



Universidad Pontificia de Comillas

# **IMPACTO DE LA GUERRA DE UCRANIA EN LAS NOTICIAS FINANCIERAS: UN ANÁLISIS DE *TEXT MINING***

Alumno: Lucía Ruiz Gavarrón

Director: Alejandro Rodríguez Gallego

## RESUMEN

En este estudio se investiga el impacto de la Guerra de Ucrania en las noticias financieras mediante el uso de técnicas de *text mining*. El objetivo consiste en analizar los cambios temáticos y el sentimiento en una muestra de noticias financieras. Para ello, se ha realizado un proceso de estructuración del *corpus*, seguido de un análisis de muestra y preprocesamiento de texto para garantizar la calidad y coherencia de los datos. A continuación, se han aplicado técnicas de *topic analysis* y *sentiment analysis* para identificar los términos clave que se han visto impactados por la guerra y evaluar la actitud asociada a las noticias. Los resultados demuestran cambios significativos en los temas tratados y en el sentimiento expresado en las noticias financieras durante el período analizado. Este trabajo proporciona una visión integral del impacto de la Guerra de Ucrania en las noticias financieras y establece las bases para futuras investigaciones en esta área.

## PALABRAS CLAVE

Guerra de Ucrania, Ucrania, Rusia, sanción, Unión Europea, petróleo, gas natural, grano, noticias financieras, análisis de tópicos, análisis de sentimiento

## ABSTRACT

This study investigates the impact of the Ukraine War on financial news using text mining techniques. The objective is to analyze thematic changes and sentiment in a sample of financial news. To achieve this, a corpus structuring process was conducted, followed by sample analysis and text preprocessing to ensure data quality and coherence. Subsequently, topic analysis and sentiment analysis techniques were applied to identify key terms that have been impacted by the war and evaluate the associated attitude in the news. The results of the study demonstrate significant changes in the topics discussed and the sentiment expressed in financial news during the analyzed period. This work provides a comprehensive insight into the impact of the Ukraine War on financial news and establishes the groundwork for future research in this area.

## KEY WORDS

Ukraine war, Ukraine, Russia, sanction, European Union, crude, natural gas, grain, financial news, topic analysis, sentiment analysis

## ÍNDICE

1. INTRODUCCIÓN .....	7
2. OBJETIVOS.....	9
3. METODOLOGÍA .....	10
4. MARCO TEÓRICO Y ESTADO DE LA CUESTIÓN.....	12
4.1. CRONOLOGÍA DEL CONFLICTO ENTRE RUSIA Y UCRANIA.....	12
4.2. CONSECUENCIAS DEL CONFLICTO .....	18
5. ANÁLISIS EXPLORATORIO DEL IMPACTO DE LA GUERRA EN LAS NOTICIAS FINANCIERAS .....	24
5.1. RECOGIDA DE DATOS Y ESTRUCTURACIÓN DE LA MUESTRA .....	26
5.2. ANÁLISIS DE LA MUESTRA .....	29
5.3. PREPROCESAMIENTO DE LA INFORMACIÓN.....	32
5.4. <i>TOPIC ANALYSIS</i> .....	36
5.5. ANÁLISIS ENFOCADO EN EL CONFLICTO DE RUSIA Y UCRANIA ..	45
5.6. SENTIMENT ANALYSIS .....	51
6. CONCLUSIONES .....	55
7. LÍMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN.....	59
8. BIBLIOGRAFÍA.....	61
9. ANEXOS.....	64

## ÍNDICE DE TABLAS

Tabla 1: Expresiones regulares comunes.....	27
Tabla 2: Clasificación de las noticias en función del número de palabras .....	29
Tabla 3: Clasificación de sentimientos más segmentada.....	54

## ÍNDICE DE FIGURAS

Figura 1: Esquema de la metodología .....	11
Figura 2: Mapa de la OTAN.....	15
Figura 3: Evolución de la ocupación rusa de Ucrania en el primer año de la guerra .....	17
Figura 4: Estimación de los ingresos de Rusia procedentes de las exportaciones de combustibles fósiles de la UE.....	20
Figura 5: La infraestructura del gas en Europa.....	21
Figura 6: Frecuencia de noticias por mes (I) .....	29
Figura 7: Histograma del número de palabras por noticia.....	30
Figura 8: Frecuencia de noticias por mes (II).....	31
Figura 9: Nube de palabras a partir de la frecuencia absoluta.....	34
Figura 10: Nube de palabras a partir de la matriz TF-IDF .....	35
Figura 11: Regla del codo para LDA .....	37
Figura 12: Top 10 palabras por topic .....	38
Figura 13: Evolución de los topics en el tiempo .....	38
Figura 14: Nube de palabras para el topic 1 .....	39
Figura 15: Evolución del topic 1 en el tiempo.....	40
Figura 16: Nube de palabras para el topic 2 .....	41
Figura 17: Evolución del topic 2 en el tiempo.....	41
Figura 18: Nube de palabras para el topic 3 .....	42
Figura 19: Evolución del topic 3 en el tiempo.....	43
Figura 20: Nube de palabras para el topic 4 .....	43
Figura 21: Evolución del topic 4 en el tiempo.....	44
Figura 22: Evolución de la frecuencia de los términos "import" y "export" en el tiempo .....	45
Figura 23: Evolución de la frecuencia de los términos "russia" y "ban" en el tiempo ...	46

Figura 24: Evolución de la frecuencia de los términos "iraq" y "turkey" en el tiempo..	47
Figura 25: Evolución de la frecuencia de los términos "crude" y "russia" en el tiempo	48
Figura 26: Evolución de la frecuencia de los términos "china" e "india" en el tiempo..	49
Figura 27: Evolución de la frecuencia del término "natural gas" en el tiempo .....	49
Figura 28: Evolución de la frecuencia del término "grain" en el tiempo .....	50
Figura 29: Evolución del sentimiento en el tiempo (I).....	53
Figura 30: Evolución del sentimiento negativo y de la frecuencia del término "russia" en el tiempo .....	53
Figura 31: Evolución del sentimiento en el tiempo (II).....	54

## 1. INTRODUCCIÓN

En un mundo cada vez más interconectado, los eventos geopolíticos y las crisis internacionales pueden tener repercusiones significativas en diversos aspectos de la sociedad. Uno de los ámbitos en los que estos acontecimientos pueden tener un impacto notable es en el mercado financiero. En este sentido, el presente trabajo de investigación se ha centrado en analizar el efecto de la Guerra de Ucrania en las noticias financieras a través de un proceso de *text mining* en el que se analizará la evolución de los temas y el sentimiento presente en una muestra de noticias de FactSet.

La Guerra de Ucrania comenzó el 24 de febrero de 2022 y ha sido uno de los conflictos más importantes y mediáticos de la última década. Sus consecuencias no se limitan únicamente al ámbito geopolítico, sino que también se extienden al mundo de las finanzas y *commodities*. La incertidumbre generada por la guerra, junto con los cambios en las relaciones internacionales y los posibles efectos económicos, han despertado el interés de los investigadores y analistas financieros.

El objetivo principal de este estudio es examinar si la Guerra de Ucrania ha tenido un impacto en las noticias financieras, tanto en los temas tratados como en el sentimiento expresado en ellas. Para lograrlo, se ha utilizado un enfoque basado en el *text mining*, una técnica que permite extraer la información importante y patrones significativos de grandes volúmenes de texto de manera automatizada.

El proceso de *text mining* empleado en esta investigación se divide en cinco etapas fundamentales: estructuración del *corpus*, análisis de la muestra, preprocesamiento de la información, *topic analysis* y *sentiment analysis*. La metodología de *topic analysis* se centra en identificar los temas más relevantes presentes en las noticias financieras durante el periodo analizado, relacionados directa o indirectamente con la Guerra de Ucrania. Por su parte, el *sentiment analysis* tiene como objetivo determinar el tono emocional de las noticias en relación con la guerra, es decir, si transmiten una perspectiva positiva, negativa o neutral.

Se ha establecido como período de análisis desde un mes antes del inicio de la guerra hasta marzo de 2023, abarcando aproximadamente un año de noticias financieras. Mediante el uso de técnicas avanzadas de procesamiento de lenguaje natural, se espera obtener resultados precisos y reveladores que ayuden a comprender mejor cómo la Guerra

de Ucrania ha influido en las noticias financieras y, por consiguiente, en la percepción y el comportamiento de los mercados financieros y sus participantes.

## 2. OBJETIVOS

A continuación, se presentan los objetivos principales de la investigación:

1. **Identificar si la Guerra de Ucrania ha tenido un impacto en los temas tratados en las noticias financieras:** El objetivo principal de esta investigación es determinar si la Guerra de Ucrania ha tenido alguna influencia significativa en los temas abordados en las noticias financieras. En el análisis, se examinará si se ha generado un desplazamiento en el enfoque de los temas financieros hacia cuestiones relacionadas con el conflicto, como sanciones económicas, impacto en los precios de las *commodities* y cambios en las actividades comerciales.
2. **Analizar los temas tratados en las noticias financieras:** El segundo objetivo de este estudio es identificar los temas más relevantes abordados en las noticias financieras durante el período comprendido entre el mes anterior al inicio de la Guerra de Ucrania en febrero de 2022 y marzo de 2023. Se busca identificar los temas más relevantes abordados en dicho período y comprender las principales preocupaciones y enfoques de los medios de comunicación financieros.
3. **Estudiar como la Guerra de Ucrania ha afectado a determinados términos y conceptos:** Otro objetivo fundamental de este trabajo de investigación es examinar cómo la guerra ha influido en los términos y palabras clave utilizados en las noticias financieras. Se llevará a cabo un análisis detallado para identificar los términos más frecuentes y relevantes asociados al conflicto, así como para evaluar cómo han evolucionado a lo largo del tiempo.
4. **Evaluar el impacto emocional de la Guerra de Ucrania en las noticias financieras:** Otro objetivo importante de esta investigación es determinar el sentimiento expresado en las noticias financieras en relación con la Guerra de Ucrania. Se empleará la técnica de *sentiment analysis* para analizar si las noticias transmiten una perspectiva mayoritariamente positiva, negativa o neutral sobre el conflicto y cómo ha evolucionado esta percepción a lo largo del tiempo. Esto permitirá comprender mejor las emociones y la actitud general que ha prevalecido en los medios financieros en relación con la guerra y si ha habido cambios significativos en el sentimiento expresado a medida que el conflicto se ha desarrollado.

### 3. METODOLOGÍA

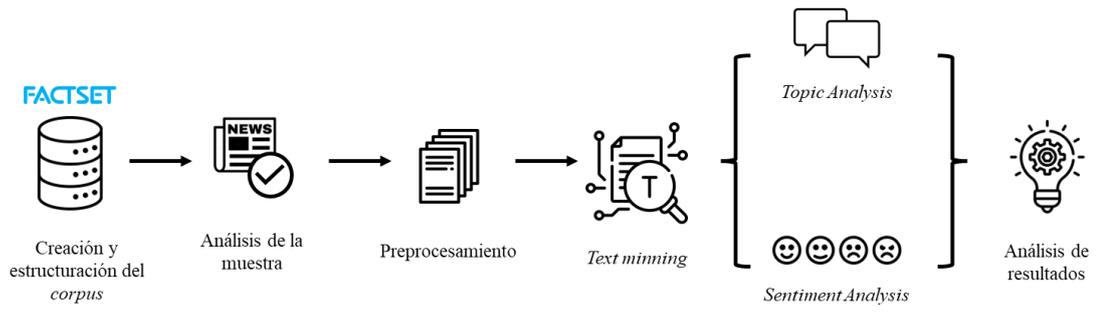
En primer lugar, se llevará a cabo un estudio teórico de las causas de la Guerra de Ucrania, el estallido del conflicto y las consecuencias desde un punto de vista económico. Para ello, se recopilará información y se establecerá un marco conceptual sólido para comprender el contexto en el que se desarrolló la guerra y sus implicaciones económicas.

En segundo lugar, se realizará un proceso de *text mining* para analizar el impacto de la Guerra de Ucrania en las noticias financieras. Se considera que esta metodología es idónea para alcanzar los objetivos, ya que permite analizar grandes volúmenes de datos no estructurados y obtener conclusiones relevantes a través de la identificación de temas y palabras clave y el análisis de sentimientos.

Este proceso se estructurará en cinco etapas: En primer lugar, se llevará a cabo la creación y estructuración del *corpus* mediante la recopilación y organización de noticias financieras obtenidas de FactSet. En segundo lugar, se realizará un análisis de la muestra, estudiando las características de los textos recopilados y eliminando aquellos que puedan perjudicar al estudio. En tercer lugar, se ejecutará un preprocesamiento del texto, que implicará la limpieza de la información para eliminar caracteres no deseados, la normalización del texto, la eliminación de *stopwords* y la lematización. Este paso es crucial para mejorar la eficacia de las técnicas de análisis.

Una vez realizado el preprocesamiento de los datos, se realizará un análisis de temas (*topic análisis*) mediante técnicas de *clustering*. Esto dará lugar a la identificación de los principales temas y términos presentes en las noticias financieras. En último lugar, se realizará un análisis de sentimiento (*sentiment análisis*) para determinar la actitud o emoción asociada a los textos analizados. Esta técnica permitirá clasificar los textos en positivos, negativos o neutros, lo que será útil para comprender la percepción y el tono de las noticias financieras en relación con la Guerra de Ucrania.

Finalmente, se llevará a cabo un estudio de los resultados obtenidos en el proceso de *text mining*. En este estudio se buscará entender las principales causas y eventos que podrían justificar los resultados. Se espera que estos resultados proporcionen una visión detallada y cualitativa del impacto de la guerra en determinados aspectos económicos y respalden la hipótesis de que la guerra ha tenido un impacto en las noticias financieras.



**Figura 1: Esquema de la metodología**

*Fuente: Elaboración propia*

## 4. MARCO TEÓRICO Y ESTADO DE LA CUESTIÓN

### 4.1. CRONOLOGÍA DEL CONFLICTO ENTRE RUSIA Y UCRANIA

En primer lugar, antes de analizar el impacto de la Guerra de Ucrania en el contexto macroeconómico, es fundamental analizar el contexto histórico y geográfico en el que se enmarca dicho conflicto. A continuación, se presenta una cronología de los eventos más significativos que han dado lugar a la situación actual:

- **1991, Independencia de Ucrania:** Tras la disolución de la URSS<sup>1</sup> en 1991, se celebró en Ucrania un referéndum en el que se aprobó la independencia de Ucrania. Esta independencia supuso una nueva era para el país, donde buscó establecer su propia identidad nacional y forjar su camino hacia la democracia y la autonomía en la arena internacional. (Mendoza, 2022).
- **2004-2005, Revolución Naranja:** La Revolución Naranja tiene su origen en las elecciones presidenciales de 2004, en las que el candidato Viktor Yanukovich, de tendencia derechista y pro-ruso, fue declarado ganador. Sin embargo, su oponente, Viktor Yushchenko, también de tendencia derechista pero pro-occidental, acusó a Yanukovich y al gobierno de haber manipulado el proceso electoral. Esto dio lugar a una serie de protestas y huelgas que terminaron con la anulación de los resultados por parte de la Corte Suprema de Ucrania y la realización de unas nuevas elecciones. En estas nuevas elecciones, el candidato Viktor Yushchenko fue declarado ganador y asumió la presidencia, con el apoyo de la que sería su primera ministra, Yulia Timoshenko. La llegada de Yushchenko al poder dio lugar a una mayor cercanía a la Unión Europea (UE) y, consecuentemente, a un deterioro de las relaciones con Rusia.

La Revolución Naranja se convirtió en un símbolo de lucha por la democracia y la libertad en Ucrania, y marcó un cambio significativo en la política del país. Sin embargo, esta revolución también generó divisiones profundas en la sociedad ucraniana, especialmente entre los partidarios del presidente Yushchenko y los partidarios de Yanukovich (Mendoza, 2022; Dorado Díaz, 2020).

---

<sup>1</sup> URSS: Unión de Repúblicas Socialistas Soviéticas

- **2010, Presidencia de Yanukovich:** En 2010, Yanukovich obtuvo la victoria en las elecciones e inició una campaña judicial contra Yulia Timoshenko que concluyó en una condena de siete años por delitos de abuso de poder y corrupción (Mendoza, 2022; Dorado Díaz, 2020).
- **2013-2014, Revolución del Euromaidan:** En 2013, el gobierno de Yanukovich anunció que no firmaría el Acuerdo de Asociación con la UE, prefiriendo en su lugar una estrecha relación con Rusia. Esto ocasionó el descontento en muchos ucranianos, ya que veían la asociación con la UE como un camino hacia la integración europea y una forma de reducir su dependencia en Rusia. Finalmente, en febrero de 2014, después de varias semanas de enfrentamientos violentos, Yanukovich huyó del país y fue destituido por el Parlamento de Ucrania. Esto provocó que muchos de sus partidarios huyesen del país, y se convocaran nuevas elecciones presidenciales, en las que fue elegido Petro Poroshenko (Mendoza, 2022; Dorado Díaz, 2020).
- **2014, Anexión de Crimea por Rusia:** El nuevo gobierno implantó una serie de medidas que afectaban a la comunidad rusoparlante, como la abolición de la ley que reconocía oficialmente el ruso como segunda lengua en las regiones donde era hablado por más del 10% de la población. Estas medidas dieron lugar a protestas en el sur y sureste del país, especialmente en las ciudades de Simferopol y Sebastopol, en Crimea.

El 27 de febrero de 2014, las tropas que Rusia albergaba en la base naval de Sebastopol, ocultando sus insignias militares, ocuparon la península de Crimea, argumentando la necesidad de proteger a los ciudadanos rusos que vivían allí y los intereses de Rusia en la región. Las tropas rusas bloquearon los puntos de entrada y salida de Crimea mientras se celebraba un referéndum en la península que, según las autoridades de Crimea y Rusia, avalaba la anexión de la región por parte de Rusia. Este referéndum se celebró el 16 de marzo de 2014 y tuvo como resultado un 95,5% de votos a favor de la reintegración de Crimea en Rusia. Sin embargo, la comunidad internacional no reconoció la legalidad del referéndum ni la anexión de Crimea por parte de Rusia, considerándola una violación del derecho

internacional y de la integridad territorial de Ucrania (Mendoza, 2022; Dorado Díaz, 2020; Cúneo, 2018a; Bebler, 2015).

