UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Generación de ciberinteligencia mediante OSINT para la detección de posturas políticas.

Trabajo Fin de Máster

Máster Universitario en Ciberseguridad y Ciberinteligencia

AUTOR/A: Carrillo Moraga, Juan Carlos

Tutor/a: Such Aparicio, José Miguel

Cotutor/a externo: BURDALO RAPA, LUIS ANTONIO

CURSO ACADÉMICO: 2022/2023

Resum

Amb motiu de les eleccions electorals celebrades al maig de 2023, aquest treball utilitza *OSINT*, en aquest cas *Twitter*, per a la generació de ciberinteligència en l'àmbit polític. Per a això pretenem detectar comunitats i investigar la homofilia política mitjançant una combinació d'aprenentatge automàtic i anàlisi de xarxes socials. L'objectiu principal és classificar als usuaris en comunitats en funció del contingut polític compartit i estudiar la variabilitat del discurs tractant de quantificar *echo chambers*. Com a objectiu secundari, es pretén detectar dins d'aquestes comunitats, *bots*, que pretenguen influenciar les campanyes electorals, així com també estudiar la toxicitat.

Paraules clau: Xarxes Socials, Xarxes Complexes, Detecció de comunitats, Agrupació de grafs, Modelatge estadístic, Aprenentatge profund, Postura, Debats Polítics, Homofilia, Twitter

Resumen

Con motivo de las elecciones electorales celebradas en mayo de 2023, este trabajo utiliza *OSINT*, en este caso *Twitter*, para la generación de ciberinteligencia en el ámbito político. Para ello pretendemos detectar comunidades e investigar la homofilia política mediante una combinación de aprendizaje automático y análisis de redes sociales. El objetivo principal es clasificar a los usuarios en comunidades en función del contenido político compartido y estudiar la variabilidad del discurso tratando de cuantificar *echo chambers*. Como objetivo secundario, se pretende detectar dentro de estas comunidades, *bots*, que pretendan influenciar las campañas electorales, así como también estudiar la toxicidad.

Palabras clave: Redes Sociales, Redes Complejas, Detección de comunidades, Agrupación de grafos, Modelado estadístico, Aprendizaje profundo, Postura, Debates Políticos, Homofilia, Twitter

Abstract

On the occasion of the electoral elections held in May 2023, this work uses *OSINT*, in this case *Twitter*, to generate cyberintelligence in the political sphere. To this end, we aim to detect communities and investigate political homophily using a combination of machine learning and social network analysis. The main objective is to classify users into communities based on the political content shared and to study the variability of discourse by trying to quantify "echo chambers". As a secondary objective, the aim is to detect within these communities, bots that aim to influence electoral campaigns, as well as to study toxicity.

Key words: Social Network, Complex Network, Community Detection, Graph Clustering, Statistical Modeling, Deep Learning, Stance, Political debates, Homophily, Twitter

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Los incidentes cibernéticos durante las elecciones estadounidenses de 2016, atribuidos al Gobierno ruso, iniciaron un nuevo capítulo en el debate sobre la ciberseguridad. Las operaciones de hackeo y filtración pusieron de relieve la cuestión de la manipulación estratégica, también denominada operaciones de influencia, como una amenaza para los procesos democráticos [133]. Aunque las operaciones de influencia no son ni mucho menos nuevas, el entorno tecnológico actual ofrece nuevas oportunidades a distintos actores.

1.1 Motivación

La motivación principal de este trabajo final de master radica en el creciente impacto de la manipulación estratégica en los procesos democráticos, especialmente en el contexto del entorno tecnológico actual.

Las elecciones estadounidenses de 2016 son un claro ejemplo de cómo las nuevas tecnologías ofrecen oportunidades sin precedentes para distintos actores que buscan influir en la opinión pública y socavar los procesos democráticos. En este contexto, es fundamental explorar y comprender los motivos detrás de estas acciones maliciosas y su impacto en nuestra sociedad.

La manipulación estratégica se ha convertido en un desafío significativo para la estabilidad de las democracias modernas. La propagación masiva de información falsa (*fake news* y *misinformation*), la creación de perfiles de usuario para dirigir contenido específico (*bots*) y la polarización de la sociedad son solo algunas de las tácticas utilizadas en estas operaciones de influencia. Estos fenómenos erosionan la confianza en las instituciones democráticas y pueden afectar negativamente la toma de decisiones informadas por parte de los ciudadanos.

Además, la rápida evolución de la tecnología ha superado en muchos aspectos las medidas de seguridad existentes, lo que dificulta aún más la detección y prevención de estas operaciones. Por lo tanto, es imperativo abordar este problema de manera integral, considerando aspectos técnicos, psicológicos, éticos y legales, a fin de desarrollar estrategias efectivas que protejan nuestros procesos democráticos.

La reflexión sobre este tema es crucial para proteger la integridad de los sistemas democráticos y garantizar la equidad en los procesos electorales. Al explorar los motivos detrás de la manipulación estratégica y sus implicaciones más amplias, se busca generar conciencia sobre esta problemática y promover la adopción de medidas adecuadas para contrarrestarla.

En resumen, la motivación detrás de este TFM es comprender mejor la manipulación estratégica (si la hay) en el contexto tecnológico actual y su impacto en los procesos democráticos. Al analizar los incidentes pasados, se busca obtener una comprensión más profunda de las motivaciones y tácticas utilizadas por los actores maliciosos, a fin de desarrollar estrategias de protección y prevención más efectivas. La reflexión sobre este tema es esencial para salvaguardar la integridad de nuestras democracias y promover la toma de decisiones informadas por parte de los ciudadanos.

1.2 Objetivos

- Recolectar y preprocesar retuits de los diputados del congreso en Twitter para obtener un conjunto de datos adecuado para el análisis.
- Clasificar usuarios según su ideología política.
- Aplicar el modelo `gdsmm` para identificar temas políticos y sus palabras representativas.
- Interpretar y analizar los resultados de Topic Modelling para obtener una panorámica general de la actualidad política en España.
- Investigar la manera de representar a los usuarios mediante sus tuits en un espacio vectorial para poder analizar la similitud de sus tuits.
- Cuantificar el "echo chamber" para cada uno de los temas seleccionados.
- Obtener una puntuación para cada uno de los usuarios que indique la probabilidad del usuario de ser un bot o no.
- Analizar cada uno de los tuits para obtener su toxicidad.
- Estudiar la correlación entre los bots y la toxicidad para ver si existe.

1.3 Estructura de la memoria

En el capítulo 2, se explorará el estado del arte para la interacción entre las redes sociales, las cámaras de eco y la polarización política. Los subapartados incluirán:

- **Redes sociales y cámaras de eco:** Se describirá cómo las redes sociales pueden dar lugar a cámaras de eco, donde los usuarios están expuestos principalmente a opiniones afines.
- **Redes sociales y polarización política:** Se analizará cómo las redes sociales pueden contribuir a la polarización política al fomentar la fragmentación de la opinión pública.
- **Conclusiones:** Este apartado resumirá las principales conclusiones sobre el impacto de las redes sociales en la polarización política y las cámaras de eco.

El capítulo 3 se centrará en el estado del arte del discurso de odio en las redes sociales. Los subapartados serán:

- **Definición de la incitación al odio en línea:** Se definirá y describirá la incitación al odio en línea, con ejemplos para clarificar el concepto.

- **Detección de la incitación al odio en línea:** Se discutirán las técnicas y herramientas utilizadas para detectar el discurso de odio en las redes sociales.
- **Lucha contra la incitación al odio en internet:** Se explorarán las estrategias y medidas para combatir el discurso de odio en línea, tanto desde un enfoque tecnológico como político.
- **Conclusiones:** Se resumirán las principales conclusiones sobre la detección y lucha contra el discurso de odio en las redes sociales.

El capítulo 4 abordará la generación del dataset y el uso del Topic Modelling. Los subapartados serán:

- **API de Twitter, Wikidata y Neo4j:** Se explicará cómo se creó el dataset utilizando fuentes como la API de Twitter, Wikidata y Neo4j.
- **Análisis de Temas Políticos en España mediante *Topic Modelling* en Twitter:** Se describirá cómo se aplicó el Topic Modelling en nuestro conjunto de datos, incluyendo la metodología usada y resultados.
- **Conclusiones:** Se presentarán las conclusiones relacionadas con la generación del dataset y el uso de Topic Modelling.

El capítulo 5 se enfocará en la cuantificación de las cámaras de eco en Twitter. Los subapartados incluirán:

- **Terminología:** Se definirán los términos clave relacionados con la cuantificación de cámaras de eco en Twitter.
- **Trabajo relacionado:** Se revisará la literatura existente sobre la detección de cámaras de eco y la representación del aprendizaje en Twitter.
- **Metodología:** Se describirá la metodología utilizada para cuantificar las cámaras de eco, incluyendo la detección de cámaras y comunidades, así como la generación de *embeddings* de usuarios.
- **Experimentos y resultados:** Se presentarán los detalles de los experimentos realizados y los resultados obtenidos.

El capítulo 6 se centrará en la cuantificación del discurso de odio en las redes sociales. Los subapartados serán:

- **Metodología:** Se explicará la metodología utilizada para cuantificar el discurso de odio, incluyendo la medición de toxicidad y la detección de bots.
- **Análisis de los resultados obtenidos:** Se presentarán los resultados de la cuantificación del odio y se analizarán las implicaciones de estos resultados.
- **Conclusiones:** Se resumirán las principales conclusiones relacionadas con la cuantificación del discurso de odio en las redes sociales.

En el último capítulo veremos cuales han sido las conclusiones respecto a los experimentos realizados, el estado del arte, cuales han sido las limitaciones que hemos tenido al realizar el trabajo junto con las propuestas para continuar con el mismo en un trabajo futuro.

CAPÍTULO 2

Redes sociales, cámaras de eco y polarización política.

El inesperado ascenso de partidos y candidatos populistas en las democracias desarrolladas han dado urgencia al debate sobre el papel que la tecnología y las redes sociales pueden estar desempeñando en la exacerbación de la polarización y la incitación a la violencia extremista. Un argumento popular que suele servir como explicación que vincula las redes sociales con la polarización política está relacionado con su capacidad para fomentar la aparición de cámaras de eco en las que se amplifican las ideas extremistas. Sunstein [125], uno de los principales defensores de este punto de vista, sostiene que la principal característica de las redes sociales es que permiten a las personas con ideas políticas afines encontrarse entre sí. En este entorno, los ciudadanos solo están expuestos a la información que refuerza sus opiniones políticas y permanecen aislados de otros individuos con opiniones opuestas, en parte debido a los efectos de los algoritmos de clasificación que generan burbujas de filtro [109] y crean incentivos para que los editores compartan contenidos clickbait e hiperpartidistas [24].

El resultado de este proceso es una sociedad cada vez más segregada por líneas partidistas y en la que es improbable llegar a acuerdos debido a la creciente desconfianza hacia el Gobierno, los medios de comunicación y ciudadanos con ideas diferentes.

A pesar de este aparente consenso, los estudios empíricos ofrecen una visión mucho más matizada de cómo los medios sociales afectan a la polarización política, cuestionando a menudo las premisas básicas de este argumento. Incluso si la mayoría de los intercambios políticos en las redes sociales tienen lugar entre personas con ideas similares, las interacciones transversales son más frecuentes de lo que comúnmente se cree [19], la exposición a noticias diversas es mayor que a través de otros tipos de medios [20] [55] y los algoritmos de *ranking* no tienen un gran impacto en el equilibrio ideológico del consumo de noticias en Facebook o Google [14] [71]. Una posible explicación de esta serie de hallazgos es que las redes sociales aumentan la exposición a la información compartida por vínculos débiles [66], como compañeros de trabajo, familiares y conocidos, que son más propensos a compartir información novedosa, incluidas noticias ideológicamente diversas [16][15]. Por supuesto, el hecho de que los medios sociales aumenten la exposición a ideas políticas diversas a partir de vínculos débiles no significa necesariamente que no tengan ningún efecto sobre la polarización política. Investigaciones anteriores muestran que la exposición repetida a información transversal conduce a la moderación política [97], lo que podría explicar por qué la polarización política en Estados Unidos ha aumentado menos entre los ciudadanos menos propensos a usar las redes sociales [29]. Sin embargo, cada vez más trabajos cuestionan este hallazgo, argumentando que es precisamente esta mayor exposición a opiniones transversales lo que

puede estar teniendo efectos polarizadores [13][124]. En el libro, Frenemies [118], Settle identifica la mayor conciencia de las identidades políticas en los medios sociales como un factor clave que impulsa la polarización afectiva.

A pesar de la creciente atención que los estudiosos prestan a este tema, aún queda mucho por saber. La mayoría de los estudios han centrado su atención en el contexto estadounidense y las pruebas empíricas sobre esta cuestión son escasas.

2.1 Redes sociales y cámaras de eco.

El debate sobre si las tecnologías digitales contribuyen a unir a personas políticamente afines o a crear comunidades ideológicas aisladas es casi tan antiguo como Internet. Posiblemente la primera vez que se esgrimió este argumento fuera en 1996 por Van Alstyne y Brynjolfsson [128], quienes advirtieron de que las tecnologías de la información podrían conducir a la ciberbalcanización. Argumentaban: ”La tecnología puede reducir las distancias geográficas y facilitar el intercambio de información, creando así una aldea global.” Pero ése era sólo un resultado posible. Su hipótesis era que un escenario alternativo sería uno en el que las sociedades estuvieran más fragmentadas y las interacciones se balcanizasen porque ”Internet facilita la búsqueda de individuos con ideas afines”, lo que ”puede facilitar y fortalecer las comunidades marginales que tienen una ideología común pero están dispersas geográficamente” y ”una vez que los individuos con ideas afines se localizan entre sí”, concluyen, ”sus interacciones posteriores pueden polarizar aún más sus puntos de vista o incluso emprender acciones en conjunto” .

En un estudio posterior Putnam [112] advierte: ”Mientras que las interacciones en el mundo real a menudo nos obligan a lidiar con la diversidad, el mundo virtual puede ser más homogéneo. Este proceso podría fomentarse con el uso de nuevas tecnologías de *filtering* que automatizan la criba de mensajes irrelevantes.” Además proponía tres hipótesis:

1. las tecnologías digitales facilitan la aparición de comunidades de individuos con ideas afines,
2. donde cada vez están más aislados de cualquier información desafiante, un proceso que se ve exacerbado por
3. la filtración de los algoritmos.

Estas tres hipótesis están presentes en la mayoría de los estudios posteriores sobre esta cuestión, a menudo analizados en el contexto de otras teorías de la comunicación política.Las siguientes secciones ofrecen un análisis más detallado de cada uno de estos componentes, con una visión general de las pruebas empíricas que los apoyan o refutan.

2.1.1. Comunidades de personas con ideas afines.

El acceso a Internet reduce drásticamente los costes del intercambio de mensajes y la búsqueda de información, independientemente de la distancia geográfica. Al no estar atados por la proximidad física, los ciudadanos adquieren la capacidad de conectarse y organizarse basándose en intereses compartidos. Estas conversaciones pueden darse en espacios públicos, a través de foros abiertos o blogs, pero también en entornos más privados, como grupos cerrados en Facebook o comunidades privadas.

Como Sunstein expone en su libro *#Republic: Divided Democracy in the Age of Social Media* [125], los espacios en línea crean oportunidades para la deliberación entre personas

con ideas afines. Esto puede no ser un punto negativo. De hecho, puede promover el desarrollo de posturas que de otro modo quedarían silenciadas u ofrecer un espacio seguro a personas que sufren discriminación.

Sin embargo, Sunstein sostiene que: "en la práctica, el resultado de la deliberación de ideas afines es la polarización del grupo, que representa un caldo de cultivo para el extremismo y puede incluso poner en riesgo la estabilidad social." Su preocupación aquí es que: "a través de los mecanismos de la influencia social y los argumentos persuasivos, los miembros se muevan hacia posiciones que carecen de mérito". Esto podría ocurrir bien porque la homogeneidad del grupo restringe el tamaño del conjunto de argumentos, bien porque los individuos serán más proclives a expresar una opinión popular dentro del grupo para obtener la aprobación del mayor número posible de miembros. La existencia de estos dos mecanismos ha quedado demostrada en [99] donde muestran que los individuos que participan en grupos de discusión homogéneos tienden a adoptar posturas más extremas tras deliberar con sus compañeros afines.

La literatura sobre redes sociales también ha abordado esta cuestión en el contexto de los estudios de homofilia; es decir, la propensión de los individuos a agruparse en función de rasgos comunes, como la ideología política [91]. Por ejemplo, Adamic et al. [2] analizaron la red de hipervínculos que conecta los 1.000 blogs políticos más activos durante las elecciones presidenciales de 2004 en EE.UU. Concluyendo que: "liberales y conservadores se vinculan principalmente dentro de sus propias comunidades". Esto se aplica no sólo a los editores de los blogs, sino también a los lectores de blogs: Lawrence et al. [82] descubrieron que los lectores de blogs gravitan hacia aquellos blogs que coinciden con su orientación política y son más propensos a la polarización que los no lectores. Aunque la mayor parte de la bibliografía inicial se centraba en Internet en general, trabajos recientes han hallado pautas similares en las redes sociales. Por ejemplo, Conover et al. [40] demostraron que los usuarios partidistas de Twitter son significativamente más propensos a difundir mensajes afines a sus posiciones ideológicas. Además, según Barberá et al. [19], cerca del 75 % de los retuits sobre temas políticos tienen lugar entre usuarios de opiniones ideológicas similares. Del mismo modo pero en Facebook, Del Vicario et al. [45] descubrieron que la información relacionada con noticias científicas y teorías conspirativas también tiende a difundirse en comunidades homogéneas y polarizadas. Más allá del ámbito político, Aiello et al. [4] demostraron que los usuarios con intereses similares tienen más probabilidades de ser amigos en las comunidades online.

Para situar este conjunto de resultados en su contexto, es importante señalar que estos patrones no son necesariamente generalizables a todos los usuarios de las redes sociales. De hecho, es probable que la percepción generalizada de polarización en las redes sociales se deba a una minoría de individuos partidistas muy activos y visibles. Barberá y Rivero [17] ofrecen pruebas que apuntan en esta dirección: "Una minoría de individuos muy activos fue responsable de la abrumadora mayoría de contenidos hiperpartidistas difundidos en Twitter antes de las elecciones presidenciales estadounidenses de 2012". Del mismo modo, un estudio sobre el intercambio de noticias en Twitter [119] reveló que solo un pequeño núcleo de la red mostraba indicios de polarización.

2.1.2. Sesgo de confirmación.

El segundo componente del argumento estándar que relaciona las tecnologías digitales y la polarización política está relacionado con la capacidad de las personas para filtrar toda la información que pueda cuestionar sus opiniones formadas. A medida que los individuos pasan más tiempo en comunidades de personas con ideas afines, no sólo están más expuestos a mensajes favorables a sus actitudes, sino que también disminuye su exposición a información contraria. Esto es lo que conduce a la aparición de cámaras

de eco, en las que los ciudadanos no ven ni oyen una amplia gama de temas o ideas, lo que limita su capacidad para llegar a un terreno común en cuestiones políticas.

Sin embargo, trabajos anteriores sobre la exposición selectiva han descubierto que la búsqueda de argumentos que refuercen la opinión propia y la evitación de retos que cuestionen dicha opinión, no están intrínsecamente relacionadas. Utilizando datos de encuestas, Garrett [62] demostró que los ciudadanos aprovechan la mayor disponibilidad de noticias en línea para aumentar la exposición a historias políticas coherentes con sus puntos de vista, pero no evitan sistemáticamente los desafíos de opinión. En un estudio de seguimiento en el que se utilizó el seguimiento del comportamiento en un entorno realista, Garrett [61] no encontró pruebas de que los individuos abandonen las noticias que contienen información con la que no están de acuerdo una vez que ya han empezado a leer esas noticias.

Los estudios sobre el consumo de noticias en línea a nivel individual también ofrecen escasas pruebas de que los ciudadanos eviten activamente la información contraria a sus actitudes. Gentzkow y Shapiro [63] mostraron que la segregación ideológica en los sitios de noticias que la gente visita regularmente es baja y no aumenta con el tiempo. Del mismo modo, Guess [69] encontró altos niveles de solapamiento en el consumo de noticias: Una mayoría de personas confía mayoritariamente en sitios web centristas, y quienes visitan los sitios web más partidistas son una minoría que tiende a ser consumidora activa de noticias y, por tanto, también visita muchos otros sitios.

Esta distinción entre la búsqueda de reforzar nuestra ideología y la evitación de cambiar de opinión cobra relevancia cuando nos centramos en las redes sociales, donde la mayoría de las noticias políticas que consumen los ciudadanos las publican sus amigos y familiares. Como demuestran Messing y Westwood [92] utilizando un experimento de laboratorio que recrea un *feed* de noticias de Facebook, es probable que las personas hagan clic y lean noticias con las que potencialmente no están de acuerdo siempre que sean recomendadas por sus amigos y conocidos. En otras palabras, las recomendaciones de noticias de los amigos son poderosas señales sociales que pueden reducir la exposición selectiva partidista a niveles indistinguibles del azar.

Por supuesto, este proceso de consumo social de información puede dar lugar a entornos de cámara de eco si las redes de ciudadanos son políticamente homogéneas y, por tanto, todo el contenido compartido por los vínculos sociales es favorable a las actitudes. De hecho, hay algunas pruebas de que los vínculos sociales en las redes en línea reproducen los vínculos *offline* en su composición y naturaleza [76] [26]. Sin embargo, un estudio tras otro demuestra que, a pesar de la capacidad de los ciudadanos para seleccionar las noticias que consumen, a partir de una gama más reducida de fuentes gracias a las tecnologías digitales, están expuestos a un conjunto diverso de puntos de vista a través de los sitios web y aún más en las redes sociales.

Este hallazgo es robusto al uso de diferentes fuentes de datos y parece mantenerse en todos los países. Flaxman, Goel y Rao [54] utilizan datos conductuales de historiales de navegación web de 50.000 adultos que consumen noticias en línea y ofrecen probablemente las mejores pruebas sobre las pautas de búsqueda de refuerzo y evitación. Por un lado, el consumo de noticias a través de las redes sociales y los motores de búsqueda contribuye a aumentar la segregación ideológica de las audiencias. Sin embargo, a nivel individual, estos canales conducen a una mayor exposición a criterios contrarios que la navegación por parte del usuario. De forma similar a lo que señala Guess [69], esta aparente paradoja puede explicarse por una minoría de individuos partidistas que son responsables de la mayor parte del consumo de noticias partidistas.

Bakshy, Messing y Adamic [14] obtienen resultados similares en Facebook. Su análisis examina el contenido ideológico de los feeds de noticias de Facebook de 10,1 millones de

usuarios estadounidenses de Facebook. Como era de esperar, descubren que la mayoría de los vínculos de amistad se establecen entre personas de un mismo grupo ideológico y que, de hecho, es más probable que los usuarios se relacionen con contenidos afines. Sin embargo, muchas amistades (el 20 % para los conservadores y el 18 % para los liberales) son transversales a los grupos ideológicos y, lo que es más importante, la exposición a noticias y opiniones ideológicamente diversas también es alta: de media, alrededor del 30 % de las noticias políticas que los usuarios ven en su feed de noticias son transversales. Esta proporción es notablemente similar a lo que Barberá [16] informa utilizando datos de Twitter: el 33 % de los tuits a los que estuvo potencialmente expuesta una muestra de usuarios con interés político son transversales.

Más allá del consumo de noticias en línea, Barnidge [20] ofrece una útil comparación de cómo los adultos estadounidenses declaran estar expuestos a desacuerdos políticos en diferentes entornos. Su estudio se basa en datos de encuestas que, a expensas de posibles sesgos de información, tienen la ventaja de permitir una comparación de las interacciones offline y el consumo offline de noticias (periódicos, televisión, etc.) con las interacciones en los medios sociales. Los resultados confirman el patrón descrito a lo largo de esta sección: Los encuestados admiten estar más expuestos al desacuerdo político en las redes sociales que en otros entornos de comunicación.

Aunque los estudios comparativos sobre el consumo de noticias en las redes sociales siguen siendo escasos, los datos presentados por Fletcher y Nielsen [56] como parte del informe anual *Reuters Institute Digital News Report* sugieren que el caso de Estados Unidos no es una excepción. Utilizando encuestas representativas realizadas en treinta y siete países, defienden que las audiencias en línea no parecen estar más fragmentadas políticamente que las audiencias *offline*, que la exposición accidental a contenido político en los medios sociales es un fenómeno generalizado y que, de hecho, las personas que utilizan sitios de redes sociales están expuestas a noticias diversas en mayor proporción que las que no lo hacen. Del mismo modo, un informe reciente del *Pew Research Center* realizado en once naciones de cuatro regiones globales (Silver et al. [121]) descubrió que los usuarios de los medios sociales tenían más probabilidades de interactuar regularmente con una red más diversa, incluidas personas de diferentes partidos políticos, que aquellos que no eran activos en los medios sociales. ¿Cómo podemos conciliar el hecho de que los ciudadanos tengan ahora una capacidad mucho mayor para filtrar cualquier desafío de opinión con el hecho de que en realidad no parezcan hacerlo? Una explicación, que ya he avanzado antes en esta sección, es que quizá la mayoría de la gente no se esfuerza realmente por evitar la información contraria a las actitudes [62] y, en ese sentido, el patrón que observamos en un entorno online replica lo que observaríamos offline. Un argumento alternativo, sin embargo, es específico de las plataformas de redes sociales: los sitios de medios sociales aumentan la exposición a opiniones diversas porque nos conectan con "vínculos débiles".

La mayoría de las personas con las que interactuamos en nuestra vida cotidiana pueden considerarse "vínculos débiles", bien porque la frecuencia con la que nos comunicamos es baja, bien porque no las percibimos como cercanas a nosotros. En cambio, "los vínculos fuertes" son esa minoría de personas a las que vemos con frecuencia y en las que solemos confiar más. Los vínculos débiles suelen ser compañeros de trabajo, amigos del colegio, parientes lejanos y otros conocidos; mientras que los "vínculos fuertes" son nuestras parejas, amigos íntimos y familiares directos. Esta distinción resulta relevante en el contexto de la difusión de información porque, como descubrió Granovetter [66], los individuos están expuestos a información novedosa a través de los vínculos débiles.

Cuando pensamos en las redes sociales, sin duda la principal forma en que influyen en nuestra vida cotidiana es facilitando el contacto con personas a las que no veríamos en persona con regularidad. En otras palabras, suponen una mayor exposición y con-

tacto con vínculos débiles que en las interacciones offline [64]. Los "vínculos débiles" son fundamentales si queremos entender la difusión de la información, porque ayudan a conectar partes periféricas de las redes en línea [18][43]. Esto se aplica no solo a los movimientos de protesta, sino también al consumo de noticias en general.

