | 0.597 | 112.50000000000013 |
| accuracy | | | 0.813 | 832.6999999999998 |
| macro avg | 0.746 | 0.675 | 0.703 | 832.6999999999998 |
| weighted avg | 0.805 | 0.813 | 0.805 | 832.6999999999998 |
| F1 macro: 0.703 F1 weighted: 0.805 | | | | |

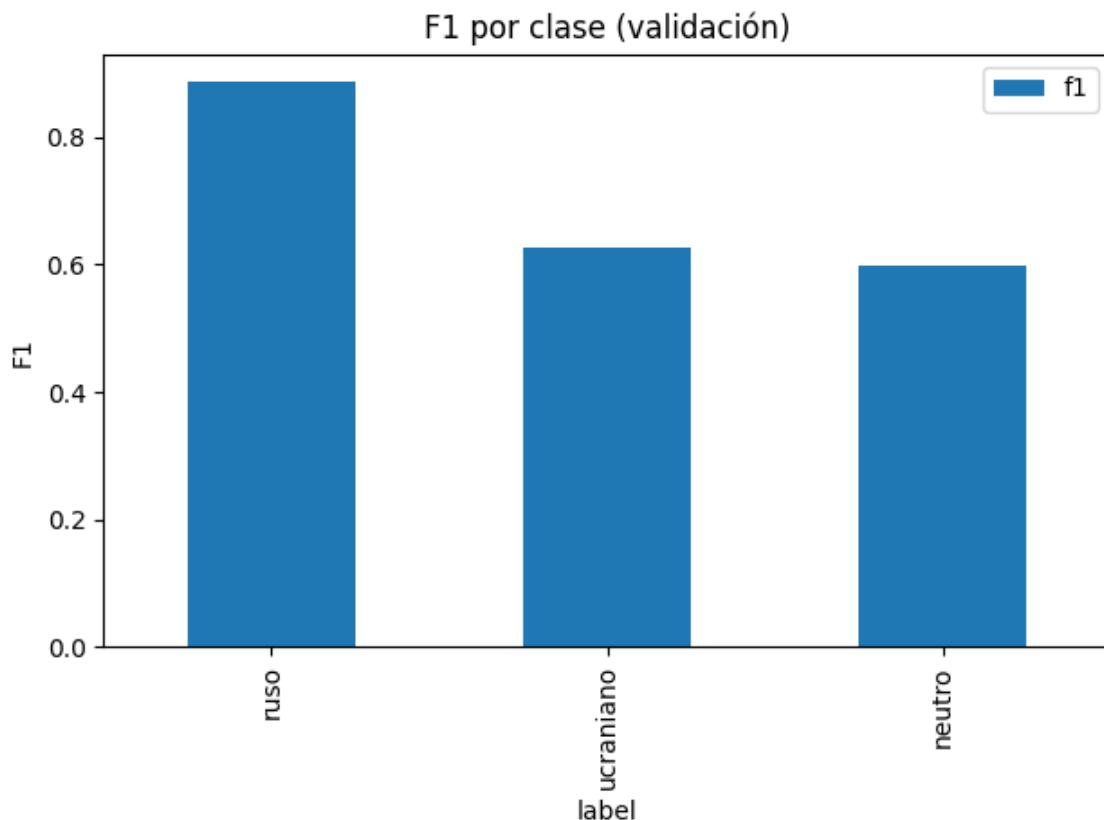
La evaluación por clase evidenció un rendimiento superior en la categoría pro-rusa (F1 = 0.886), seguida por pro-ucraniana (F1 = 0.627) y neutral (F1 = 0.597). Estas diferencias reflejan, por un lado, la asimetría en la distribución de los datos, y por otro, la mayor consistencia discursiva de los mensajes prorrusos, que suelen contener expresiones recurrentes y menos ambiguas. En cambio, los comentarios neutrales y pro-ucranianos presentan una mayor diversidad léxica y emocional, lo que reduce la capacidad de generalización del modelo en estas clases minoritarias.

Interpretación de los resultados

El modelo DistilBERT alcanzó un rendimiento general significativamente superior al de la regresión logística empleada como línea base, incrementando el F1 macro de 0,5 a +0,7 y mejorando el equilibrio entre precisión y exhaustividad en todas las categorías. Este resultado confirma la ventaja de las representaciones contextuales en tareas de clasificación ideológica, donde las señales lingüísticas son sutiles y dependientes del contexto.