- **2014-2015, Guerra del Donbás:** Tras la anexión de Crimea por parte de Rusia se produjeron disturbios en las regiones del este de Ucrania en las que la población rusoparlante se oponía al nuevo gobierno ucraniano. Los separatistas pro-rusos tomaron el control de varias ciudades y declararon la independencia de la República Popular de Donetsk y la República Popular de Lugansk, en la región del Donbás (Mendoza, 2022; Cúneo, 2018a).
- **2014-2015, Acuerdos de Minsk:** En 2014 y 2015 se firmaron los Acuerdos de Minsk bajo la supervisión de la Organización para la Seguridad y la Cooperación en Europa (OSCE) con el objetivo de poner fin al conflicto en el este de Ucrania. Sin embargo, estos acuerdos fueron violados en múltiples ocasiones (Mendoza, 2022; Cúneo, 2018b).
- **2016-2022, Continuación de la guerra:** Durante este periodo el conflicto en la zona del Donbás continuó activo, aunque con altibajos en la intensidad de las hostilidades. Mientras tanto, el gobierno de Ucrania promovió un proceso de ucranización que estableció el ucraniano como el único idioma oficial, afectando a las lenguas minoritarias como el ruso, el húngaro y el rumano. Esta política también supuso la represión de otros grupos considerados enemigos de la patria, como sindicalistas y militantes comunistas (Mendoza, 2022).
- **2020-2021, Pretensión de entrada en la OTAN<sup>2</sup>:** En 1999 la OTAN comenzó una expansión hacia el este de Europa que supuso la integración de 13 países. En 1999 se incorporaron la República Checa, Hungría y Polonia, en 2004 se produjo la incorporación de Lituania, Estonia, Letonia, Bulgaria, Rumanía, Eslovaquia y Eslovenia, en 2009 de Croacia y Albania, en 2017 Montenegro (Merino, 2022) y finalmente Macedonia en 2020.

En junio de 2020 la OTAN emitió una declaración en la que apoyaba el proceso de reforma de Ucrania y Georgia y otorgó a Ucrania el estatus de “Socio de Oportunidades Mejoradas” que respaldaba la decisión de Kiev de buscar la

---

<sup>2</sup> OTAN: Organización del Tratado del Atlántico Norte



- **2022, Estallido de la guerra:** Debido a la negativa de Estados Unidos de establecer un compromiso formal en el que declarase la neutralidad de Ucrania y no instase a Kiev a desistir su política de ucranianización y su apoyo a los ultranacionalistas, el 24 de febrero de 2022 Putin, el presidente de Rusia, inició una “operación militar especial” sobre el territorio ucraniano. Putin justificó esta operación como un acto de defensa sobre las repúblicas rusas, para detener el genocidio sobre la población filo-rusa y desmilitarizar y desnazificar Ucrania (Merino, 2022)
- **2023, Contexto actual de la guerra:** Un año después de la invasión, la artillería rusa ya se ha hecho con el control de algo menos del 75% del territorio. Sin embargo, no se ha producido la conquista acelerada que Putin esperaba por lo que, tras el fracaso y los retrocesos de Rusia en la primera fase de la guerra y a finales de 2022, el Kremlin sigue manteniendo sus objetivos políticos, aunque con una estrategia modificada. En esta estrategia Rusia continúa llevando a cabo bombardeos en diversas zonas de Ucrania, principalmente en el sureste del país, con el objetivo de controlar todo el territorio de Donbás y la costa del Mar Negro y convertir así a Ucrania en un estado continental, asfixiando su economía mediante el bloqueo de los puertos que canalizan la mayor parte de sus exportaciones. Asimismo, el enfoque actual del Kremlin parece consistir en ganar territorio al mismo tiempo que presionar a la UE en el ámbito energético y agravar las dificultades económicas producidas por el conflicto, para obligar a Ucrania a negociar la paz a cambio de territorio (Milosevich-Juaristi, 2023).

## Evolución de la ocupación rusa de Ucrania en el primer año de guerra

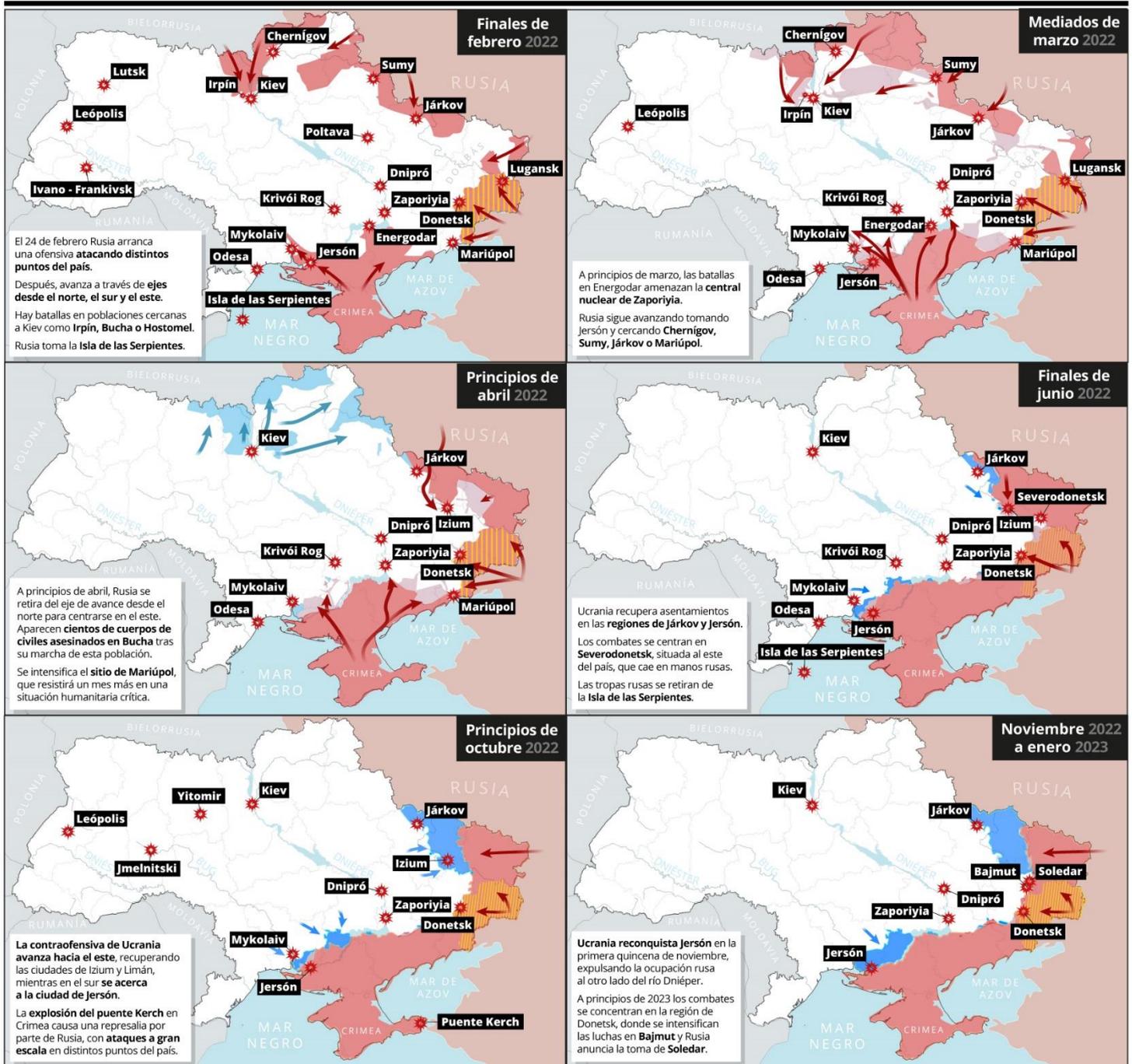


Figura 3: Evolución de la ocupación rusa de Ucrania en el primer año de la guerra

Fuente: Epdata (2023)

## 4.2. CONSECUENCIAS DEL CONFLICTO

La reciente invasión de Ucrania por parte de Rusia ha tenido un impacto significativo no solo en Europa, sino también a nivel mundial. Además de consecuencias políticas y sociales, el conflicto ha generado una gran cantidad de efectos en la economía y en las relaciones internacionales. Las relaciones entre Rusia y la UE se han visto afectadas de manera negativa y se han acercado a un punto crítico. Sin embargo, pesar de los intentos de Rusia de dividir a los países europeos, la UE ha respondido a la guerra con unidad y la invasión ha reactivado el vínculo transatlántico, que no estaba en su mejor momento (de la Corte Ibáñez, 2023).

La UE ha respondido a la guerra con dos acciones principales. En primer lugar, se han aprobado varios paquetes de sanciones económicas contra Rusia y Bielorrusia. En octubre de 2022, Rusia acumulaba ya 11.700 sanciones (de la Corte Ibáñez, 2023). Estas sanciones han sido posibles gracias a la globalización producida en los últimos años en los ámbitos económicos, financieros y tecnológicos. Dentro del paquete de sanciones impuestas por la UE<sup>4</sup> se incluían la congelación de activos de entidades de determinados sectores (bancos e instituciones financieras, compañías de defensa, organizaciones de comunicación, etc.), sanciones de movilidad y congelación de activos a altos cargos gubernamentales, militares, individuos de determinadas afiliaciones, empresarios, propagandistas y oligarcas y sanciones económicas.

Dentro de las sanciones económicas se engloban:

- **Sanciones financieras:** Exclusión del sistema SWIFT<sup>5</sup> a diez bancos rusos, restricciones de acceso a los mercados financieros y de capital de la UE, veto a las transacciones con el Banco Central Ruso, prohibición del suministro a Rusia de billetes denominados en euros y de la provisión de carteras de criptomonedas.
- **Transporte:** Cierre del espacio aéreo de la UE a las aerolíneas rusas, cierre de puertos de la UE a buques rusos, restricciones a los operarios rusos de transporte por carretera, prohibición del transporte marítimo de petróleo y de las

---

<sup>4</sup> Información obtenida de la página web del Consejo Europeo: <https://www.consilium.europa.eu/en/policies/sanctions/restrictive-measures-against-russia-over-ukraine/sanctions-against-russia-explained/>

<sup>5</sup> El Sistema SWIFT es una red global de comunicación electrónica que permite la transferencia segura y rápida de información y fondos entre bancos y otras instituciones financieras en todo el mundo.

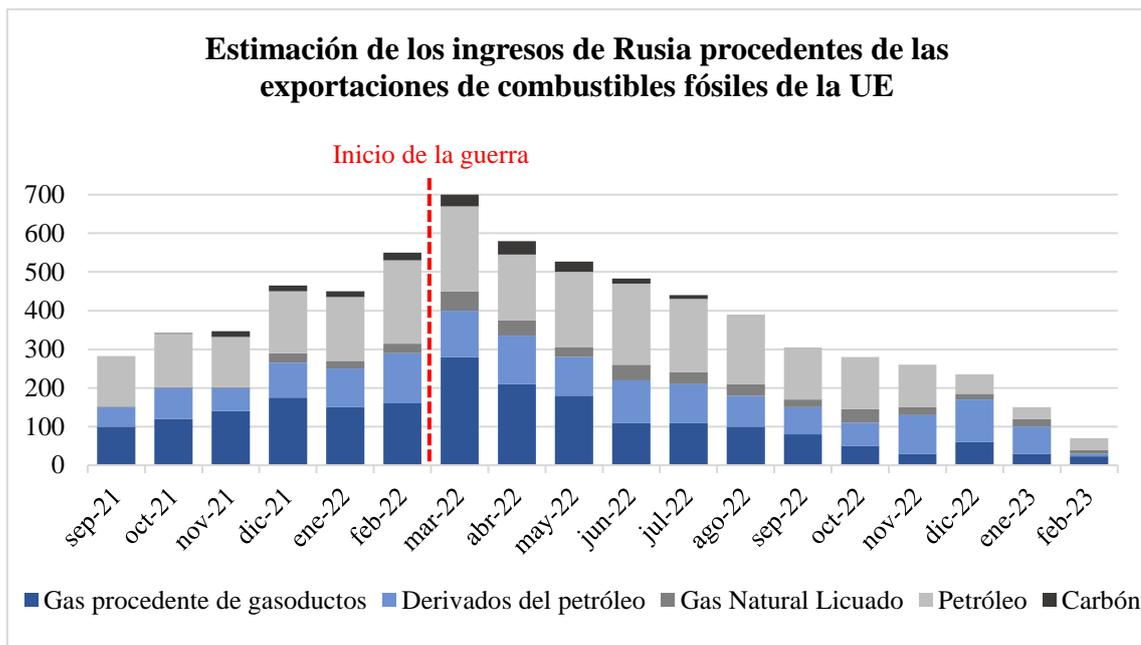
exportaciones a Rusia de productos y tecnología del sector de la aviación, marítimo y espacio.

- **Energía:** Sanciones a las importaciones de Rusia de carbón, limitaciones en el precio del petróleo ruso transportado por mar, sanciones a las exportaciones de Rusia de productos y tecnologías del sector de la refinería de petróleo y a las inversiones en el sector de la minería y la energía.
- **Defensa:** Prohibición de exportaciones a Rusia de equipos militares, armas y munición.
- **Materias primas y otros bienes:** Prohibición de exportaciones de bienes de lujo a Rusia y de importaciones de acero, hierro, cemento, asfalto, madera, papel, plástico, marisco, tabaco, cosmética, alcohol, oro y joyería por parte de la UE.
- **Servicios:** Prohibición de proveer a Rusia o a sus habitantes de servicios arquitectónicos, de ingeniería, de consultoría tecnológica, legal, de publicidad y marketing.

Dentro de estas sanciones cabe destacar las sanciones impuestas al petróleo por la UE y los países del G7<sup>6</sup>. En cuanto al sector de los combustibles, Rusia es uno de los principales exportadores de la UE. Antes de la invasión, la UE exportaba más del 60% de los combustibles fósiles de Rusia (Myllyvirta, 2022a). Sin embargo, tras el inicio de la guerra, la UE y los países del G7 se vieron forzados a implementar medidas para reducir su consumo. Por un lado, la UE y Reino Unido prohibieron las importaciones marítimas de petróleo y productos derivados a partir de diciembre de 2022 y, por otro lado, los países miembros del G7 establecieron un techo en el precio del petróleo para evitar que Rusia obtuviese más de \$60 por barril. Esta sanción es la más significativa hasta la fecha en materia de importaciones energéticas (Myllyvirta, 2022b; Sandri, 2022). En consecuencia, tras la aplicación de estas sanciones, los ingresos de Rusia en términos de combustibles fósiles cayeron en los meses de enero-febrero de 2023 un 50% en comparación con el mes de ingresos máximos de 2022, reduciendo los ingresos procedentes de la UE casi en un 90% (Ilas, 2023).

---

<sup>6</sup> G7: Organización intergubernamental compuesta por siete países líderes de la economía mundial: Canadá, Francia, Alemania, Italia, Japón, el Reino Unido y Estados Unidos.

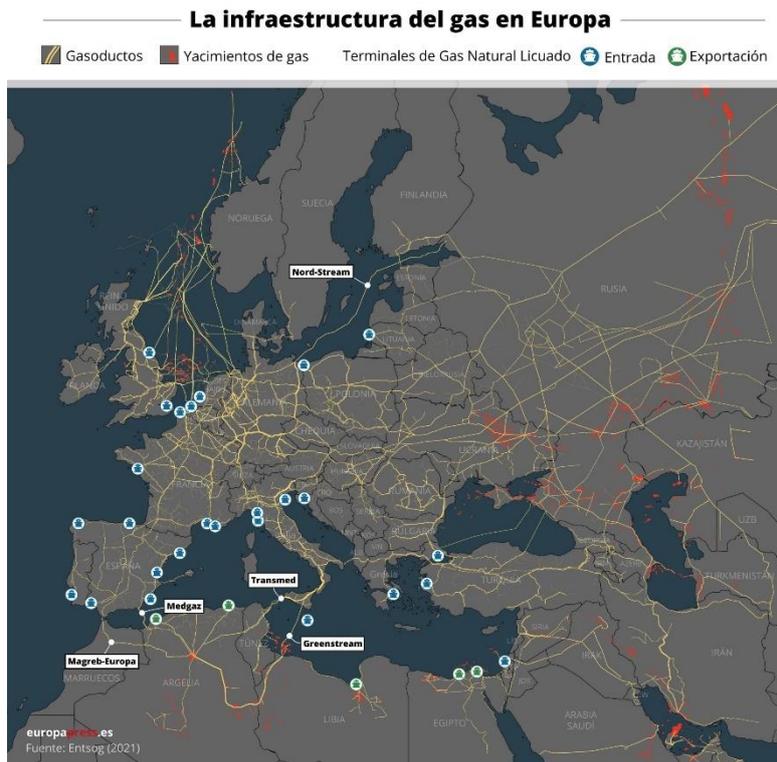


**Figura 4: Estimación de los ingresos de Rusia procedentes de las exportaciones de combustibles fósiles de la UE**

*Fuente: Adaptado de Illas (2023)*

En segundo lugar, la UE ha brindado apoyo militar a Ucrania desde el inicio de la guerra, y ha recaudado más de 3.100 millones de euros del Fondo Europeo para la Paz para respaldar militarmente a Ucrania. Teniendo en cuenta también el apoyo bilateral brindado por los Estados de la Unión, en diciembre de 2022 el total de la ayuda se acercaba a los 9.000 millones de euros (de la Corte Ibáñez, 2023). Sin embargo, debido a que Ucrania no es un país miembro de la OTAN, la ayuda por parte de los países europeos y de Estados Unidos se ha limitado a una ayuda económica, descartando el envío de tropas y la opción militar, ya que dichas acciones podrían tener grandes consecuencias, entre ellas una guerra nuclear (García Encina, 2023).

Finalmente, estas acciones han dado lugar a consecuencias desfavorables para la economía europea. Además de los gastos relacionados con la asistencia militar, las represalias rusas contra la UE por las sanciones impuestas han dado lugar a una crisis energética en Europa que ha generado grandes perjuicios económicos. Antes de la guerra, el 40% del gas importado por la UE provenía de Rusia, y este porcentaje ascendía al 100% para Bulgaria, el 80% para Polonia, alrededor del 60% para Austria y Hungría, el 50% para Alemania y el 40% para Italia (Otero-Iglesias y Steinberg, 2022).



**Figura 5: La infraestructura del gas en Europa**

*Fuente: Epdata (2023)*

En agosto de 2022, en respuesta al anuncio de las sanciones al petróleo por parte de la UE y de los países del G7 y a la suspensión de la certificación del gasoducto Nord Stream 2 por parte de Alemania, Rusia anunció la suspensión del suministro del gasoducto Nord Stream 1 de manera indefinida, debido a una supuesta fuga (Sandri, 2022). Esto provocó una gran alarma en Europa ya que las consecuencias de esta suspensión no se limitan únicamente a un invierno difícil, sino que se pronostican al menos dos o tres años de dificultades. Asimismo, no se espera que aparezcan fuentes alternativas significativas de gas en el mercado global antes de 2025, año en el que se espera la puesta en funcionamiento de las plantas de gas natural licuado de Estados Unidos, y cualquier proyecto nuevo en otros países requeriría de varios años para su puesta en funcionamiento (Vakulenko, 2022).