La otra forma en que la fuerza del vínculo es importante para este argumento está relacionada con el hecho de que la homofilia tiende a ser menor entre los vínculos débiles [91]. En otras palabras, dado que los seres humanos muestran una propensión a establecer preferentemente vínculos con otras personas que son similares a ellos, deberíamos esperar que nuestros vínculos débiles sean más diferentes de nosotros que los vínculos fuertes; y esto se aplica también a la similitud ideológica. Es esta heterogeneidad la que explica por qué los vínculos débiles son responsables de la propagación de información novedosa en las redes sociales [15].

En resumen, la explicación más probable de por qué los ciudadanos no son capaces de evitar los desafíos de opinión en las redes sociales son:

1. No tienen la capacidad de elegir qué contenidos ven porque la exposición es accidental

2. la mayoría de esos contenidos son compartidos por vínculos débiles, que tienden a ser más diversos ideológicamente que los vínculos fuertes.

2.1.3. La burbuja del filtro.

En los dos apartados anteriores se han abordado dos mecanismos que están en la base de cómo los individuos deciden consumir (o no) información política. Sin embargo, en el actual entorno *online*, éstos no son los dos únicos ingredientes que determinan la dieta mediática de los ciudadanos. A medida que aumentan el número y la heterogeneidad de las opciones, los ciudadanos no están preparados para hacer frente a la sobrecarga de información que conlleva una disponibilidad tan amplia. En este contexto, los motores de búsqueda y las redes sociales han pasado a basarse en algoritmos automatizados y personalizados en tiempo real para ayudar a los usuarios a navegar por la web.

En algunos casos, el conjunto de heurísticas y reglas que determinan cómo se muestra la información en un sitio web se conoce bien. Por ejemplo, los resultados de las búsquedas en Google se clasificaban, al menos en las primeras versiones de este sitio web, mediante el algoritmo PageRank desarrollado por sus fundadores [107]. Sin embargo, cuando se trata de redes sociales, los algoritmos que determinan cómo se clasifica la información en las noticias de los usuarios (en Facebook) o en la cronología (en Twitter) suelen considerarse una caja negra. Aparte de algunos principios generales sobre cómo estas empresas toman una serie de señales/interacciones e intentan hacer predicciones sobre la puntuación de relevancia para cada combinación de usuario y pieza de contenido, los sitios de redes sociales han publicado poca información sobre cómo funciona realmente este proceso. Recientemente Twitter ha hecho open source su algoritmo de recomendación. [1]. La razón más probable de esta opacidad es que liberar más información podría facilitar las cosas para los editores que intentan jugar con el sistema. Sin embargo, esta falta de transparencia también ha suscitado preocupación sobre hasta qué punto estos algoritmos podrían estar contribuyendo realmente a exacerbar las desigualdades y la segregación ideológica.

Sunstein [125] utiliza este concepto para justificar su preocupación por la "era del algoritmo", en la que los ciudadanos ya no controlan las noticias que consumen. Incluso si la gente no eligiera voluntariamente gravitar hacia las cámaras de eco, puede que no

tengan otra opción, ya que los sitios de medios sociales se convierten en los árbitros de lo que la gente ve y lo que no.

El libro más influyente dentro de esta línea de pensamiento es *The Filter Bubble*, de Eli Pariser [109]. Su preocupación se aplica tanto a los motores de búsqueda, que ofrecerían resultados completamente distintos en función de la predicción del sitio web sobre cuál es la intención del usuario. Como a los sitios de redes sociales, que mostrarían a los usuarios solo contenidos que probablemente les gusten. A medida que estos servicios en línea aprenden más sobre los usuarios, su precisión a la hora de predecir lo que preferirían ver es mayor, lo que acabaría desembocando en burbujas en las que los ciudadanos nunca estarían expuestos a ningún tipo de información que pudiera causarles malestar.

La preocupación de Pariser es doble. En primer lugar, expresa su preocupación por los algoritmos que aumentan la polarización política y las que sólo prefieren consumir noticias de entretenimiento. En segundo lugar, para aquellos con claras preferencias partidistas, afirma, que los algoritmos reducen la medida en que se "escucha a la otra parte". El resultado de este doble proceso, es una sociedad en la que el tipo de experiencias compartidas que son necesarias en una democracia que funcione bien simplemente no existen.

Aunque solo se ha mencionado brevemente el libro de Pariser, es importante señalar que el concepto de sesgo algorítmico ha recibido una atención mucho más amplia en la comunicación política y la informática. Los métodos de aprendizaje profundo, que sacrifican la interpretabilidad para mejorar su precisión, han sido acusados con frecuencia de tener sesgos implícitos. El hecho de que los algoritmos estén automatizados no implica que no reproduzcan comportamientos humanos comunes, ya que los conjuntos de datos de entrenamiento que utilizan para reproducir el comportamiento humano en muchos casos también incorporan dichos sesgos [33].

En el caso específico de la segregación ideológica, ¿qué sabemos sobre el impacto de los algoritmos de clasificación? La respuesta es que no mucho. Carecemos del tipo de pruebas sistemáticas que nos permitirían responder a preguntas clave en este campo. Una explicación de esta escasez de pruebas es la incapacidad de los investigadores para acceder o manipular los algoritmos de búsqueda o de clasificación social. Por eso no es de extrañar que el estudio más amplio sobre este tema lo hayan realizado los investigadores de Facebook, [14]. Sus resultados muestran que el algoritmo de Facebook induce una reducción significativa, aunque pequeña, de la exposición de los usuarios a contenidos transversales. Dicen, "las elecciones individuales limitan más que los algoritmos la exposición a contenidos que cuestionan las actitudes".

Sin embargo, este estudio tiene sus limitaciones. Como reconocen, su muestra se limita a usuarios activos que declaran su afiliación política en su perfil. Aunque esto puede llevar a subestimar hasta qué punto los usuarios están expuestos a opiniones transversales, cuando se trata de la importancia relativa de las elecciones individuales frente al algoritmo, puede ocurrir lo contrario. En otras palabras, es más probable que las personas que autodeclaran sus opiniones políticas estén políticamente interesadas y, por tanto, sean más selectivas en cuanto a los contenidos políticos que consumen. Las personas con un menor nivel de interés político pueden confiar en mayor medida en el algoritmo de Facebook para decidir a qué noticias políticas se exponen.

2.2 Redes sociales y polarización política.

La revisión de la literatura sobre redes sociales y "cámaras de eco" ha mostrado que, de forma bastante paradójica, existen pruebas empíricas convincentes que demuestran

que los sitios de redes sociales aumentan la gama de opiniones políticas a las que están expuestos los individuos. Sin embargo, eso nos dice poco sobre su impacto posterior en las actitudes individuales y la polarización. Como describo a continuación, hay buenas razones para esperar un efecto en ambas direcciones.

Los teóricos de la política han considerado durante mucho tiempo que la exposición compartida a puntos de vista diferentes es una condición necesaria para el tipo de deliberación política saludable que tiene lugar en las sociedades democráticas prósperas [94][70]. Sin embargo, más allá de este argumento normativo, la deliberación diversa también es importante porque puede tener un profundo impacto en las creencias de los ciudadanos, y en su fortaleza. Las interacciones transversales aumentan la conciencia de los fundamentos de los puntos de vista propios y contrarios [98]. La exposición a opiniones diversas también puede enmarcarse basándose en la clásica teoría del contacto intergrupal propuesta por G. Allport [7], este argumento implicaría que la exposición a miembros de grupos diferentes (en este caso, personas que apoyan a un partido diferente) debería reducir los prejuicios y fomentar la tolerancia política.

Si el uso de los medios sociales aumenta efectivamente la concienciación de los ciudadanos sobre la diversidad de puntos de vista y fomenta el contacto intergrupal, parece razonable esperar que también pueda debilitar la solidez de las creencias políticas de las personas y reducir así la polarización política. Aunque una revisión de cómo operan estos dos mecanismos en el contexto offline está más allá del alcance de este trabajo, hay de hecho amplias pruebas de que las interacciones transversales fomentan la moderación política [34], por ejemplo, aumentando el apoyo a las libertades civiles de los grupos que no gustan [97], y de que el contacto intergrupal reduce los prejuicios [110] [108].

Sin embargo, los estudios empíricos que examinan esta cuestión en el contexto de las redes sociales son escasos y ofrecen resultados dispares. Algunos estudios encuentran una relación inversa entre el uso de las redes sociales y la polarización política, como cabría esperar de la bibliografía analizada en esta sección. Probablemente, el ejemplo más destacado sea el trabajo de Boxell, Gentzkow y Shapiro [29], que hallan que, aunque el nivel de polarización política ha aumentado en todas las cohortes de edad en Estados Unidos, el cambio ha sido de menor magnitud entre el grupo de personas más jóvenes, las más propensas a ser activas en las redes sociales. Este resultado sugiere que es probable que las tecnologías digitales desempeñen un papel limitado a la hora de explicar por qué está aumentando la polarización.

A pesar de las dudas que pueda suscitar este artículo en particular, al menos otros tres estudios arrojan resultados que apuntan en la misma dirección. Heatherly, Lu y Lee [72] descubren que las personas que participan en debates transversales en las redes sociales presentan niveles más bajos de polarización política. Barberá [16] constata que los individuos en redes de Twitter ideológicamente diversas tienden a moderar sus posiciones ideológicas a lo largo del tiempo. Utilizando datos de encuestas de veintiocho países europeos, Ngyuen y Vu [102] descubren que los ciudadanos que consumen noticias políticas a través de las redes sociales no están más polarizados que los que recurren a otras fuentes. Otros estudios, en cambio, sí sugieren que la exposición a información política a través de las redes sociales podría tener efectos polarizadores. Bail et al. [13] muestran que la exposición transversal a mensajes políticos de las élites puede aumentar la polarización. Utilizando un diseño de investigación innovador, los autores reclutaron una muestra de encuestados y luego les pidieron que siguieran bots que compartían mensajes políticos que eran contrarios a sus propias opiniones. Un estudio longitudinal de las opiniones políticas de los encuestados mostró que los republicanos se volvieron significativamente más conservadores. Los demócratas también se volvieron algo más liberales, aunque el cambio no fue estadísticamente significativo. Estos resultados evidencian un efecto de reacción que podría deberse a un razonamiento motivado [87].

Del mismo modo, Suhay et al. [124] defiende que la exposición a desacuerdos políticos en entornos en línea aumenta la polarización política. Estos desacuerdos se presentan en un contexto incivil, que, en su opinión, es representativo del tipo de interacciones transversales que tienen lugar en línea. Es la crítica a las identidades partidistas, y no necesariamente las opiniones en línea sobre temas específicos, lo que impulsa la polarización.

Un último artículo que examinó los efectos polarizadores generales de las redes sociales es el de Allcott et al. [6]. Los autores estudiaron cómo el abandono de Facebook afecta a diversos resultados, incluida la polarización política. Los resultados revelan que la desactivación redujo la polarización de las opiniones sobre cuestiones políticas. Sin embargo, el conocimiento de los sujetos sobre los eventos actuales también disminuyó, lo que sugiere que los efectos despolarizantes de abandonar los medios sociales pueden explicarse por un menor nivel de exposición a la información política en general.

En conclusión, vemos que la literatura existente ofrece resultados que parecen estar en desacuerdo. ¿Cómo podemos conciliar estos resultados divergentes? Para responder a esta pregunta, es importante tener en cuenta cómo definimos la polarización política: ¿nos referimos sólo a la divergencia de opiniones políticas o posturas sobre un tema (polarización ideológica) o a la antipatía por el grupo partidista (polarización afectiva)? Además, también es posible que las diferencias entre los estudios empíricos se deban a que las redes sociales tienen efectos heterogéneos en los distintos grupos de personas, sobre todo en lo que respecta a su orientación política y a la fuerza de sus identidades partidistas.

2.3 Conclusiones

La investigación resumida en este capítulo ha hecho avanzar nuestra comprensión de cómo el éxito de las plataformas de redes sociales está transformando la estructura y heterogeneidad de las conversaciones políticas y su consiguiente impacto en la polarización política. Como hemos visto, en muchos casos las pruebas empíricas han desafiado lo que se creía a priori, mientras que en otros casos la teoría puede ayudarnos a reconciliar hallazgos aparentemente contradictorios. Aunque se han realizado avances consolidados, todavía quedan muchas cuestiones abiertas. Esta sección final ofrece una lista detallada de lo que no sabemos (todavía).

¿Son generalizables los resultados aquí descritos a otros contextos más allá de Estados Unidos? Reflejando el estado de la literatura sobre polarización política de forma más general, la mayor parte de lo que sabemos, exceptuando estudios concretos, se basa únicamente en datos de Estados Unidos. Hay una falta de teoría y evidencia empírica sobre cómo los factores contextuales median en la relación entre el uso de los medios sociales y la polarización política desde una perspectiva comparativa. Es importante mencionar que esta cuestión puede conducir a resultados mucho peores, incluida la violencia política. Por ejemplo, un documento de trabajo ha vinculado la incitación al odio difundida en las redes sociales con los ataques contra los refugiados en Alemania [95]. Esto ilustra la urgente necesidad de investigar cómo la difusión de ideas extremistas en las redes sociales podría estar motivando la violencia *offline*.

¿Existe alguna variación en el tiempo respecto a los efectos polarizadores de las interacciones en los medios sociales? Están mejorando o empeorando las cosas? Los servicios de Internet están en constante evolución, tanto en términos de qué plataformas de medios sociales o sitios web se vuelven populares como con respecto a las características de esas plataformas. Aunque Facebook existe desde hace más de diez años, ha evolucionado significativamente durante este periodo. El uso de las plataformas también puede cambiar:

Al principio, Twitter era solo un sitio web en el que los usuarios publicaban actualizaciones en tiempo real sobre sus actividades cotidianas. Actualmente se utiliza, entre otras cosas, para las noticias de última hora. Es cierto que algunos de los hallazgos de este capítulo se refieren a mecanismos que son un componente central del comportamiento humano, por ejemplo, el procesamiento sesgado de la información o la estructura de las redes sociales. Sin embargo, estudiar un entorno en constante cambio puede significar una incapacidad para comprender las condiciones de alcance en las que se mantienen esos mecanismos

Las consideraciones éticas también deben formar parte de este debate, especialmente cuando los académicos centran sus esfuerzos en estudiar cómo el extremismo alimentado por las interacciones en las redes sociales puede conducir a la violencia *offline*. ¿Cuáles son las consecuencias imprevistas de las posibles intervenciones para reducir la polarización? El trabajo de Mutz [97] proporciona una ilustración clara de los desafíos que enfrentamos en este sentido. Por un lado, la exposición a ideas contrarias puede fomentar la tolerancia política, promoviendo así un ambiente de debate saludable. Sin embargo, también existe la posibilidad de que esta exposición haga que la política sea más compleja y menos interesante para algunos usuarios, lo que podría resultar en un menor compromiso cívico y acentuar la desigualdad política.

El estudio de estas complejas relaciones multicausales es esencial para abordar la cuestión que tratamos: cómo las tecnologías digitales están impactando en la política democrática. Es crucial comprender cómo la exposición a diferentes perspectivas y la forma en que se presentan en las plataformas pueden influir en la polarización política, la participación ciudadana y la calidad del debate público.

Mediante el análisis de estas relaciones, podremos identificar posibles soluciones y estrategias para mitigar los efectos negativos de la polarización política en el contexto digital. Esto implica considerar cuidadosamente los enfoques de moderación, equilibrando la promoción de un ambiente inclusivo y respetuoso con la diversidad de opiniones, sin comprometer el interés y el compromiso de los usuarios.

En resumen, estudiar cómo las tecnologías digitales afectan a la política democrática es una cuestión de gran relevancia en la actualidad. La preocupación por la polarización política ha impulsado el debate sobre los posibles cambios normativos y las características de las plataformas digitales. Comprender las complejas relaciones entre la exposición a diferentes perspectivas políticas y sus impactos sociales es fundamental para abordar estos desafíos y promover una participación ciudadana informada y comprometida en el entorno digital.

CAPÍTULO 3
Discurso de odio en redes sociales.

El discurso del odio en línea se ha hecho cada vez más visible en las principales plataformas de medios sociales. Ante el temor de que esta retórica nociva incite a la violencia e impulse el extremismo, los gobiernos de todo el mundo están aprobando normativas y presionando a las empresas de redes sociales para que apliquen políticas que frenen la propagación del discurso del odio en línea [123].

Este capítulo pretende examinar el estado del arte (incluida la investigación científica, los estudios jurídicos y los informes políticos) sobre el discurso del odio en Internet. En particular, exploramos los debates en curso y las limitaciones de los enfoques actuales para definir y detectar el discurso de odio en línea. Proporcionamos una visión general de lo que los datos de las redes sociales y las encuestas pueden decirnos acerca de los productores, los objetivos y la prevalencia del lenguaje ofensivo. Además revisamos la evidencia empírica de las consecuencias offline del discurso de odio en línea; y ofrece ideas cuantitativas sobre qué intervenciones podrían ser más eficaces en la lucha contra el lenguaje ofensivo en línea.

3.1 Definición de la incitación al odio en línea.

No existe una única definición consensuada de incitación al odio en línea o en Internet y el tema ha sido objeto de debate entre académicos, juristas y responsables políticos similares. Por lo general, el discurso de odio se entiende como el lenguaje motivado por prejuicios, hostil y malicioso dirigido a una persona o grupo debido a sus características innatas reales o percibidas [37] [53]. Sin embargo, como sostiene Sellars [117], "a pesar de toda la extensa literatura sobre las causas, los daños y las respuestas al discurso de odio, pocos estudiosos se han esforzado por definir sistemáticamente el término."

Una amplia variedad de contenidos puede o no encajar en una definición de discurso de odio, dependiendo del contexto [117]. Por ejemplo, mientras que las calumnias y los insultos son fácilmente identificables, el lenguaje que contiene epítetos puede no ser necesariamente considerado discurso de odio por el hablante o el destinatario [47]. Por el contrario, el uso de lenguaje sutil puede pasar desapercibido ya que puede ser más difícil de identificar por alguien que no conozca la jerga. Esto es especialmente cierto en el entorno de las redes sociales, donde el discurso evoluciona rápidamente y puede ser altamente especializado. El uso de sustitutos de insultos también es común en las comunidades en línea, lo que complica aún más la definición del discurso de odio [50]. Por ejemplo, en América, entre los miembros de la ultra derecha, los periodistas han documentado el uso del término "googles" para referirse a los "negros"; "skypes" como un insulto antisemita; "yahoos" como un término despectivo para los hispanos; y "skittles"

como un término antimusulmán. De este modo, se complica definir el discurso de odio (y el discurso de odio en línea en particular).

Como resultado, las definiciones existentes de discurso de odio pueden ser extremadamente amplias o bastante limitadas. Por un lado están las definiciones que abarcan una amplia variedad de discursos dirigidos contra un individuo o grupo específico. Por otro lado se encuentran las definiciones que requieren un daño intencionado. Las definiciones más estrictas implican que el discurso del odio debe ser un "discurso peligroso", es decir, un lenguaje directamente relacionado con la incitación a la violencia masiva o al daño físico contra un grupo externo [83]. Esta tensión refleja la dificultad de desarrollar una definición que aborde adecuadamente el abanico de fenómenos que podrían considerarse incitación al odio, sin perder valiosas distinciones. El discurso del odio en línea puede implicar objetivos, motivos y tácticas dispares. El discurso que incita a la violencia es distinto del discurso que es "simplemente" ofensivo, y el uso de lenguaje dañino por un solo atacante es muy diferente de las campañas de odio coordinadas llevadas a cabo por una comunidad [117]. Trabajos recientes tratan de desarrollar definiciones más completas para identificar el discurso de odio que proporcionen contexto y tengan en cuenta las diferencias en gravedad e intención. Sin embargo, a pesar de estos avances, todavía no hay consenso en la literatura científica sobre cómo definir el discurso de odio en línea.

Tampoco hay un consenso respecto a las definiciones legales de la incitación al odio. Los Gobiernos están definiendo cada vez más la incitación al odio en sus códigos penales en un intento de regular directamente la retórica dañina tanto dentro como fuera de la red. Al igual que con las definiciones académicas, estas van desde las relativamente amplias, como la caracterización de Canadá del discurso del odio como el lenguaje que "promueve deliberadamente el odio contra cualquier grupo identificable", a definiciones más estrechas, como el marco de la Unión Europea, que define el discurso del odio como: "La incitación pública a la violencia o al odio dirigida contra un grupo de personas o un miembro de dicho grupo definido por motivos de raza, color, ascendencia, religión o creencia, u origen nacional o étnico" y "la condonación pública, negar o trivializar gravemente los crímenes de genocidio, crímenes contra la humanidad y crímenes de guerra (tal como se definen en la legislación de la UE), cuando la conducta se lleve a cabo de una manera que pueda incitar a la violencia o al odio contra dicho grupo o un miembro de dicho grupo" [117]. En el Reino Unido, incitar al odio racial o religioso es un delito penal, y existen variaciones de esta legislación. En la mayoría de las democracias desarrolladas, como Australia, Dinamarca, Francia, Alemania, India, Sudáfrica, Suecia y Nueva Zelanda [73], y en contextos autoritarios, en particular en el mundo árabe, donde las leyes que prohíben la incitación al odio en línea a menudo se agrupan con las leyes contra el extremismo [36]. Sin embargo, a pesar de la existencia de leyes que prohíben explícitamente la incitación al odio, la forma en que estas leyes deben aplicarse en la práctica, en particular en la era digital, es un tema de debate permanente.

Más recientemente, las propias plataformas en línea han desarrollado definiciones de incitación al odio con el fin de moderar los contenidos generados por los usuarios. Por ejemplo, la sección de "contenido de odio" de las Directrices de la Comunidad de YouTube establece que "no apoyamos contenidos que promuevan o condonen la violencia contra individuos o grupos por motivos de raza u origen étnico, religión, discapacidad, género, edad, nacionalidad, condición de veterano u orientación sexual/identidad de género, o cuyo propósito principal sea incitar al odio sobre la base de estas características básicas" [106]. Del mismo modo, las condiciones de servicio de Twitter establecen que la empresa prohíbe las "conductas de odio", entre las que se incluyen "promover la violencia contra otras personas o atacarlas o amenazarlas directamente por motivos de raza, etnia, origen nacional, orientación sexual, género, identidad de género, afiliación religiosa, edad, discapacidad o enfermedad." La compañía también hace hincapié en que

no permite cuentas cuyo "objetivo principal sea incitar al daño hacia otras personas en función de estas categorías" [126].

En conjunto, esta ausencia de definiciones claras y coherentes de la incitación al odio en la investigación académica, los estudios jurídicos, y entre los actores que intentan gobernar los espacios en línea ha significado que, a pesar de la amplia investigación, y las intervenciones políticas bien documentadas, nuestro conocimiento de las causas, consecuencias y medios eficaces para combatir la incitación al odio en línea sigue siendo un reto por la ambigüedad de la definición.

3.2 Detección de la incitación al odio en línea.

Al igual que no existe un consenso claro sobre la definición de discurso de odio, tampoco lo hay sobre la forma más eficaz de detectarlo en diversas plataformas. La mayoría de los enfoques automatizados para identificar el discurso de odio comienzan con una tarea de clasificación binaria en la que los investigadores se ocupan de codificar un documento como "discurso de odio o no", aunque también se han utilizado enfoques multiclase [42].

La detección automatizada del discurso del odio tiende a basarse en el procesamiento del lenguaje natural o en estrategias de minería de textos [57]. El más simple de estos enfoques son los métodos basados en diccionarios, que implican el desarrollo de una lista de palabras que se buscan y se cuentan en un texto. Los enfoques basados en diccionarios generalmente utilizan palabras de contenido para identificar el discurso de odio [86]. Reconociendo que el discurso de odio en línea puede ocultar palabras ofensivas usando errores ortográficos accidentales o intencionales, algunos investigadores han utilizado métricas de distancia, como el número mínimo de ediciones necesarias para transformar un término en otro (distancia de Levenshtein), para aumentar sus métodos basados en diccionarios [131].

Más allá de los métodos basados puramente en diccionarios, la mayoría de las técnicas más avanzadas de detección de discursos de odio implican tareas de clasificación de texto supervisadas. Estos enfoques, como el uso de clasificadores Naive Bayes, máquinas de vectores de soporte lineal (SVM), árboles de decisión o modelos de *ramdon tree*, a menudo se basan en técnicas de "bolsa de palabras" y "n-gramas". En el método de bolsa de palabras, se crea un corpus basado en las palabras que aparecen en un conjunto de datos de entrenamiento, en lugar de un diccionario predefinido. Las frecuencias de las palabras que aparecen en el texto, que ha sido anotado manualmente como "discurso de odio o no", se utilizan entonces como características para entrenar un clasificador [67] [31]. Para evitar errores de clasificación, si las palabras se utilizan en diferentes contextos o se escriben incorrectamente, algunos investigadores utilizan n-gramas, un enfoque similar a la bolsa de palabras, que combina palabras secuenciales en bigramas, trigramas o listas de palabras de longitud "n" [31][42][132]. Trabajos más recientes ha aprovechado estos enfoques para mejorar la precisión de los métodos basados en diccionarios, eliminando falsos positivos al identificar qué tuits que contienen insultos deberían clasificarse como incitación al odio [120]. También se han utilizado enfoques basados en reglas y patrones gramaticales temáticos, que incorporan la estructura de las oraciones [58]

Otros han incorporado el sentimiento a su análisis, asumiendo que el discurso del odio suele tener un tono negativo [42][46] .También se han utilizado técnicas de representación de palabras o representaciones vectoriales de texto, como doc2vec, paragraph2vec y FastText [49] [116][120], y las técnicas de aprendizaje profundo que emplean redes neuronales se han vuelto más comunes tanto para la clasificación de textos como para el análisis de sentimientos relacionados con la detección del discurso del odio [136] [137].

Reconociendo que estas técnicas pueden no ser adecuadas para identificar formas sutiles o indirectas de odio en línea, los investigadores también han empleado enfoques más motivados teóricamente. Por ejemplo, [31] Burnap et al. incorporan el concepto denomidado como "nosotros contra ellos" en su medición del discurso de odio. Encuentran que el discurso de odio a menudo utiliza pronombres en tercera persona, incluyendo expresiones como "envíenlos a todos a casa". Otros estudios han incorporado declaraciones de superioridad dentro de un grupo (además de ataques dirigidos a grupos) en sus mediciones [131]. Otro enfoque consiste en tener en cuenta los estereotipos comunes contra los grupos. Por ejemplo, el discurso contra los hispanos, podría hacer referencia al cruce de fronteras, o el lenguaje antisemita podría referirse a la banca, el dinero o los medios de comunicación [8]. Trabajos adicionales han distinguido entre el discurso de odio dirigido a un grupo (discurso de odio generalizado) y el discurso de odio dirigido a individuos (discurso de odio dirigido) para capturar matices importantes en los objetivos del discurso de odio en línea [51]. Más allá de basarse en características textuales, los investigadores también han incorporado características del usuario, incluidas características de red y recuentos de amigos/seguidores para mejorar la precisión de la detección del discurso de odio [127].