Los valores de validación muestran una leve brecha entre entrenamiento y validación, pero sin divergencia pronunciada, lo que sugiere que el modelo generaliza adecuadamente y no incurre en sobreajuste. El descenso progresivo de la función de pérdida y el aumento constante del accuracy a lo largo de las épocas corroboran una convergencia estable.



El análisis por clase también aporta evidencia interpretativa: la categoría pro-rusa exhibe mayor consistencia en los patrones semánticos, lo que explica su mayor F1. En cambio, los comentarios neutros y pro-ucranianos presentan estructuras discursivas más heterogéneas y, en muchos casos, un tono mixto (crítico y empático a la vez), lo que limita la capacidad discriminativa incluso para modelos contextuales.

Por lo tanto, la arquitectura utilizada demostró un equilibrio adecuado entre precisión, capacidad interpretativa y eficiencia computacional, consolidándose como el componente central del pipeline analítico del proyecto. La preservación de la versión entrenada en entorno local y su posible replicación en Azure Machine

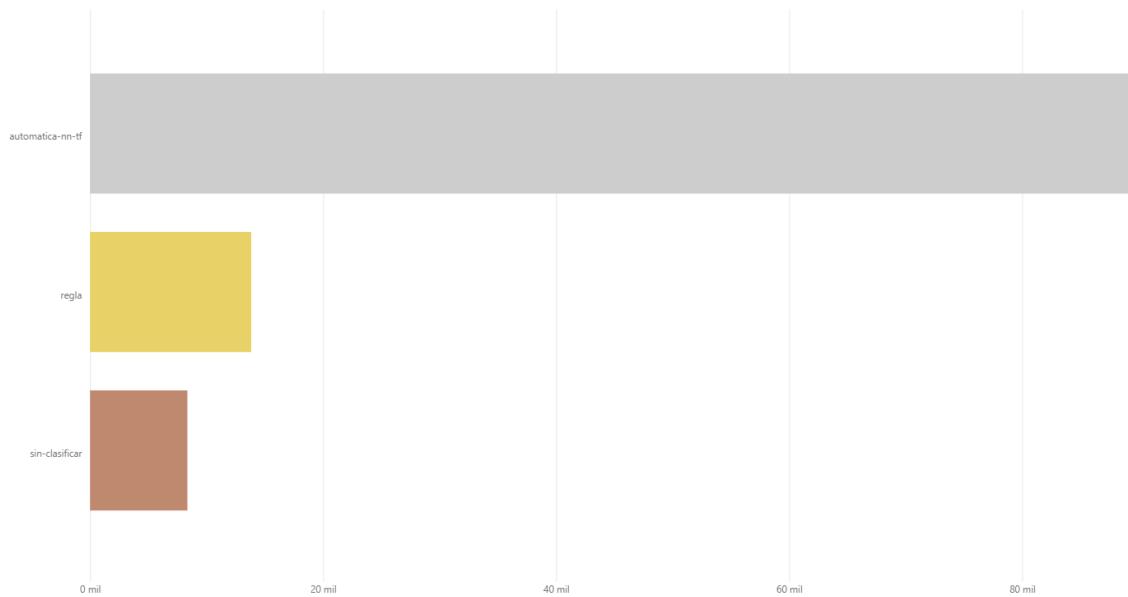
Learning (con ajustes del código) permiten la continuidad y escalabilidad futura del experimento, especialmente ante volúmenes mayores de datos o la incorporación de textos multilingües.

5.4 Integración híbrida de la clasificación

Tras el entrenamiento del modelo neuronal basado en DistilBERT, se procedió a su integración dentro del flujo de clasificación híbrido, cuyo objetivo fue ampliar la cobertura de etiquetas en el corpus y obtener una representación más completa de las posturas discursivas. Esta etapa se desarrolló principalmente en la notebook *04_final_classification_core.ipynb*, donde se combinaron los resultados del modelo automático con las etiquetas humanas y las reglas lingüísticas definidas en fases anteriores.

El esquema de integración siguió una jerarquía de decisión que priorizó la etiqueta humana cuando estaba disponible, seguida por la predicción del modelo neuronal en los casos cuya probabilidad superaba un umbral de confianza del 0,75. Los comentarios restantes fueron clasificados mediante reglas lingüísticas y expresiones regulares que detectan consignas ideológicas, referencias a líderes políticos y uso de insultos. De esta manera, se alcanzó una cobertura cercana al 80 % del corpus, manteniendo control sobre la fiabilidad de las etiquetas y preservando la trazabilidad de su origen.