En consecuencia, Europa se va a ver obligada a acelerar su autonomía energética, aumentando sus inversiones tanto en tecnología de ahorro energético como en fuentes de energía alternativas. No obstante, hasta que lo logre, deberá establecer contratos para la entrega de gas procedente de otros países en condiciones relativamente desfavorables,

con precios elevados y largas duraciones. Además, en vista de los actuales precios del gas, muchas empresas europeas, e incluso sectores completos con un alto consumo energético, se van a enfrentar a la inviabilidad económica. Como respuesta, estas organizaciones deberán optar por cerrar sus operaciones por completo, intentar resistir hasta que los precios del gas disminuyan o trasladar su producción fuera de la UE (Vakulenko, 2022).

Cabe destacar que, aunque la suspensión definitiva del gasoducto se produjo en agosto, durante el mes de junio ya se habían reducido las exportaciones de gas a través del gasoducto en un 75% y en julio se mantuvo cerrado durante 10 días por unas supuestas reparaciones, reabriendo con una capacidad reducida en un 50% (BBC, 2022).

A finales de noviembre de 2022, debido al conflicto y a la suspensión de las exportaciones de gas natural por parte de Rusia, los sobrecostos para los países de la UE ascendían a alrededor de 941.850 millones de euros, lo que llevó a los gobiernos europeos a aprobar medidas económicas por un total de aproximadamente 659.295 millones de euros. En consecuencia, en diciembre de 2022, la deuda de la mitad de los estados miembros de la UE se acercaba al 60% de su PIB. Por ello, la UE ha desarrollado una nueva estrategia para reducir su dependencia energética de Rusia (de la Corte Ibáñez, 2023).

Por otro lado, aunque el aumento en el precio de los combustibles y la electricidad ha sido significativo, también se han visto afectados otras materias primas de las que Rusia, y en menor medida Ucrania, son exportadores, como el trigo, cobre, níquel, platino, paladio y titanio (Otero-Iglesias y Steinberg, 2022). Esto supuso que, en 2022, la tasa de inflación alcanzara niveles históricos, superando el 10% en 16 de los 27 países de la UE, teniendo un impacto importante en el precio de la energía y los alimentos frescos, que aumentaron un 39,6% y 11,1% respectivamente (de la Corte Ibáñez, 2023).

Asimismo, en marzo de 2023 se produjo la amenaza por parte de Rusia de retirar su participación en la Iniciativa de Cereal del Mar Negro, pacto que facilitaba la exportación del grano y otros productos alimentarios de Ucrania (Prokopenko, 2023; Moloney, 2023). Ucrania es un importante exportador mundial de girasol, maíz, trigo y cebada, y más de la mitad del grano de trigo adquirido por el Programa Mundial de Alimentos (PMA) el año pasado provino de allí (Moloney, 2023). Por lo que, la amenaza por parte de Rusia

causó una grave preocupación en todo el mundo, ya que la continuación de este acuerdo era fundamental para garantizar la alimentación de millones de personas y, en especial, el abastecimiento de los países en desarrollo (Prokopenko, 2023).

Este acuerdo fue impulsado en julio de 2022 por la ONU y Turquía para permitir que Ucrania continuase con la exportación de grano a pesar del conflicto con Rusia, que había supuesto un bloqueo de los puertos ucranianos en el Mar Negro. Al mismo tiempo, las Naciones Unidas acordaron ayudar a Moscú a facilitar sus propios envíos agrícolas. Gracias a este acuerdo Ucrania pudo exportar 24,5 millones de toneladas de grano a través de los puertos del Mar Negro en 2022 (Prokopenko, 2023).

El acuerdo estaba estructurado para renovarse de manera automática por 120 días salvo que algunas de las partes presentasen objeciones. Sin embargo, en el mes de marzo de 2023 Rusia advirtió que únicamente aceptaría una extensión de 60 días si durante ese periodo se producían avances en la exportación de sus propios productos agropecuarios, que se habían visto afectados de forma indirecta por las sanciones occidentales (Prokopenko, 2023; Moloney, 2023).

No obstante, aunque finalmente Rusia otorgó una extensión del acuerdo por 60 días (UNCTAD, 2023), todavía se mantiene una alta preocupación respecto a este tema, ya que el conflicto también ha afectado negativamente a los agricultores ucranianos. Se teme que este año no se pueda realizar la cosecha puesto que muchas tierras de cultivo y el equipo utilizado en las granjas están minados (Moloney, 2023).

A su vez, el conflicto ha supuesto la reacción de potencias asiáticas como China e India. China ha mostrado su apoyo a Rusia en el ámbito diplomático, propagandístico y comercial. Sin embargo, su apoyo no es totalmente pleno ya que China no reconoce los argumentos rusos que justifican la invasión a Ucrania ni la independencia y anexión de las repúblicas de Donetsk y Lugansk, pero sí señala a Estados Unidos y a la OTAN como responsables de la invasión por su presión desmesurada contra Rusia (Esteban, 2023). Asimismo, las sanciones económicas impuestas por la UE en relación con el petróleo han ocasionado que Rusia busque compradores en otros países, como China e India, para suplir la demanda de la Unión de Europa (Barriocanal, 2023; El Economista, 2023).

## 5. ANÁLISIS EXPLORATORIO DEL IMPACTO DE LA GUERRA EN LAS NOTICIAS FINANCIERAS

A continuación, se va a realizar un análisis exploratorio de los temas presentes en las noticias financieras en el último año, con el objetivo de estudiar el impacto de la Guerra de Ucrania en dichos temas.

Este estudio se va a enmarcar en la disciplina de *text mining*. El *text mining* es una disciplina utilizada para extraer información relevante a partir de grandes cantidades de texto no estructurado. Su objetivo principal es descubrir patrones, tendencias, relaciones y conocimientos ocultos en los documentos de texto (Gupta y Lehal, 2009).

A continuación, se presentan las técnicas clave de esta disciplina (Gupta y Lehal, 2009):

- **Estructuración del *corpus*:** En primer lugar, es fundamental llevar a cabo una adecuada estructuración del *corpus*. El *corpus* hace referencia al conjunto o colección de textos que se utiliza como base de datos para realizar el análisis y extracción de la información. Esto implica llevar a cabo una recopilación de los textos y organizarlos en un formato coherente y accesible. El tamaño y la composición del *corpus* pueden variar dependiendo del objetivo del análisis y los recursos disponibles. Un *corpus* bien construido y representativo puede proporcionar una base sólida para extraer información, descubrir patrones, entrenar modelos de aprendizaje y realizar diversas tareas de procesamiento de lenguaje natural en el ámbito del *text mining*.
- **Análisis de la muestra:** Este análisis consiste en estudiar las características de la muestra y eliminar aquellos datos que puedan perjudicar al estudio. Dentro de este análisis se incluye un estudio de los elementos del *corpus* en función de características como el número de palabras y la fecha.
- **Preprocesamiento de texto:** Antes de aplicar las técnicas de análisis de *text mining*, es necesario realizar un preprocesamiento del texto. Esto implica limpiar el texto, eliminar caracteres no deseados, normalizar el texto (por ejemplo, convertir todo el texto a minúsculas), eliminar *stopwords* y lematizar las palabras.

- **Topic analysis:** Esta metodología, también conocida como *clustering*, consiste en agrupar documentos de texto similares en categorías o clústeres. Esta técnica permite descubrir patrones y temas dentro de un *corpus* de texto sin la necesidad de una categorización previa.
- **Sentiment analysis:** El *sentiment analysis* o análisis de sentimiento es una técnica que busca determinar la actitud o emoción asociada a un texto. En esta metodología se clasifica el texto en positivo, negativo o neutro, lo que puede ser útil para analizar opiniones, reseñas de productos, comentarios en redes sociales, etc.

Siguiendo esta estructura, en esta investigación se van a llegar a cabo los siguientes pasos: En primer lugar, se realizará una estructuración y análisis de una muestra de noticias financieras procedentes de FactSet desde enero de 2022 hasta marzo de 2023. En este paso se llevará a cabo una estructuración de los datos, separando ciertos argumentos como el título, la fecha y el contenido de la propia noticia. En segundo lugar, se realizará un preprocesamiento de la información para eliminar caracteres no relevantes y normalizar los datos. En tercer lugar, se aplicará un *topic analysis* para identificar los temas principales de las noticias de la muestra y, a continuación, se realizará un análisis enfocado en el conflicto entre Rusia y Ucrania, analizando la frecuencia de la aparición de los términos relacionados con el conflicto en el *corpus*. Finalmente, se llevará a cabo un *sentiment analysis* para evaluar el sentimiento positivo, negativo o neutro de las noticias.

## 5.1. RECOGIDA DE DATOS Y ESTRUCTURACIÓN DE LA MUESTRA

Para este estudio se ha construido una muestra con 1.834 noticias financieras en inglés procedentes de FactSet. FactSet es una empresa proveedora de servicios de información financiera y análisis para profesionales del campo financiero. Es una empresa líder en el sector y ampliamente reconocida, utilizada por profesionales de las finanzas e inversión en todo el mundo.

Para llevar a cabo este análisis se va a emplear el lenguaje de programación Python, ya que su capacidad para manejar grandes cantidades de datos y su facilidad para implementar algoritmos complejos lo convierten en la herramienta idónea para nuestra investigación.

En primer lugar, puesto que las noticias se encuentran en formato pdf (uno por cada noticia), se debe extraer la información contenida en cada uno de ellos y crear un *dataframe*, en adelante *corpus*, que albergue el texto de las noticias. Una vez creado este *corpus*, vamos a analizar si hay alguna noticia vacía o información duplicada. En el *corpus* no se ha encontrado ninguna fila vacía, pero si una noticia duplicada, por lo que una de las copias se eliminará. De esta forma nuestro *corpus* estará formado por 1833 noticias.

En segundo lugar, se va a realizar una clasificación del texto. En esta clasificación se obtendrán cinco variables: “Título”, “Fecha”, “Contenido”, “Industrias” y “Temas”. Se han elegido estas cinco variables puesto que todos los textos siguen una misma estructura en la que la primera frase corresponde con el título, la segunda hace referencia a la fecha y seguidamente se muestra el contenido. Sin embargo, hay algunos textos en los que al final del contenido se mencionan los temas o las industrias más relevantes al lado de la palabra “Industries:” o “Subjects:”, por lo que se va a crear una columna en la que se recopile esta información para aquellos textos que la identifiquen y aparezca la palabra la palabra “None” para los que no.

Para llevar a cabo esta clasificación se van a utilizar las expresiones regulares. Estas expresiones son patrones de búsqueda y manipulación de texto que se utilizan para encontrar cadenas de caracteres específicas en un documento o conjunto de datos. Esta herramienta permite realizar la extracción del texto de manera eficiente.

Algunos de los caracteres especiales más comunes en las expresiones regulares son:

**Tabla 1: Expresiones regulares comunes**

Caracter	Definición
\d	Hace referencia a caracteres numéricos (0-9)
\w	Hace referencia a caracteres alfanuméricos y el guion bajo
\s	Hace referencia a espacios en blanco (incluyendo espacios, tabulaciones y saltos de línea)
\n	Hace referencia a saltos de línea
^	Hace referencia al inicio de una cadena
\$	Hace referencia al final de una cadena
.	Hace referencia a cualquier carácter a excepción de un salto de línea
*	Indica que el elemento que le precede puede aparecer 0 o más veces
?	Indica que el elemento que le precede puede aparecer 0 o una vez
[]	Define un conjunto de caracteres que puede aparecer en una posición específica
()	Agrupar elementos y aplicar operaciones a todo el grupo
	Indica alternativas entre palabras

*Fuente: Adaptado de Kuchling (2023)*

En la estructuración de nuestro *corpus* se han empleado las siguientes expresiones regulares:

- Título:** Para encontrar la parte del texto que hace referencia al título se han seleccionado todos los caracteres que aparecen antes de la fecha. En un principio se había utilizado una expresión regular que obtenía todos los caracteres desde el inicio de la cadena hasta el primer salto de línea. Sin embargo, puesto que hay títulos que ocupan más de una línea o están formados por dos frases, la expresión regular que se ha utilizado obtiene todos los caracteres desde el inicio del texto hasta la aparición de una de las siguientes palabras (“Monday”, “Tuesday”, “Wednesday”, “Thursday”, “Friday”, “Saturday”, “Sunday”). La expresión regular utilizada ha sido la siguiente:

```
r'^(.*) (?:' + '|'.join(dias) + r')'
```

(días hace referencia a una lista de palabras que incluye los días de la semana)

- **Fecha:** La fecha se encuentra justo después del título, y en todas las noticias se encuentra en el mismo formato (e.g. “Thursday, February 09, 2023 11:22:28 AM (GMT)”). Por lo que para extraerla se va a emplear la siguiente expresión regular: `r'\w+, \s\w+\s\d{2}, \s\d{4}\s\d{2}:\d{2}:\d{2}\s\w{2}\s\ (GMT\)'`.

Este patrón busca la fecha con el formato "Día de la semana, Mes Día, Año Hora:Minutos:Segundos Zona horaria (GMT)".

- **Contenido:** Una vez obtenidos el título y la fecha, el resto del texto podría considerarse contenido. Sin embargo, como se ha mencionado anteriormente, en algunos textos tras el desarrollo de la noticia se realiza un resumen de las industrias, identificadores financieros y temas tratados bajo los apartados “Industries:”, “Related Identifiers:” y “Subjects:”. Por ello, para la extracción del contenido se va a emplear la siguiente expresión regular:

`r'\ (GMT\)\s*(.*?) (?:\s*Industries:|\s*Related Identifiers:|\s*Subjects:)'`.

En esta expresión regular se localiza el texto que se encuentra entre la cadena “(GMT)” y las palabras “Industries:”, “Related Identifiers:” o “Subjects:” en caso de haberlas.

- **Industrias:** Esta variable contiene la información del apartado de industrias que aparece en algunas noticias, y la expresión regular empleada ha sido la siguiente: `r'Industries:\s*(.*?) (?:\s*Related identifiers:|\s*Subjects:)'`.

En esta expresión se extrae el contenido que aparece entre la cadena “Industries:” y las cadenas “Related Identifiers:” y “Subjects:” en caso de haberlos.

- **Temas:** Finalmente, en esta variable se van a guardar los temas que los propios autores identifican en las noticias. Para ello se ha empleado la expresión regular `r"Subjects:(.*)"`, que extrae toda la información a partir de Subjects: hasta el final del texto.

## 5.2. ANÁLISIS DE LA MUESTRA

A continuación, se va a llevar a cabo un estudio de la muestra en el que se va a analizar la frecuencia de noticias por mes y el número de palabras por noticia.

En la siguiente figura se puede ver la distribución de noticias en el tiempo. El periodo temporal de la muestra abarca desde el 1 de enero de 2022 hasta el 3 de abril de 2023 y en el gráfico se observa como la frecuencia de noticias por mes se mantiene estable (a excepción de los meses de marzo que tienen un número mayor de noticias y abril que tan solo tiene 4 noticias). Puesto que en el análisis se van a agrupar las noticias por meses, el mes de abril no se va a tener en cuenta, ya que, en caso de hacerlo, se produciría una distorsión de los resultados.



**Figura 6: Frecuencia de noticias por mes (I)**

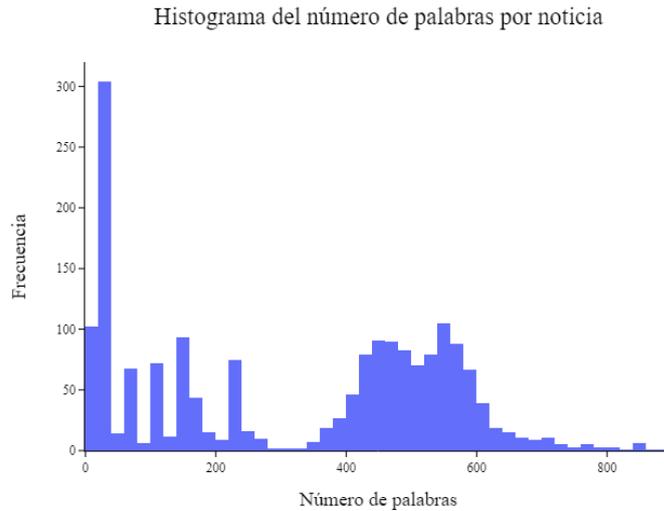
*Fuente: Elaboración propia*

A continuación, se ha realizado un análisis de las noticias en función de la cantidad de palabras que contiene cada una y se ha obtenido la siguiente información:

**Tabla 2: Clasificación de las noticias en función del número de palabras**

Rango	Frecuencia
< 100	493
100 – 300	348
> 300	987

*Fuente: Elaboración propia*

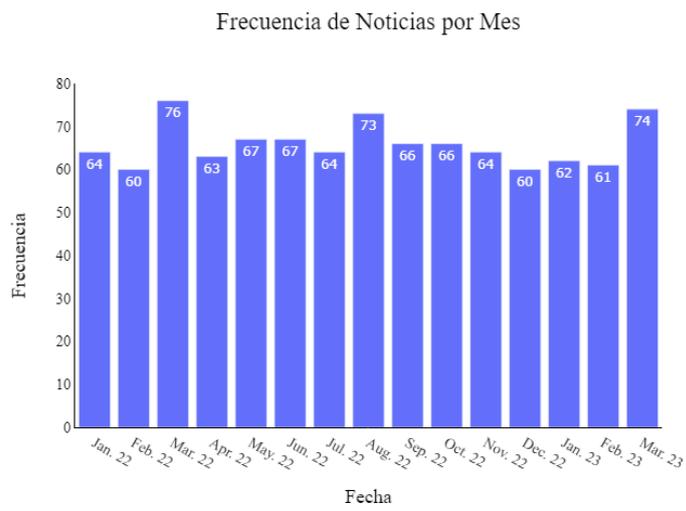


**Figura 7: Histograma del número de palabras por noticia**

*Fuente: Elaboración propia*

En la figura 7 se puede observar cómo hay dos grandes grupos de noticias: noticias con menos de 300 palabras (noticias cortas) y noticias con más de 300 palabras (noticias largas). Con el fin de utilizar noticias con información más sustantiva, se ha decidido excluir aquellas que contengan menos de 300 palabras. Este criterio se basa en que muchas noticias cortas suelen centrarse en datos numéricos, como el inventario de materias primas, sin proporcionar un contexto más amplio para entender su relevancia. Por tanto, se va a considerar que las noticias que superan las 300 palabras tienen más probabilidad de contener información más completa y contextualizada.