A pesar de estos importantes avances en la detección automática de la incitación al odio en Internet, los métodos existentes no se han probado en múltiples plataformas ni en diversos tipos de incitación al odio. Debido a la facilidad de recopilación de datos, la mayoría de los estudios existentes se han basado en datos de Twitter. Aunque otros trabajos han incorporado datos de Reddit, YouTube, Facebook, Whisper, Tumblr, Myspace, Gab, las secciones de comentarios de sitios web y blogs, son relativamente escasos [57]. Además, la gran mayoría de los estudios examinan el contenido en inglés, aunque algunos investigadores han desarrollado métodos para detectar el discurso de odio en otros idiomas.

Además, la mayoría de los estudios sobre el discurso de odio en línea buscan detectar todos los tipos de discurso de odio a la vez, o "discurso de odio general" [57]. Sin embargo, otros trabajos han examinado tipos específicos de lenguaje dañino, incluido el discurso de odio yihadista [44], el discurso de odio sectario [120], el discurso de odio antimusulmán [103], el discurso de odio contra los negros [21] o el discurso de odio misógino [21].

3.3 Lucha contra la incitación al odio en internet

La creciente preocupación por los efectos reales de la incitación al odio en Internet ha llevado a investigadores, responsables políticos y plataformas en línea a desarrollar estrategias para combatirla. Estos enfoques han adoptado generalmente dos formas: moderación de contenidos y contra-discurso.

Una estrategia para combatir el discurso del odio en línea ha sido moderar el contenido, lo que implica prohibir cuentas o comunidades que infrinjan las condiciones de servicio de las plataformas o las normas establecidas [78]. El 31 de mayo de 2016, la Comisión Europea, junto con Facebook, Twitter, YouTube y Microsoft, emitió un Código de Conducta voluntario para contrarrestar la incitación ilegal al odio en línea que exigía la eliminación de cualquier incitación al odio, tal como la define la Unión Europea (UE). Esta medida fue impulsada por el temor a un aumento del discurso intolerante contra los refugiados, así como por la preocupación de que el discurso del odio alimente los atentados terroristas [12]. Además, a partir de diciembre de 2017, ante la presión tras la marcha "Unite the Right" (Unir a la derecha) de agosto de 2017 en Charlottesville (Virginia), Twitter anunció una nueva política para prohibir cuentas afines a grupos "que

utilizan o promueven la violencia contra civiles para promover sus causas"[104]. La plataforma empezó suspendiendo varias cuentas con muchos seguidores implicadas en el nacionalismo blanco o en la organización de la marcha de Charlottesville. La empresa anunció que su prohibición de las amenazas violentas también se ampliaría para incluir cualquier contenido que glorifique la violencia [126]. Del mismo modo, en abril de 2018, Facebook anunció su conjunto de veinticinco páginas de normas que dictan qué tipos de contenido están permitidos en Facebook. La sección sobre incitación al odio afirma: "No permitimos la incitación al odio en Facebook porque crea un entorno de intimidación y exclusión y, en algunos casos, puede promover la violencia en el mundo real." El objetivo de prohibir el discurso de odio en las plataformas en línea más convencionales es reducir la probabilidad de que los usuarios cotidianos de Internet se vean expuestos incidentalmente al discurso de odio en línea.

Sin embargo, se sabe poco sobre cómo se aplican estas prohibiciones en la práctica o sobre su eficacia para reducir la incitación al odio en línea en estas plataformas o la exposición a este tipo de incitación en general. Además, el uso de la detección automática de la incitación al odio ha sido objeto de críticas en los medios de comunicación, ya que los límites de estos métodos se han puesto de manifiesto por errores embarazosos, como cuando los filtros patentados de Facebook etiquetaron un extracto de la Declaración de Independencia como incitación al odio [52]. Aunque una revisión de febrero de 2019 de la Comisión Europea sugiere que las plataformas de redes sociales, incluidas Facebook y Google, estaban eliminando con éxito el 75 % de las publicaciones etiquetadas por los usuarios que violan las normas de la UE en un plazo de 24 horas, no sabemos qué parte del discurso de odio se etiqueta o cómo esto puede estar sesgado en contra o a favor de ciertos tipos de discurso político[81].

Los trabajos empíricos sobre la eficacia de prohibir contenidos que inciten al odio arrojan resultados dispares. Al estudiar el efecto de la prohibición de los subreddits */fatpeoplehate* y */CoonTown* en Reddit en 2015, Chandrasekharan, Pavalanathan et al. [35] concluyen que la prohibición tuvo éxito. Analizando más de 100 millones de publicaciones y comentarios en Reddit, los autores descubrieron que muchas cuentas dejaron de utilizar el sitio tras la prohibición, y que las que se quedaron redujeron su uso de la incitación al odio en al menos un 80 %. Aunque muchos de estos usuarios migraron a otros subreddits, los nuevos subreddits no experimentaron un aumento en el uso de la incitación al odio, lo que sugiere que la prohibición tuvo éxito a la hora de limitar la incitación al odio en línea en Reddit. Del mismo modo, otros trabajos sugieren que la prohibición de cuentas en Twitter altera las redes sociales extremistas, ya que los usuarios que son prohibidos con frecuencia sufren importantes caídas en el número de seguidores cuando vuelven a unirse a una plataforma en particular [25].

Dicho esto, aunque es posible que las prohibiciones hayan reducido el volumen general de incitación al odio en Reddit y hayan interrumpido la actividad extremista en Twitter, es posible que dicha actividad simplemente haya migrado a otras plataformas. Newell et al. [101] descubrieron que los usuarios tóxicos descontentos por las prohibiciones buscaron plataformas alternativas como *Voat, Snapzu y Empeopled*. Los usuarios que migran a estas plataformas marginales a menudo conservan sus nombres de usuario e intentan recrear sus comunidades prohibidas en un nuevo dominio menos regulado [35]. Además de trasladar la incitación al odio de una plataforma a otra, otros trabajos sugieren que los productores de contenidos nocivos simplemente se vuelven más creativos sobre cómo seguir utilizando la incitación al odio en sus plataformas preferidas. Por ejemplo, tratando de evitar la moderación de contenidos, como se ha descrito anteriormente, los miembros de las comunidades en línea a menudo utilizan palabras sustitutivas de insultos para eludir la detección [50].

Además, los intentos de prohibir cuentas de usuarios pueden ser a veces contraproducentes, al galvanizar el apoyo de quienes simpatizan con las comunidades de odio. Cuando usuarios muy conocidos son atacados, las personas con creencias similares pueden verse motivadas a unirse en su defensa o a expresar opiniones a las que se oponen empresas u organizaciones poderosas. Por ejemplo, los estudios empíricos sobre el comportamiento extremista en línea que examinan las cuentas favorables a ISIS sugieren que los extremistas en línea ven el bloqueo de sus cuentas como un acto de violencia, una insignia de honor, y los individuos que han sido bloqueados o vetados a menudo son capaces de reactivar sus cuentas con nuevos nombres [25]. Además, la prohibición de usuarios a menudo les lleva a trasladarse a plataformas más especializadas, como *Gab* o *Voat*, que pueden radicalizar aún más a los individuos que producen odio en línea. De hecho, prohibir a los usuarios que incitan al odio los aleja de diversos entornos en los que pueden entrar en contacto con voces moderadas u opuestas, elevando sus agravios y sentimientos de persecución y empujándolos a cámaras de eco de odio en las que el extremismo y los llamamientos a la violencia *offline* se normalizan y fomentan [89]. Si bien este es un argumento teórico convincente contra la prohibición de usuarios de plataformas convencionales, se necesita más trabajo empírico para rastrear hasta qué punto los usuarios prohibidos migran a plataformas más extremas, así como si de hecho se radicalizan aún más en estas plataformas [89].

De este modo, el trabajo empírico existente sobre la eficacia de la moderación de contenidos sugiere que, si bien puede reducir la incitación al odio en determinadas plataformas, ya que los usuarios descontentos migran a otros rincones de Internet, no está claro si tales esfuerzos reducen la incitación al odio en general. Por otra parte, persisten cuestiones jurídicas, éticas y técnicas en relación con los beneficios de prohibir la incitación al odio en las plataformas de medios sociales globales, en particular fuera de las democracias occidentales. Por ejemplo, una reciente investigación de ProPublica descubrió que las normas de Facebook no son transparentes y se aplican de forma incoherente por decenas de miles de contratistas globales encargados de la moderación de contenidos. En muchos países y territorios en disputa, como los territorios palestinos, Cachemira y Crimea, activistas y periodistas han sido censurados por expresiones nocivas, ya que Facebook ha respondido a las preocupaciones del gobierno y ha trabajado para aislarse de la responsabilidad legal. El informe concluye que las normas de moderación de contenidos de incitación al odio de Facebook "tienden a favorecer a las élites y los gobiernos frente a los activistas de base y las minorías raciales". En esta línea, los gobiernos pueden declarar que el discurso de la oposición es odioso o extremista con el fin de manipular la moderación de contenidos para silenciar a sus críticos [81].

Reconociendo que la censura de la incitación al odio puede entrar en conflicto con las protecciones legales de la libertad de expresión o puede ser manipulada por los gobiernos para atacar a los críticos, los organismos internacionales como la UNESCO han mantenido generalmente que "la libre circulación de la información debe ser siempre la norma". En consecuencia, a menudo sostienen que el contra-discurso suele ser preferible a la supresión del discurso de odio [59]. El contra-discurso es una respuesta directa al odio, destinado a influir en el discurso y el comportamiento [23].

En trabajos más recientes se ha explorado el uso del contra-discurso en la esfera en línea. Por ejemplo, ante el temor a la violencia en vísperas de las elecciones keniatas de 2013, ONGs internacionales, famosos y empresas locales ayudaron a financiar "campañas de la paz" para impedir la propagación de la incitación al odio en Internet (y de la violencia online) en Kenia. Por ejemplo, una empresa ofreció dinero y tiempo de uso del teléfono móvil a los kenianos que se enviaran mensajes de paz en línea, incluyendo fotos, poemas e historias [22].

Una incipiente corriente de literatura evalúa experimentalmente qué tipos de mensajes de contra-discurso son más eficaces para reducir el discurso de odio en línea. Munger [96] muestra que el contra-discurso mediante bots automatizados puede reducir los casos de discurso racista si los instigadores son sancionados por un miembro del grupo de alto estatus, en este caso, un hombre blanco con un gran número de seguidores en Twitter. De forma similar, Siegel y Badaan [120] utilizaron una cuenta ficticia para contrarrestar el discurso de odio sectario en la Twitteresfera árabe. Descubren que el mero hecho de recibir un mensaje sancionador reduce el uso del discurso de odio, sobre todo para los usuarios de redes en las que el discurso de odio es relativamente infrecuente. Se necesita más investigación para evaluar qué tipos de contra-discurso y de qué fuentes son más eficaces para reducir el odio en línea en diversos contextos. Reconociendo el potencial de los bots para contrarrestar el discurso, Leetaru [84] propuso desplegar bots de IA en masa para luchar contra el discurso de odio en línea, aunque la viabilidad y las consecuencias de tal intervención no se conocen bien. En uno de los únicos estudios que detecta explícitamente el contra-discurso que se produce de forma natural en las redes sociales, [90] encuentran que los comentarios de contra-discurso reciben muchos más "me gusta" y participación que otros comentarios y pueden incitar a los productores de discurso de odio a disculparse o cambiar su comportamiento. Sin embargo, se necesita más trabajo empírico para ver cómo se desarrolla esta dinámica de forma más sistemática en las plataformas de redes sociales del mundo real a lo largo del tiempo.

Álvarez-Benjumea y Winter [9] comparan explícitamente la censura o la supervisión de contenidos con las intervenciones contra el discurso y comprueban si la disminución de la aceptabilidad social de los comentarios hostiles en un foro en línea reduce el uso del discurso de odio. Primero diseñaron un foro en línea e invitaron a los participantes a unirse y participar en conversaciones sobre temas sociales actuales. A continuación, manipularon experimentalmente los comentarios que los participantes observaban antes de publicar sus propios comentarios. Incluyeron un tratamiento de censura en el que los participantes no observaron ningún comentario de odio y un tratamiento de contra-discurso en el que los comentarios de incitación al odio no fueron censurados, pero se presentaron junto a mensajes que destacaban el hecho de que la incitación al odio no se consideraba aceptable en la plataforma. Al comparar el nivel de hostilidad de los comentarios y los casos de odio en las distintas condiciones de tratamiento, descubrieron que el tratamiento de censura era el más eficaz para reducir los comentarios hostiles. Sin embargo, los autores señalan que el hecho de que no observen un efecto estadísticamente significativo del tratamiento contra el discurso puede deberse al pequeño tamaño de sus muestras y a la incapacidad de controlar las interacciones repetidas a lo largo del tiempo en su configuración experimental.

En conjunto, esta revisión sobre los efectos de la censura y el contra-discurso en el discurso de odio en línea proporciona cierto optimismo, en particular con respecto al impacto de la moderación de contenidos en la reducción del discurso de odio en las plataformas convencionales y la capacidad de las campañas de contra-discurso para disminuir el alcance, la visibilidad y el daño del discurso de odio en línea. Sin embargo, sabemos muy poco sobre los posibles daños colaterales de estas intervenciones. El trabajo futuro no sólo debería proporcionar pruebas empíricas a mayor escala de este tipo de intervenciones en diversos contextos, sino que también debería tratar de evaluar los efectos a largo plazo de estos enfoques.

3.4 Conclusiones

A pesar de la creciente atención prestada al discurso de odio en línea, como demuestra este capítulo, el debate sobre cómo definir el discurso de odio en línea está lejos de resolverse. En parte como consecuencia de estos desafíos definitorios, y en parte como resultado de la naturaleza altamente contextual y evolutiva del discurso del odio en línea, detectar sistemáticamente el contenido de odio es una tarea extremadamente difícil.

Aunque las técnicas más avanzadas que emplean el aprendizaje automático, las redes neuronales y la incorporación de características contextuales han mejorado nuestra capacidad para medir y vigilar el discurso de odio en línea, la mayoría de los trabajos empíricos existentes están bastante fragmentados y a menudo detectan un único tipo de discurso de odio en una plataforma en un momento dado. Por otra parte, debido a la facilidad de recopilación de datos, la gran mayoría de los estudios se han realizado utilizando datos de Twitter en inglés y, por lo tanto, no necesariamente nos dicen mucho acerca de otras plataformas o contextos culturales. Para añadir más complicaciones, las definiciones de discurso de odio y los enfoques para detectarlo están muy politizados, sobre todo en contextos autoritarios y contextos conflictivos. Aunque algunas investigaciones han explorado múltiples tipos de discurso de odio, utilizado varios conjuntos de datos, realizado investigaciones en múltiples plataformas o examinado las tendencias en el discurso de odio a lo largo del tiempo, estos estudios son la excepción y no la regla [58]. Basándose en la rica literatura de técnicas de detección del discurso de odio en informática y ciencias sociales, el trabajo futuro debería intentar un análisis comparativo más sistemático para mejorar nuestra capacidad de detectar el discurso de odio en línea en sus diversas formas.

Los estudios científicos también han evaluado qué estrategias podrían ser más eficaces para combatir la incitación al odio en línea. La evidencia empírica sugiere que la prohibición de comunidades de odio en Reddit, por ejemplo, redujo el volumen de discurso de odio en la plataforma en general [35]. Sin embargo, otros trabajos indican que los usuarios a los que se prohíbe debatir sobre determinados temas en plataformas convencionales simplemente se trasladan a otro lugar para continuar con su discurso de odio [101]. Desde un punto de vista más optimista, las investigaciones experimentales en las que se utiliza el contra-discurso para combatir la incitación al odio en línea sugieren que recibir mensajes sancionadores de otros usuarios de Twitter (en particular de compañeros de grupo, individuos de alto estatus en la comunidad) disuade a los usuarios de tuitear contenidos de odio [96][120]. Además, los estudios empíricos a gran escala sugieren que el contra-discurso es bastante común en la esfera en línea, y los mismos eventos que desencadenan repuntes en el discurso de odio en línea a menudo desencadenan aumentos mucho mayores en el contra-discurso [103] El trabajo futuro debe continuar explorando qué tipos de contra-discurso podrían ser más eficaces en diversos contextos culturales y en diferentes plataformas, así como la forma en que se puede fomentar el contra-discurso entre los usuarios cotidianos de las redes sociales. Dadas las peligrosas consecuencias offline del discurso de odio en línea en diversos contextos globales, los académicos y los responsables políticos deben seguir construyendo sobre esta literatura existente para mejorar la detección del discurso de odio, obtener una comprensión más completa de cómo surge y se propaga el discurso de odio, desarrollar una mayor comprensión de las consecuencias offline del discurso de odio y construir mejores herramientas para combatirlo eficazmente.

CAPÍTULO 4
Dataset generado y Análisis de Temas Políticos.

Este capítulo aborda el análisis de la actualidad política en España utilizando técnicas de procesamiento de lenguaje natural, específicamente el enfoque de Topic Modelling con el modelo GDSMM (*Gibbs Sampling Dirichlet Mixture Model*). Además nos centraremos en cómo hemos realizado la recolección y análisis de retuits de los diputados del congreso en Twitter. Todo esto con el objetivo de obtener una panorámica general de los temas políticos relevantes en el país. Presentamos una metodología detallada, desde la selección de datos hasta la interpretación de resultados, con el fin de proporcionar una comprensión profunda y estructurada de la realidad política en España.

4.1 API de Twitter, Wikidata y Neo4j.

La actualidad política es un tema de interés y relevancia continua en cualquier país. Con el auge de las redes sociales, los diputados y líderes políticos utilizan plataformas como Twitter para compartir sus puntos de vista, discursos y noticias relevantes. En este contexto, el *topic modelling* con el modelo `gdsmm` se presenta como una poderosa herramienta para descubrir patrones y temáticas ocultas en grandes volúmenes de datos textuales, permitiendo obtener una visión panorámica de los temas políticos más relevantes y su interconexión. Existen varias posibilidades para obtener del API de Twitter[1] la información relevante acerca de la situación política en España. Podríamos utilizar *Hashtags* relevantes sobre la política como #solosiessi o #feminismo para obtener tuits con una temática concreta, sin embargo, con este enfoque no logramos obtener una panorámica general. Por lo tanto, se nos ocurrió monitorizar a los diputados del congreso, pues quien mejor para representar los temas más actuales sobre la política que los propios diputados. Para monitorizar en Twitter a los diputados era necesario conocer el identificador de Twitter para cada diputado, esta labor podía haberse realizado a mano. Nosotros decidimos utilizar Wikidata[2], una base de conocimientos colaborativa y de código abierto que almacena datos estructurados y enlazados para mejorar la información en la web, como fuente.

Para obtener todos los IDs de Twitter de los diputados del Congreso de España desde Wikidata, utilizamos el lenguaje `SPARQL`, el cual es un lenguaje de consulta diseñado específicamente para interactuar con bases de conocimientos que siguen el modelo de datos de grafos, como Wikidata. `SPARQL` permite realizar consultas complejas y precisas pa-

[1] https://developer.twitter.com/en/docs/twitter-api
[2] https://www.wikidata.org

ra recuperar información específica de estas bases de conocimientos.Para obtener todos los IDs de Twitter de los diputados del Congreso de España desde Wikidata utilizando SPARQL, se puede utilizar la siguiente consulta:

```
SELECT ?diputado ?diputadoLabel ?twitter_id WHERE {
  ?diputado wdt:P31 wd:Q5.
  ?diputado wdt:P39 wd:Q18907131.
  OPTIONAL { ?diputado wdt:P2002 ?twitter_id. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "es,en". }
}
```

Ahora, explicaré cada parte de la consulta:

- `SELECT ?diputado ?diputadoLabel ?twitter_id`: Esta línea indica qué variables queremos recuperar en la consulta. En este caso, estamos solicitando tres variables: ?diputado (el identificador del diputado en Wikidata), ?diputadoLabel (el nombre del diputado) y ?twitter_id (el ID de Twitter del diputado).
- `WHERE { ... }`: Aquí especificamos las condiciones que deben cumplir los datos que queremos recuperar. En este caso, estamos buscando entidades que cumplan dos condiciones:
  - `?diputado wdt:P31 wd:Q5`: El diputado debe ser una instancia de "humano"(Q5 en Wikidata) para asegurarnos de que estamos obteniendo diputados y no otros tipos de entidades.
  - `?diputado wdt:P39 wd:Q18907131`: El diputado debe ocupar la posición de "diputado del Congreso de España"(Q18907131 en Wikidata).
- `OPTIONAL { ... }`: Esta parte es opcional y se utiliza para obtener datos adicionales que pueden no estar disponibles para todas las entidades que cumplen las condiciones anteriores. En este caso, estamos obteniendo el ID de Twitter del diputado usando la propiedad P2002.
- `SERVICE wikibase:label { bd:serviceParam wikibase:language "es,en". }`: Esta línea se usa para obtener etiquetas multilingües de los diputados en español e inglés.

Al ejecutar esta consulta SPARQL en Wikidata, obtendremos una lista de diputados del Congreso de España con sus respectivos IDs de Twitter (si están disponibles). La lista final de diputados puede ser consultada en el Apéndice B al final. Es importante mencionar que la disponibilidad de los IDs de Twitter puede variar dependiendo de la información proporcionada por los usuarios en Wikidata. Algunos diputados pueden no tener IDs de Twitter registrados en la plataforma. Un punto favorable a tener en cuenta es que Wikidata es una plataforma colaborativa y la información está en constante actualización, por lo que la consulta podría utilizarse para futuros diputados. Una vez tenemos todos los IDs de Twitter de los diputados del Congreso de España que se encuentren en Wikidata, podemos utilizarlo para realizar consultas al API de Twitter.

El API de Twitter dispone de diferentes operadores[3] que podemos utilizar para obtener exactamente aquello que queremos. Nosotros queríamos obtener todos los retuits que los diputados del congreso habían realizado. Utilizamos el retuit porque el retuit, como dice la literatura, es una interacción la cual el usuario suele utilizar cuando quiere expresar algo con lo que está a favor. La consulta que utilizamos para obtener los retuits es:

[3]https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

```
'''
queries[0] = "from:2610041328 OR from:2867230876 OR from:331085228 OR
↪ from:242430168...is:retweet lang:es"
'''
import json
for i in range(len(queries)):
    rule = gen_request_parameters(queries[i],
    ↪ tweet_fields="author_id,created_at,entities,public_metrics,text,lang,
      "in_reply_to_user_id",
      expansions="author_id,referenced_tweets.id",
      results_per_call=500,
      start_time="2022-01-01",
      end_time="2023-3-12"
     )

    from searchtweets import ResultStream

    rs = ResultStream(request_parameters=rule,
                      max_tweets=1000000,
                      **search_args)

    with open('/home/gti/jcarlos/TFM/results/twitter_diputados_en-12mar_rts.jsonl', 'a',
    ↪ encoding='utf-8') as f:
        for tweet in rs.stream():
            json.dump(tweet, f)
            f.write('\n')
    print('done')
```

Una vez recopilamos todos los datos del API en formato `JSON` procedimos a almacenarlos en una base de datos, en concreto, Neo4j[4]. Neo4j es una base de datos orientada a grafos que se utiliza para almacenar y gestionar datos estructurados en forma de grafos. A diferencia de las bases de datos tradicionales que utilizan el modelo de tablas y relaciones, Neo4j permite almacenar datos como nodos (entidades) conectados entre sí mediante relaciones (aristas). Esto hace que sea especialmente útil para representar y consultar datos altamente interconectados. Almacenar los tuits de Twitter en Neo4j resulta interesante por varias razones, entre ellas:

- **Modelo de datos flexible:** Neo4j permite modelar datos complejos y relaciones entre entidades de manera más natural, lo que facilita el almacenamiento y consulta de los tuits, usuarios y relaciones entre ellos.

- **Rendimiento en consultas complejas:** Las consultas en bases de datos de grafos son altamente eficientes para encontrar patrones y relaciones en datos interconectados, lo que facilita realizar análisis avanzados en grandes volúmenes de datos.

- **Descubrimiento de patrones y comunidades** Con la estructura de grafos, es más sencillo identificar patrones y comunidades dentro de los tuits y usuarios, lo que puede ser valioso para analizar tendencias, influencias y comportamientos en Twitter.

[4]https://neo4j.com/

- **Visualización de datos:** Neo4j cuenta con herramientas de visualización que permiten ver la estructura de grafos de forma gráfica, lo que facilita la comprensión y análisis de los datos.
- **Escalabilidad:** Neo4j es altamente escalable, lo que significa que puede manejar grandes cantidades de datos y crecer según las necesidades del análisis.
- **Interacciones sociales** Al utilizar Neo4j, se pueden implementar algoritmos análisis de interacciones sociales, lo que puede ser beneficioso para entender mejor el comportamiento de los usuarios en Twitter.

En resumen, almacenar tuits de Twitter en Neo4j ofrece ventajas significativas para realizar un análisis de datos más profundo y completo. La estructura de grafos permite descubrir relaciones complejas entre los tuits, usuarios y temas, facilitando la identificación de patrones, tendencias y comunidades en la plataforma. Esta flexibilidad y capacidad de consulta hacen que Neo4j sea una opción atractiva para proyectos de análisis de datos en Twitter y otras redes sociales.