El resultado de este proceso se consolidó en la variable *label_final*, que unifica la orientación ideológica asignada a cada comentario (pro-rusa, pro-ucraniana o neutral) junto con la fuente de clasificación (humano, automatica-ml, regla).



Esta integración permitió equilibrar la precisión de los modelos con la consistencia semántica del análisis cualitativo, obteniendo así un conjunto de datos robusto para el estudio del discurso digital sobre la guerra de Ucrania.

5.5 Enriquecimiento del dataset final (bots, argumentos, país...)

Posteriormente, en la notebook *05_master_dataset_enrichment.ipynb*, se desarrolló la fase de enriquecimiento del dataset. En esta etapa se incorporaron variables contextuales adicionales, tales como el idioma del comentario, el posible comportamiento automatizado del usuario (detección de bots o cuentas coordinadas), y metadatos relativos a los eventos políticos o militares asociados. En particular me voy a detener en la etapa de la identificación de los bots. El objetivo de esta etapa fue detectar patrones de comportamiento atípicos entre los usuarios que participaban en los debates sobre la guerra de Ucrania, a fin de evaluar la posible incidencia de actividad automatizada o propagandística en la muestra. El método implementado combinó un enfoque heurístico con un modelo de detección de anomalías. En primer lugar, se calcularon métricas de comportamiento por

usuario, incluyendo la frecuencia diaria de publicación, el número de comentarios totales, la cantidad de canales diferentes en los que participaba, la repetición de mensajes idénticos (*duplicated ratio*), el uso de menciones o enlaces externos y la proporción de insultos. Estas variables fueron utilizadas para construir un índice compuesto de probabilidad de automatismo denominado *bot_score*. Dicho índice pondera los indicadores más representativos (por ejemplo, alta repetición de mensajes, escasa antigüedad de la cuenta o intervalos extremadamente cortos entre comentarios), y se normaliza entre 0 y 1. Los usuarios con un valor superior al percentil 98 del conjunto fueron marcados como sospechosos de automatización.

De forma complementaria, se aplicó un modelo *Isolation Forest* (Liu, Ting & Zhou, 2008), un algoritmo no supervisado de detección de anomalías basado en el principio de aislamiento recursivo. Este modelo analizó múltiples variables normalizadas como la frecuencia, dispersión horaria, número de canales y entropía de participación. Posteriormente etiquetó como anómalos aquellos casos con patrones de comportamiento significativamente distintos al resto de la población. La combinación de ambas aproximaciones (heurística y algorítmica) permitió minimizar falsos positivos y obtener una detección más precisa de cuentas automatizadas o coordinadas.

Además de la detección binaria (*bot_flag*), se desarrolló una segmentación de usuarios que clasifica las cuentas en cuatro grupos:

- sospecha_bot, correspondiente a los usuarios con patrones automatizados
- núcleo_duro, caracterizado por alta actividad y participación en múltiples canales
- fiel, usuarios concentrados en un único canal, pero con frecuencia sostenida
- esporádico, aquellos que realizaron un único comentario.

Esta clasificación se basó en cuantiles dinámicos de las variables de comportamiento para mejorar la precisión y distribución equitativa.

La detección de bots se complementó con el análisis de palabras clave y ejes argumentativos utilizados por los distintos segmentos. Este cruce permitió observar que los usuarios sospechosos de automatización tendían a emplear un vocabulario repetitivo, centrado en consignas ideológicas o militares, mientras que los segmentos humanos más activos presentaban mayor diversidad léxica y referencias contextuales a los eventos noticiosos. Este hallazgo refuerza la hipótesis de que parte de la conversación digital sobre la guerra está mediada por mecanismos de amplificación automatizada de mensajes.

Esta versión final, almacenada en el archivo *8_final_master_enriched.csv*, constituye la capa analítica definitiva del proyecto, desde la cual se realizaron los análisis descriptivos, comparativos y las visualizaciones en PowerBI.

6. Análisis de los resultados obtenidos

En base a los resultados expuestos en el Dashboard se puede observar de forma centralizada los patrones de interacción, orientación ideológica y dinámica temporal de los comentarios vinculados a la guerra de Ucrania en YouTube durante 2024. El dataset final incluyó 111.761 comentarios (filtro predeterminado para comentarios en 2024 para la simplicidad), provenientes de 399 videos distribuidos en nueve canales con más de 15 millones de suscriptores y 43 millones de visualizaciones en conjunto.