Tras haber realizado este filtrado, se ha vuelto a analizar el número de noticias por mes para comprobar que la cantidad de noticias por meses se mantenía estable. En la figura 8 se puede apreciar como existe una estabilidad en la frecuencia de las noticias en el tiempo, aunque durante los meses de marzo y agosto de 2022 y marzo de 2023 se puede observar un aumento en el número de noticias. Estos aumentos pueden deberse a la ocurrencia de eventos significativos durante esos meses, lo que generaría una mayor cobertura mediática y, por lo tanto, más noticias.



**Figura 8: Frecuencia de noticias por mes (II)**

*Fuente: Elaboración propia*

### 5.3. PREPROCESAMIENTO DE LA INFORMACIÓN

En la tercera parte de esta investigación se va a realizar un preprocesamiento de los datos. En este preprocesamiento, se van a llevar a cabo una serie de pasos importantes previos al análisis de texto, incluyendo la conversión del texto a minúsculas, la tokenización para dividir el texto en unidades significativas, la eliminación de *stopwords* para reducir el ruido en el texto, y finalmente la lematización para convertir cada palabra a su forma base y simplificar el análisis.

- **Conversión del texto a minúsculas:** Este paso implica transformar todas las letras mayúsculas presentes en el texto a minúsculas. Este procedimiento es fundamental en el procesamiento de lenguaje natural, ya que la mayoría de los algoritmos consideran que las palabras en mayúsculas y minúsculas son diferentes. Al convertir todo el texto a minúsculas, se asegura que las palabras sean comparables entre sí.
- **Tokenización:** La tokenización es el proceso de dividir el texto en unidades más pequeñas llamadas *tokens*. En el contexto del procesamiento del lenguaje natural, los *tokens* generalmente son palabras, pero también pueden ser frases o partes de palabras. El objetivo de la tokenización es dividir el texto en unidades significativas que puedan procesarse fácilmente. Por ejemplo, si se tuviese una frase como "El inicio de la Guerra de Ucrania", la tokenización dividiría la oración en los siguientes tokens: "El", "inicio", "de", "la", "Guerra", "en" y "Ucrania".
- **Eliminación de *stopwords*:** Las *stopwords* son palabras muy comunes que no aportan mucho significado al análisis del texto y que pueden afectar negativamente al análisis, como "the", "a", "an", "and", "of", "in", etc. Estas palabras suelen eliminarse del texto porque no aportan información relevante para el análisis. Al eliminar las *stopwords*, se reduce el ruido en el texto y el análisis se centra en las palabras más importantes.
- **Lematización:** La lematización es el proceso de transformación de cada palabra a su forma base conocida como "lemma". Por ejemplo, el lemma de "built" sería "build". La lematización es importante porque ayuda a reducir la cantidad de palabras diferentes que aparecen en el texto, lo que simplifica el análisis y hace que los modelos de lenguaje funcionen mejor. Asimismo, la lematización también

sirve para agrupar palabras relacionadas y tratar diferentes formas de la misma palabra como si fueran la misma, lo que mejora la precisión del análisis.

Una vez aplicado este preprocesamiento, se va a realizar un análisis de la variable “Contenido”, que es la que contiene el desarrollo de la noticia. El objetivo de este análisis es identificar las palabras más frecuentes y evaluar si alguna de ellas puede considerarse una *stopword*. De esta manera, se determinará si es necesario ajustar nuestra lista de *stopwords* y mejorar la calidad del análisis de texto.

Al hacer este análisis se han identificado y eliminado las siguientes *stopwords*: "week", "news", "report", "session", "data", "prior", "continue", "weekly", "expect", "bloomberg", "reuters", "yesterday", "note", "add", "api", "article", "jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov" y "dec".

Finalmente, se ha generado una nube de palabras en la que se muestran las palabras con mayor frecuencia de la variable “Contenido”. En la figura 9 se puede observar que las palabras más que se repiten son aquellas relacionadas con el petróleo, lo que indica que este es un tema importante en el contenido analizado. Por otro lado, se puede observar que las palabras "russia" y "china" tienen una frecuencia menor a la esperada. Esto se puede deber a que, al provenir las noticias de FactSet, plataforma enfocada en temas relacionados con los mercados y *commodities*, palabras como "russia" y "china" se utilizan para poner en contexto los temas principales. Es probable que, si las noticias se hubiesen obtenido de otras fuentes, como periódicos, estos términos hubiesen sido más predominantes. Esto se debe a que cada fuente puede tener un enfoque distinto y dar mayor importancia a determinados temas.





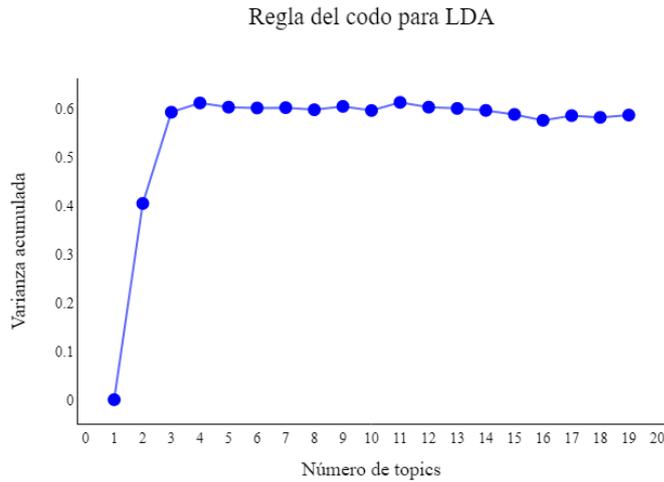
#### 5.4. TOPIC ANALYSIS

En este epígrafe se va a emplear la metodología de *topic analysis* con el objetivo de estudiar cuales son los temas más frecuentes a lo largo del periodo observado. Esta metodología, es una técnica de análisis de datos que se utiliza para identificar y analizar patrones y temas en un conjunto de datos. Para ello, se va a utilizar un algoritmo especializado en identificar y agrupar palabras y frases similares en categorías temáticas conocidas como "temas" o "*topics*". Esta técnica es especialmente útil cuando se trabaja con grandes cantidades de datos no estructurados, como textos, ya que puede ser difícil analizarlos de forma manual. Esta metodología puede aplicarse a una gran variedad de campos, incluyendo la investigación de mercado, el análisis de redes sociales, la investigación de opinión pública y muchos otros.

En esta investigación se va a utilizar el modelo LDA (*Latent Dirichlet Allocation*), que es uno de los algoritmos más populares para llevar a cabo el *topic analysis*. El algoritmo LDA asigna una distribución de probabilidad sobre los *topics* a cada documento, y una distribución de probabilidad sobre las palabras a cada tema. En otras palabras, se trata de un modelo probabilístico que busca encontrar las distribuciones de probabilidad que mejor expliquen los datos que se le proporcionan (Chang et al., 2009; Griffiths y Steyvers, 2004; Blei et al., 2003).

Para identificar el número óptimo de *topics* que deben ser identificados en el conjunto de datos, se va a ajustar un modelo LDA para diferentes números de *topics* (entre 1 y 20). En primer lugar, se va a crear una matriz de términos y documentos conocida comúnmente como *Document-Term Matrix* (DTM). En esta matriz cada celda representa la frecuencia de aparición de una palabra en un documento específico. De esta forma, la matriz DTM representa los documentos del *corpus* como vectores numéricos, permitiendo así aplicar las técnicas de análisis de datos y aprendizaje automático para extraer la información del texto. En segundo lugar, se va a ajustar un modelo LDA para diferentes números de *topics*. Una vez ajustado el modelo, se van a obtener las probabilidades de los temas para cada documento escalando los valores para que estén en el rango [0,1]. Finalmente, se va a calcular la varianza acumulada para cada escenario y elaborar un gráfico para poder aplicar la regla del codo.

La regla del codo es una técnica utilizada para determinar el número óptimo de clústeres en un conjunto de datos en el análisis de *clustering*. El objetivo de esta técnica es encontrar el punto en el cual la adición de más clústeres no proporciona un beneficio significativo en términos de la reducción de la varianza total explicada. En un gráfico se puede identificar este punto como el cambio de tendencia que se asemeja a la forma del codo.

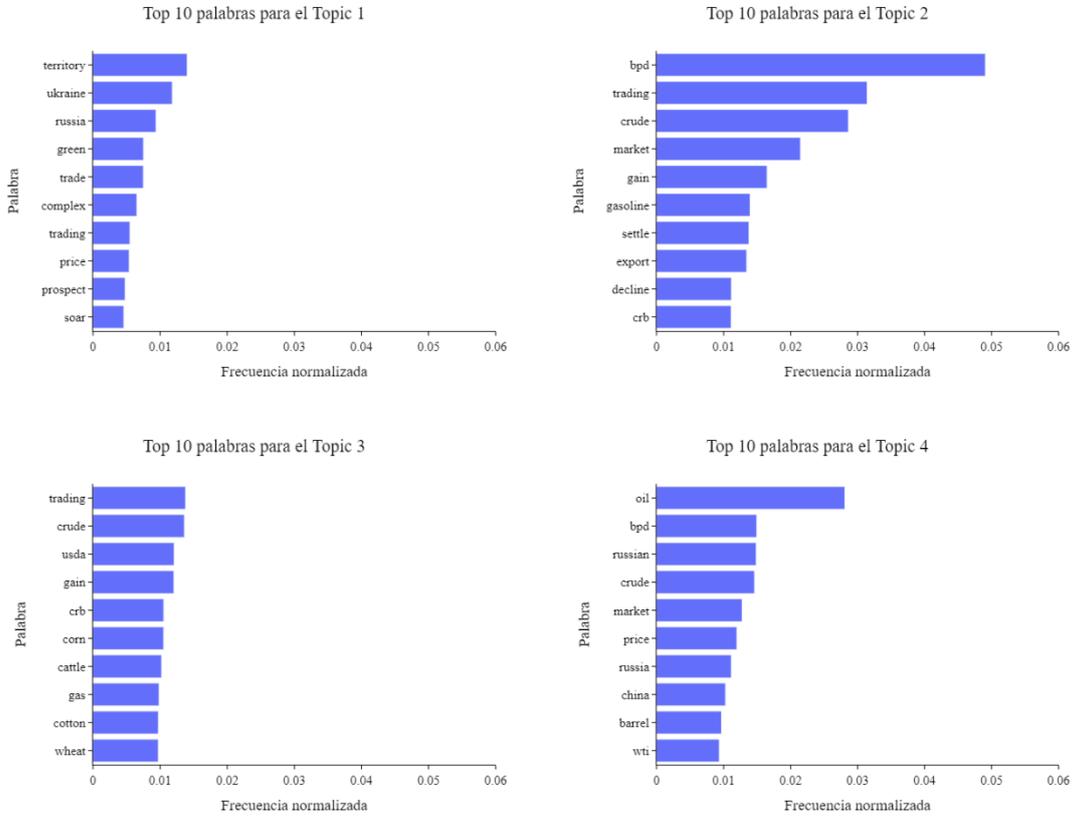


**Figura 11: Regla del codo para LDA**

*Fuente: Elaboración propia*

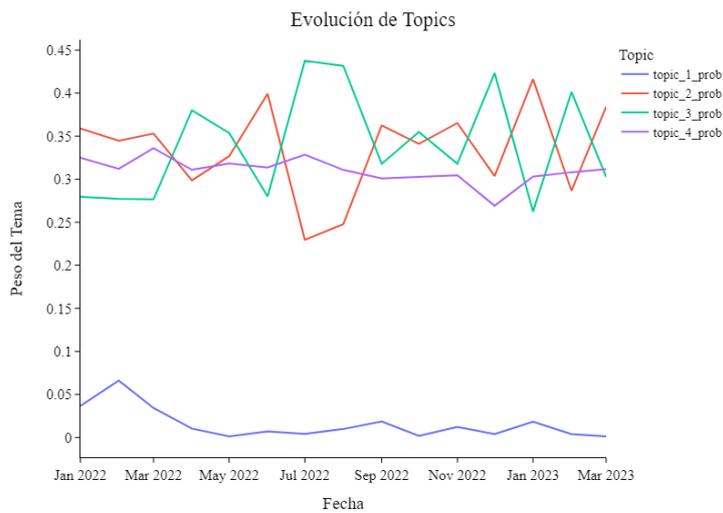
En la figura 11 se observa que el número ideal de *topics* donde se cumple la regla del codo es tres, sin embargo, con el objetivo de enriquecer la investigación y llevar a cabo un análisis más completo, se van a analizar cuatro temas. La elección de este valor se basa en el hecho de que representa el último número antes de que la varianza comience a disminuir. Además, al incluir un cuarto tema, se espera obtener una mayor diversidad y granularidad en el análisis de los datos, así como identificar relaciones adicionales o dimensiones de información que podrían no estar tan visibles con tres temas.

A continuación, se muestran las 10 palabras más frecuentes para cada uno de los *topics* obtenidos y la evolución de los cuatro *topics* en el tiempo:



**Figura 12: Top 10 palabras por topic**

*Fuente: Elaboración propia*



**Figura 13: Evolución de los topics en el tiempo**

*Fuente: Elaboración propia*

En la figura 13 se puede apreciar como el comercio del petróleo ha sido el tema principal de las noticias de FactSet en el último año, con un constante interés por parte de los medios y usuarios. No obstante, también se puede observar cómo, aunque de forma menos evidente, el conflicto entre Rusia y Ucrania ha tenido un papel subyacente en el análisis y discusión de los mercados globales, especialmente durante los meses de febrero y marzo de 2022, cuando se produjo el inicio del conflicto y se generó una mayor atención mediática hacia este tema.

A continuación, se va a realizar un análisis detallado de cada *topic* y de su evolución en el tiempo:

- **Topic 1 - Conflicto y disputas territoriales entre Rusia y Ucrania:** Este *topic* se enfoca en la tensión geopolítica entre Rusia y Ucrania, incluyendo temas como la ocupación territorial y las actividades comerciales.

Nube de palabras para el Topic 1

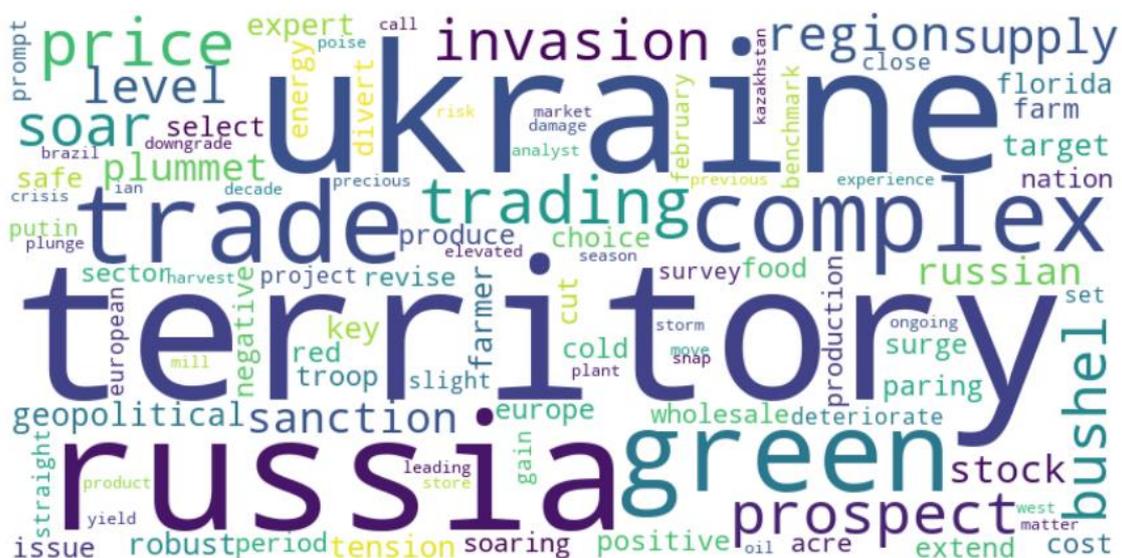


Figura 14: Nube de palabras para el topic 1

Fuente: Elaboración propia

En la figura 15 se observa la evolución del *topic* 1 en el tiempo. Como era de esperar, se observa un aumento significativo en el discurso sobre este tema durante los meses de febrero y marzo de 2022. Sin embargo, resulta interesante destacar que, a pesar de que el estallido de la guerra ocurrió el 24 de febrero de 2022, es en el mes de febrero cuando el *topic* tiene mayor peso. Esto es notable, ya que desde el inicio de la guerra hasta el final

de este mes solo hay cinco días, lo que resalta la relevancia que adquiere este tema en un período tan corto de tiempo e indica que durante las semanas previas a la invasión ya se generaban expectativas y discusiones sobre los eventos que se avecinaban.

Asimismo, en el gráfico se puede observar cómo durante el mes de enero de 2022 las noticias ya anticipaban el conflicto. Por otro lado, se puede apreciar como a medida que ha avanzado el tiempo, se ha reducido la relevancia del tema en las noticias, lo cual sugiere un desplazamiento de la atención mediática hacia otros asuntos. Sin embargo, hay algunos meses, como septiembre y noviembre de 2022 y enero de 2023, en los que el tema vuelve a ganar cierta importancia. Esto podría deberse a novedades en el conflicto o en las sanciones aplicadas a causa del mismo.

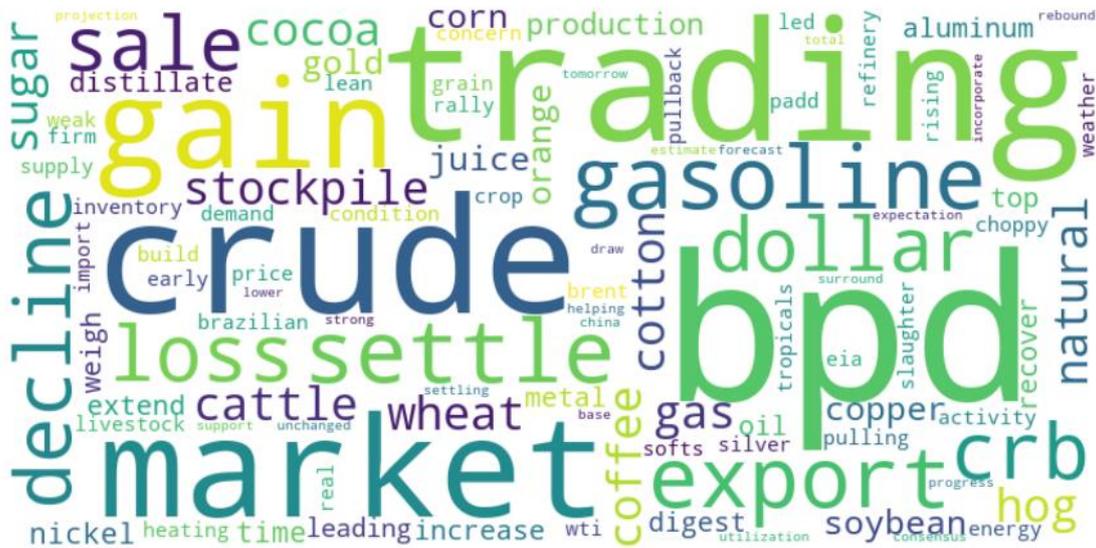


**Figura 15: Evolución del topic 1 en el tiempo**

*Fuente: Elaboración propia*

- **Topic 2 - Comercio global de petróleo y gasolina:** Este *topic* se enfoca en el comercio global de petróleo y gasolina, incluyendo temas como la producción, exportación e importación, así como los precios y las fluctuaciones del mercado.

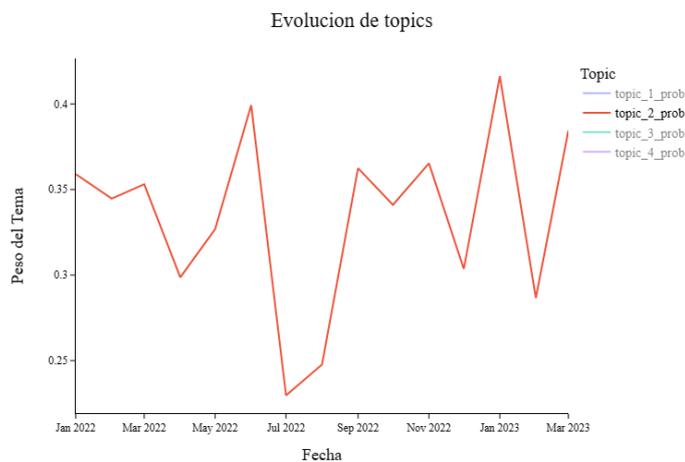
## Nube de palabras para el Topic 2



**Figura 16: Nube de palabras para el topic 2**

*Fuente: Elaboración propia*

En la figura 17 se observa la evolución del *topic 2* en el tiempo. En el gráfico se puede apreciar cómo este tema no sigue una tendencia concreta a lo largo del tiempo, sino más bien una tendencia irregular. Sin embargo, lo que si se puede observar es que durante los meses de julio y agosto de 2022 el tema tuvo una importancia menor. Esto puede deberse a que en muchas partes del mundo estos meses corresponden al verano, una época en la que la demanda de energía para la calefacción disminuye significativamente, lo que a su vez afecta a la demanda de petróleo y gasolina.



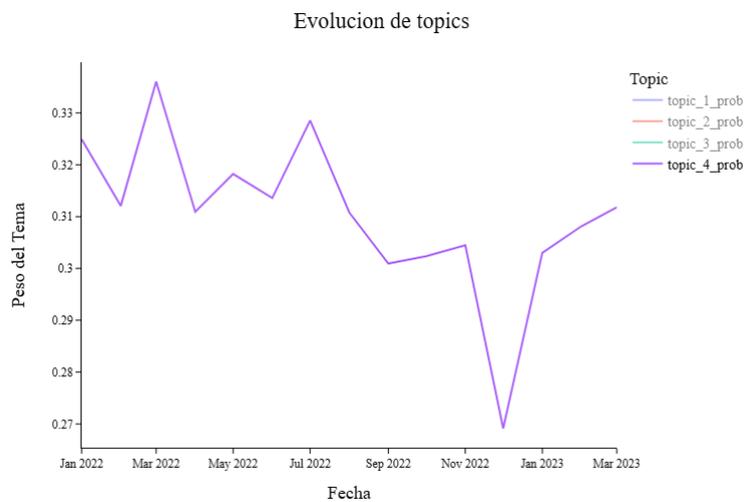
**Figura 17: Evolución del topic 2 en el tiempo**

*Fuente: Elaboración propia*





En la figura 21 se observa la evolución del *topic* 4 en el tiempo. En el gráfico se puede observar como el *topic* ha ido perdiendo peso durante el último año. Esto indica que durante los primeros meses del año la temática Rusia – China – Petróleo estaba muy presente, lo que podría deberse a las sanciones impuestas por la UE que incitaron a Rusia a forjar nuevas alianzas comerciales con otros países como China e India. Sin embargo, la tendencia a la baja que se aprecia a medida que avanza el año (y en especial durante el mes de diciembre de 2022) puede deberse a la falta de novedades significativas sobre el tema.



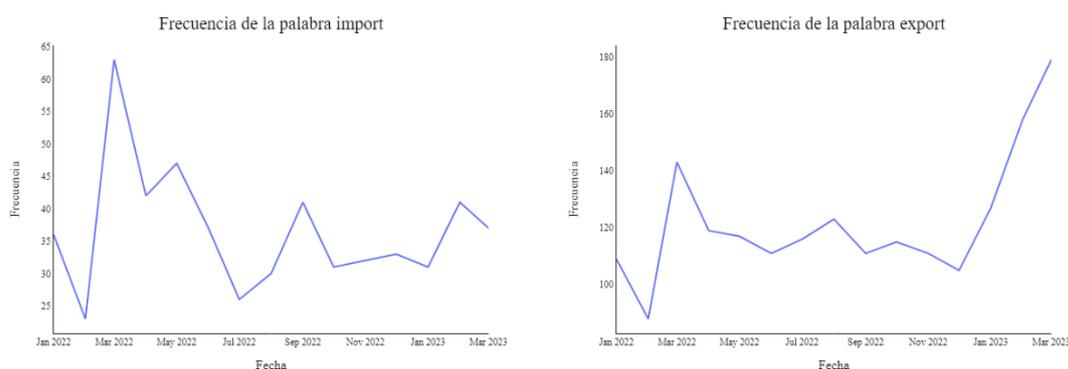
**Figura 21: Evolución del topic 4 en el tiempo**

*Fuente: Elaboración propia*

## 5.5. ANÁLISIS ENFOCADO EN EL CONFLICTO DE RUSIA Y UCRANIA

En el apartado anterior se ha realizado un análisis sobre los principales temas presentes en las noticias financieras durante el último año y se ha observado la presencia del tema del conflicto entre Ucrania y Rusia de manera subyacente en las noticias financieras. Por ello, a continuación, se va a realizar un análisis para estudiar si la Guerra de Ucrania ha tenido un impacto significativo en determinados aspectos comerciales y productos y, en caso de haberlo, comprender las razones tras él.

En primer lugar, se ha llevado a cabo un análisis de la frecuencia de los términos “import” y “export” en el *corpus*. En la figura que se presenta a continuación, se puede observar que, tras el inicio de la guerra el 24 de febrero de 2022, el discurso financiero ha prestado una mayor atención a las importaciones y exportaciones. Asimismo, es notable el incremento significativo en la frecuencia del término "export" durante los meses de febrero y marzo de 2023.



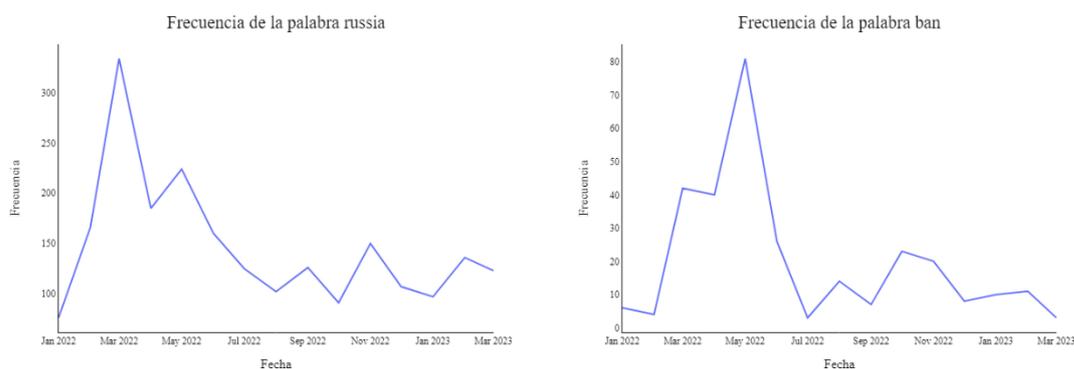
**Figura 22: Evolución de la frecuencia de los términos "import" y "export" en el tiempo**

*Fuente: Elaboración propia*

Por un lado, el aumento del uso de estos términos durante las semanas posteriores al inicio de la guerra podría deberse a las sanciones impuestas por la UE a Rusia. Como se ha mencionado anteriormente, entre estas sanciones se incluían sanciones comerciales, la exclusión financiera mediante la expulsión de algunos bancos rusos del sistema SWIFT, la congelación de gran parte de los activos del banco central ruso en el exterior (activos que el gobierno ruso pretendía utilizar para paliar el impacto económico de las sanciones), la suspensión del gasoducto Nordstream 2 por parte de Alemania, prohibiciones de

exportaciones e importaciones de determinados productos a Rusia, el cierre del espacio aéreo a las aerolíneas rusas, etc. (Otero-Iglesias y Steinberg, 2022).

Por ello, a continuación, se va a analizar si la hipótesis de la relación entre las palabras “import” y “export” con las sanciones impuestas por la UE a Rusia es correcta. En los siguientes gráficos podemos observar cómo sí existe una correlación entre los términos “import” y “export” y los términos “russia” y “ban” durante las semanas posteriores al inicio de la guerra.



**Figura 23: Evolución de la frecuencia de los términos "russia" y "ban" en el tiempo**

*Fuente: Elaboración propia*

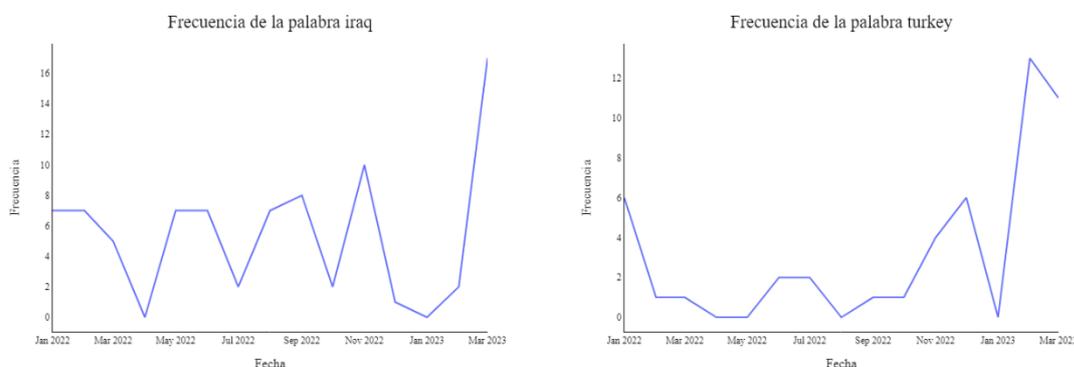
Asimismo, es interesante analizar la relación entre los términos “ban” y “russia”. En la figura 23, se puede observar cómo durante los meses de enero y febrero de 2022 apenas se utilizaba la palabra "ban". Sin embargo, en marzo de 2022, tras el estallido de la guerra, comenzó a ganar importancia. No obstante, fue en mayo de 2022 cuando esta palabra alcanzó su punto máximo. Este incremento significativo puede atribuirse al anuncio de sanciones relacionadas con el petróleo, que fueron las más relevantes hasta ese momento. Además, es importante tener en cuenta que en nuestra muestra de noticias se otorga una gran importancia al petróleo, por lo tanto, es lógico que las sanciones impuestas por la UE y el G7 a las importaciones de petróleo tuviesen un fuerte impacto en el *corpus*.

Por otro lado, se va a analizar el motivo detrás del aumento de la frecuencia del término “export” en los meses de febrero y marzo de 2023. Una de las razones podría ser que, durante los primeros meses del año 2023, a pesar de las sanciones, las exportaciones de diésel de Rusia alcanzaron niveles cercanos al récord. Durante los primeros 19 días de

marzo Rusia exportó alrededor de 1,5 millones de barriles por día de diésel. Una de las razones de este aumento se debe a que países como Turquía y Marruecos aumentaron sus compras de diésel ruso durante este último mes (Wittels y Prem, 2023).

Otra causa podría ser el conflicto entre Turquía e Irak que se produjo durante el mes de marzo de 2023. En marzo de 2023 el gobierno federal de Iraq ganó un caso de arbitraje en el Tribunal Internacional de Arbitraje, en el que llevaba inmerso desde hacía nueve años, en relación con el flujo de petróleo kurdo a través de Turquía. El tribunal acordó que el procesamiento del petróleo kurdo por parte de Turquía a través del puerto de Ceyhan violaba un acuerdo de 1973 que requería el consentimiento de Bagdad para las exportaciones. En consecuencia, Turquía tendrá que pagar a Iraq alrededor de 1.5 mil millones de dólares en concepto de daños, y la Organización Estatal de Comercialización de Petróleo de Iraq (SOMO, por sus siglas en inglés) afirmó que será la única entidad encargada de gestionar las exportaciones cargadas en Ceyhan a partir de ahora (Jalabi y Samson, 2023).

De hecho, como se puede observar en los siguientes gráficos, es muy probable que está última hipótesis sea plausible. Ya que durante esos meses se observa un mayor uso de los términos “iraq” y “turkey”.

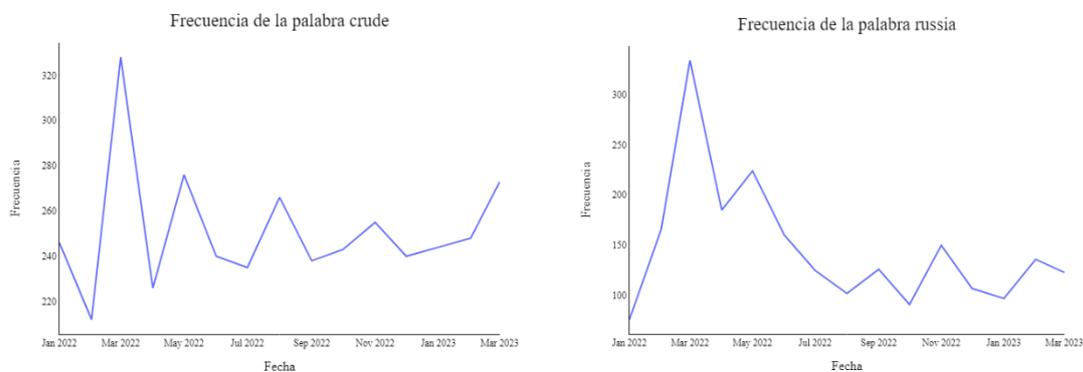


**Figura 24: Evolución de la frecuencia de los términos "iraq" y "turkey" en el tiempo**

*Fuente: Elaboración propia*

En segundo lugar, se va a analizar uno de los temas con más peso en nuestra muestra de noticias financieras: el petróleo. Por ello, se va a realizar un análisis comparativo de la

frecuencia de la palabras “crude” y “russia” para estudiar el impacto del conflicto ruso en dicho término.

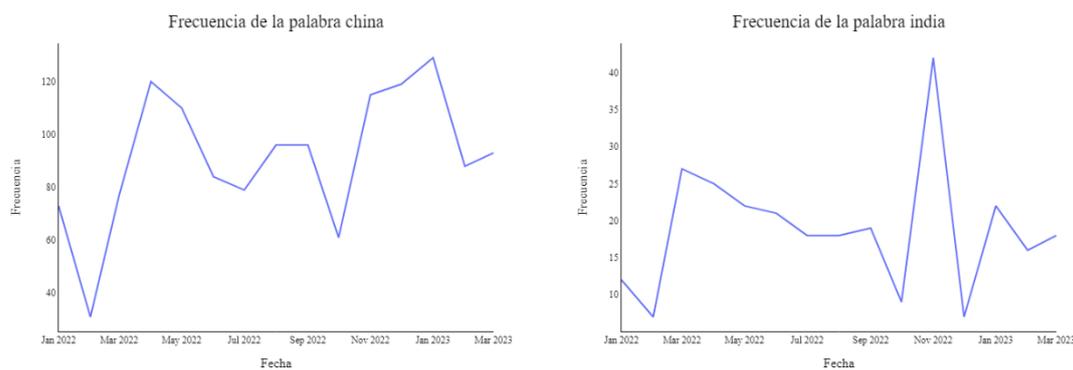


**Figura 25: Evolución de la frecuencia de los términos "crude" y "russia" en el tiempo**

*Fuente: Elaboración propia*

En este análisis se puede observar de forma muy clara como la hipótesis de la relación entre el petróleo y la Guerra de Ucrania es correcta. Como se ha mencionado anteriormente, desde la invasión de Ucrania por parte de Rusia en febrero de 2022, varios países han tomado medidas para restringir o poner fin a las importaciones de petróleo con el objetivo de debilitar el esfuerzo bélico de Rusia e incrementar su autonomía energética. En mayo de 2022, la UE y los países del G7 anunciaron las sanciones más agresivas hasta la fecha en relación a las importaciones de petróleo, que se comenzarían a aplicar en diciembre de 2022. Esto podría explicar la subida que se puede apreciar en la figura 25 en el mes de mayo y diciembre de 2022.

Estas sanciones dieron lugar a un cambio en los principales importadores de petróleo ruso, convirtiéndose China e India en los principales receptores. En la figura que se muestra a continuación, se puede observar cómo los términos “china” e “india” ganaron bastante popularidad en los meses de noviembre y diciembre de 2022.



**Figura 26: Evolución de la frecuencia de los términos "china" e "india" en el tiempo**

*Fuente: Elaboración propia*

En tercer lugar, se ha llevado a cabo un análisis del término “natural gas”. En la figura 27 se pueden observar tres anomalías durante: (i) los meses correspondientes al inicio de la guerra, (ii) los meses de junio-septiembre de 2022 y (iii) el mes de marzo de 2023.



**Figura 27: Evolución de la frecuencia del término "natural gas" en el tiempo**

*Fuente: Elaboración propia a partir de noticias de FactSet*

Por un lado, es muy probable que la primera anomalía se deba al aumento del discurso durante el inicio de la guerra sobre la dependencia energética de Europa en Rusia y el temor a una crisis energética. Por otro lado, la subida que experimentó este término durante los meses de verano puede deberse a la suspensión del gasoducto Nord Stream 1 que, aunque se produjo de manera definitiva en el mes de agosto, ya había ido reduciendo su capacidad durante los meses de junio y julio. Finalmente, no se ha encontrado ninguna noticia o evento relevante que justifique el incremento del uso del término durante el mes

de marzo de 2023, por lo que es posible que este aumento se deba a un incremento en el precio de esta *commodity* y no a un evento concreto (Expansión, 2023).