Para almacenar los datos que habíamos obtenido del API de Twitter en formato JSON en la base de datos, tuvimos que realizar diferentes consultas en el lenguaje de Neo4j, `Cypher`:

1.- Creación de restricciones para que no haya nodos repetidos de tipo `User, Tweet, Retweet y Hashtag`.

```
//CONSTRAINTS
CREATE CONSTRAINT tweetCons IF NOT EXISTS FOR (n:Tweet) REQUIRE n.idTwitter IS UNIQUE
CREATE CONSTRAINT retweetCons IF NOT EXISTS FOR (n:Retweet) REQUIRE n.idTwitter IS UNIQUE
CREATE CONSTRAINT userCons IF NOT EXISTS FOR (n:User) REQUIRE n.id IS UNIQUE
CREATE CONSTRAINT hashtagCons IF NOT EXISTS FOR (n:Hashtag) REQUIRE n.text IS UNIQUE
```

2.- Creación de los nodos de tipo `User, Tweet, Retweet y Hashtag`.

```
//Create users
CALL apoc.periodic.iterate("CALL apoc.load.json('file:///twitter.jsonl') YIELD value",
"UNWIND value.includes.users as users \r\n"+
"UNWIND users as user \r\n"+
"MERGE (u:User {id:user.id}) \r\n"+
"ON CREATE SET u.idTwitter = user.id, u.name = user.name, u.username = user.username",
{batchSize:1, batchMode: "BATCH"});

//CREATE RT AND TWEETS data
CALL apoc.load.json('file:///twitter.jsonl') YIELD value as tweets
UNWIND tweets.data as tweet
UNWIND tweet.referenced_tweets as rf
CALL apoc.do.case([rf.type = "retweeted", "MERGE (u:Retweet {idTwitter:tweet.id}) ON
↪  CREATE SET u.text = tweet.text, u.idTwitter = tweet.id, u.lang = tweet.lang,
↪  u.idAuthor = tweet.author_id, u.rtCount = tweet.public_metrics.retweet_count,
↪  u.likeCount = tweet.public_metrics.like_count, u.replyCount =
↪  tweet.public_metrics.reply_count, u.created_at = datetime(tweet.created_at) Return
↪  u"],"MERGE (u:Tweet {idTwitter:tweet.id}) ON CREATE SET u.text = tweet.text,
↪  u.idTwitter = tweet.id, u.lang = tweet.lang, u.idAuthor = tweet.author_id, u.rtCount
↪  = tweet.public_metrics.retweet_count, u.likeCount = tweet.public_metrics.like_count,
↪  u.replyCount = tweet.public_metrics.reply_count, u.created_at =
↪  datetime(tweet.created_at) RETURN u" , {tweet:tweet,rf:rf}) yield value return
↪  value,tweet,rf
```

```
14
15 //Create hashtags
16 CALL apoc.load.json('file:///twitter.jsonl') YIELD value AS tweets
17 UNWIND tweets.data as tweet
18 UNWIND tweet.referenced_tweets as rf
19 UNWIND tweet.entities.hashtags as tags
20 CALL apoc.do.when(rf.type <> "retweeted","MERGE (h:Hashtag {text:tags.tag}) ON CREATE SET
   ↪ h.text = tags.tag","",{tweet:tweet,rf:rf, tags:tags}) YIELD value RETURN count(value)
```

3.- Creación de las relaciones:

- `RT`, la cual indica que un `User` ha retuiteado un `Tweet`
- `POSTED`, la cual indica que un `User` ha publicado un `Tweet`
- `Tagged`, la cual indica que un `Hashtag` ha sido referenciado en un `Tweet`

```
1  //CREATE RT RL AND TWEETS  RL data
2  CALL apoc.load.json('file:///twitter_diputados_2022_rts-Copy1.jsonl') YIELD value AS
   ↪ tweets
3  UNWIND tweets.data as tweet
4  UNWIND tweet.referenced_tweets as rf
5  CALL apoc.do.case([rf.type = "retweeted", "MATCH (u:Retweet), (t:Tweet) WHERE u.idTwitter
   ↪ = tweet.id AND t.idTwitter = rf.id MERGE (u)-[:RT]->(t)"],"MATCH (u:User), (t:Tweet)
   ↪ WHERE u.idTwitter = tweet.author_id AND t.idTwitter = tweet.id MERGE
   ↪ (u)-[:POSTED]->(t)",{tweet:tweet,rf:rf}) YIELD value return value
6
7  //CREATE RL POSTED data
8  CALL apoc.periodic.iterate("CALL
   ↪ apoc.load.json('file:///twitter_diputados_2022_rts-Copy1.jsonl') YIELD value",
9  "UNWIND value.includes.tweets as tweet \r\n"+
10 "MATCH (u:User), (t:Tweet)\r\n"+
11 "WHERE u.idTwitter = tweet.author_id AND t.idTwitter = tweet.id \r\n"+
12 "MERGE (u)-[:POSTED]->(t)",
13 {batchSize:1, batchMode: "BATCH"});
14
15 //CREATE RL TAGGED
16 CALL apoc.periodic.iterate("CALL
   ↪ apoc.load.json('file:///twitter_diputados_2022_rts-Copy1.jsonl') YIELD value",
17 "UNWIND value.includes.tweets as tweet \r\n"+
18 "UNWIND tweet.entities.hashtags as tags \r\n"+
19 "MATCH (h:Hashtag), (t:Tweet) \r\n"+
20 "WHERE h.text = tags.tag AND t.idTwitter = tweet.id \r\n"+
21 "MERGE (h)-[:TAGGED]->(t)",
22 {batchSize:1, batchMode: "BATCH"});
```

Con esto ya tendríamos nuestra base de datos no relacional lista para interactuar con ella y realizar análisis para nuestro estudio. Como muestra de la capacidad de visualización y análisis que por defecto lleva integrado neo4j, hemos realizado un análisis de los diputados más influyentes para cada partido. Para ello hemos ejecutado el algoritmo

de Page Rank con la plataforma Zoom de Neo4j, obteniendo como resultado las siguientes figuras 4.1 4.2 4.4. Podemos apreciar un mayor tamaño e intensidad de color en las figuras para aquellos diputados con mayor influencia en la red.

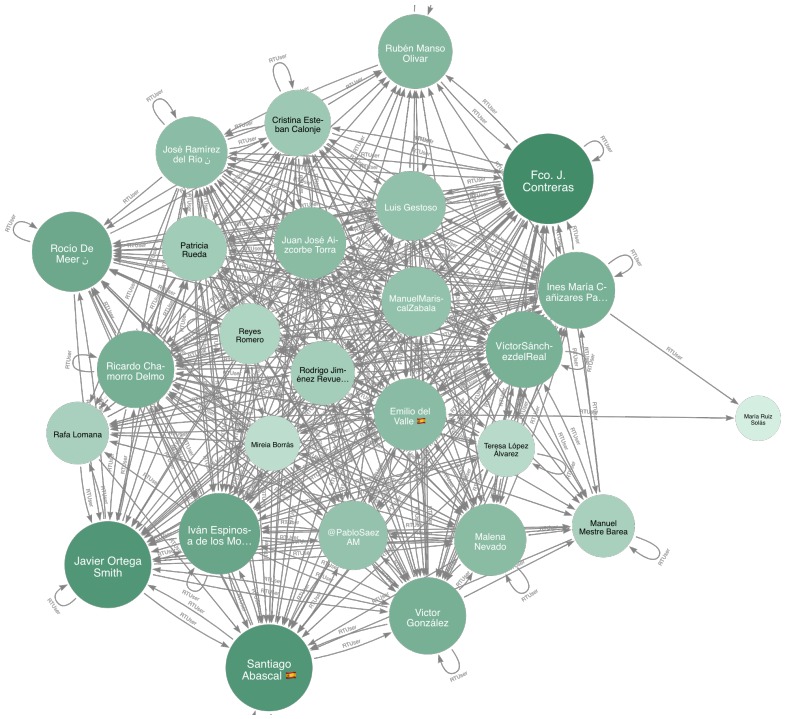

Figura 4.1: Diputados más influyentes de VOX.

4.2 Análisis de Temas Políticos en España mediante *Topic Modelling* en Twitter.

Podíamos habernos inmerso y haber estudiado el panorama político español, sin embargo decidimos utilizar un enfoque más automático. Por ello, en esta investigación realizada para este trabajo de fin de máster, abordamos el fascinante campo del análisis político a través de las redes sociales, centrándonos en Twitter como fuente de datos. El objetivo de esta sección es explicar cómo la técnica de "Topic Modelling" puede utilizarse para explorar los tuits de los diputados del Congreso de España y desentrañar los temas clave y las dinámicas políticas subyacentes. Al aprovechar el poder del procesamiento de lenguaje natural y el aprendizaje automático, buscamos identificar patrones y tendencias ocultas en la comunicación política, contribuyendo así a una comprensión más profunda del panorama político actual y su impacto en la opinión pública. De esta manera, tenemos una imagen del panorama político en España, mediante un análisis que puede replicarse en el futuro y llevar un paso más adelante este trabajo.

Topic Modelling (modelado de temas) (TM) es una técnica de procesamiento de lenguaje natural que se utiliza para descubrir patrones latentes en grandes conjuntos de datos textuales y organizarlos en temas coherentes y significativos. La idea detrás del modelado de temas es que los documentos que tratan sobre temas similares tienden a usar palabras similares o relacionadas entre sí. Por lo tanto, si podemos descubrir grupos de palabras que coocurren con frecuencia en los documentos, podemos inferir la existencia de temas subyacentes. El modelado de temas se considera una técnica de aprendizaje no supervisado, ya que no requiere ningún entrenamiento con datos ya clasificados ("etiquetados").

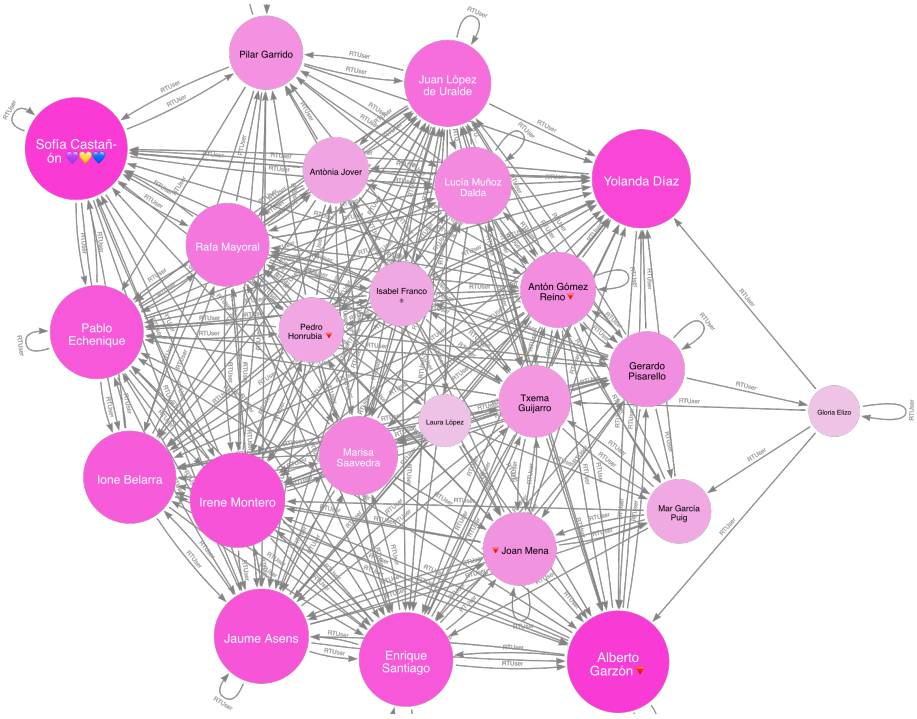

Figura 4.2: Diputados más influyentes de Podemos.

Figura 4.3: Diputados más influyentes del Partido Popular.

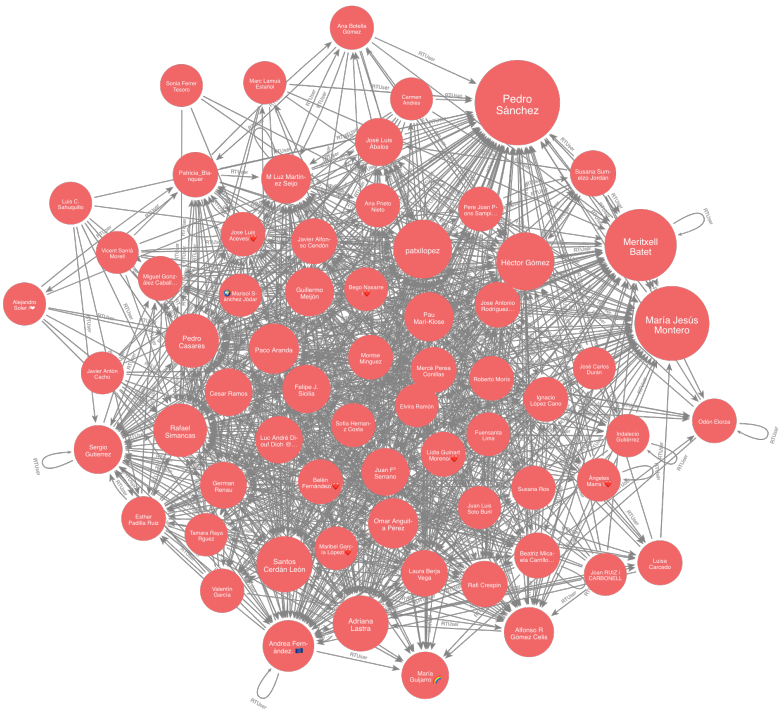

Figura 4.4: Diputados más influyentes del Partido Socialista.

Por lo tanto, el algoritmo no debe tener en cuenta ninguna característica específica de los temas antes del análisis.

Los algoritmos de TM reciben como entrada un corpus (es decir, un conjunto de documentos) y proporcionan como salida clusters (conjuntos de palabras). Cada uno de estos clusters representa un tema, es decir, un tema de discusión que aparece en los documentos.

4.2.1. Modelado de tópicos para texto corto.

Como han destacado varios investigadores (por ejemplo, [77] [113] [122]) los algoritmos de modelado de temas, (un ejemplo ampliamente utilizado, como LDA) que funcionan bien en textos de al menos unos cientos de palabras de longitud, funcionan mal cuando se aplican en textos cortos (como Tweets), ya que solo se dispone de información muy limitada de coocurrencia de palabras en textos cortos. Por lo tanto, se necesitan enfoques mejorados para abordar el problema de la escasez de información en los textos breves.

Últimamente han aparecido en la literatura métodos para tratar el "Short Text Topic Modelling" (en lo sucesivo: STTM). Jonsson y Stolee [77] informan de métodos basados en la agregación de textos (con el fin de formar textos más largos a partir de textos cortos), el *"biterm topic model"* (BTM) [134] y un modelo que agrupa vectores word2vec utilizando un modelo de mezcla gaussiana (GMM) [122]. Qiang et al. [113] presentan una revisión de varias técnicas de modelado de temas de texto corto propuestas, presentando tres categorías de métodos basados en la mezcla multinomial de Dirichlet, las coocurrencias globales de palabras y la autoagregación. Asimismo, Yin & Wang [135] presentan un método específico denominado "*Gibbs Sampling Dirichlet Mixture Model*" (`GSDMM`) para abordar el modelado de tópicos de textos cortos".

En este trabajo empleamos el algoritmo GSDMM para analizar textos cortos (tuits) y extraer los tópicos o temas sobre política de nuestro conjunto de datos.

GSDMM

El "Gibbs Sampling Dirichlet Mixture Model"(GSDMM) es un algoritmo LDA ajustado que, permite obtener mejores resultados en tareas STTM. Un supuesto inicial clave del algoritmo GSDMM es que cuando se trata de documentos cortos sólo hay un tema tratado en cada documento (supuesto de "un tema por documento").

La idea clave implementada por el algoritmo GSDMM se suele demostrar en la literatura utilizando una analogía simple reportada como "*Movie Group Process*" [135], que explicamos aquí de forma concisa.

Supongamos que tenemos un grupo de cinéfilos en un restaurante sentados aleatoriamente en K mesas. Pedimos a todos ellos que escriban en un papel una breve lista de sus películas favoritas. A continuación, los reubicamos en las K mesas con el objetivo de agruparlos de forma que los cinéfilos con las mismas películas favoritas estén sentados en la misma mesa. Para ello, les pedimos que elijan una nueva mesa -uno tras otro- aplicando las dos reglas siguientes:

1. Seleccione una mesa con más cinéfilos. Esta regla mejora la exhaustividad, todos los estudiantes que comparten el mismo interés por una película son asignados a la misma mesa.
2. Seleccionar una mesa en la que los cinéfilos compartan intereses similares. El objetivo de esta regla es aumentar la homogeneidad, ya que todos los alumnos que compartan el mismo interés por una película serán asignados a una misma mesa.

Después de aplicar repetidamente estos pasos, se espera que algunas tablas queden finalmente vacías, mientras que otras reúnen grupos más grandes de estudiantes que han coincidido con el interés de la película. Este es el proceso que implementa el algoritmo GSDMM. Yin y Wang [135] al presentar el algoritmo GSDMM informan de que "GSDMM *puede inferir el número de clústeres automáticamente con un buen equilibrio entre la exhaustividad y la homogeneidad de los resultados de la agrupación, y es rápido para converger.* GSDMM *también puede hacer frente al problema de la dispersión y la alta dimensionalidad de los textos cortos, y puede obtener las palabras representativas de cada conglomerado. Nuestro amplio estudio experimental muestra que* GSDMM *puede lograr un rendimiento significativamente mejor que otros tres modelos de agrupación.*"

4.2.2. Tópicos obtenidos con GSDMM

Metodología.

- Selección de Datos: Se recopiló un conjunto de retuits de los diputados del congreso en España desde un período determinado, considerando cuentas oficiales y líderes de los principales partidos políticos.
- Preprocesamiento de Datos: Los retuits se sometieron a un proceso de limpieza y normalización para eliminar palabras vacías, lematizar o realizar stemming, y reducir el ruido en los datos.
- Modelado de Temas con gdsmm: Se implementó el modelo gdsmm para el análisis de los retuits preprocesados. Este modelo avanzado de topic modelling permite una identificación más precisa y eficiente de los temas ocultos en el corpus.

Figura 4.5: *WordCloud* para el tópico sobre la violencia de género.

Figura 4.6: *WordCloud* para el tópico sobre la guerra de Ucrania.

- Identificación de Temas Relevantes: Se analizaron las distribuciones de palabras y retuits para identificar los temas políticos más representativos y sus palabras clave asociadas.

- Interpretación y Análisis: Los temas identificados fueron interpretados y analizados para obtener una panorámica general de la actualidad política en España. Se observaron interconexiones entre temas y se identificaron tendencias emergentes.

Realizamos pruebas con el algoritmo de `GSDMM` probando con diferentes K parámetros (tópicos finales) 10 , 20 y 100. Nos dimos cuenta que con veinte tópicos obteníamos los mejores resultados ya que con diez tópicos el algoritmo solapaba demasiado entre tópicos, con cien había muchos tópicos con cero tuits incluidos en algunos tópicos. Contamos los tópicos (usando 100 como parámetro K) que tenían un número relevante de tuits dentro de cada tópico obteniendo 20 tópicos con un gran número de tuits. Una vez tenemos el parámetro K, procedemos a obtener las palabras más relevantes para cada tópico y por lo tanto saber el tema sobre lo que versa el tópico. Nos dimos cuenta que el algoritmo es capaz de diferenciar tuits en ingles o en euskera y asignarles un tópico.

De los tópicos más relevantes nos quedamos con dos por su relevancia política y de actualidad para su posterior análisis, como son la violencia de género y la guerra de Ucrania. Mostramos dos *WordCloud*, como podemos ver en las figuras 4.5 4.6, con los temas seleccionados, no obstante se realizaron *WordCloud* para cada uno de los veinte tópicos para saber el tema de cada uno de ellos.

El análisis reveló una diversidad de temas políticos relevantes en España. También encontramos que los tópicos podrían tener subtópicos, es decir, para el tópico de la guerra de Ucrania encontramos temas de política exterior que no están relacionados con la guerra de Ucrania. Los resultados proporcionan una visión clara y estructurada de los asuntos políticos más destacados en el país durante el período analizado.

Además, el enfoque de topic modelling con el modelo `GSDMM` demostró ser una herramienta eficaz para obtener una panorámica general de la actualidad política en España a partir de retuits de los diputados del congreso en Twitter. Los resultados obtenidos ofrecen una visión profunda y objetiva de los temas políticos más relevantes, lo que contribuye a la comprensión de la realidad política en el país. Este enfoque puede ser aplicado en futuros estudios para analizar la evolución de la actualidad política y ayudar a la toma de decisiones informadas en el ámbito político.

CAPÍTULO 5
Cuantificación de cámaras de echo en Twitter.

Las cámaras de eco online son a la vez causa y efecto del polarizado ambiente político que podemos ver en distintas partes del mundo. Una cámara de eco puede concebirse como un entorno en el que las ideas se refuerzan mediante interacciones repetidas entre usuarios con tendencias y actitudes similares [114]. Las redes sociales son terreno fértil para estas interacciones repetidas polarizadoras que conducen a la formación de Cámaras de Eco [38]. Además, los usuarios sólo están expuestos al contenido con el que están de acuerdo debido a la personalización de las redes sociales [14], lo que confirma aún más sus creencias existentes (véase el sesgo de confirmación [105]) y les protege de la exposición al argumento contrario (véase la exposición selectiva [80]). Además, las cámaras de eco de las redes sociales son una fuente de creación, consumo y amplificación de información que podría conducir a una mayor polarización política [45][75].

Para comprender mejor el efecto de las Cámaras de Eco y la Polarización sobre los individuos y la sociedad en general, necesitamos detectarlas y medirlas. Los trabajos anteriores sobre la detección de cámaras de eco y polarización se basan en cuentas etiquetadas manualmente, como políticos, personas influyentes en el ámbito político, canales de noticias o conjuntos predefinidos de hashtags y palabras clave polarizadas específicas del dominio a estudiar. Otros estudios más generalistas infieren la ideología del usuario a partir de información pública del usuario (por ejemplo, usando las cuentas seguidas o los hashtags que ha usado un usuario para estimar su inclinación política) y utilizan las estimaciones de la ideología del usuario para analizar la polarización y la cámara de eco. [19].

Nuestro experimento, se puede dividir en tres partes. En primer lugar, detectamos Cámaras (o comunidades) para cada tema basándonos en la red de retuits. A continuación, seleccionamos todos los usuarios de cada comunidad y para cada usuario generamos un vector promedio de los *embeddings* obtenidos por un modelo de BERT para cada *tweet* del usuario. Por último utilizamos una métrica que explicaremos a continuación, EchoGAE [5], para para obtener el Eco en todo el tópico y para cada comunidad.

En el apartado 5.0.1, desglosamos el concepto de cámara de eco y definimos ”Eco”, ”Cámara”, ”Cámara de eco” y ”Polarización” de acuerdo con la bibliografía y nuestro método computacional. En la sección 5.2, mostramos cómo generamos *embeddings* para los usuarios utilizando *transformers* y cuantificamos el ”Eco” por ”Cámara”. En la sección 5.3, mostramos los resultados del ”echo” a dos temas controvertidos recientes: **”Violencia de Género”** y **”La Guerra de Ucrania”**. Estos temas han sido obtenidos mediante un método de *topic modelling* no supervisado, como hemos comentado en la sección 4.2.2. Para cada uno de estos temas comparamos el nivel de ”Eco” por ”Cámara”.

5.0.1. Terminología.

Los términos "Cámara de Eco" y "Burbuja de Filtrado" se utilizan a menudo de forma indistinta, aunque a veces se integran con el concepto de "Polarización". Aunque existe una idea central común sobre ellos en la literatura, es difícil encontrar una definición única y universalmente establecida para cada uno de los términos. Por ello, en esta sección proponemos las definiciones que consideramos más coherentes con los significados literales de los términos y más alineadas con nuestro método de medición.

Eco

Definimos "Eco" como el nivel de homogeneidad entre los miembros de una discusión en un foro (Cámara). Esta homogeneidad puede derivarse de similitudes en la inclinación política de sus miembros (por ejemplo, izquierda o derecha tradicional), estatus socioeconómico o características demográficas (como la edad o el sexo).

Cámara

La "Cámara" es el foro de debate donde se producen las interacciones y los usuarios comparten contenidos o ideas. En Twitter, definimos una Cámara como un grupo de usuarios vinculados por una interacción (es decir, retuits, citas, menciones y respuestas) sobre un tema. Nuestro razonamiento es que estas agrupaciones representan una red en la que los usuarios interesados en un tema específico se exponen a un debate concreto en Twitter. Un hashtag en sí mismo también puede entrar en el ámbito de las Cámaras, ya que al hacer clic en el hashtag, uno quedará expuesto a un conjunto de tweets que contienen ese hashtag.

Cámara de eco.

Definimos las "Cámaras de Eco" como las "Cámaras" con altos niveles de "Eco"; en términos de Twitter, una red de retuits con un bajo nivel de diversidad de usuarios. La lógica aquí es que cuanto más homogéneo es un foro de debate, más probable es que un usuario escuche una voz homogénea que es el eco de la voz de sus miembros. En cambio, un foro con una alta diversidad de tipos de usuarios puede acoger una voz y una perspectiva más diversas.

5.1 Trabajo relacionado.

La metodología propuesta abarca los ámbitos de la detección de cámaras de eco y la generación de *embeddings* a nivel de usuario. En esta sección se analiza el estado del arte en estas áreas.

5.1.1. Detección de Cámaras de Eco.

El problema de la detección de Cámaras de Eco es la tarea de detectar comunidades de usuarios que muestran un comportamiento de eco ideológico dentro de una Cámara cerrada de usuarios. Este comportamiento se manifiesta en temas polarizantes como la política y la religión, donde los usuarios tienden a tomar posiciones extremas sobre estos temas, ya sea oponiéndose o apoyándose [74]. Podríamos dividir los métodos de

detección de cámaras de eco en tres tipos: basados en la red [41], basados en el contenido [32] y métodos de detección híbridos [129]. Los métodos basados en redes utilizan algoritmos de detección de comunidades para detectar comunidades en los grafos de interacción creados a partir de interacciones en redes sociales, como retuits y respuestas. Los métodos basados en el contenido agrupan a los usuarios en función del contenido que utilizan extrayendo características como el sentimiento sobre un tema o utilizando un *embedding* del contenido. Por último, el enfoque híbrido incorpora el conocimiento tanto del contenido como de la topología para encontrar Cámaras Eco. En este trabajo, nos hemos centrado en el contenido para detectar cámaras de eco y en la red para detectar comunidades. Para ello utilizamos `GSDMM` [135] para obtener tópicos/Cámara de eco dentro de nuestro conjunto de datos y analizar si se forman comunidades dentro de los tópicos que hemos obtenido.

5.1.2. *Representation learning* en Twitter.

El análisis de los datos de Twitter suele adoptar la forma de dos enfoques, a menudo combinados, por un lado el basado en el contenido y por otro el basado en la red. En los enfoques basados en el contenido, los usuarios se caracterizan por los metadatos de la cuenta, los hashtags, el contenido de los tweets y otras características relacionadas con el lenguaje extraídas de sus perfiles [3][41]. En los enfoques basados en redes, los usuarios se representan en la red de retuits o en la red de menciones, siendo ambas redes dirigidas en las que las relaciones indican el flujo de comunicación [39]. El uso de redes de seguidores de usuarios es poco frecuente debido a que su recopilación de datos requiere mucho tiempo [88].

Ambos enfoques pueden beneficiarse de los recientes avances en el campo del *representation learning* y, en concreto, de los métodos de generacion de *embeddings*. Técnicas como los *word embeddings*[93], o más recientemente los *transformers* [48], han demostrado que mejoran el análisis del sentimiento en los tuits [100] y la clasificación temática de los tuits [85]. Estos modelos generan una representación vectorial del texto de forma que palabras y textos semánticamente similares comparten representaciones similares. El concepto de *word embedding* también puede aplicarse a las redes, donde las representaciones de los nodos incorporan su homofilia y similitud estructural [65]. Los *network embeddings* pueden ayudar a la detección del tipo de usuario. Por ejemplo, Ribeiro et al [115] utilizaron el aprendizaje de representaciones tanto en la estructura de la red de retuits como en el contenido de los tuits para detectar usuarios tóxicos.