En términos de orientación discursiva, el modelo híbrido que combinó la clasificación manual, neuronal y reglas lingüísticas pudo identificar una predominancia de comentarios pro-rusos (67%), frente a un 12% pro-ucranianos. Por otro lado, otro 12% de comentarios quedaron identificados como neutrales y el resto sin clasificación (dados el nivel de confianza solicitado para clasificar los comentarios). Esta asimetría

se mantiene de forma consistente entre los distintos canales analizados, con ligeras variaciones según su línea editorial. Por ejemplo, los canales Memorias de Pez (en parte por los comentarios negativos frente a los informes del canal) y Negocios TV concentraron la mayor proporción de mensajes favorables a Rusia, mientras que El Mundo, La Vanguardia y RTVE Noticias presentaron un equilibrio mayor entre las posturas. Claramente el mayor aporte de comentarios pro-rusos los hace el canal de Miguel Ruiz Calvo que sorprende con su engagement y productividad (en el año tenía más de 800 videos relacionados con la guerra).

Ranking x Engagement (Comentarios video + Comentarios 48 hs)

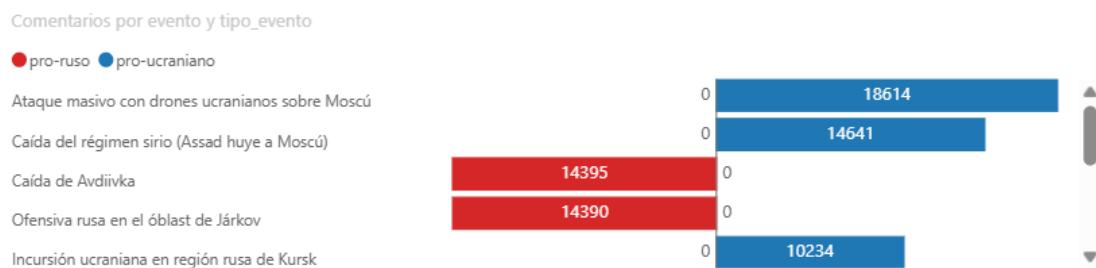
- | | |
|---|---|
|  1 | Miguel Ruiz Calvo · C/Video 846 · 48h 90% |
|  2 | Memorias de Pez · C/Video 629 · 48h 77% |
|  3 | RTVE Noticias · C/Video 260 · 48h 81% |

En cuanto a la polarización el índice medio es de 0,55, lo que indica un nivel moderado-alto de segmentación ideológica. Los canales más polarizados fueron aquellos con un discurso militante explícito, mientras que los medios tradicionales (como El País o La Vanguardia) presentaron una mayor diversidad de posturas entre sus audiencias. En estos últimos, la coexistencia de usuarios de distintas orientaciones discursivas refuerza su papel como espacios de debate público, en contraste con las “cámaras de eco” observadas en canales de opinión política o contenidos militantes.

La detección de comportamientos automatizados permitió identificar 1.131 cuentas sospechosas (alrededor del 3%), responsables de más de 16.000 comentarios (14,5%). La mayoría de estas cuentas mostraban afinidad con la narrativa pro-rusa y una elevada

repetición de consignas o eslóganes (“viva Rusia”, “culpa de la OTAN”, “mentiras de Occidente”). Sin embargo, hay que mencionar que proporcionalmente la cantidad de comentarios salientes de esas cuentas era levemente mayor en los comentarios que se identificaron como pro-ucranianos. En términos de segmentación de usuarios, predominó el grupo esporádico con un solo comentario (67,8%), seguido por el activo (20,2%), el fiel (8%) y una fracción menor del núcleo duro (2%).

El análisis léxico y semántico complementó estos resultados. Entre las palabras más frecuentes destacan rusia, ucrania, otan, guerra y estados unidos, reflejando la centralidad del eje geopolítico y militar. En la dimensión emocional, los ejes argumentativos predominantes fueron “Agradecimientos al creador”, “Festejos y memes”, e “Insultos al creador”, lo que sugiere una alta carga afectiva y un componente performativo característico del discurso en redes.