En cuarto lugar, se va a realizar un análisis de la palabra “grain” en el texto. En la figura 28 se pueden apreciar tres anomalías. En primer lugar, se puede observar un incremento de la frecuencia del término durante los meses del inicio de la guerra, probablemente debido a que Rusia y Ucrania son dos de los mayores exportadores de cereales y trigo de Europa (Otero-Iglesias y Steinberg, 2022). En segundo lugar, se ha detectado una mayor frecuencia del término “grain” durante los meses de verano. Esto puede explicarse por el hecho de que durante esta temporada se lleva a cabo la cosecha del grano, lo que hace que aumente su importancia en el mercado y, por lo tanto, en las noticias. En tercer lugar, se ha observado un gran aumento del uso del término durante el mes de marzo de 2023. Esto podría deberse a un aumento en la preocupación por las exportaciones de grano debido a la amenaza de Rusia de retirar su participación en la Iniciativa de Cereal del Mar Negro que se produjo durante ese mes.



**Figura 28: Evolución de la frecuencia del término "grain" en el tiempo**

*Fuente: Elaboración propia*

## 5.6. SENTIMENT ANALYSIS

El *sentiment analysis* o análisis de sentimiento es una técnica fundamental en el campo del procesamiento del lenguaje natural y tiene como objetivo comprender y evaluar las actitudes, emociones y opiniones expresadas en textos escritos. El análisis de sentimiento tiene diversas aplicaciones en diferentes áreas. En el ámbito empresarial, las empresas pueden utilizarlo para analizar la satisfacción del cliente a través de comentarios o reseñas en línea, realizar estudios de mercado y seguimiento de la marca. En el ámbito de las redes sociales, puede ayudar a comprender la opinión pública sobre temas específicos o evaluar la respuesta de los usuarios a una campaña o producto. También se utiliza en el campo de la investigación académica para analizar grandes cantidades de datos textuales y extraer conclusiones.

En el ámbito financiero el análisis de sentimiento busca identificar y medir el tono subyacente de los textos relacionados con el mercado, como noticias económicas, informes de empresas, publicaciones en redes sociales y comentarios de analistas. Al determinar si un texto tiene un sentimiento positivo, negativo o neutral, se pueden obtener perspectivas valiosas sobre cómo se percibe una empresa, una acción o una situación económica en particular (Mishev et al, 2020; Day y Lee, 2016)

En el ámbito financiero, el análisis de sentimiento se puede llevar a cabo utilizando diferentes enfoques. Uno de los enfoques más básicos es el uso de diccionarios de sentimiento financiero. Estos diccionarios están formados por conjuntos de palabras y términos específicos relacionados con cuestiones financieras que están asociadas con una polaridad positiva o negativa. Al asignar puntuaciones de sentimiento a las palabras en un texto, se puede calcular una puntuación global del sentimiento para evaluar si el contenido tiene un sentimiento positivo, negativo o neutro (Mishev et al, 2020).

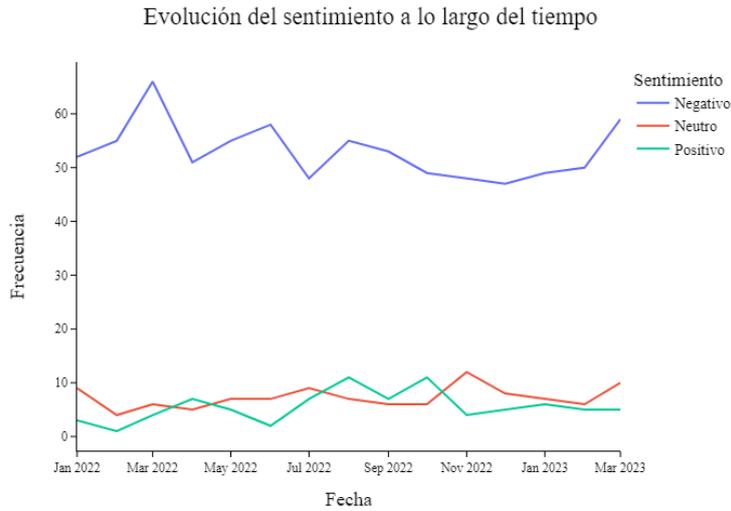
Además de los diccionarios, se han desarrollado modelos más complejos para el análisis de sentimiento financiero. Por ejemplo, Day y Lee (2016) propusieron un modelo de análisis de sentimiento basado en redes neuronales recurrentes (RNN) para identificar la polaridad de los tweets relacionados con las acciones. Para ello, utilizaron técnicas de preprocesamiento de texto y entrenaron un modelo para clasificar los tweets en categorías de sentimiento. Por otro lado, Mishev et al. (2020) propusieron un enfoque basado en el uso de redes neuronales convolucionales (CNN) para el análisis de sentimiento en noticias

financieras. Su modelo se entrenó utilizando un conjunto de datos etiquetados y filtros convolucionales para capturar características relevantes en el texto y realizar la clasificación de sentimiento.

En cuanto al propósito de este estudio, se ha decidido utilizar el método de diccionarios debido a su simplicidad y facilidad de implementación. Aunque los modelos complejos pueden ofrecer un rendimiento superior en algunas situaciones, los diccionarios financieros pueden proporcionar una aproximación inicial útil y un punto de partida para el análisis de sentimiento financiero. Además, los diccionarios financieros pueden ser más fáciles de interpretar y permiten una mayor transparencia en los resultados obtenidos.

Para nuestro estudio se va a emplear el diccionario financiero "Loughran-McDonald" que contiene términos financieros y palabras asociadas con una polaridad positiva o negativa en el contexto financiero (Day y Lee, 2016; Loughran y McDonald, 2011). Este diccionario fue elaborado en 2011 por Loughran y McDonald y se desarrolló con el propósito de analizar el contenido emocional de los informes anuales de empresas cotizadas en bolsa. El diccionario se construyó mediante un proceso de etiquetado manual, donde los investigadores asignaron a cada palabra o frase del vocabulario financiero una etiqueta que reflejaba su carga emocional. Para ello, utilizaron una muestra de informes anuales y clasificaron las palabras en tres categorías: positivas, negativas o neutras, según su efecto en el sentimiento (Loughran y McDonald, 2011).

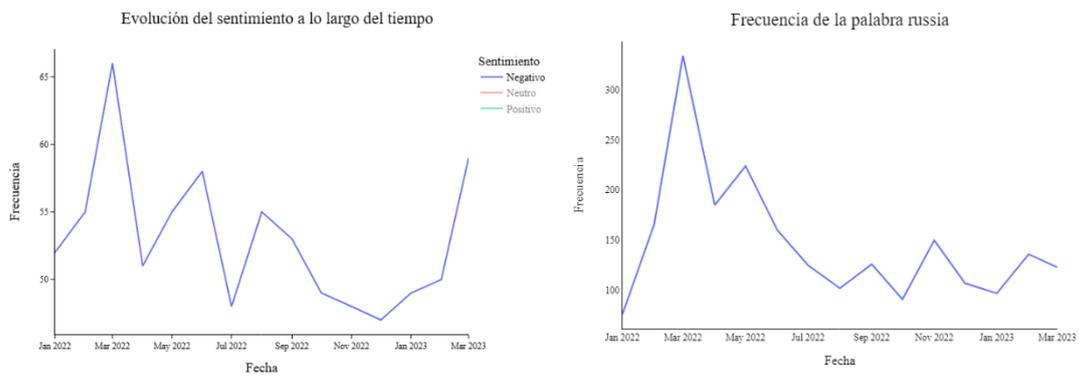
Una vez instalado el diccionario, se aplicará al *corpus*. Este asigna una puntuación a cada palabra o frase según su polaridad, es decir, si tiene una connotación positiva o negativa en el contexto financiero. Estas puntuaciones pueden variar entre -1 y 1, donde los valores negativos indican una connotación negativa, los valores positivos indican una connotación positiva y los valores cercanos a cero indican neutralidad. Luego, se suman las puntuaciones de todas las palabras y frases encontradas en el texto para obtener una puntuación global que representa el sentimiento general de la noticia. Finalmente, se realiza una clasificación en la que los valores menores que 0 serán considerados sentimiento negativo, los valores mayores que 0 sentimiento positivo y los valores iguales a 0 neutros. Una vez realizada esta clasificación se han obtenido los siguientes resultados:



**Figura 29: Evolución del sentimiento en el tiempo (I)**

*Fuente: Elaboración propia*

En la figura 29 se puede observar un predominio del sentimiento negativo a lo largo del tiempo. Esto podría indicar una percepción generalizada de incertidumbre o preocupación en el mercado financiero. Además, es interesante destacar que, al examinar el gráfico correspondiente al sentimiento negativo, se observa una correlación con la frecuencia de la palabra "russia". Esto sugiere que el diccionario está capturando de manera precisa el sentimiento negativo relacionado con el conflicto entre Rusia y Ucrania.



**Figura 30: Evolución del sentimiento negativo y de la frecuencia del término "russia" en el tiempo**

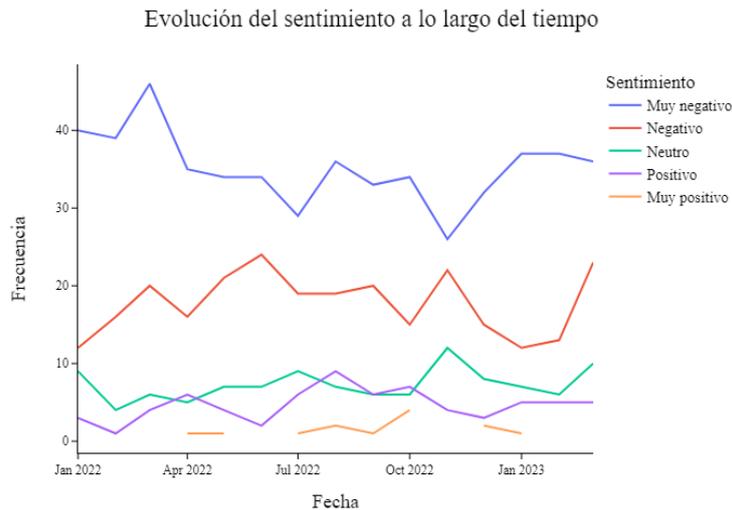
*Fuente: Elaboración propia a partir de noticias de FactSet*

A continuación, se va a realizar una clasificación más segmentada para identificar la intensidad de los sentimientos. Para ello se van a clasificar los textos de la siguiente manera:

**Tabla 3: Clasificación de sentimientos más segmentada**

Sentimiento	Puntuación
Muy negativo	menor que -0,5
Negativo	entre -0,5 y 0 (no incluido)
Neutro	igual a 0
Muy positivo	entre 0 (no incluido) y 0,5
Positivo	mayor que 0,5

*Fuente: Elaboración propia*



**Figura 31: Evolución del sentimiento en el tiempo (II)**

*Fuente: Elaboración propia a partir de noticias de FactSet*

En la figura 31 se puede apreciar cómo el sentimiento muy negativo es el predominante, especialmente durante los meses en los que se menciona la Guerra de Ucrania. Esto indica que la cobertura de noticias relacionadas con la guerra tiene un impacto significativo en el tono general de las noticias financieras. Por otro lado, el sentimiento positivo y muy positivo es escaso a lo largo del período analizado, lo que sugiere que existe una falta de eventos o noticias que generen optimismo en el contexto financiero.

## 6. CONCLUSIONES

En este estudio se ha podido comprobar como la invasión de Ucrania por parte de Rusia ha tenido un impacto significativo en las noticias financieras. Este conflicto, que se venía gestando desde 2014 con la anexión de Crimea por parte de Rusia, alcanzó su punto álgido el 24 de febrero de 2022 cuando Putin invadió Rusia bajo el pretexto de una “operación militar especial” para proteger a la población pro-rusa tras las pretensiones de Ucrania de entrar en la OTAN. Sin embargo, la verdadera razón detrás de esta decisión es mucho más antigua y está arraigada en cuestiones históricas que se remontan a la disolución de la URSS, de la cual Ucrania formaba parte, y al sentimiento imperialista de Rusia.

El objetivo principal de este estudio era validar la hipótesis del impacto de la Guerra de Ucrania en las noticias financieras mediante un proceso de *text mining*. Para ello, en primer lugar, se ha creado un *corpus* mediante la extracción de la información contenida en una muestra de noticias financieras procedentes de FactSet (herramienta de análisis de información financiera). En segundo lugar, se ha ejecutado un preprocesamiento de la información en el que se ha realizado una conversión del texto a minúsculas, tokenización, eliminación de *stopwords* y lematización. En tercer lugar, se ha llevado a cabo un proceso de *topic análisis*, utilizando el algoritmo LDA para extraer los temas principales tratados en las noticias financieras y su evolución en el periodo analizado.

En este análisis se han identificado cuatro temas y se ha podido observar que, en el último año, el comercio del petróleo ha sido el tema central de las noticias de FactSet, atrayendo un constante interés tanto de los medios de comunicación como de los usuarios. Sin embargo, se ha observado también que el conflicto entre Rusia y Ucrania ha desempeñado un papel subyacente en el análisis y la discusión de los mercados globales. Esto se evidencia especialmente en los meses de febrero y marzo de 2022, coincidiendo con el inicio del conflicto y generando una mayor atención mediática en torno a este tema.

A continuación, se van a explicar las principales conclusiones a las que se ha llegado en el análisis de los *topics*. El primer *topic* se centra en el conflicto y las disputas territoriales entre Rusia y Ucrania. En su evolución se puede observar cómo durante los meses de enero y febrero de 2022 ya se anticipaba un conflicto territorial entre Rusia y Ucrania. Sin embargo, a medida que avanzaba el tiempo se produjo una disminución de

la relevancia mediática, aunque con algunos repuntes en meses posteriores, como septiembre y noviembre de 2022 y enero de 2023, posiblemente debido a novedades en el conflicto o en las sanciones aplicadas por países de la UE o del G7.

El segundo *topic* aborda el comercio global de petróleo y gasolina. A diferencia del tema anterior, este tema no sigue una tendencia clara a lo largo del tiempo, sino que muestra fluctuaciones irregulares. Sin embargo, se ha observado que, durante los meses de julio y agosto, el tema tuvo una importancia menor, posiblemente debido a la disminución de la demanda de energía en verano, lo que afecta la demanda de petróleo.

El tercer *topic* se enfoca en el comercio de petróleo y productos agrícolas de Estados Unidos. Al compararlo con el tema anterior, se ha encontrado un comportamiento opuesto, lo que sugiere que cuando el enfoque de las noticias no está en el petróleo a nivel global, se centra en el petróleo de Estados Unidos y el comercio agrícola. También se ha observado como, durante los meses de verano, este tema ganó una mayor importancia. Esto puede deberse a las importantes cosechas agrícolas que tienen lugar en Estados Unidos durante los meses de verano, como el maíz y el algodón, lo que genera una mayor discusión y atención en torno al comercio agrícola y las fluctuaciones de precios.

El último *topic* examina la industria petrolífera de Rusia, su relación con los mercados globales y las alianzas comerciales establecidas con otros países, en especial con China. Se ha observado también un gran enfoque en el tema Rusia-China-Petróleo durante los primeros meses del año 2022. Sin embargo, a medida que fue avanzando el año este tema fue perdiendo relevancia, posiblemente debido a la falta de novedades significativas en la relación entre Rusia y China.

En cuarto lugar, se ha llevado a cabo un análisis enfocado en el conflicto para estudiar el impacto de la guerra en determinadas *commodities* y otras áreas. En este análisis se ha llegado a las siguientes conclusiones:

- **Impacto en las importaciones y exportaciones:** Por un lado, el análisis de la frecuencia de los términos "import" y "export" en el *corpus* revela que durante las semanas posteriores al inicio del conflicto se produjo un aumento significativo en la importancia de estos términos. Esto puede estar relacionado con las sanciones impuestas por la UE a Rusia, que incluían medidas comerciales y económicas que afectaban a las importaciones y exportaciones rusas. Por otro lado, se ha podido

observar un aumento en la frecuencia de la palabra “export” durante los primeros meses del año 2023. Esto puede deberse a dos razones: (i) las exportaciones de diésel ruso estaban alcanzando niveles récord (principalmente por países como Turquía y Marruecos) y (ii) el conflicto entre Turquía e Irak en relación con el arbitraje por el flujo de petróleo kurdo a través de Turquía.

- **Impacto en el petróleo:** En este análisis se ha podido observar como el estallido del conflicto tuvo un impacto significativo en el aumento del discurso del petróleo. Esto se debe a que, tras el inicio de la guerra, países de la UE y del G7 implementaron duras sanciones a las exportaciones de petróleo ruso. Asimismo, se ha podido observar como las noticias financieras reflejan el cambio que se produjo en los principales importadores de petróleo ruso, convirtiéndose China e India en los principales compradores.
- **Impacto en el gas natural:** Al analizar el impacto del conflicto en la frecuencia del término “natural gas” se han detectado tres anomalías en las que la frecuencia del término aumenta. La primera de ellas corresponde con el inicio de la guerra, lo cual puede deberse al aumento en la preocupación por una crisis energética debido a la dependencia de Europa en Rusia. La segunda se produjo durante los meses de verano (junio-septiembre) y podría deberse a la suspensión del gasoducto ruso Nord Stream 1 como represalia a la UE por las sanciones impuestas. Finalmente, no se ha encontrado un evento concreto que justifique la anomalía producida en marzo de 2023.
- **Impacto en el grano:** En cuarto y último lugar se ha analizado la evolución de la frecuencia de la palabra “grain”. En el análisis se ha detectado un mayor uso de este término en tres momentos: al inicio de la guerra, durante los meses de verano y en el mes de marzo de 2023. Por un lado, el aumento de la frecuencia durante el inicio de la guerra podría deberse a que Rusia y Ucrania son dos de los mayores exportadores de trigo en Europa. Por otro lado, la subida producida durante los meses de verano podría deberse a que durante este periodo se lleva a cabo la cosecha del grano. Finalmente, la abrupta subida producida en marzo de 2023 podría deberse a la amenaza de Rusia de retirar su participación en la Iniciativa de

Cereal del Mar Negro, iniciativa que facilitaba la exportación del grano y otros productos alimentarios de Ucrania.

En resumen, este análisis ha demostrado como las noticias financieras han captado el impacto de la guerra en las importaciones y exportaciones y en *commodities* como el petróleo, el grano y el gas natural.

Finalmente, se ha realizado un análisis de sentimiento, con el objetivo de identificar si se había producido un cambio en el sentimiento financiero a raíz de la Guerra de Ucrania. Para ello, se ha utilizado el diccionario financiero “Loughran-McDonald” y se ha identificado un sentimiento negativo predominante a lo largo del tiempo y, en especial, durante los meses del inicio del conflicto. Asimismo, se ha realizado una clasificación más segmentada para identificar la intensidad de los sentimientos, donde se observa que el sentimiento muy negativo es predominante, especialmente durante los meses en los que se menciona la Guerra de Ucrania, lo cual resalta el impacto significativo que la cobertura de noticias relacionadas con la guerra ha tenido en el tono general de las noticias financieras.

## 7. LÍMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

El presente trabajo de investigación también enfrenta algunas limitaciones que deben tenerse en cuenta al interpretar los resultados y conclusiones obtenidas. En primer lugar, el periodo de la muestra utilizado abarca solamente un mes antes del estallido de la Guerra de Ucrania hasta marzo de 2023. Esta limitación implica que no se puede realizar una comparación en base año-año de los temas tratados en las noticias financieras en un contexto de "antes y después" de la guerra. Por lo tanto, es importante tener en cuenta que los resultados ofrecen una comparación limitada ya que, existe la posibilidad que durante los meses previos al estallido del conflicto el discurso de las noticias financieras ya estuviese teniendo en cuenta el posible conflicto.

Otra limitación se relaciona con la estructura uniforme de las noticias financieras de la muestra. Al provenir de una base de datos financiera, estas noticias presentan estructuras y epígrafes uniformes. Esto puede dar lugar a la repetición de los mismos temas en varias noticias, no necesariamente porque hayan ganado importancia, sino debido al enfoque habitual de las noticias financieras. Por lo tanto, es importante considerar que la presencia de ciertos temas puede estar influenciada por la naturaleza de la fuente y no necesariamente por el impacto directo de la guerra.

Además, la naturaleza resumida de las noticias financieras también supone una limitación. Las noticias de la muestra ofrecen información sintetizada, lo que limita la cantidad de información disponible para el análisis. Dado que las noticias se centran en proporcionar una visión general de los eventos económicos y financieros, es posible que no se aborden en detalle ciertos aspectos relacionados con la Guerra de Ucrania y su impacto en la economía. Esto puede afectar a la profundidad del análisis y la capacidad de captar todos los matices del impacto de la guerra en las noticias financieras.

Estas limitaciones deben tenerse en cuenta al interpretar los resultados y conclusiones de este estudio. Aunque se han realizado esfuerzos para abordar estas limitaciones y obtener conclusiones significativas, es importante reconocer que existen ciertas restricciones que pueden afectar a nuestras conclusiones.

Sin embargo, estas limitaciones abren también paso a futuras líneas de investigación como las que se mencionan a continuación:

- **Realizar un análisis del conflicto con un periodo temporal más extenso:** Sería interesante ampliar el periodo de estudio para comprender mejor la evolución y el impacto del conflicto en Ucrania. Por un lado, se podría investigar desde 1991, fecha en la que se produjo la disolución de la Unión Soviética y la consecuente independencia de Ucrania, o desde 2014, año en el que se produjo la anexión de Crimea por parte de Rusia. Estudiar el conflicto en un contexto más amplio permitiría analizar la evolución de la frecuencia de términos muy ligados al conflicto (como el petróleo, el gas natural y el grano) en un horizonte con muchos más eventos e identificar otros aspectos económicos que se han visto afectados por este conflicto histórico.
- **Extracción de cifras numéricas para analizar el cambio en las previsiones de los precios de las *commodities* que se han visto impactadas por la guerra:** Otra línea de investigación interesante consistiría en analizar la evolución de las previsiones de los precios de las *commodities* que se han visto afectadas por la Guerra de Ucrania. Esto implicaría recopilar las previsiones numéricas presentes en las noticias sobre los precios futuros de las *commodities* clave, proporcionando así una visión más detallada sobre cómo la incertidumbre ha afectado a las perspectivas económicas. Asimismo, este análisis permitiría evaluar la precisión de las previsiones de los analistas y analizar cómo las noticias financieras reflejan y reaccionan a los cambios en las expectativas de los precios.
- **Realizar el análisis desde otras perspectivas, como la social o política:** Sería interesante explorar el impacto de la Guerra de Ucrania en las noticias desde otras perspectivas complementarias, como la social o política. Desde una perspectiva social, se podría investigar cómo la guerra ha afectado a la población ucraniana en términos de empleo, bienestar y desplazamiento, y cómo estas condiciones socioeconómicas se reflejan en las noticias. Desde una perspectiva política, se podría investigar cómo la guerra ha influido en las relaciones internacionales, las políticas gubernamentales y las decisiones de inversión en diferentes países. De esta forma se proporcionaría una comprensión más completa de los impactos multidimensionales de la guerra en diversos aspectos de la sociedad y la política y sus efectos en los medios de comunicación.

## 8. BIBLIOGRAFÍA

- Barriocanal, J. (2023, febrero 21). El gráfico que revela cuánto petróleo está comprando China a Rusia y el descuento que ofrece Moscú. *El Economista*.
- BBC. (2022, septiembre 29). Nord Stream 1: How Russia is cutting gas supplies to Europe. *BBC News*.
- Bebler, A. (2015). Crimea and the Russian-Ukrainian Conflict. *Romanian Journal of European Affairs*, 15(2).
- Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Chang, J., Boyd-graber, J., Gerrish, S., Wang, C., y Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, 22, 1-9.
- Cúneo, M. Á. (2018a). El futuro de Ucrania: Entre el conflicto de Donbas, la pérdida de Crimea y los Acuerdos de Minsk (Parte I). *CUPEA Cuadernos de Política Exterior Argentina*, 127, 103-107. <https://doi.org/10.35305/cc.vi127.39>
- Cúneo, M. Á. (2018b). El futuro de Ucrania: Entre el conflicto de Donbas, la pérdida de Crimea y los Acuerdos de Minsk (Parte II). *CUPEA Cuadernos de Política Exterior Argentina*, 128, 87-93.
- Day, M.-Y., y Lee, C.-C. (2016). Deep learning for financial sentiment analysis on finance news providers. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1127-1134. <https://doi.org/10.1109/ASONAM.2016.7752381>
- de la Corte Ibáñez, L. (2023). Europa y la guerra de Ucrania. *Global Affairs Journal*, 5, 22-29.
- Dorado Díaz, S. (2020). El conflicto ucraniano. Orígenes y causas de la última guerra en Europa. *Gladius et Scientia. Revista de Seguridad del CESEG*, 2, 1-13. <https://doi.org/10.15304>
- El Economista. (2023, marzo 29). Rusia multiplica por 22 la venta de petróleo a la India a la vez que acuerda con China el uso del yuan frente al dólar. *El Economista*.

- Epdata. (2023). *Mapas y gráficos de la situación de la guerra en Ucrania, última hora*. <https://www.epdata.es/datos/guerra-rusia-ucrania-datos-graficos/646>
- Esteban, M. (2023). ¿Cómo afecta la agresión rusa al triángulo UE-EEUU-China? En *La guerra en Ucrania un año después. Impacto global, europeo y español* (pp. 29-33). Real Instituto Elcano.
- Expansión. (2023, abril 1). El tope al gas ha permitido reducir el precio de la luz un 17% desde junio. *Expansión*.
- García Encina, C. (2023). La guerra y el liderazgo estadounidense. En *La guerra en Ucrania un año después. Impacto global, europeo y español* (pp. 23-27). Real Instituto Elcano.
- Griffiths, T. L., y Steyvers, M. (2004). Finding scientific topics. *PNAS*, *101*(1), 5228-5235. <https://doi.org/10.1073>
- Gupta, V., y Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76. <https://doi.org/10.4304/jetwi.1.1.60-76>
- Ilas, A. (2023). *One year on, who is funding Russia's war in Ukraine?* [Centre for Research on Energy and Clean Air].
- Jalabi, R., y Samson, A. (2023, marzo 25). Iraq wins landmark case against Turkey over Kurdish oil exports. *Financial Times*.
- Kuchling, A. M. (2023). *Expresiones regulares COMOS (HOWTO)*. Python Documentation. Recuperado 20 de mayo de 2023, de <https://docs.python.org/3/howto/regex.html>
- Loughran, T., y McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Mendoza, Á. M. (2022). Ucrania, tablero de una guerra fría no concluida. *Trinchera Cultural*, 1-9.
- Merino, G. E. (2022). La guerra en Ucrania, un conflicto mundial. *Revista Estado y Políticas Públicas*, *19*, 113-140.

- Milosevich-Juaristi, M. (2023). Ucrania en guerra. En *La guerra en Ucrania un año después. Impacto global, europeo y español* (pp. 11-15). Real Instituto Elcano.
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., y Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, 8, 131662-131682. <https://doi.org/10.1109/ACCESS.2020.3009626>
- Moloney, M. (2023, mayo 12). Ukraine war: WFP chief Cindy McCain says grain deal with Russia needed to feed world. *BBC News*.
- Myllyvirta, L. (2022b). *EU ban on Russian oil: Why it matters and what's next*. 1-27.
- Myllyvirta, L. (2022a). *Financing Putin's war: Fossil fuel exports from Russia in the first six months of the invasion of Ukraine* (Centre for Research on Energy and Clean Air).
- NATO. (2023a). *Enlargement and Article 10*. NATO. [https://www.nato.int/cps/en/natohq/topics\\_49212.htm](https://www.nato.int/cps/en/natohq/topics_49212.htm)
- NATO. (2023b). *Finland joins NATO as 31st Ally*. NATO. [https://www.nato.int/cps/en/natohq/news\\_213448.htm](https://www.nato.int/cps/en/natohq/news_213448.htm)
- NATO. (2023c). *Relations with the Republic of North Macedonia (Archived)*. NATO. [https://www.nato.int/cps/en/natohq/topics\\_48830.htm](https://www.nato.int/cps/en/natohq/topics_48830.htm)
- Otero-Iglesias, M., y Steinberg, F. (2022). La guerra de Ucrania en el tablero económico. *Comentario Elcano* 7/2022, 1-3.
- Prokopenko, A. (2023, marzo 17). How the Ukraine Grain Deal Went From Boon to Burden for the Kremlin. *Carnegie*.
- Sandri, P. M. (2022, septiembre 3). Rusia ejecuta su amenaza y corta el gas hacia Europa. *La Vanguardia*.
- UNCTAD. (2023, marzo 19). Black Sea Grain Initiative extended. *UNCTAD*.
- Vakulenko, S. (2022, septiembre 7). Shutting Down Nord Stream Marks the Point of No Return for Russian Gas. *Carnegie*.
- Wittels, J., y Prem, P. (2023, marzo 27). Russia's Diesel Exports Heading for Record Despite EU Sanctions. *Bloomberg*.

## 9. ANEXOS

```
# Librerías
import PyPDF2
from PyPDF2 import PdfReader
import re
from datetime import datetime
from pathlib import Path
import pandas as pd
import numpy as np
import nltk
nltk.download('punkt')
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer
import spacy
from gensim import corpora, models
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
from stopwordsiso import stopwords
from nltk.corpus import wordnet
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.exceptions import NotFittedError
from sklearn.preprocessing import MinMaxScaler
import plotly.express as px
import kaleido
from gensim.models import CoherenceModel
import plotly.graph_objects as go
from sklearn.feature_extraction.text import TfidfVectorizer
import pysentiment2 as ps
```

### PARTE I: RECOGIDA DE DATOS Y ESTRUCTURACIÓN DE LA MUESTRA

```
# Cargamos los pdfs
pdf_search = Path("Noticias/").glob("*.pdf")
pdf_files = [str(file.absolute()) for file in pdf_search]
```

```
# Extraemos el texto y creamos una lista de pdfs
pdfs = []
for i in range(len(pdf_files)):
    pdf_reader = PdfReader(pdf_files[i])
    total_pages = len(pdf_reader.pages)
    texto_pags = []
    for j in range(total_pages):
```

```

texto_pags.append(pdf_reader.pages[j].extract_text())
texto_pags = " ".join(texto_pags)
pdfs.append(texto_pags)

```

```

# Comprobamos que no hay filas vacías
datos = pd.DataFrame({'texto': pdfs})
filas_vacias = datos['texto'].isnull()
num_filas_vacias = filas_vacias.sum()
print("Hay", num_filas_vacias, "filas vacías en el DataFrame.")

```

```

# Comprobamos que no hay duplicados
datos['Duplicado'] = False

for i, row1 in datos.iterrows():
    for j, row2 in datos.iterrows():
        if i != j and row1['texto'] == row2['texto']:
            datos.at[i, 'duplicado'] = True
            datos.at[j, 'duplicado'] = True

duplicates = datos[datos['duplicado'] == True]
duplicates

```

```

# Como hay un pdf repetido eliminamos uno de la muestra
datos = datos.drop(index=1797)

```

```

# Clasificamos el texto en categorías: título, fecha, contenido,
industria y temas
pdfs2 = datos['texto']

# Título
titulo = []
dias = ['Monday,', 'Tuesday,', 'Wednesday,', 'Thursday,', 'Friday,',
'Saturday,', 'Sunday,']
patron_titulo = r'^(.*) (?:' + '|'.join(dias) + r')'

for texto in pdfs2:
    match = re.search(patron_titulo, texto, re.DOTALL)
    if match:
        titulo.append(match.group(1))

titulo = list(map(lambda s: s.replace('\n', ' '), titulo))

```

```

# Fecha
fecha = []
patron_fecha =
r'\w+, \s\w+\s\d{2}, \s\d{4}\s\d{2}:\d{2}:\d{2}\s\w{2}\s\ (GMT\)'

for texto in pdfs2:
    fecha_encontrada = re.search(patron_fecha, texto)
    if fecha_encontrada:
        fecha.append(fecha_encontrada.group(0))
    else:
        fecha.append(None)

#Comprobamos que todos tengan fecha
for i, elem in enumerate(fecha):
    if elem is None:
        print("El elemento None está en la posición {i} de la lista.")

```

```

# Contenido
contenido = []
patron_contenido = r'\ (GMT\)\s*(.*?) (?:\s*Industries:|\s*Related
Identifiers:|\s*Subjects:)'

for texto in pdfs2:
    match = re.search(patron_contenido, texto, re.DOTALL)
    if match:
        contenido.append(match.group(1))

contenido = list(map(lambda s: s.replace('\n', ' '), contenido))

```

```

# Industrias
industria = []
patron_industria = r'Industries:\s*(.*?) (?:\s*Related
identifiers:|\s*Subjects:)'

for texto in pdfs2:
    industria_encontrada = re.search(patron_industria, texto)
    if industria_encontrada:
        industria.append(industria_encontrada.group(1))
    else:
        industria.append(None)

```

```

# Temas
tema = []
patron_tema = r"Subjects:(.*)"

for texto in pdfs2:
    tema_encontrado = re.search(patron_tema, texto)
    if tema_encontrado:
        tema.append(tema_encontrado.group(1))
    else:
        tema.append(None)

```

```

# Creamos el dataframe
df = pd.DataFrame(list(zip(titulo, fecha, contenido, industria,
tema)), columns=["Titulo", "Fecha", "Contenido", "Industrias",
"Temas"])
df

```

## PARTE II: ANÁLISIS DE LA MUESTRA

```

# Analizamos la frecuencia de noticias por mes
df['Fecha'] = pd.to_datetime(df['Fecha'], format='%A, %B %d, %Y
%i:%M:%S %p (%Z)')
df['Fecha']=pd.to_datetime(df['Fecha'], format='%Y-%m-%d')
df['Fecha'] = df['Fecha'].dt.date
df['Fecha'] = df['Fecha'].apply(lambda x: x.replace(day=1))
df['Fecha']=df['Fecha'] + pd.offsets.MonthBegin(0)

```

```

# Creamos el gráfico
freq_por_mes = df.groupby('Fecha')['Titulo'].count().reset_index()
freq_por_mes['Fecha_abr'] = freq_por_mes['Fecha'].dt.strftime('%b.
%y')
fig = px.bar(freq_por_mes, x='Fecha_abr', y='Titulo', text='Titulo')

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white',font_family='Times New Roman',
font_color='black',
xaxis=dict(title='Fecha',showline=True,linewidth=1,linicolor='black',
title_font=dict(size=18),ticks='outside',tickcolor='black',
tickwidth=1,ticklen=5,tickmode='linear'),
yaxis=dict(title='Frecuencia', showline=True, linewidth=1,
linicolor='black',title_font=dict(size=18),ticks='outside',
tickcolor='black',tickwidth=1,ticklen=5),title=dict(text='Frecuencia de
Noticias por Mes',x=0.5,y=0.98,xanchor='center',yanchor='top',
font=dict(size=22)))

```

```

fig.update_layout(showlegend=False)

# Guardamos el gráfico
fig.write_image("frecuencia_noticias_por_mes.png", engine="kaleido")

fig.show()

```

```

# Eliminamos el mes de abril
df = df.loc[df['Fecha'] != '20230401']
# Analizamos el número de palabras
df['Num_palabras'] = df['Contenido'].apply(lambda x: len(x.split()))
bins = [0, 100, 300, float("inf")]
labels = ["<100", "100-300", ">300"]
df['Rango_palabras'] = pd.cut(df['Num_palabras'], bins=bins,
labels=labels)

tabla =
pd.DataFrame(df['Rango_palabras'].value_counts()).reset_index()
tabla.columns = ['Rango de palabras', 'Número de noticias']

print(tabla)

```

```

# Creamos el gráfico
fig = px.histogram(df, x='Num_palabras', nbins=50)

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white',font_family='Times New Roman',
font_color='black',
xaxis=dict(title='Número de palabras',showline=True,linewidth=1,
linecolor='black',title_font=dict(size=18),ticks='outside',
tickcolor='black',tickwidth=1,ticklen=5),
yaxis=dict(title='Frecuencia',showline=True,linewidth=1,
linecolor='black',title_font=dict(size=18),ticks='outside',
tickcolor='black',tickwidth=1,ticklen=5),
title=dict(text='Histograma del número de palabras por noticia',
x=0.5,y=0.98,xanchor='center',yanchor='top',font=dict(size=22))

fig.update_layout(showlegend=False)

# Guardamos el gráfico
fig.write_image("histograma_palabras.png", engine="kaleido")

fig.show()

