Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Análisis de debates partidarios en redes sociales con Procesamiento de Lenguaje Natural

Tesis de Licenciatura en Ciencias de la Computación

María Victoria Zolezzi

Director: Esteban Feuerstein
Codirector: Federico Albanese
Buenos Aires, 2024

Análisis de debates partidarios en redes sociales con Procesamiento de Lenguaje Natural

Abstract. En este trabajo nos dedicamos a analizar distintos aspectos sobre debates partidarios de Estados Unidos en redes sociales. En particular, en Twitter examinamos la evolución temporal del flujo de los usuarios en comunidades políticas. Estas últimas son halladas mediante la aplicación de algoritmos de detección de comunidades en grafos. Por otro lado, en Reddit analizamos el debate del aborto y el debate del control de armas. Clasificando posteos y comentarios por sentimiento, estudiamos diferencias entre las distintas clasificaciones, como ser el sentimiento de sus respuestas, su toxicidad, su engagement, entre otras. Además, analizamos la evolución temporal del sentimiento y del engagement de los debates, prestando especial atención a aquellos momentos en donde el interés por los mismos crece como consecuencia de algún hecho de la realidad. Para intentar determinar qué fenómenos dentro del debate podemos atribuir a estos eventos (por ejemplo un aumento de posteos, de comentarios, de score, etcétera), acudimos a métodos de inferencia causal.

Palabras claves: detección de comunidades, análisis de sentimiento, inferencia causal, toxicidad, redes sociales.

Índice general

1. INTRODUCCIÓN

En los últimos años, ha habido un cambio radical en la forma en la que las personas interactúan y se comunican en línea. Las redes sociales se han convertido en el principal espacio para expresar ideas, participar en debates, consumir y compartir información, y todo a escala global. Con miles de millones de usuarios participando activamente en diversas plataformas, las redes sociales se han convertido en una parte integral de la vida diaria, moldeando no solo las relaciones personales, sino también las dinámicas sociales y el intercambio cultural. Este aumento en la interacción digital ha traído como consecuencia un incremento exponencial en la generación de contenido por parte de los usuarios, desde publicaciones y comentarios basados en texto, hasta contenido multimedia como fotos y videos. Como resultado, hoy en día las redes sociales proporcionan un vasto conjunto de datos que permite a los investigadores explorar patrones, dinámicas y tendencias sobre el comportamiento humano a una escala sin precedentes.

Se han realizado numerosos estudios sobre las redes sociales abordando una amplia variedad de aspectos. A modo de ejemplo, y sin abarcar enteramente la enorme lista de temáticas posibles, podemos mencionar a las llamadas *echo chambers* (comunidades homogéneas y cerradas en las cuales los usuarios tienden a agruparse según sus creencias) [42, 70, 17]; la polarización (formación de grupos opuestos basados en las creencias y opiniones de las personas)[61, 8, 86, 27]; la toxicidad (un comentario grosero, irrespetuoso o poco razonable que probablemente induciría a una persona a abandonar la discusión) [47, 78, 63]; entre otros. Además, existen aplicaciones en múltiples áreas, como ser la política, la psicología o el mercado de valores. Kušen y Strembeck [46] realizan un análisis de sentimiento sobre la discusión en Twitter acerca de las elecciones austríacas de 2016, donde entre otros hallazgos, encuentran diferencias en la sentimentalidad de los candidatos (siendo más neutral en el caso del ganador y más emocional en el del perdedor), que la información negativa sobre ambos candidatos se extendió por más tiempo que la neutral o positiva, y que hubo una clara polarización en términos de los sentimientos difundidos por los seguidores de Twitter de los dos candidatos presidenciales. Dang et al. [20] analizan cómo Twitter es usado para la comunicación política durante los períodos de elección, centrándose en las características y el comportamiento comunicativo de las cuentas influyentes. Tadesse et al. [81] desarrollaron un clasificador para predecir signos de depresión utilizando técnicas de procesamiento de lenguaje natural y machine learning en posteos de Reddit. Pagolu et al. [66] aplican análisis de sentimiento (junto con otras técnicas de machine learning) en tweets para estudiar su relación con los movimientos del mercado de valores, hallando que la positividad en las redes sociales puede influir en la inversión y los precios de las acciones, y que existe una correlación fuerte entre los aumentos y descensos en los precios de las acciones con los sentimientos públicos en los tweets. En esta misma área, Xu y Cohen [90] introducen un modelo que utiliza redes neuronales para predecir movimientos en los precios de las acciones a partir de tweets y precios de acciones históricos.

Dentro de la amplia variedad de herramientas computacionales aplicables a las redes sociales, el Procesamiento de Lenguaje Natural (NLP) resulta extremadamente útil, teniendo en cuenta la enorme cantidad de texto disponible para analizar. En particular,

podemos mencionar el análisis de sentimiento, el cual tiene como objetivo estudiar las opiniones, actitudes y emociones de las personas hacia una entidad, expresadas en un texto. Melton et al. [59] analizan contenido en Reddit relacionado con la vacunación contra el Covid-19, y encuentran que los sentimientos expresados en estas comunidades de redes sociales son en general más positivos que negativos. Li et al. [49] utilizan esta herramienta para analizar los sentimientos, preocupaciones y su variación en el tiempo con respecto a las clases online popularizadas durante la pandemia, por parte de estudiantes y educadores en Reddit. Zhang et al. [92] proponen un sistema para identificar las debilidades de un producto, con el fin de asistir a los fabricantes en la detección de puntos de mejora de sus productos mediante el análisis de sentimiento. Naf'an et al. [62], desarrollan un clasificador de comentarios con el objetivo de identificar aquellos que contengan elementos de cyberbullying. Veletsianos et al. [85] analizan la sentimentalidad en respuestas a videos de YouTube, específicamente sobre charlas TED, en donde encuentran diferencias según el tópico y el presentador. Entre sus hallazgos, encuentran que las respuestas resultan más polarizadas cuando la presentadora es mujer que cuando es hombre, o que el uso de animación neutraliza tanto la positividad como la negatividad.

Asimismo, el análisis de emoción también ha sido aplicado en numerosos estudios sobre las redes sociales. Sailunaz y Alhajj [77] detectan y analizan sentimiento y emoción en tweets con el objetivo de generar recomendaciones personalizadas para los usuarios, basadas en su actividad. Kodati y Tene [44] proponen un modelo que detecta emociones al mismo tiempo que reconoce cuáles son las más comunes en textos relacionados con el suicidio, utilizando posteos en redes sociales.

Existe una amplia variedad de métodos que pueden ser empleados para el análisis de redes sociales además de las mencionadas técnicas de NLP. Un ejemplo de esto sería la inferencia causal. Esta última tiene como objetivo determinar si una asociación observada realmente refleja una relación causa-efecto. Chandrasekharan et al. [15] estudian la prohibición de varios subreddits de discurso de odio y las consecuencias que esta medida trajo al sitio web utilizando inferencia causal. Olteanu et al. [65] examinan cómo los eventos violentos impactan en el discurso de odio en las redes sociales, centrándose especialmente en los ataques que involucran a árabes y musulmanes en países occidentales. Observan que dichos eventos tienden a aumentar el discurso de odio en línea, especialmente los mensajes que promueven la violencia. Empleando técnicas de inferencia causal, Welbers y Opgenhaffen [88] evalúan el impacto de las páginas de Facebook de los periódicos en la difusión de sus noticias, concluyendo que estas páginas ejercen una influencia significativa en la propagación de sus artículos en esta red social.

Más allá de las técnicas, resulta interesante analizar qué aspectos de las redes sociales se han visto modificados a través del tiempo. A continuación mencionaremos algunos de los tantos trabajos que abordan esta temática, sin pretender proporcionar una lista exhaustiva de los mismos. Garimella y Weber [30] investigan si la polarización en Twitter aumenta a lo largo de ocho años, y encuentran que sí lo hace y que varía entre un 10 y un 20 % según la métrica utilizada. Liu et al. [51] analizan siete años de datos de Twitter para cuantificar cómo han evolucionado los usuarios, su comportamiento y la plataforma en su conjunto. Albanese et al. [3] presentan un framework que incluye técnicas de procesamiento de lenguaje natural y algoritmos de machine learning para grafos con el objetivo de identificar "shifting users", es decir usuarios que pueden cambiar de opinión a lo largo del tiempo. Efstratiou et al. [25] analizan trece años de data de Reddit, con el objetivo

de estudiar la relación entre las echo chambers y las interacciones hostiles dentro de una comunidad, donde encuentran que la polarización y la toxicidad son más dominantes entre comunidades del mismo lado del espectro político. Browarnik et al. [40], desarrollan diversas metodologías para entrenar modelos de NLP capaces de identificar comunidades en redes sociales basándose únicamente en su jerga a lo largo del tiempo. Flamino et al. [27] analizan los cambios en el panorama de los medios de comunicación en Twitter entre las elecciones presidenciales de Estados Unidos de 2016 y 2020. Encuentran una disminución en la proporción de contenido falso y extremadamente sesgado entre ambos períodos, y un incremento en los comportamientos de cámara de eco, como así también una polarización ideológica latente durante ambas elecciones tanto a nivel de usuario como de influencer. Crupi et al. [19] investigan cómo evolucionó el debate sobre la vacunación en Italia durante la pandemia del COVID-19 utilizando tres años de data, centrándose principalmente en la polarización del mismo. Ribeiro et al. [74] presentan una caracterización data-driven (basada en el análisis de datos) de la "Manosphere" (un conglomerado de comunidades online misóginas centradas en temas relacionados a la masculinidad) en un período de catorce años, encontrando que las comunidades más antiguas están volviéndose menos populares y activas, mientras que las comunidades más nuevas están prosperando, y que estas últimas resultan ser más tóxicas y misóginas que las antiguas. Garimella et al. [29] estudian la evolución de debates controversiales en Twitter durante seis años, céntrandose en cómo la estructura de interacciones y el contenido de las discusiones varían según el nivel de "collective attention" (volumen de actividad relacionada con ellas en redes sociales). Sus hallazgos revelan que los picos en el interés están asociados con un aumento en la controversia de la discusión y con una convergencia del léxico utilizado por los bandos opuestos. Por otra parte no encontraron ninguna tendencia consistente a largo plazo en la polarización de las discusiones.

Entre los numerosos aspectos de interés vinculados al análisis de redes sociales mediante técnicas de procesamiento de lenguaje natural, así como al efecto del tiempo en dichos análisis, en la presente tesis nos abocaremos a dos análisis principales:

1. Evolución temporal de las interacciones entre usuarios en Twitter: como primer acercamiento, tomamos tweets relacionados con Donald Trump en un período de tres meses, con el objetivo de analizar cómo evoluciona el flujo de los usuarios en las comunidades presentes en esta red a lo largo del tiempo.

2. Análisis de debates controversiales en Reddit: utilizando siete años de datos correspondientes a posteos y comentarios relacionados con el debate del control de armas y el debate del aborto, nos proponemos estudiar la evolución del sentimiento y del engagement de los mismos, como así cambios en estas dimensiones en aquellos momentos en donde aumenta el interés en los debates.

Dentro de los resultados más destacables observamos una consistencia en la forma en la que los usuarios interactúan en ambas redes sociales. En Twitter, los usuarios tienden a mantenerse en sus comunidades políticas a lo largo del tiempo, y por lo tanto a relacionarse con los mismos individuos. En Reddit, con respecto al sentimiento y al engagement, no se identificaron cambios significativos en sus tendencias a lo largo del tiempo, ni tampoco cuando se produce un aumento en el interés en los debates. Por otra parte, en Reddit encontramos que los posteos positivos tienden a recibir más respuestas negativas en comparación a las respuestas de los posteos negativos o neutrales. Sin embargo, esto

no se corresponde con la toxicidad de las mismas, dado que los comentarios de los posteos negativos resultaron ser más tóxicos que los de los posteos positivos.

El resto de esta tesis se estructura de la siguiente manera. En el capítulo 2, se aborda el análisis de la evolución temporal de las interacciones entre usuarios en Twitter, presentando tanto los datos y métodos utilizados, como los resultados obtenidos. El capítulo 3 se centra en el análisis de debates controversiales en Reddit. Allí se describen los datasets y métodos empleados, además de los resultados de la experimentación realizada. Finalmente, en el capítulo 4 se exponen las conclusiones finales de la tesis.

2. EVOLUCIÓN TEMPORAL DE LAS INTERACCIONES ENTRE USUARIOS EN TWITTER

La mayoría de los usuarios de las redes sociales suelen estar expuestos principalmente a contenido que refuerza sus posiciones y los aísla de otras comunidades ideológicas [42]. Como consecuencia, la diversidad de opiniones online no se traduce en debates enriquecedores entre usuarios con diferentes ideologías, sino que los usuarios tienden a agruparse según sus creencias, constituyendo comunidades homogéneas y cerradas conocidas como cámaras de eco (echo chambers) [70]. Si bien algunos autores cuestionan esta idea, dando lugar a opiniones que la matizan un poco [83], existe voluminosa literatura y evidencia de que el fenómeno se presenta en numerosas situaciones. Múltiples estudios analizaron las características de estas comunidades en distintos contextos políticos como el debate sobre el cambio climático [35],el movimiento #BlackLivesMatter [80] o elecciones en Estados Unidos [18], Canadá [32] o Argentina [3]. Otros trabajos científicos caracterizaron los efectos negativos de dichas comunidades cerradas, encontrando que las mismas aumentan el discurso de odio [21], el extremismo político [50], la confirmación de bias [9] o la difusión de fake news [16].

En este contexto, resulta de interés analizar aquellos usuarios y grupos de usuarios en redes sociales que rompen con dicha dinámica e interactúan en múltiples comunidades ideológicamente opuestas. En particular, esos usuarios pueden participar en varias comunidades simultáneamente o variar a través del tiempo. Con este objetivo, analizaremos el flujo de usuarios en comunidades políticas de Estados Unidos a lo largo de tres meses. Las mismas serán halladas mediante la aplicación de un algoritmo de detección de comunidades. Los datos, métodos y resultados serán detallados en las próximas secciones.

2.1. Data Collection

El set de datos utilizado se corresponde con el dataset 2020US del previamente mencionado paper de Albanese et al. [3]. El mismo consiste en tweets públicos que contengan la keyword “realDonaldTrump” (usuario de Twitter del expresidente Donald Trump) datados entre mayo y julio de 2020.

2.2. Métodos

Para poder analizar el flujo de los usuarios en las distintas comunidades, primero es necesario hallarlas. En esta sección describiremos lo métodos utilizados para alcanzar este objetivo.

2.2.1. Red de Retweets

Representamos la interacción entre individuos mediante un grafo dirigido y no pesado, donde los nodos son los usuarios y las aristas los retweets entre ellos (una o más). Pondremos una arista del nodo u al nodo v si el usuario v retweeteó a u. No se consideraron nodos aislados (usuarios que nunca retweetean ni son retweeteados) para este análisis. Cabe aclarar que utilizar los retweets resulta relevante ya que el hecho de retweetear no

solo indica interés en el mensaje, sino también confianza en el mismo y en su autor, así como acuerdo con su contenido [60].