Finalmente, el análisis temporal evidenció picos de actividad coincidentes con eventos simbólicos clave del lado de los comentarios pro-rusos (como la caída del dictador sirio Assad o el ataque con drones sobre Moscú). En el canal pro-ucraniano se desactivó el evento de la Caída de Avdiivka. Esto sugiere que no sólo los momentos de avance militar ruso actúan como catalizadores discursivos, sino también los momentos simbólicos y geopolíticos amplifican la participación de cuentas afines o automatizadas.

En conjunto, los resultados apuntan a una configuración polarizada, emocionalmente intensa y parcialmente automatizada del debate digital sobre la guerra de Ucrania. Si

bien la mayoría de los usuarios parecen responder de forma espontánea, la evidencia de actividad coordinada y la homogeneidad de ciertos mensajes refuerzan la hipótesis de una estrategia comunicacional orientada a influir en la narrativa pública internacional.

7. Conclusiones y planes de mejora

7.1 Conclusiones generales

El presente trabajo tuvo como propósito analizar los comentarios de YouTube relacionados con la guerra entre Rusia y Ucrania durante 2024, aplicando técnicas de análisis de lenguaje natural (NLP) y modelos neuronales de clasificación. Los resultados permiten afirmar que la plataforma constituye un espacio de alto contenido político y emocional, donde los discursos espontáneos, militantes y automatizados conviven e interactúan de forma dinámica.

La aplicación del modelo híbrido —que combinó clasificación manual, reglas lingüísticas y redes neuronales basadas en **DistilBERT**— permitió identificar con solidez tres ejes de posicionamiento: pro-ruso, pro-ucraniano y neutral. De acuerdo con el análisis realizado, el 67 % de los comentarios presentaron afinidad con narrativas prorrusas, un 12 % con posturas pro-ucranianas y otro 12 % se mantuvo en un tono neutral. Esta distribución refuerza la hipótesis de una **asimetría discursiva** en la esfera digital, donde el relato pro-ruso exhibe mayor capacidad de amplificación y permanencia en el tiempo.

En línea con los objetivos iniciales, se observó que los canales con una orientación ideológica definida (como *Miguel Ruiz Calvo* o *Negocios TV*) son los principales generadores de discurso polarizado, mientras que los medios tradicionales (*RTVE Noticias*, *El País*, *La Vanguardia*) operan como espacios de mayor diversidad

discursiva y debate público. En conjunto, el índice medio de polarización (0,55) refleja un entorno de interacción política con alto grado de carga emocional. El trabajo también permitió identificar comportamientos automatizados, asociados principalmente a la narrativa pro-rusa (cuantitativamente), aunque con presencia proporcional equitativa en el discurso pro-ucraniano. Un total de 1.131 cuentas (3 % del total) fueron catalogadas como sospechosas, responsables de aproximadamente 14,5 % de los comentarios del corpus. Este hallazgo sugiere la existencia de estrategias de amplificación discursiva coordinadas, orientadas a reforzar narrativas dominantes o deslegitimar posturas opuestas.

En cuanto al análisis semántico, se constató que las palabras más frecuentes giran en torno a *rusia, ucrania, otan, guerra y estados unidos*, lo que confirma la **centralidad geopolítica** del conflicto en el debate digital. A su vez, los ejes argumentativos más recurrentes –“Agradecimientos al creador”, “Festejos y memes” e “Insultos al creador”– muestran la **dimensión afectiva** del intercambio discursivo, donde el humor, la ironía o la indignación funcionan como mecanismos de pertenencia identitaria y refuerzo de grupo.

Por último, el análisis temporal evidenció que los picos de interacción coinciden con **eventos simbólicos y militares relevantes**, como la caída de Avdiivka, los ataques con drones sobre Moscú o la huida de Bashar al-Ásad a Rusia. Estos momentos no solo impulsan la participación espontánea, sino que también activan redes coordinadas de cuentas afines. En suma, el trabajo permitió responder de forma positiva a las preguntas planteadas: se identificaron los actores discursivos predominantes, los eventos de mayor resonancia mediática y los niveles de radicalización y automatización presentes en el debate online.

7.2 Posibles mejoras y consideraciones

Como línea de mejora, se propone ampliar el alcance temporal y temático

del estudio, incorporando la totalidad de los videos publicados durante el periodo 2022–2025 y extendiendo el análisis a otras plataformas sociales como X (Twitter), Telegram y TikTok. Esto permitiría comparar la circulación de narrativas y evaluar el grado de coherencia entre ecosistemas mediáticos distintos, aportando evidencia más robusta a nivel comparativo.