```

```

# Eliminamos las noticias con menos de 300 palabras
df2 = df[df['Num_palabras'] > 300]

# Creamos el gráfico
freq_por_mes = df2.groupby('Fecha')['Titulo'].count().reset_index()
freq_por_mes['Fecha_abr'] = freq_por_mes['Fecha'].dt.strftime('%b.%y')
fig = px.bar(freq_por_mes, x='Fecha_abr', y='Titulo', text='Titulo')

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New Roman',
font_color='black',
xaxis=dict(title='Fecha', showline=True, linewidth=1, linecolor='black',
title_font=dict(size=18), ticks='outside', tickcolor='black',
tickwidth=1, ticklen=5, tickmode='linear'),
yaxis=dict(title='Frecuencia', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
title=dict(text='Frecuencia de Noticias por Mes', x=0.5, y=0.98,
xanchor='center', yanchor='top', font=dict(size=22)))

fig.update_layout(showlegend=False)

# Guardamos el gráfico
fig.write_image("frecuencia_noticias_por_mes_300palabras.png", engine="
kaleido")

fig.show()

# Guardamos los datos en un Excel para no tener que repetir el proceso
de extracción
df.to_excel('datos_completos.xlsx', index=False)
df2.to_excel('datos_filtrados.xlsx', index=False)

```

### PARTE III: PREPROCESAMIENTO DE LA INFORMACIÓN

```
# Cargamos el dataframe
df = pd.read_excel('datos_filtrados.xlsx')

# Función de preprocesamiento

# Stop words
stop_words = stopwords("en")

# Lemmatizador
lemmatizer = WordNetLemmatizer()

def lemmatize_word(word):
    pos = get_wordnet_pos(word)
    lemmatized_word = lemmatizer.lemmatize(word, pos=pos)
    return lemmatized_word

def get_wordnet_pos(word):
    wordnet_pos = wordnet.synsets(word)
    if not wordnet_pos:
        return wordnet.NOUN
    nltk_pos = {
        'n': wordnet.NOUN,
        'v': wordnet.VERB,
        'a': wordnet.ADJ,
        's': wordnet.ADJ_SAT,
        'r': wordnet.ADV
    }

    pos = nltk_pos.get(wordnet_pos[0].pos()[0].lower(), wordnet.NOUN)
    return pos

def preprocesamiento(texto):
    if pd.isnull(texto):
        return ''
    else:
        # Convertimos el texto a minúsculas
        texto = texto.lower()
        # Tokenizamos el texto en palabras individuales
        palabras = nltk.word_tokenize(texto)
        # Eliminamos stopwords
        palabras_filtradas = [palabra for palabra in palabras if
            palabra not in stop_words and palabra.isalpha()]
        # Aplicamos la lematización
        palabras_lemmatizacion = [lemmatize_word(palabra) for palabra
            in palabras_filtradas]
```

```

# Unimos las palabras procesadas en una sola cadena de texto
texto_procesado = ' '.join(palabras_lemmatizacion)
return texto_procesado

```

```

# Aplicamos el preprocesamiento a los datos
df_proc = pd.DataFrame()
df_proc['Titulo_proc'] = df['Titulo'].apply(preprocesamiento)
df_proc['Fecha'] = df['Fecha']
df_proc['Contenido_proc'] = df['Contenido'].apply(preprocesamiento)
df_proc['Temas_proc'] = df['Temas'].apply(preprocesamiento)
df_proc['Industrias_proc'] = df['Industrias'].apply(preprocesamiento)

```

```

# Analizamos las palabras con mayor frecuencia
palabras = []
for doc in df_proc['Contenido_proc']:
    palabras.extend(doc.split())

palabras_contadas = Counter(palabras)

print(palabras_contadas.most_common(50))

# Eliminamos las palabras que consideramos stopwords
def eliminar_palabras(corpus, palabras_a_eliminar):
    nuevo_corpus = []
    for doc in corpus:
        palabras = [palabra for palabra in doc.split() if palabra not
                    in palabras_a_eliminar]
        nuevo_corpus.append(' '.join(palabras))
    return nuevo_corpus

```

```

palabras_a_eliminar = ['week', 'news', 'report', 'session', 'data',
                       'prior', 'continue', 'weekly', 'expect', 'bloomberg', 'reuters',
                       'yesterday', 'note', 'add', 'api', 'article', 'jan', 'feb', 'mar',
                       'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'dec']
df_proc['Contenido_proc'] =
eliminar_palabras(df_proc['Contenido_proc'], palabras_a_eliminar)

```

```

# Creamos la nube de palabras
texto = ' '.join(df_proc['Contenido_proc'])
nube_palabras = WordCloud(width=800,height=500,
background_color='white', min_font_size=10,
collocations=False).generate(texto)

plt.figure(figsize=(8,8), facecolor=None)
plt.imshow(nube_palabras)

```

```
plt.axis("off")
plt.tight_layout(pad=0)
plt.savefig('nube_de_palabras.png', dpi=300, bbox_inches='tight')
```

```
# Creamos la TF-IDF
corpus = df_proc['Contenido_proc'].tolist()
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(corpus)
feature_names = vectorizer.get_feature_names_out()
tfidf_scores = tfidf_matrix.mean(axis=0).tolist()[0]
terms_scores = list(zip(feature_names, tfidf_scores))
sorted_term_tfidf_scores = sorted(terms_scores, key=lambda x: x[1],
reverse=True)
top_20_terms = sorted_term_tfidf_scores[:20]
```

```
# Creamos la nube de palabras
terms = {term: score for term, score in sorted_term_tfidf_scores}
nube_palabras = WordCloud(width=800, height=500,
background_color='white').generate_from_frequencies(terms)
plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(nube_palabras)
plt.axis("off")
plt.tight_layout(pad=0)
plt.savefig('nube_de_palabras_tfidf.png', dpi=300,
bbox_inches='tight')
```

## PARTE IV: TOPIC ANALYSIS

```
# Vectorizamos el texto de la columna "Contenido_proc"
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(df_proc['Contenido_proc'])

# Identificamos el número óptimo de topics
n_topics_range = range(1, 20)

# Aplicamos el modelo LDA y calculamos la varianza acumulada para cada
topic
variance = []
for n_topics in n_topics_range:
    lda = LatentDirichletAllocation(n_components=n_topics,
random_state=42)
    lda.fit(X)
    topic_probs = lda.transform(X)
    scaler = MinMaxScaler()
    topic_probs_scaled = scaler.fit_transform(topic_probs)
    variance.append(topic_probs_scaled.var(axis=0).sum())
```

```

# Creamos el gráfico
n_topics_range = list(range(1, 20))

fig = px.line(x=n_topics_range, y=variance, hover_data={'Número de
topics':n_topics_range, 'Varianza acumulada':variance})

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New Roman',
font_color='black',
xaxis=dict(title='Número de topics', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5, tickmode='linear'),
yaxis=dict(title='Varianza acumulada', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
title=dict(text='Regla del codo para LDA', x=0.5, y=0.98,
xanchor='center', yanchor='top', font=dict(size=22)))

fig.add_trace(go.Scatter(x=n_topics_range, y=variance, mode='markers',
marker=dict(size=10, color='blue', symbol='circle', opacity=1,
line=dict(color='blue', width=2))))

fig.update_layout(showlegend=False)

# Guardamos el gráfico
fig.write_image("regla_del_codo.png", engine="kaleido")
fig.show()

# Seleccionamos el número de topics
n_topics = 4

# Vectorizamos el texto de la columna "Contenido_proc"
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(df_proc['Contenido_proc'])

# Aplicamos el modelo LDA para 4 topics
lda = LatentDirichletAllocation(n_components=n_topics,
random_state=42)
lda.fit(X)

topic_probs = lda.transform(X)
for i in range(n_topics):
    df_proc[f'topic_{i+1}_prob'] = topic_probs[:, i]

df_proc['max_topic'] = topic_probs.argmax(axis=1) + 1

```

```

# Creamos el gráfico
def plot_topic_timeseries(catalog, time_column, regex, title):

    df = catalog[[time_column]].join(catalog.filter(regex=regex),
    how='left')
    df2 = df.groupby(time_column).mean().reset_index()
    result = df2.melt(id_vars=time_column, var_name='Topic',
    value_name='Weight')
    fig = px.line(result, x=time_column, y='Weight', color='Topic')

    # Formateamos el gráfico
    fig.update_layout(plot_bgcolor='white',font_family='Times New
    Roman',font_color='black',
    xaxis=dict(title='Fecha',showline=True,linewidth=1,
    linecolor='black',title_font=dict(size=18),ticks='outside',
    tickcolor='black',tickwidth=1,ticklen=5),
    yaxis=dict(title='Peso del Tema',showline=True,linewidth=1,
    linecolor='black',title_font=dict(size=18),ticks='outside',
    tickcolor='black',tickwidth=1,ticklen=5),
    legend=dict(title='Topic', font=dict(size=14)),
    title=dict(text=title,x=0.5,y=0.98,xanchor='center',yanchor='top',
    font=dict(size=22)))

    # Guardamos el gráfico
    fig.write_image("topic_analysis.png", engine="kaleido")

    fig.show()

plot_topic_timeseries(df_proc, 'Fecha', 'topic_\d+_prob', 'Evolucion de
topics')

```

```

# Obtenemos el top 10 de palabras
vocab = np.array(vectorizer.get_feature_names_out())

for i in range(n_topics):
    topic_words = vocab[lda.components_[i].argsort()[::-1][:10]]
    # Calculamos la frecuencia normalizada para cada palabra
    topic_word_probs=lda.components_[i][lda.components_[i].argsort()
    [::-1][:10]]/lda.components_[i].sum()
    topic_word_probs,topic_words=zip(*sorted(zip(topic_word_probs,
    topic_words),reverse=False))

    # Creamos el gráfico
    fig=go.Figure()fig.add_trace(go.Bar(x=topic_word_probs,
    y=topic_words, orientation='h'))

```

```

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New
Roman', font_color='black',
xaxis=dict(title='Frecuencia normalizada', showline=True,
linewidth=1, linecolor='black', title_font=dict(size=18),
ticks='outside', tickcolor='black', tickwidth=1, ticklen=5,
range=[0,0.06]),
yaxis=dict(title='Palabra', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
title=dict(text=f'Top 10 palabras para el Topic {i+1}', x=0.5,
y=0.9, xanchor='center', yanchor='top', font=dict(size=22))
# Guardamos el gráfico
fig.write_image(f"topic_{i+1}_palabras.png", engine="kaleido")
fig.show()

```

```

# Creamos la nube de palabras
for i in range(n_topics):
    topic_words = vocab[lda.components_[i].argsort()[::-1][:100]]
    word_freq = {word:lda.components_[i][vocab == word][0] for word in
    topic_words}
    wc = WordCloud(width=800, height=400, background_color='white')
    .generate_from_frequencies(word_freq)
    plt.figure(figsize=(12,6))
    plt.imshow(wc, interpolation='bilinear')
    plt.axis('off')
    plt.title(f'Nube de palabras para el Topic {i+1} \n', fontdict=
    {'family': 'Times New Roman', 'size': 22, 'horizontalalignment':
    'center'})
    plt.savefig(f'nube_de_palabras_topic_{i+1}.png', dpi=300,
    bbox_inches='tight')
    plt.show()

```

## PARTE V: ANÁLISIS ENFOCADO EN EL CONFLICTO DE RUSIA Y UCRANIA

```

# Creamos una función de conteo de palabras
def count_word(column, word, flag=False):
    def detection(x):
        hits1 = re.findall(word, x)
        return len(hits1)
    result = column.apply(detection)
    if flag:
        result = result.clip(upper=1)
    return result
# Creamos una función para el gráfico
def grafico_frecuencia_palabra(dcf, palabra):

```

```

fig = px.line(dcf)

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New
Roman', font_color='black',
xaxis=dict(title='Fecha', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
yaxis=dict(title='Frecuencia', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
title=dict(text=f'Frecuencia de la palabra {palabra}', x=0.5,
y=0.98, xanchor='center', yanchor='top', font=dict(size=22)))

fig.update_layout(showlegend=False)

# Guardar el gráfico
fig.write_image(f"frecuencia_{palabra}.png", engine="kaleido")
fig.show()

```

```

# Análisis del término "import"
df_proc['Import'] = count_word(df_proc['Contenido_proc'], 'import')
df_Import = df_proc[['Fecha', 'Import']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Import, "import")

```

```

# Análisis del término "export"
df_proc['Export'] = count_word(df_proc['Contenido_proc'], 'export')
df_Export = df_proc[['Fecha', 'Export']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Export, "export")

```

```

# Análisis del término "russia"
df_proc['Russia'] = count_word(df_proc['Contenido_proc'], 'russia')
df_Russia = df_proc[['Fecha', 'Russia']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Russia, "russia")

```

```

# Análisis del término "ban"
df_proc['Ban'] = count_word(df_proc['Contenido_proc'], ' ban ')
df_Ban = df_proc[['Fecha', 'Ban']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Ban, "ban")

```

```

# Análisis del término "iraq"
df_proc['Iraq'] = count_word(df_proc['Contenido_proc'], 'iraq')
df_Iraq = df_proc[['Fecha', 'Iraq']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Iraq, "iraq")

```

```
# Análisis del término "turkey"
df_proc['Turquia'] = count_word(df_proc['Contenido_proc'], 'turkey')
df_Turkey = df_proc[['Fecha', 'Turquia']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Turkey, "turkey")
```

```
# Análisis del término "crude"
df_proc['Petroleo'] = count_word(df_proc['Contenido_proc'], 'crude')
df_Petroleo = df_proc[['Fecha', 'Petroleo']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Petroleo, "crude")
```

```
# Análisis del término "china"
df_proc['China'] = count_word(df_proc['Contenido_proc'], 'china')
df_Petroleo = df_proc[['Fecha', 'China']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Petroleo, "china")
```

```
# Análisis del término "india"
df_proc['India'] = count_word(df_proc['Contenido_proc'], ' india ')
df_Petroleo = df_proc[['Fecha', 'India']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Petroleo, "india")
```

```
# Análisis del término "natural gas"
df_proc['Gas_Natural'] = count_word(df_proc['Contenido_proc'],
'natural gas')
df_Gas_Natural=df_proc[['Fecha', 'Gas_Natural']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Gas_Natural, "natural gas")
```

```
# Análisis del término "grain"
df_proc['Grano'] = count_word(df_proc['Contenido_proc'], 'grain')
df_Grano = df_proc[['Fecha', 'Grano']].groupby('Fecha').sum()
grafico_frecuencia_palabra(df_Grano, "grain")
```

## PARTE VI: SENTIMENT ANALYSIS

```
# Importamos el diccionario Loughran-McDonald
lm = ps.LM()

# Realizamos el proceso de sentiment analysis
df_proc['Sentimiento'] = 0.0
df_proc['Clasificacion'] = ''
for index, row in df_proc.iterrows():
    contenido = row['Contenido_proc']
    palabras = contenido.split()
    score = lm.get_score(palabras)
    sentimiento = score['Polarity']
    df_proc.at[index, 'Sentimiento'] = sentimiento
    if sentimiento < 0:
        df_proc.at[index, 'Clasificacion'] = 'Negativo'
    elif sentimiento == 0:
        df_proc.at[index, 'Clasificacion'] = 'Neutro'
    else:
        df_proc.at[index, 'Clasificacion'] = 'Positivo'
resumen = df_proc['Clasificacion'].value_counts().reset_index()
resumen.columns = ['Clasificacion', 'Cantidad']
resumen
```

```
# Creamos el gráfico
resumen_tiempo = df_proc.groupby(['Fecha',
    'Clasificacion']).size().unstack().reset_index()
fig = px.line(resumen_tiempo, x='Fecha', y=['Negativo', 'Neutro',
    'Positivo'])

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New Roman',
    font_color='black',
    xaxis=dict(title='Fecha', showline=True, linewidth=1, linecolor='black',
    title_font=dict(size=18), ticks='outside', tickcolor='black',
    tickwidth=1, ticklen=5),
    yaxis=dict(title='Frecuencia', showline=True, linewidth=1,
    linecolor='black', title_font=dict(size=18), ticks='outside',
    tickcolor='black', tickwidth=1, ticklen=5),
    title=dict(text='Evolución del sentimiento a lo largo del tiempo',
    x=0.5, y=0.98, xanchor='center', yanchor='top', font=dict(size=22))
    legend=dict(title='Sentimiento', font=dict(size=14)),

# Guardamos el gráfico
fig.write_image("sentiment_analysis.png", engine="kaleido")
fig.show()
```

```

# Repetimos el proceso de sentiment analysis con una clasificación más
segmentada
df_proc['Sentimiento'] = 0.0
df_proc['Clasificacion2'] = ''
for index, row in df_proc.iterrows():
    contenido = row['Contenido_proc']
    palabras = contenido.split()
    score = lm.get_score(palabras)
    sentimiento = score['Polarity']
    df_proc.at[index, 'Sentimiento'] = sentimiento
    if sentimiento < -0.5:
        df_proc.at[index, 'Clasificacion2'] = 'Muy negativo'
    elif sentimiento < 0:
        df_proc.at[index, 'Clasificacion2'] = 'Negativo'
    elif sentimiento == 0:
        df_proc.at[index, 'Clasificacion2'] = 'Neutro'
    elif sentimiento < 0.5:
        df_proc.at[index, 'Clasificacion2'] = 'Positivo'
    else:
        df_proc.at[index, 'Clasificacion2'] = 'Muy positivo'
resumen = df_proc['Clasificacion2'].value_counts().reset_index()
resumen.columns = ['Clasificacion2', 'Cantidad']
resumen

```

```

# Creamos el gráfico
resumen_tiempo = df_proc.groupby(['Fecha',
'Clasificacion2']).size().unstack().reset_index()
fig = px.line(resumen_tiempo, x='Fecha', y=['Muy negativo', 'Negativo',
'Neutro', 'Positivo', 'Muy positivo'])

# Formateamos el gráfico
fig.update_layout(plot_bgcolor='white', font_family='Times New Roman',
font_color='black',
xaxis=dict(title='Fecha', showline=True, linewidth=1, linecolor='black',
title_font=dict(size=18), ticks='outside', tickcolor='black',
tickwidth=1, ticklen=5),
yaxis=dict(title='Frecuencia', showline=True, linewidth=1,
linecolor='black', title_font=dict(size=18), ticks='outside',
tickcolor='black', tickwidth=1, ticklen=5),
title=dict(text='Evolución del sentimiento a lo largo del tiempo',
x=0.5, y=0.98, xanchor='center', yanchor='top', font=dict(size=22))
legend=dict(title='Sentimiento', font=dict(size=14)),

# Guardamos el gráfico
fig.write_image("sentiment_analysis_2.png", engine="kaleido")
fig.show()

```