En este trabajo, hemos trabajado con el contenido generado por un usuario a través de sus tuits para representarlo vectorialmente mediante *sentence embeddings* y también usamos la topología de la red para entrenar un *Graph Auto-Encoder* (GAE)[79].

Generación de embeddings a nivel de usuario.

Los *embeddings* a nivel de usuario se utilizan para modelar el comportamiento de los usuarios en diversas tareas. Los métodos más recientes utilizan redes neuronales para codificar los datos de comportamiento de los usuarios (por ejemplo, tweets recientes en redes sociales o consultas recientes en motores de búsqueda) en vectores de baja dimensión. Estos enfoques reducen la cantidad de trabajo de ingeniería y extracción manual de características al automatizar las relaciones entre los propios datos del usuario, así como su relación con los datos de otros usuarios.

Los datos específicos del usuario en las redes sociales pueden dividirse en cuatro categorías diferentes:

- información del perfil del usuario
- actividad del usuario
- conectividad de red del usuario
- contenido generado por el usuario.

En el análisis del comportamiento de los usuarios en las redes sociales, los investigadores han utilizado diferentes combinaciones de las categorías mencionadas para crear representaciones de usuario específicas para cada tarea y universales [68]. La mayor parte de la investigación modela el comportamiento del usuario a través de su contenido generado. Estos textos generados por cada usuario pueden modelarse utilizando distintos métodos, como la Asignación de Latente de Dirichlet (LDA) [27][111], la Red Neuronal Convolucional (CNN) [10] o los *word embeddings* [111], entre otros.

Además, la conectividad de red de los usuarios también es común en el modelado de los atributos de los usuarios. Estos métodos intentan mapear las redes sociales en representaciones de baja dimensión de forma que se preserven las estructuras topológicas locales y globales. Los algoritmos de detección de comunidades y los modelos de redes neuronales con grafos son algunos de los métodos más utilizados para modelar redes sociales [130].

La información auxiliar, como la información de perfil, también ayudaría a modelar el comportamiento del usuario y a mejorar los métodos [137]. Sin embargo, todos los métodos de generación de *embeddings* a nivel de usuario para la detección de Echo Chamber se basan en un conjunto de usuarios políticos, palabras clave y hashtags etiquetados y seleccionados. Esto los haría menos robustos, más exigentes en cuanto a esfuerzo manual y menos generalizables a tareas posteriores de análisis de redes sociales, ya que los métodos supervisados son vulnerables al *concept drift* [60]. En otras palabras, a medida que pase el tiempo, las celebridades políticas, los hashtags políticos y el uso del lenguaje cambiarán.

En la sección 5.2, explicamos cómo generamos los *embeddings* de los usuarios basándonos en *sentence-transformers*.

5.2 Metodología.

En esta sección, explicaremos primero nuestro método de análisis de texto y de redes para detectar las cámaras y las comunidades más importantes para el debate en torno a cada tema. A continuación, mostramos cómo realizamos los embeddings para los usuarios de cada cámara y utilizamos los embeddings de los usuarios para medir el Eco de cada Cámara. Además queríamos estudiar también la correlación entre la toxicidad y el uso de bots, para ver si con nuestro conjunto de datos podíamos ver alguna correlación que indicase que existe dicha correlación y que por lo tanto pudiera haber un interés por una tercera persona de "toxificar" un argumento o qué partido político es más toxico. Este análisis será explicado en el capítulo 6. La figura 5.1 ofrece una perspectiva de nuestra metodología propuesta.

5.2.1. Detección de cámaras (*short-text clustering*) y comunidades (*network clustering*)

Utilizando un algoritmo de detección de comunidades, detectamos comunidades de usuarios que podrían estar en Cámaras de Eco en un grafo compuesto por retuits. En es-

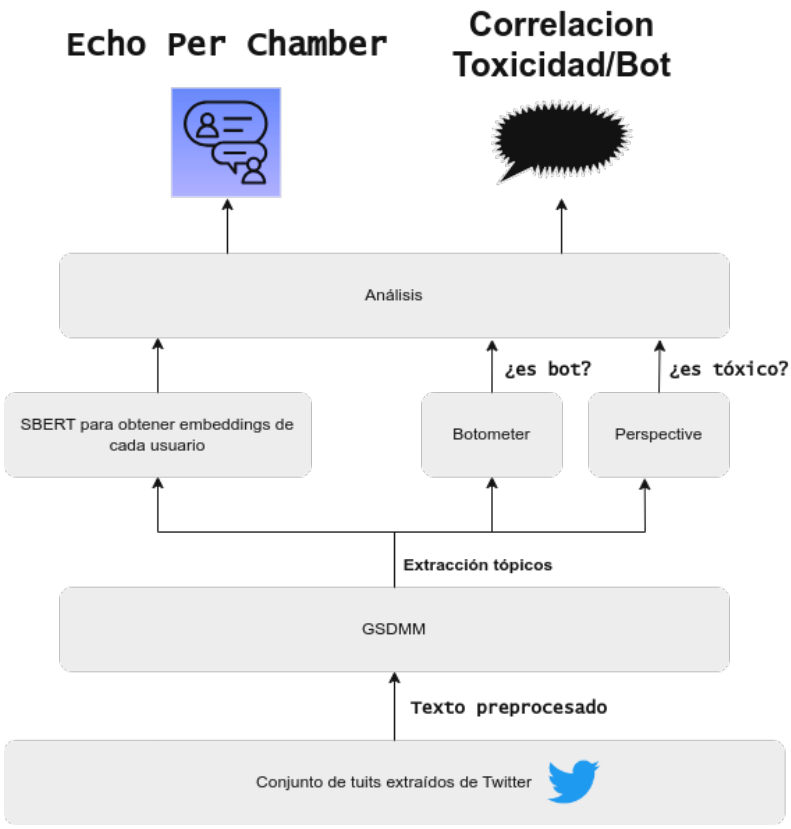

Figura 5.1: Metodología propuesta para este TFM.

te artículo, utilizamos el algoritmo Louvain [28] para encontrar Cámaras de Eco en redes de retuits sobre tres temas controvertidos. Tras realizar un rastreo de todos los retuits de los diputados del congreso, seleccionamos un subconjunto de tuits utilizando un algoritmo de *topic modelling* para texto corto que nos extrae tuits sobre un tema de carácter relevante. Utilizamos los tuits que han sido retuiteados que pertenecen a ese tópico para construir el grafo de retuits, en el que los nodos son usuarios y el enlace representa que el usuario A retuiteó al usuario B. No utilizamos todo el conjunto de datos porque no todos los tuits se retuitean y no todos los tuits son sobre el mismo tema por la forma en que han sido recopilados. Seleccionamos el retuit como interacción porque muestra el apoyo del usuario [30] al tweet de otro usuario. Hemos comprobado que el grafo de retuits suele mostrar comunidades distintas que representan los lados del debate. Para comprobarlo, seleccionamos un pequeño conjunto de usuarios, en nuestro caso los diputados de cada partido político. Y quedó claro que se formaban comunidades representando cada partido político. Nótese que no informamos del resultado de este pequeño experimento debido a su limitado alcance, no obstante en la figura 5.2 podemos ver cómo se forman las comunidades de un vistazo. Esto lo hacemos, más bien, como una forma de verificar que estamos encontrando comunidades significativas que podrían ayudarnos en nuestro análisis. Observamos que otros gráficos de interacción (gráficos de respuestas o menciones) no muestran comunidades ideológicamente segregadas que sean fáciles de detectar con un algoritmo de detección de comunidades basado en la topología. En este trabajo, utilizamos únicamente las características topológicas para detectar las comunidades. Por ello, optamos por utilizar la red de retuits. Después de construir la red de retuits, utilizamos el algoritmo de Louvain [28] para encontrar las comunidades. El algoritmo de Louvain es un algoritmo muy utilizado para detectar comunidades en grafos. Es rápido y puede encontrar rápidamente comunidades significativas en grafos grandes.

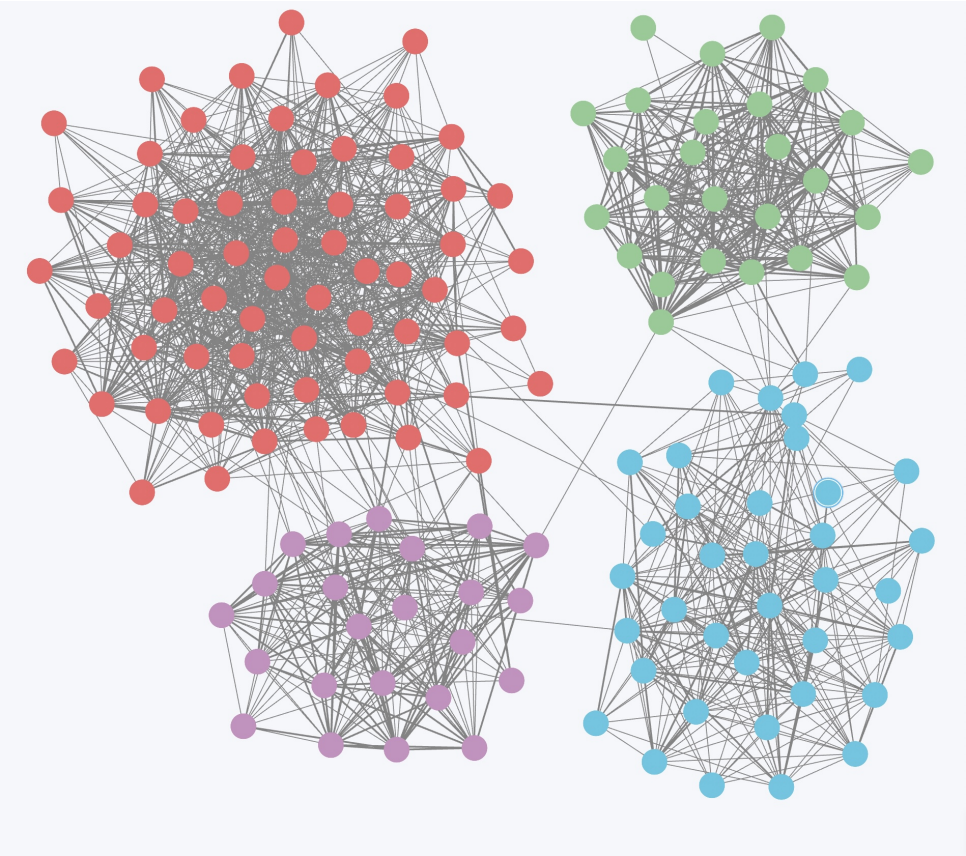

Figura 5.2: Comunidades que se forman tomando los retuits de los diputados de cada partido político. Los nodos rojos representan al PSOE, los verdes a Vox, los violetas a Podemos y los azules al PP. Los enlaces muestran la interacción de retuit.

5.2.2. Generación de *embeddings* de usuarios

El siguiente paso en nuestro análisis consiste en representar a los usuarios de Twitter en términos de cualidades extraídas de su comportamiento semántico en línea (tweets). Para ello, comenzamos extrayendo las características de los tweets que el usuario ha generado recientemente. Tras preprocesar el texto de los tweets (eliminando menciones, URL, etc.), los convertimos en vectores utilizando un modelo de *transformers* denominado (distiluse-base-multilingual-cased-v1) proporcionado por *Hugging-Face* en una biblioteca de *Python* llamada sentence-transformers. El modelo se ajusta con precisión para asignar frases y párrafos cortos a un espacio vectorial denso de 768 dimensiones de forma que se preserven las características semánticas del texto, de modo que los vectores puedan utilizarse para tareas como el *clustering* o la búsqueda semántica. A continuación, representamos a cada usuario basándonos en la media del vector de los *embeddings* de sus tuits. En cuanto a la velocidad del método, el proceso completo para 200 tuits de un usuario, transformarlos en vectores y promediar todos los vectores tardó aproximadamente 3 segundos por usuario en la GPU de Google-Colab.

El promedio de los *embeddings* se ha probado anteriormente para obtener *embeddings* de una frase promediando los *embeddings* de cada una de las palabras de la frase [11], sin embargo, hasta donde sabemos, no se ha aplicado en *embeddings* de frases múltiples (tuits) para representar al autor de las frases; que es lo que hacemos aquí. Aunque promediar *embeddings* de palabras con el fin de codificar frases prácticamente supera a muchos codificadores de frases basados en RNN, sabemos lógicamente que las palabras de una frase son elementos que dependen secuencialmente unos de otros y su orden debería tenerse en cuenta en un modelo de NLP. Sin embargo, podemos argumentar intuitivamente que promediar tendría mucho más sentido cuando estamos tratando con *embeddings* de tuits que son los elementos independientes de la mentalidad del usuario y el orden apenas significaría mucho en este caso. Por lo tanto, esperamos que promediar los *embeddings* de las frases (tuits) independientes arroje resultados significativos.

```
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('distiluse-base-multilingual-cased-v1')

def embed_user_tweets(tweets):
    emb = model.encode(tweets)
    emb = np.mean(emb, axis=0)
    return emb
```

5.2.3. Cuantificando el echo.

Esta sección presenta el enfoque utilizado para cuantificar las cámaras de eco en las conversaciones en línea. Nuestro objetivo es evaluar si el debate en torno a un tema determinado si las comunidades formadas por los usuarios pueden caracterizarse como cámaras de eco o comprenden un grupo diverso de individuos con ideologías distintas. Para ello, construimos un grafo $G_{Rt} = (V, E)$, donde V representa el conjunto de usuarios de las redes sociales y E representa las aristas que denotan interacciones homófilas, como los retweets. Además, obtenemos un conjunto de comunidades a partir de un algoritmo de detección de comunidades, donde cada comunidad consiste en un grupo de usuarios.

Nuestro objetivo principal es medir el nivel de "echo" dentro de todo el grafo calculando la "Puntuación de la Cámara de Eco" (ECS) para cada comunidad. En consecuencia, esta sección presenta la métrica ECS para cuantificar las cámaras de eco. Para poder

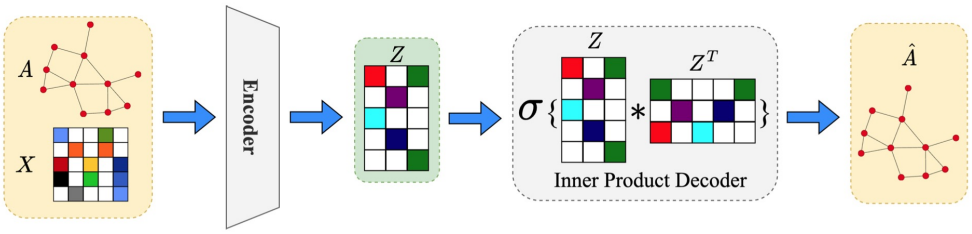

Figura 5.3: El modelo EchoGAE consta de dos componentes principales: un *coder* y un *decoder*. El *coder* utiliza las representaciones del contenido del usuario (X) y la matriz de adyacencia (A) para generar las representaciones del usuario (Z). A continuación, el *decoder* reconstruye la matriz de adyacencia $(\hat{A})$ utilizando las representaciones del usuario.

calcular la ECS necesitamos tanto los *embeddings* de usuarios como la topología de la red. Para ello se entrena un modelo denominado EchoGAE[5], que permite la representación de usuarios basada en su similitud ideológica.

EchoGAE es una adaptación del modelo Graph Auto-Encoder (GAE) [79], hecho a medida para generar *embeddings* de los usuarios basada en tuits e interacciones. Como modelo *self-supervised*, EchoGAE elimina la necesidad de etiquetado ideológico del usuario. Emplea dos capas convolucionales de grafos para codificar el grafo en una representación latente, que posteriormente se decodifica para reconstruir la estructura del grafo. El objetivo de EchoGAE es minimizar la entropía cruzada binaria entre las matrices de adyacencia real y reconstruida. Podemos ver la arquitectura en la figura 5.3. Este modelo será el que utilicemos para obtener los *embeddings* para calcular el ECS.

ECS (*Echo Chamber Score*) utiliza la distancia en el espacio vectorial como aproximación a estos factores, reflejando lo estrechamente relacionados que están los usuarios dentro de una comunidad (cohesión) y lo distinta que es una comunidad de otras (separación).

Imaginemos que $Z \in \mathbb{R}^{nxd}$ represente las incrustraciones/*embeddings* de usuario, donde n es el número de usuarios y d es la dimensión de los *embeddings*. Además, $\Omega = \{\omega_1\omega_2, ..., \omega_M\}$ el conjunto de comunidades, donde $\omega_i \subset V$ representa la i^{th} comunidad formada por usuarios. Para un usuario $u \in \omega$, calculamos el valor de cohesión λ_u como la distancia media entre u y otros usuarios de la misma comunidad mediante la ecuación 5.1.

$$\lambda_u = \frac{1}{|\omega|} \sum_{v \in \omega, v \neq u} dist(u, v) \tag{5.1}$$

Aquí, $|\omega|$ denota el número de usuarios en la comunidad ω y $dist(u, v)$ representa la distancia (por ejemplo, euclídea) entre los usuarios u y v en el espacio vectorial $Z^{(u)}$ y $Z^{(v)}$ respectivamente. Del mismo modo, calculamos el valor de separación $\triangle_u$ como la distancia media entre u y la comunidad más cercana distinta de ω mediante la ecuación 5.2.

$$\triangle_u = min_{\omega \in \Omega, u \neq v} \left[\frac{1}{|\omega|} \sum_{v \in \omega} dist(u, v) \right] \tag{5.2}$$

Para calcular la puntuación de cámara de eco (ECS) para una comunidad $\Omega = \{\omega_1\omega_2, ..., \omega_N\}$ utilizamos la ecuación 5.3 (fórmula inspirada en la puntuación de silueta, muy usada en técnicas como el clustering) que produce una puntuación entre 0 y 1, donde una puntua-

ción más alta indica una mayor probabilidad de un efecto de cámara de eco dentro de la comunidad.

$$ECS^*(\omega) = \frac{1}{|\omega|} \sum_{u \in \omega} \frac{max(\lambda_u, \triangle_u) + \lambda_u - \triangle_u}{2 \cdot max(\lambda_u, \triangle_u)} \tag{5.3}$$

Si queremos calcular para el grafo entero (y así diferenciar entre temas, cuales son más proclives al echo), usamos la fórmula 5.4.

$$ECS(\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} ECS^*(\omega) \tag{5.4}$$

5.3 Experimentos y resultados.

A continuación, utilizamos el contenido generado por los usuarios para medir el efecto de cámara de eco y verificar si las comunidades detectadas son realmente cámaras de eco.

5.3.1. Datos .

Utilizamos un gran conjunto de datos de Twitter (como hemos comentado en el capítulo 4), recopilados utilizando el API, con datos del 1 de enero de 2022 al 12 de marzo de 2023. Todos los tweets recopilados contienen tweets que han sido retuiteados por los diputados del congreso de la XIV legislatura. Los tweets pueden ser un tweet original y retuits. No contemplamos ni tweets citados (retuits con comentarios) ni respuestas. La lista de los diputados que se ha tenido en cuenta puede verse en la tabla B.1 en el apéndice. Unicamente crearemos dos grafos G_{RT} uno para cada tema (violencia de genero y guerra de Ucrania) para realizar el análisis. Estos tópicos han sido seleccionados como hemos visto en la sección 4.2.2 por lo solo utilizaremos un grafo que contenga los usuarios y los tuits que formen parte del tópico. Para el tópico de la violencia de género contamos con 20322 tuits y para la guerra de ucrania 21066.

5.3.2. Análisis de los resultados obtenidos.

Como hemos comentado en la sección 4.2.2, nos quedamos con dos tópicos para cuantificar el "echo" de cada una de las comunidades y ver si hay una diferencia significativa siguiendo la metodología propuesta. Para ver las comunidades nos ayudamos de la herramienta de visualización *Gephi*[1] para ver las relaciones de *RT* entre los usuarios. Además la propia herramienta lleva integrada el algoritmo de Louvain [28] como hemos mencionado en la sección 5.2.1 que era el algoritmo a utilizar para detectar comunidades ya que es ampliamente utilizado en la literatura actual. La figura 5.4 representa la red de retuits que obtenemos para el tópico de la violencia de género. Como vemos los nodos/usuarios se representan cada uno con un color dependiendo de la comunidad a la que pertenece. Como además teníamos a los diputados anotados, podemos inferir si la comunidad es del Partido Popular, Socialista, etc, asignando el color representativo para cada comunidad. Podemos ver como podríamos detectar en base a los retuits de un usuario el partido político con el que el usuario más se representa. La figura 5.5 representa la red de retuits que obtenemos para el tópico de la guerra de Ucrania.

[1] https://gephi.org

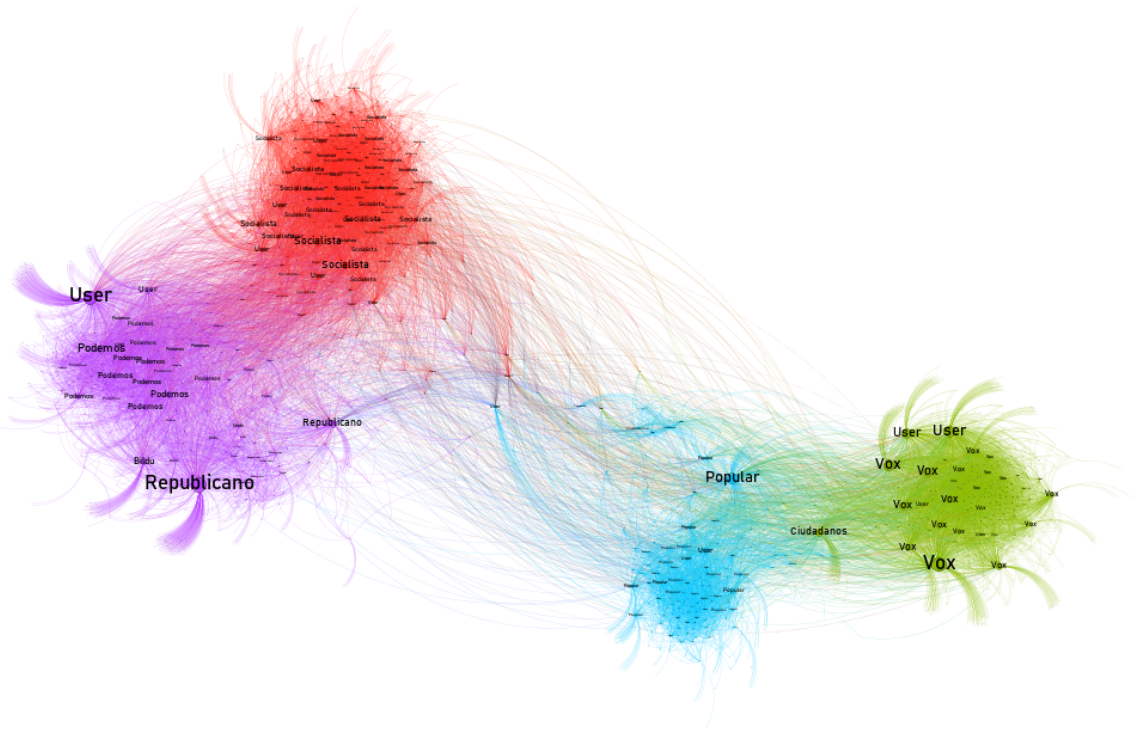

Figura 5.4: Comunidades que se forman para el tópico de la violencia de género usando *Gephi* como software de visualización.

Entrenamos el modelo de EchoGAE usando tanto la red de interacción (figuras 5.4 5.5) como los embeddings de cada usuario (hemos mostrado en la seccion 5.2.2 como obtenerlos). Una vez entrenado el modelo hemos obtenido las siguientes puntuaciones ECS para cada tópico.

Para el tópico de la violencia de género obtenemos:

- $ECS(VGenero) = 0{,}545$
- $ECS^*(VOX) = 0{,}570$
- $ECS^*(PODEMOS) = 0{,}576$
- $ECS^*(PSOE) = 0{,}511$
- $ECS^*(PP) = 0{,}469$

Para la guerra de Ucrania obtenemos:

- $ECS(Guerra) = 0{,}597$
- $ECS^*(VOX) = 0{,}577$
- $ECS^*(PODEMOS) = 0{,}628$
- $ECS^*(PSOE) = 0{,}594$
- $ECS^*(PP) = 0{,}537$

Para la violencia de género los datos sugieren que las comunidades de Podemos y Vox, con una perspectiva política más extremista que las otras obtienen un valor de ECS^*

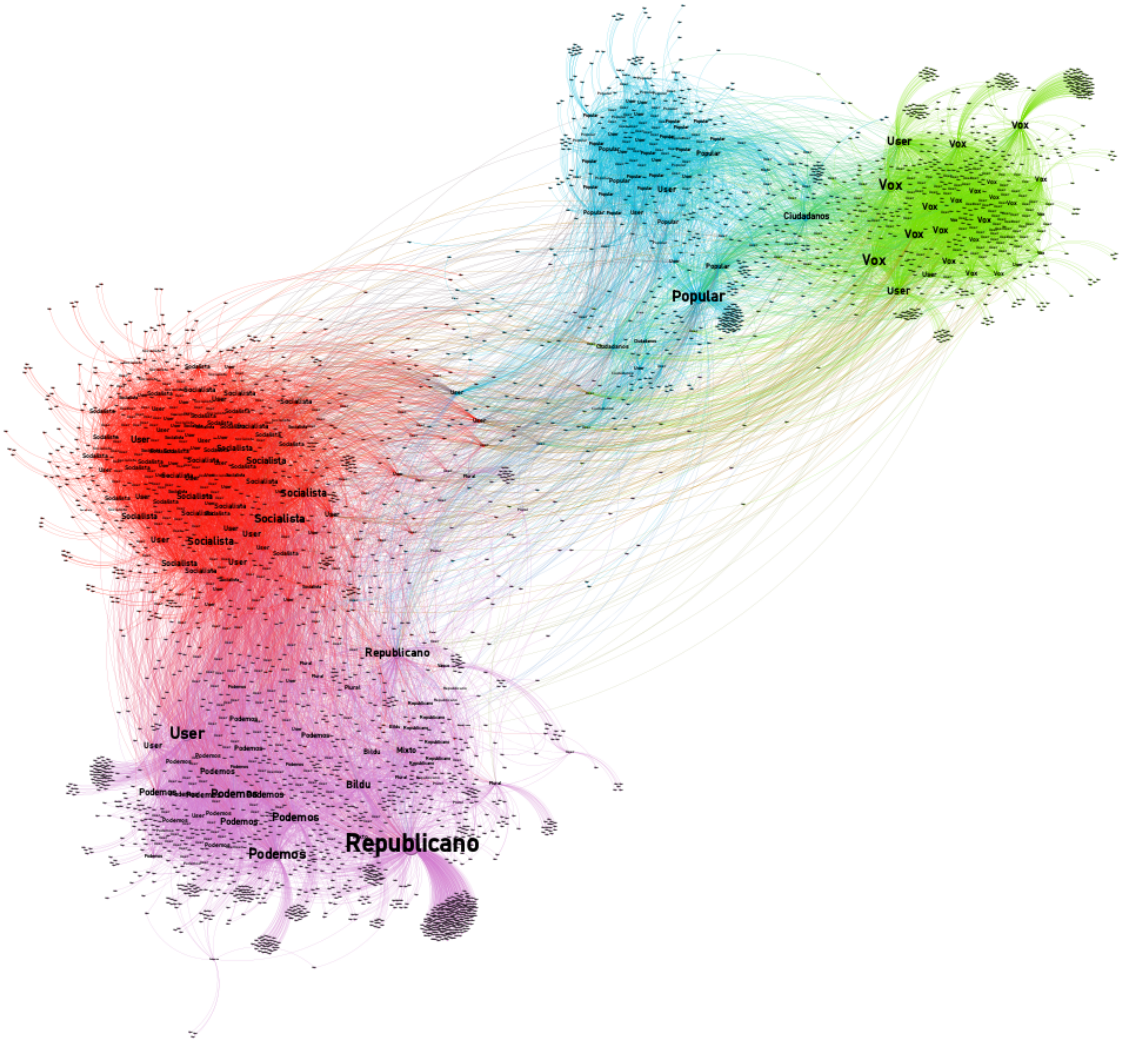

Figura 5.5: Comunidades que se forman para el tópico de la guerra de Ucrania usando *Gephi* como software de visualización..

mayor. Mientras que el valor ECS^* para el Partido Socialista y el Partido Popular es menor.

Para la guerra de Ucrania vemos valores ECS^* mayores para el PSOE y luego Podemos, seguido de VOX y por último el PP.

Podemos decir también que el tópico de la guerra de Ucrania $ECS(Guerra)$ tiene mayor echo que el tópico de la violencia de género $ECS(VGenero)$

CAPÍTULO 6
Cuantificación del hate.

En el entorno actual de las redes sociales, Twitter ha emergido como una plataforma prominente para la discusión y el intercambio de ideas, especialmente entre figuras políticas y sus seguidores. Sin embargo, esta interacción no siempre es constructiva y respetuosa. La presencia de comportamientos tóxicos, como insultos, ataques personales y amenazas, plantea preocupaciones sobre la calidad del debate en línea y su impacto en la participación ciudadana. Paralelamente, la proliferación de bots en Twitter ha sido objeto de discusión debido a su capacidad para influir en la opinión pública y alterar la percepción de los acontecimientos. Este capítulo se centra en analizar si existe una correlación entre la toxicidad presente en los retuits de los diputados del congreso durante los años 2022 y parte de 2023 en España, y la posible presencia de cuentas automatizadas (bots) en esos retuits.

6.1 Metodología.

Para realizar el análisis mencionado, utilizamos nuestro conjunto de retuits de los diputados del congreso. Decidimos centrarnos para este análisis únicamente en los tópicos de la guerra de Ucrania y la violencia de género en lugar de analizar todo el conjunto de datos ya que las APIs que utilizamos para el análisis tienen un límite de uso. Para estos dos temas, analizamos por un lado la toxicidad de todos los tuits asignando a cada tuit una puntuación que nos indica su nivel de toxicidad, por otro lado analizamos a los usuarios asignando una puntuación que nos indica cómo de probable es que ese usuario sea un bot. Podemos ver en la figura 5.1 el análisis que vamos a realizar.

6.1.1. Medición de Toxicidad

Para evaluar la toxicidad de los comentarios, se utilizó el API de Google Perspective[1]. El API proporciona una puntuación que indica la probabilidad de que un comentario sea tóxico en una escala del 0 al 1, donde 0 representa la no toxicidad y 1 representa una alta probabilidad de toxicidad. El API utiliza modelos de aprendizaje automático, en particular modelos de procesamiento de lenguaje natural (NLP), para analizar y comprender los patrones lingüísticos en el texto. Estos modelos han sido entrenados en grandes conjuntos de datos que contienen ejemplos de comentarios tóxicos y no tóxicos para aprender a reconocer los patrones asociados con la toxicidad. Además de la puntuación general de toxicidad, el API Perspective también puede proporcionar puntuaciones en dimensiones específicas de toxicidad, como TOXICIDAD, TOXICIDAD GRAVE, ATAQUE DE IDEN-

[1] https://perspectiveapi.com

TIDAD, INSULTO, PROFANIDAD y AMENAZA. Cada retuit recibió una puntuación en cada una de estas categorías, lo que nos permitió cuantificar y categorizar el grado de toxicidad presente en los retuits. Google Perspective define cada categoría de toxicidad como:

- **TOXICIDAD**: Comentario grosero, irrespetuoso o poco razonable que puede hacer que la gente abandone una discusión.
- **TOXICIDAD GRAVE**: Un comentario muy odioso, agresivo, irrespetuoso o que puede hacer que un usuario abandone una discusión o renuncie a compartir su punto de vista. Este atributo es mucho menos sensible a formas más leves de toxicidad, como comentarios que incluyen usos positivos de palabras malsonantes.
- **ATAQUE DE IDENTIDAD**: Comentarios negativos o de odio dirigidos a alguien por su identidad.
- **INSULTO**: Comentario insultante, incendiario o negativo hacia una persona o un grupo de personas.
- **PROFANIDAD**: Palabrotas, palabras malsonantes u otro lenguaje obsceno o profano.
- **AMENAZA**: Describe la intención de infligir dolor, lesiones o violencia contra un individuo o grupo.

6.1.2. Detección de Bots

Para determinar si un usuario es un bot o no, se utilizó el API de Botometer[2], que asigna a cada cuenta un valor entre 0 y 1 que indica la probabilidad de que sea un bot. Se estableció un umbral de 0.8 para clasificar a un usuario como bot, basándonos en nuestra evaluación que sugieren que este valor es un buen límite para diferenciar entre cuentas automatizadas y cuentas humanas.

El API Botometer, también conocido como "BotOrNot", es una herramienta desarrollada por la Universidad de Indiana que se utiliza para determinar la probabilidad de que una cuenta de usuario en Twitter sea un bot. La principal función del Botometer es analizar diferentes características y patrones en la actividad de una cuenta para evaluar su nivel de automatización y, por lo tanto, predecir la probabilidad de que sea un bot en lugar de una cuenta manejada por un usuario humano.

Para realizar esta evaluación, Botometer utiliza un enfoque basado en aprendizaje automático (*machine learning*) que se entrena en una amplia gama de datos provenientes de cuentas de bots y cuentas humanas reales. Estos datos incluyen patrones de comportamiento, características de actividad y contenido de los tweets. A continuación, se describen algunos de los aspectos clave que Botometer utiliza para determinar si un usuario es un bot o no:

- **Características de la Cuenta**: El Botometer analiza metadatos de la cuenta, como la fecha de creación, el número de seguidores, el número de seguidos, la frecuencia de tweets y la existencia de una foto de perfil y una descripción de usuario. Los bots a menudo tienen perfiles incompletos o incoherentes en comparación con las cuentas humanas genuinas.

[2]https://botometer.osome.iu.edu

- **Comportamiento de Actividad**: El análisis incluye el patrón de actividad de la cuenta, como la cantidad y el intervalo entre tweets, retweets y menciones. Los bots tienden a mostrar una actividad constante y repetitiva, mientras que los usuarios humanos suelen tener patrones más variables y contextuales.
- **Contenido de los Tweets**: El contenido de los tweets se analiza para identificar patrones de lenguaje, uso de enlaces y menciones a otros usuarios. Los bots a menudo comparten enlaces de manera indiscriminada y utilizan lenguaje genérico o repetitivo.
- **Redes Sociales**: Botometer también examina la red social del usuario, incluidas las relaciones y conexiones entre cuentas. Si una cuenta está conectada con muchas otras cuentas identificadas previamente como bots, esto podría aumentar su probabilidad de ser considerada un bot.
- **Características Textuales**: El contenido textual de los tweets es analizado en busca de patrones de lenguaje que puedan indicar la automatización. Esto incluye la detección de repeticiones de mensajes, uso excesivo de palabras clave o frases específicas, y la generación de contenido no coherente.
- **Características Semánticas**: El análisis también considera la semántica de los tweets y cómo se relacionan con temas y eventos actuales. Los bots a menudo carecen de la capacidad de comprender y responder adecuadamente a eventos cambiantes en tiempo real.

Al combinar y ponderar estas diferentes características, Botometer genera una puntuación de probabilidad que indica la posibilidad de que una cuenta sea un bot. Una puntuación más alto sugiere una mayor probabilidad de automatización, mientras que una puntuación más bajo indica una mayor probabilidad de que la cuenta sea manejada por un usuario humano.

Es importante destacar que aunque Botometer es una herramienta útil para evaluar la probabilidad de que una cuenta sea un bot, no es infalible y puede haber falsos positivos o falsos negativos. La detección de bots es un desafío en constante evolución debido a las tácticas cambiantes utilizadas por los desarrolladores de bots y la mejora continua de las técnicas de detección.

6.2 Análisis de los resultados obtenidos.

Una vez hemos obtenido los valores de toxicidad para cada tópico, realizamos la media de los valores para cada partido obteniendo la figura 6.1. Como podemos ver, obtenemos valores de toxicidad bajos, pero más altos en general para los partidos más radicales.

Sin embargo, solo con la media no tenemos una visión completa de la distribucion de los datos por lo que decidimos realizar un gráfico de cajas y bigotes que mostramos en la figura 6.2. Podemos ver cómo existen muchos valores atípicos que indican casos aislados de toxicidad en cada partido.

Si hablamos de los valores que hemos obtenido con Botometer, mostramos un histograma en la figura 6.3. Desde nuestra experiencia y estudio el límite humano-bot está en torno a 0.81 o 0.80, pero no hay ningún consenso sobre cual es el valor umbral para considerar que un usuario es un bot.

Para ver la correlación entre los valores de toxicidad y los valores de Botometer decidimos realizar un gráfico de dispersión (ver Figura 6.4). Si los puntos se agrupan de

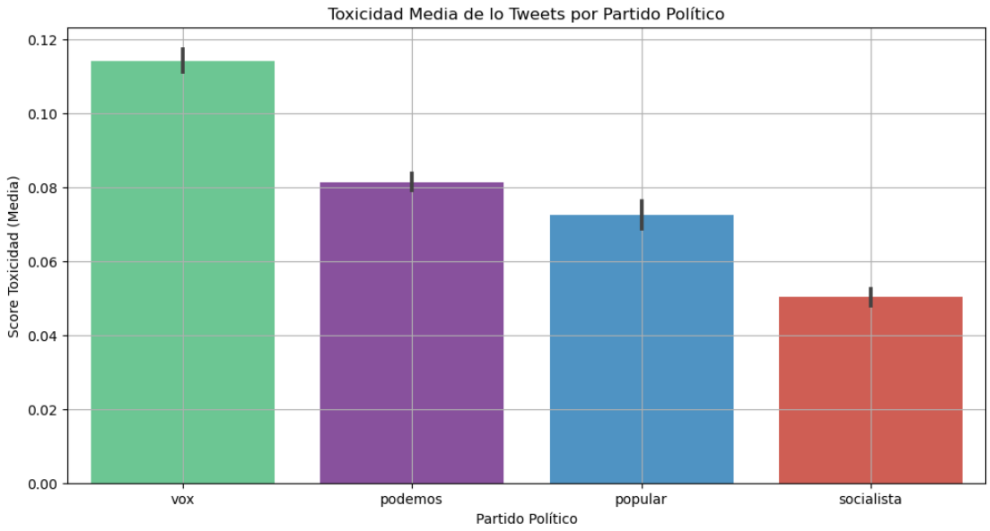

Figura 6.1: Toxicidad media de los tuits por partido político.

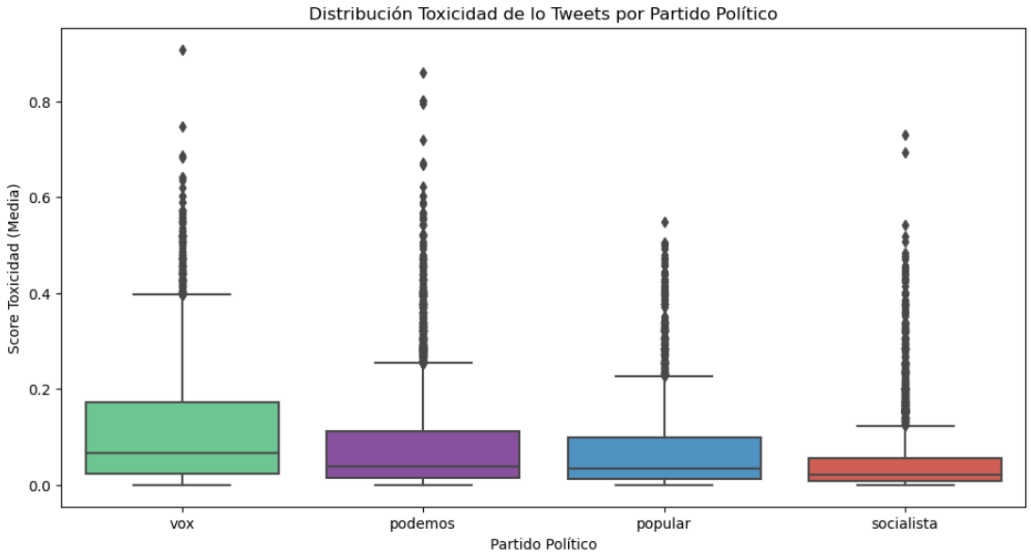

Figura 6.2: Gráfico de cajas y bigotes de la distribución de toxicidad en los tuits por partido político.

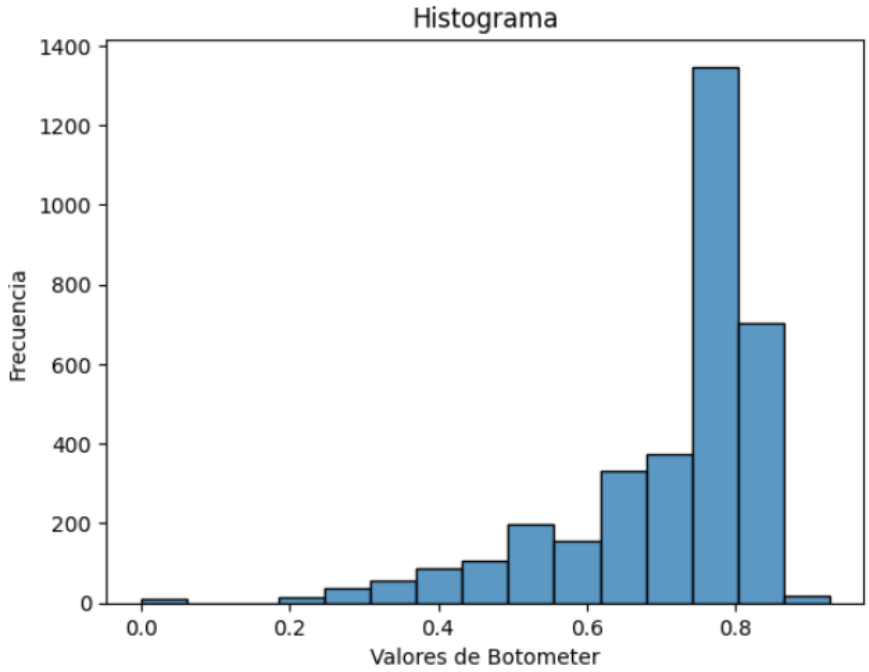

Figura 6.3: Histograma con el valor de botometer que indica la probabilidad de ser un bot.

idTwitter	text	TOXICITY	INSULT	SEVERE_TOXICITY	IDENTITY_ATTACK	PROFANITY	THREAT
1560999966495199232	@haselparis Si, sois unos machistas. Machistas de mierda.	0.9	0.9	0.5	0.4	0.9	0.0
1574891388289916930	@RubenSanchezTW @sanchezdelreal Pedazo de escoria, que hay que ser muy hdgp para decir lo que has dicho. De niño en el colegio cuando llegaba el jefe de estudio a clase era para informar de un atentado y decirnos que nuestros padres estaban bien, colegio militar. Eres un pedazo de mierda infecta.	0.9	0.9	0.6	0.2	0.8	0.0
1570750029501169664	@CanalUGR Sois una mierda vergonzante! La universidad pública no debe dar cabida al fascismo. Dais pena y asco!	0.9	0.9	0.5	0.2	0.7	0.0
1615086766825898008	@Tuareg555 Estamos denunciando tu mierda de cuenta para que no puedas generar más odio contra las mujeres. Ser inmundo	0.9	0.9	0.4	0.2	0.8	0.0
1593155200906240000	Un Padrastro HDLP le mete la Po... A su Hijastra de 14 años, un juez dice que no hay abuso de superioridad y le baja la condena... Y para la derecha clasista y la progresia mediática la culpa es de Irene Montero #FerrerasGate #Ascodepaís #YoConIreneMontero	0.9	0.9	0.7	0.5	0.8	0.1
1577936635219656706	"Putas, ninfómanas y os vamos a follar a todas..." ESTO ES ALGO QUE NO SE IMPROVISA. `https://t.co/XqvaR9VRfD`	0.9	0.8	0.6	0.3	0.9	0.4
1547293992865468416	¿Estamos en València o en el infierno por maric4s y feminaz1s? `https://t.co/to5rjFsQ1p`	0.9	0.8	0.6	0.7	0.6	0.0

Tabla 6.1: Tuits más tóxicos para el tópico de la violencia de género.

manera ascendente o descendente en una línea, hay una correlación. Si están dispersos al azar, no hay correlación. Como vemos no existe ninguna correlación aparente y por lo tanto descartamos en nuestro conjunto de datos al menos, una correlación.

En la tabla 6.1 se muestran los diez tuits más tóxicos para el tópico de la violencia de género. Junto con los valores asignados mediante Google Perspective.

6.3 Conclusiones

En este capítulo, se exploró tanto la toxicidad como la correlación entre la toxicidad presente en los retuits de los diputados del congreso en Twitter y la presencia de cuentas automatizadas (bots). Los resultados sugieren que no existe una relación entre la actividad de los bots y la toxicidad en los retuits políticos.

A pesar de que no hayamos encontrado una correlacion en nuestra pequeña muestra (de unos 20000 tuits para la violencia de género) no significa que no exista una correlación ya que creemos que al solo estar contemplando retuits de los diputados del congreso, nuestros resultados no son representativos de todo lo que puede darse más allá de las opiniones de los diputados, sin embargo consideramos importante realizar el análisis

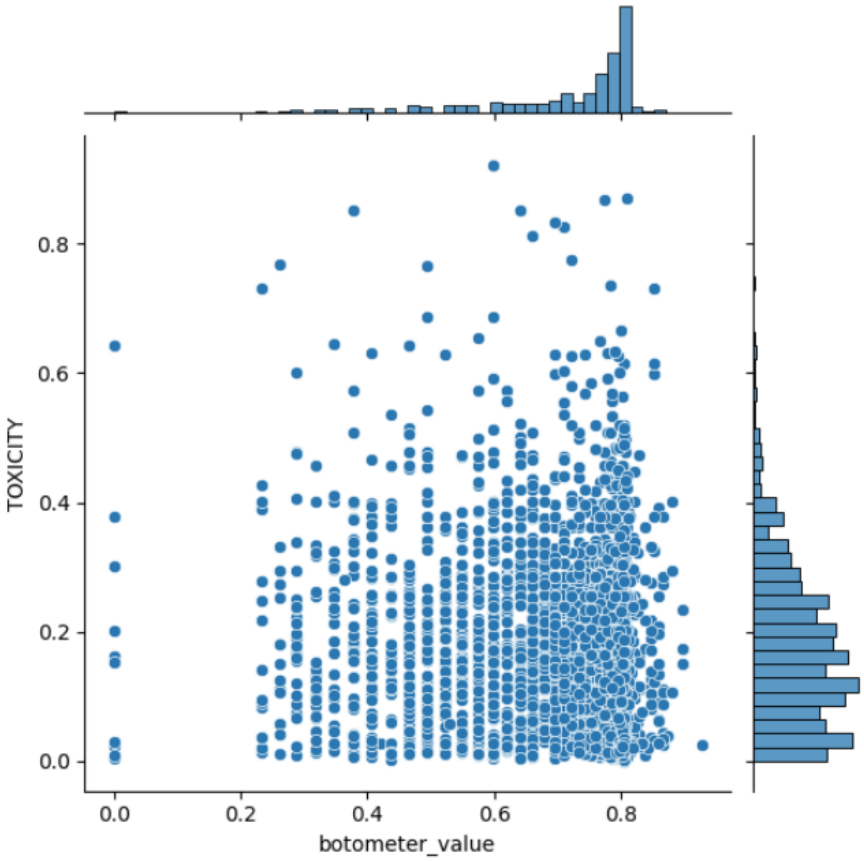

Figura 6.4: Gráfico de dispersión entre la media de la toxicidad de los tuits de un usuario y el valor de botometer que indica la probabilidad de ser un bot.

puesto que ya teníamos los datos y podía ser útil de cara a un trabajo futuro donde se extrajera una muestra más representativa con un conjunto de datos menos encorsetado.

Este estudio, aunque breve, contribuye a la comprensión de la interacción entre bots y la toxicidad presente en Twitter en el ámbito político en España.

CAPÍTULO 7
Conclusiones, Limitaciones y Trabajo Futuro.

7.1 Conclusiones

En este trabajo final de máster se planteaba en un principio como objetivo principal clasificar a los usuarios como votantes de la izquierda o votantes de la derecha en función del contenido político compartido. Hemos visto que el retuit es una interacción que indica que el usuario está a favor de lo retuiteado, según Barberá et al. [19], cerca del 75 % de los retuits sobre temas políticos tienen lugar entre usuarios de opiniones ideológicas similares. En nuestros experimentos hemos visto, por ejemplo en las figuras 5.4 5.5 cómo a partir de los retuits y cuentas etiquetadas por partido político, podemos inferir que un usuario es de izquierdas o de derechas a partir del algoritmo de Louvain [28].

Una de nuestras ideas descartadas al empezar a trabajar en este trabajo fue entrenar un modelo de BERT que a partir de texto clasificase el tuit a favor o en contra de la ideologia de izquierdas o de derechas y monitorizar, por ejemplo, debates para saber quien estaba teniendo mejor representación en redes como Twitter en tiempo real, idea fuertemente inspirada de este articulo. Esto es un problema complejo ya que existen problemas como la brevedad del texto y la enorme variación en el vocabulario de los tuits que dificultan enormemente este problema conocido como *Stance Detection*. En la literatura se expresa que únicamente contemplando los retuits no podemos determinar la posición política de un usuario. Deberíamos tener en cuenta mas variables para tener resultados fidedignos. Los algoritmos actuales de detección de posturas requieren datos de entrenamiento específicos para cada acontecimiento o tuits anotados y, por tanto, son difíciles de adaptar a nuevos acontecimientos. Al no contar con un conjunto de datos en español con el que entrenar dicho modelo y ver fuera del alcance de este trabajo esta idea fue descartada.

Como objetivo secundario, también planteamos detectar dentro de estas comunidades, bots, que pretendan influenciar las campañas electorales, así como también cuantificar echo chambers y estudiar como difieren los discursos, por ejemplo a favor o en contra del aborto. Por lo tanto, al no poder centrarnos tanto en la deteccion de posturas políticas mediante modelos del lenguaje, decidimos centrar nuestros esfuerzos en el estudio de la toxicidad y la cuantificación de echo chamber.

La investigación presentada en este trabajo final de master sobre redes sociales y cámaras de eco ha avanzado en nuestra comprensión de cómo el éxito de las plataformas de redes sociales está transformando las conversaciones políticas y su impacto en la pola-

rización política. Aunque se han logrado avances significativos, existen varias cuestiones pendientes.

1. **Generalización de resultados:** La mayoría de los estudios se basan en datos de Estados Unidos, lo que plantea interrogantes sobre la generalización de los hallazgos a otros contextos internacionales, especialmente en relación con la violencia política.

2. **Variación en el tiempo:** La evolución constante de las plataformas de medios sociales y su uso plantean la pregunta de si los efectos polarizadores están mejorando o empeorando con el tiempo, lo que dificulta la comprensión del estado actual.

3. **Consideraciones éticas:** La investigación sobre extremismo y violencia relacionada con las redes sociales requiere una consideración ética profunda, ya que las intervenciones para reducir la polarización pueden tener consecuencias imprevistas en la participación cívica y el debate político.

4. **Complejas relaciones multicausales:** Comprender cómo la exposición a diferentes perspectivas políticas en las plataformas digitales afecta a la polarización, la participación ciudadana y la calidad del debate es esencial. Esto implica encontrar un equilibrio entre la moderación y la promoción de la diversidad de opiniones.

En última instancia, el estudio de cómo las RRSS impactan en la política democrática es de gran relevancia en la actualidad, y es fundamental para abordar los desafíos relacionados con la polarización política y promover una participación ciudadana informada y comprometida en el entorno digital.

Respecto a la problemática del discurso de odio en línea nuestro estado del arte destaca varios aspectos cruciales:

1. **Definición y detección desafiantes:** A pesar de la creciente atención, la definición precisa del discurso de odio en línea sigue siendo un tema debatido. Detectar sistemáticamente este tipo de contenido es extremadamente difícil debido a su naturaleza contextual y en constante evolución.

2. **Fragmentación en la investigación:** Aunque se han empleado técnicas avanzadas como el aprendizaje automático y las redes neuronales para detectar el discurso de odio, la mayoría de los estudios existentes se centran en un solo tipo de discurso de odio en una plataforma específica. Además, la mayoría de los datos provienen de Twitter en inglés, lo que limita la comprensión de otras plataformas y contextos culturales.

3. **Politización y variedad de enfoques:** Las definiciones y enfoques para detectar el discurso de odio están políticamente cargados, especialmente en contextos autoritarios y conflictivos. Aunque algunos estudios han explorado múltiples aspectos del discurso de odio, estos son la excepción.

4. **Estrategias de mitigación:** Se ha demostrado que la prohibición de comunidades de odio reduce el discurso de odio en algunas plataformas, pero también existe evidencia de que los usuarios trasladan su discurso a otros lugares. El contra-discurso, donde los usuarios sancionan el discurso de odio, parece ser una estrategia efectiva, y su eficacia varía según el contexto cultural y la plataforma.

Hemos conseguido crear una base de datos en Neo4j con el conjunto de retuits de los diputados del congreso, contando con 29436 usuarios (personas, periódicos...) que

han retuiteado los diputados y 192 diputados y un total de 611771 tuits. Esta base de datos puede ser utilizada para obtener los tuits mas retuiteados en un periodo de tiempo, cuales son los hashtags mas usados, etc. Sobre estos tuits hemos realizado una tarea de *Topic Modelling*, obteniendo una variedad de temas relevantes de la actualidad política en España y quedándonos con dos tópicos interesantes para el estudio. Para estos dos tópicos, La Guerra de Ucrania y la violencia de género, hemos realizado análisis para la cuantificación de *echo chamber*.

También hemos analizado la correlación entre la toxicidad y la existencia de bots para estos tópicos. Obteniendo que para nuestro conjunto de datos, que solo contempla los retuits de los diputados, no estamos teniendo una visión completa de la toxicidad que pudiera haber en Twitter y obteniendo que no existe una correlación entre la toxicidad y la existencia de bots.

7.2 Limitaciones

En primer lugar, se ha encontrado que la API de Twitter ha cerrado un endpoint necesario para que Botometer funcione, lo que ha dificultado la detección de bots, provocando que no podamos realizar el análisis para el tópico de la Guerra de Ucrania. En segundo lugar, se ha identificado la necesidad de mejorar la detección de bots en Twitter para saber si existen bots con la capacidad de influir en las campañas electorales. En tercer lugar, se ha encontrado que los algoritmos de modelado de temas, como LDA, funcionan mal cuando se aplican en textos cortos, como los tweets, ya que solo se dispone de información muy limitada de co-ocurrencia de palabras en textos cortos. En tercer lugar, se ha encontrado también, una escasez de conjuntos de entrenamiento para la detección de posturas políticas en el español.

Durante el desarrollo del trabajo encontramos una herramienta que podía ayudarnos a sacar más resultados, IBM Debater. El API de IBM Debater es una herramienta de procesamiento de lenguaje natural que se enfoca en el análisis y generación de argumentos basados en texto. La probamos y nos dimos cuenta que solo era válida para texto en inglés. Esta API de IBM Debater, puede ser una herramienta interesante para el análisis en un futuro cuando esté adaptada para el español.

Relevante para un análisis de toxicidad en Twitter tendríamos los siguientes endpoints. **Claim Detection**, este endpoint se utiliza para detectar afirmaciones en un texto. Es útil para identificar las declaraciones clave hechas en un discurso o artículo. Útil para identificar afirmaciones potencialmente tóxicas o agresivas en los tweets. **Claim Boundaries**, complementario al anterior, este endpoint ayuda a identificar los límites de una afirmación dentro de un texto, es decir, dónde comienza y termina una afirmación. Ayuda a definir los límites de las afirmaciones tóxicas en un tweet. **Evidence Detection**, detecta evidencia o argumentos de apoyo dentro de un texto. Puede ayudar a identificar datos o información que respalda una afirmación. Puede identificar la evidencia o argumentos utilizados en los tweets para respaldar afirmaciones tóxicas. **Argument Quality**, este endpoint evalúa la calidad de un argumento en función de su lógica y estructura. Puede ayudar a determinar la solidez de un argumento. Permite evaluar la calidad de los argumentos tóxicos presentes en los tweets.

Relevante para un análisis de echo chamber en Twitter: **Pro / Con**, identifica los argumentos a favor (pro) y en contra (con) de un tema o declaración en un texto. Es útil para analizar diferentes perspectivas. Ayuda a identificar los argumentos a favor y en contra de un tema en los tweets, lo que puede revelar la existencia de cámaras de eco en las que solo se presentan argumentos unilaterales. **Theme Extraction,** identifica los temas principales que se tratan en un conjunto de textos. Es útil para resumir y categorizar con-

tenido textual. Identifica los temas principales en los tweets, lo que es fundamental para entender los temas predominantes dentro de una echo chamber.

7.3 Trabajo futuro

A pesar de las limitaciones encontradas, se han identificado áreas de mejora y se han propuesto líneas de investigación futura para continuar avanzando en este campo. En primer lugar, se propone continuar con la investigación en la detección de bots y la mejora de la detección de posturas políticas en otras redes sociales. En segundo lugar, se sugiere estudiar las implicaciones éticas de la generación de ciberinteligencia en el ámbito político. En tercer lugar, se propone investigar en la mejora de los algoritmos de modelado de temas para textos cortos.

Respecto al discurso de odio, en el futuro, es crucial llevar a cabo análisis comparativos sistemáticos para mejorar la detección del discurso de odio en línea en sus diversas formas. Además, se deben explorar estrategias más efectivas para combatir la incitación al odio en línea, teniendo en cuenta el contexto cultural y las plataformas específicas. Dada la gravedad de las consecuencias offline del discurso de odio, académicos y responsables políticos deben seguir construyendo sobre esta base para mejorar la detección, comprensión y mitigación de este fenómeno en línea.

Aprovechando que tenemos una forma automática de obtener los identificadores de los diputados y tambien podemos automatizar la ingesta de los datos en la base de datos, se plantea como trabajo futuro la monitorización de la actualidad política construyendo una pagina web usando Streamlit[1]. Desde esta pagina podríamos acceder directamente a los retuits que mas impacto están teniendo, el sentimiento de los tuits o ver el volumen de tuits que esta teniendo la red social. Podemos ver un ejemplo de la idea en la figura 7.1, sacada de un articulo de Medium.

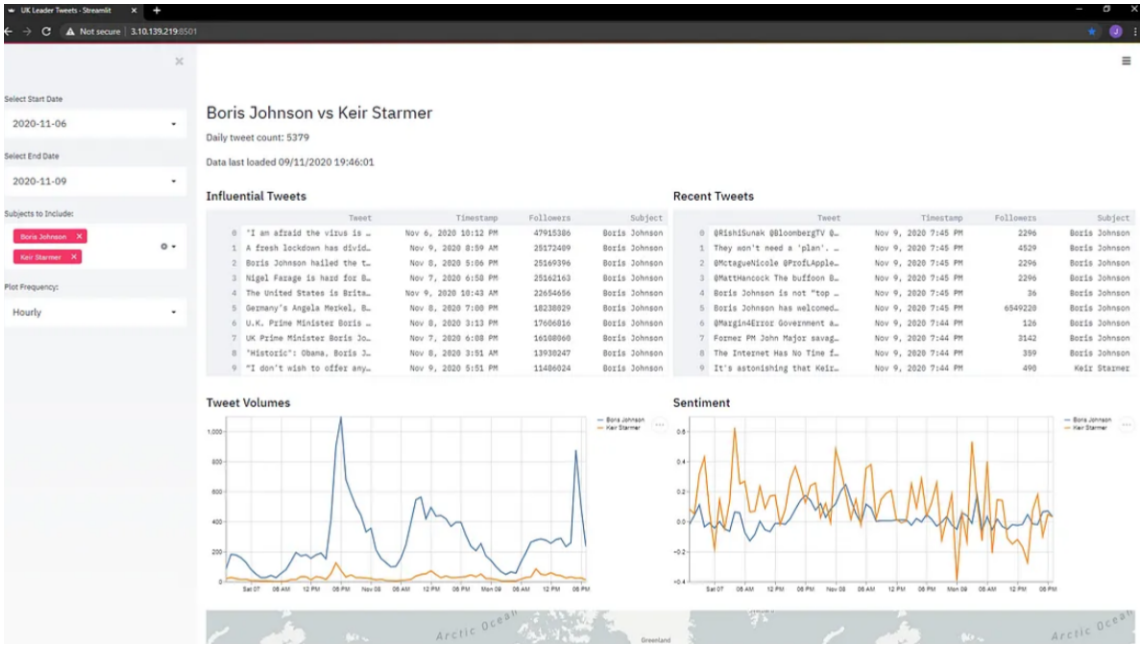

Figura 7.1: Aplicación construida con Streamlit que monitoriza a los politicos Boris Johnson y Keir Starmer.

[1]https://streamlit.io/

APÉNDICE A

OBJETIVOS DE DESARROLLO SOSTENIBLE

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**		X		
ODS 4. **Educación de calidad.**		X		
ODS 5. **Igualdad de género.**		X		
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades.**	X			
ODS 11. **Ciudades y comunidades sostenibles.**		X		
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**			X	
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**		X		
ODS 17. **Alianzas para lograr objetivos.**		X		

Tabla A.1: Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

La investigación académica y científica contemporánea debe abordar desafíos complejos y cambiantes que afectan a la sociedad global. Uno de estos desafíos es comprender la dinámica de las redes sociales, especialmente en plataformas como Twitter, donde la información fluye de manera constante y donde se pueden observar fenómenos como la formación de "echo chambers" (cámaras de eco) y la propagación de contenido tóxico. Este tipo de investigación es esencial para abordar cuestiones importantes relacionadas con la comunicación en línea y sus implicaciones en la sociedad actual.

En este contexto, es relevante explorar cómo la investigación sobre echo chambers y toxicidad en Twitter puede relacionarse con los Objetivos de Desarrollo Sostenible (ODS). La relación entre la investigación en redes sociales y los ODS puede no ser inmediatamente evidente, pero a través de un análisis más profundo, podemos identificar conexiones significativas.

En primer lugar, es importante reconocer que algunos de los ODS tienen una relación más directa con la investigación sobre redes sociales que otros. A continuación, exploraremos la relevancia de los ODS en el contexto de la investigación sobre echo chambers y toxicidad en Twitter, asignando niveles de relevancia (alto, medio, bajo o no procede) a cada uno de ellos.

1. **Fin de la pobreza (No procede):** No es evidente una relación directa entre la investigación sobre redes sociales y la erradicación de la pobreza. Si bien la comunicación en línea puede influir en las percepciones sobre la pobreza, nuestra investigación generalmente se centra en otros aspectos.

2. **Hambre cero (No procede):** De manera similar al ODS 1, la investigación en redes sociales rara vez aborda temas relacionados con la seguridad alimentaria y el hambre.

3. **Salud y bienestar (Relevancia media):** La salud mental y el bienestar en línea pueden estar relacionados con la investigación sobre echo chambers y toxicidad en Twitter. El acoso en línea y la difusión de información falsa pueden tener un impacto en la salud emocional de las personas, lo que hace que este objetivo sea relevante en cierta medida.

4. **Educación de calidad (Relevancia media):** La calidad de la información en línea es un factor importante en la investigación sobre redes sociales, especialmente cuando se trata de cómo se forman las echo chambers y cómo se propagan mensajes tóxicos. La educación en línea y la información que circula en Twitter pueden ser relevantes para entender estos fenómenos.

5. **Igualdad de género (Relevancia media):** La igualdad de género es un tema relevante en la investigación sobre toxicidad en línea. La discriminación de género y el acoso en línea son problemas serios que pueden estar relacionados con la formación de echo chambers y la propagación de contenido tóxico.

6. **Agua limpia y saneamiento (No procede):** Este objetivo se centra en el acceso a agua limpia y saneamiento básico, lo cual no tiene una relación directa con nuestra investigación sobre redes sociales y toxicidad en Twitter.

7. **Energía asequible y no contaminante (No procede): La energía no está directamente relacionada con la investigación en redes sociales y toxicidad en Twitter.**

8. **Trabajo decente y crecimiento económico (No procede):** Aunque se puede relacionar con la automatización y la economía digital en el uso de bots en Twitter, no se relaciona de manera significativa con la investigación sobre las dinámicas de las redes sociales y la toxicidad en línea. Por lo tanto, en este contexto específico, consideramos que este ODS no procede.

9. **Industria, innovación e infraestructuras (Relevancia media):** La investigación sobre el uso de bots en Twitter y cómo afecta a la información y la comunicación se relaciona directamente con la innovación, lo que hace que este objetivo sea relevante.

10. **Reducción de las desigualdades (Relevancia alta):** La investigación sobre echo chambers y toxicidad en Twitter está relacionada con la polarización y la creación de burbujas de información, lo que podría contribuir a la ampliación de desigualdades de opinión.

11. **Ciudades y comunidades sostenibles (Relevancia media):** La investigación se centra en comunidades en línea y cómo se comportan en Twitter, este objetivo podría ser relevante en términos de sostenibilidad en la comunicación digital.

12. **Producción y consumo responsables (No procede):** Este objetivo generalmente se aplica más a la producción física de bienes y servicios, por lo que no se relaciona de manera significativa con la investigación en línea.

13. **Acción por el clima (No procede):** Nuestra investigación no aborda la difusión de información falsa sobre el cambio climático o la influencia de bots en la opinión pública sobre el medio ambiente, si lo hiciera este objetivo podría ser relevante.

14. **Vida submarina (No procede):** No existe una relación evidente entre la vida submarina y la investigación en redes sociales y toxicidad en Twitter.

15. **Vida de ecosistemas terrestres (No procede):** Al igual que el ODS 14, este objetivo no guarda relación con la investigación en línea.

16. **Paz, justicia e instituciones sólidas (Relevancia media):** La toxicidad en línea y la formación de echo chambers pueden tener implicaciones en la paz social y la justicia en la medida en que afecten la calidad del discurso público y la polarización.

17. **Alianzas para lograr objetivos (Relevancia media):** En este trabajo la UPV ha colaborado junto con S2 Grupo para llevar a cabo las investigaciones pertinentes.

APÉNDICE B

Listado de diputados a los que realizamos un seguimiento.

Tabla B.1: Listado de diputados del congreso de la XIV legislatura con twitter.

username	twitterID	Link
Santi_ABASCAL	260788584	https://twitter.com/Santi_ABASCAL
JuanjoAizcorbe	2832034762	https://twitter.com/JuanjoAizcorbe
MertxeAizpurua	555326714	https://twitter.com/MertxeAizpurua
alfonsocendon	901414373050241024	https://twitter.com/alfonsocendon
aalmodobar	155350950	https://twitter.com/aalmodobar
cayetanaAT	303168830	https://twitter.com/cayetanaAT
bea_fanjul	320145391	https://twitter.com/bea_fanjul
carmenandres_	372852334	https://twitter.com/carmenandres_
AnguitaOmar	875222185	https://twitter.com/AnguitaOmar
arandapaco	76285597	https://twitter.com/arandapaco
gemmaraujo	526173273	https://twitter.com/gemmaraujo
Jaumeasens	212377097	https://twitter.com/Jaumeasens
joanbaldovi	378944786	https://twitter.com/joanbaldovi
BassaMontse	1084408297	https://twitter.com/BassaMontse
meritxell_batet	725700028392689664	https://twitter.com/meritxell_batet
ionebelarra	344739325	https://twitter.com/ionebelarra
abeltran_ana	2300055003	https://twitter.com/abeltran_ana
lauraberja86	269472408	https://twitter.com/lauraberja86
Patri_Blanquer	399278810	https://twitter.com/Patri_Blanquer
GenisBoadella	541483505	https://twitter.com/GenisBoadella
AnaBotellaPSOE	4258411577	https://twitter.com/AnaBotellaPSOE
albertbotran	2978274629	https://twitter.com/albertbotran
buenopin	130906485	https://twitter.com/buenopin
Caballerohelena	2177035457	https://twitter.com/Caballerohelena
jacallejascano	403536088	https://twitter.com/jacallejascano
PabloCamPiq	2679135387	https://twitter.com/PabloCamPiq
ZaidaCantera	2610041328	https://twitter.com/ZaidaCantera
conceptermens	2867230876	https://twitter.com/conceptermens
capdevilajoan	331085228	https://twitter.com/capdevilajoan
educarazo	242430168	https://twitter.com/educarazo

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
Antonia_Jover_	230807056	https://twitter.com/Antonia_Jover_
_Maria_Dantas_	87042194	https://twitter.com/_Maria_Dantas_
pedro_casares	378525757	https://twitter.com/pedro_casares
googlemars_	793518916005990400	https://twitter.com/googlemars_
santicl	141183245	https://twitter.com/santicl
OscarClavell	540111388	https://twitter.com/OscarClavell
fjconpe	192834709	https://twitter.com/fjconpe
rafi_crespin	3348787462	https://twitter.com/rafi_crespin
solcruzguzman	789120175	https://twitter.com/solcruzguzman
edelvallerod	1122839391096004608	https://twitter.com/edelvallerod
CelsoDelgadoOU	755798880118243328	https://twitter.com/CelsoDelgadoOU
DioufLuc	4276393785	https://twitter.com/DioufLuc
JCDuran_	222048521	https://twitter.com/JCDuran_
GloriaElizo	301403189	https://twitter.com/GloriaElizo
odonelorza2011	284488502	https://twitter.com/odonelorza2011
xavieritja	243593550	https://twitter.com/xavieritja
jmespejosaav	405430824	https://twitter.com/jmespejosaav
ivanedlm	48976307	https://twitter.com/ivanedlm
CrisEstebanVox	1193908437362577408	https://twitter.com/CrisEstebanVox
AntidioF	1228373473	https://twitter.com/AntidioF
BelenFCasero	2718789058	https://twitter.com/BelenFCasero
SofCastanon	3495603017	https://twitter.com/SofCastanon
RafaLomana	821157253	https://twitter.com/RafaLomana
soniafetesoro	264761667	https://twitter.com/soniafetesoro
Isabel_Franco_	302453497	https://twitter.com/Isabel_Franco_
Juanb0462	379117713	https://twitter.com/Juanb0462
cucagamarra	83573034	https://twitter.com/cucagamarra
oscargamazo	170340172	https://twitter.com/oscargamazo
rafajosevelez	504365221	https://twitter.com/rafajosevelez
MarcLamua	171506399	https://twitter.com/MarcLamua
Adrilastra	200194757	https://twitter.com/Adrilastra
SebastianLede15	1179187428	https://twitter.com/SebastianLede15
MikelLegarda	4178226827	https://twitter.com/MikelLegarda
TeresaGdVinuesa	2286278800	https://twitter.com/TeresaGdVinuesa
Quin1954	267248683	https://twitter.com/Quin1954
valentingarciag	353808961	https://twitter.com/valentingarciag
belitagl	864392892	https://twitter.com/belitagl
egnieto3	1634794128	https://twitter.com/egnieto3
margpuig	2921983427	https://twitter.com/margpuig
PilarGarrido_	3377815605	https://twitter.com/PilarGarrido_
agarzon	11904592	https://twitter.com/agarzon
LuisGestoso	932321656416108544	https://twitter.com/LuisGestoso
Hectorgomezh	60114737	https://twitter.com/Hectorgomezh
AntonGomezReino	365392319	https://twitter.com/AntonGomezReino
vicpiedra	49607100	https://twitter.com/vicpiedra
CarmenGGuinda	4385137295	https://twitter.com/CarmenGGuinda
Aglezterol	164274777	https://twitter.com/Aglezterol

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
MartaGlezVzqz	3436686957	https://twitter.com/MartaGlezVzqz
sandrage76	199850925	https://twitter.com/sandrage76
SonyaGuerraLpz	701090171865976832	https://twitter.com/SonyaGuerraLpz
TxemaGuijarro	752439542859304960	https://twitter.com/TxemaGuijarro
lidiaguinart	440281129	https://twitter.com/lidiaguinart
Sergio_GP	162055437	https://twitter.com/Sergio_GP
indasalinas	293944355	https://twitter.com/indasalinas
Hernanzsofia	393300798	https://twitter.com/Hernanzsofia
herrerobono	2849798849	https://twitter.com/herrerobono
Eselkaos	302024398	https://twitter.com/Eselkaos
BelenHoyo	239779545	https://twitter.com/BelenHoyo
AntonioHurtado	20585781	https://twitter.com/AntonioHurtado
JonInarritu	402593346	https://twitter.com/JonInarritu
miqueljerez	385882186	https://twitter.com/miqueljerez
rodrijr111	1681133749	https://twitter.com/rodrijr111
nasholop	224199272	https://twitter.com/nasholop
Lauralopezd	193228093	https://twitter.com/Lauralopezd
GemaL_Somoza	223167040	https://twitter.com/GemaL_Somoza
AndresLorite	222267978	https://twitter.com/AndresLorite
pepelosadaf	225094442	https://twitter.com/pepelosadaf
Roser_Maestro	118453663	https://twitter.com/Roser_Maestro
rubenmansolivar	386796948	https://twitter.com/rubenmansolivar
pilarmarcosd	144981542	https://twitter.com/pilarmarcosd
pmklose	1165362025	https://twitter.com/pmklose
gmariscalanaya	196204374	https://twitter.com/gmariscalanaya
AngelesMarra	487856790	https://twitter.com/AngelesMarra
valentinam	9242202	https://twitter.com/valentinam
luzseijo	270356445	https://twitter.com/luzseijo
OskarMatute	27675374	https://twitter.com/OskarMatute
JoseMariaMazon	1109041676042018816	https://twitter.com/JoseMariaMazon
guillermomeijon	118033828	https://twitter.com/guillermomeijon
javier_merino	82652418	https://twitter.com/javier_merino
mestremanuel	93471453	https://twitter.com/mestremanuel
sergimiquel	176191893	https://twitter.com/sergimiquel
onofremiralles	263982632	https://twitter.com/onofremiralles
mjmonteroc	310187114	https://twitter.com/mjmonteroc
IreneMontero	372812630	https://twitter.com/IreneMontero
TristanaMg	789723368	https://twitter.com/TristanaMg
MoroMjesus	2365318187	https://twitter.com/MoroMjesus
luciadalda	2578198139	https://twitter.com/luciadalda
mariadelamiel	316058009	https://twitter.com/mariadelamiel
Begonasarre	88802757	https://twitter.com/Begonasarre
NavalpotroJulio	409108065	https://twitter.com/NavalpotroJulio
pedronavarrol	116701487	https://twitter.com/pedronavarrol
malenanevado	274481242	https://twitter.com/malenanevado
jaimedeolano	1637483929	https://twitter.com/jaimedeolano
anioramas	479394745	https://twitter.com/anioramas

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
inmaoria	475802494	https://twitter.com/inmaoria
Ortega_Smith	1118941682	https://twitter.com/Ortega_Smith
EstherPadillaR	202540944	https://twitter.com/EstherPadillaR
jpagesm	1347580676	https://twitter.com/jpagesm
anapastorjulian	307570052	https://twitter.com/anapastorjulian
raquel_pedraja	112223206	https://twitter.com/raquel_pedraja
estherpcamarero	266683385	https://twitter.com/estherpcamarero
MercePerea	214794461	https://twitter.com/MercePerea
AuxiPD	192883412	https://twitter.com/AuxiPD
mercedes18601	1071468288465612800	https://twitter.com/mercedes18601
meripita44	1369105908	https://twitter.com/meripita44
perejoanpons	22079985	https://twitter.com/perejoanpons
anaprietonieto	289844440	https://twitter.com/anaprietonieto
MargaProhens	361322719	https://twitter.com/MargaProhens
Normapujol	109332337	https://twitter.com/Normapujol
joseramirezdel2	1113813322170863616	https://twitter.com/joseramirezdel2
ElviraRamon	390217747	https://twitter.com/ElviraRamon
tamarayar	3082102098	https://twitter.com/tamarayar
mdelaoredondo	3083771273	https://twitter.com/mdelaoredondo
germanrenau	153563085	https://twitter.com/germanrenau
JDRequenaRuiz	527564981	https://twitter.com/JDRequenaRuiz
AlbertoRA_VOX	96874050	https://twitter.com/AlbertoRA_VOX
gomezdcelis	22473944	https://twitter.com/gomezdcelis
CarlosRojas_PPA	158846236	https://twitter.com/CarlosRojas_PPA
cromeroher	461900036	https://twitter.com/cromeroher
rromerovilches	3074491774	https://twitter.com/rromerovilches
susana_ros	365957901	https://twitter.com/susana_ros
_patricia_rueda	2613310915	https://twitter.com/_patricia_rueda
gabrielrufian	2904896141	https://twitter.com/gabrielrufian
jruizcarbonell	163541727	https://twitter.com/jruizcarbonell
RuizSolas	3368070699	https://twitter.com/RuizSolas
isabanes	110442665	https://twitter.com/isabanes
isauralealf	555395673	https://twitter.com/isauralealf
lcsahuquillo	364183185	https://twitter.com/lcsahuquillo
asalvav	1093032954	https://twitter.com/asalvav
sanchezdelreal	53517943	https://twitter.com/sanchezdelreal
marianos_ugt	1264670185	https://twitter.com/marianos_ugt
sanchezcesar	28385397	https://twitter.com/sanchezcesar
sanchezcastejon	68740712	https://twitter.com/sanchezcastejon
asanchog	1373591724	https://twitter.com/asanchog
sancho_herminio	711969278149308416	https://twitter.com/sancho_herminio
EnriqueSantiago	293012562	https://twitter.com/EnriqueSantiago
SarriMorell	615560129	https://twitter.com/SarriMorell
SeguraJust	1308490587520282624	https://twitter.com/SeguraJust
danielsenderos	2790857133	https://twitter.com/danielsenderos
JFrSerrano	337929140	https://twitter.com/JFrSerrano
yolandaalcaldes	4409892215	https://twitter.com/yolandaalcaldes

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
SimancasRafael	402219133	https://twitter.com/SimancasRafael
AlejandroSolerM	251096605	https://twitter.com/AlejandroSolerM
jlsteeg	593952938	https://twitter.com/jlsteeg
eloysuarezl	241080178	https://twitter.com/eloysuarezl
RicardoTarno	3017650371	https://twitter.com/RicardoTarno
caroltelechea	2234780184	https://twitter.com/caroltelechea
eledhmel	3050371833	https://twitter.com/eledhmel
EdurneUriarte	250092838	https://twitter.com/EdurneUriarte
RoberUriarte	3019568698	https://twitter.com/RoberUriarte
PilarVallugera	1724984834	https://twitter.com/PilarVallugera
anadebande	397867366	https://twitter.com/anadebande
Mireia_veca	312297259	https://twitter.com/Mireia_veca
MartinaVelardeG	3300574151	https://twitter.com/MartinaVelardeG
Viondi	15667049	https://twitter.com/Viondi
AinaVS	53350394	https://twitter.com/AinaVS
czambranogr	2385882402	https://twitter.com/czambranogr
AurelioZapata5	1107322810001690624	https://twitter.com/AurelioZapata5
AnaZurita7	518445571	https://twitter.com/AnaZurita7
abalosmeco	202372417	https://twitter.com/abalosmeco
JLAceves	132666884	https://twitter.com/JLAceves
fjosealcaraz	266066285	https://twitter.com/fjosealcaraz
JAngelVillalon	386439973	https://twitter.com/JAngelVillalon
GerardAlv7	1427847870	https://twitter.com/GerardAlv7
javieranton	119170897	https://twitter.com/javieranton
InesArrimadas	552561770	https://twitter.com/InesArrimadas
carmen_banos	264362379	https://twitter.com/carmen_banos
Ferran_Bel	352930189	https://twitter.com/Ferran_Bel
VicenteBetoret	141173322	https://twitter.com/VicenteBetoret
_mireiaborras	2348453556	https://twitter.com/_mireiaborras
evapatriciab	2675863029	https://twitter.com/evapatriciab
tcabcas	516834154	https://twitter.com/tcabcas
Graciacanales3	844285125248651264	https://twitter.com/Graciacanales3
InesCanizares	2561709684	https://twitter.com/InesCanizares
luisacarcedo	925676922838929408	https://twitter.com/luisacarcedo
BeaMCarrillo	621790334	https://twitter.com/BeaMCarrillo
mcastellonPP	265938163	https://twitter.com/mcastellonPP
rchamode	279233001	https://twitter.com/rchamode
mariocortesc	1303352534204059648	https://twitter.com/mariocortesc
MeerRocio	4525423215	https://twitter.com/MeerRocio
Yolanda_Diaz_	761862806	https://twitter.com/Yolanda_Diaz_
PabloEchenique	25555639	https://twitter.com/PabloEchenique
ierrejon	482389606	https://twitter.com/ierrejon
AITOR_ESTEBAN	158442670	https://twitter.com/AITOR_ESTEBAN
afernb	702298741651447808	https://twitter.com/afernb
pedro_fhz	805508913261146112	https://twitter.com/pedro_fhz
FigaredoJoseM	1113036780813672448	https://twitter.com/FigaredoJoseM
DiegoGagoB	267707645	https://twitter.com/DiegoGagoB

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
MarioGarcesSan	808948514155855872	https://twitter.com/MarioGarcesSan
Montsechavar	837874657	https://twitter.com/Montsechavar
TeoGarciaEgea	224074203	https://twitter.com/TeoGarciaEgea
MorisSiero	917967385	https://twitter.com/MorisSiero
AliciaGarcia_Av	495288613	https://twitter.com/AliciaGarcia_Av
SaraGimnez	479740555	https://twitter.com/SaraGimnez
Migonzalezcaba	3391734845	https://twitter.com/Migonzalezcaba
AriagonaGP	3221713971	https://twitter.com/AriagonaGP
InesGranollers	357586041	https://twitter.com/InesGranollers
Maritxu30	270901461	https://twitter.com/Maritxu30
otazu35	983101658	https://twitter.com/otazu35
MGutierrezCs	2250812558	https://twitter.com/MGutierrezCs
HispanPablo	3095873030	https://twitter.com/HispanPablo
MarionaID	441732574	https://twitter.com/MarionaID
beajimenez2023	288349257	https://twitter.com/beajimenez2023
fuensantalima	256656869	https://twitter.com/fuensantalima
patxilopez	16084460	https://twitter.com/patxilopez
juralde	69060709	https://twitter.com/juralde
JILopezBas	3246171929	https://twitter.com/JILopezBas
pmanglano	244101525	https://twitter.com/pmanglano
joanmargall	15071264	https://twitter.com/joanmargall
MariscalZabala	257669729	https://twitter.com/MariscalZabala
imoblanca	3108277737	https://twitter.com/imoblanca
MayoralRafa	1269484171	https://twitter.com/MayoralRafa
joanmena	61822734	https://twitter.com/joanmena
montseminguez	847169983	https://twitter.com/montseminguez
DiezMoneo	1255527496772173824	https://twitter.com/DiezMoneo
MackMontesinos	1197128623364423680	https://twitter.com/MackMontesinos
DiegoMovellan	96331240	https://twitter.com/DiegoMovellan
MariloNarvaez	467682272	https://twitter.com/MariloNarvaez
CnLacoba	4172373034	https://twitter.com/CnLacoba
miriamnoguerasM	343507680	https://twitter.com/miriamnoguerasM
jr0rtega	1233298924368269312	https://twitter.com/jr0rtega
vejer_ortiz	724965496391634944	https://twitter.com/vejer_ortiz
mapaniagua	90467570	https://twitter.com/mapaniagua
juan_pedreno	3208501132	https://twitter.com/juan_pedreno
abellas_p	1006180942237626368	https://twitter.com/abellas_p
G_Pisarello	3039268715	https://twitter.com/G_Pisarello
JsPostigo	2457797184	https://twitter.com/JsPostigo
arnauramirez	92407680	https://twitter.com/arnauramirez
CesarJRamos	11086042	https://twitter.com/CesarJRamos
NestorRego	435720619	https://twitter.com/NestorRego
CarmenRiolobos	106811959	https://twitter.com/CarmenRiolobos
JoseantonioJun	21361705	https://twitter.com/JoseantonioJun
rosaromerocr	253968932	https://twitter.com/rosaromerocr
MartaRosiq	572925673	https://twitter.com/MartaRosiq
RuizdePinedo	596531845	https://twitter.com/RuizdePinedo

Continúa en la siguiente página

Tabla B.1 – continuación de la página anterior

username	twitterID	Link
MarisaSaavedraM	897945036402327552	https://twitter.com/MarisaSaavedraM
PabloSaezAM	559273864	https://twitter.com/PabloSaezAM
Idosagasti	250635207	https://twitter.com/Idosagasti
jsalvadorduch	326105216	https://twitter.com/jsalvadorduch
msolsj	613233224	https://twitter.com/msolsj
J_Sanchez_Serna	300969753	https://twitter.com/J_Sanchez_Serna
LuisStamaria	227819294	https://twitter.com/LuisStamaria
sergiosayas	144600462	https://twitter.com/sergiosayas
dvserrada	271436089	https://twitter.com/dvserrada
felipe_sicilia	246982544	https://twitter.com/felipe_sicilia
juanluissotoadd	1916660826	https://twitter.com/juanluissotoadd
SSumelzo	520609496	https://twitter.com/SSumelzo
uxia_tizon	1244014932250832896	https://twitter.com/uxia_tizon
JulioUtrilla	1309893771165868032	https://twitter.com/JulioUtrilla
Rubendariove	246713858	https://twitter.com/Rubendariove
Evelascomorillo	987327876084191232	https://twitter.com/Evelascomorillo
nvillagrasa	146397973	https://twitter.com/nvillagrasa
J_Zaragoza_	715890348569010176	https://twitter.com/J_Zaragoza_

Bibliografía

[1] URL: https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

[2] Lada A Adamic y Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". En: *Proceedings of the 3rd international workshop on Link discovery*. 2005, págs. 36-43.

[3] Aseel Addawood et al. "Linguistic cues to deception: Identifying political trolls on social media". En: *Proceedings of the international AAAI conference on web and social media*. Vol. 13. 2019, págs. 15-25.

[4] Luca Maria Aiello et al. "Friendship prediction and homophily in social media". En: *ACM Transactions on the Web (TWEB)* 6.2 (2012), págs. 1-33.

[5] Faisal Alatawi, Paras Sheth y Huan Liu. "Quantifying the Echo Chamber Effect: An Embedding Distance-based Approach". En: *arXiv preprint arXiv:2307.04668* (2023).

[6] Hunt Allcott et al. "The welfare effects of social media". En: *American Economic Review* 110.3 (2020), págs. 629-676.

[7] Gordon Willard Allport, Kenneth Clark y Thomas Pettigrew. "The nature of prejudice". En: (1954).

[8] Wafa Alorainy et al. ""The enemy among us" detecting cyber hate speech with threats-based othering language embeddings". En: *ACM Transactions on the Web (TWEB)* 13.3 (2019), págs. 1-26.

[9] Amalia Álvarez-Benjumea y Fabian Winter. "Normative change and culture of hate: An experiment in online environments". En: *European Sociological Review* 34.3 (2018), págs. 223-237.

[10] Silvio Amir et al. "Modelling context with user embeddings for sarcasm detection in social media". En: *arXiv preprint arXiv:1607.00976* (2016).

[11] Sanjeev Arora, Yingyu Liang y Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings". En: *International conference on learning representations*. 2017.

[12] Evelyn Aswad. "The Role of US Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban". En: *HRLR Online* 1 (2016), pág. 1.

[13] Christopher A Bail et al. "Exposure to opposing views on social media can increase political polarization". En: *Proceedings of the National Academy of Sciences* 115.37 (2018), págs. 9216-9221.

[14] Eytan Bakshy, Solomon Messing y Lada A Adamic. "Exposure to ideologically diverse news and opinion on Facebook". En: *Science* 348.6239 (2015), págs. 1130-1132.

[15] Eytan Bakshy et al. "The role of social networks in information diffusion". En: *Proceedings of the 21st international conference on World Wide Web*. 2012, págs. 519-528.

[16] Pablo Barberá. "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US". En: *Job Market Paper, New York University* 46 (2014), págs. 1-46.

[17] Pablo Barberá y Gonzalo Rivero. "Understanding the political representativeness of Twitter users". En: *Social Science Computer Review* 33.6 (2015), págs. 712-729.

[18] Pablo Barberá et al. "The critical periphery in the growth of social protests". En: *PloS one* 10.11 (2015), e0143611.

[19] Pablo Barberá et al. "Tweeting from left to right: Is online political communication more than an echo chamber?" En: *Psychological science* 26.10 (2015), págs. 1531-1542.

[20] Matthew Barnidge. "Exposure to political disagreement in social media versus face-to-face and anonymous online settings". En: *Political communication* 34.2 (2017), págs. 302-321.

[21] Valerio Basile et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". En: *Proceedings of the 13th international workshop on semantic evaluation*. 2019, págs. 54-63.

[22] Susan Benesch. "Countering dangerous speech to prevent mass violence during Kenya's 2013 elections". En: *Final Report* (2014), págs. 1-26.

[23] Susan Benesch. "Defining and diminishing hate speech". En: *State of the world's minorities and indigenous peoples* 2014 (2014), págs. 18-25.

[24] Yochai Benkler, Robert Faris y Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[25] Jonathan M Berger y Heather Perez. *The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters*. George Washington University, 2016.

[26] James Bisbee y Jennifer M Larson. "Testing social science network theories with online network data: An evaluation of external validity". En: *American political science review* 111.3 (2017), págs. 502-521.

[27] David M Blei, Andrew Y Ng y Michael I Jordan. "Latent dirichlet allocation". En: *Journal of machine Learning research* 3.Jan (2003), págs. 993-1022.

[28] Vincent D Blondel et al. "Fast unfolding of communities in large networks". En: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[29] Levi Boxell, Matthew Gentzkow y Jesse M Shapiro. "Greater Internet use is not associated with faster growth in political polarization among US demographic groups". En: *Proceedings of the National Academy of Sciences* 114.40 (2017), págs. 10612-10617.

[30] Danah Boyd, Scott Golder y Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter". En: *2010 43rd Hawaii international conference on system sciences*. IEEE. 2010, págs. 1-10.

[31] Pete Burnap y Matthew L Williams. "Us and them: identifying cyber hate on Twitter across multiple protected characteristics". En: *EPJ Data science* 5 (2016), págs. 1-15.

[32] Fernando H Calderón et al. "Content-based echo chamber detection on social media platforms". En: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, págs. 597-600.

[33] Aylin Caliskan, Joanna J Bryson y Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". En: *Science* 356.6334 (2017), págs. 183-186.

[34] Michael X Delli Carpini, Fay Lomax Cook y Lawrence R Jacobs. "Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature". En: *Annu. Rev. Polit. Sci.* 7 (2004), págs. 315-344.

[35] Eshwar Chandrasekharan et al. "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech". En: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), págs. 1-22.

[36] Naganna Chetty y Sreejith Alathur. "Hate speech review in the context of online social networks". En: *Aggression and violent behavior* 40 (2018), págs. 108-118.

[37] Raphael Cohen-Almagor. "Fighting hate and bigotry on the Internet". En: *Policy & Internet* 3.3 (2011), págs. 1-26.

[38] Elanor Colleoni, Alessandro Rozza y Adam Arvidsson. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". En: *Journal of communication* 64.2 (2014), págs. 317-332.

[39] Michael Conover et al. "Political polarization on twitter". En: *Proceedings of the international aaai conference on web and social media*. Vol. 5. 1. 2011, págs. 89-96.

[40] Michael D Conover et al. "Partisan asymmetries in online political activity". En: *EPJ Data science* 1.1 (2012), págs. 1-19.

[41] Michael D Conover et al. "Predicting the political alignment of twitter users". En: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, págs. 192-199.

[42] Thomas Davidson et al. "Automated hate speech detection and the problem of offensive language". En: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1. 2017, págs. 512-515.

[43] Pasquale De Meo et al. "On Facebook, most ties are weak". En: *Communications of the ACM* 57.11 (2014), págs. 78-84.

[44] Tom De Smedt, Guy De Pauw y Pieter Van Ostaeyen. "Automatic detection of online jihadist hate speech". En: *arXiv preprint arXiv:1803.04596* (2018).

[45] Michela Del Vicario et al. "The spreading of misinformation online". En: *Proceedings of the national academy of Sciences* 113.3 (2016), págs. 554-559.

[46] Fabio Del Vigna12 et al. "Hate me, hate me not: Hate speech detection on facebook". En: *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*. 2017, págs. 86-95.

[47] Richard Delgado. "Words that wound: A tort action for racial insults, epithets, and name-calling". En: *Harv. CR-CLL Rev.* 17 (1982), pág. 133.

[48] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". En: *arXiv preprint arXiv:1810.04805* (2018).

[49] Nemanja Djuric et al. "Hate speech detection with comment embeddings". En: *Proceedings of the 24th international conference on world wide web*. 2015, págs. 29-30.

[50] Natasha Duarte, Emma Llanso y Anna Loup. "Mixed messages? The limits of automated social media content analysis". En: (2017).

[51] Mai ElSherief et al. *Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media*. 2018. arXiv: `1804.04257 [cs.CL]`.

[52] *Facebook flagged Declaration of Independence as hate speech — nypost.com*. `https://nypost.com/2018/07/05/facebook-flagged-declaration-of-independence-as-hate-speech/`. 2018.

[53] Robert Faris et al. "Understanding harmful speech online". En: *Berkman Klein Center Research Publication* 2016-21 (2016).

[54] Seth Flaxman, Sharad Goel y Justin M Rao. "Filter bubbles, echo chambers, and online news consumption". En: *Public opinion quarterly* 80.S1 (2016), págs. 298-320.

[55] Richard Fletcher y Rasmus Kleis Nielsen. "Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication". En: *Journal of Communication* 67.4 (2017), págs. 476-498.

[56] Richard Fletcher y Rasmus Kleis Nielsen. "Are people incidentally exposed to news on social media? A comparative analysis". En: *New media & society* 20.7 (2018), págs. 2450-2468.

[57] Paula Fortuna y Sérgio Nunes. "A survey on automatic detection of hate speech in text". En: *ACM Computing Surveys (CSUR)* 51.4 (2018), págs. 1-30.

[58] Paula Cristina Teixeira Fortuna. "Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes". En: (2017).

[59] Iginio Gagliardone et al. *Countering online hate speech*. Unesco Publishing, 2015.

[60] João Gama et al. "A survey on concept drift adaptation". En: *ACM computing surveys (CSUR)* 46.4 (2014), págs. 1-37.

[61] R Kelly Garrett. "Echo chambers online?: Politically motivated selective exposure among Internet news users". En: *Journal of computer-mediated communication* 14.2 (2009), págs. 265-285.

[62] R Kelly Garrett. "Politically motivated reinforcement seeking: Reframing the selective exposure debate". En: *Journal of communication* 59.4 (2009), págs. 676-699.

[63] Matthew Gentzkow y Jesse M Shapiro. "Ideological segregation online and offline". En: *The Quarterly Journal of Economics* 126.4 (2011), págs. 1799-1839.

[64] Homero Gil de Zúñiga y Sebastián Valenzuela. "The mediating path to a stronger citizenship: Online and offline networks, weak ties, and civic engagement". En: *Communication Research* 38.3 (2011), págs. 397-421.

[65] Palash Goyal y Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey". En: *Knowledge-Based Systems* 151 (2018), págs. 78-94.

[66] Mark S Granovetter. "The strength of weak ties". En: *American journal of sociology* 78.6 (1973), págs. 1360-1380.

[67] Edel Greevy y Alan F Smeaton. "Classifying racist texts using a support vector machine". En: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, págs. 468-469.

[68] Jie Gu et al. "Exploiting behavioral consistence for universal user representation". En: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5. 2021, págs. 4063-4071.

[69] Andrew M Guess. "Media choice and moderation: Evidence from online tracking data". En: *Unpublished manuscript* (2016).

[70] Jurgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991.

[71] Mario Haim, Andreas Graefe y Hans-Bernd Brosius. "Burst of the filter bubble? Effects of personalization on the diversity of Google News". En: *Digital journalism* 6.3 (2018), págs. 330-343.

[72] Kyle A Heatherly, Yanqin Lu y Jae Kook Lee. "Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites". En: *New Media & Society* 19.8 (2017), págs. 1271-1289.

[73] Jeffrey W Howard. "Free speech and hate speech". En: *Annual Review of Political Science* 22 (2019), págs. 93-109.

[74] Bohan Jiang et al. "Mechanisms and Attributes of Echo Chambers in Social Media". En: *arXiv preprint arXiv:2106.05401* (2021).

[75] Julie Jiang, Xiang Ren, Emilio Ferrara et al. "Social media polarization and echo chambers in the context of COVID-19: Case study". En: *JMIRx med* 2.3 (2021), e29570.

[76] Jason J Jones et al. "Inferring tie strength from online directed behavior". En: *PloS one* 8.1 (2013), e52168.

[77] Elıas Jónsson. "An Evaluation of Topic Modelling Techniques for Twitter". En: 2016. URL: https://api.semanticscholar.org/CorpusID:53680644.

[78] Sara Kiesler et al. "Regulating behavior in online communities". En: *Building successful online communities: Evidence-based social design* 1 (2012), págs. 4-2.

[79] Thomas N Kipf y Max Welling. "Variational graph auto-encoders". En: *arXiv preprint arXiv:1611.07308* (2016).

[80] Silvia Knobloch-Westerwick, Cornelia Mothes y Nick Polavin. "Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information". En: *Communication Research* 47.1 (2020), págs. 104-124.

[81] Zachary Laub. *Hate Speech on Social Media: Global Comparisons — cfr.org*. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons. 2019.

[82] Eric Lawrence, John Sides y Henry Farrell. "Self-segregation or deliberation? Blog readership, participation, and polarization in American politics". En: *Perspectives on Politics* 8.1 (2010), págs. 141-157.

[83] Jonathan Leader Maynard y Susan Benesch. "Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention". En: *Genocide Studies and Prevention* 9.3 (2016).

[84] Kalev Leetaru. *Fighting Social Media Hate Speech With AI-Powered Bots — forbes.com*. https://www.forbes.com/sites/kalevleetaru/2017/02/04/fighting-social-media-hate-speech-with-ai-powered-bots/?sh=d579a4b27b10. [Accessed 12-Jun-2023].

[85] Joseph Lilleberg, Yun Zhu y Yanqing Zhang. "Support vector machines and word2vec for text classification with semantic features". En: *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE. 2015, págs. 136-140.

[86] Shuhua Liu y Thomas Forss. "Combining n-gram based similarity analysis with sentiment analysis in web content classification". En: *Special Session on Text Mining*. Vol. 2. SCITEPRESS. 2014, págs. 530-537.

[87] Milton Lodge y Charles S Taber. *The rationalizing voter*. Cambridge University Press, 2013.

[88] VenkataSwamy Martha, Weizhong Zhao y Xiaowei Xu. "A study on Twitter user-follower network: A network based analysis". En: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2013, págs. 1405-1409.

[89] Alice E Marwick y Rebecca Lewis. "Media manipulation and disinformation online". En: (2017).

[90] Binny Mathew et al. "Analyzing the hate and counter speech accounts on twitter". En: *arXiv preprint arXiv:1812.02712* (2018).

[91] Miller McPherson, Lynn Smith-Lovin y James M Cook. "Birds of a feather: Homophily in social networks". En: *Annual review of sociology* 27.1 (2001), págs. 415-444.

[92] Solomon Messing y Sean J Westwood. "Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online". En: *Communication research* 41.8 (2014), págs. 1042-1063.

[93] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". En: *Advances in neural information processing systems* 26 (2013).

[94] John Stuart Mill. *On liberty, utilitarianism, and other essays*. Oxford University Press, USA, 2015.

[95] Karsten Müller y Carlo Schwarz. "Fanning the flames of hate: Social media and hate crime". En: *Journal of the European Economic Association* 19.4 (2021), págs. 2131-2167.

[96] Kevin Munger. "Tweetment effects on the tweeted: Experimentally reducing racist harassment". En: *Political Behavior* 39 (2017), págs. 629-649.

[97] Diana C Mutz. "Cross-cutting social networks: Testing democratic theory in practice". En: *American Political Science Review* 96.1 (2002), págs. 111-126.

[98] Diana C Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.

[99] David G Myers y Helmut Lamm. "The group polarization phenomenon." En: *Psychological bulletin* 83.4 (1976), pág. 602.

[100] Usman Naseem et al. "Transformer based deep intelligent contextual embedding for twitter sentiment analysis". En: *Future Generation Computer Systems* 113 (2020), págs. 58-69.

[101] Edward Newell et al. "User migration in online social networks: A case study on reddit during a period of community unrest". En: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10. 1. 2016, págs. 279-288.

[102] An Nguyen y Hong Tien Vu. "Testing popular news discourse on the "echo chamber" effect: Does political polarisation occur among those relying on social media as their primary politics news source?" En: *First Monday* 24.5 (2019).

[103] Alexandra Olteanu et al. "The effect of extremist violence on hateful speech online". En: *Proceedings of the international AAAI conference on web and social media*. Vol. 12. 1. 2018.

[104] *Organizer of 'Unite the Right' rally loses Twitter verification — whsv.com*. `https://www.whsv.com/content/news/White-nationalist-rally-organizer-loses-Twitter-verification-458189563.html`. 2017.

[105] Margit E Oswald y Stefan Grosjean. "Confirmation bias". En: *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* 79 (2004).

[106] *Our ongoing work to tackle hate — blog.youtube*. `https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/?sjid=15561995000596385536-EU`. [Accessed 05-Jun-2023].

[107] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web.* Inf. téc. Stanford InfoLab, 1999.

[108] Elizabeth Levy Paluck, Seth A Green y Donald P Green. "The contact hypothesis re-evaluated". En: *Behavioural Public Policy* 3.2 (2019), págs. 129-158.

[109] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

[110] Thomas F Pettigrew y Linda R Tropp. "A meta-analytic test of intergroup contact theory." En: *Journal of personality and social psychology* 90.5 (2006), pág. 751.

[111] Daniel Preoţiuc-Pietro et al. "Beyond binary labels: political ideology prediction of twitter users". En: *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*. 2017, págs. 729-740.

[112] Robert D Putnam. "Bowling alone: America's declining social capital: originally published in journal of democracy 6 (1), 1995". En: *Culture and Politics: A Reader* (2000), págs. 223-234.

[113] Jipeng Qiang et al. "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey". En: *CoRR* abs/1904.07695 (2019). arXiv: 1904.07695. URL: http://arxiv.org/abs/1904.07695.

[114] W Quattrociocchi et al. "The echo chamber effect on social media". En: *Proceedings of the National Academy of Sciences Mar* 118.9 (2021).

[115] Manoel Ribeiro et al. "Characterizing and detecting hateful users on twitter". En: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1. 2018.

[116] Anna Schmidt y Michael Wiegand. "A survey on hate speech detection using natural language processing". En: *Proceedings of the fifth international workshop on natural language processing for social media*. 2017, págs. 1-10.

[117] Andrew Sellars. "Defining hate speech". En: *Berkman Klein Center Research Publication* 2016-20 (2016), págs. 16-48.

[118] Jaime E Settle. *Frenemies: How social media polarizes America*. Cambridge University Press, 2018.

[119] Jesse Shore, Jiye Baek y Chrysanthos Dellarocas. "Network structure and patterns of information diversity on Twitter". En: *arXiv preprint arXiv:1607.06795* (2016).

[120] Alexandra A Siegel y Vivienne Badaan. "# No2Sectarianism: Experimental approaches to reducing sectarian hate speech online". En: *American Political Science Review* 114.3 (2020), págs. 837-855.

[121] Laura Silver, Christine Huang y Kyle Taylor. "In emerging economies, smartphone and social media users have broader social networks". En: *Pew Research Center* (2019).

[122] Vivek Kumar Rangarajan Sridhar. "Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words". En: *VS@HLT-NAACL*. 2015. URL: https://api.semanticscholar.org/CorpusID:556501.

[123] Reuters Staff. *Facebook, Google, Twitter agree to delete hate speech in 24 hours: Germany — reuters.com*. https://www.reuters.com/article/us-germany-internet-idUSKBN0TY27R20151215. [Accessed 05-Jun-2023].

[124] Elizabeth Suhay, Emily Bello-Pardo y Brianna Maurer. "The polarizing effects of online partisan criticism: Evidence from two experiments". En: *The International Journal of Press/Politics* 23.1 (2018), págs. 95-115.

[125] Cass R. Sunstein. *Republic.com 2.0*. Princeton University Press, 2007. ISBN: 9780691143286. URL: http://www.jstor.org/stable/j.ctt7tbsw (visitado 07-02-2023).

[126] *Twitter's policy on hateful conduct | Twitter Help — help.twitter.com*. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. [Accessed 05-Jun-2023].

[127] Elise Fehn Unsvåg y Björn Gambäck. "The effects of user features on Twitter hate speech detection". En: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. 2018, págs. 75-85.

[128] Marshall Van Alstyne y Erik Brynjolfsson. "Electronic Communities: Global Villages or Cyberbalkanization?(Best Theme Paper)". En: *ICIS 1996 Proceedings* (1996), pág. 5.

[129] Giacomo Villa, Gabriella Pasi y Marco Viviani. "Echo chamber detection and analysis: a topology-and content-based approach in the COVID-19 scenario". En: *Social Network Analysis and Mining* 11.1 (2021), pág. 78.

[130] Xiao Wang et al. "Community preserving network embedding". En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[131] William Warner y Julia Hirschberg. "Detecting hate speech on the world wide web". En: *Proceedings of the second workshop on language in social media*. 2012, págs. 19-26.

[132] Zeerak Waseem. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter". En: *Proceedings of the first workshop on NLP and computational social science*. 2016, págs. 138-142.

[133] Christopher Whyte. "Cyber conflict or democracy "hacked"? How cyber operations enhance information warfare". En: *Journal of Cybersecurity* 6.1 (sep. de 2020). tyaa013. ISSN: 2057-2085. DOI: 10.1093/cybsec/tyaa013. eprint: https://academic.oup.com/cybersecurity/article-pdf/6/1/tyaa013/33746011/tyaa013.pdf. URL: https://doi.org/10.1093/cybsec/tyaa013.

[134] Xiaohui Yan et al. "A Biterm Topic Model for Short Texts". En: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, págs. 1445-1456. ISBN: 9781450320351. DOI: 10.1145/2488388.2488514. URL: https://doi.org/10.1145/2488388.2488514.

[135] Jianhua Yin y Jianyong Wang. "A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering". En: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: Association for Computing Machinery, 2014, págs. 233-242. ISBN: 9781450329569. DOI: 10.1145/2623330.2623715. URL: https://doi.org/10.1145/2623330.2623715.

[136] Shuhan Yuan, Xintao Wu y Yang Xiang. "A Two Phase Deep Learning Model for Identifying Discrimination from Tweets." En: *EDBT*. 2016, págs. 696-697.

[137] Zhen Zhang et al. "ANRL: attributed network representation learning via deep neural networks." En: *Ijcai*. Vol. 18. 2018, págs. 3155-3161.