Por otro lado, actualmente los datos se gestionan en formato CSV y Excel en un entorno local, lo cual resulta adecuado para la etapa exploratoria pero limitado para volúmenes crecientes de información. Se recomienda migrar el flujo de trabajo a una arquitectura de datos más robusta (por ejemplo, DuckDB, BigQuery o PostgreSQL) que permita optimizar la lectura, consulta y actualización de los datasets, además de facilitar la integración con herramientas de visualización y versionamiento.

Finalmente, se plantea como línea de trabajo futura la comparación internacional de las narrativas sobre la guerra en Ucrania, extendiendo el análisis a otros países europeos y latinoamericanos. Este enfoque comparativo permitirá evaluar la influencia de los medios nacionales, las redes sociales y los actores políticos en la construcción del discurso público, contribuyendo al estudio de la geopolítica digital y la comunicación de conflictos contemporáneos.

8. Bibliografía

- Agencia EFE. (2025, 17 de febrero). ONU hace balance de 3 años de crímenes en Ucrania: 12.600 civiles muertos, 29.300 heridos. SWI swissinfo.ch. Recuperado de <https://www.swissinfo.ch>
- BBC News. (2024, 16 de febrero). Alexei Navalny: Russian opposition leader dies in Arctic prison. <https://www.bbc.com/news>
- Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulación, desinformación y radicalización en la política estadounidense. Oxford University Press.
- Bruns, A., & Highfield, T. (2018). Public Sphere 2.0: Rethinking Political Communication in the Digital Age. Routledge.
- Calvo, E. (2015). Anatomía política de Twitter en Argentina: Tuiteando la brecha, fake news y nuevas redes de poder. Teseo Press.
- Calvo, E., & Aruguete, N. (2020). Fake news, trolls y otros encantos: Cómo funcionan (para bien y para mal) las redes sociales. Siglo XXI Editores.
- Cardón, D. (2010). La democracia internet: Promesas y límites. Seuil.
- Castells, M. (2012). Redes de indignación y esperanza: Los movimientos sociales en la era de Internet. Alianza Editorial.
- Council on Foreign Relations. (2025). Ukraine: Conflict tracker.
<https://www.cfr.org/global-conflict-tracker/conflict/conflict-in-ukraine>
- Databricks. (2022). The Medallion Architecture: A modern approach to building Lakehouse systems. Databricks Technical Paper. <https://www.databricks.com>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Diez-Gracia, A. (2024). Agenda y demanda informativa sobre la guerra de Ucrania en la prensa internacional. Revista de Comunicación, 23(2), 1-21.

<https://revistadecomunicacion.com>

Fairclough, N. (1995). *Media Discourse*. Edward Arnold.

Google Developers. (2024). YouTube Data API v3.

<https://developers.google.com/youtube/v3>

Google Developers Policy. (2024). API Usage Limits.

<https://developers.google.com/youtube/terms>

Gramsci, A. (1971). *Antología*. Siglo XXI Editores.

Hall, S. (1980). Estudios culturales: Dos paradigmas. *Media, Culture & Society*, 2(1), 57-72.

Hasan, M. (2024). Russia-Ukraine Propaganda on Social Media: A Bibliometric Analysis. *Journalism and Media*, 5(3), 980-992.

<https://doi.org/10.3390/journalmedia5030062>

House of Commons Library. (2024). Russia's war on Ukraine: Developments in 2024. UK Parliament Research Briefing CBP-9847.

<https://commonslibrary.parliament.uk/research-briefings/cbp-9847>

Human Rights Watch. (2022). Guerra entre Rusia y Ucrania (nota descriptiva, 24 de febrero de 2022). <https://www.hrw.org>

Institute for the Study of War. (2024). Russia-Ukraine War Campaign Assessment (annual summary). <https://www.understandingwar.org>

Iyengar, S., Sood, G., & Lelkes, Y. (2019). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 79(S1), 221-245.

Kahn, G. (2023, 31 de marzo). Bloqueada en Occidente, la propaganda rusa prospera en español en TV y redes sociales. Radio Televisión Martí / Reuters Institute.

<https://www.martinoticias.com>

Kirby, P. (2025, 15 de mayo). Why did Putin's Russia invade Ukraine? BBC News.