2.2.2. Detección de comunidades

El problema de detección de comunidades implica la partición de un grafo en conjuntos de nodos densamente conectados, donde los nodos pertenecientes a diferentes comunidades tienen conexiones escasas [12]. Dentro de la gran variedad de algoritmos que pueden ser utilizados para la detección de comunidades, analizamos las siguientes cuatro alternativas:

- Stochastic Block Model [39]: se trata de un modelo generativo para bloques, grupos o comunidades en grafos. Cada uno de los n vértices es asignado a uno de los K bloques (comunidades en nuestro caso). Los parámetros del SBM son el número de nodos, un vector de probabilidades de dimensión K y una matriz que expresa la cantidad de aristas entre cada uno de los grupos de dimensión $K \times K$. Las aristas son dispuestas aleatoriamente entre los nodos con probabilidades que dependen de la pertenencia de estos a los grupos. Los nodos de la misma comunidad tienen la misma probabilidad de ser conectados entre sí.

- Nested Stochastic Block Model [68]: este modelo viene a sortear uno de los problemas del SBM: la incapacidad de encontrar comunidades pequeñas en grafos grandes, ya que la cantidad máxima de comunidades que puede hallar es del orden de $O(\sqrt{N})$, con N la cantidad de nodos del grafo. Bajo este modelo, las comunidades se agrupan en un nivel superior constituyendo un multigrafo de bloques, también modelado mediante SBM. En este nuevo grafo, los nodos representan comunidades o bloques, y se añade una arista entre dos bloques si alguno de los nodos que los componen estaba conectado en el nivel anterior. Dado que puede haber múltiples aristas entre dos bloques, el resultado es un multigrafo. Este proceso se repite recursivamente hasta obtener un grafo con un único bloque, creando un nested Stochastic Block Model (NSBM).

- Louvain [12]: se trata de un algoritmo goloso para maximizar la modularidad, la cual se define como la cantidad de aristas que caen dentro de un grupo menos la cantidad esperada en un grafo equivalente cuyas aristas son colocadas aleatoriamente [64]. Está basado en una heurística que equilibra la calidad de la solución, medida por la modularidad, y la complejidad computacional, que escala aproximadamente de manera lineal con el número de aristas [26].

- Infomap [75]: este algoritmo está basado en Random Walks [67] (cadena de Markov de nodos seleccionados aleatoriamente [55]). La idea es que si un grafo tiene una estructura de comunidades significativa, un random walker tiende a estar atrapado en una comunidad por un tiempo considerable antes de viajar a la siguiente [56].

Emmons et al. [26] probaron que la performance de Louvain supera a la de Infomap, por lo que descartamos esta opción. Por otra parte, Peixoto [69] expone varias limitaciones que presentan los métodos que denomina "descriptivos", dentro de los cuales se encuentra el algoritmo Louvain, y cómo son superadas por los métodos de inferencia (SBM y sus variantes). Un ejemplo de esto es el límite de la resolución, que establece que la optimización de la modularidad falla en encontrar comunidades de tamaños menores a $\sqrt{2E}$, con

E la cantidad de aristas del grafo [28]. En contraste, los métodos de inferencia jerárquicos, como ser NSBM, no tienen un límite de resolución significativo. Otro problema de los métodos descriptivos es que pueden encontrar particiones con alta modularidad incluso en grafos aleatorios que no contienen comunidades [33]. La inferencia Bayesiana del SBM es diseñada específicamente para evitar este problema. Por otro lado, Peixoto [69] también menciona que los métodos descriptivos son muy populares debido a las eficientes heurísticas que permiten su aplicación en redes muy grandes, y que existe la creencia de que los métodos basados en SBM son mucho más lentos. Sin embargo, demuestra que esto no es cierto, gracias a los métodos de inferencia modernos que resultan muy competitivos.

Considerando estas ventajas, decidimos utilizar NSBM como nuestro algoritmo de detección de comunidades. Utilizamos la implementación de graph-tool [1].

2.3. Resultados

Separando nuestros datos en tres conjuntos (uno por cada mes de data disponible), creamos una red de retweets para cada uno de ellos. Filtramos los usuarios para considerar solo a aquellos que participaron del debate en los tres meses analizados y descartamos los nodos aislados. De esta forma obtenemos redes de 630704, 884957 y de 851663 nodos para los períodos de mayo, junio y julio respectivamente. Recordemos que los nodos representan a los usuarios, y las aristas los retweets entre ellos.

Luego, corrimos NSBM para intentar hallar la comunidad demócrata y la republicana en cada red. Inicialmente corrimos el algoritmo para que encuentre dos comunidades (este número resulta ser un parámetro del modelo). Obtenida la clasificación, restaba identificar a qué partido pertenecía cada comunidad. Para ello, utilizamos un dataset de Kaggle [2] que contiene usuarios verificados de Twitter junto con su afiliación política. El porcentaje de nodos bien clasificados fue de 97,75 %, 97,42 % y de 97.66 % para los datos de mayo, junio y julio respectivamente. Cabe mencionar que este set de datos solo contiene información sobre la afiliación de 1655 usuarios, y en nuestro caso, contamos con números de usuarios que rondan los 700000. Con lo cual, hay una enorme cantidad de nodos cuya orientación política real no pudimos validar, sin embargo, sabemos que estos interactúan en comunidades en donde participan los usuarios partidarios.

En la figura 2.1 podemos ver cómo fue la clasificación para el mes de mayo. La misma fue generada utilizando Gephi [10], empleando el algoritmo de ForceAtlas2 [41] para el layout, y coloreando los nodos según su comunidad.

[1] `https://graph-tool.skewed.de/static/doc/autosummary/graph_tool.inference.minimize_nested_blockmodel_dl.html`

[2] (`https://www.kaggle.com/datasets/mrmorj/us-politicians-twitter-dataset/data`)

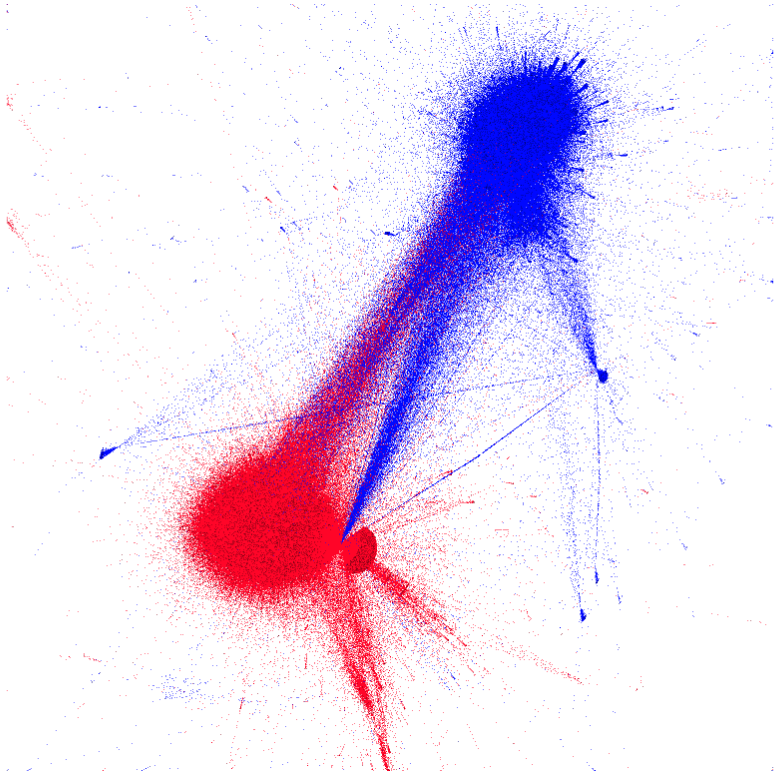


Fig. 2.1: Resultado de la detección de comunidades en el digrafo de retweets del mes de mayo. En rojo podemos encontrar la comunidad republicana (58 %) y en azul la comunidad demócrata (41 %).

Por otra parte, replicamos el experimento utilizando una red de retweets no dirigida. En este caso, conservamos únicamente la componente conexa más grande dentro de los mismos, ya que nuevamente no estamos interesados en los usuarios aislados. De esta forma obtenemos redes de 613259, 858449 y de 830295 nodos para los períodos de mayo, junio y julio respectivamente. La clasificación del mes de mayo puede encontrarse en la figura 2.2.

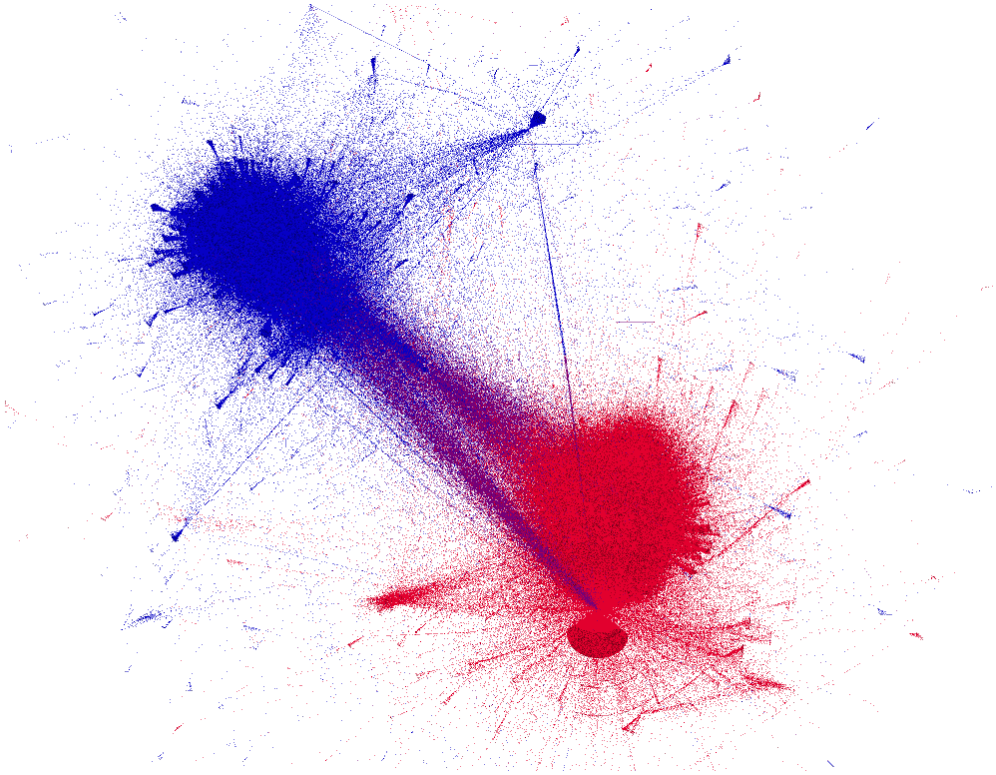


Fig. 2.2: Resultado de la detección de comunidades en el grafo de retweets del mes de mayo. En rojo podemos encontrar la comunidad republicana (59 %) y en azul la comunidad demócrata (40 %).

Viendo la figura 2.2, ambas clasificaciones parecen ser muy similares y de hecho lo son: aproximadamente el 1 % de los nodos fue clasificado en comunidades distintas al aplicar el algoritmo NSBM sobre el grafo y el digrafo.

Una vez obtenidas nuestras clasificaciones, procedimos a analizar el flujo de usuarios entre las tres comunidades a lo largo del período, utilizando los resultados correspondientes al digrafo de retweets. Para visualizar este proceso, construimos un diagrama de Sankey. El mismo se encuentra en la figura 2.3.

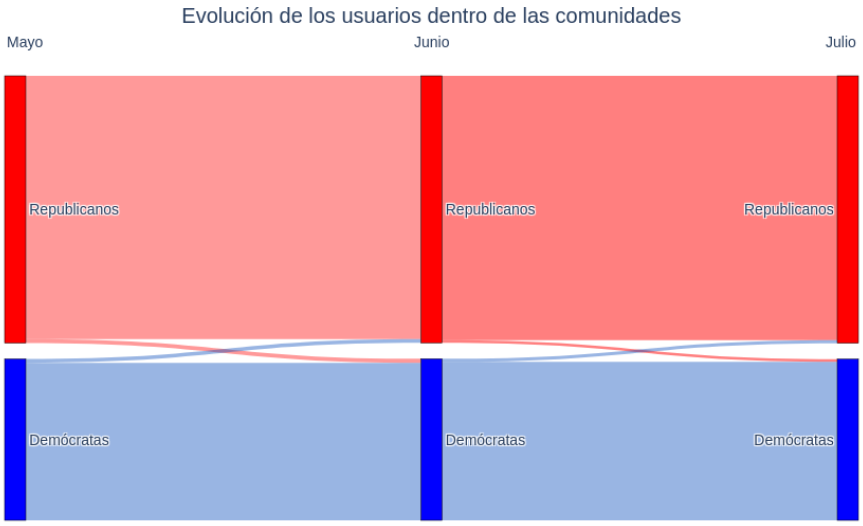


Fig. 2.3: Diagrama de Sankey indicando el flujo de usuarios en las distintas comunidades a lo largo del período estudiado.

En el diagrama 2.3 podemos observar que la gran mayoría de los usuarios se mantiene en su comunidad durante todo el período. Encontramos que 4193 usuarios cambiaron de comunidad entre mayo y junio, y que 3150 lo hicieron de junio a julio. Hubo 1539 usuarios que cambiaron tanto de mayo a junio, como de junio a julio, y todos ellos retornaron a su comunidad original al final del período. Estos resultados nos indican que las personas se mantienen en su grupo con el pasar del tiempo, aunque existan pequeñas fluctuaciones en el medio.

2.4. Conclusiones

Con el objetivo de analizar movimientos dentro de comunidades de eco, en este capítulo del trabajo nos hemos dedicado a clasificar por comunidad política a los usuarios de Twitter cuyo contenido (generado o retweeteado) está relacionado con Donald Trump, en un lapso de tres meses. A partir de esta clasificación observamos cómo los usuarios se trasladan entre las comunidades a lo largo del período. Los resultados señalaron que los mismos tienden a mantenerse en su grupo original, lo que sugiere que que la personas con las que un usuario interactúa se mantienen en el tiempo.

Definitivamente existen limitaciones en nuestra investigación, principalmente relacionadas con la falta de datos. Nuestra intención inicial era replicar este análisis en períodos de tiempo considerablemente más amplios, dado que no resulta esperable que una persona cambie de ideología de un mes a otro. Sin embargo, a causa de dificultades relacionadas con la disponibilidad de los datos de Twitter, por nuestra parte hemos optado por detener el análisis en este punto y continuar analizando debates partidarios en Reddit, ya que esta red social provee una mayor accesibilidad a su data.

Como trabajo futuro, sería interesante repetir el análisis en un marco temporal más extenso, idealmente abarcando varios años, para determinar si este resultado se replica

en dicho contexto. Este enfoque nos proporcionaría una comprensión más completa de los patrones de comportamiento a lo largo del tiempo. Además, se tendría mayor evidencia para catalogar a un usuario como "shifting" (aquellos que rompen con la dinámica de las cámaras de eco e interactúan en múltiples comunidades ideológicamente opuestas), y se podría intentar caracterizar a este grupo, no solo desde el punto de vista de sus interacciones, sino también del contenido que que generan en términos de sentimiento, toxicidad, agresividad y temática.

3. ANÁLISIS DE DEBATES CONTROVERSIALES EN REDDIT

En los últimos años las redes sociales se han posicionado como el espacio por excelencia para debatir una incontable variedad de aspectos de la realidad, convirtiéndose en un foro político, social y cultural abierto las 24 horas del día, los 365 días del año. Este fenómeno ha llevado a la proliferación de debates controversiales en las distintas plataformas. Vamos a decir que un debate es controversial cuando trata sobre un tema específico y existen dos o más posturas bien definidas que se oponen entre sí y que son discutidas ampliamente. En esta sección, nos enfocaremos en analizar diversas características de la evolución temporal de estos debates en Reddit. En particular, nos interesa investigar acerca de los posibles cambios en la manera en que los usuarios participan de los mismos.

Como mencionamos en la introducción, Garimella et al. [29] estudian la evolución de debates controversiales en Twitter a lo largo de seis años, centrándose en cómo la estructura de las interacciones y el contenido de las discusiones cambian según el nivel de "atención colectiva"(volumen de actividad relacionada a las mismas en las redes sociales). Inspirados en este paper, decidimos analizar debates controversiales, pero esta vez en Reddit.

La red social Reddit está estructurada en torno a la participación comunitaria y el intercambio de contenido, donde los usuarios pueden publicar enlaces, imágenes, videos y textos, así como interactuar con otros a través de comentarios y votaciones. Lo que distingue a Reddit es que está organizado en comunidades llamadas subreddits, que comparten temáticas específicas y que además cuentan con reglas y moderadores propios. Esto fomenta la creación de microcomunidades con culturas y normas particulares. Los usuarios pueden votar positiva o negativamente las publicaciones y comentarios, lo que afecta su visibilidad en la plataforma y permite medir su reputación.

Es importante destacar que Reddit y Twitter son dos redes sociales distintas, con dinámicas distintas. Quizás la diferencia más pertinente en este punto resulta ser que no existe un equivalente a la Red de Retweets que podemos armar con datos de Twitter en Reddit, ya que no contamos con ningún mecanismo equivalente que pueda ser tan fuerte en sus implicancias (interés, confianza y acuerdo con el contenido) en esta red social. Por esta razón, el enfoque de nuestra experimentación estará basado en el contenido del debate, medido en sentimiento, emoción y toxicidad, y también en las medidas de engagement: score (número de votos positivos menos los negativos) y cantidad de comentarios).

Los datasets y métodos utilizados para la experimentación serán descriptos en las secciones 3.1 y 3.2 respectivamente. A grandes rasgos, la experimentación consistió en clasificar posteos y comentarios por sentimiento, para luego analizar la evolución del mismo a lo largo del tiempo y posibles cambios en él con el aumento de popularidad de los debates. También se replicó ese análisis para las medidas de engagement. Luego, dejando de lado la evolución temporal, se examinaron las diferencias en las respuestas de los posteos de cada clasificación, tanto en términos de sentimiento, como de toxicidad y engagement. Los detalles y resultados de la experimentación podrán encontrarse en la sección 3.3. Por último, las conclusiones de la misma están en la sección 3.4.

3.1. Data Collection

El dataset de posteos utilizado se corresponde con el dataset "full dataset" del trabajo de Demarco, de Zárate y Feuerestein [22]. En dicho estudio, los autores proponen una técnica basada en texto para cuantificar la alineación de comunidades online a través de diversas dimensiones sociales. El set de datos consiste en los posteos realizados entre 2012 y 2018 en Reddit que contenían texto, ya sea en su cuerpo o en su título.

Dado que solo estamos interesados en el debate del control de armas y en el debate del aborto, seleccionamos una serie de keywords relacionadas con estas temáticas para luego filtrar el dataset de forma tal de conservar únicamente los posteos que mencionen al menos una de estas palabras. Las mismas se componen por las keywords propuestas por Lu et al. [54], a las cuales sumamos algunas otras seleccionadas manualmente, y pueden encontrarse en la tabla 3.1.

Debate	Keywords	#Posteos	#Usuarios
Aborto	abortion, prolife, prochoice, anti-abortion, pro-abortion, planned parenthood	18178	12912
Control de armas	gun control, gun right, pro gun, progun, pro-gun, anti gun, antigun, anti-gun, gunfree, gun law, gun safety, gun violence	8001	5137

Tab. 3.1: Keywords utilizadas para cada debate.

El dataset de comentarios, que puede encontrarse en el siguiente link de HuggingFace `https://huggingface.co/datasets/fddemarco/pushshift-reddit-comments`, está compuesto por todos los comentarios realizados en Reddit entre 2012 y 2016. Los mismos fueron filtrados para conservar únicamente aquellos comentarios que respondían los posteos seleccionados previamente.

Debate	#Comentarios	#Usuarios
Aborto	425237	80651
Control de armas	185877	29336

Tab. 3.2: Cantidad de comentarios y usuarios presentes en el dataset de comentarios.

3.2. Métodos

En esta sección describiremos los métodos utilizados en la investigación.

3.2.1. Natural Language Processing

Antes de presentar los modelos de NLP que empleamos en la experimentación, explicaremos brevemente algunos conceptos importantes para los mismos.

TF-IDF

Sus siglas significan Term frequency - Inverse document frequency. Se trata de una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Un primer approach podría ser buscar las palabras que más se repiten, sin

embargo estas suelen ser las llamadas "stop words" ("el", "la" , "y", "en", etc.) que contrariamente no aportan relevancia semántica, y que de hecho, suelen ser removidas antes de realizar el análisis. Luego, se podría llegar a pensar que son las palabras más raras las que más significado aportan, aunque esto no es del todo cierto, ya que no necesariamente una palabra poco común lo haga. TD-IDF viene a resolver este problema mediante la combinación de dos estadísticas:

- term frequency: mide la frecuencia de un término en un documento. Es calculado dividiendo la cantidad de apariciones de la palabra en un documento sobre la cantidad total de palabras del mismo. Si $f_{t,d}$ es la frecuencia del término t en el documento d, $TF(t,d)$ se define como:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

- inverse document frequency: mide qué tan importante es un término en todos los documentos del conjunto. Se calcula tomando el logaritmo del número total (N) de documentos del corpus (definido como conjunto de documentos) D dividido por la cantidad de documentos que contienen el término en cuestión:

$$IDF(t,D) = log\left(\frac{N}{|\{d : d \in D/t \in d\}|}\right)$$

Luego, TD-IDF es calculado multiplicando estas dos estadísticas:

$$TD\text{-}IDF(t,d,D) = TF(t,d) * IDF(t,d)$$

Algunas de las aplicaciones de esta herramienta son:

- En el campo de la recuperación de información, por ejemplo en los motores de búsqueda. TD-IDF se utiliza para rankear los documentos basados en qué tan relevantes resultan para una query.
- En el resumen de textos y la extracción de palabras claves, ya que el puntaje obtenido para cada término indica qué tan importante resulta para un documento.
- En la vectorización de textos, en donde se convierte a un conjunto de documentos en una matriz de TD-IDF features. Esto permite que los textos crudos se representen con un formato amigable para los modelos de machine learning.

Transformers

Se trata de un modelo de transducción de secuencias basado principalmente en *Attention* [84]. Previo a su aparición, los modelos de transducción de secuencias estaban basados en redes convolucionales o recurrentes. Los Transformers presentan varias ventajas frente a estos modelos tradicionales. Entre ellas, permiten un mejor modelado de las dependencias a largo plazo entre elementos de la secuencia de entrada y soportan procesamiento en paralelo. Su aplicación ha demostrado un rendimiento sobresaliente en diversas tareas

lingüísticas, tales como la clasificación de texto, traducción y *question answering*. También han sido aplicados en otros campos, como ser *computer vision* o *speech processing* [43].

Los Transformers siguen la arquitectura encoder-decoder. Esta consta de dos pasos: la codificación, en la cual se mapea una secuencia de entrada de representaciones simbólicas $(x_1, ..., x_n)$ a una secuencia de representación continua $z = (z_1, ..., z_n)$; y la decodificación en donde, dado z, el decodificador genera una secuencia de salida $(y_1, ..., y_n)$ de símbolos, uno a la vez. En cada paso, el modelo es auto-regresivo, es decir que consume los símbolos generados previamente como input adicional al momento del generar el próximo. En la sección 5.1 se encuentra una explicación más detallada sobre esta arquitectura.

BERT

BERT (Bidirectional Encoder Representations from Transformers) [24] es un modelo de machine learning para procesamiento de lenguaje natural. Este modelo obtuvo resultados de estado del arte en una amplia variedad de tareas de procesamiento del lenguaje natural, como ser Respuesta a Preguntas (SQuAD v1.1 [71]), Inferencia de Sentido Común (SWAG [91]), Análisis de Sentimiento (GLUE [87]), entre otras. Su arquitectura es un Transformer bidireccional multicapa.

El modelo es primero preentrenado con data no etiquetada en diferentes tareas de preentrenamiento. Una de ellas es el llamado "masked language model" (MLM). Este enmascara aleatoriamente algunos de los tokens del input, y el objetivo es predecir el id original del token faltante basándose únicamente en su contexto. Esto permite que la representación fusione el contexto izquierdo y derecho, lo que permite preentrenar un Transformer bidireccional. Otra de las tareas es "predicción de la siguiente oración" que preentrena conjuntamente representaciones de pares de textos. Los datos utilizados para este procedimiento son el BookCorupus de Google y la Wikipedia (en inglés).

Luego el modelo pasa por una etapa de fine-tuning. BERT es inicializado con los parámetros preentrenados, y estos son ajustados utilizando datos etiquetados para las tareas de interés.

RoBERTa

RoBERTa (Robustly optimized BERT approach) [52] es un modelo basado en BERT. La mayor diferencia con respecto a BERT está en su entrenamiento. RoBERTa fue entrenado con un set de datos mucho mayor al de BERT y con mecanismos más efectivos. En particular, fue entrenado con 160GB de texto (supera más de 10 veces el tamaño del dataset usado para entrenar BERT). Además, durante el entrenamiento usa una técnica de enmascarado dinámica, mediante el cual genera un patrón de enmascarado cada vez que una secuencia es pasada al modelo, que ayuda al mismo a aprender representaciones de las palabras más robustas y generalizables. RoBERTa obtuvo resultados de estado del arte en GLUE [87], RACE [48] y SQuAD [72, 91].

3.2.1.1 Análisis de Sentimiento y Emoción

El análisis de sentimiento es una técnica de procesamiento del lenguaje natural utilizada para determinar el sentimiento expresado en un fragmento de texto. Involucra identificar y extraer información subjetiva de datos de texto, categorizándola como positiva, negativa o neutral.

Por su parte, el análisis de emoción tiene como objetivo identificar en un texto distintas emociones humanas, como por ejemplo la felicidad, la tristeza o la preocupación.

En el presente trabajo, para llevar a cabo la clasificación de los textos según sentimiento y emoción utilizamos dos modelos basados en RoBERTa, ambos entrenados con texto únicamente en inglés. En particular, para el análisis de sentimiento usamos el modelo Twitter-roBERTa-base for Sentiment Analysis, que puede encontrarse en `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest`. Está basado en el trabajo de Loureiro et al. [53], en el cual se presenta un conjunto de modelos de lenguaje especializados en datos diacrónicos de Twitter. El modelo clasificará cada texto entre tres labels posibles: *negative, positive y neutral*. Su output consiste en un score asociado a cada etiqueta que representa su probabilidad.

Para el análisis de emoción, elegimos el modelo roberta-base-go_emotions (`https://huggingface.co/SamLowe/roberta-base-go_emotions`), el cual está basado en el paper de Demszky et al. [23]. Los labels son *admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral*. Al igual que en el modelo anterior, vienen acompañados por un score que representa su probabilidad.

3.2.1.2 Detección de Tópicos

La detección de tópicos implica la identificación de la estructura temática subyacente en un corpus de texto. Su propósito es condensar un conjunto de documentos en una serie de temas que sean los más prevalecientes en el mismo.

Para detectar tópicos utilizamos dos modelos:

BERTopic

BERTopic [31] es una herramienta de modelado de tópicos. Esencialmente consta de tres pasos. Con la premisa de que los documentos que contienen el mismo tópico son semánticamente similares, primero convierte cada documento en un embedding que lo represente usando un modelo de lenguaje preentrenado, cuya arquitectura está basada en Transformers (en particular utilizan el framework Sentence-BERT (SBERT) [73]). Estas representaciones vectoriales permitirán comparar a los documentos semánticamente. Luego, para abordar el problema de la "maldición de la dimensionalidad", que surge en espacios de alta dimensión donde el concepto de localidad espacial no está bien definido y las medidas de distancia apenas difieren, se reduce la dimensionalidad de los embeddings utilizando UMAP [58]. Una vez finalizado este proceso, los embeddings son clasificados con HDBSCAN [57]. En el tercer paso las representaciones de los tópicos son modeladas a partir de los documentos de cada grupo, en donde a cada uno de ellos se le asignará un tópico. Para esto, se modifica TD-IDF de forma tal que en vez de representar la importancia de una palabra para un documento, represente la importancia de un término para un tema. Esta variación se denomina c-TD-IDF. Por último, para obtener la cantidad de temas requeridos por el usuario, se combinan iterativamente el tópico menos común con aquel que le sea más similar.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [11] es un modelo probabilístico generativo de un corpus. La idea es modelar cada documento como distribuciones probabilísticas sobre tópicos latentes, donde cada tópico es una distribución probabilística sobre palabras. Una suposición importante que LDA hace, es que documentos con tópicos similares usan grupos de palabras similares. El objetivo es que este modelo asigne probabilidades altas a los documentos del corpus, como así también a documentos similares a ellos. El detalle del modelo generativo puede encontrarse en la sección 5.2.1 del anexo.

En detección de tópicos lo que buscamos no es generar documentos, sino determinar, dado un documento, qué tópicos le corresponden junto con sus probabilidades. Esto lo podemos obtener revertiendo el proceso generativo y aprendiendo la distribución a posteriori de las variables latentes del modelo, dada la data observada. El problema es que esta distribución en general no se puede calcular de manera exacta mediante inferencia. Sin embargo, existen muchos algoritmos de aproximación que se pueden utilizar, como por ejemplo la aproximación de Laplace, Markov chain Monte Carlo, o Inferencia Variacional Bayesiana. Explicaremos brevemente esta última técnica en la sección 5.2.2 del anexo.

En particular, en este trabajo emplearemos la implementación de LDA de Scikit-Learn, que utiliza Inferencia Variacional Bayesiana. Esta implementación está basada en los papers de Hoffman et al. [37] y Hoffman et al. [38].

3.2.1.3 Análisis de Toxicidad

Según Borkan et al. [13], la toxicidad se define como cualquier comportamiento que sea grosero, irrespetuoso o irrazonable, que haga que alguien quiera abandonar una conversación.

En este trabajo, para analizar la toxicidad de los posteos utilizamos Detoxify ([34]). Se trata de una biblioteca open source de Python que identifica comentarios que contienen lenguaje tóxico en siete idiomas (inglés, italiano, francés, ruso, portugués, español y turco). En particular, usamos la versión *original-small*, la cual está basada en BERT y fue entrenada con un dataset de comentarios de Wikipedia manualmente etiquetados para comportamientos tóxicos.

3.2.2. Inferencia Causal

Inferencia causal refiere al proceso de inferir los efectos de cualquier tratamiento, efecto, intervención, etc. Permite establecer relaciones de causa y efecto en lugar de meras asociaciones, y diseñar estrategias para intervenir un sistema. Algunos ejemplos de aplicaciones de inferencia causal son determinar el efecto de algún tratamiento en una enfermedad [76], el efecto de las políticas de cambio climático en las emisiones [2], o el efecto de las redes sociales en la salud mental [82].

Coloquialmente se utiliza el término "correlación" como un sinónimo de dependencia estadística. Formalmente, esta refiere una relación estadística entre dos variables. La técnica más conocida para calcular correlación, y la que utilizaremos a lo largo del trabajo, es el Coeficiente de Correlación de Pearson.
Este mide la relación lineal entre dos variables mediante un coeficiente calculado como una normalización de la covarianza entre ellas, que va de -1 a 1 (fórmula 3.1). Es importante destacar que la correlación no implica causalidad [4]. Esto también ocurre para las relaciones no lineales, las cuales pueden ser medidas por la Información Mutua (fórmula 3.2). Esta cuantifica la

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \quad (3.1)$$

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y} P(x,y)log\frac{P(x,y)}{P(x)P(y)} \quad (3.2)$$

información o reducción de la incertidumbre (entropía) de una variable aleatoria X, frente al conocimiento del valor de otra variable aleatoria Y.

Para cualquier grado de asociación, no necesariamente toda la asociación es causal, o por el contrario, ninguna parte de ella lo es. Puede ocurrir que la cantidad de asociación y la cantidad de causalidad sean distintas. Existen casos en donde dos variables pueden tener una relación de causa y efecto explicada por una causa común entre ellas. Por ejemplo, podríamos llegar a creer que tener un encendedor en el bolsillo causa cáncer de pulmón, cuando en realidad ambas son causadas por fumar. Esta variable se denomina "de confusión" (*confounding variable*).

Para poder determinar el verdadero efecto de una variable X sobre una variable Y se utiliza el contrafáctico, es decir, cómo habría evolucionado la métrica de respuesta después de la intervención si esta nunca hubiera ocurrido: si no ocurrió X, entonces ocurrió Z. Queremos comparar el escenario Y contra el escenario Z para identificar qué efecto tiene X en Y.

3.2.2.1 CausalImpact

Brodersen et al. [14] propusieron un enfoque para estimar el efecto causal de una intervención diseñada en una serie temporal. Dada una serie temporal de respuesta (por ejemplo clicks) y un conjunto de series temporales de control (por ejemplo clicks en mercados no afectados o clicks en otros sitios), se construye un modelo bayesiano de series temporales estructurales. Este modelo se utiliza para predecir el contrafáctico, que luego es comparado con la serie de respuesta para determinar si la intervención tuvo un efecto significativo en esta.

Esta herramienta está originalmente implementada en R. En este trabajo utilizamos su versión en Python.

3.3. Resultados

A continuación presentamos de manera resumida la experimentación realizada.

Inicialmente nos preguntamos si existieron cambios en el sentimiento los debates a lo largo de los años. Para esto, comenzamos por clasificar los posteos por sentimiento y analizar su evolución temporal. El detalle de este experimento y sus resultados pueden encontrarse en la sección 3.3.1.

También nos preguntamos si el sentimiento de los debates presentaba variaciones cuando la popularidad de los mismos aumentaba. En este caso creíamos que era probable que los mismos se vuelvan más negativos en tales circunstancias, o al menos menos neutrales. Este mismo interrogante nos surgió con respecto a las medidas de engagement. Hipotetizamos que los usuarios interactúan más cuando el interés en los debates aumenta. Teniendo en cuenta que consideramos un incremento en la popularidad como un incremento en la cantidad de posteos diarios, en ambos casos necesitábamos determinar si existía alguna relación entre esa cantidad y los posteos de cada clasificación, como así con las medidas de engagement. Para esto, acudimos a la correlación de Pearson y a la Información Mutua. Estos experimentos se encuentran en las secciones 3.3.2 y 3.3.3.

Por otra parte, nos preguntamos si había alguna diferencia en el sentimiento de las respuestas, toxicidad y medidas de engagement de los posteos, dependiendo si estos eran

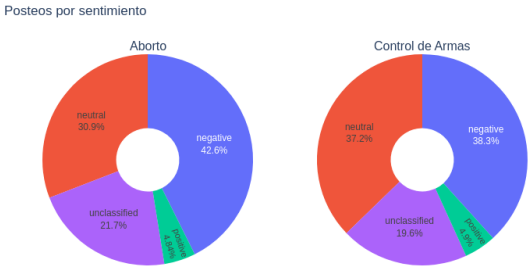
positivos, negativos o neutrales. Esperábamos que esto sí suceda, dado que las personas solemos reaccionar de distinta manera ante mensajes positivos que negativos. En las secciones 3.3.4 3.3.5 se encuentran detalladas las metodologías y resultados de estos experimentos.

Por último, sabiendo que la popularidad de los debates aumenta cuando ocurre algún hecho de la realidad relacionado con los mismos, utilizamos inferencia causal para intentar establecer a estos hechos como lo causantes de ciertas variaciones en los debates, por ejemplo aumentos en la cantidad de posteos o comentarios diarios. La sección 3.3.6 contiene el detalle de este experimento.

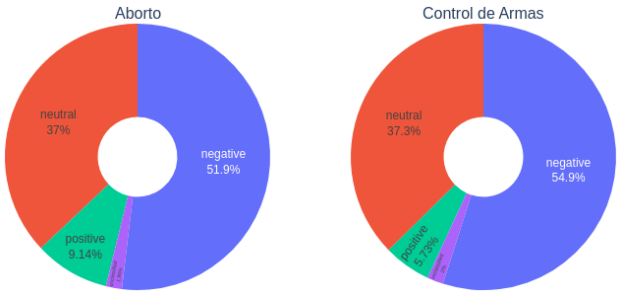
3.3.1. Evolución del sentimiento de los debates

Lo primero que nos gustaría determinar es si podemos encontrar un cambio en el sentimiento del debate a lo largo de los años. En particular, analizar si el debate se volvió más negativo, o menos neutral con el pasar del tiempo.

El primer paso es clasificar los posteos y comentarios. Para eso, tomamos el campo de texto del dataset de posteos y de los comentarios, y lo clasificamos por sentimiento utilizando el modelo mencionado en la sección 3.2.1.1. De cada resultado arrojado por el modelo para cada texto (probabilidades de cada label), nos quedamos con la etiqueta cuyo score fue máximo. En caso de que no se haya podido clasificar el texto debido a su largo, le asignamos la etiqueta "unclassified". Las clasificaciones pueden verse en la figura 3.1.



Posteos por sentimiento.



Comentarios por sentimiento.

Fig. 3.1: Clasificaciones por sentimiento y emoción.

Para analizar la evolución temporal de las clasificaciones por sentimiento, agrupamos la data clasificada de forma tal de quedarnos con la cantidad de posteos (o comentarios) diarios pertenecientes a cada clase. De esta forma obtenemos series temporales como ser

cantidad de posteos *negative* por día, y las normalizamos para obtener la proporción de posteos *negative* por día con respecto al total de posteos para esa fecha.

Al graficar estas series temporales, podemos ver que no hay evidencia de que los debates hayan evolucionado con respecto al sentimiento: las proporciones siguen la misma tendencia a lo largo de todo el período de estudio. Dejamos en la figura 3.4 la evolución temporal del sentimiento de los posteos y de los comentarios para el debate del aborto. En el anexo se encuentran las figuras equivalentes para el debate del control de armas 5.10.

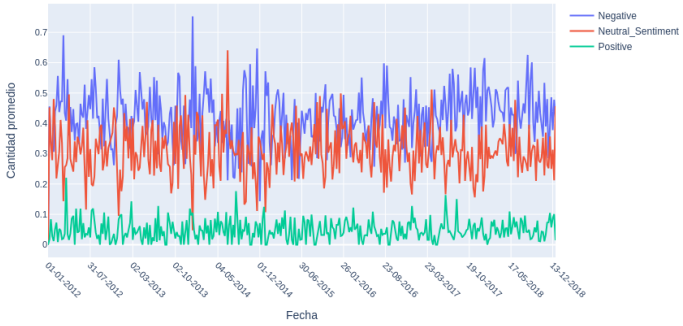


Fig. 3.2: Posteos.

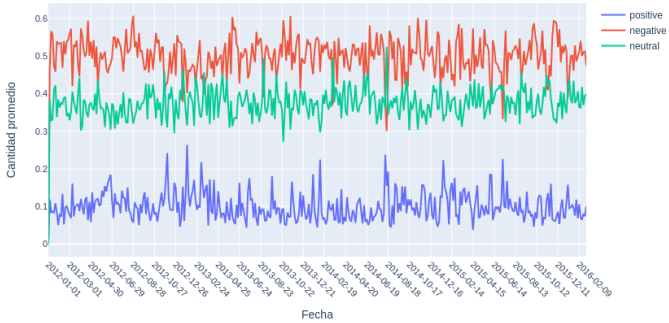


Fig. 3.3: Comentarios.

Fig. 3.4: Evolución temporal del sentimiento de los posteos y comentarios del debate del aborto.

3.3.2. Sentimiento de los debates cuando aumenta el interés

Después de no haber hallado cambios en el sentimiento de los debates a lo largo de los años, nos preguntamos si el mismo se veía modificado en aquellos momentos en donde incrementa el interés en el debate. Hay más de una razón por la cual esto puede suceder, pero generalmente se da cuando ocurre algún hecho de la realidad relacionado con la temática del debate, donde la gente suele comentar al respecto, avivando el mismo. En el caso del debate del control de armas, un ejemplo de estos eventos podría ser un tiroteo.

Vamos a medir el interés utilizando la cantidad de posteos, por lo cual un aumento de estos últimos es relacionado con un aumento del interés. Luego, lo que nos interesa ver es si la serie temporal de posteos diarios tiene alguna relación con las series proporcionales de los posteos por sentimiento. Por ejemplo, si viéramos que la cantidad de posteos diarios aumenta cuando aumenta la proporción de posteos negativos (o viceversa), podríamos decir que el aumento en el interés en el debate está relacionado con un aumento en la negatividad en el mismo.

Para intentar encontrar esas relaciones utilizamos la correlación de Pearson, la cual indica si existe una relación lineal entre dos variables. En nuestro caso, las variables serán la cantidad de posteos diarios y la proporción de posteos de cada clasificación.

Debate	Sentimiento	Corr. con posteos	p-value (corr)	IM con posteos
Aborto	Negative	-0.043	0.030	0.028
	Positive	0.011	0.563	0.002
	Neutral	-0.005	0.792	-0.055
Control de Armas	Negative	0.021	0.339	0.211
	Positive	-0.001	0.974	-0.026
	Neutral	-0.012	0.596	0.221

Tab. 3.3: Correlación e Información Mutua de la serie temporal de posteos diarios y las series proporcionales de cada clasificación.

Como podemos ver en la tabla 3.3, la mayoría de las correlaciones no son significativas (p-valor $> 0{,}05$), y las que sí lo son, tienen valores muy cercanos a 0. Esto quiere decir que no se encontró ninguna relación lineal entre la cantidad de posteos diarios y las proporciones diarias de cada clasificación para ninguno de los dos debates. Frente a estos resultados, intentamos calcular la correlación agrupando las series por semana en vez de por día, sin éxito. También probamos shifteándolas, es decir, adelantando o atrasando las distintas series temporales. Esto nos permitiría ver, por ejemplo, si los posteos diarios del día i tenían alguna relación con la proporción de posteos de las clasificaciones del día $i+j$, con j la cantidad de veces que shifteamos la serie. En todos los casos obtuvimos resultados similares.

Luego de haber descartado la existencia de una relación lineal entre las variables, nos preguntamos si estaban relacionadas de forma no lineal. Para eso procedimos a calcular su Información Mutua.Podemos ver los resultados en la tabla 3.3. Nuevamente, los valores obtenidos son muy cercanos a 0, con lo cual descartamos la existencia de una información mutua significativa que avale la relación entre la serie temporal diaria y las proporcionales por clasificación. Una vez más, intentamos shifteando las series y agrupándolas semanalmente, aunque no encontramos mayores hallazgos.

Con estos resultados, concluimos que no existe relación entre las series temporales estudiadas, y con lo cual, no podemos decir que cuando aumenta el interés en un debate (medido en el volumen de posteos) ocurre un cambio en el sentimiento del mismo, o viceversa. Originalmente, queríamos probar la existencia de una relación causal entre la cantidad de posteos y el sentimiento de los mismos, es decir que el incremento en la popularidad del debate provocaba cambios en su contenido (medido en sentimentalidad). Sin embargo, dado que para que exista una relación causal entre dos variables su información mutua debe ser alta, descartamos la existencia de la misma.

3.3.3. Efectos del aumento del interés en los debates en las medidas de engagement

En esta sección, nos abocaremos a estudiar los posibles efectos del aumento del interés en las medidas de engagement de los debates. Estas son el número de comentarios de los posteos y su score (calculado como la cantidad de votos positivos menos la cantidad de votos negativos). En particular, qué sucede con ellas cuando aumenta en el interés en los debates: ¿los usuarios comentan más?, ¿premian más a los posteos?

Para determinarlo, repetimos el procedimiento de la sección 3.3.2: construimos las series temporales con los promedios diarios del score y del número de comentarios de los posteos, y calculamos su correlación e información mutua con la serie temporal de posteos diarios. La idea nuevamente es intentar ver si encontramos alguna relación entre los posteos diarios y estas nuevas series temporales de engagement, de forma tal de determinar si el interés en el debate (medido como un aumento de los posteos) guarda relación con estas medidas.

Debate	Medida	Corr. con posteos	p-value (corr.)	IM con posteos
Aborto	Comentarios	0.036	0.07	0.803
	Score	0.064	0.00	0.806
Control de Armas	Comentarios	-0.015	0.49	0.851
	Score	0.003	0.91	0.753

Tab. 3.4: Correlación e Información Mutua (IM) de la serie temporal de posteos diarios y las series de engagement.

Los resultados del cálculo de la correlación pueden encontrarse en la tabla 3.4, en la cual podemos observar que (en los casos en la que es significativa), resulta ser muy cercana a 0, indicando que no existe una relación lineal entre las series.

Con respecto a la información mutua, los resultados también se encuentran expuestos en la tabla 3.4. Podemos ver que los valores resultan ser altos (como referencia la información mutua entre la serie de posteos y ella misma es 3.817). Esto indica que existe información mutua entre la serie temporal de posteos diarios y las series diarias de engagement, lo que sugiere que cambios en las medidas de engagement pueden ser explicados parcialmente por la serie temporal de posteos diarios.

Por otra parte, calculamos la información mutua entre las series temporales de cantidad promedio de comentarios y de score, la cual fue de 2.346 y 2.835 en el debate del aborto y en el del control de armas respectivamente. Es probable que tanto estos valores como los valores expuestos en la tabla 3.4 puedan ser explicados por el funcionamiento del algoritmo que utiliza Reddit para ordenar el contenido en el feed por default. El mismo utiliza la fecha y el score para ubicar los posteos en el feed, otorgándole una mayor visibilidad a los posteos más recientes mejor puntuados [1]. Es esperable que aquellos posteos que más se muestran sean los que más comentarios reciban, y que esta sea la razón de los valores altos de información mutua hallados.

Lo que nos interesaría ver ahora es si los hechos de la realidad que producen un aumento de interés en el debate (aumento en la cantidad de posteos) son los causantes de estas fluctuaciones en las medidas de engagement, y no el algoritmo. Para probar esto formalmente, en una próxima sección 3.3.6, analizaremos la existencia de una relación causal entre hechos de la realidad que avivan el debate del control de armas y sus posibles efectos en estas medidas.

3.3.4. Sentimiento y toxicidad de los comentarios de los posteos positivos, negativos y neutrales

Considerando que vimos que el sentimiento de los debates no se vio modificado significativamente con aumentos en su popularidad ni con el paso del tiempo, en esta sección

[1] https://support.reddithelp.com/hc/es-es/articles/23511859482388-El-enfoque-de-Reddit-con-respecto-a-las-recomendaciones-de-contenido

nos propondremos responder la pregunta "¿hay alguna diferencia en el sentimiento de las respuestas a los posteos negativos, positivos y neutrales?". Es decir, ¿el sentimiento de las respuestas depende en algún punto del sentimiento del posteo al que contestan?

Para responder esta pregunta, tomamos nuestros posteos clasificados y analizamos las clasificaciones de sus respuestas. Cabe aclarar que en este caso, utilizamos únicamente los posteos y comentarios cuyas clasificaciones hayan obtenido un score mayor a 0,8. Esto nos permite trabajar solo con aquellos textos que con una alta probabilidad pueden considerarse mayoritariamente negativos, positivos o neutrales, y no una combinación de estos.

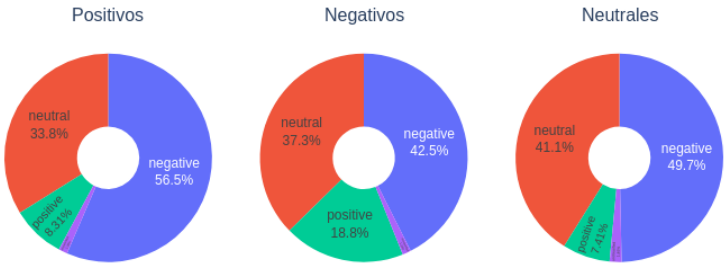


Fig. 3.5: Clasificación de los comentarios de los posteos del aborto.

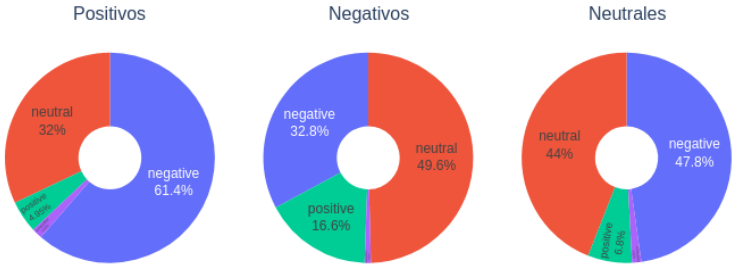


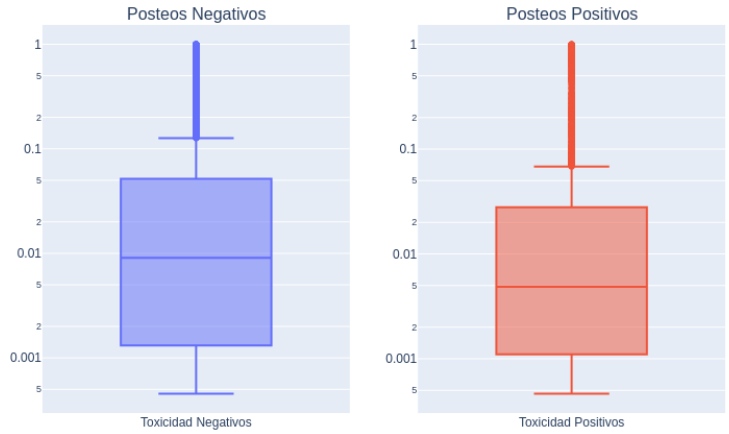
Fig. 3.6: Clasificación de los comentarios de los posteos del control de armas.

Fig. 3.7: Clasificación de las respuestas a los posteos según su sentimiento.

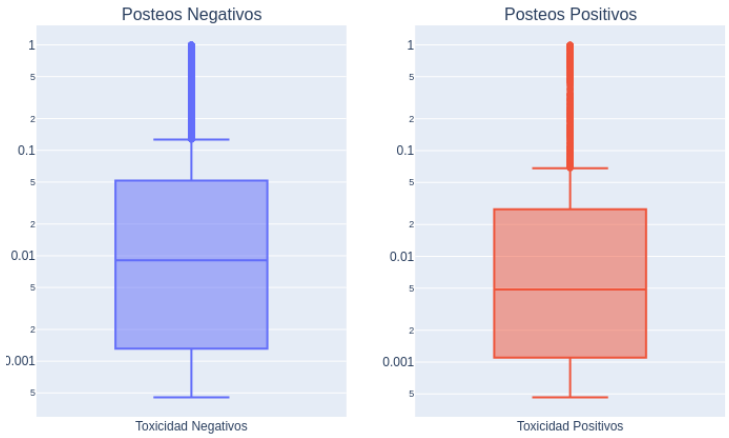
Posteos Positivos

Como podemos ver en la figura 3.7, en ambos debates encontramos que los posteos positivos reciben una mayor cantidad de respuestas negativas comparados con las respuestas de los posteos negativos y neutrales. Esto nos hace preguntarnos si los posteos positivos generan también más comentarios tóxicos que los negativos. Recordemos que un texto negativo no es necesariamente tóxico, podría ser triste, expresar preocupación, miedo, etcétera.

Utilizando la biblioteca Detoxify, clasificamos los comentarios utilizados en la sección anterior según su toxicidad. Los dividimos según la clasificación en sentimiento del posteo al que respondían, con el objetivo de analizar si encontrábamos diferencias en la toxicidad de los comentarios de los posteos negativos y positivos.



Debate del aborto.



Debate del control de armas.

Fig. 3.8: Boxplots sobre el score de la toxicidad de los comentarios a posteos negativos y positivos de los debates.

Luego, utilizamos el test U de Mann-Whitney (prueba no paramétrica de la hipótesis nula de que la distribución subyacente de la muestra x es la misma que la distribución subyacente de la muestra y) para comparar las distribuciones de la toxicidad de las respuestas de cada grupo. En la figura 3.8 se encuentran los boxplots correspondientes al score de toxicidad de los mismos, para ambos debates. Para ambos debates pudimos determinar que la distribución del nivel de toxicidad de los comentarios de posteos negativos era mayor que la de los posteos positivos. Esto confirma que, aunque los posteos positivos reciban respuestas más negativas, estas no resultan ser más tóxicas. En la tabla 3.5 mostramos el promedio y la mediana de la toxicidad de los comentarios de los posteos positivos y negativos de ambos debates.

Debate	Resp. a posteos	Tox. Promedio	Mediana de la Tox.
Aborto	Negative	0.111	0.009
	Positive	0.066	0.005
Control de Armas	Negative	0.126	0.007
	Positive	0.078	0.003

Tab. 3.5: Promedio y mediana de la toxicidad de los posteos negativos y positivos.

Posteos Negativos

Por otra parte, encontramos una diferencia entre los debates: los posteos negativos del debate del control de armas reciben en su mayoría respuestas neutrales, mientras que en el debate del aborto, negativas. Esto nos dice que el sentimiento de las respuestas no solo depende del sentimiento del posteo al que responden. Nos preguntamos si esta diferencia también se traduce en que los posteos negativos del debate del aborto generen respuestas más tóxicas que los del debate del control de armas. Para intentar generar una intuición sobre esto, comenzamos por clasificar estos comentarios por emoción, utilizando el modelo explicado en la sección 3.2.1.1.

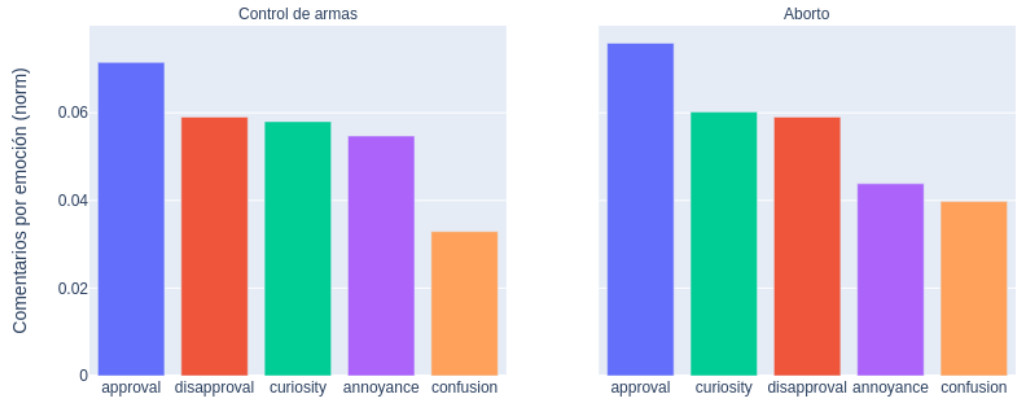


Fig. 3.9: Emociones predominantes en los comentarios a posteos negativos de cada debate. En estos gráficos no se muestra la emoción neutral, que es mayoritaria en ambos debates (56 % y 49 % respectivamente).

En la figura 3.9 podemos ver las principales emociones de las respuestas a posteos negativos de ambos debates. Estas resultan ser prácticamente iguales, con lo cual, no nos dan ningún indicio sobre nuestra pregunta original. Frente a esto, decidimos clasificar los posteos de ambos debates.

En la figura 3.10 tenemos dos gráficos de torta mostrando las proporciones de cada emoción de los posteos de los debates. Dejando de lado las clases “other” y “unclassified”, notamos que las emociones predominantes en el debate del aborto son tristeza, curiosidad, gratitud y miedo; mientras que en el debate del control de armas estas son curiosidad, gratitud, confusión y fastidio. Si ahora nos concentramos en los posteos negativos, las emociones predominantes en el debate del aborto son tristeza, miedo y confusión; y en el debate del control de armas fastidio, curiosidad y desaprobación. Estas pueden verse gráficamente en la figura 3.11.

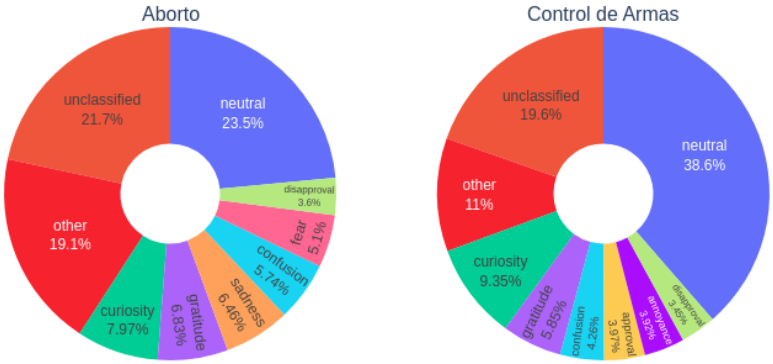


Fig. 3.10: Clasificación de los posteos de los debates según emoción. En el diagrama añadimos la clase "other" para representar a las emociones menos comunes que no representaban una proporción significativa del total de posteos.

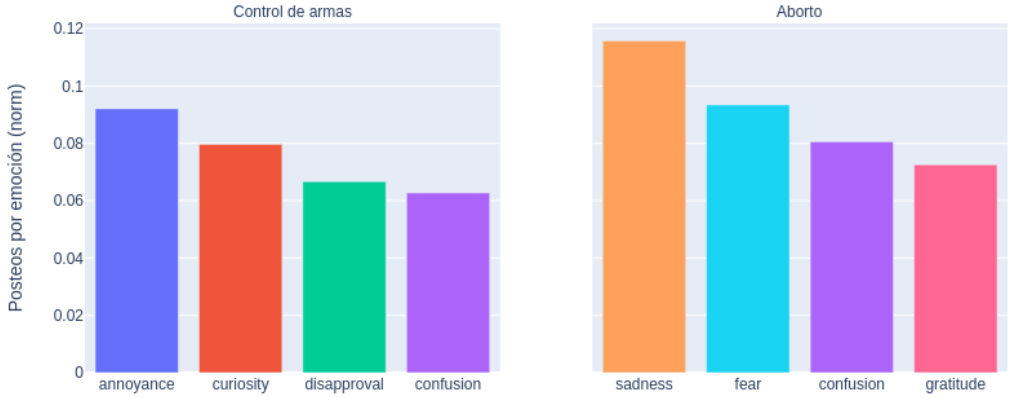


Fig. 3.11: Emociones predominantes en los posteos negativos de cada debate. En estos gráficos no se muestra la emoción neutral, que es mayoritaria en ambos debates (49 % y 26 % respectivamente).

Viendo las emociones predominantes, creemos que resulta natural pensar que los posteos de fastidio y desaprobación están más relacionados con la toxicidad que los de tristeza o miedo. Para probar esto formalmente, clasificamos los posteos negativos según toxicidad y aplicamos el test U de Mann-Whitney sobre las distribuciones resultantes. Obtuvimos resultados significativos que indicaron que el nivel de toxicidad de los posteos negativos del debate del control de armas es superior al del aborto. En la figura 3.12 se exponen los boxplots del nivel de toxicidad de los posteos negativos de ambos debates.

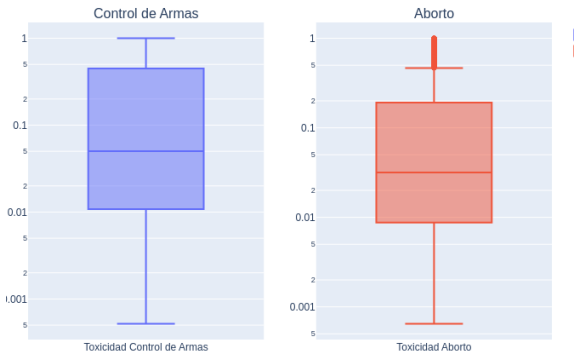


Fig. 3.12: Boxplot sobre el nivel de toxicidad de los posteos negativos de ambos debates.

Con este resultado entonces sabemos que los posteos negativos del aborto son menos tóxicos que los del control de armas. Esperamos que las respuestas del primero también sean menos tóxicas que las del segundo.

Finalmente, para responder a nuestra pregunta clasificamos los comentarios de los posteos negativos por toxicidad. Podemos ver las distribuciones del nivel de toxicidad de los mismos en la figura 3.13.

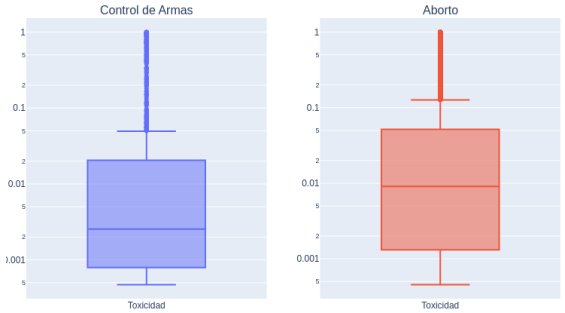


Fig. 3.13: Boxplot sobre los niveles de toxicidad de los comentarios de los posteos negativos de ambos debates.

Utilizando el test U de Mann-Whitney, probamos que el nivel de toxicidad de las respuestas a posteos negativos del control de armas es menor a los del aborto, contrario a lo que creíamos. Este resultado nos indica que, a pesar de compartir temática, la toxicidad en los posteos se comporta de manera distinta a la de sus respuestas.

Luego de estos experimentos pudimos confirmar que el sentimiento de los posteos no determina ni se comporta de la misma forma que el sentimiento de sus respuestas, y que lo mismo ocurre para la toxicidad.

3.3.5. Medidas de engagement de los posteos negativos, positivos y neutrales

Ahora nos preguntamos si existe alguna diferencia entre las medidas de engagement de los posteos según si son negativos, positivos o neutrales. En la figura 3.16 podemos ver los histogramas correspondientes a cada medida para el debate del aborto. En la figura 5.13 del anexo, se encuentran los histogramas respectivos para el debate del control de armas.

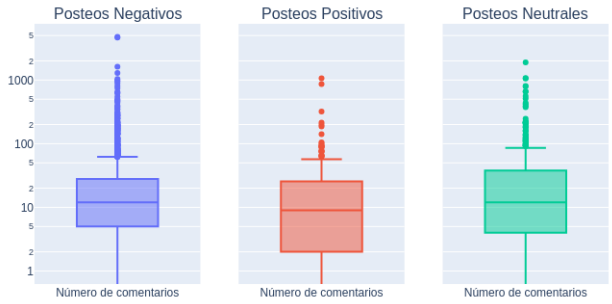


Fig. 3.14: Boxplot del número de comentarios por posteo.

Fig. 3.15: Boxplot del score por posteo.

Fig. 3.16: Boxplots sobre el score y el número de comentarios del debate del aborto, por cada sentimiento.

Una vez más, acudimos al test U de Mann-Whitney para comparar las distribuciones de la cantidad de comentarios y el score de cada una de las clasificaciones entre sí.

En ambos debates, el test concluyó significativamente que los posteos negativos tienden a ser más contestados que los positivos. En la sección 3.3.4 habíamos determinado que los posteos negativos presentan niveles de toxicidad superiores a los positivos. Juntando ambos resultados, obtenemos que los posteos negativos, que resultan más tóxicos, obtienen una mayor cantidad de respuestas que los positivos, que son menos tóxicos. Esto parece no coincidir con la idea de que un posteo tóxico induce a las personas a abandonar la conversación. Sin embargo, en varios trabajos previos se encuentra evidencia de que el número de comentarios aumenta con el nivel de toxicidad [89, 79, 78].

Por otra parte, los posteos positivos reciben mejor score que los negativos y que los neutrales.

Replicamos este análisis con los comentarios positivos, negativos y neutrales (solo para el score). El test determinó que los comentarios positivos son mejor puntuados que los neutrales y negativos en ambos debates. La diferencia está en que en el debate del aborto, los comentarios negativos son mejor puntuados que los neutrales, y en el debate del control de armas, los negativos son peor puntuados que los neutrales. Esto nos indica que el score no depende solamente del sentimiento del texto, si no que la temática tiene incidencia en él también.

3.3.6. Efectos de hechos de la realidad en el debate del control de armas

En esta sección, nos planteamos la pregunta de qué efectos tienen en el debate los hechos que aumentan el interés en el mismo. En particular, tomaremos como eventos a tiroteos, ya que cuando estos ocurren, reavivan el debate del control de armas [29]. Intentaremos ver qué efectos podemos atribuirles sobre la cantidad de posteos, su sentimiento y las medidas de engagement. Para esto, utilizaremos CausalImpact, descripto en la sección 3.2.2.1. Además de la serie temporal de respuesta, la herramienta requiere que le pasemos un conjunto de series temporales de control que no deben haber sido afectadas por el evento. En nuestro caso elegimos las series temporales de posteos diarios de subreddits más generales, como ser *r/news*,*r/climate*, *r/science*, *r/environment* y *r/health*. También utilizaremos sus series diarias proporcionales de score y cantidad de comentarios.

Empezamos analizando el efecto de los tiroteos en la cantidad de posteos diarios. Antes de utilizar CausalImpact realizamos algunas pruebas preliminares. Primero nos concentramos en identificar si existió alguna variación en la cantidad de posteos que le pueda ser atribuido al tiroteo en cuestión. En 6 de los 19 tiroteos seleccionados (que pueden encontrarse en la tabla 3.6) no hallamos ninguna, con lo cual los descartamos. En segundo lugar, utilizamos BERTopic y LDA (ambos descritos en la sección 3.2.1.2) para intentar determinar si los posteos hablaban sobre temáticas relacionadas al tiroteo en períodos cercanos al mismo, ya que podría haber ocurrido que los usuarios estén discutiendo sobre otros tópicos. En particular, 3 de los 19 tiroteos coincidieron con períodos de elección, y las publicaciones contemporáneas a ellos abordaban predominantemente esta temática en lugar del tiroteo.

Fecha	#Heridos	Estado
20-07-2012	82	Colorado
14-12-2012	26	Connecticut
16-09-2013	15	Washington DC
23-05-2014	20	California
18-09-2014	8	Florida
17-05-2015	27	Texas
17-06-2015	19	South Carolina
01-10-2015	17	Oregon
02-12-2015	35	California

Fecha	#Heridos	Estado
20-02-2016	8	Michigan
22-04-2016	8	Ohio
12-06-2016	102	Florida
14-06-2017	6	Virginia
01-10-2017	+900	Nevada
05-11-2017	48	Texas
14-02-2018	34	Florida
18-05-2018	23	Texas
27-10-2018	17	Pennsylvania
07-11-2018	28	California

Tab. 3.6: Tiroteos a analizar. Estos fueron obtenidos del Gun Violence Archive (`https://www.gunviolencearchive.org/`). Los tiroteos marcados con rojo son aquellos para los que no se identificó ninguna variación en la serie de posteos diarios cercanos a esa fecha. Con naranja marcamos los tiroteos contemporáneos a un aumento de posteos que hablaban sobre otro tema.

Una vez filtrados los tiroteos, conservamos 10. CausalImpact nos pide establecer un período previo y un período posterior al evento. La extensión de los mismos fue elegida manualmente para cada caso, teniendo en cuenta evitar períodos con fuerte presencia de outliers en la cantidad diaria de posteos, y varió de entre 15 días a 2 meses dependiendo del tiroteo. Por otra parte, en algunos casos las series presentaban demasiadas variaciones, por lo que utilizamos rolling window de 5 días para sustituir cada valor por el promedio de ese día y los 4 días anteriores, y de esa forma suavizar las series. En todos los casos, la herramienta determinó que el tiroteo causó un aumento en la cantidad de posteos, es

decir que incrementó el volumen del debate.

A modo de ejemplo, tomamos el tiroteo del 14 de diciembre de 2012, conocido como Sandy Hook Elementary School shooting. En la tabla 3.7 se exponen los tópicos hallados por BERTopic. En la figura 3.17 y en la tabla 3.8 se encuentran la visualización del output de CausalImpact y su resumen.

Nombre del Tópico	Representación	Cantidad de posteos
0_gun_guns_would_people	gun, guns, would, people, control, weapons, could, im, think, background	38
1_gun_like_control_school	gun, like, control, school, people, could, health, mental, way, dont,	24
2_gun_guns_people_control	gun, guns, people, control, amendment, would, weapons, dont, government, defend,	24
3_gun_control_like_blaming	gun, control, like, blaming, time, talk, know, want, tragedy, im,	22
4_school_age_children_students	school, age, children, students, people, one, killed, guns, shot, two	12

Tab. 3.7: Output de BERTopic correspondiente a los posteos del debate del control de armas posteados entre el 14 y el 29 de diciembre de 2012.

En la siguiente figura (3.17) encontramos la visualización del output de CausalImpact, conformado por tres paneles. El primero muestra los datos y una predicción contrafáctico para el período posterior al tratamiento. El segundo, muestra la diferencia entre los datos observados y las predicciones contrafácticas. Este es el efecto causal puntual, según lo estimado por el modelo. El tercer panel suma las contribuciones puntuales del segundo panel, dando como resultado un gráfico del efecto acumulativo de la intervención.

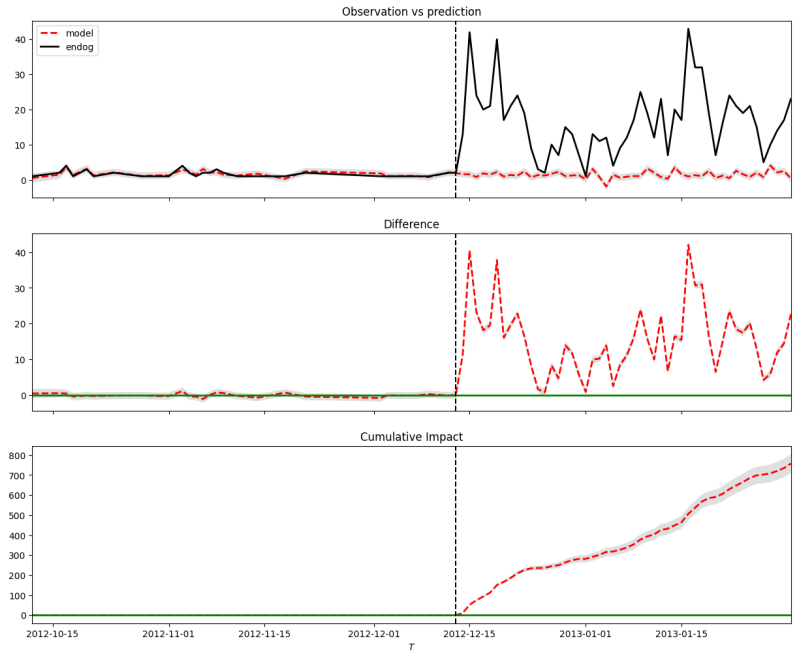


Fig. 3.17: Visualización del output de CausalImpact.

	Average	Cumulative
Actual	16	827
Predicted	1	69
95 % CI	[0, 2]	[22, 116]
Absolute Effect	15	757
95 % CI	[16, 14]	[804, 710]
Relative Effect	1089.6 %	1089.6 %
95 % CI	[1157.4 %, 1021.7 %]	[1157.4 %, 1021.7 %]

Tab. 3.8: Resumen del output de CausalImpact.

En la tabla 3.8 encontramos el resumen del output de CausalImpact. La columna "Average" se refiere al promedio (a lo largo del tiempo) durante el período siguiente al tiroteo. La columna "Cumulative" suma los puntos de tiempo individuales, que en nuestro caso es la cantidad de posteos. Aunque no se muestra en la tabla, el resumen también contiene el p-valor (0.2 %) y la probabilidad del efecto causal (99.8 %).

Además de analizar la cantidad de posteos, utilizamos la herramienta para intentar establecer a los tiroteos como causas de cambios en el sentimiento de los posteos y comentarios publicados cercanos a la fecha en la que ocurrieron. Sin embargo, este no se vio modificado en ningún caso luego de un tiroteo. Este resultado se condice con los de la sección 3.3.2, en donde no habíamos podido encontrar ninguna relación lineal ni no lineal entre las series temporales de posteos diarios y las series proporcionales de sentimiento.

También tomamos como efecto cambios en las medidas de engagement (cantidad de comentarios y score). La herramienta determinó que ningún tiroteo tuvo efecto en ellas. Volviendo a la teoría acerca del por qué de los valores de información mutua hallados en la sección 3.3.3, podemos confirmar que los cambios en las medidas de engagement no son causados por los eventos que aumentan la popularidad del debate, y que probablemente sí se relacionen con el algoritmo de Reddit.

Cabe aclarar que intentamos replicar este experimento para el debate del aborto, sin embargo, no encontramos hechos particulares relacionados con el mismo que se hayan podido ver reflejados en la cantidad de posteos, al menos no un número de eventos significativo para poder generalizar los resultados.

3.4. Conclusiones

En este capítulo de la tesis nos abocamos al análisis de dos debates controversiales en Reddit, utilizando técnicas de NLP e inferencia causal. Primeramente, nos dedicamos a analizar cambios en el debate a lo largo 6 años, y también en aquellos momentos en donde el interés por los debates aumenta. En particular, estudiamos el sentimiento de los debates y su engagement. Ambas dimensiones resultaron no verse alteradas en ninguno de los dos casos. Estos resultados nos indican que la forma en la que las personas se expresan (a nivel de sentimiento) e interactúan con otros usuarios (cómo votan, qué tanto comentan) no varía con el pasar de los años, como así tampoco cuando los debates ganan popularidad.

Por otra parte, estudiamos cómo se diferencia el contenido generado como respuesta a los posteos, dependiendo de su sentimiento. En cuanto al engagement, utilizamos el test

U de Mann-Whitney para comparar las distribuciones de cantidad de comentarios y score de los posteos de cada clasificación. El test determinó en ambos debates que los posteos negativos son más contestados que los positivos, y que los positivos son más votados que los negativos.

Además, comparamos las proporciones de respuestas negativas, positivas y neutrales de los posteos de cada clasificación. Notamos que en ambos debates los posteos positivos reciben mayoritariamente respuestas negativas. Por un lado, esto nos indica que no existe una relación directa entre el sentimiento de los posteos y el de sus respuestas. Por otro, nos hizo preguntarnos si los posteos positivos también recibían respuestas más tóxicas. Empleando el test U de Mann-Whitney para comparar las distribuciones de toxicidad de las respuestas de los posteos de cada clasificación, pudimos determinar que esta hipótesis no era cierta, dado que los posteos negativos recibían respuestas más tóxicas que los positivos.

Asimismo, observamos que las proporciones de respuestas negativas, positivas y neutrales de los posteos de cada clasificación no se presentan de igual manera en ambos debates. En particular, el sentimiento predominante de las respuestas de los posteos negativos del debate del aborto fue negativo, y en el del control de armas neutral. Esto nos indica que la temática de los posteos es una variable que influye en el sentimiento de sus respuestas.

Por último, utilizando inferencia causal analizamos qué efectos tuvieron hechos de la realidad, específicamente tiroteos, en el debate del control de armas. En múltiples ocasiones encontramos que cuando ocurre un tiroteo los usuarios aumentan su participación en el debate, aunque no en la totalidad los casos. Con respecto al sentimiento y al engagement de los debates, pudimos determinar que los eventos no tienen incidencia en ellos, lo que está en línea con lo hallado al estudiar su evolución temporal.

Como trabajo futuro, se podría agregar al análisis el nivel de controversia de los posteos y de los comentarios. Tanto para estudiar su evolución temporal como para determinar si los tiroteos tienen alguna incidencia en él. Así mismo, resultaría interesante analizar los árboles de respuestas y ver cómo impactan en ellos los niveles de toxicidad y el sentimiento para intentar comprender mejor cómo reaccionan los usuarios ante posteos tóxicos o negativos. Este análisis no fue realizado en la presente tesis dado que el dataset elegido no contaba con información que indicara el padre de los comentarios, haciendo que el armado del árbol no fuera posible.

En cuanto a las limitaciones, nuestros resultados dependen en gran medida de los modelos elegidos. Por un lado, podrían utilizarse otras variantes de los mismos, por ejemplo muchos trabajos utilizan Perspective API[2] para realizar sus análisis de toxicidad [5, 25, 89]. Por otro, es importante tener en cuenta que en todos los casos empleamos modelos ya entrenados a partir de datos distintos a los utilizados en esta tesis. En el futuro, también se podría intentar entrenar modelos propios. En nuestro caso, esta alternativa fue descartada debido a que resulta una tarea compleja que además implica perder una gran cantidad de data para destinar al entrenamiento y validación del modelo. Por otra parte, los datasets disponibles, que por lo general son escasos, también resultan una limitación. En nuestro caso, solo pudimos acceder a sets de datos correspondientes a Estados Unidos, con lo cual, nuestros resultados se limitan únicamente a ese territorio. Aunque es posible que los mismos sean similares en otros países y contextos, no podemos confirmarlo. Lo mismo ocurre con otras redes sociales.

[2] `https://www.perspectiveapi.com/`

4. CONCLUSIONES FINALES

A lo largo de la tesis, analizamos debates partidarios en las redes sociales Twitter y Reddit. En la primera a través de algoritmos de detección de comunidades, y en la segunda mediante técnicas de NLP y de inferencia causal. Dejando de lado las conclusiones particulares de ambos capítulos, que pueden encontrarse en las secciones 2.4 y 3.4 respectivamente, en ambos análisis sobre la evolución temporal hallamos tendencias que se mantenían en el tiempo sobre las dimensiones estudiadas. En el caso de Twitter fue acerca de la pertenencia de los usuarios en comunidades, y en el caso de Reddit, el sentimiento y el engagement del debate. Estos resultados indican cierta consistencia en la forma en la que los usuarios interactúan en las redes.

Como trabajo futuro sería interesante clasificar también por comunidades a los debates partidarios de Reddit. A partir de esto, se podría determinar si los usuarios tienden a mantenerse en su comunidad a lo largo del tiempo como vimos en el análisis de Twitter, aunque esta vez en un período considerablemente más largo. Esta clasificación además permitiría identificar usuarios que interactúan con ambos extremos. Puesto que no es lo mismo un usuario que interactúa positivamente en dos comunidades que uno que interactúa positivamente en una y negativamente en otra, se podrían utilizar las técnicas de NLP empleadas en este trabajo para intentar caracterizar el contenido que generan, con el objetivo de entender el comportamiento de aquellos usuarios que fomentan un intercambio de opiniones con bajos niveles de toxicidad. Así mismo, resultaría de interés observar si la estructura en comunidades cambia en los momentos de mayor popularidad de los debates, para ver si lo usuarios tienden a cerrarse aún más en ellas, o si estos promueven un intercambio más diverso. En línea con esto último, también podría analizarse la polarización de los debates: cómo evoluciona a través del tiempo, cómo responde ante incrementos en el interés en el debate y si eventos de la realidad tienen efectos en ella.

5. ANEXO

5.1. Transformers

En la figura 5.1 se encuentra el esquema de esta arquitectura.

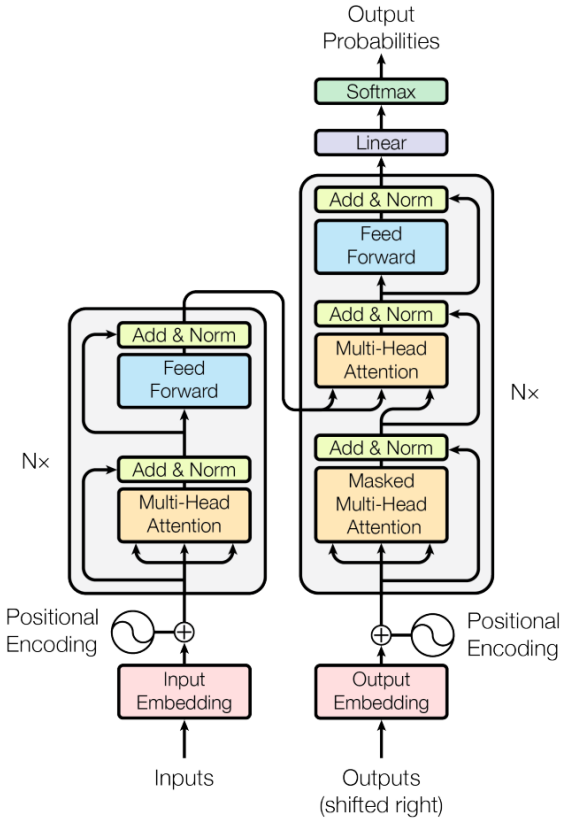


Fig. 5.1: Arquitectura del modelo Transformer. Tomado del paper "Attention is all you need" de Vaswani et al. [84].

Encoder

Lo primero que encontramos en el encoder es el input embedding, cuyo objetivo es convertir los inputs, compuestos por palabras, a números. La secuencia de entrada es convertida primero en una secuencia de tokens. A cada uno de ellos se les asigna un número que se corresponde con su identificador en el vocabulario del modelo. El problema con estos ids es que pueden escalar muy rápidamente, ya que los lenguajes poseen una enorme cantidad de palabras. Para solucionar esto, cada identificador es convertido en un vector de dimensión d_{model}. Esta conversión es parte del entrenamiento, ya que el modelo debe aprender a generar estos vectores de forma tal que representen el significado de la palabra que representa el token en cuestión.

Luego, encontramos el positional encoding, que se ocupa de agregar información sobre la posición del token de entrada dentro de la oración. Tienen la misma dimensión que

los embeddings (d_{model}) de forma tal que puedan ser sumados. En el paper, se utilizan funciones seno y coseno de distintas frecuencias que toman como parámetros de entrada la posición del token y la dimensión, haciendo que cada dimensión del positional encoding corresponda a una función sinusoidal.

El encoder está compuesto por $N = 6$ capas idénticas, donde cada una posee dos subcapas. La primera consiste en un mecanismo multi-headed self-attention, y la segunda es una red feed forward. Alrededor de ambas se emplea una conexión residual, seguida de una capa de normalización.

El mecanismo de atención [7] es una técnica que permite a las redes neuronales enfocarse en partes específicas de la secuencia de entrada. Resumidamente, se le asigna un peso a cada parte de la secuencia, siendo este mayor para aquellas secciones que resultan más relevantes. Permite que cada palabra "vea" y se relacione con todas las demás palabras en la secuencia, proporcionando una comprensión contextual más profunda. En particular, self-attention es un mecanismo de atención que relaciona diferentes posiciones de una única secuencia para calcular una representación de la misma.

En el paper se utiliza el mecanismo de atención denominado "Scaled Dot-Product Attention" (figura 5.2). Dada Q una matriz de queries, K una matriz de claves, ambas de dimensión d_k, y V una matriz de valores de dimensión d_v, la atención se calcula como

$$Attention(Q, V, K) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (5.1)$$

En el caso de self-attention, Q, K y V provienen de la misma fuente, es decir, los embeddings de entrada.

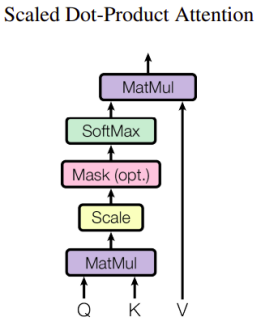


Fig. 5.2: Scaled Dot-Product Attention, Vaswani et al. [84].

Como mencionamos previamente, la arquitectura de Transformers utiliza multi-head attention. En vez de usar una única función de atención con claves, valores y queries de dimensión d_{model}, se realizan h proyecciones lineales de las mismas con proyecciones de dimensión d_k, d_k, d_v respectivamente, aprendidas por el modelo. En cada una de las versiones proyectadas de las queries, claves y valores se aplica la función de atención en paralelo, produciendo un output de dimensión d_v. Estos son concatenados, y una vez más proyectados, resultado en los valores finales. En la figura 5.3 podemos ver gráficamente este mecanismo.

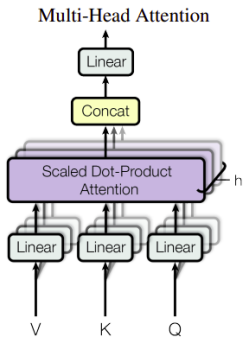


Fig. 5.3: Multi-Head Attention, Vaswani et al. [84].

La multi-head attention permite que el modelo reciba conjuntamente la información de distintos subespacios de representación en diferentes posiciones. Con una sola cabeza de atención, el promedio inhibe esto.

Además de la subcapa de self-attention, encontramos una subcapa compuesta por una red neuronal feed-forward que consiste en dos transformaciones lineales y que utiliza una función ReLU como función de activación entre ellas.

Ambas subcapas emplean conexiones residuales [36] que se saltean una o más capas, permitiendo que la información fluya directamente desde una capa anterior a una posterior, sin pasar por las capas intermedias. Así mismo, ambas están seguidas por una capa de normalización (layer normalization [6]), dado que normalizar la actividad de las neuronas es

una forma de reducir los tiempos del entrenamiento de la red.

Decoder

En el caso del decodificador, tenemos el output embedding y el positional encoding, que funcionan esencialmente de la misma manera que sus equivalentes en el encoder. Luego encontramos una capa llamada "masked multi-head attention", que recibe como query, clave y valor la misma oración de entrada del decoder, que ya pasó por las etapas del output embedding y del positional encoding. El objetivo en este punto es asegurar que el modelo no está utilizando palabras que se encuentren después que la palabra actual en la oración (es por esta razón también el output embedding es shifteado una posición). El procedimiento es muy similar al de la multi-head attention, la modificación se realiza antes de calcular la función softmax sobre la matriz $\left(\frac{QK^T}{\sqrt{d_k}}\right)$ (5.1), cuyos valores que estén por encima de la diagonal serán reemplazados por $-\infty$.

El decoder también recibe la información proveída por el encoder en forma de claves y valores. Estos son tomados por un capa de multi-head attention, que también recibe como query el resultado de la capa masked multi-head attention explicada arriba. Dado que su input no es una única oración, ya no tenemos self-attention, si no que el mecanismo de atención utilizado es cross-attention.

Al igual que el encoder, se emplean conexiones residuales alrededor de ambas subcapas de atención, seguidas de una capa de normalización.

Por último, se utiliza una transformación lineal que debe ser aprendida, y una función softmax para convertir la salida del decodificador en las probabilidades predichas del próximo token.

5.2. LDA

5.2.1. LDA: modelo generativo

Explicaremos su versión más simple.

Primero definimos:

- palabra: es la unidad básica de data discreta. Se define como un ítem del vocabulario $\{1, ..., V\}$. Las palabras se van a representar como vectores con una única componente en 1 y el resto en 0.
- documento: secuencia de N palabras denotado por $\mathbf{w} = (w_1, ..., w_N)$.
- corpus: colección de M documentos denotado por $\mathbf{D} = (\mathbf{w}_1, ..., \mathbf{w}_M)$.

Se asumen los siguientes procesos generativos para cada documento $\mathbf{w}$ en un corpus D.

1. Elegir $N \sim Poisson(\xi)$.
2. Elegir $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$.
3. Por cada una de las N palabras w_n:
   a. Elegir un tópico $z_n \sim Multinomial(\boldsymbol{\theta})$
   b. Elegir una palabra w_n de $p(w_n|z_n, \boldsymbol{\beta})$, una probabilidad multinomial condicionada en el tópico z_n.

Dentro de las simplificaciones de esta versión del modelo, mencionamos que la dimensión k de la distribución Dirichlet se asume conocida y fija. Además, las probabilidades de las palabras son parametrizadas mediante una matriz $\boldsymbol{\beta}$ de $k \times V$, donde $\boldsymbol{\beta}_{ij} = p(w^j = 1|z^i = 1)$. Esta cantidad a estimar también se asumirá fija.

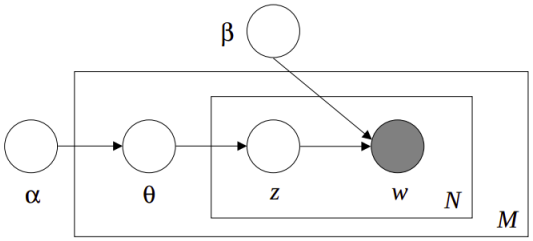


Fig. 5.4: Modelo gráfico de LDA.

La figura 5.4 muestra el modelo gráfico de LDA:

- El parámetro $\boldsymbol{\alpha}$ representa la prior de la distribución de Dirichlet sobre los tópicos por documento. Un valor alto de $\boldsymbol{\alpha}$ indica que cada documento contiene una mezlca de todos los temas. Por el contrario, un valor bajo indica que cada documento toca pocos tópicos.
- El parámetro $\boldsymbol{\beta}$ representa la prior de la distribución de Dirichlet sobre las palabras por tópico. Un valor alto indica cada tópico utilizará muchas palabras, un valor bajo, que utilizará pocas.
- Los tópicos se distribuyen con una multinomial de prior $\boldsymbol{\theta}$.
- $\mathbf{z}$ es el vector de N tópicos.
- $\mathbf{w}$ un conjunto de N palabras.

Tanto $\boldsymbol{\alpha}$ como $\boldsymbol{\beta}$ son parámetros a nivel de corpus, las variables $\boldsymbol{\theta}_d$ son a nivel de documento, y las variables z_{dn} y w_{dn} son a nivel de palabra.

Dados los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$, la distribución conjunta de una mixtura de tópicos $\boldsymbol{\theta}$, un conjunto de N tópicos $\mathbf{z}$ y un conjunto de N palabras $\mathbf{w}$ está dada por:

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}),$$

donde $p(z_n|\boldsymbol{\theta})$ es $\boldsymbol{\theta}_i$ con i es el único tal que $z_n^i = 1$. Integrando sobre $\boldsymbol{\theta}$ y sumando sobre z obtenemos la distribución marginal de un documento:

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}$$

Tomando el producto de la probabilidad marginal de cada documento, obtenemos la probabilidad del corpus:

$$p(\mathbf{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_d$$

Como mencionamos en la sección 3.2.1.2, no buscamos generar documentos, sino determinar, dado un documento, qué tópicos le corresponden junto con sus probabilidades. Esto lo podemos obtener revertiendo el proceso generativo que explicamos anteriormente y aprendiendo la distribución a posteriori de las variables lantentes del modelo, dada la data observada. En LDA esto implica resolver la siguiente ecuación:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}$$

A continuación veremos cómo calcular esta distribución mediante Inferencia Variacional Bayesiana.

5.2.2. LDA: Inferencia Variacional Bayesiana

La Inferencia Variacional Bayesiana aproxima la verdadera distribución a posteriori $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ mediante una distribución más sencilla llamada "variacional". Esta tiene sus propios parámetros variacionales y se elige de forma tal que sea la más "cercana" a p. Para esto, se utiliza la divergencia de Kullback-Leibler (KL) [45] que permite medir la similitud entre dos distribuciones de probabilidad.

Esta distribución variacional debe ser tomada de una familia de distribuciones. Antes de definirla, notemos que existe una dependencia entre $\boldsymbol{\theta}$ y $\boldsymbol{\beta}$, debido a las aristas entre $\boldsymbol{\theta}$, $\mathbf{z}$ y $\boldsymbol{\beta}$ en la figura 5.5. En el paper [11] se trabaja sobre un modelo simplificado, que es producto de eliminar estas aristas y los nodos $\mathbf{w}$, y agregar parámetros variacionales (figura 5.6).

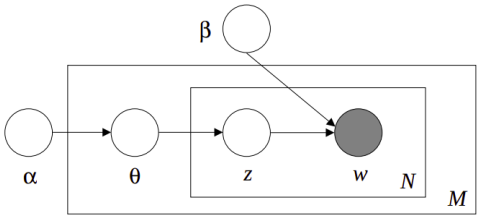


Fig. 5.5: Modelo gráfico de LDA.

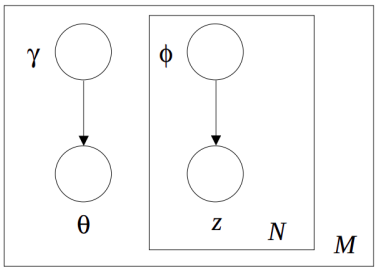


Fig. 5.6: Modelo gráfico simplificado de LDA.

Fig. 5.7: Modelos gráficos de LDA.

A partir de este modelo se obtiene una familia de distribuciones sobre las variables latentes. Esta familia está caracterizada por

$$q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\gamma) \prod_{n=1}^{N} q(z_n|\boldsymbol{\phi}_n)$$

donde el parámetro de Dirichlet γ y el parámetro multinomial $(\phi_1, ..., \phi_N)$ son los parámetros variacionales libres. Notar que $\boldsymbol{\beta}$ no tiene una distribución variacional, y esto se debe a que en el paper se la trata como un parámetro del modelo.

Luego, queremos encontrar los valores de $\boldsymbol{\phi}$ y γ que minimicen la distancia entre $q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi})$ y $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$(\gamma^*, \boldsymbol{\phi}^*) = \underset{(\gamma,\boldsymbol{\phi})}{\arg\min}\, KL(q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}))$$

Como se demuestra en el apéndice A.3 del paper de Blei et al. [11], minimizar la divergencia de KL entre la distribución variacional y la distribución a posteriori es equivalente a maximizar la cota inferior L definida como

$$L(\gamma, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{E}[\log\ p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})] - \mathbb{E}[\log\ q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi})]$$

La función L puede ser maximizada respecto de γ y $\boldsymbol{\phi}$ mediante un algoritmo iterativo de punto fijo. En particular, al derivar la divergencia de KL e igual a 0 con respecto a cada parámetro se obtienen las siguientes ecuaciones de actualización:

$$\boldsymbol{\phi}_{ni} \propto \boldsymbol{\beta}_{iw_n} exp\left\{\boldsymbol{\Psi}(\gamma_i) - \boldsymbol{\Psi}\left(\sum_{j=1}^{k} \gamma_j\right)\right\}$$

$$\gamma_i = \boldsymbol{\alpha_i} + \sum_{n=1}^{N} \boldsymbol{\phi}_{ni},$$

donde $\boldsymbol{\Psi}$ es la primera derivada de la función $\log\Gamma$ que es computable mediante la aproximación de Taylor [1].

A partir de estas ecuaciones se define el siguiente pseudocódigo:

Algorithm 1 LDA con IVB

1: inicializar $\boldsymbol{\phi}^0_{ni} := 1/k$ para todo i y n.
2: inicializar $\gamma_i := \boldsymbol{\alpha}_i + N/k$ para todo i.
3: **while** ! convergencia **do**
4: **for** $n = 1, ..., N$ **do**
5: **for** $i = 1, ..., K$ **do**
6: $\boldsymbol{\phi}^{t+1}_{ni} := \boldsymbol{\beta}_{iw_n} \exp(\boldsymbol{\Psi}(\gamma^t_i))$
7: **end for**
8: Normalizar $\boldsymbol{\phi}^{t+1}_{ni}$ para que sume 1.
9: **end for**
10: $\gamma^{t+1} := \boldsymbol{\alpha} + \sum_{n=1}^{N} \boldsymbol{\theta}^{\boldsymbol{t+1}}_{\boldsymbol{n}}$
11: **end while**

5.3. Evolución del sentimiento de los debates

En la sección 3.3.1, construimos series temporales con la proporción diaria de posteos de cada clasificación. A partir de ellas, generamos gráficos para observar su evolución temporal. El gráfico correspondiente al debate del aborto puede encontrarse en la sección previamente mencionada, en la figura 3.4. En la figura 5.10 encontramos su equivalente para el debate del control de armas.

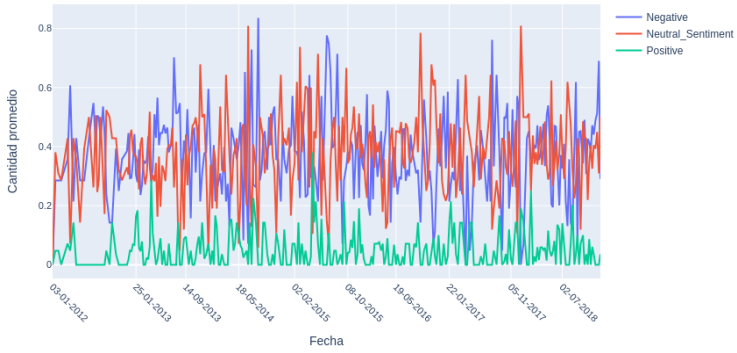


Fig. 5.8: Posteos.

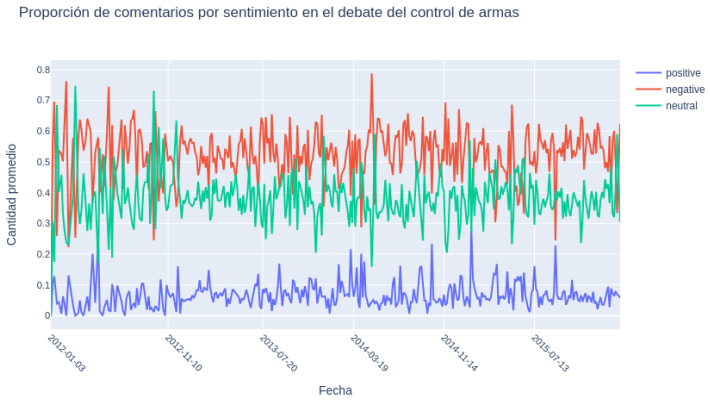


Fig. 5.9: Comentarios.

Fig. 5.10: Evolución temporal del sentimiento de los posteos y comentarios del debate del control de armas.

5.4. Medidas de engagement de los posteos negativos, positivos y neutrales

En la sección 3.3.5 construimos boxplots para la cantidad de comentarios y el score de los posteos positivos, negativos y neutrales de ambos debates. En la figura 5.13 encontramos los correspondientes al debate del control de armas.

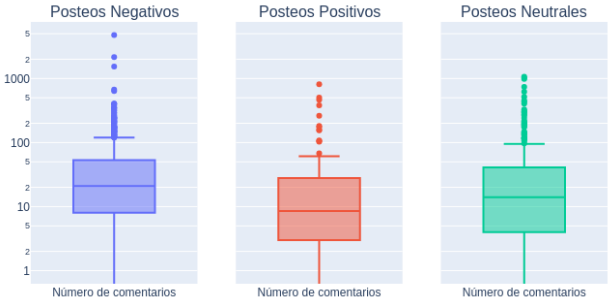


Fig. 5.11: Boxplot del número de comentarios por posteo.

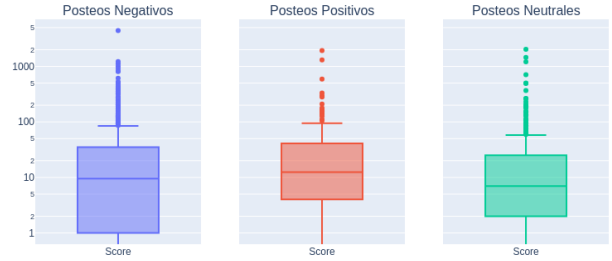


Fig. 5.12: Boxplot del score por posteo.

Fig. 5.13: Boxplots sobre el score y el número de comentarios del debate del control de armas, por cada sentimiento.

Bibliografía

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.

[2] L. Aihua, P. P. Miglietta, and P. Toma. Did carbon emission trading system reduce emissions in china? an integrated approach to support policy modeling and implementation. *Energy Systems*, pages 1–23, 2021.

[3] F. Albanese, L. Lombardi, E. Feuerstein, and P. Balenzuela. Predicting shifting individuals using text mining and graph machine learning on twitter. *arXiv preprint arXiv:2008.10749*, 2020.

[4] J. Aldrich. Correlations genuine and spurious in pearson and yule. *Statistical science*, pages 364–376, 1995.

[5] H. Almerekhi, S. b. B. J. Jansen, and c.-s. b. H. Kwak. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020*, pages 294–298, 2020.

[6] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[8] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[9] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[10] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[13] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[14] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 2015.

[15] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–22, 2017.

[16] D. Choi, S. Chun, H. Oh, J. Han, and T. T. Kwon. Rumor propagation is amplified by echo chambers in social media. *Scientific reports*, 10(1):310, 2020.

[17] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

[18] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

[19] G. Crupi, Y. Mejova, M. Tizzani, D. Paolotti, and A. Panisson. Echoes through time: evolution of the italian covid-19 vaccination debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 102–113, 2022.

[20] L. Dang-Xuan, S. Stieglitz, J. Wladarsch, and C. Neuberger. An investigation of influentials and the role of sentiment in political communication on twitter during election periods. In *Social Media and Election Campaigns*, pages 168–198. Routledge, 2017.

[21] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):37825, 2016.

[22] F. Demarco, J. M. O. de Zarate, and E. Feuerstein. Measuring ideological spectrum through nlp. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023)*, 2023.

[23] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[25] A. Efstratiou, J. Blackburn, T. Caulfield, G. Stringhini, S. Zannettou, and E. De Cristofaro. Non-polar opposites: analyzing the relationship between echo chambers and hostile intergroup interactions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 197–208, 2023.

[26] S. Emmons, S. Kobourov, M. Gallant, and K. Börner. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7):e0159161, 2016.

[27] J. Flamino, A. Galeazzi, S. Feldman, M. W. Macy, B. Cross, Z. Zhou, M. Serafino, A. Bovet, H. A. Makse, and B. K. Szymanski. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nature Human Behaviour*, pages 1–13, 2023.

[28] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41, 2007.

[29] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 43–52, 2017.

[30] V. R. K. Garimella and I. Weber. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, pages 528–531, 2017.

[31] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[32] A. Gruzd and J. Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & internet*, 6(1):28–45, 2014.

[33] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[34] L. Hanu. Unitary team. detoxify. github, 2020.

[35] T. Häussler. Heating up the debate? measuring fragmentation and polarisation in a german climate change hyperlink network. *Social Networks*, 54:303–313, 2018.

[36] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

[38] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[39] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[40] M. Igal Browarnik, J. M. Ortíz de Zárate, and E. Feuerstein. Identificación de comunidades en intervalos de tiempo a traves del lenguaje. In *VI Simposio Argentino de Ciencia de Datos y GRANdes DAtos (AGRANDA 2020)-JAIIO 49 (Modalidad virtual)*, 2020.

[41] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[42] K. H. Jamieson and J. N. Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[43] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[44] D. Kodati and R. Tene. Identifying suicidal emotions on social media through transformer-based deep learning. *Applied Intelligence*, 53(10):11885–11917, 2023.

[45] S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[46] E. Kušen and M. Strembeck. Politics, sentiments, and misinformation: An analysis of the twitter discussion on the 2016 austrian presidential elections. *Online Social Networks and Media*, 5:37–50, 2018.

[47] H. Kwak, J. Blackburn, and S. Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3739–3748, 2015.

[48] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[49] S. Li, Z. Xie, D. K. Chiu, and K. K. Ho. Sentiment analysis and topic modeling regarding online classes on the reddit platform: educators versus learners. *Applied Sciences*, 13(4):2250, 2023.

[50] L. Lima, J. C. Reis, P. Melo, F. Murai, L. Araujo, P. Vikatos, and F. Benevenuto. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, pages 515–522. IEEE, 2018.

[51] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 305–314, 2014.

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[53] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*, 2022.

[54] H. Lu, J. Caverlee, and W. Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222, 2015.

[55] V. Martínez, F. Berzal, and J.-C. Cubero. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, 49(4):1–33, 2016.

[56] N. Masuda, M. A. Porter, and R. Lambiotte. Random walks and diffusion on networks. *Physics reports*, 716:1–58, 2017.

[57] L. McInnes, J. Healy, S. Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[58] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[59] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad. Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512, 2021.

[60] P. Metaxas, E. Mustafaraj, K. Wong, L. Zeng, M. O'Keefe, and S. Finn. What do retweets indicate? results from user survey and meta-review of research. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 658–661, 2015.

[61] J. Mills, E. Aronson, and H. Robinson. Selectivity in exposure to information. *The Journal of Abnormal and Social Psychology*, 59(2):250, 1959.

[62] M. Z. Naf'an, A. A. Bimantara, A. Larasati, E. M. Risondang, and N. A. S. Nugraha. Sentiment analysis of cyberbullying on instagram user comments. *Journal of Data Science and Its Applications*, 2(1):38–48, 2019.

[63] J. A. Neto, K. M. Yokoyama, and K. Becker. Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the international conference on web intelligence*, pages 26–33, 2017.

[64] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[65] A. Olteanu, C. Castillo, J. Boy, and K. Varshney. The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[66] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE, 2016.

[67] K. Pearson. The problem of the random walk. *Nature*, 72(1865):294–294, 1905.

[68] T. P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[69] T. P. Peixoto. *Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths*. Cambridge University Press, 2023.

[70] W. Quattrociocchi, A. Scala, and C. R. Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

[71] P. Rajpurkar. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[72] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[73] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[74] M. H. Ribeiro, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, S. Long, S. Greenberg, and S. Zannettou. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207, 2021.

[75] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[76] K. J. Rothman and S. Greenland. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150, 2005.

[77] K. Sailunaz and R. Alhajj. Emotion and sentiment analysis from twitter text. *Journal of computational science*, 36:101003, 2019.

[78] M. Saveski, B. Roy, and D. Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, pages 1086–1097, 2021.

[79] V. Shankaran and R. Sharma. Analyzing toxicity in deep conversations: A reddit case study. *arXiv preprint arXiv:2404.07879*, 2024.

[80] L. G. Stewart, A. Arif, and K. Starbird. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, volume 70, 2018.

[81] M. M. Tadesse, H. Lin, B. Xu, and L. Yang. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893, 2019.

[82] M. Tromholt. The facebook experiment: Quitting facebook leads to higher levels of well-being. *Cyberpsychology, behavior, and social networking*, 19(11):661–666, 2016.

[83] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[85] G. Veletsianos, R. Kimmons, R. Larsen, T. A. Dousay, and P. R. Lowenthal. Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on youtube ted talk comments. *PloS one*, 13(6):e0197331, 2018.

[86] I. Waller and A. Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268, 2021.

[87] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[88] K. Welbers and M. Opgenhaffen. Social media gatekeeping: An analysis of the gatekeeping influence of newspapers' public facebook pages. *New Media & Society*, 20(12):4728–4747, 2018.

[89] Y. Xia, H. Zhu, T. Lu, P. Zhang, and N. Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, 2020.

[90] Y. Xu and S. B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, 2018.

[91] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

[92] W. Zhang, H. Xu, and W. Wan. Weakness finder: Find product weakness from chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, 39(11):10283–10291, 2012.