<https://www.bbc.co.uk>

Le Grand Continent. (2023, 30 de diciembre). Las tendencias de búsqueda en Google

en 2023. <https://www.legrandcontinent.eu>

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *Proceedings of the*

2008 IEEE International Conference on Data Mining (ICDM) (pp. 413-422). IEEE.

<https://doi.org/10.1109/ICDM.2008.17>

Marinero, I. (2022, 7 de diciembre). Lo más buscado en Google en 2022 en España: de Ucrania y el Wordle a Tamara Falcó y Will Smith. *El Español - Omicrono*.

<https://www.elespanol.com>

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176-187.

McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.

O'Leary, P. (2023). *Modern data architecture: From data lakes to lakehouses*.

O'Reilly Media.

Pariser, E. (2011). *El filtro burbuja: Cómo la nueva web personalizada está cambiando lo que leemos y cómo pensamos*. Penguin Press.

Raasveldt, M., & Mühleisen, H. (2022). *DuckDB: An Embeddable Analytical Database*. VLDB Endowment.

Reuters. (2024a, 21 de enero). Shelling kills civilians in Donetsk market.

<https://www.reuters.com>

Reuters. (2024b, 8 de diciembre). Syrian President Bashar al-Assad flees to Moscow as regime collapses. <https://www.reuters.com>

Reuters (Guy Faulconbridge). (2024, 17 de febrero). Russia says its forces move forward after Ukraine withdraws from Avdiivka. Reuters. <https://www.reuters.com>

Reuters (Max Hunder & Guy Faulconbridge). (2023, 31 de marzo). Russia vs Ukraine: The biggest war of the fake news era. *Reuters / Forbes México*.

<https://www.forbes.com.mx>

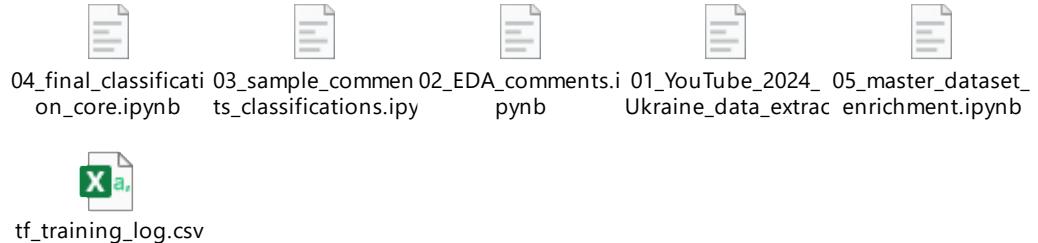
Sampedro, V. (2016). *El cuarto poder en red: Por un periodismo (de código) libre*.

Icaria Editorial.

- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
- Sarsfield, R., & Abuchanab, Z. (2024). Populist storytelling and negative affective polarization: Social media evidence from Mexico. Cambridge University Press.
- Statista. (2024). Global Google search interest in “Ukraine war” (2022-2024). Statista Research. <https://www.statista.com>
- Stoltenberg, J. (2023, 18 de febrero). Speech by NATO Secretary General at the Munich Security Conference. NATO.int.
- The Economist. (2023, 20 de mayo). How Ukraine is fighting the propaganda war. The Economist.
- The New York Times. (2023, 22 de agosto). Troop deaths and injuries in Ukraine war near 500,000, U.S. officials say. NYTimes.
- U.S. Congressional Research Service. (2024). U.S. security assistance to Ukraine: Background, recent developments, and issues for Congress. CRS Report R47652. <https://crsreports.congress.gov>
- van Dijk, T. A. (2003). Critical Discourse Analysis. Routledge.
- Verón, E. (1984). Construir el acontecimiento: Los medios frente a lo inédito. Gedisa.
- Wilches-Tinjacá, F., Guerrero-Sierra, H. F., & Niño, C. (2024). Emociones políticas y narrativas prototípicas: TikTok en las campañas políticas: Estudio de caso. Revista Latina de Comunicación Social, 82, 1-29. <https://doi.org/10.4185/rcls-2024-2234>
- Zelenski, V. (2022, 16 de marzo). Discurso ante el Congreso de los EE. UU. El País.

9. Anexo con el código fuente desarrollado.

Notebooks & Logs:



Repositorio público:

https://github.com/vladyslavbdhm/analisis_guerra_ucrania_youtube

Dashboard in OneDrive y PDF:

