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

Modelos para el análisis de la polarización a través de PLN

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires
en el área de Ciencias de la Computación

Juan Manuel Ortiz de Zarate

Director de tesis:	Dr. Esteban Zindel Feuerstein
Consejero de Estudios:	Dr. Diego Fernández Slezak
Lugar de Trabajo:	Departamento de Computación Facultad de Cs. Exactas y Naturales Universidad de Buenos Aires

Buenos Aires, 2023

MODELOS PARA EL ANÁLISIS DE LA POLARIZACIÓN A TRAVÉS DE PLN

La polarización, entendida como la intensificación de contrastes ideológicos y sociales, ha emergido con fuerza como un tema central de preocupación en el ámbito político y académico desde el comienzo del siglo XXI. Varios países han evidenciado este fenómeno, especialmente con el ascenso de regímenes post neoliberales. Estas divisiones, en algunos casos, han despertado inquietudes sobre una posible erosión de la democracia, ya que se teme que algunas facciones puedan optar por estructuras no democráticas en lugar de ceder el poder a un grupo rival.

Por otro lado, diversos trabajos de la literatura señalan una posible influencia de las redes sociales en la polarización. Estas plataformas, mediante su diseño y algoritmos de recomendación, podrían crear "cámaras de eco", donde las opiniones se refuerzan mutuamente, lo que podría potenciar la división y el aislamiento ideológico. Resultando en un potencial aumento de conflictos y malentendidos entre diferentes grupos sociales y políticos. A su vez, otros estudios indican la dificultad de poder hacer este análisis causal al mismo tiempo que sugieren que los motivos de la polarización podrían ser multicausales o incluso ajenos a las redes sociales. Por ello, es necesario contar con nuevas herramientas para analizar mejor este fenómeno complejo y aprovechar la masividad de datos digitales que estas plataformas nos brindan.

Los avances recientes en el procesamiento del lenguaje natural (PLN), combinados con la digitalización de las discusiones a través de redes sociales, ofrecen un panorama prometedor para comprender y abordar la polarización desde una perspectiva computacional. Estas herramientas y técnicas avanzadas permiten un análisis más profundo de las conversaciones y discusiones, además nos brindan la oportunidad de diseñar intervenciones más informadas y efectivas, con el objetivo final de promover un mayor entendimiento y diálogo en nuestra sociedad contemporánea.

En esta tesis, presento técnicas, modelos y herramientas innovadoras para analizar textualmente la polarización. Introduzco dos técnicas II únicas para cuantificar la polarización en discusiones basadas en el contenido de los posteos y las interacciones de los usuarios. Estas técnicas demuestran una eficiencia y precisión superiores a los métodos

previos [83]. También, las aplico en un estudio comparativo entre EEUU y Argentina 5.1, para evaluar su efectividad en diferentes contextos culturales y políticos.

Sin embargo, también existen usuarios que, aunque se encuentran en contextos polarizados, no se alinean estrictamente con ninguno de los principales grupos y actúan como "puentes"de entendimiento. Aplicando variantes de las técnicas de cuantificación sobre focus groups 7, identifico que estos usuarios alternan posturas según el tema y, basándome en un análisis cualitativo realizado junto a un grupo de sociólogos, determinamos que buscan una perspectiva más reflexiva y pluralista sobre los temas.

Uno de los sectores más relevante en los contextos de polarización está dado por la clase política. Es por eso que también analizo 7.7 el comportamiento de los políticos en las principales redes sociales, destacando las diferencias de interacción y discusión entre Oficialismo y Oposición en Argentina, cómo utilizan distintas estrategias de comunicación dependiendo de la red social y evidenciando la relación entre toxicidad y repercusión en plataformas como Twitter.

Si bien la polarización se basa en dos grandes grupos antagónicos, dentro de cada polo, existen matices ideológicos. Para analizar con mas precisión esto, desarrollo una técnica 8.4 para asignar puntajes a textos basándose en su inclinación ideológica. Encuentro una alta correlación entre nuestros resultados y los obtenidos por métodos previos [204] donde se utilizaron las interacciones. Además comparo la eficacia de dos tipos de embeddings, utilizando Fasttext y un modelos de lenguaje de gran tamaño (LLM), donde el último demostró ser superior. Nuestro método, al centrarse en el texto, ofrece una flexibilidad significativamente mayor al de las interacciones, permitiendo evaluar cualquier conjunto textual en varios espectros ideológicos, como posturas sobre el aborto o economía. Esta técnica aporta una nueva dimensión al análisis de polarización.

Los LLMs están ganando terreno en aplicaciones, a menudo proporcionando opiniones subjetivas en sus respuestas, como se observa en ejemplos de DeepMind y Anthropic. En el último capítulo de mi tesis IV, utilizando la encuesta Latinobarómetro 2020 sobre la población argentina, examino tres LLMs (GPT, Cohere, Bard) y su alineación con las respuestas de la encuesta. Descubro que los LLMs reflejan opiniones del sector más masculinizado y politizado de Argentina, con Bard y GPT inclinándose hacia la población educada y adulta, y GPT mostrando afinidad con posturas de derecha. Estos hallazgos enfatizan la necesidad de ser conscientes de las inclinaciones de estos modelos al considerar

sus opiniones en temas controvertidos o subjetivos.

En esta tesis, presento técnicas computacionales innovadoras para analizar la polarización, colaborando estrechamente con expertos en ciencias sociales para validar nuestras herramientas. Concluyo que la polarización está fuertemente ligada al lenguaje, permitiéndonos cuantificarla y tomar medidas para abordarla. Mientras que algunos actores mantienen posturas ambivalentes, otros, como los políticos, adaptan sus estrategias según el público. También destacamos que las inteligencias artificiales, como los chatbots, presentan inclinaciones en debates contemporáneos, lo que subraya la necesidad de estar informados sobre sus posiciones.

Palabras claves: Polarización, Procesamiento del lenguaje natural, Redes Sociales, PLN Social.

MODELS FOR THE ANALYSIS OF POLARIZATION THROUGH NLP

Polarization, understood as the intensification of ideological and social contrasts, has forcefully emerged as a central concern in the political and academic spheres since the beginning of the 21st century. Various countries have witnessed this phenomenon, especially with the rise of post-neoliberal regimes. In some cases, these divisions have raised concerns about a potential erosion of democracy, as it is feared that some factions may opt for non-democratic structures rather than cede power to a rival group.

On the other hand, many works in the literature point out a possible influence of social networks on polarization. These platforms, through their design and recommendation algorithms, could create "echo chambers", where opinions reinforce each other, which could enhance division and ideological isolation. Resulting in a potential increase in conflicts and misunderstandings between different social and political groups. At the same time, other studies indicate the difficulty of being able to do this causal analysis while suggesting that the reasons for polarization could be multi-causal or even unrelated to social networks. Therefore, it is necessary to have new tools to better analyze this complex phenomenon and take advantage of the massive digital data that these platforms provide us.

Recent advances in natural language processing (NLP), combined with the digitization of discussions through social networks, offer a promising outlook for understanding and addressing polarization from a computational perspective. These advanced tools and techniques allow for a deeper analysis of conversations and discussions, and also give us the opportunity to design more informed and effective interventions, with the ultimate goal of promoting greater understanding and dialogue in our contemporary society.

In this thesis, I present innovative techniques, models, and tools for textually analyzing polarization. I introduce two unique techniques II for quantifying polarization in discussions based on the content of posts and user interactions, demonstrating efficiency and precision superior to previous methods [83] and applying them in a comparative study between the USA and Argentina 5.1. I also analyze the influence of polarization on the specific jargon of groups, developing a clustering method to identify and classify these groups and proposing strategies to adapt the model to the evolution of language.

However, there are also users who, although they exist in polarized contexts, do not strictly align with any of the main groups and act as 'bridges' of understanding. Applying variants of the quantification techniques to focus groups, I identify 7 that these users alternate stances depending on the topic and, based on a qualitative analysis with sociologists, we determine that they seek a more reflective and pluralistic perspective on issues.

One of the most relevant sectors in contexts of polarization is politicians. That's why I also analyze the behavior of politicians on major social networks 7.7, highlighting the differences in interaction and discussion between the Ruling Party and the Opposition in Argentina, how they use different communication strategies depending on the social network, and demonstrating the relationship between toxicity and impact on platforms like Twitter.

Although polarization is based on two major antagonistic groups, within each pole, there are ideological nuances. To analyze this more precisely, I developed a technique for assigning scores to texts based on their ideological inclination 8.4. I found a high correlation between our results and those obtained by previous methods [204] where interactions were used. In addition, I compare the efficacy of two types of embeddings, using Fasttext and an LLM, with the latter proving to be superior. Our method, by focusing on the text, offers significantly greater flexibility than interactions, allowing the evaluation of any textual set on various ideological spectrums, such as stances on abortion or the economy. This technique adds a new dimension to the analysis of polarization.

Large language models (LLMs) are gaining ground in applications, often providing subjective opinions in their responses, as observed in examples from DeepMind and Anthropic. In the last chapter of my thesis IV, using the Latinobarómetro 2020 survey on the Argentine population, I examined three LLMs (GPT, Cohere, Bard) and their alignment with the survey responses. I found that the LLMs reflect opinions from the more masculinized and politicized sector of Argentina, with Bard and GPT leaning towards the educated and adult population, and GPT showing affinity with right-wing stances. These findings emphasize the need to be aware of these models' inclinations when considering their opinions on controversial or subjective topics.

In this thesis, I have presented innovative computational techniques for analyzing polarization, working closely with social science experts to validate our tools. I concluded that polarization is strongly linked to language, allowing us to quantify it and take measures to address it. While some actors maintain ambivalent stances, others, such as politicians,

adapt their strategies according to the audience. We also highlight that artificial intelligence, such as chatbots, present inclinations in contemporary debates, underscoring the need to be informed about their positions.

Keywords: Polarization, Natural Language Processing, Social Networks, Social NLP.

AGRADECIMIENTOS

En primer lugar quiero agradecer a mi director Esteban Feuerstein por guiarme y apoyarme en todo mi trayecto. Durante estos años atravesé momentos muy difíciles que me hubieran sido imposible sobrellevar sin su compañía y eso es algo que atesoraré para siempre. Muchas veces escuché que las relaciones con los directores son complicadas e incluso casos en los que abandonaron sus carreras debido a eso. Tuve la suerte de que mi experiencia sea totalmente opuesta. Si pude disfrutar cada etapa y momento de este doctorado fue gracias a su dirección, tuve libertad para elegir las líneas de investigación que más me motivaban a la vez que me enseñaba como realizar mi trabajo rigurosa y profesionalmente. Termino este trayecto académico siendo un mucho mejor profesional y eso es algo que me llena de felicidad.

Quiero agradecer también a todos los colegas con los que colaboré a lo largo de estos años: Gabriel Kessler, Gabriel Vommaro, Brenda Focas, Marco Di Giovanni, Juan Manuel Dias, Alejandro Avenburg, Martín Browarnik y Franco Demarco. De ellos aprendí muchísimo y también esta tesis se debe a las investigaciones que pudimos hacer en conjunto. Quiero agradecer especialmente a Federico Albense, con quien comenzamos casi a la par nuestros doctorados y nos acompañamos a lo largo de todo el trayecto. Gracias a él este doctorado no fue tan solitario, además, me ayudó a mejorar muchísimo la comunicación de mis trabajos, lo que me abrió muchas puertas que hoy me permiten estar atravesando un gran momento. Gracias Fede.

A Fundar, donde, una vez terminada mi beca, pude encontrar un hermoso lugar en el cual continuar desarrollando mis líneas de investigación acompañado de excelentes profesionales. Especialmente Dani Yankelevich, quien siempre me apoyó y acompañó en todas mis iniciativas y de quién aprendí a cómo mantenerme al tanto en lo último del estado del arte.

A mis amigos que soportaron mis catarsis y monotemas todo este tiempo: Moncho, Maxi, Alesi, Ini, Tincher y Leo. A Alexis Soifer, con quien atravesamos paralelamente las dificultades académicas los últimos 15 años, su ayuda, consejos y experiencias fueron fundamentales para equivocarme un poco menos.

A Betina y Octavio que fueron quienes me enseñaron el amor por la ciencia y la educación. Disfruté mucho todos estos años académicos porque aprender cosas nuevas siempre me resultó muy satisfactorio, y sin eso, no habría llegado hasta acá. También Pili fue indispensable, sin el amor de todos ellos no hubiese tenido la valentía y seguridad para no dejar nunca de insistir en esta aventura.

A toda mi familia y especialmente a Virginia, Eva, Javi, Emma, Gala, Facundo, Roberto, Marco, Guido y Bruno. Haber tenido redes de contención donde pude recibir amor incondicionalmente me salvó muchas veces de no bajar los brazos. También a quienes ya no están pero siempre me acompañan Estela y Abuelito.

Por último a mi compañera de vida, Yami. No sólo soportó mis crisis, miedos, catarsis y frustaciones, siempre empujándome y dándome confianza para seguir intentándolo. También me enseñó a ser mejor persona, superar muchos miedos y aprender a compartir toda una vida con una persona maravillosa al lado. Te amo, gracias.

TBD

Índice general

Parte I

INTRODUCCIÓN

1. ¿POR QUÉ ES IMPORTANTE TENER HERRAMIENTAS PARA EL ANÁLISIS DE LA POLARIZACIÓN?

La polarización social, definida como la intensificación de opiniones opuestas y la disminución de puntos de vista intermedios en la sociedad, es en forma creciente objeto de preocupación política y académica [191]. En particular, desde el ascenso al poder de los regímenes post neoliberales o nacional-populares a comienzos del siglo XXI, la polarización se ha extendido en distintos países y coyunturas electorales en al menos Argentina, Bolivia, Brasil, Colombia, Costa Rica, Ecuador, Estados Unidos, España, Reino Unido y Nicaragua [11, 20, 30, 74, 82, 96, 143].

Diversas investigaciones de opinión pública destacan la preocupación por la relación entre polarización y una creciente erosión a la democracia en toda la región [116, 130, 151], por ejemplo porque se podría optar por preferir un régimen no democrático antes que la llegada al poder del color político rival. Además, Levitsky et al. [130] argumentan que la erosión de las normas democráticas y la polarización partidista pueden ser precursores de la decadencia democrática. A través de un análisis de casos históricos y contemporáneos, sostienen que las democracias no mueren necesariamente a manos de golpes militares o revoluciones violentas, sino que a menudo se desintegran gradualmente desde dentro, especialmente cuando las élites políticas no respetan las normas no escritas de tolerancia mutua y autocontención. En este contexto, la polarización extrema puede llevar a que los actores políticos vean a sus oponentes no como rivales legítimos, sino como amenazas existenciales, justificando así acciones antidemocráticas. Por otro lado, [151] sostiene que la democracia liberal está en crisis debido a la combinación de factores como el estancamiento económico, el cambio cultural y la creciente polarización. Argumenta que esta confluencia de factores está llevando a un descontento generalizado con la democracia liberal y al surgimiento de populismos autoritarios. La polarización, en particular, alimenta la desconfianza en las instituciones democráticas y crea un terreno fértil para líderes que prometen soluciones simples a problemas complejos, a menudo a expensas de las libertades democráticas.

La irrupción de las redes sociales digitales[68] dió lugar a nuevas formas de intervenir intencionadamente sobre la polarización para sacar algún provecho [47, 189]. Adicionalmente, un creciente cuerpo de literatura sugiere que estas plataformas pueden fomentar la división de la sociedad al crear "cámaras de eco" o "burbujas de filtro" en las que los usuarios se exponen principalmente a opiniones que refuerzan sus propias creencias preexistentes [166]. Este fenómeno puede intensificar los conflictos y aumentar la incomprensión entre diferentes grupos sociales y políticos.

Además, las características intrínsecas de las redes sociales, como los algoritmos de recomendación, pueden exacerbar aún más la polarización. Estos sistemas suelen priorizar el contenido que es probable que genere interacción, lo que puede incentivar el uso de tácticas de división y enfrentamiento [198]. Los efectos de las redes sociales en la polarización de la sociedad no sólo son un tema de interés académico, sino que también tienen implicaciones prácticas para la democracia y la cohesión social. Los puntos de vista muy contrastantes en algunos grupos tienden a provocar conflictos que conducen a ataques de una comunidad a otra acosándola, insultándola o "trolleándola" [123]. La literatura existente muestra diferentes problemas que plantea la controversia en redes sociales, como la división de las comunidades, la información sesgada, los discursos de odio y los ataques entre grupos.

Por lo tanto la influencia de estas plataformas en la división de la sociedad es un área de interés creciente para la investigación científica. Las redes sociales se han convertido en una fuerza ineludible en la política y la formación de la opinión pública, permitiendo a las personas expresar y consumir información de maneras anteriormente inimaginables. Al entender estos procesos, los investigadores, los responsables de políticas y los ingenieros pueden diseñar intervenciones más eficaces para mitigar los aspectos negativos de las redes sociales y promover el diálogo y el entendimiento.

Por otro lado, en los últimos años, el campo del procesamiento del lenguaje natural (PLN) ha logrado grandes avances en el modelado del lenguaje a través de técnicas como Word2Vec, Attention y Transformers. Esto, junto con la digitalización de las discusiones producto del surgimiento de las redes sociales, nos pone ante un escenario ideal para el análisis de estos fenómenos desde una perspectiva computacional. De este modo, podemos aportar nuevos métodos, hallazgos y visiones para enfrentar esta problemática de forma más holística.

1.1. Polarización y redes sociales

Después de un período relativamente breve de euforia sobre la posibilidad de que las redes sociales inauguraran una era dorada de democratización global, ahora existe una preocupación generalizada en muchos sectores de la sociedad —incluidos los medios de comunicación, académicos, la comunidad filantrópica, la sociedad civil e incluso los propios políticos— de que las redes sociales puedan estar socavando la democracia[196]. Este temor no se limita sólo a las democracias nuevas o inestables, que a menudo son susceptibles al retroceso democrático, sino también a algunas de las democracias más consolidadas del mundo, incluyendo a los Estados Unidos. Efectivamente, en un lapso de poco más de cinco años, hemos sido testigos de un notable cambio en la percepción académica sobre la tecnología y la democracia. En 2010, el Journal of Democracy publicó un influyente artículo titulado “Tecnología de Liberación” [63]. Sin embargo, el mismo periódico, reflejando una creciente preocupación, incluyó en un foro sobre las elecciones estadounidenses de 2016 un artículo titulado “¿Puede la democracia sobrevivir a Internet?”[169]. Durante esos 5 años (y la tendencia continúa hasta el día de hoy), no sólo se han multiplicado las opiniones y resultados de los trabajos sino que el interés por este tema ha crecido notablemente [121].

Hoy en día existe un gran debate entorno a la influencia de las redes sociales en la polarización, habiendo distintas posiciones sobre su implicancia. Tucker et al. [197] hace una gran revisión de la literatura relacionada a este tópico donde señala varias limitaciones inherentes a este análisis que limitan y dificultan su estudio. Entre ellas, cabe destacar las siguientes:

- **Definiciones ambiguas y falta de consenso:** Una de las primeras limitaciones es la falta de consenso sobre qué constituye “política”, “desacuerdo” y “conversación”. Esta variedad de definiciones lleva a inconsistencias en los hallazgos y dificulta la comparación entre estudios [72]. Los investigadores no siempre coinciden en estos términos fundamentales, lo que afecta la medición y el análisis de las interacciones políticas.

- **Problemas metodológicos:** Los estudios sobre discusiones políticas enfrentan varios desafíos metodológicos. Estos incluyen cómo medir la frecuencia y naturaleza de las conversaciones políticas, cómo muestrear adecuadamente a los participantes (por

ejemplo, muestreo aleatorio tradicional vs. técnicas como el muestreo *Snowball Sampling*[88]), y cómo diseñar investigaciones que distingan adecuadamente entre causa y efecto. Muchos estudios dependen de métodos con *self-reporting* (casos donde los participantes se auto evalúan o dan información no estrictamente cuantificable de ellos mismos), los cuales pueden estar sesgados y resultar en estimaciones infladas de comportamientos como la participación en discusiones y el uso de medios.

- **Cambios rápidos en el entorno en línea:** La literatura también enfrenta el problema de la rapidez con la que evoluciona el entorno de las discusiones en línea. Esto significa que los estudios de alta calidad pueden volverse obsoletos casi tan pronto como se publican, dado que el panorama de las redes sociales y las formas de comunicación continúan cambiando.

- **Calidad de las conversaciones *online*:** La calidad de las conversaciones políticas en medios digitales y su civilidad son puntos de preocupación [165]. Mientras algunos estudios[13] encuentran niveles bajos de incivilidad y sarcasmo en plataformas como Twitter, otros notan un aumento significativo de la incivilidad en sitios como Reddit [158], especialmente en contextos políticos polarizados. Esto tiene implicaciones para entender cómo las interacciones en línea pueden influir en la polarización.

- **Efectos de la exposición a puntos de vista contrapuestos:** La exposición a argumentos de ambos lados del espectro político no necesariamente mitiga la polarización. En algunos casos, puede incluso fortalecer las convicciones propias, sugiriendo que simplemente aumentar la exposición a puntos de vista divergentes no es suficiente para reducir la polarización [85].

Es debido a las dificultades que plantean estas limitaciones y complejidades que aún no hay un consenso respecto al efecto que producen las redes sociales sobre la polarización de la sociedad. Así como mencionamos en la introducción diversos trabajos que señalan los potenciales efectos negativos, existen otros que los matizan y cuestionan. A continuación mencionaremos algunos de estos estudios.

Bakshy et al-[25] hallaron que, si bien los algoritmos de Facebook y las elecciones de los usuarios pueden crear un cierto grado de homogeneidad ideológica en las noticias que las personas ven, los usuarios todavía están expuestos a una variedad más amplia de puntos de vista de lo que algunos críticos de las "cámaras de eco" en línea sugieren [174]. Sin embargo, también destacó que esta exposición varía considerablemente entre los usuarios, dependiendo de la diversidad de sus redes de amigos y de cómo interactúan con el contenido que se les presenta.

En el trabajo "Like-minded sources on Facebook are prevalent but not polarizing" [160] concluyeron que, aunque la exposición a contenido de fuentes afines es común, reducir esta exposición no necesariamente disminuye la polarización política. El estudio no encontró efectos medibles en la polarización afectiva, el extremismo ideológico, las evaluaciones de candidatos o la creencia en afirmaciones falsas, aunque aumentó la exposición a contenido de fuentes diversas y disminuyó la exposición a lenguaje no cívico.

El estudio "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" [26] explora cómo las estructuras de comunicación en línea son flexibles y específicas a la situación, y cómo el nivel de polarización política en estas plataformas varía considerablemente según el tema en cuestión. Las conclusiones del trabajo

revelan que ciertos eventos políticos importantes, tienden a generar discusiones que se asemejan a "cámaras de eco". En contraste, eventos significativos pero no políticos muestran un patrón de conversación nacional, donde individuos de diferentes ideologías interactúan frecuentemente leyendo y retuiteando los mensajes de los demás. También menciona casos donde las discusiones comienzan siendo nacionales y luego colapsan en "cámaras de eco". Finalmente, sugieren que mientras las redes sociales pueden funcionar como "cámaras de eco" en ciertos contextos políticos, también tienen la capacidad de facilitar conversaciones más abiertas y menos polarizadas dependiendo del tipo de evento y del contexto general en que se desarrolle la discusión.

Boxwell et al. [41] estudian la evolución de la polarización en 12 países de la OECD junto con la tendencia de distintas variables que podrían contribuir a agudizar la división. Según sus resultados, Estados Unidos es el país en el que más creció la polarización junto con otros 5 países que también experimentaron aumentos de la misma. Sin embargo, en los otros 6 países detectaron una disminución de este fenómeno. Por último, cabe destacar que no hallaron una relación entre la penetración de internet en la sociedad y el crecimiento de la polarización.

Estos análisis subrayan que aún queda un considerable camino por recorrer para alcanzar un consenso sobre el papel que desempeñan las redes sociales en la fragmentación de la sociedad. Es crucial determinar si realmente intensifican esta división, si su influencia es negativa o positiva, si sus efectos dependen de variables terceras o adicionales, si generan "cámaras de eco", o si contribuyen a romper o diversificar las que ya existen fuera del ámbito digital. Con el objetivo de superar las limitaciones mencionadas y arrojar algo de luz sobre estas incógnitas aún sin resolver, esta tesis desarrolla diversas técnicas y métodos para analizar estos fenómenos desde nuevas perspectivas.

1.2. Algunos casos resonantes

Para hacer énfasis en la necesidad de desarrollar herramientas que puedan ayudar al análisis de la polarización, comentamos algunos casos puntuales que han tenido lugar en los últimos años en los cuales han ocurrido simultáneamente picos de polarización con eventos de extrema importancia en la vida real. Se ha observado en todos estos casos una fuerte intervención en el debate digital, donde se aprovecha el contexto polarizado para influir en intereses sectoriales a través de la diseminación de *fake-news*, cuentas *fake* y ataques coordinados.

1.2.1. Elecciones 2016 en EEUU

La campaña de Trump en 2016 estuvo plagada de denuncias y acusaciones cruzadas desde un principio. Empezando él mismo por denuncias sistemáticas a los principales medios de comunicación de su país por mentir u ocultar información[1], pasando por los mails de Hillary[2] hasta la afirmación de injerencia Rusa en los debates[189]. Sin embargo pocas de estas denuncias tuvieron una repercusión comparable al caso de Cambridge analítica[3] el cual llegó a sentar en el banquillo de acusados a Mark Zuckerberg para dar explicaciones

[1] https://www.cronista.com/internacionales/Trump-disparo-contra-los-medios-de-comunicacion-por-envenenar-cabezas-20161015-0008.html

[2] https://en.wikipedia.org/wiki/Hillary_Clinton_email_controversy

[3] https://www.nytimes.com/2016/11/20/opinion/cambridge-analytica-facebook-quiz.html

ante el congreso, la prensa y la sociedad internacional[4].

El servicio brindado por Cambridge logró una segmentación comunicacional sumamente precisa, mostrando a cada usuario el mensaje más efectivo posible según su perfil psicológico, logrando así una gran afinidad hacia el candidato.

El método

En 2013 un profesor de la Universidad de Cambridge llamado Aleksander Kogan desarrolló, como un proyecto personal ajeno a la universidad, una aplicación que proponía a los usuarios descubrir su personalidad. Cuando un usuario quería hacer la prueba llamada "Thisisyourdigitallife" ("ésta es tu vida digital", en español), la app solicitaba permisos para acceder a su información personal y también a la de su red de amigos[5].

De esta forma, los individuos que hacían el test y aceptaban las condiciones para ello estaban proporcionando todos sus datos al desarrollador de la app, al que, a la vez, le permitían recolectar la información de todos sus contactos. Este último dato es de gran importancia ya que, a pesar de que la aplicación la descargaron algunos cientos de miles, gracias a la información de sus contactos lograron recolectar información de más de 50 millones de personas.

Luego, según contó el ex-trabajador de Cambridge Analytica Christopher Wylie[6], Kogan vendió la información que había recabado con su app a Cambridge, que con todos estos datos en su poder hizo uso del OCEAN score[7]. Este es un test que mediante diversas preguntas analiza la composición de cinco dimensiones de personalidad: O (Openness o apertura a nuevas experiencias), factor C (Conscientiousness o responsabilidad), factor E (Extraversion o extraversión), factor A (Agreeableness o amabilidad) y factor N (Neuroticism o inestabilidad emocional). Es decir este test permite estimar el perfil psicológico de la persona.

Con el enorme volumen de información otorgado por Facebook (páginas de interés, grupos de pertenencia, referentes sociales a los que sigue, "me gusta", etc) les fue posible estimar mediante métodos de IA el puntaje de los 50 millones de usuarios sin tener que hacerles realmente el cuestionario. Este trabajo lo hicieron en base a un famoso paper [213] en el que lograron, con técnicas de machine learning, mejores resultados de la estimación de la personalidad mediante los likes de Facebook que con cuestionarios y se superó además la capacidad humana, especialmente cuando la predicción estaba relacionada a posturas políticas.

Finalmente y, una vez más, gracias a Facebook y su funcionalidad de segmentación de campañas de marketing[8] les fué posible generar distintos mensajes de campaña para cada tipo de personalidad y así maximizar la efectividad de la misma. Incluso, se intentó persuadir a demócratas de no ir a votar, según se indica en [44].. También es muy probable, aunque no hay información certera al respecto, que se hayan utilizado esos datos para la campaña fuera de Facebook y sus mensajes en los medios.

[4] `https://www.youtube.com/watch?v=mZaec_mlq9M`

[5] `https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump`

[6] `https://www.theguardian.com/uk-news/video/2018/mar/17/cambridge-analytica-whistleblower-we-spent-1m-harvesting-millions-of-facebook-profiles-video`

[7] `https://www.nytimes.com/2016/11/20/opinion/cambridge-analytica-facebook-quiz.html`

[8] `https://www.facebook.com/business/ads/ad-targeting`

1.2.2. Brexit

Según una extensa entrevista hecha por Carole Cadwalladr del diario The Guardian a Christopher Wylie (el ex trabajador de Cambridge)[45], la campaña por la desafiliación de la UE del Reino Unido estuvo fuertemente influenciada por la intervención de Cambridge en la campaña de redes gracias a la influencia de Robert Mercer, un científico de la computación experto en el análisis de datos que se hizo multimillonario mediante la creación de una empresa que utiliza algoritmos de IA para intervenir en los mercados financieros.

Mercer además tiene una importante actividad política a favor de los partidos conservadores siendo uno de los principales donantes a sus causas políticas (entre ellas la de Trump)[146] y es accionista de Cambridge. En 2016 jugó un papel clave en la campaña a favor de la salida del Reino Unido de la Unión europea, también conocida como Brexit. Según Andy Wigmore, director de comunicaciones de Leave.EU[9], Mercer cedió de forma gratuita los servicios de Cambridge Analytica a Nigel Farage, líder del Partido de la Independencia de Reino Unido (UKIP). Mediante la recogida de datos en perfiles de Facebook, la empresa pudo asesorar a Leave.EU sobre cómo influir de forma individualizada en estas personas para que apoyaran el Brexit.

El método fue muy parecido al utilizado en el caso de la campaña de Trump haciendo foco en la interpelación a personas que solían estar lejos de las discusiones políticas pero que gracias a sus análisis de personalidad sabían que podían ser potenciales seguidores del Brexit, si se los interpelaba de la forma correcta. Por ejemplo, y según Wylie, descubrieron patrones del tipo: la gente que dió like "I Hate Israel" también tiende a likear "Nike Shoes" y "Kit Kats". Esto les indicaba que haciendo una segmentación a través de estas dos últimas páginas podían llegar a un público potencialmente afin a sus ideas y más amplio.

1.2.3. Hostigamiento en discusiones

En noviembre de 2017, luego de especulaciones sobre la interferencia rusa en las elecciones presidenciales de los Estados Unidos de 2016 a través de las redes sociales, Twitter [10] publicó una lista de 2,752 cuentas[11] vinculadas a los esfuerzos de propaganda rusa. Twitter no especifica cómo se identificaron las cuentas, pero afirma que están afiliadas a la Agencia de Investigación de Internet (RU-IRA), una entidad conocida como una granja de trolls que opera cuentas de redes sociales falsas para provocar controversias y conflictos[12]. Twitter también señala que las cuentas de RU-IRA utilizaron estrategias automatizadas y no automatizadas y que algunas cuentas "parecen haber intentado organizar mítines y manifestaciones, y varias de ellas cometieron actos abusivos y hostigamiento". Twitter estima que el 9 % de los tweets de las cuentas de RU-IRA estaban relacionados con las elecciones.

En un reciente trabajo[189] se estudió la relación entre la homofilia[13] política y el organizado hostigamiento (o "trolleo") de las discusiones en torno al movimiento #Blac-

[9] `https://en.wikipedia.org/wiki/Leave.EU`

[10] A lo largo de esta Tésis hago referencia a Twitter aunque actualmente fue renombrado a X. Esto es debido a que durante el desarrollo de esta Tésis ese fué su nombre oficial.

[11] `https://democrats-intelligence.house.gov/uploadedfiles/Exhibit_b.pdf`

[12] `https://intelligence.house.gov/uploadedfiles/prepared_testimony_of_sean_j._edgett_from_twitter.pdf`

[13] Tendencia de las personas a relacionarse con sus semejantes, es decir, individuos que son similares en una o más características

kLivesMatter[14]. En el mismo se observa que las discusiones estaban polarizadas en torno a posiciones políticas y que el trolleo de estas discusiones buscó tomar ventaja de dicha "grieta". Hallaron que las cuentas denunciadas por Twitter como pertenecientes a RU-IRA tuvieron una importante participación en ambos lados de esta discusión, buscando amplificar la homofilia y aumentar así el enfrentamiento entre ambos bandos de la sociedad.

Por otra parte en nuestro país una investigación de la revista El Gato y La Caja [15] evidenció una actividad similar en torno a la discusión sobre el recorte en CONICET en 2016. A diferencia del caso anterior esta vez el trolleo tuvo como objetivo denostar y disciplinar a un sector que estaba quejándose de las políticas en Ciencia y Tecnología del gobierno.

Mediante análisis estructurales del grafo de discusiones y los contenidos de los tweets pudieron demostrar una clara diferencia de comportamiento entre las comunidades pro y anti-CONICET. La primera mostraba una actividad más "espontánea" y desestructurada, mientras que la segunda una línea mucho más específica empujada por un número reducido de cuentas altamente retuiteadas por cuentas "robots".

Casos como estos fueron ampliamente estudiados en diversos contextos [93, 212], evidenciando un problema actual de nuestra vida en las redes sociales, donde diversas fuerzas intervienen mediante técnicas computacionales en las discusiones, con el objetivo de favorecer sus intereses influenciando nuestro comportamiento u opinión sin que nos demos cuenta.

1.2.4. Medidas frente al Covid19

El artículo de Calvo et al. [20] aborda la interacción entre la política, los medios de comunicación y la percepción pública del Covid19 en Argentina. Los autores se preguntan cómo las preferencias políticas afectan las percepciones de riesgo sanitario y laboral y cómo diferentes enfoques comunicativos influyen en la transmisión de estos mensajes políticos.

Para responder a estas preguntas, se llevó a cabo un experimento utilizando tuits apareados para analizar cómo diferentes encuadres o presentaciones de la misma información sobre la Covid19 afectaban la percepción y respuesta del público. Al variar aleatoriamente elementos como el autor del tuit, el mensaje político, las imágenes, y la aceptación y apoyo por parte de otros usuarios, los investigadores pudieron medir el impacto de cada uno de estos factores en la propensión de las personas a compartir o responder a estos mensajes.

Los resultados del estudio revelan que los mensajes negativos y polarizantes tienden a disminuir la tasa de propagación de información entre los encuestados, particularmente cuando estos mensajes activan identidades partidarias. Esto sugiere que las personas pueden ser más reacias a compartir o creer en información que contradice su identidad o creencias políticas, incluso cuando se trata de una crisis de salud pública. Además, se encontró que la interpretación del riesgo por parte de las personas y sus respuestas a preguntas relacionadas están fuertemente influenciadas por estas narrativas mediáticas.

El trabajo destaca la profunda interacción entre la polarización política, los encuadres mediáticos y la percepción del riesgo en el contexto de la pandemia de Covid19 en Argentina, subrayando la importancia de considerar estos factores al formular respuestas de políticas públicas y estrategias de comunicación.

[14] `https://es.wikipedia.org/wiki/Black_Lives_Matter`
[15] `https://elgatoylacaja.com.ar/jugada-preparada/`

1.3. Avances en el procesamiento del lenguaje natural

El campo del Procesamiento de Lenguaje Natural (PLN) ha experimentado avances revolucionarios en la última década. A continuación describiremos brevemente algunos de los más destacados en relación al modelado computacional del lenguaje.

1.3.1. Hipótesis distribucional

La hipótesis distribucional es un principio fundamental en la lingüística y el procesamiento del lenguaje natural (PLN). Fue popularizada en el siglo XX y se basa en la idea de que las palabras que ocurren en contextos similares tienden a tener significados similares. En otras palabras, el significado de una palabra se puede inferir a partir del conjunto de contextos en los que aparece habitualmente.

Zellig Harris, un prominente lingüista estructuralista, es a menudo acreditado por formalizar esta hipótesis en los años 50 [94]. El resumió la idea en la frase: "Las palabras que ocurren en el mismo contexto tienden a tener significados similares".

Esta hipótesis es la base sobre la cual se centran los mas importantes hitos de PLN, que desarrollaremos en las siguientes secciones, en torno al modelado del lenguaje.

1.3.2. Word Embeddings

Uno de los primeros hitos significativos fue la introducción de "word2vec" por Mikolov et al. en 2013[147], que permitió representar palabras como vectores en un espacio continuo, capturando relaciones semánticas y sintácticas entre ellas. Para esto, entrenaron una red neuronal con el objetivo de predecir la probabilidad de co-ocurrencia entre las palabras. Una vez que esta red finaliza su entrenamiento la capa oculta es quien "aprendió"la representación vectorial de cada palabra. Esta técnica logró posicionar vectores de palabras semánticamente similares cercanos en el espacio vectorial (ver figura 1.1) a la vez que permitió realizar operaciones matemáticas entre ellos obteniendo resultados semánticamente congruentes. Como por ejemplo el famoso caso de restarle al vector de "Rey", el vector de "Hombre" y sumarle el vector de "Mujer" lo que devolvía un vector cuya posición era muy cercana al vector de "Reina".

Poco después, Facebook AI Research introdujo "FastText", que extendía el concepto de word2vec para representar no solo palabras, sino también sub-palabras (n-grams), lo que mejoraba la representación de palabras raras o mal escritas y permitía representar palabras en diferentes idiomas con un solo modelo[111].

1.3.3. Atención

Word2Vec y Fasttext inferían el significado de las palabras a través de los contextos en los que aparecían. Sin embargo, una misma palabra puede tener significados muy distintos dependiendo el contexto (Ver figura 1.3), por lo cual no sería correcto representar dos acepciones de una misma palabra con un único vector.

El mecanismo de atención, introducido por Bahdanau et al. en 2014 [22] solucionó este problema utilizando una red neuronal que se encarga de generar nuevos vectores contextualizados para cada palabra en base a los embeddings de word2vec. Esto permitió a los modelos tener un "entendimiento" más profundo del texto ya que se pudo detallar mejor el significado de cada palabra y frase según el contexto correspondiente.

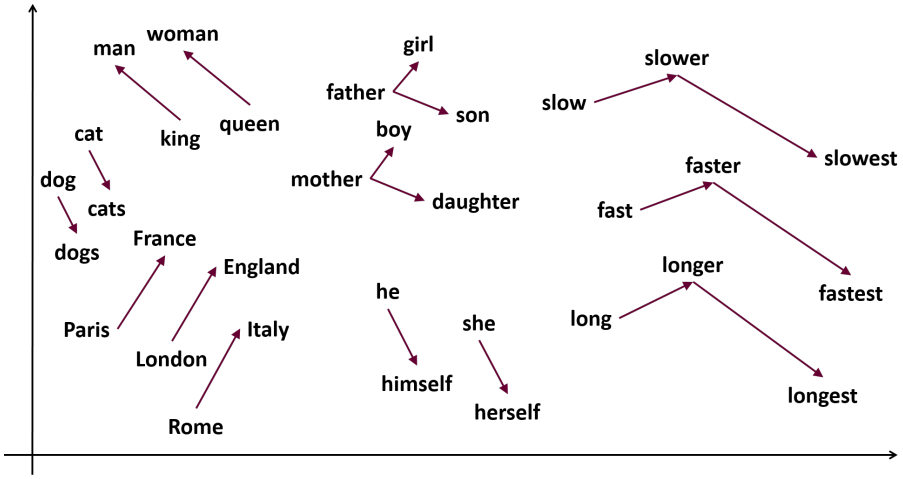


Fig. 1.1: Ilustración a forma de ejemplo sobre cómo word2vec posiciona a los vectores de palabras según su significado y similaridad

1.3.4. Transformers

Inspirado en el mecanismo de atención, Vaswani et al.[202] introdujeron en 2017 la arquitectura "Transformer", que se basa completamente en mecanismos de Atención para procesar la entrada y producir la salida . Los Transformers eliminaron la necesidad de recurrencia, ofreciendo mejoras significativas en la eficiencia y el rendimiento en una amplia variedad de tareas de PLN. Esto permitió el surgimiento de los grandes modelos de lenguaje (LLM).

1.3.5. BERT y Modelos Pre-entrenados

En 2018, Google AI presentó "BERT" (Bidirectional Encoder Representations from Transformers), un modelo basado en la arquitectura Transformer que se pre-entrena en grandes corpus de texto y luego se ajusta a tareas específicas[60]. BERT marcó una nueva era en PLN, ya que muchos modelos posteriores adoptaron el enfoque de pre-entrenamiento y *fine-tuning*, obteniendo resultados sin precedentes en numerosas tareas.

1.3.6. LLMs y Avances Recientes

Finalmente, modelos de lenguaje de gran escala como "GPT" de OpenAI, "Flan" de Google y "LlaMa" de Meta representan algunos de los modelos de vanguardia en el campo del PLN [42] a la fecha de escritura de esta tesis. Estos modelos, con cientos de miles de millones de parámetros, son capaces de realizar tareas de generación de texto, traducción, resumen y muchas otras con una precisión y fluidez notables. Son el testimonio del poder de combinar grandes cantidades de datos con arquitectura que poseen miles de millones de capas de Atención, lo que le proporciona a los modelos un gran capacidad para modelar el lenguaje.

Los LLMs son, básicamente, implementaciones a escalas enormes (en términos de cantidad de parámetros a ajustar y datos sobre los que se los entrena) de las tecnologías anteriormente mencionadas. La diferencia en la capacidad de los distintos modelos mencionados al inicio radica en como se los entrenó.

Fig. 1.2: Distintos significados de la palabra “Rey”

(a) Descripción de Luis XIV de Francia

(b) Un tweet sobre dos miembros de una banda de K-Pop

Fig. 1.3: Del lado izquierdo ’Rey’ es una palabra semánticamente cercana a ’Monarca’ o ’Príncipe’ mientras que del lado derecho está más cercana a ’Amigo’

El entrenamiento de esta tecnología implica dos fases principales: Pre entrenamiento y ajuste fino o (*fine-tuning*).

1. **Pre Entrenamiento:** Durante esta fase, el modelo desarrolla habilidades esenciales, como entender y generar lenguaje. El proceso demanda una vasta cantidad de datos, equipo computacional de alta potencia y extensas horas de procesamiento. Hasta la fecha, solo gigantes tecnológicos como Google, Meta y OpenAI han llevado a cabo esta fase con éxito. Otros actores, en su mayoría, se benefician de modelos ya preentrenados por estas empresas.

2. **Ajuste Fino (Fine-Tuning):** En esta etapa, se define el comportamiento del modelo. Esto incluye detalles como su tono conversacional, capacidad de respuesta, límites en sus respuestas y cordialidad, entre otros. A diferencia de la fase anterior, el ajuste fino es menos intensivo en recursos y es más accesible para actores medianos y pequeños. Sin embargo, demanda una mayor participación humana, sobre todo en la curación de datos, para garantizar que las respuestas se alineen con estándares específicos de conocimiento (como salud, programación, cocina) y éticos (evitando insultos, discriminación y otros comportamientos no deseados).

En suma, todos estos avances han permitido atacar numerosas tareas que eran problemáticas para PLN. Algunas que adolecían de pobres performances o sistemas realmente

complejos y difíciles de mantener. Otras, que simplemente estaban fuera del radar del estado del arte, como por ejemplo tareas de *Common Sense Reasoning*. En esta tesis me propongo hacer uso de todos estos avances para contribuir con nuevas herramientas computacionales al estudio de la polarización en las discusiones digitales. A su vez también planteo el problema que los LLMs representan dentro de este área, ya que son nuevos agentes que operan sobre esta realidad digital mediante la generación masiva de texto y la consulta sobre temas sensibles o subjetivos.

1.4. La polarización y su relación con el lenguaje

El artículo "We Don't Speak the Same Language: Interpreting Polarization through Machine Translation"[118] examina la creciente polarización observada entre partidos políticos, medios de comunicación y élites en los EE.UU., especialmente en las redes sociales. El estudio se centra en cómo diferentes comunidades perciben y utilizan el lenguaje de manera divergente, llevando a la premisa de que estas comunidades están esencialmente "hablando diferentes idiomas".

Para abordar este fenómeno, los autores introducen una metodología innovadora que utiliza la traducción automática como herramienta de análisis. La idea central es que si dos subcomunidades están utilizando el lenguaje de formas significativamente diferentes, las técnicas de traducción automática pueden ser aplicadas para identificar y traducir estas diferencias, proporcionando así una visión única de la polarización del lenguaje.

El estudio utiliza un corpus sustancial de 86.6 millones de comentarios hechos por 6.5 millones de usuarios para demostrar la aplicabilidad y eficacia de este enfoque. A través de la traducción automática, se identifican palabras y frases que tienen diferentes connotaciones y significados entre las comunidades polarizadas, lo que destaca las diferencias fundamentales en la percepción y el uso del lenguaje.

Una de las principales ventajas de este enfoque es que proporciona un marco sencillo pero potente para analizar grandes conjuntos de datos a nivel de palabras, permitiendo a los investigadores identificar y comprender las sutilezas del lenguaje que contribuyen a la polarización. Esto es especialmente relevante en la era actual, ya que como mencionamos anteriormente, las redes sociales desempeñan un papel crucial en la formación de opiniones y percepciones.

El trabajo resalta la importancia del lenguaje en la polarización y ofrece una herramienta innovadora para analizar y comprender este fenómeno a nivel granular. La traducción automática, tradicionalmente utilizada para convertir un idioma a otro, aquí se utiliza para descifrar las diferencias de lenguaje intrínsecas entre comunidades polarizadas, ofreciendo una nueva perspectiva sobre cómo el lenguaje refleja y amplifica las divisiones sociales y políticas.

Otros trabajos recientes [126, 177, 194] demuestran que las comunidades pueden expresarse con diferentes términos o formas de hablar, y usar diferentes jergas, que pueden detectarse con el uso de técnicas relacionadas con el texto. Ramponi et al.[176, 177] construyen clasificadores y predictores muy eficientes de la pertenencia de cuentas dentro de una comunidad dada, inspeccionando el vocabulario utilizado en tweets, para muchas comunidades heterogéneas de Twitter, como jugadores de ajedrez, diseñadores de moda y miembros y seguidores de partidos políticos[61]. En [194], Tran et al. descubrieron que el estilo del lenguaje, caracterizado usando un modelo híbrido de lenguaje de n-gramas de palabras y etiquetas de partes del habla, es un mejor indicador de la identidad de la comu-

nidad que el tema, incluso para comunidades organizadas en torno a temas específicos. Por último, Lahoti et al. [126] modelan el problema de aprender el espacio ideológico liberal-conservador de los usuarios de medios sociales y las fuentes de medios como un problema de factorización de matriz no negativa restringida. Validan su modelo y solución en un conjunto de datos reales de Twitter que consiste en temas controvertidos, y muestran que son capaces de separar a los usuarios por ideología con más del 90 % de pureza.

En resumen, estos estudios destacan la significativa relación entre el lenguaje y la polarización política, lo que nos permite explorar este fenómeno utilizando técnicas de PLN. Por ello, a lo largo de esta tesis, emplearemos y validaremos diversos análisis, técnicas y enfoques en distintos contextos para examinar la hipótesis de que la polarización influye en el lenguaje, generando diferentes variantes del mismo idioma según el grupo ideológico al que se pertenezca.

En este contexto, el PLN se ha convertido en una herramienta invaluable para estudiar la polarización. A medida que las fuentes de datos textuales aumentan en número y tamaño, el PLN está ganando terreno en muchos subcampos de las ciencias sociales, incluida la investigación sobre la polarización política. A diferencia de los votos o encuestas, los textos permiten a sus autores expresar una opinión más matizada y profunda. Los datos textuales de Internet reflejan comportamientos observados en contraste con las encuestas, y los métodos computacionales proporcionan acceso a estas vastas cantidades de datos.

Renáta Németh ha publicado recientemente una revisión sobre el uso del PLN en el análisis de la polarización política [154]. En la misma, señala que el campo del PLN ha mostrado un enorme crecimiento en su aplicación al estudio de este fenómeno, especialmente en la última década. Entre los trabajos analizados hallaron un predominio de las investigaciones centradas en contextos estadounidenses, con un notable 59 % de los estudios analizados orientados hacia este país. Esta concentración geográfica destaca una posible limitación en la generalización de los resultados a nivel internacional, una preocupación reforzada por el hecho de que el 80 % de los estudios en EE. UU. no contaban con autores de otros países, lo que subraya una marcada centralidad estadounidense en la temática de investigación.

Además, la metodología predominante en estos estudios ha sido la utilización de datos provenientes de Twitter y el empleo de técnicas de aprendizaje automático supervisado, representando el 43 % y el 33 % respectivamente. Aunque estas herramientas han permitido el análisis de grandes volúmenes de datos textuales, solo una minoría de estudios ha integrado conocimientos especializados para profundizar en la interpretación de las características más relevantes. Esta falta de integración de conocimiento experto señala un desafío significativo en el campo, evidenciando la necesidad de un acercamiento más interdisciplinario que no solo abarque las herramientas computacionales, sino también los fundamentos teóricos y metodológicos de las ciencias sociales.

También señala que la polarización lingüística puede aparecer en diferentes niveles de la esfera pública política, incluidos los canales oficiales de comunicación política (por ejemplo, discursos parlamentarios); los diferentes tipos de medios; los contenidos generados por los usuarios (por ejemplo, redes sociales), y la capa de expertos (por ejemplo, textos escritos por jueces). La cobertura de estas diferentes capas, ha sido un tema de interés en los estudios relevados por ellos. Sin embargo, marcan que rara vez los estudios abarcaron más de una de estas capas, lo que podría ofrecer *insights* valiosos sobre los procesos de difusión entre estos grupos.

El encuentro entre las ciencias computacionales y las ciencias aplicadas ha llevado a

una convergencia de diferentes paradigmas metodológicos. Este encuentro, sin embargo, ha enfrentado desafíos significativos, como una crisis de reproducibilidad en los campos que adaptan la inteligencia artificial. Este problema resalta aún más la importancia de un enfoque integrado y metodológicamente riguroso que pueda satisfacer los estándares de calidad en la investigación.

La evaluación detallada de la literatura revela que, aunque el PLN ostenta un potencial significativo para contribuir al estudio de la polarización política, la efectividad y pertinencia de estos esfuerzos dependerán en gran medida de la capacidad de los investigadores para amalgamar enfoques predictivos y explicativos. Asimismo, la colaboración interdisciplinaria e internacional jugará un papel crucial en la ampliación de la comprensión y aplicación de estas técnicas en estudios futuros.

1.5. Aportes de este trabajo

En esta tesis, se han realizado diversos esfuerzos para abordar y contribuir al estudio de la polarización mediante el uso de PLN. Se ha buscado no solo ampliar el entendimiento de la polarización en distintos contextos sociopolíticos, sino también desarrollar herramientas y técnicas que faciliten un análisis más profundo y riguroso. A continuación, se detallan las principales contribuciones de este trabajo, organizadas en dos secciones: la primera, centrada en cómo estos aportes abordan las limitaciones existentes y la segunda, que describe concretamente los avances y hallazgos presentados en cada capítulo de la tesis.

1.5.1. Contribuciones al uso de PLN en el estudio de la Polarización

En esta tesis, abordo varias de las limitaciones destacadas anteriormente. A diferencia de muchos estudios centrados predominantemente en Estados Unidos, mi investigación extiende su análisis a regiones menos representadas, como Argentina y Brasil, proporcionando así nuevas perspectivas sobre la dinámica de la polarización en contextos sociopolíticos diversos.

La colaboración interdisciplinaria es un pilar fundamental de este trabajo, realizando investigaciones conjuntas con expertos en ciencias sociales y colegas internacionales. Esta integración enriquece la interpretación de los datos y aporta para un enfoque holístico en el análisis. Además, la metodología adoptada en esta tesis combina técnicas de aprendizaje automático supervisadas y no supervisadas, lo cual permite una exploración más diversa de los conjuntos de datos.

En cuanto a las fuentes de datos, he optado por un enfoque más amplio que el habitual, centrado en Twitter, incorporando plataformas como Facebook, Instagram, Reddit y desgrabaciones de grupos focales. Este enfoque diversificado ayuda a capturar una gama más extensa de expresiones y opiniones, proporcionando una visión más completa de la polarización.

Otro aspecto destacado de este estudio es el análisis de la polarización en varios niveles de la esfera pública, cómo las élites y el público general. Además, se aborda la emergente influencia de los LLMs en el discurso público, explorando los posibles sesgos y discriminaciones que estos pueden introducir en la discusión política, lo que puede generar nuevas fuentes de controversia.

Finalmente, esta tesis contribuye a la reproducibilidad de la investigación en PLN y ciencias políticas mediante la publicación abierta de los conjuntos de datos y los códigos de programación utilizados, facilitando así la verificación y extensión del trabajo por parte

de otros investigadores. Este enfoque transparente no solo fortalece la integridad de la investigación, sino que también fomenta un diálogo académico más abierto y colaborativo.

1.5.2. Detalle de los Aportes por Capítulo

En el capítulo II desarrollamos dos técnicas distintas para cuantificar la polarización de una discusión en base al texto de los posteos que realizan sus usuarios y las interacciones entre ellos. Estás técnicas, a diferencia de las anteriormente conocidas en el estado del arte, tienen un funcionamiento computacionalmente mas eficiente, además de una mayor precisión a la hora de detectar si una discusión esta polarizada o no. Estas técnicas fueron publicadas en los congresos internacionales ICCS 2020 [162] y SPIRE 2020 [163]. Luego utilizamos estas (y otras) técnicas para hacer un análisis comparado de la polarización de EEUU y Argentina. Este análisis fue publicado como artículo en la revista anfibia [16].

Básandonos en la hipótesis de que la polarización tiene un efecto sobre la jerga de quienes componen sus polos, en el capítulo 5.6 desarrollamos una técnica de clustering para identificar dichos polos en base a las discusiones de los mismos. Esto permite por un lado validar la hipótesis y por otro aprovechar el contexto polarizado para identificar mas eficientemente a que polo pertenece cada individuo. Adicionalmente, como la jerga es parte del lenguaje vivo y está en constante mutación, proponemos técnicas para mantener la eficiencia del modelo a lo largo del tiempo. Una versión preliminar de este trabajo fue publicado en [100], la versión completa fue presentada como tesis de licenciatura de Martín Browarnik [17] a quien codirigí junto con Esteban Feuerstein.

Si bien en los contextos polarizados la mayoría de los usuarios se concentran en dos grandes grupos, algunos quedan ajenos a ellos. Dichos usuarios pueden representar un "puente" de entendimiento entre las visiones antagónicas de los polos. Por eso, en el capítulo 7 nos proponemos analizar a estos actores aplicando una variación de la técnica de cuantificación de la polarización sobre desgrabaciones de focus groups. Detectamos que ciertos sujetos cambian de polo en base al tópico, es decir, en algunos temas opinan como el "polo A" y en otros como el "polo B". Luego, en conjunto con un grupo de sociólogos, analizamos su comportamiento de forma cualitativa. Así hallamos que prefieren mantener una "distancia reflexiva" respecto de los medios, que valoran el pluralismo, que consideran la realidad compleja y/o difícil de comprender y en consecuencia intentan formarse un juicio propio a partir de distintas fuentes. Este trabajo fue publicado en la revista SAAP[117].

En el capítulo 7.7 nos proponemos analizar a otro tipo de actor muy relevante en los contextos polarizados: los políticos. Para esto desarrollamos técnicas y herramientas para analizar y comparar su comportamiento en las 3 principales redes del momento: Twitter, Facebook e Instagram. Hallamos que, el Oficialismo y la Oposición en argentina son más interpelativos en Twitter que en las otras dos redes, y que en esa red los políticos (independientemente del sector de pertenencia) tienden a discutir sobre temas comunes. En Twitter, además, encontramos una fuerte correlación entre el grado de toxicidad de los mensajes y su repercusión. Por el contrario, en las otras dos redes, Oficialismo y Oposición hablan principalmente de temas diferentes, sobre los que tienen más propiedad, y el nivel de toxicidad es bajo. También detectamos temas en común entre Oficialismo y Oposición en los que no hay confrontación. Esta investigación fue publicada en la revista Cuadernos

[16] `https://www.revistaanfibia.com/twitter-el-laboratorio-politico/`
[17] `http://gestion.dc.uba.ar/media/academic/grade/thesis/Tesis_Martin_Browarnik_Final.pdf`

Info [7].

Si bien la polarización política tiende a agrupar a la gente en dos grandes sectores antagónicos, siempre hay matices. Es decir, algunos usuarios son más de derecha (o izquierda) que otros dentro de su mismo "polo". Para poder medir estos matices, en el capítulo 8.4, desarrollamos una técnica que asigna un puntaje a corpus de texto en base a si su contenido esta más cercano a una ideología de derecha o de izquierda. Básandonos en el trabajo de Waller et al. [204], donde proponen una técnica con este mismo objetivo pero utilizando las interacciones de los usuarios con las comunidades, hallamos que existe una fuerte correlación en los resultados obtenidos usando las interacciónes y usando el texto de los posteos. Además, comparamos la performance de dos tipos de embbedings distintos: Fasttext (Word Embeddings) y Cohere (un LLM basado en Attention). Descubrimos que el LLM es muy superior a Fasttext, aún habiendo entrenado a Fasttext sobre el corpus de texto y utilizando pre-trained word vectors y no aplicando ningún tipo de re-entrenamiento sobre Cohere. Finalmente, la técnica propuesta por Waller requiere que se elijan dos comunidades que representen las posiciones antágonicas para poder estimar la dimensión ideológica. Nuestra técnica, al estar basada solamente en el texto permitiría poder usar cualquier tipo de semilla, no necesariamente los posteos de una determinada comunidad. También podría posicionarse cualquier tipo de corpus de texto, no sólo posteos de reddit. Por lo que habilita la posibilidad de posicionar cualquier conjunto de textos sobre cualquier espectro ideológico. Por ejemplo, posicionar los discursos de políticos sobre el espectro a favor y en contra del aborto, o heterodoxia o ortodoxia económica. Por lo tanto, esta técnica representa un gran aporte al análisis de las posiciones polarizadas de forma mucho mas general. Este trabajo fué publicado en el workshop NL4AI [58].

Los grandes modelos de lenguaje (LLM) se están volviendo omnipresentes en aplicaciones abiertas, como agentes de diálogo y asistentes de escritura. En estos entornos, se ha observado que los LLM ofrecen opiniones en respuesta a consultas subjetivas: por ejemplo, DeepMind's Sparrow dice que la pena de muerte no debería existir (Glaese et al., 2022), mientras que los modelos de Anthropic afirman que la IA no es una amenaza existencial para humanidad (Bai et al., 2022). Es por esto que resulta pertinente analizar a qué sectores de la población se parecen estos LLMs cuando opinan sobre cuestiones controvertidas. Para esto, en la última parte IV de esta tesis, utilizamos la encuesta de Latinobarométro 2020 hecha sobre la población argentina para interrogar a 3 LLMs distintos: GPT, Cohere y Bard. Dado que las preguntas eran en formato multiple chocie y tenían una relación ordinal ("Muy de acuerdo", "Muy en desacuerdo"), estimamos la distancia de cada encuestado a cada LLM como el promedio de las diferencias absoluta de cada respuesta. Luego mediante análisis multivariado analizamos la significatividad de cada variable demográfica respecto a esta distancia. Fué así que hallamos que los 3 modelos tienen correlación con el sector más masculinizado y politizado de la población argentina. Adicionalmente, Bard y GPT tuvieron correlación con la población más adulta y con altos niveles de estudio y GPT también con aquellos más de derecha. Es importante poder evidenciar estos posicionamientos de los LLMs para genera conciencia en su uso y tener una mejor noción de a quiénes representan sus opiniones sobre cuestiones controvertidas o subjetivas. Este estudio es encuentra publicado[18] en la web de Fundar[19], fundación con la que colaboramos en dicho proyecto.

[18] `https://fund.ar/publicacion/sesgos-algoritmicos-y-representacion-social-en-los-modelos-de-lenguaje-generativo`

[19] `https://fund.ar/`

A lo largo de esta tesis hemos desarrollado diversas técnicas y herramientas computacionales que permiten analizar de forma cuantitativa y computacionalmente eficiente el fenómeno de la polarización desde distintos enfoques. Además, hemos trabajado en equipos multidisciplinarios con colegas de las ciencias sociales junto con quienes hemos podido construir y, más importante aún, validar la utilidad de estas herramientas. Este enfoque conjunto subraya la importancia del trabajo colectivo para afrontar un problema tan grande y complejo como es la división de nuestra sociedad.

Como principal conclusión hallamos que la polarización tiene una fuerte correlación con el lenguaje de las personas. Lo que nos permite cuantificar de forma computacional y a través del texto si una discusión esta siendo polarizada, posicionar sobre el espectro ideológico a las comunidades en base a sus discusiones y diferencias a un polo de otro en el tiempo en base a sus posteos. Esto nos permite tener una mejor cuantificación sobre la situación de polarización en un contexto específico y tomar acciones concretas y rápidas para poder mitigar estas situaciones.

También, evidenciamos que no todos los actores necesariamente se polarizan, algunos prefieren tener posiciones ambivalentes según el tema manteniendo una distancia reflexiva. Además, mostramos como los políticos actúan sobre estos contextos, tomando distintos comportamientos en base a si están en un entorno endogámico con el público o más heterogéneo. Finalmente, evidenciamos que los nuevos actores de la vida cotidiana (los chatbos de AI), también tienen posicionamientos en estos debates y es necesario ser consciente de ellos.

Parte II

TÉCNICAS PARA LA CUANTIFICACIÓN DE LA POLARIZACIÓN

2. ¿PARA QUE CUANTIFICARLA?

La controversia en las redes sociales es un fenómeno con un alto impacto social y político. Se han realizado análisis interesantes sobre elecciones presidenciales [189], decisiones del congreso [89], propagación del odio [47] y acoso [123]. Este fenómeno ha sido ampliamente estudiado desde la perspectiva de diferentes disciplinas, desde el análisis seminal de conflictos entre los miembros de un club de karate [215] hasta problemas políticos en tiempos modernos [2, 5, 52, 57, 149].

La irrupción de las redes sociales digitales [68] dio lugar a nuevas formas de intervención intencionada para obtener ventajas [47, 189]. Además, puntos de vista altamente contrastantes en grupos tienden a provocar conflictos que llevan a ataques de una comunidad a otra, como acoso, "brigading" o "trolling"[123]. La literatura existente informa de un gran número de problemas relacionados con la controversia, desde la división de comunidades y la propagación de información sesgada, hasta el aumento de discursos de odio y ataques entre grupos. Por ejemplo, Kumar, Srijan, et al.[123] analizan muchas técnicas de defensa contra ataques en *Reddit* [1] mientras que Stewart, et al. [189] insinúan que hubo interferencia externa en *Twitter* durante las elecciones presidenciales de EE. UU. de 2016 para beneficiar a un candidato.

Como se muestra en [122, 125], detectar la controversia también proporciona la base para mejorar la "dieta de noticias" de los lectores, ofreciendo la posibilidad de conectar a los usuarios con diferentes puntos de vista recomendándoles contenido personalizado para leer [152]. Otros estudios sobre "puentes entre cámaras de eco" [84] y los efectos positivos del diálogo entre grupos [10, 171] sugieren que la participación directa es efectiva para mitigar conflictos.

Un clasificador preciso y automático de temas controversiales, por lo tanto, ayuda a desarrollar estrategias rápidas para prevenir desinformación, peleas y sesgos. Además, la identificación de los principales puntos de vista y la detección de usuarios semánticamente más cercanos también es útil para llevar a las personas a discusiones más saludables. *Medir* la controversia es aún más poderoso, ya que se puede usar para establecer niveles de controversia. Con este propósito, proponemos un proceso basado en contenido para medir la controversia en las redes sociales, recopilando el contenido de las publicaciones sobre un tema fijo (un hashtag o una palabra clave) como entrada principal.

La cuantificación de la controversia a través del análisis del vocabulario también abre varias vías de investigación, como el análisis de si la polarización está siendo creada, mantenida o aumentada a través de la forma de hablar de la comunidad.

Nuestra principal contribución puede resumirse como el diseño de un proceso de detección de controversia y su aplicación a 30 conjuntos de datos heterogéneos de Twitter. Superamos a los enfoques más avanzados, tanto en términos de precisión como de velocidad computacional.

Nuestros métodos se prueban en conjuntos de datos de Twitter. Esta plataforma de microblogging ha sido ampliamente utilizada para analizar discusiones y polarización [149, 175, 195, 206, 211]. Es una elección natural para esta tarea, ya que representa uno de los principales foros de debate público [206], es un destino común para expresiones afiliativas [98] y a menudo se utiliza para informar y leer noticias sobre eventos actuales [187]. Una

[1] https://www.reddit.com/

ventaja adicional es la disponibilidad de datos en tiempo real generados por millones de usuarios. Otras plataformas de redes sociales ofrecen servicios similares de compartición de datos, pero pocas pueden igualar la cantidad de datos y la documentación proporcionada por Twitter. Una última ventaja de Twitter para nuestro trabajo es dada por los *retweets* (compartir un tweet creado por otro usuario), que típicamente indican respaldo [32] y, por lo tanto, ayudan a modelar discusiones ya que pueden señalar "quién está con quién".

2.1. Trabajos previos

Debido a su alta importancia social, muchos trabajos se centran en medidas de polarización en redes sociales en línea y medios sociales [4, 52, 54, 83, 91]. La principal característica que conecta estos trabajos es que las medidas propuestas se basan en las características estructurales del grafo social subyacente. Entre ellos, destacamos el trabajo de Garimella et al. [83] que presenta una extensa comparación de medidas de controversia, diferentes enfoques de construcción de grafos y fuentes de datos, logrando un rendimiento de vanguardia. Usamos este enfoque como referencia para comparar nuestros resultados.

En [83] los autores proponen muchas métricas para medir la polarización en Twitter. Sus técnicas, basadas en la estructura del grafo de respaldo, pueden detectar con éxito si una discusión (representada por un conjunto de tweets) es controvertida o no, independientemente del contexto y, lo más importante, sin la necesidad de experiencia en el dominio. También incluyen dos métodos para medir la controversia basados en el análisis del contenido de las publicaciones, ambos fallidos. El primero de estos métodos comienza con la incrustación de tweets en vectores, la agrupación de estos vectores en dos grupos y un cálculo final de la divergencia KL^2 como medida de distancia entre clusters, y de la medida I2 [109] para cuantificar la heterogeneidad del cluster. El segundo método se basa en el análisis de sentimientos. Su hipótesis es que las discusiones controvertidas tienen una varianza mayor que las no controvertidas. Este enfoque está limitado por el hecho de que depende de herramientas específicas del idioma que no funcionan de manera confiable para idiomas distintos al inglés.

Matakos et al. [141] también desarrollan un *índice de polarización* con un enfoque basado en grafos, sin incluir características relacionadas con el texto, modelando opiniones como números reales. Su medida captura con éxito la tendencia de las opiniones a concentrarse en comunidades de redes, creando cámaras de eco.

Otros trabajos para la detección de controversias a través del contenido se han realizado sobre Wikipedia [65, 108] mostrando que los contenidos textuales son buenos indicativos para estimar la polarización. Estos trabajos dependen en gran medida de Wikipedia y no pueden ser extrapolados a las redes sociales.

En su tesis [107], Jang explica la controversia generando un resumen de dos posturas en conflicto que construyen la controversia. Su trabajo muestra que un subconjunto específico de tweets es suficiente para representar las dos posiciones opuestas en un debate polarizado.

En esta tesis proponemos dos técnicas nuevas para la cuantificación de la controversia. La principal diferencia entre las técnicas propuestas en 4 y 3 es que las técnicas presentadas en 4 son menos dependientes de la estructura del grafo, teniendo un proceso basado en contenido que introduce la posibilidad de definir y detectar conceptos como la "frontera semántica" de un cluster. Esto abre una posible aplicación de los resultados de esta inves-

[2] La divergencia de Kullback-Leibler es una medida de cuánto difiere una distribución de probabilidad de una distribución de probabilidad de referencia.

tigación en términos de intervención sobre la polarización. Las mejoras sobre 3 (utilizado como una segunda referencia en este trabajo), incluyen una comparación más amplia de modelos de PNL y medidas de distancia, una mayor heterogeneidad de los conjuntos de datos utilizados, y resultados en mejores rendimientos tanto en términos de puntuaciones AUC ROC como en tiempos computacionales.

2.2. Definición de tópico

En la literatura, un tópico suele definirse por un único hashtag[3]. Sin embargo, esto podría ser demasiado restrictivo en muchos casos. A veces, una discusión en un momento particular podría no tener un hashtag definido, pero podría girar en torno a una cierta *palabra clave*, es decir, una palabra o expresión que no es específicamente un hashtag pero que se utiliza ampliamente en el tópico. Por ejemplo, durante las elecciones presidenciales de Brasil en 2018, capturamos la discusión mediante las menciones a la palabra *Bolsonaro*, que es el apellido del principal candidato. En nuestro enfoque, un tópico se operacionaliza como un hashtag específico o *palabra clave*. Por lo tanto, para cada tópico, recuperamos todos los tweets que contienen uno de sus hashtags o la *palabra clave* y que se generaron durante la ventana de observación. También nos aseguramos de que el tema seleccionado esté asociado con un volumen de actividad lo suficientemente grande.

2.3. Conjuntos de datos

En esta sección detallamos las discusiones que utilizamos para probar nuestra métrica y cómo determinamos el *ground-truth* (es decir, si la discusión es realmente controversial o no). Utilizamos treinta discusiones diferentes que tuvieron lugar entre marzo de 2015 y junio de 2019, la mitad de ellas con controversia y la otra mitad sin ella. Consideramos discusiones en cuatro idiomas diferentes: inglés, portugués, español y francés, que ocurren en cinco regiones del mundo: América del Sur y del Norte, Europa Occidental, Asia Central y del Sur. También estudiamos estas discusiones tomando primero 140 caracteres y luego 280 de cada tweet para analizar la diferencia en el rendimiento y el tiempo de cálculo con respecto a la longitud de las publicaciones.

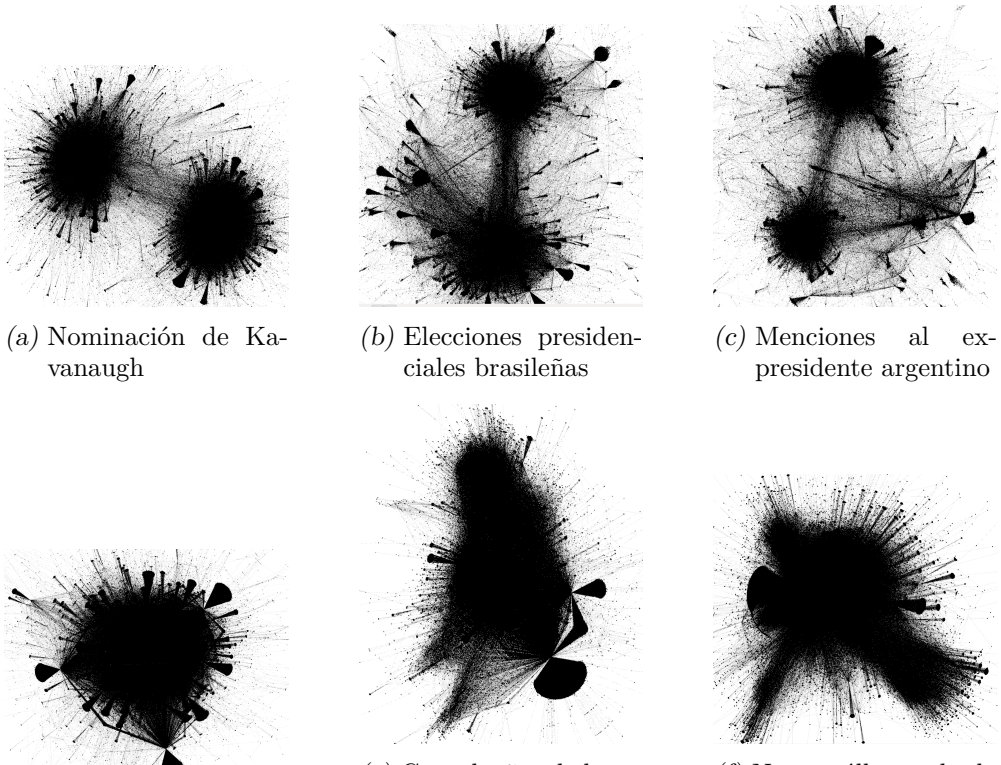
Para definir la cantidad de datos necesarios para ejecutar nuestro método, establecimos que el modelo Fasttext debe predecir al menos un usuario de cada comunidad con una probabilidad mayor o igual al 0.9 durante diez entrenamientos diferentes. Si ese no es el caso, no podemos usar los métodos aquí desarrollados. Esta decisión nos hizo considerar solo un subconjunto del total de datos utilizados en [83], porque debido al tiempo transcurrido desde su trabajo, muchos tweets habían sido eliminados y, en consecuencia, el volumen de los datos no era suficiente para nuestro marco. Para ampliar nuestra base de experimentos, agregamos nuevos debates; más información detallada sobre cada uno está disponible en el repositorio de código[4]. Para seleccionar nuevas discusiones y determinar si son controvertidas o no manualmente, buscamos temas ampliamente cubiertos por los medios de comunicación convencionales y que hayan generado una amplia discusión, tanto

[3] Un hashtag es una palabra o frase precedida por el símbolo "#" en las redes sociales que se utiliza para etiquetar y categorizar contenido relacionado en línea, facilitando su búsqueda y descubrimiento por otros usuarios.

[4] El código y las redes utilizadas en este trabajo están disponibles aquí:
`http://github.com/jmanuoz/Vocabulary-based-Method-for-Quantify-Controversy`

en línea como fuera de línea. Para las discusiones no controvertidas, nos centramos en las "noticias suaves" y el entretenimiento, pero también en eventos que, aunque impactantes y/o dramáticos, no generaron grandes controversias. Para validar esa intuición, revisamos manualmente una muestra de tweets, sin poder identificar ninguna instancia clara de controversia. Por otro lado, para los debates controvertidos, nos centramos en eventos políticos como elecciones, casos de corrupción o decisiones judiciales.

Para establecer aún más la presencia o ausencia de controversia de nuestros conjuntos de datos, visualizamos las redes correspondientes a través de ForceAtlas2 [105]. Esta es una técnica ampliamente utilizada que consiste en simular fuerzas de atracción y repulsión entre los nodos basándose en sus conexiones. Esto permite posicionarlos en la imagen de tal manera que los nodos altamente conectados se encuentren cercanos entre sí, mientras que los nodos poco conectados se sitúan a una mayor distancia. Se ha descubierto recientemente que este algoritmo es muy útil para visualizar interacciones comunitarias [203], ya que representa usuarios más cercanos interactuando entre sí y usuarios más lejanos interactuando menos. La Figura 2.1 muestra ejemplos de cómo se ven las discusiones no controvertidas y controvertidas respectivamente con el diseño de ForceAtlas2. Como podemos ver en estas figuras, en una discusión controvertida, el diseño muestra dos grupos bien separados, mientras que en una no controvertida genera un gran clúster.



(a) Nominación de Kavanaugh

(b) Elecciones presidenciales brasileñas

(c) Menciones al expresidente argentino

(d) Concierto de Halsey

(e) Cumpleaños de la estrella del pop

(f) Nuevo álbum de la banda EXO

Fig. 2.1: Diseño ForceAtlas2 para diferentes discusiones. (a), (b) y (c) son controvertidos mientras que (d), (e) y (f) son no controvertidos.

Se proporciona más información sobre los conjuntos de datos utilizados en nuestros experimentos en la siguiente Tabla 2.1.

Tab. 2.1: Estadísticas de los conjuntos de datos; el grupo superior representa temas controvertidos, mientras que el inferior representa temas no controvertidos

Hashtag/Palabras clave	**#Idioma**	**#Tweets**	**Descripción y período de recolección**
#LeadersDebate	EN	250 000	Debate entre candidatos, 11-21 Nov,2019
pelosi	EN	252 000	Impeachment a Trump, 06 Dic,2019
@mauriciomacri	ES	108 375	Menciones a Macri, 1–11 Ene,2018
@mauriciomacri	ES	120 000	Menciones a Macri, 11-18 Mar,2018
@mauriciomacri	ES	147 709	Menciones a Macri, 20-27 Mar,2018
@mauriciomacri	ES	309 603	Menciones a Macri, 05–11 Abr,2018
@mauriciomacri	ES	254 835	Menciones a Macri, 05–11 May,2018
Kavanaugh	EN	260 000	Nominación de Kavanaugh, 03 Oct,2018
Kavanaugh	EN	259 999	Nominación de Kavanaugh, 05 Oct,2018
Kavanaugh	EN	260 000	Nominación de Kavanaugh, 08 Oct,2018
Bolsonaro	PT	170 764	Elecciones brasileñas, 27 Oct,2018
Bolsonaro	PT	260 000	Elecciones brasileñas, 28 Oct,2018
Bolsonaro	PT	260 000	Elecciones brasileñas, 30-10-2018
Lula	PT	250 000	Menciones a Lula el día de las noticias sobre los chats de Moro, 11-10 Jun,2019
Dilma	PT	209 758	Impeachment a Roussef, 06-11-2015
EXODEUX	EN	179 908	Nuevo álbum de EXO, 07 Nov,2019
Thanksgiving	EN	250 000	Día de Acción de Gracias, 28 Nov,2019
#Al-HilalEntertainment	AR	221 925	Al-Hilal campeón, 01 Dic,2019
#MiracleOfChristmasEve	KO	251 974	Cumpleaños del cantante Segun Woo, 23-12-2019
Feliz Natal	PT	305 879	Deseos de Feliz Navidad, 24 Dic,2019
#kingjacksonday	EN	186 263	Cumpleaños de estrella del pop, 24–27 Mar,2019
#Wrestlemania	EN	260 000	Evento Wrestlemania, 08 Abr,2019
Notredam	FR	200 000	Incendio en Notre-Dame, 16 Abr,2019
Nintendo	EN	203 992	Lanzamiento de Nintendo, 19–28 May,2019
Halsey	EN	250 000	Concierto de Halsey, 07–08 Jun,2019
Bigil	EN	250 000	Cumpleaños de Vijay, 21–22 Jun,2019
#VanduMuruganAJITH	EN	250 000	Fans de Ajith, 23 Jun,2019
Messi	ES	200 000	Cumpleaños de Messi, 24 Jun,2019
#Area51	EN	178 220	Bromas sobre el Área 51, 13 Jul,2019
#OTDirecto20E	ES	148 061	Evento de un programa de música en España, 20 Ene,2020

3. CUANTIFICANDO LA POLARIZACIÓN A TRAVÉS DE LAS INTERACCIONES Y EL TEXTO

Este capítulo está basado en el trabajo desarrollado junto a Esteban Feuerstein (Universidad de Buenos Aires). El mismo fue presentado en el congreso ICCS 2020 [162].

3.1. Método

Nuestro enfoque para medir la controversia se basa en una forma sistemática de caracterizar la actividad de las redes sociales a través del contenido. Empleamos un proceso con cinco etapas, a saber: *construcción del grafo*, *identificación de la comunidad*, *entrenamiento del modelo*, *predicción* y *medición de la controversia*. La salida final del proceso es un valor que mide cuán controversial es un tema, siendo los valores más altos indicativos de mayores grados de controversia. El método se basa en analizar el contenido de las publicaciones a través de Fasttext [111], una biblioteca para el aprendizaje eficiente de representaciones de palabras y clasificación de frases desarrollada por el equipo de investigación de Facebook. En resumen, nuestro método funciona de la siguiente manera: a través de Fasttext entrenamos un modelo independiente del idioma que puede predecir a qué comunidad pertenecen los usuarios según su jerga. Luego tomamos sus predicciones y calculamos una puntuación basada en la noción física de *Momento Dipolar*[1] utilizando un enfoque lingüístico para identificar usuarios centrales o característicos y establecer la polaridad a través de ellos. A continuación, proporcionamos una descripción detallada de cada etapa.

3.1.1. Construcción del Grafo

Este párrafo ofrece detalles sobre el enfoque utilizado para construir grafos a partir de datos brutos. Como mencionamos en la Capítulo 2, extraemos nuestras discusiones de Twitter. Nuestro objetivo es construir un grafo de conversación que represente la actividad relacionada con un solo tema de discusión, es decir, un debate sobre un evento específico.

Para cada tema, construimos un grafo G donde asignamos un vértice a cada usuario que contribuye a él y añadimos un arista dirigida del nodo u al nodo v siempre que el usuario u retuitee un tweet publicado por v. Los retweets típicamente indican respaldo [32]: los usuarios que retuitean señalan su respaldo a la opinión expresada en el tweet original propagándola más. Los retweets no están limitados a ocurrir solo entre usuarios que están conectados en la red social de Twitter, sino que los usuarios pueden retuitear publicaciones generadas por cualquier otro usuario. Como muchos otros trabajos en la literatura [34, 47, 75, 124, 149, 189], establecemos que se necesita un retweet entre un par de usuarios para definir una arista entre ellos.

3.1.2. Identificación de la Comunidad

Para identificar la jerga de una comunidad, necesitamos ser muy precisos al definir sus miembros. Si en nuestro deseo de encontrar dos comunidades principales forzamos la

[1] En física, el momento dipolar eléctrico es una medida de la separación de cargas eléctricas positivas y negativas dentro de un sistema, es decir, una medida de la polaridad general del sistema

partición del grafo en ese número preciso de comunidades, podríamos estar añadiendo ruido en la jerga de las principales comunidades que se enfrentan entre sí. Por ello, agrupamos el grafo utilizando dos algoritmos populares: Walktrap [172] y Louvain [37]. Ambos son algoritmos basados en estructuras que tienen un muy buen rendimiento con respecto a la medida de Modularidad Q[2]. En este contexto la modularidad Q podría entenderse como la detección de comunidades en base a qué tanto se comunican los usuarios dentro de ellas en comparación con cuánto lo hacen con usuarios fuera de la comunidad. Es decir, si en una red, Louvain identifica dos comunidades A y B, esto indica que los usuarios de la comunidad A se comunican mucho más entre ellos que con los de la comunidad B y viceversa. Estas técnicas no detectan un número fijo de grupos; su resultado depende de la optimización de la Modularidad Q, lo que resulta en comunidades menos "ruidosas". Las principales diferencias entre los dos métodos, en lo que respecta a nuestro trabajo, son que Louvain es un algoritmo heurístico mucho más rápido pero produce grupos con peor Modularidad Q. Por lo tanto, para analizar la compensación entre el tiempo de cálculo y la calidad, decidimos probar ambos métodos. En esta etapa queremos capturar los tweets de las principales comunidades para crear el modelo que podría diferenciarlas. Por lo tanto, tomamos las dos comunidades identificadas por el algoritmo de agrupación que tienen el mayor número de usuarios y las utilizamos para el siguiente paso de nuestro método.

3.1.3. Entrenamiento del Modelo

Después de detectar las principales comunidades, creamos nuestro conjunto de datos de entrenamiento para alimentar el modelo. Para ello, extraemos los tweets de cada grupo y los sometemos a algunas transformaciones. Primero, eliminamos los tweets duplicados, es decir, los retweets sin texto adicional. En segundo lugar, eliminamos del texto los nombres de usuario, enlaces, puntuación, tabulaciones, espacios iniciales y finales, espacios generales y "RT", el texto que indica que un tweet es de hecho un retweet.
Como se muestra en trabajos anteriores, los emojis[3] están correlacionados con el sentimiento [159]. Además, dado que pensamos que las comunidades expresarán diferentes sentimientos durante la discusión, es previsible que los emojis desempeñen un papel importante como separadores de tweets que diferencian entre los dos lados. Por lo tanto, decidimos agregarlos al conjunto de entrenamiento traduciendo cada emoji en una palabra diferente. Por ejemplo, el emoji :) se traducirá como *feliz* y :(como *triste*. Las relaciones entre emojis y palabras están definidas en la biblioteca R *textclean*[4].

Finalmente, agrupamos los tweets por usuario concatenándolos en una cadena y etiquetándolos con la comunidad del usuario, a saber, con las etiquetas *C1* y *C2*, correspondientes al grupo más grande y al segundo grupo más grande respectivamente. Es importante destacar que tomamos el mismo número de usuarios de cada comunidad para evitar sesgos en el modelo. De este modo, utilizamos el número de usuarios de la comunidad principal más pequeña.

El conjunto de entrenamiento construido de esta manera se utiliza para alimentar el modelo. Como dijimos, usamos Fasttext [111] para este entrenamiento. Para definir los valores de los hiperparámetros utilizamos los hallazgos de [210], donde se encuentran los mejores hiperparámetros para entrenar modelos de *embeddings* usando Fasttext y datos

[2] Q(G)=$\sum_{C \in G}(e_c - a_c)$, donde G es el grafo, C cada una de sus comunidades, e_c la fracción de aristas internas y a_c la fracción de aristas en el borde

[3] https://emojipedia.org/twitter/

[4] https://cran.r-project.org/web/packages/textclean/textclean.pdf

de Twitter. También cambiamos el valor predeterminado del hiperparámetro *epoch* a 20 en lugar de 5 porque queremos más convergencia previniendo tanto como sea posible la variación entre diferentes entrenamientos. Estos valores podrían cambiar en otros contextos o redes sociales donde tengamos más texto por usuario o diferentes dinámicas de discusión.

3.1.4. Predicción

La siguiente etapa consiste en identificar a los *usuarios característicos* de cada lado de la discusión. Estos son los usuarios que mejor representan la jerga de cada lado. Para ello, los tweets de los usuarios pertenecientes al componente conectado más grande del grafo se desinfectan y transforman exactamente como en el paso de Entrenamiento.

Decidimos restringirnos al componente conectado más grande porque en todos los casos contiene más del 90 % de los nodos. El 10 % restante de los usuarios no participa en la discusión desde un punto de vista colectivo, sino más bien de forma aislada y este tipo de intervención no añade información interesante a nuestro enfoque. Luego, eliminamos de este componente a los usuarios con grado menor o igual a 2 (es decir, usuarios que fueron retuiteados por otro usuario o retuitearon a otra persona menos de tres veces en total). Su participación en la discusión es marginal, por lo que no son relevantes en cuanto a la controversia ya que añaden más ruido que información al momento de medir. Este paso podría ajustarse de manera diferente en una red social diferente. Nombramos a este componente resultante *grafo-raíz*.

Finalmente, clasificamos a los usuarios. Considerando que Fasttext devuelve para cada clasificación tanto la etiqueta predicha como la probabilidad de la predicción, clasificamos a cada usuario del componente resultante con nuestro modelo entrenado, y tomamos a los usuarios que fueron etiquetados con una probabilidad mayor o igual al 0.9. Estos son los *usuarios característicos* que se utilizarán en el siguiente paso para calcular la medida de controversia.

3.1.5. Medida de Controversia

Esta sección describe las medidas de controversia utilizadas en este trabajo. Este cálculo se inspira en la medida presentada por Morales et al. [149] y se basa en la noción de momento dipolar que tiene su origen en la física.

Primero, asignamos a los *usuarios característicos* la probabilidad devuelta por el modelo, negativizando el valor si la etiqueta predicha era *C2*. Por lo tanto, a estos usuarios se les asignan valores en el conjunto [-1,-0.9] ∪ [0.9,1]. Luego, establecemos valores para el resto de los usuarios del *grafo-raíz* mediante propagación de etiquetas [217], un algoritmo iterativo para propagar valores a través de un grafo por vecindad del nodo.

Definimos a n^+ y n^- cómo el número de vértices con valores positivos y negativos respectivamente, a V cómo el número total de vértices, y a $\Delta A = \frac{\mid n^+ - n^- \mid}{\mid V \mid}$ cómo la diferencia absoluta de su tamaño normalizado. Además, definimos a gc^+ y gc^-) como los valores promedio entre los vértices n^+ y n^- respectivamente y establecemos τ como la mitad de su diferencia absoluta, $\tau = \frac{\mid gc^+ - gc^- \mid}{2}$. La medida de controversia del contenido del momento dipolar se define como: $DMC = (1 - \Delta A)\tau$.

La justificación para esta medida es que si los dos lados están bien separados, entonces la propagación de etiquetas asignará diferentes valores extremos a las dos particiones,

donde los usuarios de una comunidad tendrán valores cercanos a 1 y los usuarios de la otra a -1, lo que lleva a valores más altos de la medida *DMC*. También se debe tener en cuenta que las diferencias más grandes en el tamaño de las dos particiones (reflejadas en el valor de ΔA) llevan a valores más pequeños para la medida, que toma valores entre cero y uno.

Cabe preguntarse aquí si distintas distribuciones de probabilidad en la predicción del modelo entrenado afectarían al método. Es decir si una distribución normal podría tener un funcionamiento distinto que una bimodal en la aplicación de esta técnica. Dado que este umbral lo definimos sólo para seleccionar a los *usuario característicos*, y poder así anotarlos en el grafo antes de correr la propagación de etiquetas[217], tener distintas distribuciones sólo cambiaría la cantidad de nodos etiquetados previos a dicha propagación. Por lo cual, sólo se vería afectado el tiempo de cómputo, no la precisión del método.

Sí podría surgir un problema si el modelo funcionara mal y una predicción con una probabilidad mayor a 0.9 no fuera fiable. Discutiremos sobre esto en la sección 3.3.

3.2. Experimentos

Ejecutamos el método anterior sobre diferentes discusiones, obteniendo los siguientes resultados.

3.2.1. Resultados

Entrenar un modelo Fasttext no es un proceso determinístico, ya que diferentes ejecuciones pueden producir diferentes resultados incluso utilizando el mismo conjunto de entrenamiento en cada una. Para analizar si estas diferencias son significativas, decidimos calcular 20 puntuaciones para cada discusión. Las desviaciones estándar entre estas 20 puntuaciones fueron bajas en todos los casos, con una media de 0.01 y un máximo de 0.05. En consecuencia, decidimos informar en esta sección el promedio entre las 20 puntuaciones; en la práctica, tomar el promedio entre 5 ejecuciones sería suficiente. La Figura 3.1 muestra las puntuaciones calculadas por nuestra medida en cada tema para los dos métodos de agrupamiento.

El beanplot muestra la función de densidad de probabilidad estimada para una medida calculada en los temas, las observaciones individuales se muestran como pequeñas líneas blancas en un gráfico de dispersión unidimensional, y la mediana como una línea negra más larga. El beanplot se divide en dos grupos, uno para temas controvertidos (izquierda/oscuro) y otro para temas no controvertidos (derecha/claro). Por lo tanto, el grupo negro muestra la distribución de puntuaciones en discusiones controvertidas y el grupo gris en discusiones no controvertidas. Una mayor separación de las dos distribuciones indica que la medida es mejor para capturar las características de los temas controvertidos, porque una buena separación permite establecer un umbral en la puntuación que separa las discusiones controvertidas de las no controvertidas.

Como podemos ver en la figura, las medianas están bien separadas en ambos casos, con poca superposición. Para cuantificar mejor esta superposición, medimos la sensibilidad [137] de estas predicciones midiendo el área bajo la curva ROC (AUC ROC), obteniendo un valor de 0.98 para el agrupamiento Walktrap y 0.967 para Louvain (donde 1 representa una separación perfecta y 0.5 significa que son indistinguibles).

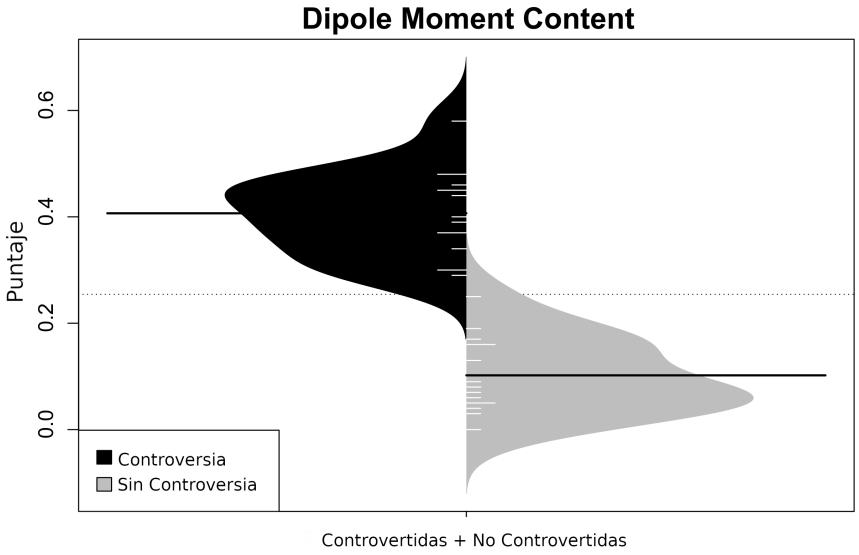


Fig. 3.1: Puntuaciones promedio de controversia en 20 ejecuciones en conjuntos de datos con 280 caracteres

Dado que Garimella et al. [83] han hecho público su código[5], reproducimos su mejor método, *Randomwalk*[6], en nuestros conjuntos de datos y medimos el AUC ROC, obteniendo una puntuación de 0.935. Un hallazgo interesante fue que su método tuvo un rendimiento deficiente en sus propios conjuntos de datos. Esto se debió al hecho (ya explicado en la Sección 3.2) de que no fue posible recuperar las discusiones completas; además, en ningún caso pudimos recuperar más del 50 % de los tweets. Por lo tanto, decidimos eliminar estas discusiones y medir nuevamente el AUC ROC de este método, obteniendo un valor de 0.99. Nuestra hipótesis es que el rendimiento de ese método se vio gravemente afectado por la falta de integridad de los datos. También probamos nuestro método en estos conjuntos de datos, obteniendo un AUC ROC de 0.99 con Walktrap y 0.989 con el agrupamiento Louvain.

Concluimos que nuestro método funciona mejor, ya que en la práctica ambos enfoques muestran el mismo rendimiento, especialmente con Walktrap. Pero en presencia de información incompleta, nuestra medida es más robusta. El rendimiento de Louvain es ligeramente peor, pero, como mencionamos en la Sección 3.1, este método es mucho más rápido. Por lo tanto, decidimos comparar el tiempo de ejecución de nuestro método con ambas técnicas de agrupamiento y también con el algoritmo *Randomwalk*. En la Figura 3.3 podemos ver la distribución de los tiempos de ejecución de todas las técnicas a través de box plots. Ambas versiones de nuestro método son más rápidas que *Randomwalk*, mientras que Louvain es más rápido que Walktrap.

Luego analizamos el impacto de la longitud del texto considerado en nuestro método. La Figura 3.2 muestra los resultados de un experimento similar al de la Figura 3.1, pero considerando solo 140 caracteres por tweet. Como podemos ver, aquí la superposición es mayor, teniendo un AUC de 0.88. En cuanto al impacto en el tiempo de cálculo, en la práctica observamos un crecimiento lineal en función del tamaño del texto. Resul-

[5] https://github.com/gvrkiran/controversy-detection
[6] Esta es una medida basada en caminatas aleatorias sobre la estructura del grafo

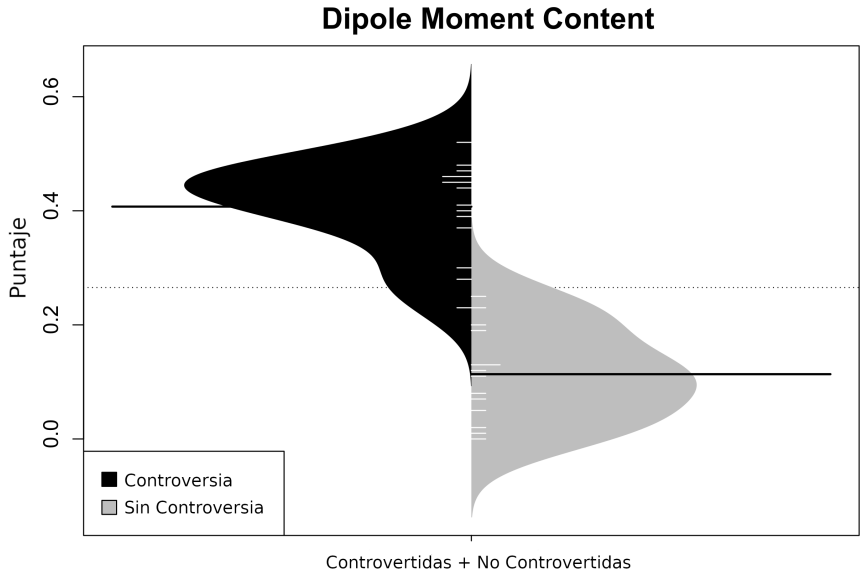


Fig. 3.2: Puntuaciones promedio de controversia en 20 ejecuciones en conjuntos de datos de 140 caracteres por tweet

tados anteriores de [111] informaron una complejidad de O(h log_2(k))[7] en las tareas de entrenamiento y prueba.

Medimos los tiempos de ejecución de las fases de entrenamiento y predicción (las dos fases relacionadas con el texto de nuestro método), los tiempos resultantes se informan en la Figura 3.4, que muestra el tiempo de ejecución en función del tamaño del conjunto de texto. También incluimos la mejor función estimada que aproxima el tiempo de cálculo en función del tamaño del conjunto de texto. Como se puede ver, el tiempo crece casi linealmente, variando desde 30 segundos para un conjunto de 111 KB hasta 84 segundos para un conjunto de 11941 KB[8]. Finalmente, medimos los tiempos de ejecución para todo el método en cada conjunto de datos con 280 caracteres. Los tiempos estuvieron entre 170 y 2467 segundos con una media de 842, lo que en la práctica representa una cantidad razonable de tiempo.

3.3. Discusiones

La tarea que abordamos en esta parte de la tesis ciertamente no es fácil, y nuestro estudio tiene algunas limitaciones, las cuales discutimos en esta sección. Nuestro trabajo nos lleva a algunas conclusiones sobre la posibilidad general de medir la controversia a través del texto y qué aspectos deben considerarse para profundizar nuestro trabajo.

3.3.1. Limitaciones

Dado que nuestro enfoque de la controversia es similar al de Garimella et al. [83], compartimos algunas de sus limitaciones con respecto a varios aspectos: *Evaluación* -

[7] Donde k es el número de clases y h la dimensión de la representación del texto

[8] Comparamos modelos polinómicos de grado 1 a 5 y logmodel, el modelo lineal tiene el error RMSE más bajo entrenando con validación cruzada de 10-fold.

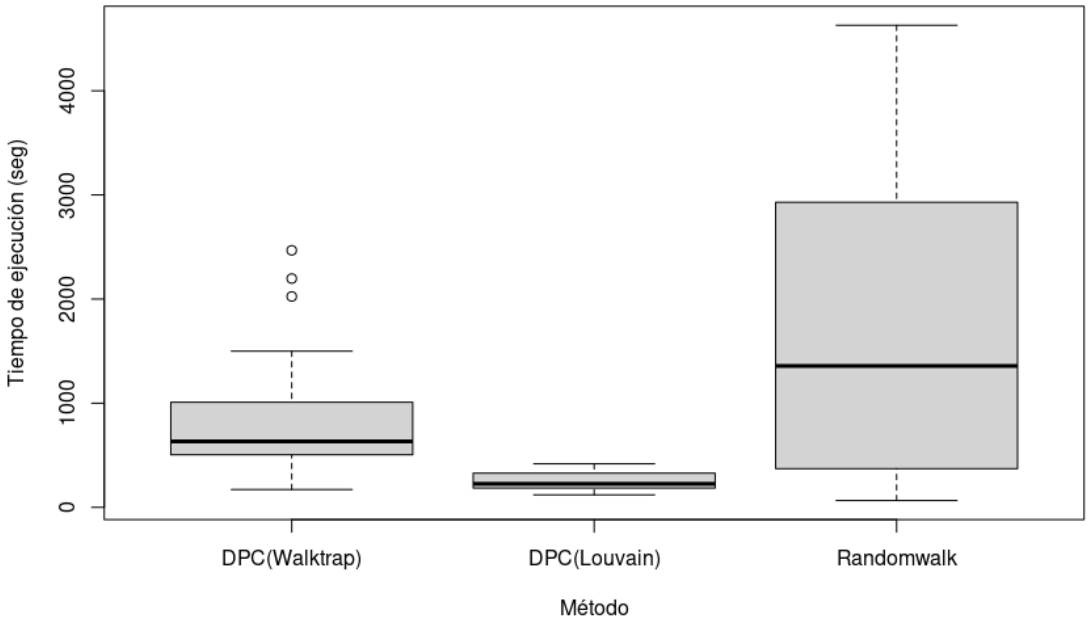


Fig. 3.3: Medidas del tiempo de ejecución de nuestro método con cada tipo de agrupamiento y algoritmo Randomwalk

dificultades para establecer el *ground-truth*, *Controversias multisided* -controversia con más de dos polos, *Elección de datos* - selección manual de temas, y *Sobreajuste* - pequeño conjunto de experimentos.

Aunque tenemos más discusiones respecto a [83], todavía es un conjunto pequeño desde un punto de vista estadístico. Aparte de eso, nuestro enfoque basado en el lenguaje tiene otras limitaciones que mencionamos a continuación, junto con sus soluciones o mitigación.

Tamaño de los datos. Entrenar un modelo de NLP que pueda predecir etiquetas con una probabilidad mayor o igual que 0.9 requiere una cantidad significativa de texto, por lo que nuestro método funciona solo para discusiones "grandes". La mayoría de las controversias interesantes son aquellas que tienen consecuencias a nivel de sociedad, en general lo suficientemente grandes para nuestro método.

Discusiones en varios idiomas. Cuando varios idiomas participan en una discusión, es común que los usuarios tiendan a retuitear más tweets en su propio idioma, creando subcomunidades. En estos casos, nuestro modelo tenderá a predecir puntuaciones de controversia más altas. Este es el caso, por ejemplo, de #germanwings, donde los usuarios tuitean en inglés, alemán y español y tiene la puntuación más alta en temas no controvertidos. Sin embargo, la polarización que abordamos en este trabajo es normalmente parte de una célula de la sociedad (una nación, una ciudad, etc.), y por lo tanto se desarrolla en un solo idioma. Creemos que limitar la efectividad de nuestro análisis a discusiones en un solo idioma no es una limitación seria.

Solo Twitter. Nuestros hallazgos se basan en conjuntos de datos provenientes de Twitter. Aunque esto es ciertamente una limitación, Twitter es uno de los principales lugares para la discusión pública en línea y uno de los pocos para el cual hay datos disponibles. Por

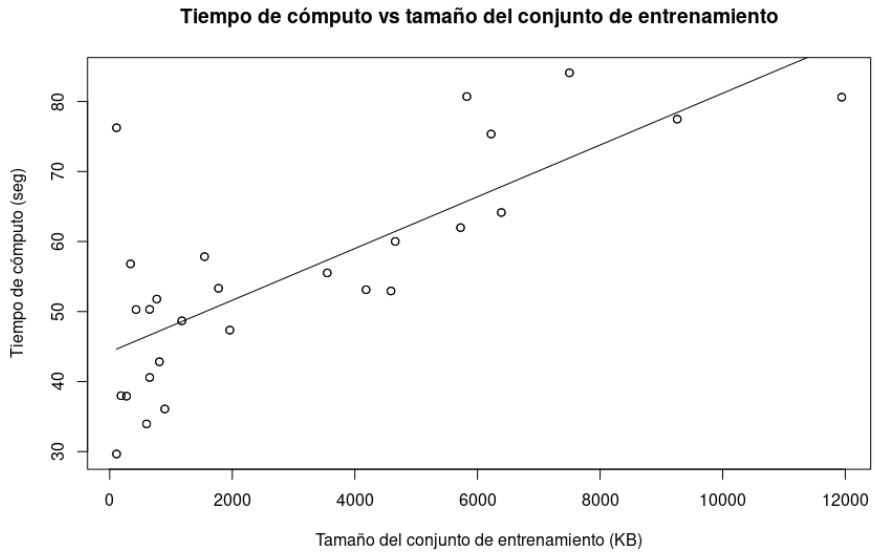


Fig. 3.4: Medidas de tiempo de las ejecuciones relacionadas con el texto en función del tamaño del conjunto de texto

lo tanto, Twitter es una elección natural. Sin embargo, el límite característico de Twitter de 280 caracteres por mensaje (140 hasta hace poco tiempo) es una limitación intrínseca de esa red. Creemos que en otras redes sociales como Facebook o Reddit, nuestro método funcionará aún mejor, ya que tener más texto por usuario podría redundar en un mejor modelo de NLP, como verificamos al comparar los resultados con 140 y 280 caracteres por publicación.

Precisión de de la predicción. Si bien Fasttext es un modelo ampliamente utilizado en el estado del arte, la probabilidad de sus predicciones podrían no ser correctas, lo cual generaría *usuarios característicos* erróneos lo que produciría un mal funcionamiento del método. En trabajos futuros podría implementarse una mejora en este sentido mediante la calibración[156] del modelo y la utilización de modelos más modernos y eficientes como los LLMs.

3.3.2. Conclusiones

En este capítulo, presentamos el primer método sistemático a gran escala para cuantificar la controversia en las redes sociales a través del contenido. Hemos demostrado que este método funciona en español, inglés, francés y portugués, es independiente del contexto y no requiere la intervención de un experto en el dominio.

Hemos comparado su rendimiento con medidas de controversia basadas en estructura de vanguardia, mostrando que tiene el mismo rendimiento y también es más robusto. También hemos demostrado que más texto implica un mejor rendimiento sin aumentar significativamente el tiempo de cálculo, por lo tanto, el método podría usarse en otros contextos como otras redes sociales como Reddit o Facebook. Planeamos probarlo en esas redes como trabajo futuro.

Entrenar el modelo no es una tarea costosa ya que Fasttext tiene un buen rendimiento en esto. Sin embargo, el mejor rendimiento para detectar comunidades principales se

obtiene mediante Walktrap. La complejidad de ese algoritmo es O(mn^2)[172], donde m y n son el número de aristas y vértices respectivamente. Esto hace que este método sea bastante costoso de calcular en redes grandes. Sin embargo, hemos demostrado que con Louvain el método aún obtiene un AUC ROC muy similar (0.99 con Walktrap y 0.989 con Louvain). Con información incompleta, su rendimiento empeora, pero sigue siendo bueno (0.96) y mejor que el estado anterior del arte.

Este trabajo abre varias vías para investigaciones futuras. Una es identificar qué palabras, semánticas/conceptos o expresiones de lenguaje hacen que una comunidad se diferencie de la otra. Hay varias formas de hacerlo, por ejemplo, a través de los embeddings de palabras que Fasttext devuelve después del entrenamiento [111]. También podríamos usar técnicas de interpretabilidad en modelos de aprendizaje automático [66]. Finalmente, podríamos probar otras técnicas para medir la controversia a través del texto, usando otro modelo de NLP como la red neuronal preentrenada BERT [60] o, en un enfoque completamente diferente, midiendo el índice de dispersión de los embeddings de palabras de las discusiones [176]. En el siguiente capítulo desarrollamos otra técnica utilizando esta última idea.

4. CUANTIFICANDO LA POLARIZACIÓN A TRAVÉS DE EMBEDDINGS

Este capítulo está basado en el trabajo desarrollado junto a Marco Di Giovanni (Politecnico di Milano), Esteban Feuerstein (Universidad de Buenos Aires) y Marco Brambilla (Politecnico di Milano). El mismo fue presentado y publicado en el congreso SPIRE 2020 [163].

4.1. Metodología

A diferencia de la técnica anterior esta se enfoca más en el texto de las conversaciones que en las interacciones. Sin embargo, es necesario utilizar las interacciones para hallar a los dos principales comunidades de la potencial controversia.

Por esto es que las dos primeras fases de este enfoque son las mismas que las descriptas en 3.1.1 y 3.1.2 y la diferencia radica en los siguientes pasos: la *fase de embedding* y la *fase de cálculo del puntaje de controversia*.

Al igual que en el capítulo anterior, el resultado final del proceso es un valor positivo que mide la controversia de un tema, donde valores más altos corresponden a grados más bajos de controversia.

Nuestra hipótesis es que, utilizando los embeddings generados por un modelo de PNL, podemos distinguir diferentes formas de hablar; cuanto más controversial es la discusión, mejor diferenciación obtenemos.

4.1.1. Embeddings

En esta fase, nuestro propósito es vectorizar a cada usuario en su embedding correspondiente. Estos vectores codifican propiedades sintácticas y semánticas de las publicaciones de las cuentas correspondientes. Se utilizarán en la siguiente fase para calcular el puntaje de controversia, ya que necesitamos vectores semánticamente significativos de dimensión fija para realizar los cálculos posteriores.

En primer lugar, los tweets pertenecientes a los usuarios de las dos comunidades principales seleccionadas en la etapa anterior se agrupan por usuario y se sanean. Eliminamos duplicados y, de cada tweet, eliminamos nombres de usuarios, enlaces, puntuación, tabulaciones, espacios en blanco iniciales y finales, espacios generales y la palabra clave de retweet “RT”, la cadena que indica que un tweet es en realidad un retweet. Se han desarrollado muchas técnicas de vectorización de frases en la literatura, desde modelos simples de bolsa de palabras hasta complejos modelos de lenguaje profundo. Para realizar este paso, seleccionamos dos modelos entre los más avanzados, a saber, Fasttext y BERT, que transforman textos en vectores de dimensión fija que codifican significado y significancia semántica.

Fasttext

[111] Es una herramienta basada en el modelo skipgram, donde cada palabra se representa como una bolsa de n-gramas de caracteres. A cada n-grama de caracteres se le asocia una representación vectorial; las palabras se representan como la suma de estas

representaciones. Este es un método rápido que permite entrenar rápidamente modelos en grandes corpus y calcular representaciones de palabras también para palabras que no aparecen en los datos de entrenamiento. Entrenamos este modelo con datos etiquetados, de acuerdo con el resultado de Louvain (etapa anterior), representando la comunidad del usuario. Para definir los valores de los hiperparámetros utilizamos los hallazgos de [210], donde los autores investigan los mejores hiperparámetros para entrenar modelos de embedding de palabras usando Fasttext y datos de Twitter. Usamos el modelo entrenado para calcular el embedding de texto.

BERT

Representaciones Bidireccionales de Codificadores de Transformadores o en inglés Bidirectional Encoder Representations from Transformers(BERT) [59] es un modelo de representación de lenguaje profundo del estado del arte basado en Transformers [202] preentrenado de forma no supervisada en el volcado completo de Wikipedia para más de 100 idiomas. El modelo está diseñado para el aprendizaje por transferencia, por lo que debe ser re-entrenado durante algunos *epochs* para tareas específicas, insertando una capa completamente conectada adicional en la parte superior, sin modificaciones sustanciales específicas de la tarea en la arquitectura. Usamos la versión BASE de BERT (12 capas, dimensión oculta de 768, 12 cabezas por capa, para un total de $110M$ parámetros).

Dado un conjunto de datos de tweets etiquetados de acuerdo con el resultado de Louvain (etapa anterior), re-entrenamos BERT en una tarea de clasificación de 2 clases durante 6 *epochs* (tasa de aprendizaje establecida en 10^{-5}). Dado que nuestro objetivo es obtener embeddings de tweets, después del procedimiento de entrenamiento eliminamos la capa completamente conectada y usamos las salidas de BERT como embeddings. En detalle, BERT primero divide una frase en tokens, añadiendo el token *[CLS]* al principio. Luego, vectorizamos cada token en un vector de 786 dimensiones. Dado que necesitamos un único vector de longitud fija para calcular nuestro puntaje, seleccionamos como agregador el embedding del token *[CLS]*. Esta es la misma estrategia seleccionada durante la etapa de *fine-tunning*. Realizamos esta etapa usando el repositorio de GitHub bert-as-service [208].

Para entrenar Fasttext y BERT de manera supervisada, necesitamos crear un conjunto de entrenamiento con sus etiquetas. Etiquetamos cada usuario con su comunidad, es decir, con las etiquetas C_1 y C_2, correspondientes respectivamente a los grupos más grande (Comunidad 1) y segundo más grande (Comunidad 2). Es importante repetir que, para evitar sesgos en el modelo, tomamos el mismo número de usuarios de cada comunidad, reduciendo la primera comunidad principal al número de usuarios de la segunda.

4.1.2. Cálculo del Puntaje de Controversia

Para calcular el puntaje de controversia, seleccionamos algunos usuarios como los mejores representantes del punto de vista principal de cada lado. Los estimamos mediante el algoritmo HITS [119] para estimar el puntaje autoritativo y de hub de cada usuario. HITS (Hypertext Induced Topic Search) asigna dos puntajes a cada nodo en una red: uno para medir su autoridad, que representa su importancia como fuente de información confiable, y otro para medir su hubness, que mide su capacidad para conectar nodos autoritativos. Estos puntajes se calculan iterativamente considerando las relaciones de enlace entre los nodos, con los nodos autoritativos siendo aquellos que reciben enlaces de nodos hub y viceversa.

Una vez corrido este algoritmo, tomamos el 30 % de los usuarios con el puntaje autoritativo más alto y el 30 % con el puntaje de hub más alto y los llamamos *usuarios centrales*.

Finalmente, calculamos el puntaje de controversia r, utilizando los embeddings de los usuarios centrales $x_i \in \mathbb{R}^k$ y las etiquetas $y_i \in \{1, 2\}$, imponiendo su pertenencia al grupo C_1 o C_2, calculado durante la fase de identificación de la comunidad.

Calculamos los centroides de cada grupo j con la ecuación 4.1, donde $|C_j|$ es la magnitud del grupo C_j, y un centroide global c_{glob} con la ecuación 4.2.

$$c_j = \frac{1}{|C_j|} \sum_{i:y_i=j} x_i \tag{4.1}$$

$$c_{glob} = \frac{1}{|C_1| + |C_2|} \sum_{i} x_i \tag{4.2}$$

Definimos D_j como la suma de las distancias entre los embeddings x_i y sus centroides c_j usando la ecuación 4.3 para $j = 1{,}2$, donde $dist$ es una función de distancia genérica. De manera similar, D_{glob} es la suma de distancias entre todas los embeddings y el centroide global.

$$D_j = \sum_{i:y_i=j} dist(x_i, c_j) \tag{4.3}$$

Debido a la *maldición de la dimensionalidad* [29], medir distancias sobre un gran número de dimensiones no es una tarea trivial y la utilidad de una medida de distancia depende de los subespacios a los que pertenece el problema [184]. Por esta razón, seleccionamos y probamos cuatro medidas de distancia: L_1 (Manhattan), L_2 (Euclidiana), Cosine y Mahalanobis [55] (particularmente útil cuando el espacio vectorial no es interpretable ni homogéneo, ya que también tiene en cuenta las correlaciones del conjunto de datos y se reduce a la distancia euclidiana si la matriz de covarianza es la matriz identidad).

El puntaje de controversia r se define en la ecuación 4.4.

$$r = \frac{D_1 + D_2}{D_{glob}} \tag{4.4}$$

Intuitivamente, representa cuánto están separados los grupos. Esperamos que, si el conjunto de datos es una única nube de puntos, este valor debería estar cerca de 1 ya que los dos centroides c_1 y c_2 estarán cerca el uno del otro y cerca del centroide global c_{glob}. Por el contrario, si los embeddings dividen con éxito el conjunto de datos en dos grupos claramente separados, sus centroides estarán lejos y cerca de los puntos que pertenecen a sus propios grupos. Cabe destacar que r es, por definición, positivo, ya que D_1, D_2 y D_{glob} también lo son.

Los conjuntos de datos y el código completo están disponibles en github[1] y los resultados discutidos en la siguiente sección son completamente reproducibles.

4.2. Resultados

En esta sección recopilamos los resultados obtenidos con las diferentes técnicas descritas anteriormente y los comparamos con el método estructurado de vanguardia "RW" [83]

[1] https://github.com/jmanuoz/Measuring-controversy-in-Social-Networks-through-NLP

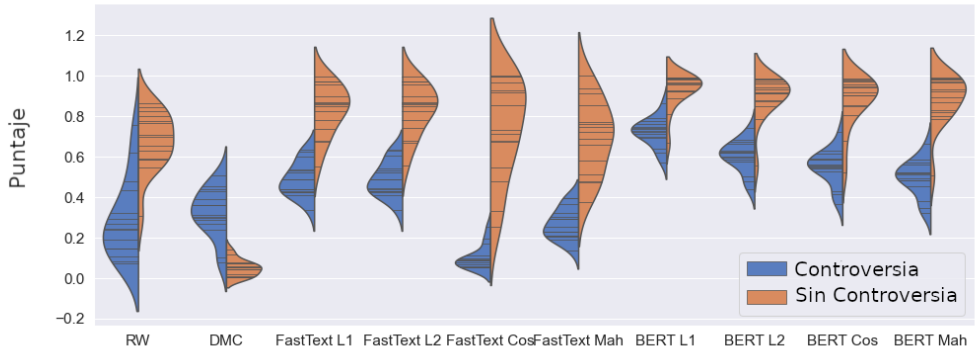


Fig. 4.1: Comparación de distribuciones de puntajes

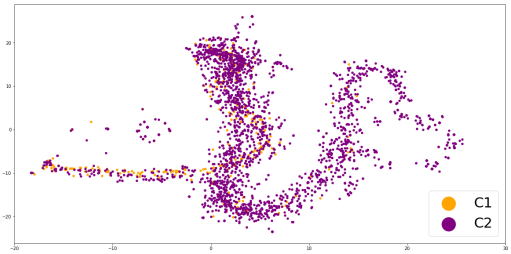
Tab. 4.1: Comparación de puntajes ROC AUC

Método	L1	L2	Coseno	Mahalanobis	Referencia
FastText	0.987	0.987	**0.996**	0.991	-
BERT	0.942	0.947	0.942	**0.964**	-
DMC	-	-	-	-	**0.982**
RW	-	-	-	-	**0.924**

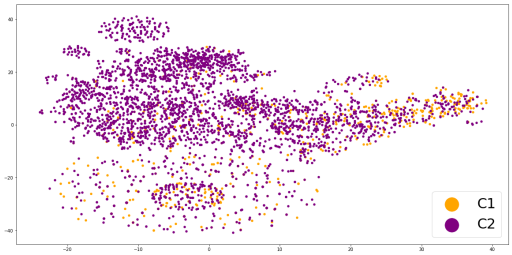
y nuestro trabajo anterior "DMC"3, un enfoque basado en estructura y texto. En la Figura 4.1 mostramos las distribuciones de puntajes de Fasttext y BERT, usando las cuatro diferentes distancias descritas anteriormente, comparadas con las referencias "RW" y "DMC". Lo graficamos mediante beanplots con puntajes de conjuntos de datos controvertidos en el lado izquierdo y los no controvertidos en el lado derecho. Nota que, dado que por definición el enfoque "DMC" da puntajes más altos para conjuntos de datos controvertidos y puntajes más bajos para los no controvertidos, las dos distribuciones están invertidas.

Cuanto menos se superponen las dos distribuciones, mejor funciona el pipeline. Por lo tanto, para cuantificar el rendimiento de los diferentes enfoques, calculamos el ROC AUC. Por definición, este valor está entre 0 y 1, donde 0,5 significa que las curvas están perfectamente superpuestas (es decir, puntuación aleatoria), mientras que los valores de 0 y 1 corresponden a distribuciones perfectamente separadas. La comparación entre las diferentes medidas de distancia se informa en la Tabla 4.1. Como podemos ver, el mejor puntaje (el valor más alto) es obtenido por el modelo Fasttext con distancia coseno, superando a los métodos de vanguardia [83, 162].

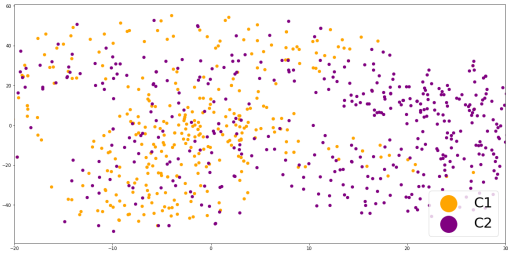
Aunque BERT ha alcanzado muchos resultados de vanguardia en diferentes tareas de PLN [59], FastText se adapta mejor a nuestro pipeline. Analizando los casos puntuados erróneamente observamos que BERT falla principalmente con los conjuntos de datos no controvertidos, por ejemplo el conjunto de datos *Feliz Natal* (puntaje de controversia 0,51). Nuestra hipótesis es que, dado que BERT es un modelo más grande y complejo que FastText, a veces sobreajusta los datos. BERT es capaz de separar las dos formas de hablar de las comunidades incluso cuando son muy similares, no opuestas en una controversia, explotando diferencias que no somos capaces de percibir. Para verificar cualitativamente este comportamiento, representamos los embeddings producidos por cada técnica reduciendo su dimensión a 2 con el algoritmo t-SNE [200] para fines de visualización.



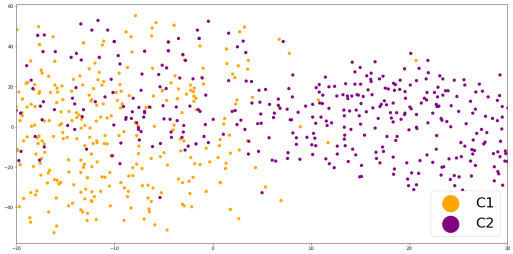
(a) Embeddings de Fasttext del conjunto de datos *Kingjacksonday*



(b) Embeddings de BERT del conjunto de datos *Kingjacksonday*



(c) Embeddings de Fasttext del conjunto de datos *Feliz Natal*



(d) Embeddings de BERT del conjunto de datos *Feliz Natal*

Fig. 4.2: Embeddings reducidos por t-SNE producidas por Fasttex y Bert

En la figura 4.2 mostramos los embeddings reducidos obtenidos por cada método para dos conjuntos de datos no controvertidos, *Cumpleaños de Jackson* y *Feliz Natal*. El primer conjunto de datos es correctamente predicho como no controvertido por ambos métodos y podemos ver que sus embeddings están altamente mezclados, como se esperaba. Sin embargo, las embeddings de *Feliz Natal* están mezcladas cuando se utiliza Fasttext, mientras que BERT aún puede dividirlas en dos grupos separados. Esto muestra que, para el caso de *Feliz Natal*, BERT todavía está diferenciando dos formas de hablar.

Tiempo Computacional

La Figura 4.3 muestra los diagramas de caja sobre los 30 conjuntos de datos de los tiempos computacionales totales (en segundos) de nuestros dos mejores algoritmos, desde el inicio (etapa de construcción del gráfico) hasta el final (etapa de cálculo del índice de controversia), en comparación con los valores de referencia. Nuestros enfoques son más rápidos que el método basado en gráficos de referencia (RW), mientras que el enfoque DMC es sólo más rápido que nuestra variante BERT. El enfoque Fastext supera a ambos valores de referencia, permitiendo un análisis más rápido cuando se utiliza desde una perspectiva en tiempo real, ya que podría ser necesaria una intervención para prevenir comportamientos maliciosos, ya descritos en la Sección 2.

4.3. Conclusiones

En este capítulo diseñamos un proceso basado en PNL (Procesamiento de Lenguaje Natural) para medir la controversia. Probamos algunas variantes, como dos técnicas de vectorización (utilizando los modelos de lenguaje Fasttext y BERT) y cuatro medidas de distancia. Aplicamos estos enfoques en 30 conjuntos de datos heterogéneos de Twitter, y

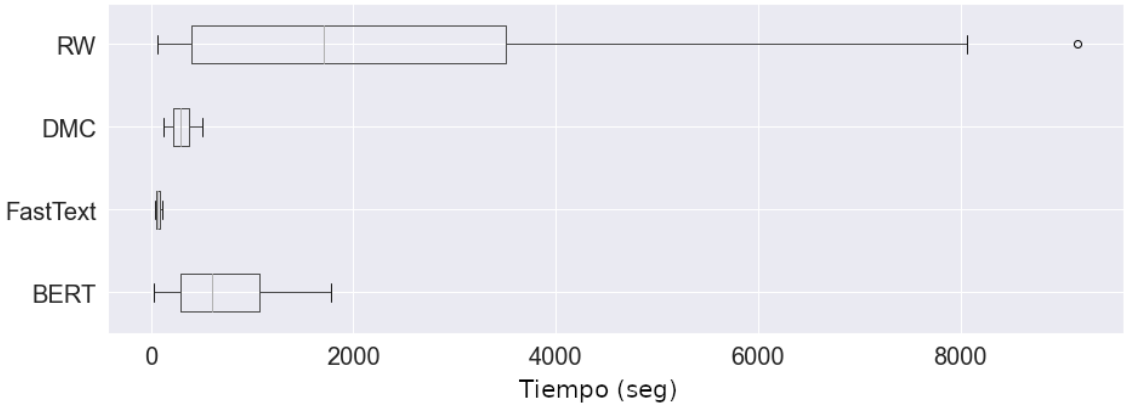


Fig. 4.3: Comparación del tiempo computacional

comparamos los resultados. Nuestro mejor enfoque, utilizando FastText y la distancia del coseno, supera no solo al método basado en gráficos de última generación [83], donde los autores afirman que las técnicas basadas en contenido no funcionan tan bien como las basadas en estructura, sino también al capítulo anterior 3 publicado en [162], en términos de puntuación ROC AUC y velocidad, debido a la menor dependencia de la estructura del gráfico y la inserción de una contribución semántica.

Nuestro proceso involucra a FastText, un modelo rápido para codificar oraciones, o BERT, un modelo de lenguaje más preciso, más lento debido al complejo proceso de ajuste fino requerido. Fasttext obtiene el mejor rendimiento en general, alcanzando una puntuación ROC AUC de 0,996. Como informamos en la sección anterior, esto es probablemente porque BERT es tan potente que podría diferenciar formas de hablar incluso cuando no están en controversia. Debido a la naturaleza de nuestro proceso, Fasttext funciona mejor y también tiene un tiempo de cálculo mucho más rápido. Estos resultados abren una nueva perspectiva para el análisis de redes sociales que puede ayudar a las personas a participar en discusiones más saludables, ya que estos enfoques nos permiten detectar de manera más rápida y eficiente los diferentes puntos de vista.

Al igual que el enfoque del capítulo anterior y lo publicado en [162], esta técnica también tiene algunas limitaciones: *Evaluación*, dificultades para establecer la verdad básica, *Controversias multisectoriales*, controversia con más de dos lados, *Elección de datos*, recolección manual de temas y *Sobreajuste*, pequeño conjunto de experimentos, aunque ahora tenemos 10 discusiones más, aún no es lo suficientemente grande desde un punto de vista estadístico.

Nuestro enfoque basado en lenguaje tiene otras limitaciones. En primer lugar, entrenar un modelo de PNL que pueda tener un buen rendimiento requiere una cantidad significativa de texto, por lo tanto, nuestro método funciona solo para discusiones “grandes” en términos de volumen. Sin embargo, las controversias más interesantes son aquellas que tienen consecuencias a nivel de sociedad, generalmente lo suficientemente grandes para nuestro método. Respecto a q En segundo lugar, nuestros hallazgos se basan en conjuntos de datos provenientes únicamente de Twitter. Al igual que en el capítulo anterior, esto se debe a la disponibilidad de datos que nos otorgaba esta red (desde que es X esto ya no es tan sencillo) y a que es una de las principales plataformas elegidas por los usuarios a la hora de debatir. De todas formas, este método también puede ser aplicado a otras fuentes de datos (como lo haremos en el capítulo 7), ya que sólo necesita las comunidades de cada

usuario y los respectivos textos de sus intervenciones en la discusión.

Otra posible línea de investigación en base a esta nueva técnica podría involucrar análisis relacionados con el usuario, como la detección de usuarios que están en la "frontera semántica", en casos controvertidos, y cómo se comportan con el tiempo. Esto podría ser útil para encontrar si hay actores que pueden ayudar a prevenir la polarización. También analizaremos qué usuarios se sitúan en lados semánticos opuestos para detectar rápidamente las principales diferencias entre dos comunidades.

Finalmente, también detectaremos y analizaremos los comportamientos de los usuarios que realizan intervenciones mixtas en un debate polarizado, por ejemplo, publicando opiniones de ambos lados de la controversia.

Parte III

SOCIEDADES DIVIDIDAS

5. COMPARACIÓN DE LA POLARIZACIÓN ENTRE ARGENTINA Y ESTADOS UNIDOS

5.1. Introducción

¿Cómo entender la polarización en América Latina? ¿Cuáles son las particularidades en relación al caso norteamericano, que se usa como referencia en los estudios? ¿Debemos pensarla como una reacción a los avances en términos de derechos e igualdad en el período posneoliberal y/o se asienta en tendencias políticas de más larga data? Esta tesis sostiene que es necesario adaptar los conceptos y perfeccionar los métodos para captar las particularidades de un fenómeno que existe y contribuye a la erosión democrática pero que tiene rasgos propios en cada país en particular y en la región en general. Específicamente, estudiamos las dinámicas de la polarización en la esfera digital en Argentina, con el objeto de contribuir a definir este proceso en nuestro subcontinente. Para tal fin desarrollamos los métodos de la parte part II de esta tesis y los utilizamos en este capítulo para medir las dinámicas de polarización que se usan en este trabajo.

Tal como se dijo, todavía no sabemos suficiente de las particularidades de la polarización en la región, ya que la literatura del tema proviene sobre todo de Estados Unidos[154], donde, entre otras diferencias centrales, hay dos identidades políticas estables (Demócratas y Republicanos) en torno de las cuales se configura el mapa de la polarización. En tal sentido, sin identidades tan fijas, es muy probable que en nuestros países la polarización sea más dinámica y tenga otros actores. En efecto, en el Proyecto Polder [1] (Polarización y Derechos) en curso con Gabriel Vommaro sobre distintos países de la región notamos la necesidad de diferenciar entre polarización entre los medios, en las distintas plataformas digitales, las dirigencias políticas y la sociedad. Efectivamente, la dinámica de polarización en cada uno de ellos, en los distintos países y en distintos momentos adquiere configuraciones distintas. Así las cosas, hay temáticas que polarizan, hay coyunturas y hay estrategias que pueden mostrar momentos de polarización para una parte de la sociedad pero ser indiferente para el resto. Esto nos lleva a la necesidad de afinar nuestra comprensión de la polarización, para lo cual es preciso emplear metodologías innovadoras, como las utilizadas en este capítulo. De esta manera, contribuimos a comprender las singularidades de la polarización a nivel local.

¿Qué concluyen los estudios realizados en Estados Unidos que marcan la agenda sobre polarización y por ende son un contrapunto necesario a la hora de pensar las particularidades de los casos locales?. En dicho país hay consenso en sostener un aumento de la polarización política en las últimas dos décadas [64, 77]. Las y los especialistas han debatido si la polarización es un fenómeno sólo de las élites [78] o también del público [1]. Pero lo cierto es que a la mayoría de los norteamericanos de manera creciente les disgustan o desconfían de aquellos identificados con el partido opuesto (Demócratas vs Republicanos), aún si sus posiciones sobre los diversos tópicos no distan tanto [140]. Una serie de factores se han conjugado para exacerbar la proclividad de los identificados con cada partido para dividir el mundo en un valorado "in-group" y un despreciado "ex-group" [102]. Uno de dicho factores es que la proliferación de medios hiperpartisanos ha reforzado identidades políticas y consecuentemente sentimientos exacerbados hacia los partidos políticos. Así las

[1] https://polarizacion.net/

cosas, cuando hay polarización afectiva es más probable evaluar a los miembros del outgroup como más radicalizados e ideológicamente distantes de lo que realmente son [129]. En esta misma dirección, Mason distingue entre la polarización de identidades políticas de la polarización en tópicos. Para la autora, el peso de las identidades políticas lleva a los individuos a adherir a partidos o líderes cuyos programas tienen ideas más extremas que las propias.

¿Qué ha sucedido entretanto en la Argentina? Una investigación reciente con datos del Barómetro de las Américas muestra un aumento de la polarización en Argentina comparando las elecciones presidenciales de 2015 y 2019 [133]. En relación a sus características diferenciales, nuestro trabajo centrado en Twitter encuentra tres rasgos importantes: la conformación de comunidades dentro de los polos, los desplazamientos entre comunidades y de un polo a otro, y la existencia de ciertos tópicos de discusión que se correlacionan con tales desplazamientos. Esto nos permite estar atentos a (al menos) tres hechos: que las identidades no son tan fijas, sino que como veremos comparado a USA hay mayor posibilidad de cambios, cuáles son los temas que llevan a los cambios y finalmente la existencia de acuerdos primarios y diferendos secundarios entre las comunidades que conforman tales polos. Esto gravitará en la existencia de debates internos, así como en la potencialidad que desacuerdos crecientes entre comunidades de un polo lleven a rupturas. Si nuestro argumento es cierto, hay una noticia optimistas y otra pesimista para las democracias de la región: la optimistas es que la posibilidad de cambios es mayor, lo que cuestiona la idea de la polarización como un rasgo estable e inamovible, más bien se trataría de una condición potencial que se motoriza en ciertas coyunturas y en otras no. A modo de ejemplo, el análisis que presentamos fue realizado durante el período de la pandemia Covid19. Se evidenciará que la división persiste en temas de política y manejo de la pandemia. No obstante, es importante señalar que no surge un polo abiertamente contrario a las medidas sanitarias ni un sector que niegue la existencia de la pandemia, como sucede en países con mayor polarización, como Estados Unidos y Brasil. Lo inquietante es que también introduce una mayor incertidumbre sobre lo que podría transformarse en el acuerdo primario en una coyuntura, relegando otros temas. Tal como presumiblemente ocurrió en Brasil, donde el voto de enojo anti-PT desempeñó un papel crucial en la consolidación temporal del respaldo a Bolsonaro por parte de electores que no necesariamente comparten todos los principios de extrema derecha de dicho mandatario.

Este capítulo esta basado en el trabajo desarrollado junto a Federico Albanese (Universidad de Buenos Aires), Gabriel Kessler (Universidad Nacional de San Martín) y Esteban Feuerstein(Universidad de Buenos Aires) y se encuentra publicado en la Revista Anfibia [2].

5.2. Datos y metodología

Cómo muestra de la discusión política argentina, descargamos millones de tweets en español que mencionaban a la cuenta @alferdez (cuenta oficial del presidente argentino) del 9 de mayo al 9 de junio de 2020. Para comparar con la coyuntura estadounidense, descargamos tweets en inglés que hacían mención a la cuenta @realdonaldtrump (cuenta oficial del presidente estadounidense) en el mismo período. Con la intención de observar las modificaciones en la polarización a través del tiempo, en ambos casos tomamos 2 “fotos”: la primera (a la que denominamos T0) corresponde a los tweets emitidos entre el 9 y el 16

[2] https://www.revistaanfibia.com/twitter-el-laboratorio-politico/

de Mayo de 2020, mientras que la segunda (llamada T1), comprende el período del 1 al 9 de Junio de 2020. Para cada una de esas "fotos" construimos el grafo de interacciones de la mimsa forma que en 3.1.1.

Dado que la cantidad de usuarios que aparecen es del orden de las decenas o centenas de miles, para visualizarlos de manera efectiva es necesaria alguna herramienta. Con ese fin utilizamos el ya mencionado 2.3 layout ForceAtlas2 [105].

Por último, nos propusimos identificar a las principales comunidades en cada foto (entendiendo como comunidades a grupos de usuarios que tienen una alta interacción entre sí y poca con usuarios de otras comunidades). Para esto utilizamos el algoritmo Louvain [51], al igual que en la sección II].

Es importante notar que el algoritmo utilizado detecta, en general, múltiples comunidades, la mayoría con muy pocos individuos cada una. Por ello nos enfocaremos en las "comunidades principales", es decir las más grandes en cantidad de usuarios, siempre que la sumatoria de esas cantidades alcance, como mínimo, el 70 % de los usuarios.Además de descubrir cómo se agrupan los usuarios, resulta de interés saber sobre qué temas o tópicos discuten. En particular, buscamos conocer de qué se habla principalmente en cada comunidad y qué temas les interesan a los individuos que cambian de comunidad [3]. A continuación presentamos los principales hallazgos, la comparación con Estados Unidos y las consecuencias para las políticas públicas y la comprensión de la polarización en Argentina, así como posiblemente en otros países de la región.

5.3. Los polos se componen de comunidades

El gráfico 5.1 nos muestra dos polos claramente distanciados, la famosa grieta y según nuestro algoritmo el 70 % de los usuarios se agrupa en cuatro comunidades principales, de las cuales dos se encuentran en lo que podríamos llamar el polo oficialista y dos en el polo opositor. En la siguiente imagen podemos observar el resultado de aplicar Forceatlas2 sobre la foto T0 (9 al 16 de mayo de 2020) de la discusión Argentina.

[3] Para ello realizamos una reducción dimensional del espacio de palabras usando lo que se conoce como Factorización no negativa de matrices. Este algoritmo consiste en describir al corpus como una matriz (en donde cada columna representa a un término, cada fila al texto de un tweet y cada celda la frecuencia de dicho término en el tweet) y luego achicar esa matriz buscando combinaciones de columnas que describen el espacio con una menor dimensión pero preservando la mayor cantidad de información posible.

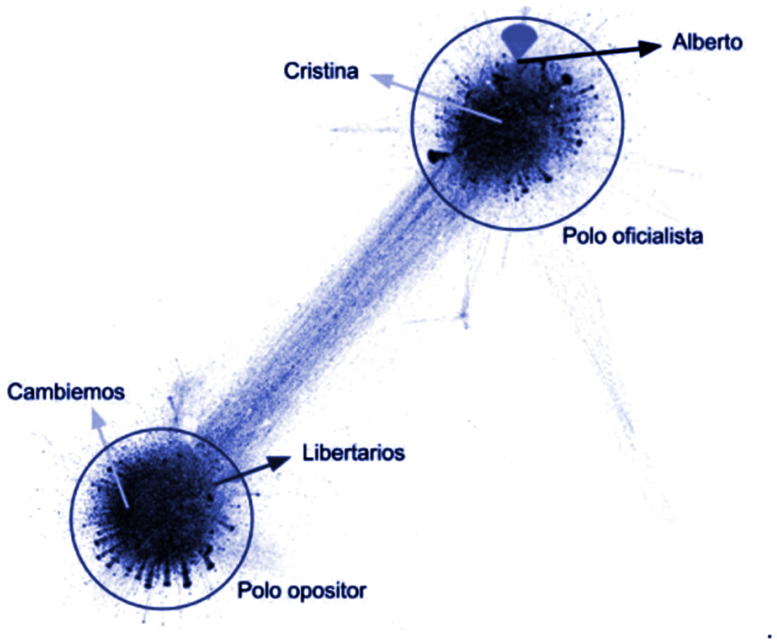


Fig. 5.1: Discusión en Twttier sobre @alferdez

En las redes sociales en general, y en Twitter en particular, existen usuarios que se destacan por tener una cantidad de seguidores o "followers" mayor que la media. Estos usuarios, a los que se llama "autoridades" suelen ser personas o entidades conocidos o notables fuera de Twitter, en el "mundo real" y nos brinda claves de cómo se configura el discurso político en una sociedad. Se trata en su mayoría de políticos, miembros de la farándula o medios de comunicación masivos. En las tablas 5.1,5.3,5.2,5.4 listamos las cuatro comunidades principales, con los usuarios más conocidos de cada una.

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@alferdez	Alberto Fernandez	M	61	Presidente de la nación
@paulitachaves	Paula Chavez	F	35	Conductora de TV
@telefenoticias	Telefe Noticias	N/A	N/A	Programa de noticias de TV
@elchatoprada	Pablo Prada	M	55	Productor de TV
@danielscioli	Daniel Scioli	M	63	Excandidato a presidente de la nación
@minsaurralde	Martín Insaurralde	M	49	Intendente de Lomas de Zamora
@MatiasLammens	Matías Lammens	M	40	Ministro de Turismo y Deportes de la Nación y excandidato a Jefe de Gobierno de la Ciudad Autónoma de Buenos Aires

Tab. 5.1: Autoridades de la comunidad .ªlbertoçompuesta de 9.461 usuarios

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@CFKArgentina	Cristina Kirchner	F	67	Ex-presidenta de la nación y actual vice presidenta de la nación
@Kicillofok	Axel Kicillof	M	48	Gobernador de la provincia de Buenos Aires y ex-ministro de economía de Cristina Kirchner
@rialjorge	Jorge Rial	M	58	Conductor de TV
@C5N	C5N	N/A	N/A	Canal de Noticias
@FernandezAnibal	Aníbal Fernandez	M	63	Ex-jefe de gabinete de Cristina Kirchner
@titifernandez1	Miguel Angel Fernandez	M	68	Periodista deportivo
@eldestape_radio	El Destape Radio	N/A	N/A	Radio de noticias
@VHMok	Víctor Hugo Morales	M	72	Periodista político

Tab. 5.2: Autoridades de la comunidad Çristinaçompuesta de 19.051 usuarios

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@FerIglesias	Fernando Iglesias	M	63	Diputado nacional de Cambiemos
@herlombardi	Hernan Lombardi	M	60	Ex-ministro de cultura de Cambiemos
@gracielaocana	Graciela Ocaña	F	59	Diputada nacional de Cambiemos
@infobae	Infobae	N/A	N/A	Portal de noticias
@connieansaldi	Constanza Ansaldi	F	46	Conductora de TV
@perfilcom	Perfil	N/A	N/A	Diario de noticias
@majulluis	Luis Majul	M	59	Periodista político
@psirven	Pablo Sirven	M	63	Periodista político

Tab. 5.3: Autoridades de la comunidad Çambiemosçompuesta de 19.850 usuarios

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@CarlosMaslaton	Carlos Maslaton	M	61	Abogado libertario
@jlespert	Jose Luis Espert	M	58	Economista libertario y excandidato a presidente de la nación
@lilianafranco20	Liliana Franco	F	64	Periodista especializada en economía
@luisrosalesARG	Luis Rosales	M	55	Excandidato a vicepresidente de Jose Luis Espert

Tab. 5.4: Autoridades de la comunidad "Libertariosçompuesta de 2.388 usuarios

¿Quiénes son las autoridades de cada comunidad y qué nos dice esto sobre la política en Twitter? Las 5 comunidades son bastante similares: compuestas en su mayoría por políticos en actividad, los oficialistas con cargos públicos y periodistas, son en su mayoría hombres (sobre 27 hay sólo 5 mujeres) cuyo promedio de edad es superior a los 40 años. Es decir, hay una conjunción entre autoridades de sus espacios políticos y periodistas afines. Cada una de las personas de estas comunidades habla principalmente con otros usuarios de su misma comunidad y muy poco con los usuarios de otras comunidades, fomentando la denominada "cámara de eco" al menos en cuanto a la existencia de agendas comunitarias. No obstante, como muestran investigaciones recientes, esto no significa que no se vaya a mirar lo que comentan otras comunidades, pero sin establecer diálogo con ellas y, por supuesto, sin compartirlo. Ahora bien, ¿de qué habla principalmente cada comunidad en el epicentro de la epidemia de Covid19 en Argentina y América Latina? La comunidad de Cristina se centra en la deuda externa y el estado de la macroeconomía. En cuanto a la de Alberto, se focaliza en temas de Salud, el estado de los hospitales y en pedirle a la sociedad que "cuide lo conseguido" haciendo referencia al mejor desempeño sanitario de la Argentina en la pandemia, al compararla con otros países de la región. Por su parte, los activistas de la comunidad de Cambiemos reclaman poder trabajar, haciendo foco en que las pymes (Pequeñas Y Medianas Empresas) están teniendo una baja facturación, y la necesidad de ciertas medidas sanitarias para poder ir a trabajar de forma segura y evitar así la quiebra. Por último, los miembros de la comunidad Libertaria también hablan de

la situación de las pymes y la situación precarizada de los médicos durante las guardias en hospitales, la extensión de sus jornadas laborales y el incumplimiento de los pagos salariales.

En Estados Unidos 5.2 también observamos dos polos pero las comunidades son distintas. Aquí las 3 principales comunidades abarcan al 73 % de los usuarios, lo que a priori denota una mayor cohesión (o menor fragmentación) en la polarización. Las llamamos: Trump, Republicanos y Demócratas y al igual que en el caso anterior estos nombres fueron decididos en base a los usuarios más famosos o influyentes de cada una. Al igual que en Argentina, hay mayoría de varones sobre mujeres, en general a partir de los 40 años pero perfiles un poco má diversos, pero siempre desde una élite: CEO de empresas, recaudadores de partidos, sacerdote anti Maduro, algunos actores y dirigentes de ONGs o think tanks y el Primer Ministro de Israel. A continuación damos un detalle de los mismos:

Cuenta	**Persona/Entidad**	**Sexo**	**Edad**	**Descripción**
`@HillaryClinton`	Hillary Clinton	F	72	Excandidata a presidenta de EEUU por el partido demócrata
`@SenSanders`	Bernie Sanders	M	78	Exprecandidato a presidente de EEUU por el partido demócrata
`@CaslerNoel`	Noel Casler	M	31	Comediante
`@ProjectLincoln`	Project Lincoln	N/A	N/A	Comité de republicanos y ex republicanos contra la reelección de Trump
`@eugenegu`	Eugene Gu	M	57	CEO de Coolquit
`@itsJeffTiedrich`	Jeff Tiedrich	M	63	Militante político anti Trump
`@TheDailyShow`	The Daily Show	N/A	N/A	Noticiero de TV
`@kenolin1`	Ken Olin	M	65	Actor y productor
`@JohnBrennan`	John Brennan	M	64	Ex-Director de la CIA de Obama
`@funder`	Scott Dworkin	M	Desc.	Recaudador de fondos demócrata

Tab. 5.5: Autoridades de la comunidad "Demócrataçompuesta de 100.593 usuarios

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@realDonaldTrump	Donald Trump	M	73	Presidente de EEUU
@POTUS	US President	N/A	N/A	Cuenta del presidente de EEUU
@MbuyiseniNdlozi	Mbuyiseni Ndlozi	M	35	Político liberal de Sudáfrica
@PadreJosePalmar	Jose Palmar	M	58	Sacerdote Católico opositor a Maduro y residente en EEUU
@chuckwoolery	Chuck Woolery	M	79	Conductor de TV y músico
@FoxNews	Fox News	N/A	N/A	Canal de noticias
@Lord_Sugar	Alan Sugar	M	73	Empresario y asesor político
@IngrahamAngle	Laura Ingraham	F	56	Conductora de noticiero en Fox News
@IsraeliPM	PM of Israel	N/A	N/A	Cuenta del primer ministro de Israel

Tab. 5.6: Autoridades de la comunidad "Trumpçompuesta de 87.662 usuarios

Cuenta	Persona/Entidad	Sexo	Edad	Descripción
@TomFitton	Tom Fitton	M	51	Presidente de Judicial Watch y militante de Trump
@WhiteHouse	White House	N/A	N/A	Cuenta de la Casa Blanca
@ScottPresler	Scott Presler	M	32	Militante republicano
@PressSec	Kayleigh McEnany	F	32	Secretaria de prensa de la Casa Blanca
@thebradfordfile	thebradfordfile	N/A	N/A	Sitio de alt-right
@DeAnna4Congress	DeAnna Lorraine	F	36	Política Republicana
@SidneyPowell1	Sidney Powell	F	65	Abogada y escritora
@BurgessOwens	Burgess Owens	M	68	Político Republicano
@BernardKerik	Bernard B. Kerik	M	64	Ex-secretario nacional de seguridad de Bush
@GOPChairwoman	Ronna McDaniel	F	47	Presidenta del comité republicano

Tab. 5.7: Autoridades de la comunidad Republicanosçompuesta de 53.184 usuarios

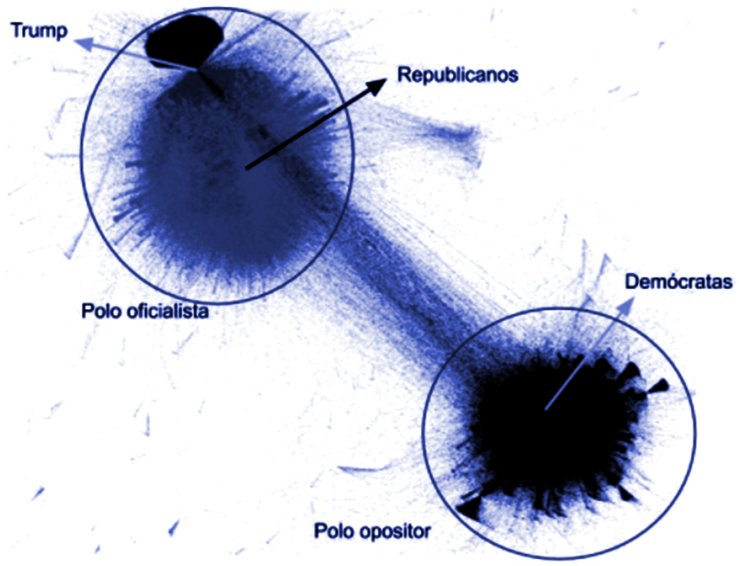


Fig. 5.2: Discusión en Twttier sobre @realDonaldTrump

Es interesante notar que en ambos casos los presidentes tienen sus propias comunidades, las cuales se encuentran muy cercanas (y en el mismo polo) al sector político que lo apoyó y forma parte del gobierno. Esto puede deberse a que el presidente, además de participar en el debate político, tiene gran cantidad de contenido en redes destinado a la gestión presidencial. El mismo comportamiento fue hallado en un trabajo donde analizaron datos de Twitter en 2017 alrededor de la figura de Mauricio Macri [6].

En cuanto a los temas que se discuten en cada comunidad, los demócratas twittean principalmente sobre la presidencia de Donald Trump y sobre el concepto de fake news. Es importante remarcar que esto no quiere decir que compartan fake news, sino que hablan del concepto y en la mayoría de los casos relacionándolo con Trump. Esta relación la hacen en dos sentidos, por un lado acusando a Trump de difundir fake news y por el otro criticándolo por acusar a otros de difundir fake news. En cuanto a los usuarios de la comunidad de Trump, hablan en su mayoría sobre la salida de Estados Unidos de la Organización Mundial de la Salud (WHO por sus siglas en inglés) y, por otro lado, también le agradecen a su presidente por las medidas que está tomando. Finalmente, la comunidad Republicana habla principalmente de las presidencias de Trump y Obama y del "Obamagate" (la acusación a Barack Obama de estar conspirando contra Donald Trump).

¿Qué podemos concluir sobre Twitter y el espacio público hoy? Considerando las autoridades y los temas, parece ser una plataforma donde se configuran los discursos "desde arriba" y se difunden hacia la fracción más politizada de la sociedad. ¿Existen estos mismos

polos y comunidades en la sociedad?, ¿Los tópicos de Twitter son también sus preocupaciones? En otras palabras, la grieta de Twitter existe en la sociedad de un modo más o menos similar. Estudios en curso en distintos países muestran una creciente segmentación de la actividad política según la plataforma: los adultos a partir de 50 años son muy activos en Facebook, los más jóvenes en Instagram y Youtube y esta última plataforma es muy utilizada por la derecha y la ultra derecha en los países centrales. Twitter por su parte parece ser un lugar de mayor configuración de la grieta, tanto por las autoridades que la lideran como por la dinámica de alineamiento intra comunitario y de disputa intercomunitaria. Pero el interrogante que resta es si en cada país hay una línea de continuidad entre la polarización en la esfera digital, en los medios tradicionales, en la dirigencia política y en la sociedad. En ese escenario a multi escala, la polarización en Twitter nos habla sobre todo de la oferta política más que de la demanda. En otras palabras, esto dice menos que eso sea un reflejo de la polarización en la sociedad y que sean esos temas los que de manera menos visible preocupen, dividan o cohesionan a la sociedad. En todo caso, puede ser así en ciertas coyunturas electorales, pero eso no significa que esto sea asi en la vida cotidiana.

¿Qué cambios se evidencian un mes después, en la segunda captura de datos? Como es de esperar, se mantienen en ambos países dos polos bien diferenciados y las mismas comunidades conformándose. Sin embargo, los usuarios que las componen no son exactamente los mismos, sino que algunos cambian de comunidad, produciéndose desplazamiento como veremos a continuación.

5.4. Hay polarización pero con desplazamientos intrapolares e interpolares

¿Por qué alguien se mueve de una comunidad a otra? Para acercanos a una respuesta aplicamos el método desarrollado en [8] que nos permite estimar qué usuarios cambiarán de comunidad política. Esta técnica consiste en entrenar un modelo de aprendizaje automático mediante XGBoost [49] sobre distintos atributos tales como: tópicos discutidos (detectado mediante NMF[209]) por cada usuarios y métricas correspondientes al grafo de retweets (por ejemplo, grado, *PageRank*, *Betweenness Centrality* y otros). El objetivo del modelo es predecir en base a estos atributos que usuarios cambiarán de comunidad.

Una vez entrenado el modelo y corroborada su eficacia en la predicción, buscamos identificar los temas que les importan a los usuarios que se desplazan. Para esto aplicamos el mismo enfoque que en [8], un análisis de importancia de los atributos mediante una permutación aleatoria de los valores de los mismos. Cabe diferenciar un tema que es ampliamente hablado en twitter (los llamados *trending topics*) de un tema hablado ampliamente por los usuarios que cambian de comunidad, ya que éstos son una pequeña minoría y no necesariamente sus temas son los más tratados en general.

En el caso de Argentina los tópicos son los “Créditos hipotecarios”, las “guardias médicas” y “salud y hospitales”, temas de debate coyuntural en plena pandemia. En efecto, “creditos hipotecarios” hace referencia a los créditos hipotecarios llamados UVA. En particular, en el período estudiado los bancos incumplieron las medidas del Banco Central de la Republica Argentina (BCRA) que permiten a los deudores trasladar cuotas impagas hasta el final del crédito. En cuanto al tópico “Salud”, se refiere a la dicotomía “salud” vs “economía” en tiempos de pandemia, mencionando la cantidad de muertes por COVID-19 en Argentina y otros países y estableciendo un debate sobre si las medidas de aislamiento no estaban profundizando la recesión, un tema central en el discurso de la oposición al

gobierno. Es interesante notar cómo dichos tópicos están más ligados a la vida cotidiana, afectan a la vida "real" y el temas de las guardias hospitalarias y de los créditos han sido menos recuperado por el debate político entre gobierno y oposición que los otros tópicos discutidos por comunidades. Dicho en otras palabras, los temas menos polarizados pero que importan, son los que generan desplazamientos.

En la siguiente figura 5.3 se observa una representación gráfica de lo antes descripto, donde cada burbuja representa a una comunidad y con una nube de palabras que describe de qué tópicos habla principalmente. Las flechas muestran los desplazamientos de usuarios entre mayo y junio de 2020. Los porcentajes de los mismos indican la cantidad de usuarios respecto al total de la comunidad. Además, dentro de la fecha remarcamos el tópico de mayor interés para dichos individuos que migran.

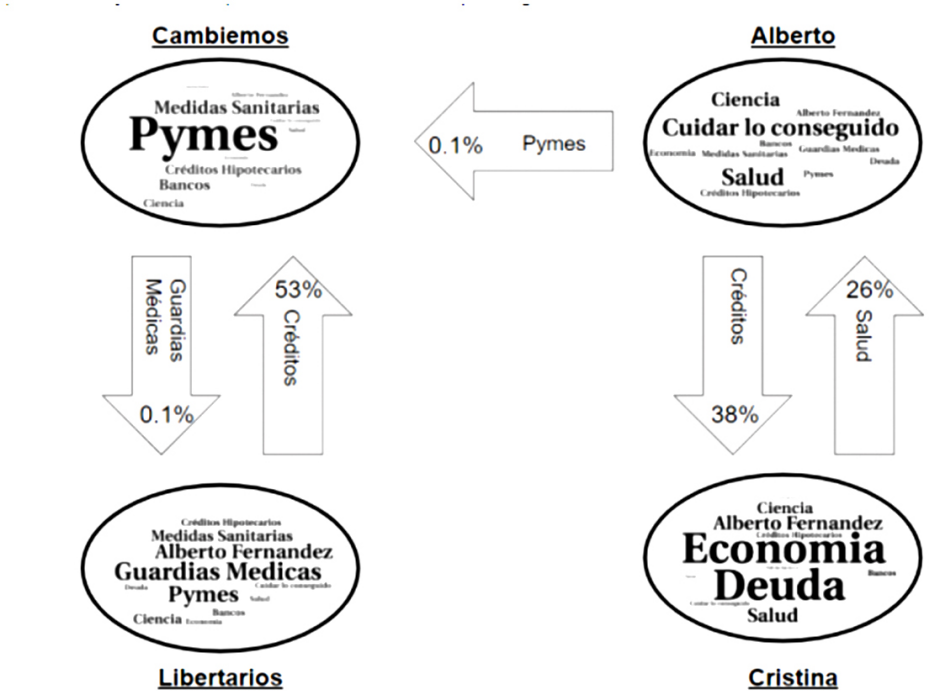


Fig. 5.3: Cambios de comunidad según tema en la discusión sobre @alferdez

Como dijimos, la polarización es dinámica y las comunidades pueden cohesionarse o fragmentarse en momentos electorales o coyunturas críticas. Al analizar los tweets correspondientes a las elecciones parlamentarias en Argentina del 20171 (los tweets de una semana antes de las PASO y los tweets de una semana antes de las elecciones generales), pudimos ver que el tema "Santiago Maldonado" era usado por los usuarios de las comunidades de 1Pais (Partido encabezado por Sergio Massa, líder de la oposición al Kirchnerismo en 2013, presidente de la cámara de diputados por el Frente de Todos en 2020 y ministro de economía 2022-2023) y PJ (Lista encabezada por el ex ministro de interior y trans-

porte, Florencio Randazzo, de la gestión de Cristina Kirchner) que migraron a Unidad Ciudadana (Cristina Kirchner, ex presidenta y vicepresidenta), mientras que el tópico sobre la "venezuelización" de Argentina, era usado por los individuos de 1Pais y PJ que migraron a la comunidad de Cambiemos (Partido político liderado por Mauricio Macri). También se observó cómo la comunidad de Randazzo se desarmó entre las elecciones y sus miembros pasaron a otras comunidades más grandes (principalmente hacia 1Pais y Unidad Ciudadana).

Por otro lado, cuando hicimos un análisis equivalente para las elecciones presidenciales de Argentina del 2019, encontramos una dinámica similar con los partidos que sacaron menos de 3 % de los votos: el Frente de Izquierda, el Frente NOS (Partido nacionalista conservador liderado por Gomez Centurión, director de aduanas y vicepresidente del Banco Nacion de Argentina durante la gestión de Mauricio Macri) y UNITE (Partido liberal encabezado por el economista José Luis Espert). Los usuarios que en un primer momento estaban en las comunidades correspondientes a dichos partidos, luego migraron hacia comunidades más grandes. En particular, los desplazamientos fueron hacia Frente de Todos (Alberto Fernandez) el caso del Frente de Izquierda y hacia Juntos por el Cambio (Mauricio Macri) para los otros dos. De esta forma, el método planteado no solo permite ver que temas son los de interés para distintos grupos y tipos de individuos, sino también permite ver la cercanía entre distintos partidos políticos, al detectar cómo los usuarios cambian de comunidad política.

¿Qué sucede en Estados Unidos? el tema principal para los usuarios que cambian desde la comunidad republicana a la demócrata es el Obamagate 5.4. Lo que muestra la reacción de cierto sector republicano a las acusaciones dichas por Trump contra el ex Presidente. Por otro lado, el principal tópico para los individuos que migran de la comunidad demócrata hacia la republicana y la de Trump es justamente el tópico que denominamos "thank you", en donde los usuarios le agradecen a Trump por distintas medidas que está tomando.

Por último, se deben destacar las diferencias y similitudes entre Estados Unidos y Argentina. En ambos casos, la cantidad de usuarios que migran de un polo a otro es pequeña. Sin embargo, en Argentina se logran diferenciar distintos subgrupos dentro de un mismo polo. Es decir, una corriente más Kirchnerista, con Cristina, y otra corriente más Albertista. Y dentro de la oposición, una corriente más libertaria y otra de Cambiemos. Esto muestra que si bien la famosa "grieta" está presente en Twitter, hay distintas comunidades con intereses y discursos distintos. En contraposición, el método de detección de comunidades encuentra una sola comunidad principal opositora en Estados Unidos: la demócrata. En conclusión, en ambos casos encontramos una polarización marcada (tal como se muestra en los gráficos de la red de tweets), pero en Argentina observamos que sigue habiendo una pluralidad de discursos y diversidad de temas que está menos presente en el bipartidismo estadounidense.

5.5. Mayor polarización en EEUU vs Argentina

Si bien ambos países muestran altos niveles de polarización, el fenómeno es de mayor magnitud en EEUU, al menos en la imagen de la sociedad proyectada en Twitter. El método desarrollado en el capítulo 4 de esta Tesis, permite medir mediante técnicas de procesamiento del lenguaje natural lo que podríamos denominar la "polarización semántica" de un grupo de usuarios. A través de un índice que analiza el lenguaje o jerga que utilizan las comunidades detectadas en ese conjunto de tweets cuantificamos el nivel de

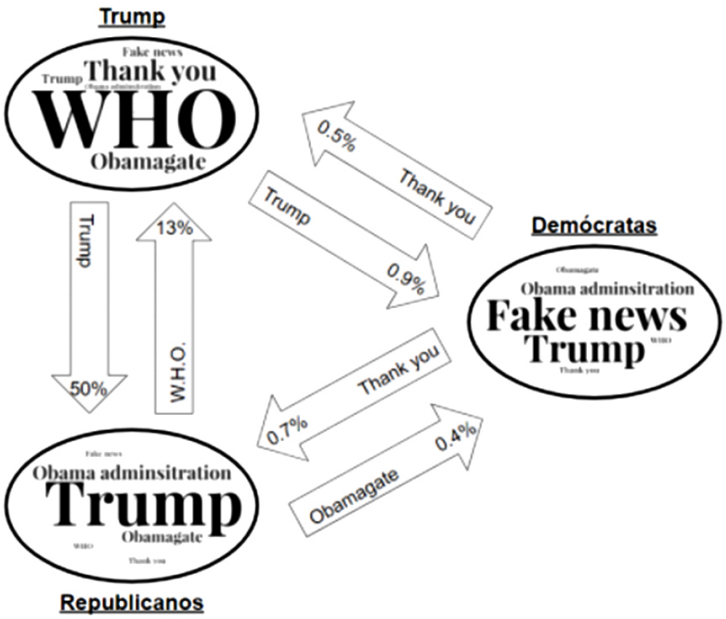


Fig. 5.4: Cambios de comunidad según tema en la discusión sobre @realDonaldTrump

polarización de la discusión. El índice toma valores entre 1 y 0, un mayor valor del índice denota un mayor nivel de polarización. Aplicamos nuestra medida día por día en ambos países en el período del 8 de mayo al 23 de junio de 2020, obteniendo un promedio de 0.85 en EEUU y de 0.74 en Argentina.

¿Qué sucede si se compara con otros países y coyunturas? Para ello analizamos las polarizaciones en distintos contextos: entre ellos el Impeachment a Trump[4], el debate electoral en Reino Unido en Noviembre de 2019[5], la nominación de Kavanaugh a la corte suprema estadounidens[6], las elecciones presidenciales en Brasil en 2018 y las discusiones en torno a Mauricio Macri también durante 2018. Si bien todos estos análisis nos dieron métricas que denotan una fuerte polarización, los casos estadounidenses también resultaron los mas polarizados con un score promedio de 0.80, seguidos por Brasil con un 0.76, luego UK con 0.71 y finalmente Argentina con 0.69. Cabe destacar que si bien podemos decir que todas son discusiones controversiales (según lo desarrollado en el capítulo 4.2) es necesario realizar un análisis de significatividad de estas diferencias para poder afirmar mediante estos únicos datos que un caso es más controversial que el otro.

En resumen esto indica un mayor antagonismo semántico entre ambos polos en el caso

[4] `https://www.nytimes.com/2019/09/24/us/politics/democrats-impeachment-trump.html`
[5] `https://www.bbc.com/news/election-2019-50268753`
[6] `https://www.nytimes.com/2018/10/06/us/politics/brett-kavanaugh-supreme-court.html`

estadounidense a lo largo del tiempo. Es decir que las ideas que expresan en sus tweets (semántica) y la forma en que lo hacen (jerga) son claramente distintas. Adicionalmente, como marcamos en la sección anterior, notamos una mayor inestabilidad en términos de retención de usuarios de cada comunidad y polo. Lo que podría marcar menos antagonismos y una mayor predisposición al intercambio. Por último podemos notar que Brasil también parece tener una mayor polarización mientras que UK una mas similar a la Argentina, aunque estos son indicios para analizar más en profundidad en futuros trabajos.

5.6. ¿Qué nos dice esto de la polarización en Argentina?

Sosteníamos en el comienzo la necesidad de captar las particularidades de la polarización en América Latina y, en nuestro caso, en Argentina. Vimos que existe en Twitter una grieta compuesta por dos polos conformadas por comunidades. Que esta grieta es menor que en Estados Unidos y que hay más comunidades en cada polo que en dicho país. Asimismo, que acorde con la literatura, las identidades son más estables en el Norte, visto que los desplazamientos dentro de las comunidades es menor. También que los desplazamientos se hacen por los tópicos menos polarizados y de algún modo que tocan más de cerca la vida cotidiana, como los créditos o las guardias médicas. En Estados Unidos, por su parte, el nivel de debate político-ideológico es mayor.

¿Qué nos dice Twitter sobre la polarización en ambas sociedades pero también de la polarización política en general? Los famosos, autoridades o influencers de Twitter son políticos y periodistas, de edad media y más, sobre todo varones, en el caso argentino del área metropolitana de Buenos Aires y con mayor diversidad de perfiles en Estados Unidos, pero sin duda elites, como CEO de empresas, actores famosos o recaudadores de partidos. De esta manera, el dispositivo evidencia el entramado discursivo de la política en cada país, donde figuran políticos y periodistas bien posicionados. Si nos guiamos por otras investigaciones que realizamos en Argentina, si bien es una polarización de elites, es cierto que allí se generan los discursos y relatos que luego circulan en las sociedades. En efecto, aun los no polarizados, en lugar de articular un discurso con tópicos distintos, lo que hacían era sobre todo enhebrar tópicos y argumentos de cada polo, con un nivel de intensidad afectiva menor o considerando a la polarización como una forma de pluralismo, por ende, un rasgo positivo o en todo caso, no negativo 7.

¿Qué implica una polarización más dinámica en el caso argentino? Por un lado, debe ser entendida en una tendencia más general de la política en América Latina. Como sostiene J.P. Luna [132] los ciclos políticos, la "luna de miel" con los presidentes electos y las coaliciones partidarias tienden a ser cada vez de menor duración en toda la región. Esta inestabilidad es motivo de un estado de perpetua amenaza a la legitimidad de los gobiernos. Y las redes son parte de ese juego: la hiper exposición y el escrutinio constante por medios como Twitter y otros favorecen esta aceleración de los ciclos políticos. Así las cosas, los diferendos y desplazamientos entre comunidades son parte de esta aceleración y cambios. Quizás sea un rasgo propio de una política con partidos más débiles que los del pasado. Ni una anomalía ni una crisis temporaria, sino una mutación en la lógica de las democracias de la región.

Vemos también que la polarización es un concepto cuando menos polisémico. Pareciera ser un recurso potencialmente disponible, activado desde arriba y a veces también desde abajo ante temas polarizantes o eventos electorales claves. Sin duda, más allá de los discursos en pos de acabar con la grieta, hoy ya no cabe dudas que la polarización

tiene beneficios electorales. El hecho de que sea una polarización más dinámica es quizás un potencial de diferendos y fisuras entre las comunidades de cada polo pero también posibilita una conversación más plural y democrática, más deliberativa, con acuerdos primarios y diferendos secundarios. Lo inquietante es, que si los acuerdos primarios suelen ser polarizantes, pues acrecientan la oposición con el otro polo; eso puede erocionar la democracia en la región, como lo señalamos al comienzo. El caso brasileño nos vuelve una y otra vez: en una coyuntura electoral polarizada, puede traccionar más quien se muestre más intransigente con el contrincante, puesto que si se marca bien la grieta, todos los de su vertiente lo acompañarán, aún con sus diferendos secundarios. Pero también, los desplazamientos muestran plasticidad, reflexividad, capacidad de moverse según las propias convicciones y, como vimos en el caso argentino, con temas ligados a problemas públicos y concretos, como los créditos o la atención hospitalaria. Quizás esto sea el mayor potencial para que la polarización, que en la literatura de redes es el motor del cambio, de ruptura de status quo, sea un dispositivo coyuntural de la dinámica de las democracias más que una amenaza a su continuidad. O quizás sea ambas cosas.

6. IDENTIFICACIÓN DE COMUNIDADES POLARIZADAS EN EL TIEMPO

La identificación de comunidades es una tarea ampliamente estudiada desde diversas disciplinas hace ya largos años. En el famoso estudio de los 70' "An information flow model for conflict and fission in small groups" Zachary [214] analiza las comunidades que se encontraban latentes dentro de un grupo de Karate basándose únicamente en la estructura del grafo generado por sus relaciones.

Hoy en día, la interacción humana se ha trasladado en gran medida a las redes sociales, que facilitan la comunicación virtual y el intercambio de contenido. Con la aparición del SARS-CoV-2 y el aislamiento subsiguiente, la dependencia de las redes sociales aumentó significativamente, convirtiéndolas en un componente central de la vida cotidiana.

Como mencionamos en la parte II de esta tesis las redes sociales, a través de sus algoritmos de segregación, pueden crear "Filter Bubbles" [167] que refuerzan la homofilia [145] , mostrando a los usuarios solo información que coincide con sus intereses y ocultando lo contrario. Esto resulta en un "mundo de confort" para cada usuario, donde solo se encuentran con puntos de vista, opiniones y gustos similares a los suyos. Como consecuencia, estas burbujas podrían aumentar la intolerancia hacia opiniones diferentes y profundizar las divisiones en las relaciones sociales [50], incluso extendiéndose más allá del ámbito virtual.

Como se mencionó en la introducción 1, nuestra hipótesis es que en comunidades entre las que se observan grandes polarizaciones, como en el ámbito de la política, las jergas se vuelven sumamente específicas y distinguibles dentro de cada comunidad, las cuales utilizan distintos términos o frases que son propias de ellas y no suelen ser utilizadas por otras. En estos casos, mediante algoritmos de PLN, se podrían generar modelos que permitan identificar, con una buena probabilidad, la comunidad de pertenencia de un usuario basándonos exclusivamente en la forma de escribir.

Para desarrollar estos modelos necesitamos analizar la jerga que utilizan los usuarios a lo largo del tiempo y abstraernos del vocabulario puntual utilizado durante la duración de un tópico (eventos específicos que se discuten durante un cierto tiempo). Independientemente de la mayor o menor duración de un tópico, estos son siempre temporales y, eventualmente, se diluirán dando lugar a un nuevo tópico de conversación. La jerga, por el contrario, se mantiene más estable a lo largo del tiempo, trascendiendo la aparición de varios tópicos. A pesar de su naturaleza temporal, en algunos casos el tópico puede modificar la jerga y los usuarios de las comunidades podrían adoptar algunos de los términos introducidos por los tópicos que pasarían a formar parte de la jerga de esa comunidad.

En este capítulo, nuestro objetivo es encontrar la mejor estrategia para desarrollar modelos que permitan predecir la pertenencia de un usuario a una comunidad a lo largo del tiempo a través del texto de sus posteos, analizando la jerga propia de las comunidades de manera independiente de los tópicos.

Consideramos que en el área de la política y, más específicamente, en el contexto de las elecciones presidenciales, los cambios de tópico son algo frecuente, lo cual puede deteriorar la capacidad de predicción de los modelos generados. Por ese motivo, desarrollamos distintas estrategias de predicción buscando que la jerga sea lo más robusta posible para que los modelos desarrollados funcionen adecuadamente a lo largo del tiempo.

Para validar nuestra hipótesis, tomamos como casos de estudio las discusiones en Twitter en torno a las elecciones presidenciales del 2019 en Argentina, las cuales se desarrollaron en un contexto de una muy alta polarización entre los partidos políticos de Cambiemos y el Frente de Todos. Podemos ver en estos tweets ejemplos de la jerga utilizada por las 2 principales comunidades en donde el partido del *Frente de Todos* utiliza palabras como "Macrisis" 6.1 y "MacriGato" 6.2 mientras que el partido de *Cambiemos* utiliza palabras como "Kretina" 6.3 y "Alberso" 6.4.

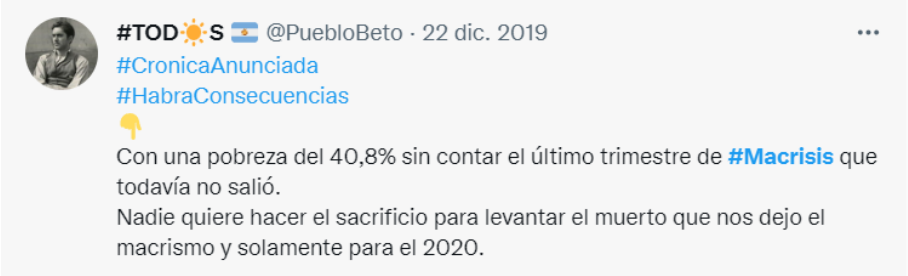


Fig. 6.1: La comunidad del frente de todos utiliza palabras como "Macrisis"

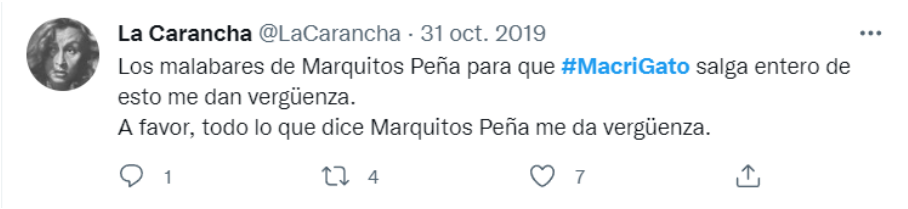


Fig. 6.2: La comunidad del frente de todos utiliza palabras como "MacriGato"

Fig. 6.3: La comunidad de Cambiemos utiliza palabras como "Kretina"

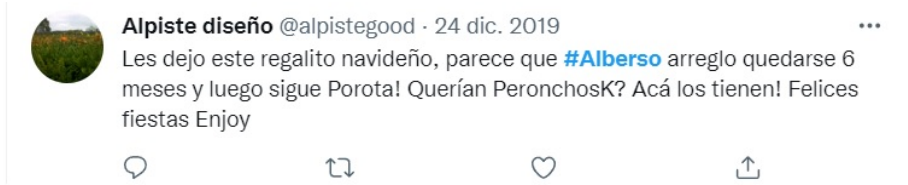


Fig. 6.4: La comunidad de Cambiemos utiliza palabras como "Alberso"

Como mencionamos anteriormente, a lo largo del tiempo ocurren cambios de tópico que pueden alterar el lenguaje por tiempos prolongados o de manera permanente. Entrenamos diversos modelos de PLN mediante Fasttext [111] capaces de predecir la comunidad de pertenencia de los usuarios, utilizando únicamente los tweets publicados entre los meses

de agosto y diciembre del 2019, es decir, desde las PASO[1] hasta la asunción de Alberto Fernandez como Presidente en diciembre. Para expandir lo más posible el análisis, buscamos que el método sea agnóstico del lenguaje y analizamos también la posición de la población brasilera en torno al Presidente Jair Bolsonaro en el contexto de la Pandemia del SARS-CoV-2 durante los meses de junio-julio del 2020.

Utilizamos Walktrap [173] y Louvain [51] como ground-truth a la hora de medir el accuracy de los modelos desarrollados tanto para el caso de Argentina como para el caso de Brasil. Si bien es cierto que ninguno de estos métodos puede ser considerado un ground-truth real, para obtenerlo necesitaríamos, para cada dataset de entrenamiento, preguntarle a cada usuario con qué comunidad se siente identificado, lo cual nos resulta imposible para este trabajo. Por este motivo, para aproximarnos a la verdad, consideramos el resultado de estos métodos como ground truth.

Este trabajo lo dividimos en 2 etapas. En la primera, nuestro objetivo es encontrar la mejor manera para generar buenos modelos predictivos en un intervalo de tiempo determinado, y descubrimos que la mejor forma de entrenar estos modelos es utilizar los datos de 2 días no consecutivos de la semana (por ejemplo, martes y jueves). En la segunda etapa buscamos la mejor estrategia para conseguir que esos modelos se mantengan efectivos a lo largo del tiempo, para lo cual definimos una métrica que nos permite detectar cuándo es necesario reentrenar los modelos. Con esta métrica pudimos obtener un accuracy superior a 0.77.

En la sección 6.2 desarrollaremos las metodologías utilizadas para realizar los experimentos y detallaremos las dos etapas que mencionamos en el párrafo anterior. Ahondaremos también en cada una de esas etapas y en los procesos que se siguieron en cada una de ellas. En el capítulo 6.5 mostraremos los resultados obtenidos tanto para el entrenamiento como para el reentrenamiento de los modelos y detallando lo que observamos en cada una de las estrategias que se probaron. Finalmente, el capítulo 6.6 expone los resultados y las conclusiones de todo lo realizado. Indicamos que para lograr un modelo predictivo eficiente, es clave evitar el uso de datos de días contiguos en el entrenamiento, minimizando la influencia de tópicos breves. Es esencial reentrenar los modelos regularmente para mantener su eficacia, incluso si actualmente predicen bien. Esta práctica asegura que los modelos permanezcan robustos y adaptables a tópicos que persisten por períodos más extensos. Observamos también que, si bien los resultados obtenidos son positivos, aún no hay un método cuya precisión de predicción sea definitiva, con lo cual este trabajo deja abierta la posibilidad de seguir perfeccionando la métrica definida para el reentrenamiento y así seguir trabajando en una mejor estrategia para la predicción de comunidades.

6.1. Trabajos previos

Las comunidades en las redes sociales han sido objeto de estudio de diversos trabajos y se las ha analizado desde diferentes perspectivas. Los estudios que mencionamos a continuación tratan la interacción entre los usuarios en distintas redes sociales y los efectos que estas interacciones pueden tener. Esto permite analizar temas como la polarización o la jerga, entre otros, todos ellos de interés para nuestro estudio y para el análisis que desarrollaremos más adelante. Dividimos esta sección en 2 categorías centrales que nos marcan los principios para realizar este trabajo:

[1] Elecciones Primarias, Abiertas, Simultaneas y Obligatorias en Argentina

- La homofilia en las redes sociales, es decir, la tendencia de las personas a la atracción por sus semejantes, la cual nos permite establecer comunidades en relación a determinado eje e identificar sus jergas (Sec. 6.1.1).

- Estabilidad de las comunidades en el tiempo, ¿Las comunidades se mantienen en el tiempo o tienden a cambiar? (Sec. 6.1.2).

6.1.1. Homofilia en las redes sociales

Aruguete et al. [19] estudian que los usuarios de Twitter comparten exclusivamente los eventos políticos que son congruentes con sus puntos de vista. En este estudio, muestran que las redes sociales crean burbujas con los usuarios que se interesan por los mismos temas y, de esta manera, los exponen a información relacionada con sus mismos puntos de vista. Al compartir eventos, aumenta la frecuencia con la que usuarios de una misma burbuja reciben posteos del mismo tema. Como consecuencia, los tópicos de los posteos que esos usuarios realizan se vuelven "tendencia", alcanzando nuevas audiencias dentro de la misma comunidad.

Bessi et al. [33] analizan el consumo de la información de los usuarios de Facebook en Italia, haciendo foco sobre todo en las teorías conspirativas y en cómo los usuarios que creen en una teoría conspirativa son más propensos a creer en otras y a confiar mucho más en la información brindada dentro de una comunidad de una teoría conspirativa de la que participan. Para esto, crean un grafo de interacciones de los usuarios utilizando los posts de los mismos, siendo un "me gusta" un feedback positivo, el "compartir", la intención de ampliar la visibilidad del post y un comentario, la intención de crear un debate, aunque este último puede tener un refuerzo positivo o negativo respecto al posteo.

Estos dos últimos trabajos refuerzan el concepto de homofilia en las redes sociales, cómo los usuarios se suelen mover dentro de su mismo círculo generando comunidades más cerradas, interactuando mayormente entre usuarios que comparten los mismos intereses. En el trabajo de Bessi, utilizan la red social Facebook para poder crear los gráfos de interacciones, mientras que en nuestro trabajo utilizamos Twitter, red social con la cual podemos hacer una correlación de las interacciones, siendo los "me gusta", "compartir" y "comentar" de Faceebok los "me gusta", "retweet" y "responder" de Twitter. Utilizando esta idea, asumimos que los retweets entre los usuarios son un apoyo a sus puntos de vista, permitiendo relacionarlos dentro de la misma comunidad.

En el trabajo anteriormente presentado en el capítulo 4 analizamos las discusiones en las redes sociales con el objetivo de encontrar una buena forma de cuantificar el nivel de controversia de estas a partir del vocabulario. Entre otras cosas, mostramos que las discusiones en Argentina y en Brasil están altamente polarizadas. Esto lo utilizamos como base de nuestro trabajo para analizar la jerga utilizada en estos dos países.

En dicho capítulo también mostramos como las jergas de las comunidades en contextos de polarización pueden diferenciarse mediante técnicas de PLN. En esta capítulo retomaremos este concepto para el desarrollo de un método de detección de comunidades polarizadas en el tiempo.

6.1.2. Estabilidad de las comunidades en el tiempo

Dado que en este capítulo nos proponemos identificar comunidades a lo largo del tiempo cabe preguntarse si las mismas se mantienen estables en este sentido. En el trabajo de

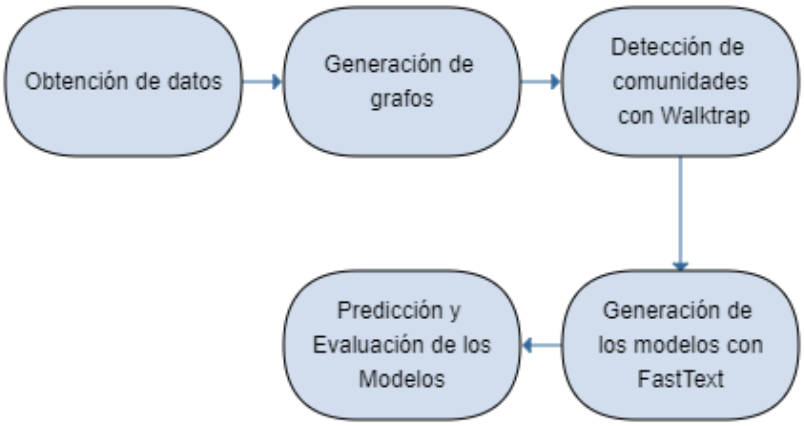


Fig. 6.5: Flujo de pasos para el entrenamiento de los modelos durante la etapa 1 para encontrar la mejor estrategia de entrenar modelos.

Albanese et al. [9] hablan sobre cómo las comunidades se modifican, ya que las personas cambian su opinión y pueden pasar de una comunidad a otra. Para eso, estudian los cambios de las relaciones de los usuarios en Twitter utilizando técnicas de PLN. Presentan un framework de machine learning para clasificar usuarios de redes sociales como "shifting users", es decir, usuarios que pueden llegar a cambiar de opinión a lo largo del tiempo basados en las propiedades topológicas de un grafo y el texto de las discusiones de Twitter. Si bien este trabajo está más enfocado en los usuarios que cambian de opinión a lo largo del tiempo que en la identificación de las comunidades en sí, los resultados demuestran que estos son muy pocos, lo que lleva a la conclusión de que las comunidades son estables y tiene sentido analizar su jerga para poder identificarlas a lo largo del tiempo

6.2. Metodología

La metodología se encuentra dividida en 2 etapas. En la primera etapa (Fig 6.5), que se desarrolla en la Sección 6.3 nuestro objetivo es encontrar la mejor manera para generar buenos modelos predictivos en un intervalo de tiempo determinado y en la segunda (Fig 6.6), buscamos la mejor estrategia para mantener a esos modelos efectivos a lo largo del tiempo.
A su vez, la primera etapa se divide en las siguientes sub etapas

Primero descargamos todos los tweets que utilizaron determinados keywords en las fechas que estudiamos segmentándolos por día. Al separar los datos de esta manera, obtenemos una unidad de tiempo que es lo suficientemente pequeña para identificar la aparición de un nuevo tópico y lo suficientemente grande como para analizar la jerga de nuestro predictor. Una vez que tenemos los datasets, generamos los grafos de interacción entre los usuarios para cada uno de ellos y utilizamos el algoritmo de Walktrap [173] para identificar las principales comunidades. Luego utilizamos los grafos para generar los modelos que vamos a utilizar para predecir la pertenencia de los usuarios a las 2 principales comunidades. Realizamos las predicciones con los modelos y analizamos los resultados.

La segunda etapa 6.6 también se divide en distintas sub etapas: En la primera seleccio-

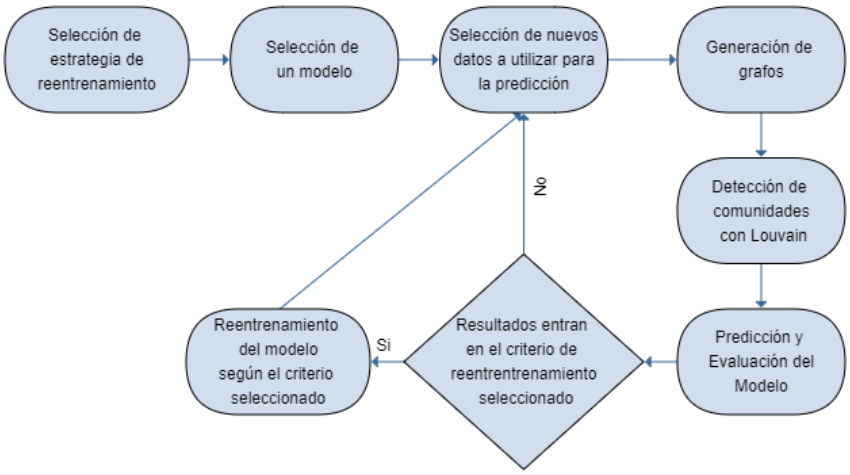


Fig. 6.6: Flujo de pasos para el reentrenamiento diario de los modelos, se repite el proceso hasta que no se encuentren datos nuevos para predecir

namos un criterio para decidir en qué momento se debe realizar un reentrenamiento [2]. En base al criterio elegido, se vuelven a generar grafos de interacciones incorporando nuevos datos para el reentrenamiento. Luego utilizamos los grafos para generar los nuevos modelos y realizamos las predicciones sobre datos nuevos utilizando estos modelos. Repetimos este proceso cada vez que los resultados entren en el criterio seleccionado inicialmente.

6.3. Etapa 1

6.3.1. Obtención de datos

A fin de obtener los datos que vamos a utilizar para generar nuestros modelos, utilizamos la red social Twitter como fuente, dado que en ella, los usuarios expresan libremente sus opiniones y pueden interactuar con otros a través de retweets. Además, Twitter provee distintas APIs[3] para descargar libremente esos tweets. Elegimos como criterio de descarga de los tweets los *keywords* "*Macri*" y "*Bolsonaro*", ya que son importantes figuras políticas y ejes de discusión.

Los *keywords* son un conjunto de palabras relacionadas con un tema que utilizan los usuarios de manera inconsciente al escribir. Los *hashtags* por su parte, los escriben los usuarios con la intención de destacar determinados tópicos, por ejemplo #NoVuelvenMas o #Macrisis. Entonces, si hubiésemos elegido los *hashtags* como criterio de búsqueda no hubiéramos obtenido los tweets que no fueron etiquetados manualmente, perdiéndonos gran parte de la discusiones en la red. Además, dado que los *keywords* no están relacionados a un tópico especifico, es posible identificar mejor a las comunidades por sobre el contexto.

De esta manera, descargamos los tweets en formato de dataframes. Los dataframes son estructuras que permiten almacenar tablas de datos, en las que cada columna puede

[2] Con reentrenamiento nos referimos a ajustar los paramteros ya estimados en los entrenamientos anteriores, no a entrenar el modelo de 0.

[3] https://developer.twitter.com/en/docs/twitter-api

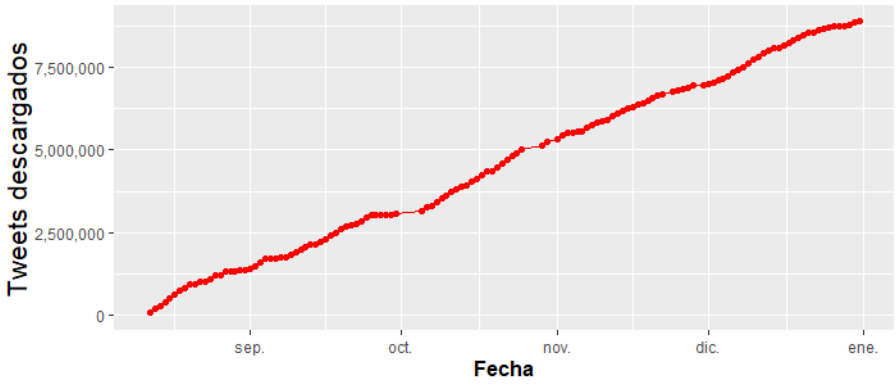


Fig. 6.7: Cantidad de tweets descargados a lo largo del tiempo para el dataset de las PASO

contener un tipo de dato distinto. En la tabla 6.1 podemos ver algunas de las columnas que contienen los datasets obtenidos.

Parámetro	**Definición**
screen_name	Usuario que realizó el tweet
text	Texto del tweet
reply_to_screen_name	Usuario al que le contestó
is_retweet	Si es un retweet o no
retweet_count	Cantidad de retweets que tiene
hashtags	Los hashtags asociados
mentions_screen_name	Menciones a otros usuarios
retweet_text	Texto del retweet

Tab. 6.1: Parámetros obtenidos de la librería de twitter

Una vez descargados los tweets, utilizamos la misma metodología que se aplicó en el trabajo [56] para normalizar los datos, pasando su codificación a ASCII, convirtiendo los emojis a texto con la librería text_clean de R[4], eliminando links, caracteres especiales y textos duplicados.

Para este trabajo descargamos entonces, 2 datasets distintos. El primero, del 12 de agosto del 2019, día posterior a las PASO, al 31 de diciembre del mismo año utilizando como criterio el uso de la palabra *Macri*, dado que como Presidente en este período, es uno de los ejes más importantes de la discusión electoral. Aproximadamente se descargan entre 50.000 y 80.000 tweets por día, teniendo un total aproximado de 9.5 millones de tweets como se ve en la figura 6.7.

En el segundo dataset, se descargaron los tweets de la población brasilera desde el 1 de junio hasta el 30 de julio del 2020 para poder analizar la polarización de la población en torno a *Bolsonaro* en el contexto de la pandemia del SARS-CoV-2. En este caso, se descargaron aproximadamente unos 100.000 tweets por día obteniendo un total aproximado de 6 millones de tweets, como se ve en la figura 6.8.

[4] `https://www.rdocumentation.org/packages/textclean/versions/0.9.3/topics/replace_emoji`

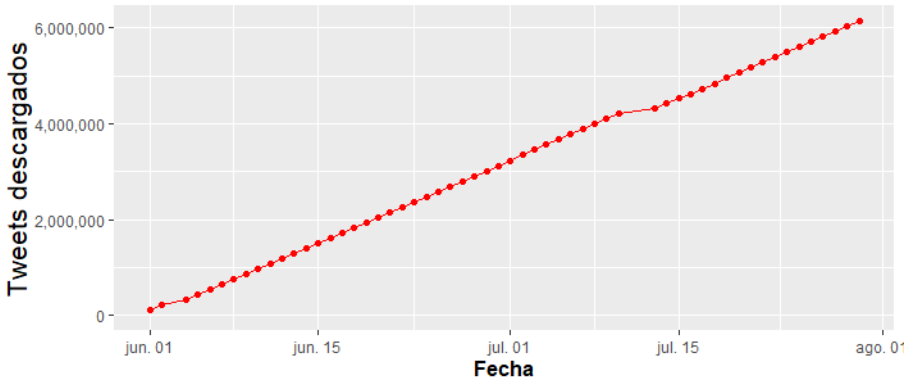


Fig. 6.8: Cantidad de tweets descargados de junio a agosto en brasil 2020 para el dataset de Bolsonaro

6.3.2. Generación del grafo

Para generar el grafo e identificar a las comunidades del mismo seguimos los mismo pasos que lo descripto en el capítulo 3.1 de esta Tesis.

A modo de ejemplo, en el grafo de la figura 6.9 se puede observar la polarización entre las 2 principales comunidades en los dias del 15 y 17 de julio del 2019, cada nodo representa a un usuario, el tamaño de los nodos representa el grado de entrada del nodo, es decir, crece en función de los retweets. Utilizando a Walktrap para identificar a las comunidades, coloreamos con el color celeste a la comunidad del *Frente de Todos* y con el amarillo al partido de *Cambiemos*. Se puede ver cómo Anibal Fernandez (@FernandezAnibal) y Mariano Recalde (@marianorecalde) pertenecen a la comunidad del *Frente de Todos* mientras que Fernando Iglesias (@FerIglesias), Mario Raúl Negri (@marioraulnegri) e Ignacio Montes de Oca (@nachomdeo) pertenecen a *Cambiemos* y cómo las interacciones de los usuarios se dan, en su mayoría, entre miembros de la misma comunidad.

De la misma manera, se generó un grafo, como se puede ver en la figura 6.10, para analizar la polarización en Brasil, en este caso, el color verde hace referencia al partido oficialista, mientras que el color rojo hace referencia al *Partido de los trabajadores*, se puede ver a Lula Da silva (@LulaOficial) en la comunidad del PT y en oposición, a Eduardo Bolsonaro (@BolsonaroSP) hijo de Jair Bolsonaro y diputado oficialista.

6.3.3. Detección de comunidades con Walktrap

Para poder clasificar los resultados obtenidos por los algoritmos de identificación de comunidades, es decir, asignarle una semántica a cada cluster identificado, previamente seleccionamos manualmente un conjunto de usuarios de cada comunidad identificando los usuarios que más fueron retweeteados y los usuarios con mayor cantidad de seguidores. Utilizamos este criterio de selección de usuarios dado que estos son muy influyentes. En algunos casos puede tratarse de gente famosa o de renombre y referentes dentro de cada comunidad. Como dijimos anteriormente, un retweet lo consideramos un apoyo a la opinión del usuario.

A modo de ilustración, analizamos, entre muchos otros, los grafos de polarización, como los de la figura 6.9. En este grafo podemos ver que en la comunidad celeste se encuentran

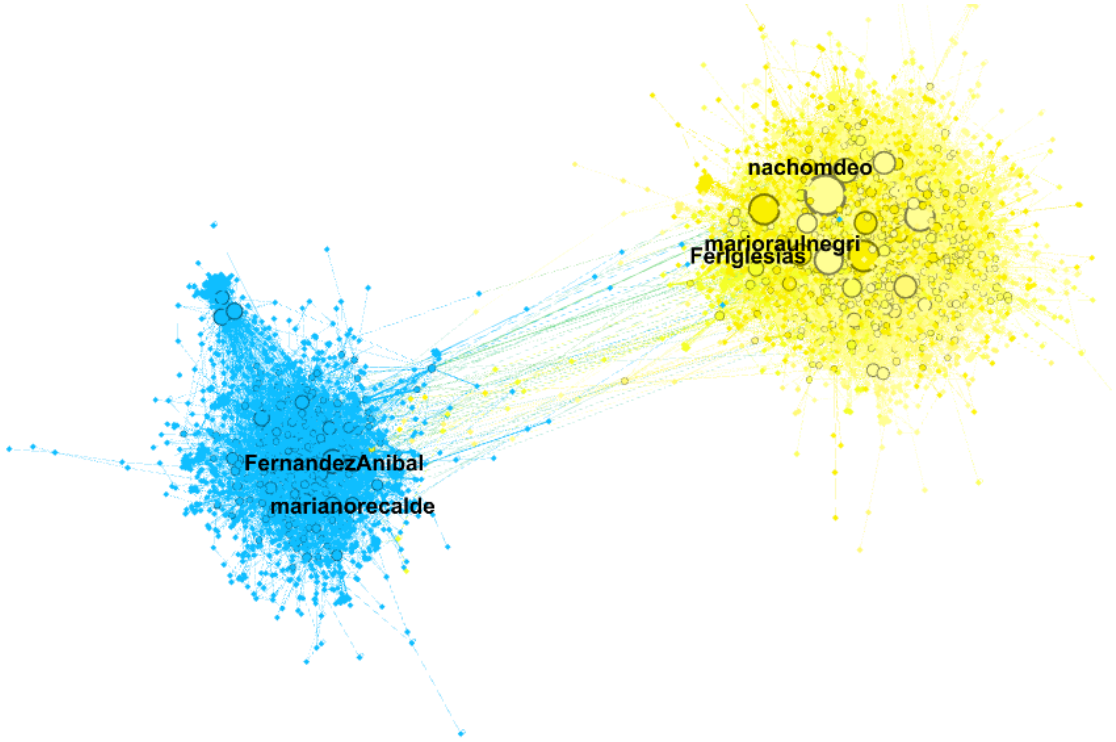


Fig. 6.9: Grafo de retweets del día 13 de octubre del 2019 visualizado a través de ForceAtlas2 y coloreado mediante Walktrap

nombres de usuario como los de *FernandezAnibal*, Secretario General de la presidencia, Jefe de Gabinete durante la presidencia de Cristina Fernandez de Kirchner en el 2015 y Ministro de Seguridad de la Argentina al momento de realizar estos experimentos [5], y *marianorecalde*, candidato a Jefe de Gobierno de la Ciudad de Buenos Aires en el 2015 por el Frente para la Victoria, Legislador de la Ciudad de Buenos Aires en representación de Unidad Ciudadana entre el 2017 y 2019 y actual Senador [6]. Por lo tanto, podemos inferir que la comunidad celeste apoya al partido del *Frente de Todos*. Por otra parte, en la comunidad amarilla figuran *marioraulnegri*, actual Diputado Nacional de la provincia de Córdoba, Presidente del interbloque parlamentario de Juntos por el Cambio [7], *FerIglesias*, Diputado de la Ciudad Autónoma de Buenos Aires por el partido de Cambiemos [8], y *nachomdeo*, periodista y escritor, antiperonista y opositor del gobierno de Cristina Fernandez de Kirchner [9]. De esta manera, podemos clasificar a la comunidad amarilla como la que apoya al partido de *Cambiemos*. Este análisis nos permite clasificar las comunidades obtenidas con Walktrap.

[5] https://www.argentina.gob.ar/seguridad
[6] https://www.senado.gob.ar/senadores/senador/498
[7] https://www.diputados.gov.ar/diputados/mnegri
[8] https://www.diputados.gov.ar/diputados/faiglesias
[9] https://www.penguinlibros.com/es/1692-ignacio-montes-de-oca

Fig. 6.10: Grafo de retweets de los días 16 y 17 de junio del 2020 en Brasil visualizado a través de ForceAtlas2 y coloreado mediante Walktrap

6.3.4. Generación de los modelos con FastText

Para generar los modelos, seleccionamos los días a partir de los cuales queremos entrenarlos y con el texto de los tweets de esos días, ejecutamos FastText [111], en modo supervisado utilizando los hiperparametros definidos en la tabla 6.2

Hiperparámetro	Configuración usada	Definición
dim	300	Tamaño de los vectores de las palabras
wordNGrams	2	Longitud máxima de la palabra en n-grams
ws	5	Tamaño de ventana de contexto, es decir, cantidad de palabras para adelante y para atras que mira posicionandose en una palabra.
epoch	20	Cantidad de veces que se ve cada muestra

Tab. 6.2: Hiperparámetros utilizados para ejecutar FastText

Al igual que en los capítulos anteriores utilizamos Fasttext para la generación de modelos de PLN y para la predicción de datos.

Para mejorar los resultados, utilizamos también *pretrained vectors* en español para los datasets de las elecciones en Argentina y en portugués para los datasets de Brasil, provistos por FastText[10], los cuales son modelos ya entrenados con el vocabulario de cada idioma que contienen información de la lengua utilizada en los tweets para mejorar la eficacia de nuestro modelo, creemos que esto resulta muy útil para modelos entrenados con pocos datos porque nos permite tener una mejor interpretación del lenguaje.

6.3.5. Predicción y evaluación con los modelos

Utilizamos los modelos entrenados para predecir la pertenencia de los usuarios a las principales comunidades en un nuevo conjunto de datos, luego los clasificamos con Walktrap, como ground-truth, y calculamos la métrica de ROC [178] para verificar la efectividad de estos modelos.

Decidimos utilizar el área bajo la curva AUC de ROC para medir el accuracy/precisión de los modelos y poder seleccionar los buenos modelos y descartar los no tan buenos. En este trabajo consideramos una buena predicción cuando el AUC es superior a 0.7. Elegimos este valor luego de realizar distintos experimentos para los datasets de las elecciones del 2019 en Argentina. Con estos experimentos detectamos que, si utilizábamos un umbral más alto, necesitábamos realizar un entrenamiento cada 2 días o, en algunos casos, todos los días. Esto resultaba perjudicial para el modelo ya que, de ese modo, podíamos estar entrenando sobre un tópico y no tanto sobre la jerga.

[10] `https://fasttext.cc/docs/en/crawl-vectors.html`

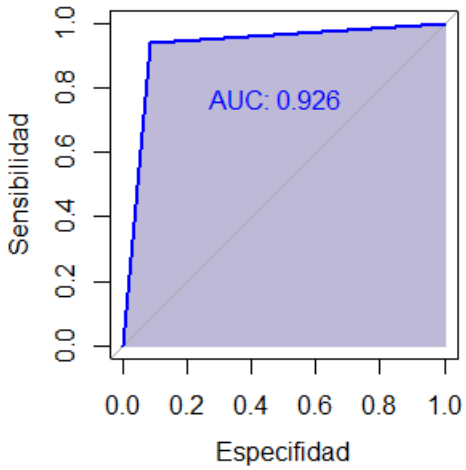


Fig. 6.11: Ejemplo de una curva de ROC y su AUC

La curva de *ROC* presenta la sensibilidad o verdaderos positivos (*vp*) en función de la especificidad o falsos positivos (*fp*) para distintos puntos de corte. Se entiende por *vp* al resultado positivo de un experimento que efectivamente era positivo. Mientras que el *fp* hace referencia al resultado positivo que se obtiene en un experimento pero que debería haber resultado negativo. Utilizamos nuestro ground truth para identificar nuestros *vp* y nuestros *fp*, nos fijamos qué tasa de *vp* y *fp* obtenemos y la marcamos como un punto de la curva ROC. Si la predicción fuera perfecta, es decir, sin solapamiento, la curva pasaría por el punto (0,1) dejando por debajo de ella el cuadrante de ejes positivos y generando así un área igual a 1 bajo ella. Si la predicción fuera inútil (ambas tasas coincidieran y la sensibilidad fuese igual a la proporción de falsos positivos), la curva sería la diagonal de (0,0) a (1,1). En este ejemplo, que se puede ver en la figura 6.11, realizamos el cálculo del AUC entrenando un modelo con los días 6 y 7 de octubre del 2019 y realizando una predicción sobre el día 8 de octubre del mismo año obteniendo un AUC de 0.926.

6.4. Etapa 2

6.4.1. Selección de una estrategia de reentrenamiento de los modelos

Al pasar el tiempo, la jerga se va modificando producto de los tópicos que surgen. Es por eso que, para que los modelos nos sigan resultando útiles para clasificar a los usuarios dentro de ciertas comunidades, es necesario un reentrenamiento de los mismos. Para esto definimos distintas estrategias en base al paso del tiempo y a la experiencia obtenida en los experimentos de la etapa anterior, para reentrenar los modelos. Una vez seleccionada la estrategia que vamos a utilizar para reentrenar a nuestro modelo, seleccionamos los nuevos datos de entrada y realizamos los experimentos.

6.4.2. Generación de los grafos y modelos

Según la estrategia seleccionada en el paso anterior, definiremos cómo vamos a reentrenar nuestro modelo, ya sea agregando días nuevos, eliminando días anteriores, etc. En

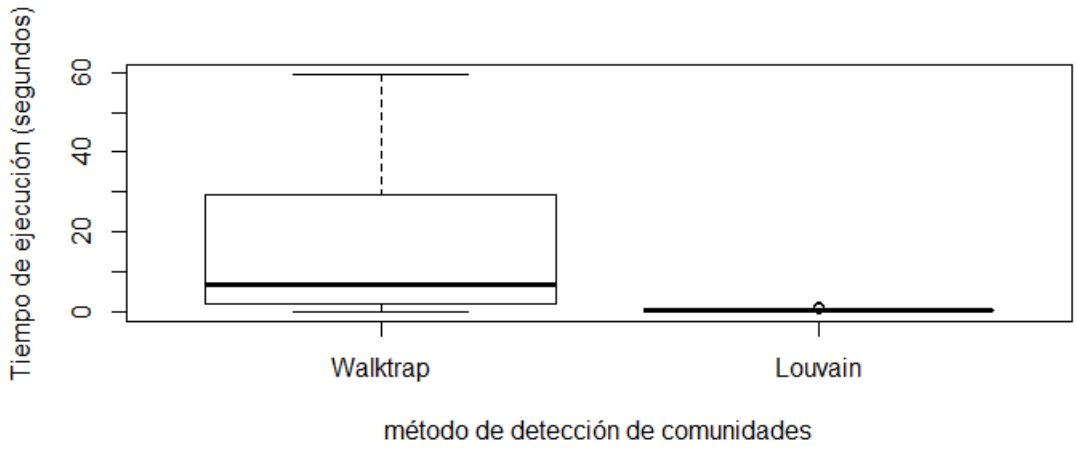


Fig. 6.12: Comparación de tiempos de ejecución de Walktrap vs Louvain en todo el período entre agosto y diciembre del 2019

cualquier caso, necesitamos volver a generar un nuevo grafo de interacciones para poder generar los modelos predictivos.

Para esta etapa decidimos utilizar a Louvain [51] como ground truth ya que al agregar cada vez más datos el método de Walktrap se vuelve muy costoso de ejecutar y Louvain logra identificar a las comunidades con un menor costo computacional y los resultados son bastante similares.

En el boxplot de la figura 6.12 se observa la comparación entre los tiempos de ejecución de ambos algoritmos en cada uno de los días desde el 12 de agosto del 2019 al 31 de diciembre del 2019. Este diagrama representa una serie de datos a través de sus cuartiles. La caja abarca el rango intercuartil de la distribución, es decir, el tramo que va desde el primer cuartil hasta el tercero, lo que incluye el 50 % de las observaciones centrales. La linea negra representa la mediana de los tiempos de ejecución, las lineas al principio y el final son los valores mínimos y máximos obtenidos. Se puede ver claramente cómo el algoritmo de Walktrap necesita de mucho más tiempo para ejecutarse.

6.4.3. Predicción y evaluación de los modelos

Una vez generado el modelo, se lo utiliza para predecir en el día siguiente al último día del entrenamiento. Para esto ejecutamos FastText para obtener la probabilidad de pertenencia de un usuario a cada comunidad y utilizamos Louvain como ground-truth para poder identificar las predicciones correctas. Luego calculamos la métrica de ROC y analizamos el AUC para determinar la eficacia de la predicción. Ya con este resultado, se utiliza la estrategia seleccionada previamente y se determina si el resultado obtenido dictamina que es necesario un reentrenamiento, y de ser así, se vuelve a la etapa anterior para reentrenar el modelo; de lo contrario, se seleccionan los datos del siguiente día y se vuelve a predecir, repitiendo este proceso hasta que se haya analizado toda la información disponible. Como mencionamos anteriormente, en este trabajo dividimos los conjuntos de datos en tweets diarios para para utilizar tanto de en el entrenamiento como en la predicción. Esto nos proporciona una unidad de tiempo lo suficientemente pequeña para

identificar la aparición de un nuevo tópico y lo suficientemente grande como para analizar la jerga con nuestro predictor.

6.5. Experimentos y resultados

6.5.1. Generación de los modelos

Nuestro objetivo en esta etapa es encontrar la mejor estrategia para crear modelos predictivos que permitan clasificar a las personas en las principales comunidades en base a la jerga utilizada. Para eso analizamos distintos experimentos hasta obtener el método de entrenamiento que consideramos que funciona mejor.

En una primera etapa, generamos modelos con pocos días de entrenamiento para identificar más fácilmente los posibles motivos de un buen o mal resultado en una predicción.

Analizamos modelos en intervalos de 7 días. Para eso generamos modelos de 1 día utilizando el 90 % de los tweets de ese día para predecir los 3 días siguientes, los 3 anteriores y el 10 % de los tweets no utilizados del día seleccionado. Nuestra hipótesis es que, como la jerga se mantiene a lo largo del tiempo, es indistinto predecir en días anteriores o posteriores a la generación del modelo. Este experimento no nos dio buenos resultados, pero nos hizo entender que entrenar en un día solo es una mala idea. Si bien los días contiguos al día de entrenamiento dieron un buen resultado, al alejarse 1 o 2 días más la predicción empieza a disminuir drásticamente, como se puede ver en la figura 6.13. Además, dependiendo del día seleccionado, se podría estar hablando de un tópico en particular y existe la posibilidad de que la predicción estuviera captando este tópico y no la jerga, como era nuestra intención. En consecuencia, decidimos ampliar este experimento incrementando la cantidad de días a utilizar para la generación del modelo. Utilizamos entonces n días contiguos para generar el modelo, pero los resultados obtenidos fueron muy similares. Analizando las conversaciones de los tweets, reconfirmamos nuestra teoría de que en el ámbito de la política y, en particular, en el contexto de las elecciones presidenciales, en una misma semana pueden aparecer diversos cambios de tópico producto de las campañas, discursos, etc. Así, entendimos que teníamos que buscar una estrategia que nos permitiera capturar la jerga abstrayéndonos de los tópicos conversados durante unos pocos días. Es en este contexto que decidimos modificar la forma de entrenar los modelos, bajo la hipótesis de que, al seleccionar una cierta cantidad de días no contiguos de cada semana, podríamos sacar el foco de lo que se trató unos días en particular y concentrarnos en la jerga utilizada por las personas, obteniendo mejores resultados en las predicciones.

Intervalos cerrados

La intención de este experimento fue la de evaluar la capacidad de predicción dentro de un intervalo cerrado utilizando días de este como entrenamiento. Si bien la jerga puede mutar, es algo que perdura a lo largo del tiempo. Es por esto que, si generamos un modelo con información de algunos días dentro de un intervalo cerrado, este modelo debería poder predecir con una buena probabilidad dentro de este intervalo, dado que este capturaría la jerga utilizada en este período. Para tener una buena estimación de la performance, el ROC AUC de los días de entrenamiento se estimó mediante cross-validation [31] utilizando el 90 % de los datos para entrenar y el 10 % restante para testear.

En este experimento predijimos dentro del mes de diciembre, utilizando solamente 8 días para el entrenamiento. Entrenamos un modelo A, con los días 1 al 4 y 28 al 31 de

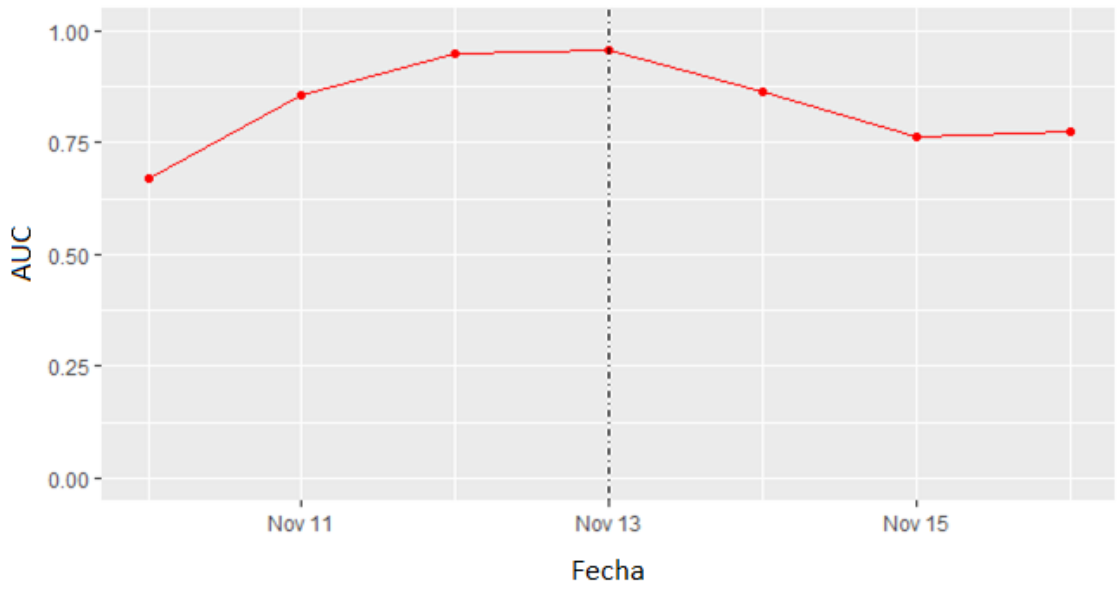


Fig. 6.13: AUC de las predicciones del modelo entrenado con 1 solo día

diciembre, es decir con los primeros y últimos días del intervalo y otro B utilizando 2 días de cada semana del mes es decir, 3,5,10,12,17,19,24,26 de diciembre. En la figura. 6.14, graficamos el ROC AUC por día de cada modelo. Se observa que, para el modelo A, el AUC promedio es de 0.78 y su desvío estándar(SD, según sus iniciales en inglés) de 0.13 mientras que para el modelo B, el AUC promedio es de 0.82 y su SD de 0.129

Los resultados de este experimento nos muestran que es más beneficioso entrenar con distintos días de la semana en un intervalo cerrado que entrenar solamente con los extremos, ya que al entrenar con días contiguos podemos estar entrenando sobre un tópico, mientras que al generar el modelo utilizando distintos momentos, podemos minimizar este riesgo. Seleccionamos solamente este ejemplo como ilustración de que entrenar 2 días de la semana no contiguos obtiene en general mejores resultados, pero este mismo experimento se realizó también para los datasets de septiembre, octubre y noviembre con resultados similares.

Predecir los días futuros

En el experimento pasado, pudimos ver que seleccionar 2 días de cada semana nos permite obtener una mejor interpretación de la jerga que seleccionando los primeros 4 días y los últimos 4. En este experimento, nuestro objetivo es probar si esta misma estrategia nos resulta efectiva para predecir los días futuros. Para probar esto, entrenamos 2 modelos de 8 días cada uno con el objetivo de predecir los 7 días siguientes. El primer modelo lo entrenamos con 8 días anteriores a la semana que queremos predecir y el segundo, lo entrenamos utilizando 2 días de cada semana de las 4 anteriores.

En este caso, el Modelo A, utiliza los 8 días anteriores a la semana que queremos predecir, es decir, del 7 al 14 de noviembre y el Modelo B, utiliza 2 días de cada semana de las 4 anteriores, siendo estos los días 15, 17, 22 y 24 de octubre y 5,7,12 y 14 de noviembre y evaluamos sobre los siguientes 7 días. En la figura. 6.15, graficamos el ROC AUC por

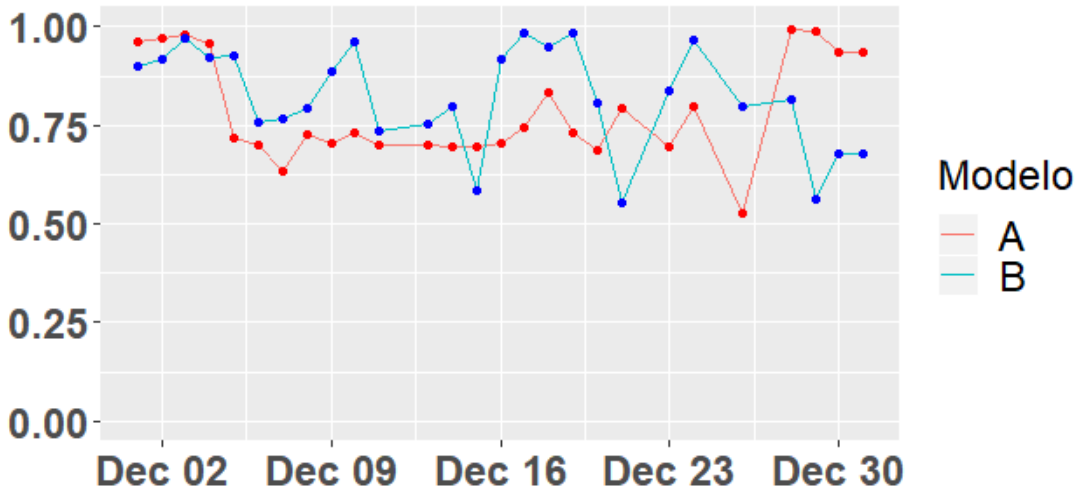


Fig. 6.14: AUC por día entrenando dentro del intervalo de predicción

día de cada modelo. Se observa que, para el modelo A, el AUC promedio es de 0.73 y su SD de 0.105 mientras que para el modelo B, el AUC promedio es de 0.796 y su SD de 0.023

Los resultados obtenidos demuestran que utilizar los 2 días de cada semana para entrenar nuestro modelo nos permite predecir con mayor precisión los días futuros, lo cual respalda nuestra hipótesis inicial.

Aumentando la cantidad de datos

Este experimento consistió en verificar si aumentar la cantidad de días de entrenamiento ayuda a mejorar la capacidad de predecir del modelo.

Para esto entrenamos un modelo A, con todos los días de septiembre y octubre e intentamos predecir en noviembre y lo comparamos con otro B, usando solamente 2 días de cada semana de esos mismos meses, es decir los días 3,6,10,12,17,19,24,27 de septiembre y los días 8,10,15,17,22,24 y 29 de octubre. En la figura. 6.16, graficamos el ROC AUC por día de cada modelo. Se observa que, para el modelo A, el AUC promedio es de 0.716 y su SD de 0.054 mientras que para el modelo B, el AUC promedio es de 0.711 y su SD de 0.08.

Este experimento nos demuestra que agregar más días para entrenar el modelo no necesariamente redundará en mejores resultados. Se observa que utilizando una diferencia de 44 días entre los dos modelos, la diferencia obtenida fue tan solo de 0.005 a favor del modelo A.

6.5.2. Reentrenamiento de modelos

Como pudimos ver en los experimentos anteriores, al pasar el tiempo o intentar predecir las comunidades fuera del intervalo cerrado, la eficiencia del modelo empieza a disminuir. Asimismo, a lo largo del tiempo suelen aparecer cambios de tópicos, estos eventos modifican la jerga para siempre o simplemente por uno o varios días, el problema es que es imposible

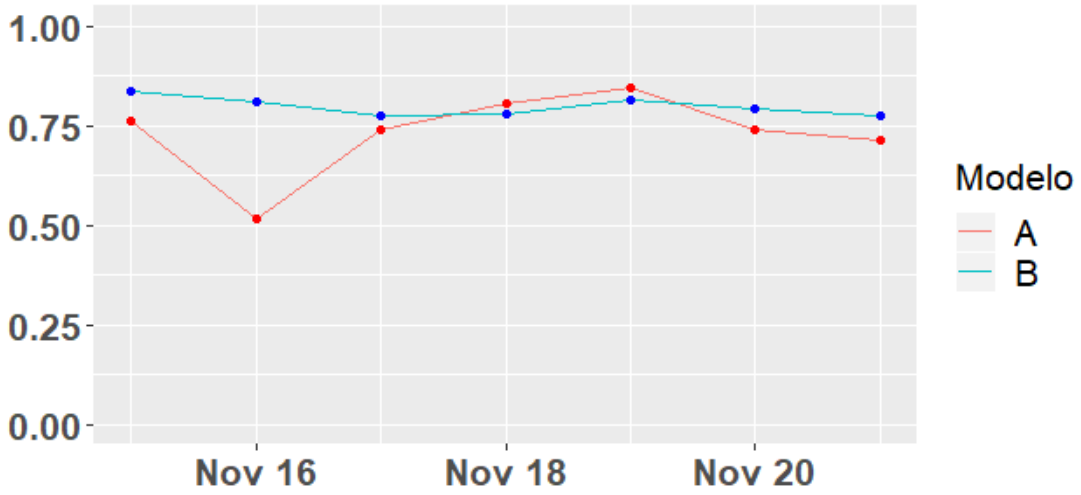


Fig. 6.15: AUC por día entrenando dentro del intervalo de predicción

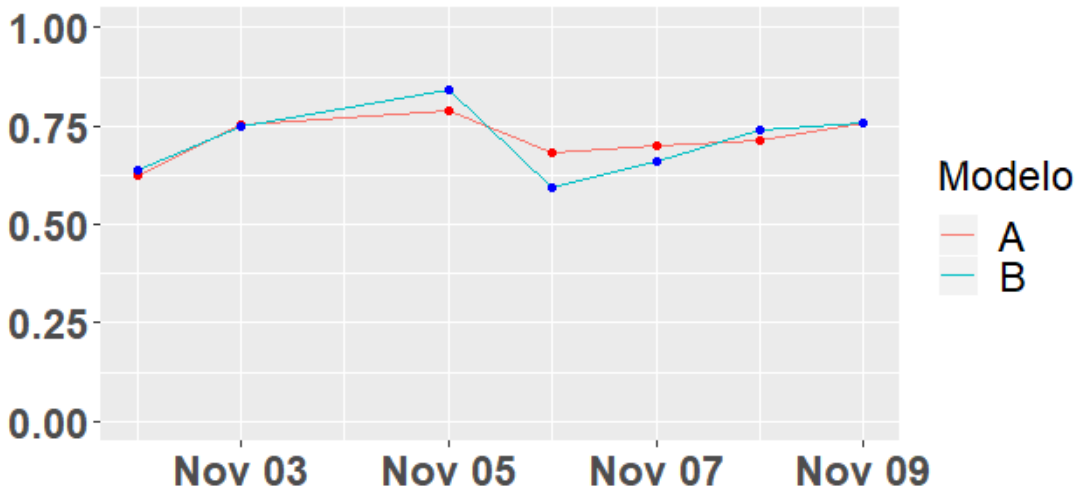


Fig. 6.16: AUC por día entrenando dentro del intervalo de predicción

determinar cuando van a ocurrir. Bajo esta premisa, necesitamos definir cuándo es que un modelo empieza a ser malo y necesita ser reentrenado para mantener la calidad de los resultados obtenidos. Decimos que un modelo predice mal cuando el AUC obtenido es inferior a 0.7.

En esta etapa realizaremos diversos métodos para determinar qué estrategia puede resultar más efectiva. En cada método se experimentará con los mismos datasets, estos están detallados en la tabla 6.3.

Dataset	**Definición**
DS-Septiembre	Contiene los tweets bajo el criterio *Macri* de los días de septiembre y octubre del 2019.
DS-Octubre	Contiene los tweets bajo el criterio *Macri* de los días de octubre y noviembre del 2019.
DS-Noviembre	Contiene los tweets bajo el criterio *Macri* de los días de noviembre y diciembre del 2019.
DS-Bolsonaro	Contiene los tweets bajo el criterio *Bolsonaro* de los días de junio y julio del 2020.

Tab. 6.3: Datasets utilizados para los experimentos de reentrenamiento.

Aprendiendo de los buenos y malos resultados

En este método, procuramos mitigar la posibilidad de confusiones causadas por la disminución en la calidad de la predicción, ya sea debido a la introducción de un nuevo tema o a cambios en la jerga. Por esta razón, determinamos que el modelo no está realizando predicciones adecuadas cuando se observan dos días consecutivos con un AUC menor a 0.7 En este caso, conservamos el criterio de entrenar con días no contiguos, utilizamos entonces el último día que predijo correctamente como refuerzo de la jerga que se viene utilizando e incorporamos el segundo día con AUC menor a 0.7 a nuestro modelo, estableciendo que si la mala predicción se mantiene más de un día, es necesario incorporar la nueva terminología para poder seguir prediciendo correctamente.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.17 y 6.18, para DS-Septiembre este método no pudo ser aplicado porque en este período no ocurrió que 2 días consecutivos tengan un AUC menor a 0.7. Lo mismo ocurrió en los próximos 3 métodos que utilizan una lógica similar. Para el DS-Octubre se logró un AUC promedio de 0.719, un SD de 0.109 en 5 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.72, un SD de 0.11 en 2 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.754, un SD de 0.07 en 2 reentrenamientos.

Experimento 6.5.2, Quitándole relevancia a los datos más viejos.

Al pasar el tiempo, puede que la jerga empiece a mutar producto de los cambios de tópico, es por eso que creemos que la forma de hablar en el pasado no sea la misma que en el futuro, nuestra intención entonces en esta etapa, es la de ir disminuyendo en un 10 % el porcentaje de datos utilizados de los días más viejos al agregar un día nuevo, es decir, al

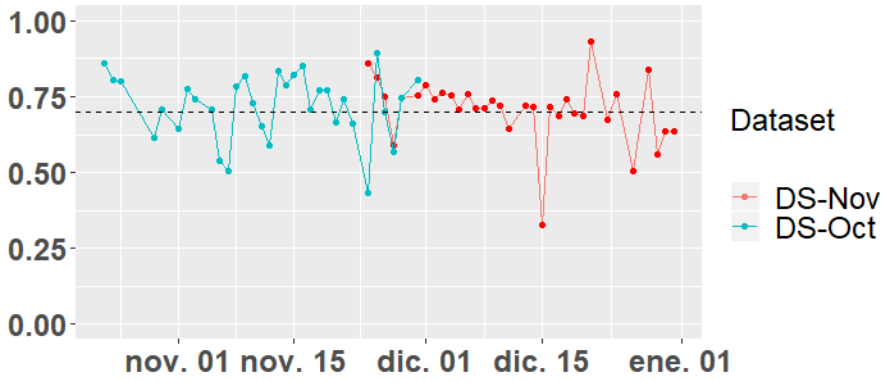


Fig. 6.17: AUC por día con los Datasets de las PASO del 2019 en Argentina

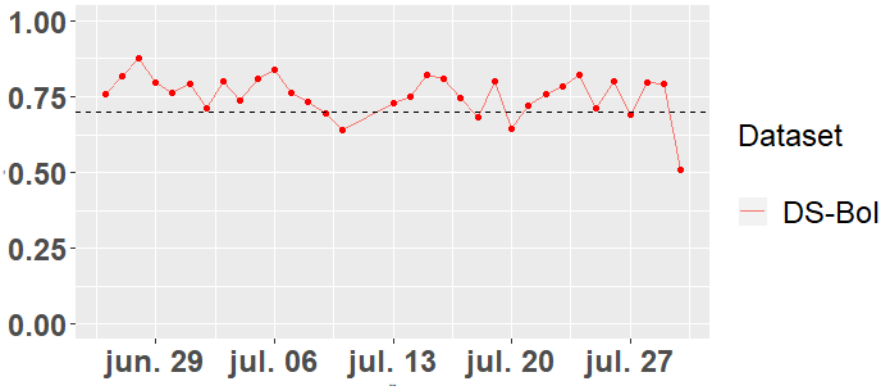


Fig. 6.18: AUC por día con el Dataset de Brasil en el 2020

agregar 10 días el primer día de entrenamiento dejaría de utilizarse y así sucesivamente. En este experimento, utilizamos el mismo criterio de reentrenamiento que en el experimento pasado, es decir, utilizar para el reentrenamiento el último día cuyo AUC fue mayor al 0.7 y el segundo consecutivo menor a este valor, pero a diferencia del método anterior, disminuimos el porcentaje de datos utilizados de los tweets más viejos en un 10 % por día agregado.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.19 y 6.20. Para el DS-Octubre se logró un AUC promedio de 0.735, un SD de 0.088 en 4 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.707, un SD de 0.078 en 1 reentrenamiento. Para DS-Bolsonaro, un AUC de 0.731, un SD de 0.096 en 3 reentrenamientos. Como mencionamos anteriormente, para el DS-Septiembre este método no pudo ser aplicado.

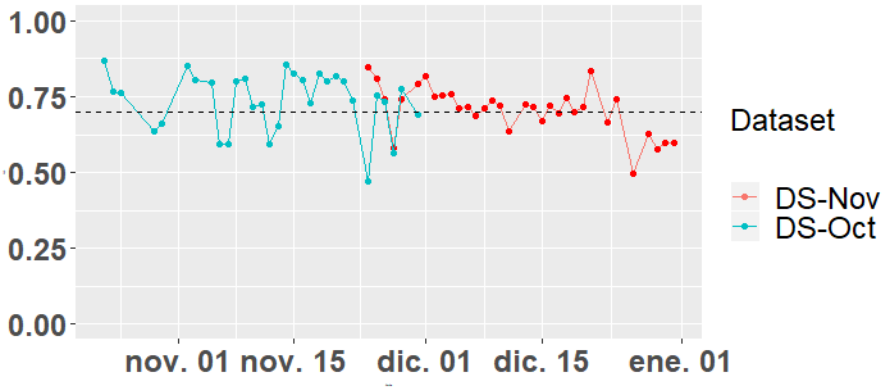


Fig. 6.19: AUC por día con los Datasets de las PASO del 2019 en Argentina

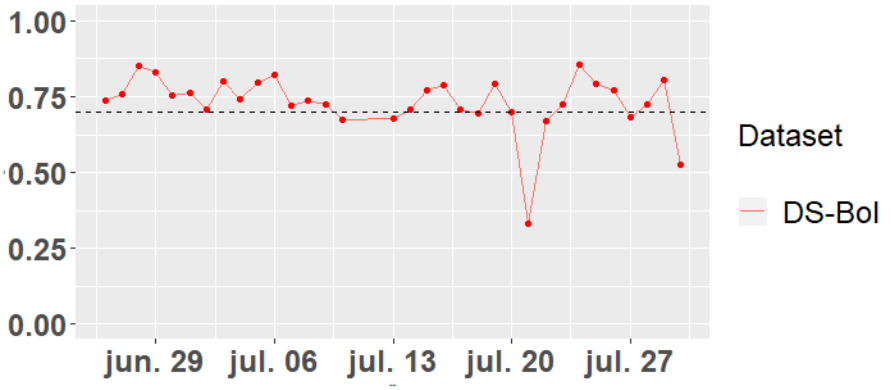


Fig. 6.20: AUC por día con el Dataset de Brasil en el 2020

Utilizando los primeros 2 días consecutivos con mal resultado

Para este método, nos inspiramos en la técnica de *Boosting* ya que ponderamos los datos mal clasificados del modelo anterior para generar mejores clasificadores. Basándonos

en esto, al encontrar 2 días consecutivos con un AUC menor a 0.7, decidimos utilizar a estos 2 días para el reentrenamiento, para reforzar el modelo con los días en los que la predicción fue más débil.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.21 y 6.22. Para el DS-Octubre se logró un AUC promedio de 0.744, un SD de 0.095 en 4 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.735, un SD de 0.088 en 4 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.746, un SD de 0.089 en 2 reentrenamientos. Como mencionamos en el punto 6.5.2, para el DS-Septiembre este método no pudo ser aplicado.

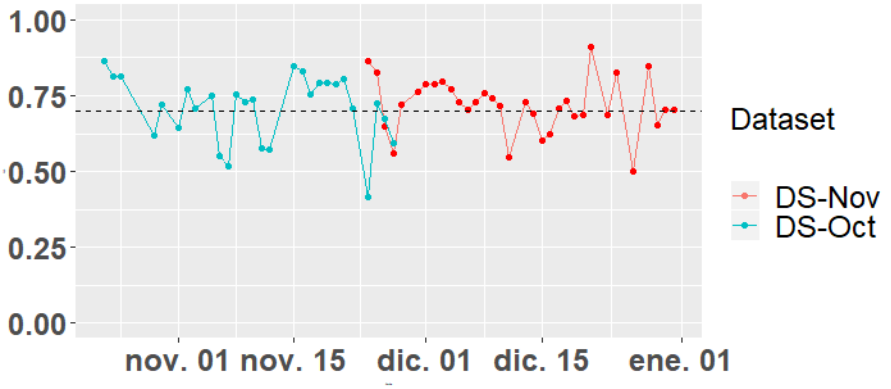


Fig. 6.21: AUC por día con los Datasets de las PASO del 2019 en Argentina

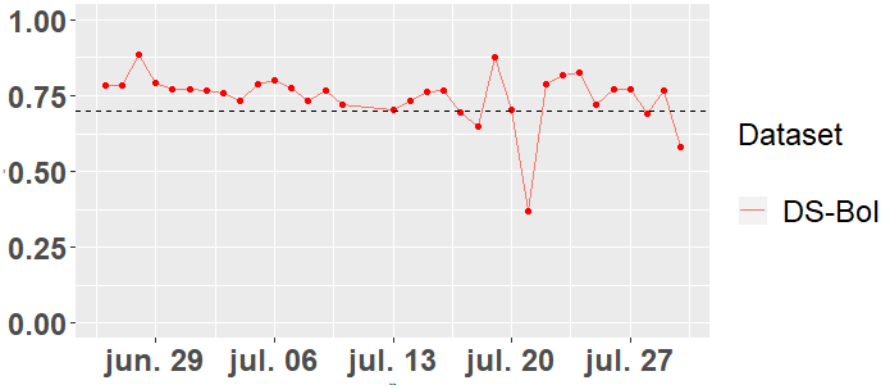


Fig. 6.22: AUC por día con el Dataset de Brasil en el 2020

Experimento 6.5.2, Quitándole relevancia a los datos más viejos.

Al igual que en el método anterior, decidimos utilizar la técnica de reentrenar nuestro modelo incorporando los últimos 2 días consecutivos con AUC menor a 0.7 y disminuir un 10 % el porcentaje de datos utilizados de los días más viejos por cada día agregado.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.23 y 6.24. Para el DS-Octubre se logró un AUC promedio de 0.714, un SD de

0.116 en 4 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.728, un SD de 0.072 en 2 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.741, un SD de 0.108 en 2 reentrenamientos. Como mencionamos en el punto 6.5.2, para el DS-Septiembre este método no pudo ser aplicado.

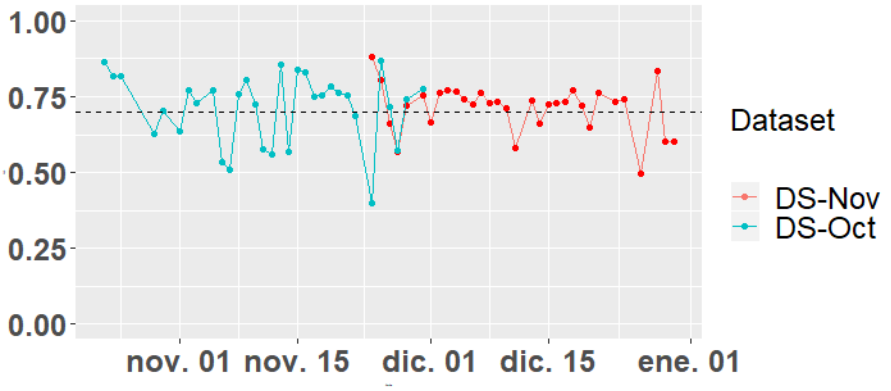


Fig. 6.23: AUC por día con los Datasets de las PASO del 2019 en Argentina

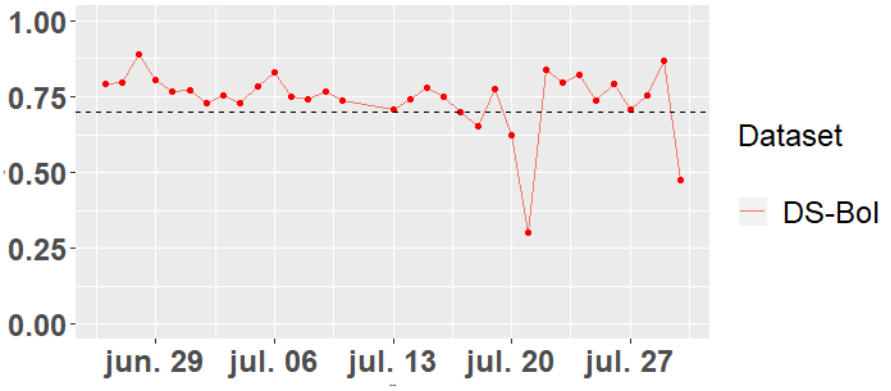


Fig. 6.24: AUC por día con el Dataset de Brasil en el 2020

Sliding Window los martes y jueves

En este experimento nos inspiramos en el concepto de Sliding Window. En este método se establece una ventana de x envíos simultáneos para no sofocar al receptor con demasiada información y se espera la respuesta para continuar la transmisión, nosotros en nuestro caso tenemos una ventana de 4 días, siendo estos los martes y jueves de cada semana, para no sobrecargar al modelo con información pasada, vamos eliminando la información del día más viejo al agregar los datos del nuevo día. El objetivo es verificar si entrenar siempre con los mismos días de la semana y eliminando los datos más viejos, se logre generar un modelo que esté más actualizado a lo que está ocurriendo y pueda ser menos sensible a los cambios de tópico.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.25 y 6.26. Para el DS-Septiembre se logró un AUC promedio de 0.764, un SD

de 0.084 en 9 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.725, un SD de 0.111 en 11 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.742, un SD de 0.082 en 11 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.758, un SD de 0.083 en 13 reentrenamientos.

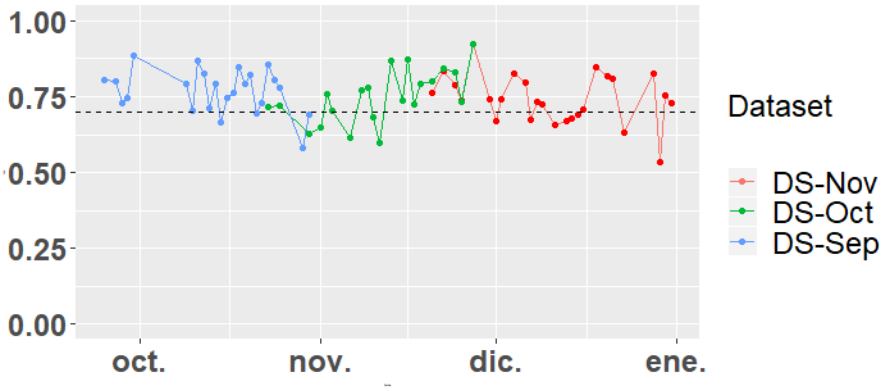


Fig. 6.25: AUC por día con los Datasets de las PASO del 2019 en Argentina

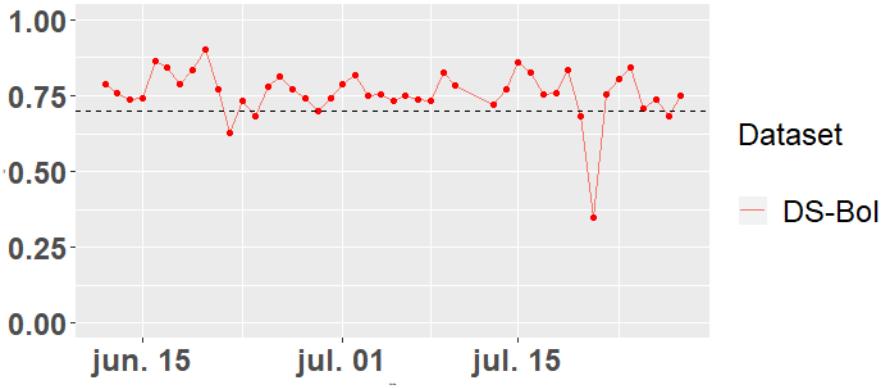


Fig. 6.26: AUC por día con el Dataset de Brasil en el 2020

Forzando reentrenamientos semanales

Cuando va pasando el tiempo, la calidad de la predicción empieza a disminuir, no siempre ocurre que 2 días seguidos el AUC sea menor que 0.7, muchas veces ese valor se mantiene muy cerca de esta linea, es por eso que decidimos realizar un reentrenamiento si al pasar una semana no fue necesario realizar uno, de esta manera buscamos mejorar el AUC obtenido y evitar un posible reentrenamiento futuro por un mal resultado. Para este reentrenamiento agregamos el último día de esta semana.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.27 y 6.28. Para el DS-Septiembre se logró un AUC promedio de 0.741, un SD de 0.054 en 4 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.726, un SD de 0.102 en 5 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.726,

un SD de 0.085 en 5 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.758, un SD de 0.091 en 6 reentrenamientos.

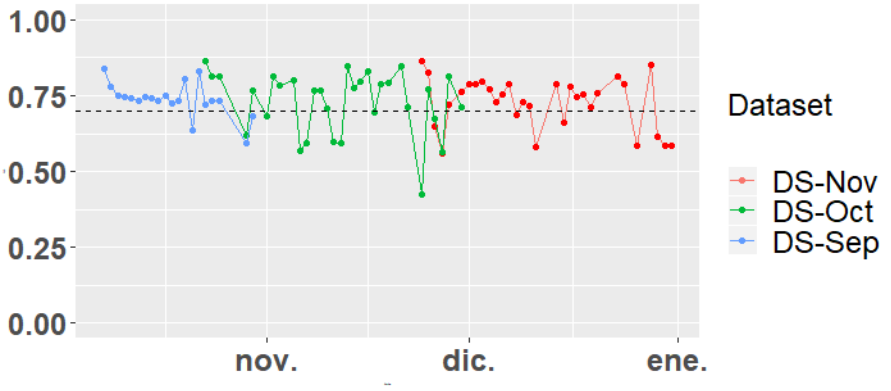


Fig. 6.27: AUC por día con los Datasets de las PASO del 2019 en Argentina

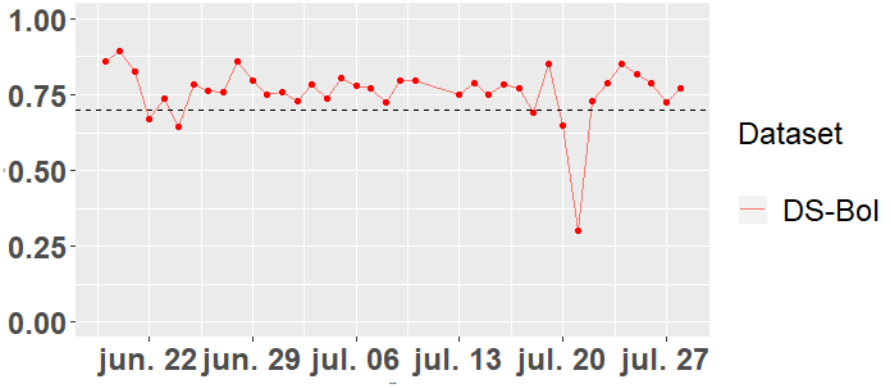


Fig. 6.28: AUC por día con el Dataset de Brasil en el 2020

Nueva métrica de performance a lo largo del tiempo

Siguiendo con el espíritu del caso anterior, no siempre ocurre que el AUC sea menor que 0.7 durante 2 días contiguos pero igual se puede detectar que la calidad del predictor empieza a disminuir y los valores de la predicción empiezan a oscilar entre el valor del punto de corte, obteniendo malos resultados pero no entrando en los planteados hasta el momento para reentrenar. Por otro lado, aunque en ocasiones pueda arrojar resultados positivos, si durante dos días consecutivos presenta mal desempeño, nos lleva a una fase de reentrenamiento. En algunos casos esto se debe a que un tópico se extendió más allá de lo habitual, incorporando información que afecta negativamente el rendimiento futuro de nuestro modelo. Bajo esta lógica, introducimos una métrica que determina qué tanto está disminuyendo la calidad. Para eso, empezamos a almacenar la diferencia que hay entre el valor del AUC obtenido y nuestro umbral de 0.7, es decir, $\Delta = AUC - 0{,}7$. Cuando la sumatoria de los valores almacenados es menor que 0.05 realizamos un reentrenamiento

agregando a nuestro modelo el último día que hizo que sea necesario el reentrenamiento y reseteamos el acumulador. Por ejemplo, si nuestro AUC fue de 0.76, al día siguiente 0.74 y al siguiente 0.63, nuestro $\Delta_1 = 0{,}06$ $\Delta_2 = 0{,}04$ $\Delta_3 = -0{,}07$ y finalmente $\sum_{i=0}^{3} \Delta_i = 0{,}03$ Como en este caso el valor resultante es menor a 0.05, realizamos un reentrenamiento.

El Δ se almacena hasta una semana para evitar que muy buenas predicciones del pasado hagan que el predictor prediga muy mal durante muchos días hasta que el valor del umbral sea menor a 0.05

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.29 y 6.30. Para el DS-Septiembre se logró un AUC promedio de 0.762, un SD de 0.070 en 3 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.744, un SD de 0.106 en 6 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.741, un SD de 0.070 en 7 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.743, un SD de 0.095 en 4 reentrenamientos.

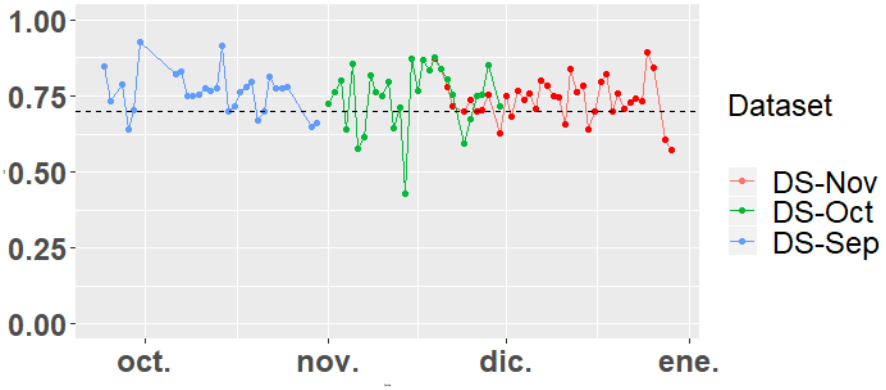


Fig. 6.29: AUC por día con los Datasets de las PASO del 2019 en Argentina

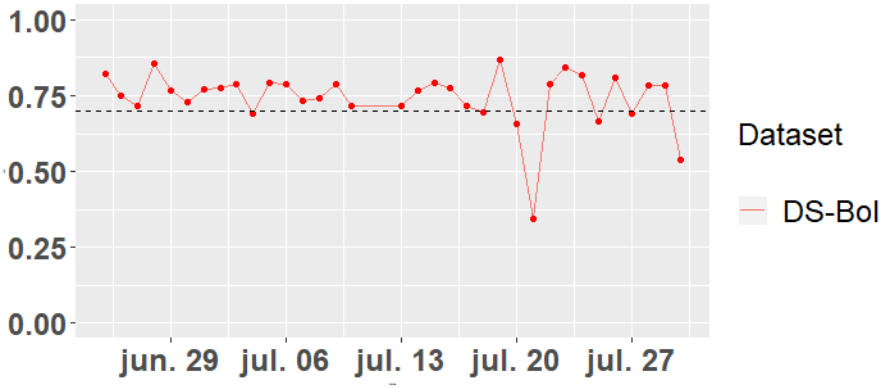


Fig. 6.30: AUC por día con el Dataset de Brasil en el 2020

Experimento 6.5.2, Quitándole relevancia a los datos más viejos

Al igual que en experimentos pasados, la idea es repetir el mismo proceso que en el caso anterior disminuyendo en un 10 % el porcentaje de datos utilizados en los días más

viejos al agregar un día nuevo al conjunto de entrenamiento

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.31 y 6.32. Para el DS-Septiembre se logró un AUC promedio de 0.759, un SD de 0.079 en 3 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.737, un SD de 0.115 en 6 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.751, un SD de 0.089 en 8 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.743, un SD de 0.088 en 3 reentrenamientos.

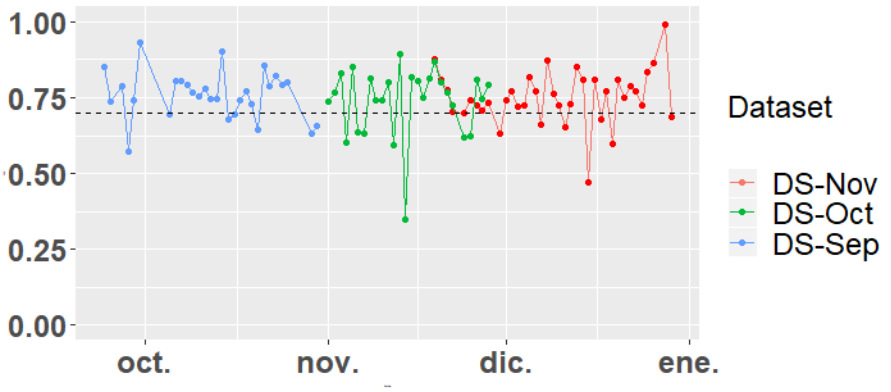


Fig. 6.31: AUC por día con los Datasets de las PASO del 2019 en Argentina

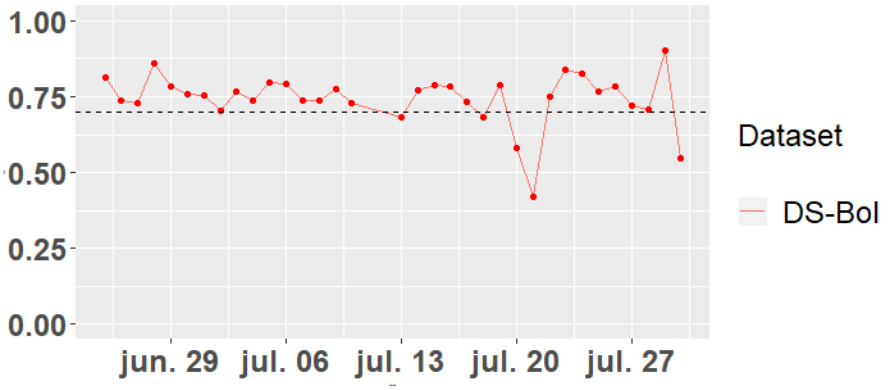


Fig. 6.32: AUC por día con el Dataset de Brasil en el 2020

Se hace más rigurosa la métrica establecida en el experimento 6.5.2

Siguiendo la misma línea que en experimento 6.5.2, se sube el umbral de 0.05 a 0.07 con el objetivo de aumentar la calidad de la predicción.

Ejecutamos este método en nuestros datasets definidos anteriormente y generamos los gráficos 6.33 y 6.34. Para el DS-Septiembre se logró un AUC promedio de 0.771, un SD de 0.066 en 3 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.746, un SD de 0.113 en 6 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.751, un SD de 0.078 en 9 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.763, un SD de 0.089 en 8 reentrenamientos.

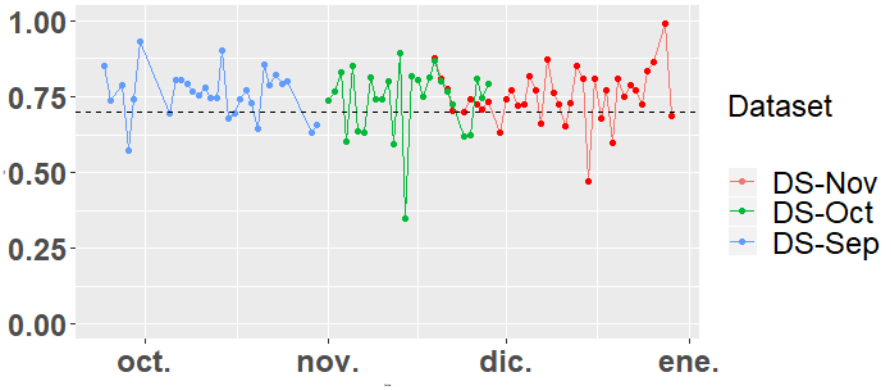


Fig. 6.33: AUC por día con los Datasets de las PASO del 2019 en Argentina

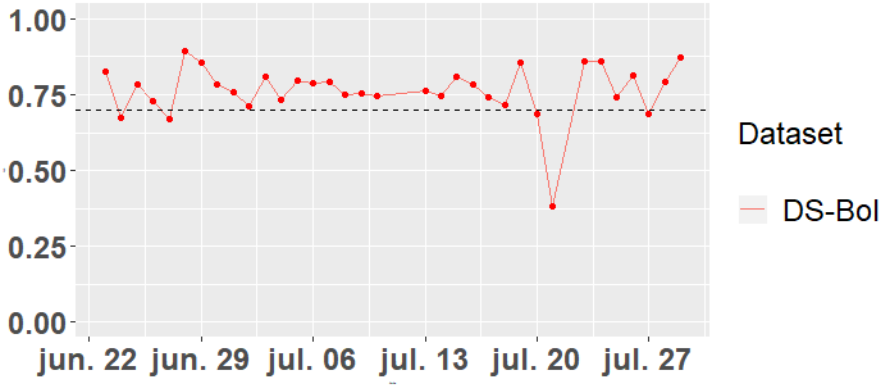


Fig. 6.34: AUC por día con el Dataset de Brasil en el 2020

Experimento 6.5.2, Quitándole relevancia a los datos más viejos

Volvemos a realizar el experimento anterior pero disminuyendo en un 10 % el porcentaje de los datos más viejos utilizados al agregar un nuevo día

En este caso aplicando este método a nuestros datasets previamente establecidos, hemos creado los gráficos 6.35 y 6.36. Para el DS-Septiembre se logró un AUC promedio de 0.768, un SD de 0.071 en 3 reentrenamientos. Para el DS-Octubre se logró un AUC promedio de 0.745, un SD de 0.101 en 6 reentrenamientos. Para DS-Noviembre un AUC promedio de 0.754, un SD de 0.09 en 10 reentrenamientos. Para DS-Bolsonaro, un AUC de 0.774, un SD de 0.05 en 7 reentrenamientos.

Resultados de los reentrenamientos

Para analizar los resultados obtenidos en los distintos experimentos, dividimos el análisis en 3 ejes:

- Cantidad de iteraciones necesarias para procesar todos los datos.

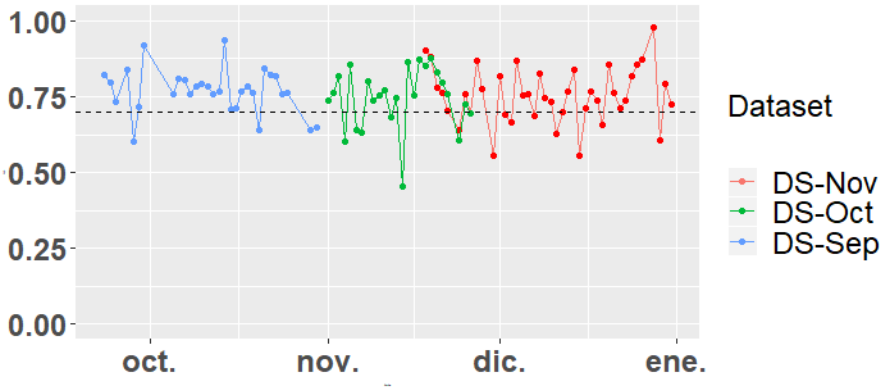


Fig. 6.35: AUC por día con los Datasets de las PASO del 2019 en Argentina

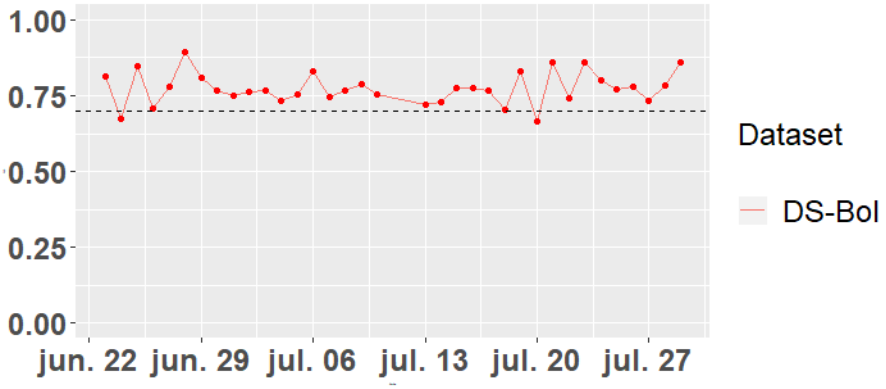


Fig. 6.36: AUC por día con el Dataset de Brasil en el 2020

- AUC promedio obtenido para cada uno de los experimentos.
- Desvío standard obtenido para cada uno de los experimentos.

Cantidad de iteraciones

En la tabla 6.37 comparamos la cantidad de reentrenamientos que fueron necesarios para poder analizar la totalidad de los datos de cada dataset. Cabe destacar que la cantidad de reentrenamientos necesarios no es algo necesariamente bueno o malo para la calidad de la predicción, sino que simplemente depende de la estrategia de reentrenamiento seleccionada. En algunos casos siempre se reentrena utilizando la misma cantidad de días y en otros es incremental. A medida que se aumenta la cantidad de datos, reentrenar modelos se puede volver más costoso, tanto en recursos como en tiempo, con lo cual hay que encontrar un equilibrio entre la cantidad de iteraciones, tamaño del conjunto de datos y precisión deseada para cada modelo.

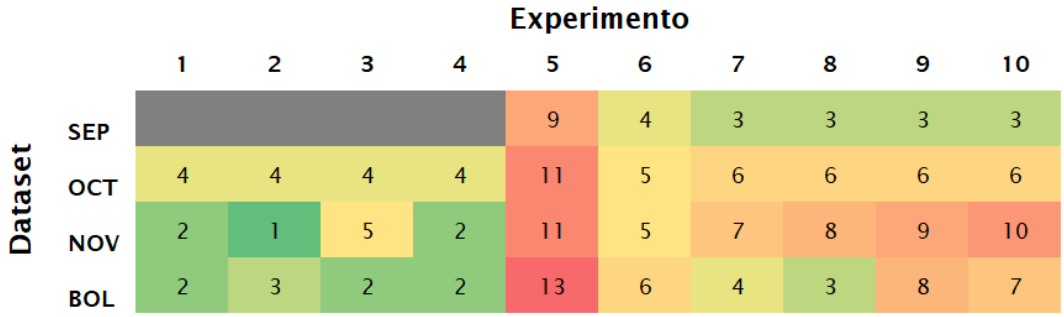


Fig. 6.37: Comparación de cantidad de reentrenamientos que fueron necesarios en los distintos experimentos

AUC promedio

En la tabla 6.38 comparamos el AUC promedio obtenido en cada uno de los experimentos. Se puede ver como los experimentos que utilizan la métrica introducida en el experimento 6.5.2 y ajustada en el experimento 6.5.2 y 6.5.2 son las que mejores resultados consiguieron.

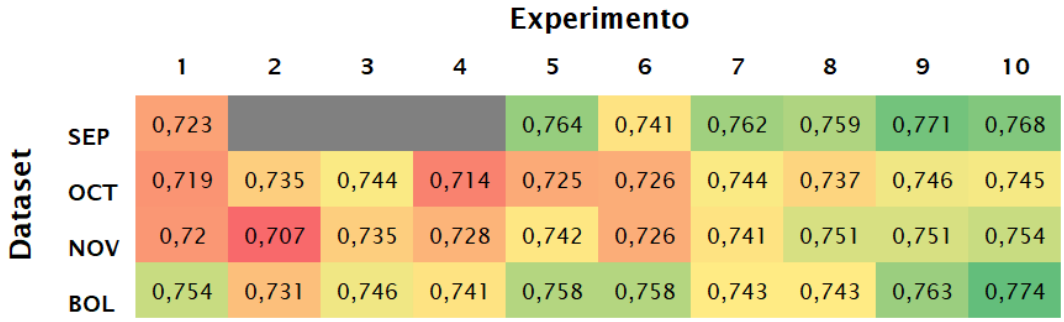


Fig. 6.38: Comparación de AUC obtenido en los distintos experimentos

Desvío standard

En la tabla 6.39 comparamos el desvío standard promedio de los resultados de los experimentos. Mientras más pequeño sea este valor, más estable va a ser nuestro predictor a lo largo del tiempo. Podemos ver que no hay ningún método más efectivo que otro en este punto. Esto se debe a que día a día puede haber mucha volatilidad en la forma de comunicarse de las personas .

6.6. Conclusiones

En la sección de entrenamiento de los modelos pudimos verificar nuestra hipótesis inicial al analizar los tres experimentos. Si bien aumentar la cantidad de días es beneficioso

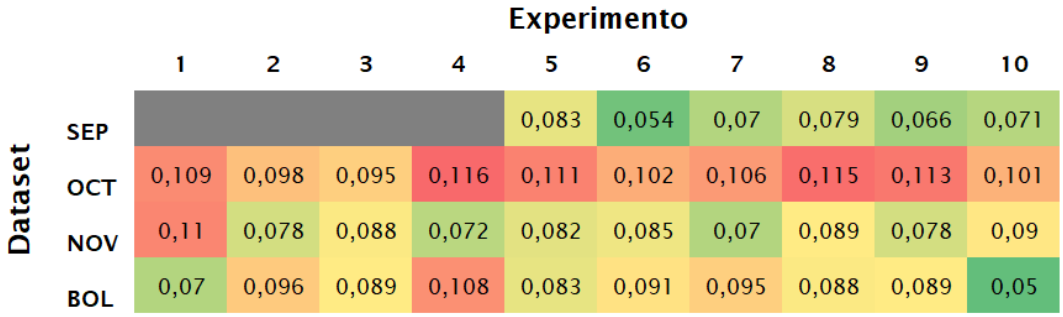


Fig. 6.39: Comparación del desvío standard de los distintos experimentos

para mejorar la performance del modelo, es importante elegir una buena estrategia para hacerlo. Esto lo podemos ver en los resultados de los primeros dos experimentos donde utilizando la misma cantidad de datos, la estrategia de seleccionar los días de entrenamiento de semanas distintas permite obtener un mejor resultado en la predicción y en el tercero, que a pesar de tener más datos y días de entrenamiento, el resultado no varía mucho.
En el experimento 6.5.1 notamos que predecir dentro del período de entrenamiento nos permite obtener un mejor resultado. Lo atribuimos a que los cambios de tópico no afectan tanto al predictor debido a que las jergas utilizadas en los tópicos son capturadas para el entrenamiento de los modelos dentro de este intervalo.
En la sección de reentrenamiento de modelos, pudimos ver que no hay una bala de plata para elegir el mejor método, es decir, los métodos que parecen obtener una mejor performance suelen requerir una mayor cantidad de reentrenamientos. Esto puede ser muy costoso en modelos muy grandes ya que los algoritmos de detección de comunidades empiezan a requerir muchos más recursos a medida que aumenta la cantidad de datos. En particular, Walktrap se vuelve muy costoso de ejecutar muy rápidamente, lo cual obliga a utilizar algoritmos de optimización como Louvain. Por otro lado, en los experimentos también se puede ver que, al igual que en la sección de entrenamiento, tener más reentrenamientos o más datos no indican necesariamente que la performance del método vaya a ser superior. Consideramos que la métrica introducida en el método 6.5.2 es la que nos permite obtener un mejor resultado global. Esta, al tener una "memoria" de las predicciones pasadas, nos ayuda a evitar el reentrenamiento cuando aparece un nuevo tópico que dura unos pocos días y nos permite introducir nuevos datos cuando la calidad de la predicción empieza a disminuir, permitiendo encontrar un método estable a lo largo del tiempo.

En conclusión, nuestro trabajo nos sugiere que para obtener un modelo que pueda predecir con una buena probabilidad, debemos seleccionar una estrategia de entrenamiento de modelos que no utilicen días contiguos para la selección de datos y evitar así darle demasiada entidad a tópicos "cortos" tratados durante pocos días y para mantener a los modelos efectivos a lo largo del tiempo. Debemos ir reentrenando los modelos, incluso si estos no están prediciendo mal, para mantenerlos lo mas eficientes y más robustos posible ante la aparición de tópicos que se mantienen durante períodos más largos.

Como trabajo futuro se pueden seguir buscando nuevas estrategias para el entrenamiento y reentrenamiento de los modelos. En este trabajo, nosotros nos enfocamos en el ámbito de la política, en el cual los cambios de tópico son algo muy frecuente, motivo por

el cual consideramos que luego de un tiempo iban a ser necesarios reentrenamientos para mantener el modelo útil a lo largo del tiempo. Sin embargo, la polarización en las redes sociales es un fenómeno que afecta a otros ámbitos de la sociedad con menos alteraciones de la jerga, por lo que se podría analizar si los métodos utilizados en este trabajo logran una mayor precisión en estos contextos. Teniendo esto en cuenta, un trabajo futuro podría ser el análisis de la aplicación de los métodos que desarrollamos para este trabajo al ámbito del fútbol en torno a discusiones polarizadas como "quién es mejor: Messi o Cristiano Ronaldo, Messi o Maradona, Maradona o Pelé, Argentina o Brasil, etc."

Se puede seguir trabajando en los experimentos para encontrar si existe una cota óptima para el método 6.5.2 que maximice la calidad de la predicción. Podríamos analizar el uso de otros clasificadores de texto como por ejemplo BERT[59] que suele comportarse mejor al aumentar la cantidad de dimensiones utilizadas para los vectores de palabras o GPT[43]. En esta tesis utilizamos los algoritmos de clustering de Walktrap y Louvain, pero se podría comparar los resultados al utilizar otros como por ejemplo infomap[180], fast greedy[155] o label propagation[217]. Como ya se ha explicitado, experimentamos con la red social Twitter, se podría también analizar el comportamiento de las comunidades en distintas redes sociales, como Reddit o Facebook. Se podrían analizar también otros criterios para reentrenar los modelos, como por ejemplo, se podría estudiar qué palabras se utilizan más en una comunidad, cuándo aparecen y cuándo dejan de usarse para identificar cuándo conviene eliminarla de nuestros modelos predictivos. Se podrían analizar también los cambios de tópico durante determinado tiempo para identificar si hay alguna manera de detectar si un tópico nuevo va a introducir algún término nuevo a la jerga que va a perdurar a lo largo del tiempo o por el contrario, si va a ser algo pasajero.

7. LOS DIVERGENTES

Una incipiente línea de estudios ha mapeado la polarización en las redes en la medida que se han transformado en uno de los canales principales de participación en la región [134] y en particular ha mostrado el rápido proceso de configuración de comunidades ideológicamente polarizadas en el Brasil [161] y, en el caso argentino, la activación de la "grieta" en coyunturas críticas, como frente a la muerte del fiscal Alberto Nisman [1] y la desaparición y posterior muerte de Santiago Maldonado [2] [46]. En un marco más general, la polarización se expresa en temas de género, aborto, corrupción y estado, política internacional (en particular en torno a Venezuela y la "chavización") así como en relación a temas nacionales, por ejemplo, el proceso de paz en Colombia [74], delito, medio ambiente e incluso frente a las medidas contra el Covid19 como muestra el trabajo de Calvo y Aruguete [20]. No obstante, todavía no sabemos suficiente de las particularidades de la polarización en la región, ya que la literatura del tema proviene sobre todo de Estados Unidos, donde, entre otras diferencias centrales, hay identidades políticas estables (Demócratas y Republicanos) en torno de los cuales se configura el mapa de la polarización. Por lo demás, tampoco se ha indagado en los "no polarizados", sobre quienes se enfoca este trabajo.

Como se dijo, los estudios realizados en Estados Unidos en la medida que marcan la agenda sobre polarización son un contrapunto necesario a la hora de pensar las particularidades de los casos locales. En dicho país hay consenso en sostener un aumento de la polarización política en las últimas dos décadas [64, 71, 77]. Las y los especialistas han debatido si la polarización es un fenómeno sólo de las élites [78], o también del público [1]. Pero lo cierto es que a la mayoría de los norteamericanos de manera creciente les disgustan o desconfían de aquellos identificados con el partido opuesto (Demócratas vs Republicanos), aun si sus posiciones sobre los diversos tópicos no distan tanto [140]. Una serie de factores se han conjugado para exacerbar la proclividad de los identificados con cada partido para dividir el mundo en un valorado *"in-group"* y un despreciado *"ex-group"* [102]. Uno es el incremento del "*political sorting"* (identificación política) ligado a religión, raza y lugar de residencia (Mason, 2015). Otro es que la proliferación de medios hiperpartisanos ha reforzado identidades políticas y consecuentemente sentimientos exacerbados hacia los partidos políticos [128]. En efecto, cuando hay polarización afectiva es más probable evaluar a los miembros del *out-group* como más radicalizados e ideológicamente distante de lo que realmente son [129]. En esta misma dirección, [140] distingue entre la polarización de identidades políticas de la polarización en tópicos. Para la autora, el peso de las identidades políticas lleva a los individuos a adherir a partidos o líderes cuyos programas tienen ideas más extremas que las propias.

Ahora bien, sabemos mucho menos de quienes se ubican en una zona intermedia en

[1] Alberto Nisman era el Fiscal de la Unidad Amia que investigaba el atentado a la Mutual Judía de Buenos Aires en 1994. Apareció muerto en su casa en 2015 días antes de presentar un informe contra la Presidente Cristina Fernández de Kirchner. Su muerte es considerada como un suicidio para una parte de la sociedad y como un homicidio ligada al poder para otra.

[2] Santiago Maldonado era un joven argentino que en 2017 desaparece y luego es encontrado muerto después de la represión por parte de Gendarmería de una protesta indígena en la Patagonia. Su desaparición mantuvo en vilo al país durante meses y generó posiciones encontradas sobre el rol de Gendarmería. La justicia consideró que murió por hipotermia al caer a un río y cerró la causa, pero la familia apeló el fallo.

un proceso de polarización. No es sorprendente ya que las ciencias sociales siempre han tenido dificultades en tratar el *"entre deux"*. El caso paradigmático son los debates sobre las clases medias, que se han resistido a una definición consensuada, si la comparamos con los mayores acuerdos en torno a la definición de clase obrera o de clase alta (ver [53]). En relación con los no polarizados, lo cierto es que han recibido un menor interés, ya que la atención ha estado dirigida a los polos. El interrogante inicial es si los no polarizados son un grupo o sólo una categoría residual cuando se analizan procesos de polarización. En otras palabras, ¿tienen características comunes que permitan identificar sus atributos en la vida real, o simplemente son quienes sin ningún rasgo que los asemeje dejan de participar de una parte del juego social como quienes no se interesan por un deporte y sus rivalidades internas, sin por eso compartir rasgos en común entre ellos?

Este capítulo tiene también un objetivo metodológico para avanzar en el estudio de la polarización. Para ello se utilizan métodos mixtos en la fase del análisis de los datos. Los métodos mixtos son particularmente válidos cuando el objetivo es la complementariedad en el análisis y la posibilidad de expandir la compresión de un problema complejo [142]. Un primer punto es que la polarización es un proceso que se genera entre dos grupos en el marco de algún tipo de interacción, sea real y/o imaginaria. Una una limitación de la mayoría los estudios sobre polarización, es que suelen analizar encuestas de opinión, análisis de bases electorales y, en pocos casos, entrevistas cualitativas individuales; pero por lo general sin estudiar las interacciones. Consideramos que las interacciones son escenarios necesarios para observar la polarización, dado que allí se generan debates, controversias e intercambios entre personas con distintas posiciones que pueden, justamente, generar polarización.

El corpus consta de grupos focales que se tratan con métodos cualitativos y luego con un método computacional de medición de polarización semántica. En rigor, el trabajo es producto de la confluencia entre dicha investigación sobre noticias de delito en los 8 noticieros prime time de la Argentina [3] y una variante de la técnica descripta en 4. El método ha sido entrenado para captar polarización entre ambas comunidades en la medida en que se produzcan en ciertas controversias, más allá de la temática. Justamente parte del interés del mismo es que su aplicación permite observar la extensión de la polarización a temáticas que en principio podrían considerarse no intensamente politizadas, como el delito.

El aporte de este capítulo es el siguiente: luego de analizar la conformación de controversias en cuatro tópicos (consumo de medios, inseguridad, violencia de género y corrupción), encontramos casos de individuos que llamamos "divergentes". Definimos a un sujeto como divergente si (a) en las distintas controversias planteadas adopta en algunas posiciones de la Comunidad 1 y en otras de la Comunidad 2 y/o (b) en una misma controversia acuerda con tópicos de ambos grupos. Por lo tanto, los sujetos son considerados divergentes en relación con ambas comunidades. Una vez identificados mediante el análisis informático, encaramos el análisis cualitativo de las intervenciones e interacciones de cada uno de los sujetos divergentes en los grupos focales en los que participaron con el objetivo de elucidar si poseían atributos en común. Nuestra afirmación es que un rasgo común

[3] Proyecto de Investigación Orientado CONICET- Defensoría del Público: "De la propiedad a la recepción. Estudio integral del circuito productivo de las noticias sobre delito e inseguridad en los noticieros televisivos de mayor audiencia de la Argentina", dirigido por Gabriel Kessler. En el proyecto trabajamos con 12 grupos focales conformados por 10 personas que fueron distribuidas según edad, clase social y lugar de residencia. Por tal motivo, los criterios de delimitación de clase eran tanto por nivel socioeconómico como por lugar de residencia.

consiste en mantener una "distancia reflexiva" con los medios. Los divergentes valoran el pluralismo, consideran la realidad compleja y/o difícil de comprender y en consecuencia intentan formarse un juicio propio a partir de distintas fuentes [80]. En ocasiones, esta "distancia reflexiva" los hace desistir de adoptar posiciones muy definidas sobre un tema (*"uno nunca sabe…"*). No necesariamente todos los divergentes son muy politizados, en varios la distancia reflexiva con los medios va de la mano de una concentración en la esfera privada y poco interés por los asuntos públicos. Sin embargo, en todos los casos, esta posición hacia los medios es concomitante con la adhesión a tópicos de ambos campos a la hora de debates controversiales. Los convergentes, por su parte, seguían más las posturas de algún grupo y denotaban más exposición selectiva (*selective exposure*) con los medios afines [101].

Una pregunta central es si esta posición divergente es causa, consecuencia o resultado de un proceso de retroalimentación: es decir, si la distancia reflexiva explica la divergencia en los temas; si la divergencia en torno a los temas los lleva a no adherir a ningún medio en particular o, nos encontramos frente a un proceso de retroalimentación entre una y otra variable, sin que sea simple encontrar la génesis del proceso. Como veremos en las páginas siguientes, nos inclinamos por esta tercera posibilidad por una serie de razones que explicamos a lo largo del texto.

El recorrido de este capítulo, que como dijimos fue publicado en [117] junto a Gabriel Kessler (Universidad de San Martín), Brenda Focás (Universidad de San Martín) y Esteban Feuerstein (Universidad de Buenos Aires) , es el siguiente. Luego de explicar la metodología y los resultados generales de los grupos focales, se presenta el método informático de polarización semántica desarrollado por dos de los autores de este trabajo. Allí se incorpora la idea de divergencia y se observan los resultados para tres grupos focales. Luego se presentan las características generales de la divergencia y a continuación, se reconstruyen perfiles de tres tipos de divergentes. Esta tipología nos ayudará a la reflexión sobre la potencialidad de los divergentes para mitigar la polarización.

Para este análisis utilizamos una variación de la técnica presentada en el capítulo 4, que detallamos en la sección 7.2.

7.1. Debates y controversias sobre delitos en grupos focales

El acceso al consumo de medios es posible a partir de investigaciones empíricas en pequeña escala, y los resultados se analizan entendiendo las interpretaciones de los receptores dentro de un sistema sociocultural circundante [110]. En los estudios de recepción se utilizan distintas técnicas metodológicas, una de ellas es trabajar con grupos focales. Ellos permiten observar las interacciones grupales donde se negocian sentidos de lo recibido [16] y se presta para el análisis de recepción mejor que las encuestas o entrevistas individuales. Para esta investigación se conformaron, en cada lugar, grupos definidos con cierta homogeneidad de edad, clase y sexo, de modo de poder normalizar estas variables y estar atentos a la emergencia de otras que no son fácilmente previstas antes de realizar los grupos.

Para este capítulo nos centramos en 3 grupos focales realizados en la Ciudad Autónoma de Buenos Aires en septiembre de 2017, dado que eran los que más se adaptaban al lenguaje en el que está entrenado el método computacional, tal como se explicará en el próximo apartado. Uno de los grupos estaba conformado por personas de 18 a 30 años de nivel socioeconómico medio y medio-alto, que vivían en distintos barrios de CABA. Con la

excepción de los barrios del sur, que alberga a los sectores de menor nivel socio-económico y experimenta tasas de delito, particularmente homicidios, más elevadas. El segundo tenía las mismas características, pero su rango etario era de 40 a 65 años. El último se trataba de personas de 20 a 50 años, de un nivel socioeconómico medio bajo y bajo, de barrios del sur de la capital (Villa Lugano, Soldati, Barracas, Boca, Constitución, Balvanera). Cabe aclarar que la conformación de los grupos no tenía ninguna pauta respecto de ubicación política, dado que el objetivo buscado era analizar la recepción de delitos en general.

Para el análisis se elaboró un corpus con los noticieros transmitidos durante el horario central por los canales 11 y 13 emitidos desde la Ciudad Autónoma de Buenos Aires, 10 y 12 de Córdoba, 3 y 5 de Rosario y 9 y 7 de Mendoza, en el período de dos meses (mayo y junio de 2016). Dentro de los noticieros señalados, seleccionamos tres noticias de delito, violencia e inseguridad para exponer a los participantes. Se procuró que hubiera diversidad de temas vinculados con el delito y la inseguridad y de canales de televisión, para analizar si alguna de esas variables incidía en los modos de interpretación de las noticias.

En relación con el consumo de medios, las y los participantes mostraban cierta predilección sobre mirar/leer noticias de algunos medios en los que confiaban más, pero tampoco demasiado marcada, sino que solían consultar otros medios no necesariamente afines. También observamos que, en los grupos de sectores medios, tenía lugar un "efecto de la tercera persona", ya que en general había acuerdo en que los medios ejercían incidencia en la opinión pública, pero ellos/as por su mirada crítica y ciertas competencias culturales, quedarían exceptuados de esta influencia [4].

Como explicamos, los y las participantes miraron tres noticias durante los grupos focales. La primera era una noticia sobre un robo en un "supermercado chino"[5] en el Gran Buenos Aires, emitida por Telefé Noticias[6], donde un policía, que custodiaba el lugar, mataba al presunto ladrón. Lo particular de esa noticia es que aparecía el cuerpo del asesinado. La controversia en los grupos estaba dada en torno a quienes eran las víctimas y los victimarios. Para algunos/as las víctimas eran la cajera, los dueños del supermercado, y el custodio. Solo una minoría mencionó al supuesto delincuente como víctima. En relación con el victimario, la mayoría señaló a la persona que entró a robar. La muerte estaba también en el centro del debate ya que mientras para algunos/as "no era necesario matarlo" y mucho menos mostrarlo por televisión, otros/as opinaban que "si entró a robar podía morir" y que mostrarlo servía para que el resto aprenda.

La segunda noticia trataba sobre el caso de corrupción que tenía a Lázaro Báez[7], un empresario ligado a la familia Kirchner que es procesado por distintas causas de corrupción cuya causa es muy mediática, en el momento de la noticia se encontraba en prisión preventiva y era emitida por Telenoche[8]. Si bien era una noticia que a priori podría marcar

[4] Para Davinson (1983), el "efecto de tercera persona" tiene lugar cuando alguien tiende a creer que otras personas son más influenciables por los mensajes de los medios masivos de comunicación que ellos mismos. El razonamiento sería: "No seré influenciado, pero ellos (las terceras personas) pueden serlo".

[5] Robo y muerte en "supermercado chino" Canal: Telefé. Fecha de emisión: 02/08/2016.Duración: 4 minutos 59 segundos. Primer bloque

[6] https://telefenoticias.com.ar/actualidad/impactante-video-del-momento-en-que-un-policia-mato-a-un-ladron-en-un-supermercado-chino/ En el habla cotidiana en las grandes urbes y en particular en Buenos Aires se denominaen forma genérico "supermercado chino" a supermercados en barrios cuyos propietarios provienen de China.

[7] Lázaro Báez. Canal: Canal 13. Fecha de emisión 1/08/2016. Duración: 6 minutos 1 segundo. Primer bloque

[8] https://www.infobae.com/politica/2016/07/03/la-entrevista-completa-a-lazaro-baez-desde-la-carcel/ Para una semblanza de Lázaro Baez ver Salinas (2013)

un sesgo claro de polarización, lo cierto es que los y las participantes en general criticaron la cobertura del caso en tono sensacionalista y marcaban cierta distancia entre la noticia y su vida cotidiana. La polarización en este caso se centraba sobre todo en que algunos/as atribuían un fuerte peso de la corrupción en el gobierno de Néstor y Cristina F. de Kirchner mientras que otros/as al ponerlo en duda, lo equiparaban con la corrupción en el gobierno de Mauricio Macri, pero en todo caso, la controversia central era que se oponían a la asociación férrea o exclusiva entre kirchnerismo y corrupción.

La tercera noticia, emitida por canal 10 de Córdoba[9], abordaba un caso de femicidio (Carina Drigani)[10]. En este caso es donde hubo una mayor convergencia en torno a señalar a Carina como la víctima y su muerte como moralmente inaceptable. Sin embargo, notamos cierta divergencia en relación con el lugar de la mujer en la sociedad hoy y algunas polémicas en torno a los motivos que podrían haber terminado en el femicidio. Es decir, hay matices, incluso ciertos cuestionamientos, aunque es preponderante una mirada férreamente condenatoria.

7.2. Metodología

En esta sección detallaremos el procedimiento para cuantificar computacionalmente la polarización sobre los distintos tópicos discutidos en cada grupo focal. Como dijimos anteriormente, este procedimiento está basado en la técnica desarrollada en el capítulo 4 de esta tesis.

El método que utilizamos en este trabajo propone un enfoque lingüístico de la cuantificación de la controversia, a través de los cuales es posible establecer un score que represente el nivel de polarización en la discusión mediante el análisis del texto. Al estar basado en el lenguaje permite explorar otras aristas de la controversia, por ejemplo, qué usuarios expresan los discursos más polarizados o cuáles se encuentran más en las "fronteras" de cada comunidad. Por último, se pueden implementar en contextos distintos a las redes sociales como en este caso, que se usa para interacciones entre un grupo de personas, en la medida que hayan sido posibles de grabar y luego transcribir los intercambios verbales.

La técnica cuenta con 3 etapas que se resumen de la siguiente manera: primero se genera un modelo de procesamiento del lenguaje natural capaz de clasificar textos como oficialistas u opositores en Argentina y mapear texto a espacios vectoriales según ese mismo criterio. Es importante notar que los modelos de Fasttext (la herramienta que usamos para generar los modelo) [38] no son fácilmente interpretables, es decir que es muy difícil poder dilucidar si la diferencia de jerga entre comunidades que encuentra el modelo corresponde a modismos, palabras particulares, frecuencias de términos o algún otro concepto desconocido. Luego utilizamos este modelo para estimar los embeddings de cada intervención de los participantes del focus.

Finalmente utilizamos los embeddings para estimar las distancias semánticas de las intervenciones y de este modo el score de polaridad de los tópicos. De acuerdo a la hipótesis de este modelo, si estamos ante una conversación polarizada, tendremos dos grandes grupos representando las principales posturas que a su vez tendrán semánticas muy distintas (o completamente opuestas). Si esto sucede, tendríamos a nuestros embeddings agrupados

[9] Femicidio Carina Drigani. Canal 10 de Córdoba. Fecha de irradiación: 2/08/2016. Duración: 11 minutos 11 grupos segundos. Primer bloque

[10] https://www.lanacion.com.ar/seguridad/horror-en-cordoba-hallan-muerta-a-la-fisioterapeuta-que-estaba-desaparecida-nid1895965

principalmente en dos zonas distintas de nuestro espacio vectorial, lo que nos permitirá medir la distancia semántica de las comunidades. El score entonces lo definimos como la distancia entre los embeddings de las dos principales comunidades.

7.2.1. Generación del modelo

Como mencionamos en el capítulo 4, para entrenar al modelo que clasifique a cada texto como perteneciente al polo 1 o 2 es necesaria una gran cantidad de texto. Como no contamos con la cantidad de intervenciones suficientes en las desgrabaciones de los focus group, utilizamos discusiones en Twitter en torno a las menciones a @mauriciomacri sucedidas durante septiembre de 2017, mismo mes y año en el que se realizaron los grupos focales.

Entonces, el modelo lo entrenamos sobre dichos datos de Twitter aplicando los mismo pasos anteriormente mencionados: identificación de comunidades en el grafo de retweets y entranamiento de un modelo de Fasttext mediante los tweets de cada comunidad. El entrenamiento lo realizamos utilizando vectores pre-entrenados sobre Wikipedia en español.

Si bien en le técnica original entrenábamos al modelo usando los mismos datos sobre los que mediríamos la polarización, consideramos que las discusiones de twitter ocurridas en la misma fecha podrían tener una gran similaridad con las ocurridas en las desgrabaciones. Además, es importante destacar que el volumen de datos descargados supera el millón de tweets. Como se demostró en la sección 5.6 de esta tesis, estos modelos son eficaces en identificar a los usuarios pertenecientes a cada comunidad dentro de periodos de tiempo próximos a los datos de entrenamiento.

7.2.2. Estimación de los embeddings

Una vez entrenado el modelo lo utilizamos para mapear cada intervención (desgrabada) de los participantes del grupo focal a espacios vectoriales. Fasttext luego del entrenamiento genera un modelo que además de clasificar textos puede retornar los embedding mediante los cuales interpreta cada cadena de texto. Los mismos poseen una dimensión de 300 ya que es la dimensión con la cual se entrenó la red pre-entrenada sobre Wikipedia.

Para este experimento usamos los grupos focales hechos en CABA, ya que es el público que más se parece al usado para entrenar el modelo, dado que los usuarios de Twitter son sobre todo de niveles educativos medios o altos, un proxy de clase media. Sanitizamos las desgrabaciones quitando las aclaraciones hechas entre paréntesis (que indican acciones de los participantes), transformamos el texto a minúsculas, quitamos el nombre de cada participante y eliminamos las intervenciones de los moderadores.

Finalmente separamos las intervenciones según los 4 tópicos trabajados: Consumo de medios, Inseguridad, Femicidios y Corrupción y mapeamos los textos de cada uno de ellos a los correspondientes embeddings de 300 dimensiones calculados por Fasttext.

7.2.3. Cálculo del score

En el método original inferíamos la pertenencia de cada usuario a su comunidad a través del grafo de retweets utilizando Walktrap o Louvain. Sin embargo en este caso no tenemos dicho grafo por lo que decidimos inferir la comunidad de cada uno aplicando el algoritmo de clusterización Agglomerative Clustering [12] sobre los embeddings de las intervenciones. Agglomerative Clustering es un algoritmo de agrupamiento jerárquico que

comienza tratando cada punto de datos como un cluster individual. Luego, iterativamente, fusiona los clusters más cercanos en uno solo, basándose en la distancia entre ellos. Este proceso continúa hasta que se alcanza un número deseado de clusters o hasta que todos los puntos de datos se han fusionado en un solo cluster. Por lo tanto, nos permite agrupar vectores (en vez de nodos de un grafo) en dos conjuntos disjuntos de forma rápida y eficiente.

Finalmente, una vez identificados los embeddings de cada comunidad calculamos el score de polarización de la misma forma que en el capítulo 4.1.2.

7.3. *Score* de polarización e identificación de divergentes

Como mencionamos anteriormente, mediante este método decidimos medir la controversia en los tópicos discutidos en los grupos focales de CABA, estos son: Corrupción, Inseguridad, Consumo de medios y Femicidios. Para aplicar el método tomamos las intervenciones de los participantes de los tres grupos focales. Es decir, por cada uno agrupamos en un mismo archivo las intervenciones realizadas en los tres focus y sobre eso calculamos el score de controversia. Los resultados fueron los siguientes 7.1:

Corrupción	Inseguridad	Consumo de Medios	Femicidios
0.32	0.33	0.35	0.42

Tab. 7.1: Polarización por tema

La metodología utilizada en estos experimentos es análoga a la utilizada en experimentos previos 4, en este caso nuestro corpus lo constituyen las transcripciones de los grupos focales mientras que en los anteriores analizamos las discusiones en plataformas digitales como Twitter. En los experimentos anteriores realizados 4 el umbral de separación entre polarizado y no polarizado se encuentra entre 0.37 y 0.4. Por lo que el único tópico no polarizado sería Femicidios mientras que para los sí polarizados el más controvertido es Corrupción, seguido por Inseguridad y por último Consumo de medios. Puede observarse que los valores de los scores se encuentran bastante cerca, siendo la diferencia máxima de 0.1 entre Corrupción y Femicidios. En los experimentos realizados con otros tipos de controversias las diferencias de score entre tópicos, sobre todo entre los polarizados y no, fue mucho mayor. Por ejemplo, la diferencia entre debates altamente polarizados en la Cámara de Representantes de Estados Unidos, y eventos no polarizados como la celebración de cumpleaños de personajes del mundo del espectáculo alcanzó el valor de 0.8 4.

Es importante aclarar que, al ser esta una adaptación de la técnica 4 para ser implementada en un tipo diferente de corpus de datos (desgrabaciones de focus groups en lugar de publicaciones en Twitter), sería necesario aplicarla sobre un mayor número y diversidad de desgrabaciones para corroborar su robustez y obtener una mejor estimación de la significancia de las diferencias de puntuación obtenidas. Dada la dificultad de obtener este tipo de corpus de datos, en este capítulo nos limitamos al análisis de este caso y corroboraremos la eficacia de la técnica mediante análisis cualitativos realizados por los expertos en ciencias sociales que son coautores de este trabajo.

En el siguiente gráfico 7.1 puede observarse el posicionamiento de cada participante según su nivel de divergencia y cantidad de intervenciones, quienes tienen distintos símbolos según la cantidad de tópicos en los que intervinieron. Cada usuario tiene el sufijo _x en referencia al grupo focal en el que participó, esto es para poder diferenciar a participantes

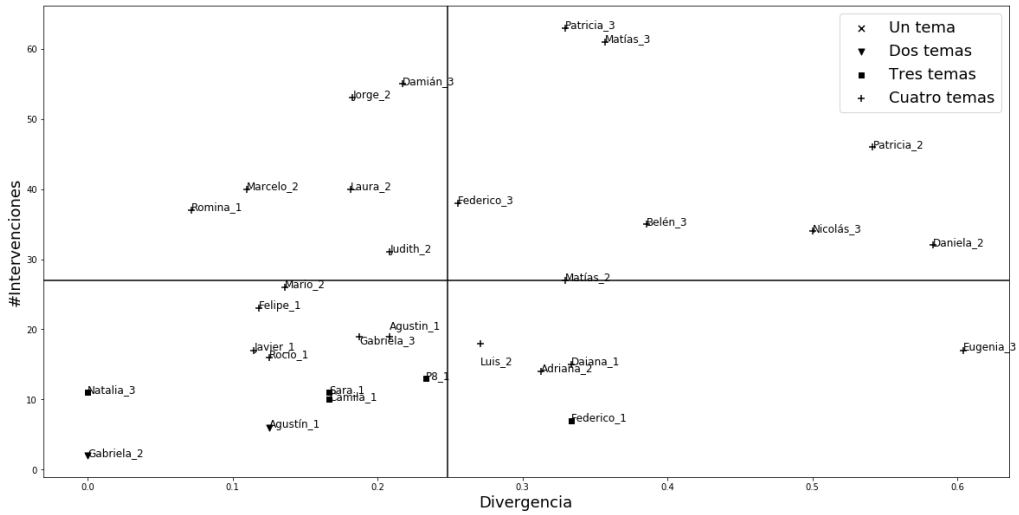


Fig. 7.1: En el eje Y marcamos la cantidad de intervenciones de cada participante y en el eje x su nivel de divergencia

con el mismo nombre de distintos grupos focales. Las rectas que salen de los ejes y y x están en los valores medios (de todos los participantes de los 3 grupos) de intervenciones y divergencia respectivamente. Mediante este gráfico puede caracterizarse a los participantes por áreas; así aquellos por sobre la recta del eje y son los más extrovertidos, mientras que los que están por debajo de ella intervinieron menos en las discusiones respecto a sus pares y, por lo tanto, los caracterizamos como más introvertidos. Por otro lado, los participantes a la derecha del eje x pueden considerarse los más divergentes, es decir aquellos que más pasaron en sus intervenciones de una comunidad a otra. Por último, también podríamos caracterizar a los participantes por cuadrante, donde el cuadrante superior derecho contiene a los participantes extrovertidos y divergentes y a los del cuadrante inferior izquierdo a los introvertidos y convergentes.

7.4. Análisis de embeddings

Para tener una mejor interpretación del cálculo de los scores de polarización decidimos analizar los embeddings de las intervenciones de cada tópico producidos por nuestro modelo. Para esto optamos por visualizarlos a través de una reducción dimensional mediante t-SNE [136]. t-SNE (t-distributed stochastic neighbor embedding) es un algoritmo de aprendizaje automático para visualización de datos. Se trata de una técnica de reducción de dimensionalidad no lineal muy adecuada para visualizar vectores de alta dimensión en un espacio de baja dimensión (de dos o tres dimensiones). Específicamente, modela cada vector de alta dimensión como puntos bi o tridimensionales de tal manera que vectores similares son modelados como puntos cercanos y vectores diferentes son modelados como puntos distantes con una alta probabilidad.

En los siguientes gráficos figs. 7.2 to 7.5 podemos ver los embeddings reducidos a 2 dimensiones mediante t-SNE de cada tópico. Cada embedding representa la intervención (desgrabada) de cada participante durante la discusión del tópico y está coloreado según la comunidad de pertenencia detectada.

Se puede notar que, aunque una comunidad sea mayoritaria que la otra en los casos

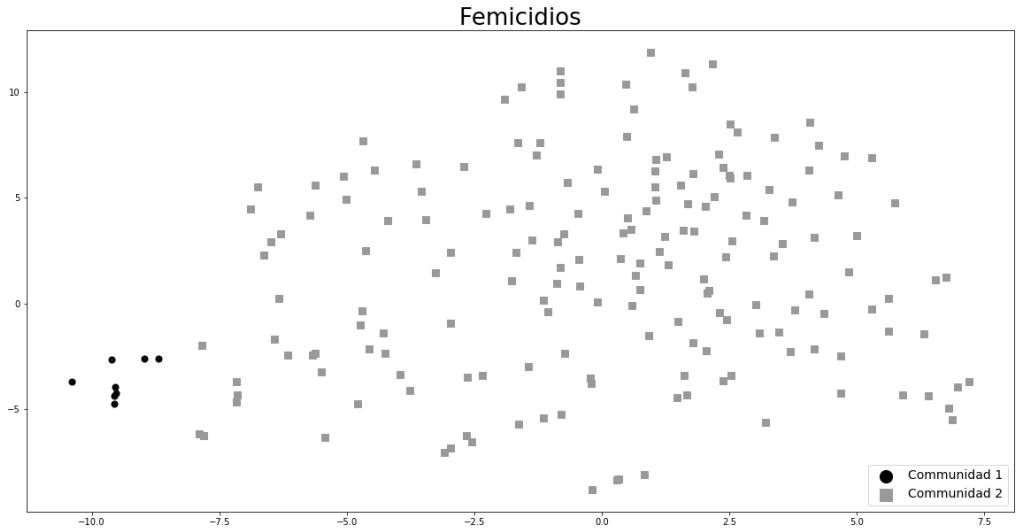


Fig. 7.2: Embeddings de las intervenciones sobre femicidios

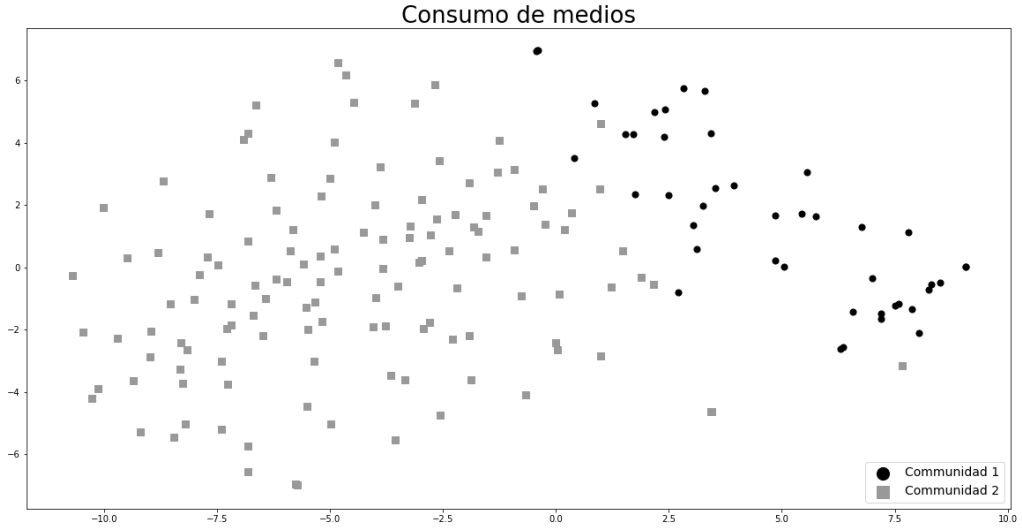


Fig. 7.3: Embeddings de las intervenciones sobre consumo de medios

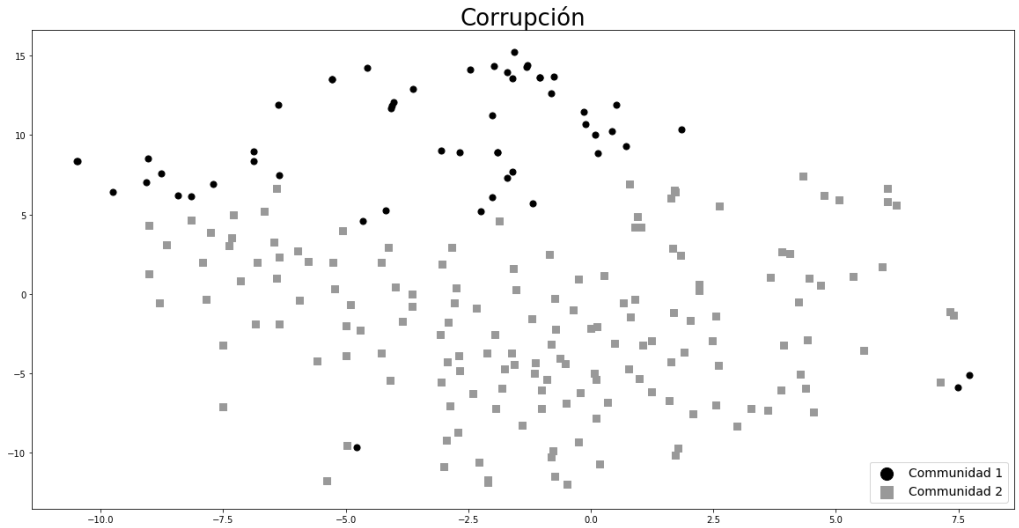


Fig. 7.4: Embeddings de las intervenciones sobre corrupción

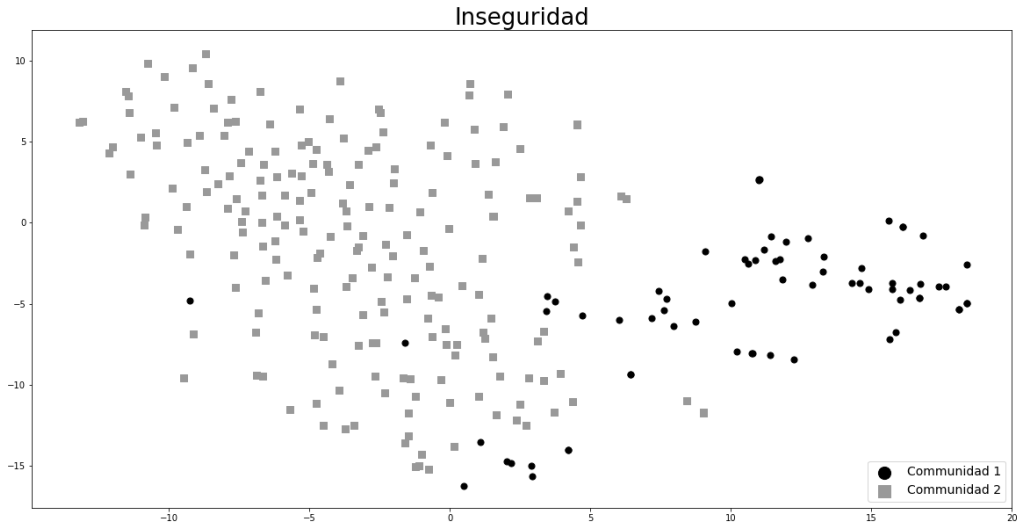


Fig. 7.5: Embeddings de las intervenciones sobre inseguridad

de Inseguridad, Corrupción y Consumo de medios, ambas presentan una cantidad significativa de puntos. Sin embargo, en el caso de Femicidios tan sólo 8 puntos corresponden a la Comunidad 2, mientras que las otras 180 intervenciones restantes pertenecen a la Comunidad 1. Este gran desbalance de magnitudes entre los tamaños de las comunidades, es uno de los factores que influyen en un score menos polarizado ya que D_1 (ver 4.3) será muy similar a D_tot debido a que se calculó casi con la misma cantidad de *embeddings*, por lo tanto r tendrá un valor mayor. Esta relación además va de la mano con la noción de que para que haya controversia en el debate, ambos polos deben tener un tamaño significativo, de lo contrario no sería polarización lo que sucede sino el desacuerdo de un pequeño grupo y una gran mayoría con la misma posición.

7.5. ¿Qué es ser divergente?

Una vez categorizados por el método computacional, nos centramos en las participaciones de los casos que se encontraban en el cuadrante de divergentes extrovertidos (ya que teníamos mayor número de intervenciones para reconstruir sus interacciones). Para ello reconstruimos en los grupos las distintas intervenciones de los ubicados en el polo divergente. El objetivo era determinar si tenían características en común. Como dijimos, el desafío era saber si eran más cercanos a un grupo que a una categoría sin existencia como realidad sociológica. Mediante la construcción de tipología inductiva quisimos observar los atributos en común, luego características secundarias que introdujeran matices y poder elaborar una tipología. El rasgo en común hallado era la actitud hacia los medios: les molesta, aburre, están en desacuerdo con la forma en que los medios se situaban en un panorama polarizado. A algunos la polarización los "aburre" porque es muy previsible lo que va a decir u opinar cada uno de los medios según su posición: lejos de favorecer la consonancia cognitiva, valoran lo novedoso en tanto que un medio lo sorprenda por su posición frente a un tema. Pero más en detalle, los motivos de la incomodidad eran tres, lo que nos permitió elaborar tres tipos de divergentes. El primero valoraba muy especialmente la objetividad, por lo cual les molestaba lo que llamaban "la falta de objetividad": cada medio o periodista teñía las noticias con su ideología volviendo a la información poco creíble y así "uno tiene que trabajar mucho" para formarse un juicio cabal de los sucesos. A diferencia de las hipótesis que la "cámara de eco" reduce la complejidad del mundo, porque nos guía acerca de lo que debemos opinar en cada tema, para estos divergentes el imperativo es lo contrario [67]: cada información está tan tamizada por la ideología que es preciso balancear entre distintas fuentes para llegar a formarse alguna idea de "la verdad".

En segundo lugar, a algunos/as divergentes les incomodaba el tono de burla, agresión o descalificación por parte de los periodistas hacia quien piensa distinto. De este modo, rechazaban el tipo de odio que genera la polarización afectiva. En otros divergentes, la polarización no incomodaba, sino que, por el contrario, se la valoraba al considerar que era lógico tener posturas distintas y hasta muy opuestas sobre un mismo tema, lo consideraban un indicador de "pluralismo" de la esfera pública. En este tipo algunos consideraban útil la pluralidad de opiniones desde un punto de vista pragmático: eran quienes hacían un uso pedagógico de los medios a la hora de acumular conocimientos para manejarse en la vida cotidiana: los medios actúan como "mapas cognitivos" para moverse en la ciudad, para saber qué hacer y, por ende, cuanto mayor pluralidad, más disponibilidad potencial de conocimientos diferentes. Así, las tres posiciones o subtipos que veremos en detalle en el apartado siguiente eran: valoración de la objetividad, disgusto con la polarización afectiva

y valoración del pluralismo por motivos expresivos o pragmáticos.

Como planteamos en la introducción, la pregunta es si este tipo de relación con los medios es causa, consecuencia o es un proceso de retroalimentación entre medios y opiniones. No podemos saberlo a ciencia cierta, pero nos inclinamos por la tercera posibilidad. Esto es, lo que vemos es una cierta "distancia reflexiva" con los medios: no es una crítica feroz, no es una situación de escepticismo extremo. De alguna manera es una vuelta a un público autónomo, que cree que la realidad es compleja, que la verdad existe, pero tampoco les desvela encontrarla. Consideran que los medios proponen un contrapunto de verdades absolutas y opuestas entre sí, un juego con el que no sienten afinidad, más como una actitud general sobre los medios que resultado de una evaluación pormenorizada de los argumentos de cada polo. En rigor, se puede ubicar a los participantes en general y a los divergentes en particular, en un continuo entre individuos más y menos politizados y es sobre todo para los divergentes más politizados muy importante la tarea de formarse un juicio propio a partir de distintas fuentes. Así, revisan distintos medios y hay un cierto orgullo de no ser *"víctimas"* de la exposición selectiva ni de la cámara de eco. Otros se caracterizan más por considerar que *"uno nunca sabe"* muy bien lo que pasa: la realidad es muy compleja y excede la posibilidad de un medio para captarla. Por último, el grupo menos politizado, consume pocos medios, en general pertenecen a sectores populares y están abrumados por una vida cotidiana compleja. Para ellos, la principal razón del distanciamiento es que los medios nunca hablan de lo que pasa cerca de uno. No se trata de un juicio novedoso, sino que es una crítica clásica de los sectores populares sobre las noticias (Kessler, 2009; Fonseca y Sandoval, 2006). En resumen, el juicio de varios de los divergentes es que los medios por lo general *"nos muestran una realidad que no existe."*

Ahora bien, más allá de esta distancia reflexiva ¿Hay alguna pauta en común en las opiniones concretas sobre los tópicos?, ¿existe un tercer polo discursivo, una "ancha avenida del medio"[11] con su propio relato? No pareciera ser el caso, los polos discursivos continúan siendo dos, pero los individuos los enhebran de modos muy distintos, al punto tal que puedan parecer contradictorios si se los evalúa teniendo como parámetros los relatos organizados en kirchnerismo-anti kirchnerismo. Al fin de cuentas, si se trata de una audiencia autónoma, es lógico formarse una opinión propia. A modo de ejemplo: en relación con la corrupción, pueden considerar que la noticia presentada está *"inflada"*, que no se sabe nada, que el gobierno de Macri es corrupto, pero al mismo tiempo detestar a Cristina Fernández de Kirchner. O decir que roban en ambos gobiernos. Alguien puede exculpar al gobierno kirchnerista, pero igualmente estar sumamente enojado y sentirse víctima en tanto miembros de la sociedad por los robos de la corrupción. En otros casos, en particular mujeres, pueden tener un compromiso muy fuerte con el femicidio como tema y al mismo tiempo celebrar la muerte del presunto ladrón en una línea de continuidad con una alta sensibilidad a la violencia. Algunos divergentes sobre todo expresan dudas, se afirman cuestiones tales como *"lo que veo son imágenes, no realidades"*; *"no están aclaradas, cuáles son las fuentes"*, *"habría que escuchar la otra campana"*. Como concluye Luis, 65 años, del barrio de Balvanera, *"cada uno tiene su versión de lo que ve, de lo que escucha, y cómo lo contás cambia también, cambia la visión del otro sobre la verdad, entonces "la" verdad es medio como subjetiva, ¿no?"*.

Por último, la pregunta es si tienen un rol activo de disminuir la polarización en las

[11] Nos referimos a la famosa frase de Sergio Massa de intentar captar a aquellos votantes que no se inclinaban ni por el kirchnerismo ni por el macrismo. Ver sobre la expresión señalada https://www.lanacion.com.ar/politica/sergio-massa-nid2011565

interacciones. En ciertas situaciones eran introvertidos y los más extrovertidos, solían expresar dudas, plantear la necesidad de escuchar otras versiones y sobre todo sostener que nada que provenga de un medio abiertamente polarizado o partisano puede ser *"toda la realidad"* pero veremos en el próximo apartado como cada tipo de divergente parece tener un rol disímil a la hora de mitigar la polarización. Un interrogante es si hay algún tipo de disposición del orden de la psicología política para explicar esta situación de distanciamiento. En general, la literatura anglosajona ha analizado diferencias cognitivas y emocionales entre conservadores y liberales. En un artículo acerca de por qué los conservadores son más felices que los liberales, Napier y Jost (2008) señalan que los liberales tienden a disfrutar de reflexiones extendidas y de este modo prolongan la clausura cognitiva, mientras que los conservadores suelen preferir respuestas relativamente simples, nada ambiguas sobre las cuestiones de la vida. Y por estas razones, los liberales podrían encontrarse menos satisfechos con la situación, debido al efecto deletéreo de la rumiación y la introspección. En dichos casos, entonces, se podría esperar que, las diferencias ideológicas en la necesidad de cognición, contribuirían a explicar la brecha en el bienestar subjetivo. La situación de los divergentes parece ser ligeramente diferente; no obstante, se requieren investigaciones adicionales para profundizar en este aspecto. Ciertamente se alejan de la clausura cognitiva de los conservadores, aunque no siempre están interesados en desplegar grandes disquisiciones. De todos modos, tienen una predisposición a la apertura cognitiva, más allá de que no siempre la emprendan activamente. Sin duda son aliados para mitigar la polarización, los más activos se oponían activamente a las miradas unilaterales, intentaban establecer matices, podían evitar el odio o la descalificación del otro, ya que no estaban inmersos en situaciones de polarización afectiva. Esto también llevaba a que el resto de los participantes, al no poder ubicarlos en el polo contrario, eran más propensos a escucharlos, a tomar en cuenta su posición y a no enojarse como lo harían con alguien muy polarizado.

7.6. Tipología de divergentes

A fin de comprender mejor el perfil de los divergentes, determinamos tres tipos paradigmáticos según la característica de su divergencia para desarrollar sus intervenciones y preguntarnos sobre su potencial rol para mitigar la polarización.

7.6.1. Valoración de la objetividad

"Tomar todo con pinzas": Matías tiene 42 años, vive en Barracas y es empleado técnico en YPF. El centro de su divergencia está en su posición respecto de los medios. En efecto, consulta distintos medios para hacer "su propio balance", aunque con mayor asiduidad medios kirchneristas o afines. El eje de su crítica es la falta de objetividad de muchos periodistas, quienes de manera subrepticia tratan de transmitir su ideología, pero sin explicitarla, parecería que esto lo enoja y también lo aburre:

Bueno, yo veo de todo. Yo llego a las 6 de la mañana, porque yo entro a trabajar a las 6 de la mañana, y 6 de la mañana ya me estoy fumando Infobae, TN, Página 12, o sea hago mi propia realidad. Pongo todo en la balanza, todo, hasta leo Telam, todo. Y después 6.30, 7 arranco (…) Yo veo TN, C5N, canal de Venezuela, todo, todo.

Su principal demanda es la "objetividad" en los periodistas.

El tema de que si un tipo es periodista tiene que presentar cierta imparcialidad, o sea si se presenta como periodista tiene que presentarse como, tiene que ser imparcial.

En cambio, vos ves periodistas de C5N que se caen de maduro que no son imparciales, ponele Navarro que ahora no está más. O Víctor Hugo, no son imparciales (…) pero si vos me decías mirá, yo soy, tengo esta preferencia política y aparte doy notas…o sea, sacá la conclusión porque me doy cuenta. Pero no blanquean su inclinación.

¿Qué sería ser objetivo?

Objetivo es decir, para mí es cuando agarran y dicen bueno, tengo información de que, pero cuál es la información, quién te dio esa información, cuál es la fuente

Su crítica está extendida a todos los medios y periodistas, por eso su actitud, que reitera varias veces en el grupo, es "tomar con pinzas".

Tomo con pinzas para dónde rumbea la noticia. De repente dicen no, porque se murió, no sé, tal y fue por culpa de Cristina, y ya cambió. No, fue por culpa de Macri que se murió, ya cambio. Cuando deja de ser serio…

Lo que pasa en Facebook yo también lo tomo con pinzas. …Dicen compartir que esta nena (desapareció de su casa), y capaz que estás viendo atrás hay unas letras árabes, esta nena es de Arabia, no es de acá. Entonces no sé el objetivo si hay algo marketinero…

Esta postura de diferenciar lo que muestran los medios o las imágenes con la realidad, es un hilo conductor en distintos temas y es un factor de divergencia en los tópicos. Por ejemplo, ante la pregunta de si aumentó la inseguridad, su afirmación es propia de la comunidad 1| (oficialista o kirchnerista) que no sabe *"si hay realmente más robos o es que hay más cámaras"* que filman, pero se aleja de dicha comunidad con sus intervenciones ante la noticia del femicidio. De ninguna manera deja de condenar el femicidio, pero se pregunta *"qué es lo que puede pensar un hombre para volverse violento"*. Esto suscita una reacción muy adversa de parte del grupo, lo que lo lleva a aclarar enfáticamente que de ningún modo estaba justificando o atenuando la responsabilidad, sino que *"intentaba ponerse en la cabeza"* de un hombre violento. Pero lo cierto que la forma en que interviene lo aleja de la posición mayoritaria de condena cerrada. Otra divergencia se observa frente a la imagen en los medios de un presunto delincuente asesinado por un policía de civil. Antes de esa noticia había afirmado que cuando ve una noticia de un robo violento en la televisión, siente lo siguiente:

Sí, que dios me perdone, que le vuelen la cabeza. Que lo revienten, yo me mato laburando que lo maten, eso me agarra.

Pero más adelante, ante la imagen del presunto ladrón asesinado, no duda en afirmar que el muerto es la víctima y qué no se sabe la verdad del hecho, ya que sólo por una imagen de una cámara de seguridad no podemos probar que la persona estaba cometiendo un delito. Finalmente, en la noticia de corrupción, por un lado, critica que la nota trata de inducir que Cristina está implicada caso Báez pero que nunca lo dice (Un juicio del grupo 1).

Lo que pasa que ahí el noticiero interpreta una frase que Lázaro dice bueno, todo se corta en Lázaro y no va más para arriba, interpreta que va para arriba está Cristina, más para arriba no sé, los empresarios…

Pero luego si se siente afectado por la corrupción y asume que es un tema existente e importante.

A mí sí me afecta (…) Y yo pago mis impuestos y quiero saber esa plata a dónde va. No es que pago los impuestos porque …pago el 21 % del IVA y quiero saber esa plata a dónde va. Si alguien se la robó quiero saber quién se la robó.

Finalmente, ante la pregunta si la noticia le genera interrogantes, culmina diciendo que *"te da interrogantes, pero interrogantes no de la noticia sino de toda la situación del*

país.", lo cual de algún modo resume su posición general frente a la relación entre medios y realidad.

En resumen, lo que él llama objetividad es el intento de restablecer confianza en un medio de referencia a partir de que aclaren las fuentes, esto lo lleva a un trabajo de cotejo y constatación. Matías tiene un rol activo en las discusiones polarizadas, en la medida que está bien informado con las fuentes de ambos polos, por lo cual, tienden a intervenir cuando consideran que alguien expone una posición muy alejada de la objetividad y en general su opinión se impone o no es rebatida. Es decir, hay posibilidad de escucha cuando a una postura polarizada no se opone la contraria en forma contundente, sino una apelación a la objetividad, la información fundada en datos y en evidencia.

7.6.2. Polarización es pluralismo

"Si todos dijeran lo mismo, sería totalmente sospechoso" sostiene Nicolás tiene 28 años, es docente, vive en Floresta y tiene algunas similitudes con el caso anterior en la relación con los medios, pero a diferencia de Matías no mira muchas noticias y no tiene televisor. Su característica principal es que aprueba que haya miradas diversas, le parece bien que no todos opinen lo mismo y por ello, no hay una crítica a los periodistas o canales de distintas tendencias. Es un defensor del pluralismo mediático y mantiene un cierto escepticismo, al decir que no está convencido por algún medio en particular, pero como se dijo, a diferencia de Matías, no dedica mucho tiempo a cotejar distintas fuentes.

A mí me parece bien. Si todos dijeran lo mismo sería totalmente sospechoso. No puede ser que todos los periodistas de un país, de los distintos estratos de agencias grandes de noticias, de diarios independientes digan exactamente lo mismo sobre algo…

Principalmente Facebook, links de noticias de (…), no busco mucho, pero porque también siento que buscar es como, es un trabajo demasiado arduo. A lo que voy es, es esa cosa que a donde vayas hay como tanta parcialidad que es…tipo nada me termina convenciendo ni de un lado ni del otro, entonces leer la misma nota de cinco diarios o de cinco lugares distintos para decir a ver, qué pasó. En casos de cosas importantes o que me llaman la atención a veces lo hago, pero no en todo.

También en relación con la noticia sobre Báez por un lado, defiende en cierta medida la inocencia de Néstor Kirchner pero también cree que el tema finalmente no se va a resolver nunca.

Una cuestión especulativa. Es como bueno, estamos diciendo esto, pero no sabés bien… esto va a terminar probablemente en la nada después de muchísimo tiempo. Probablemente sucedió y no se vaya a resolver. Como alguien se llevó un montón de guita…ya se la llevaron la guita, no la vas a encontrar probablemente porque sabemos cómo funciona el tema judicial y tiene una cosa de es como bueno, a mí me genera mucha sospecha cuando una sola noticia llena todos los medios. Es como, me están mostrando una es porque están pasando como cinco por atrás que no.

A diferencia del anterior, no cree ni le parece deseable la objetividad, sino que considera que no hay una realidad por fuera de las interpretaciones. En este caso el rol de mitigar la polarización es un poco menor que el anterior, sobre todo actúa en algunas controversias restándole gravedad a las posturas de los periodistas y los medios que podría denunciar, por ejemplo, Matías, al tener una visión positiva de lo que considera pluralismo. Su rol es más bien quitar un poco de tensión a las miradas que cuestionan a fondo la postura del otro, considerando que es un rasgo normal y hasta esperable en una escena mediática plural.

7.6.3. Disgusto emocional y distanciamiento

"Cosas que por ahí pasan en mi barrio y no lo veo ahí" sostiene Belén de 37 años frente a los delitos que presentan los noticiera. Ella vive en Villa Soldati, trabaja como empleada en una casa. No tiene mucha exposición a medios, pareciera si acceder al "consumo incidental" [148], en cuanto se entera lo que pasa por internet, ya sea por whatsapp o porque alguien comparte una noticia, pero sin ir activamente a la noticia, sino que la noticia llega a ella.

Ponele en Facebook, vos vas pasando las noticias y te aparece el video así, vos ya ves el video ahí y lo tenés ahí. No hace falta que vos entres, a veces te pasa directamente y vos ya lo estás viendo

Cuando pasa algo lo primero que es inmediato es Internet, uno agarra el teléfono y enseguida se entera lo que pasó antes que en la tele o en el noticiero.

Lo primero que agarrás es Internet. Claro, como el momento en que no estoy trabajando casi siempre tengo el teléfono por algo y bueno. Es interesante, uno está todo el tiempo con el teléfono por eso se entera de todo lo que pasa…

Uno de los rasgos de divergencia principal se basa en que le molesta el tono de burla o agresivo de los medios para quien no piensa igual.

A mí lo que me molesta es cuando veo canal 13 que por ahí ponen dibujitos burlándose de esa persona por más que digan es un corrupto, es una información que ellos tienen, que ellos dan y ponen un dibujito con la cara de esa persona o de un político para informar que para ellos es un corrupto o le encontraron cuentas que esto que el otro, se burlan de la persona. Eso no me gusta.

¿Por qué te molesta? Por la burla, porque es mucha agresión. Si no pensás igual que yo sos…

La segunda distancia con los medios, una crítica reiterada en sectores populares: los medios no muestran mi realidad, en el propio barrio pasan cosas graves (algunas similares a las que se ven en los noticieros y otras distintas), pero nunca aparecen en las noticias.

No se ve lo que pasa en el barrio de uno. Yo lo que veo que por ahí también veo en Facebook que hacen campañas entonces la foto se comparte, se comparte, y vos por ahí ves y por ahí no la viste en el noticiero la chica esa… O cosas que por ahí pasan en mi barrio y no lo veo ahí.

En relación con los tópicos, no interviene mucho o al menos no despliega mucho sus opiniones en varios temas, sí en relación con la violencia. Aquí podría encontrarse otra divergencia entre grupos, aunque desde la lógica de Belén hay una continuidad: al mismo tiempo que está muy comprometida con el tema de los femicidios y es donde interviene con mayor firmeza, luego celebra la muerte del presunto ladrón. En ambos casos el plexo convergente es una alta sensibilidad frente a la victimización: fue víctima de un robo, eso la dejó muy sensible al tema y también conoce muchos casos de violencia de género en su entorno.

(sobre la violencia de género) *gente se compromete más, antes uno miraba para otro lado, sabía que pasaba.*

(sobre la imagen del joven muerto) *Cuando es así lo podés mirar. Cuando es al revés que matan a un inocente te da…claro, o adelante cuando hay criaturas, cosas así por ahí sí te da cosa. A mí eso no me pone mal porque podría haber sido al revés y podría haber*

matado al policía (...) me da un poco de impresión obvio, pero al ver que el cayó no fue el policía o la chica…No me dio alegría pero…

A diferencia de los dos casos anteriores, Belén en ninguna de los temas controversiales tuvo un rol de mitigación de la polarización, si bien opinaba de casi todos los temas, su divergencia no la llevaba a un cuestionamiento de las posiciones más polarizadas.

7.7. Conclusiones

Este capítulo aporta a los estudios de medios y polarización desde el punto de vista metodológico y de los contenidos. En relación con lo primero presenta un método novedoso para detectar la polarización que puede utilizarse en debates y controversias, en este caso aplicado a grupos focales donde se discuten distintos temas vinculados con delitos. Las potencialidades de este método para el estudio de la polarización son muy amplias. En nuestro caso específico, nos permitió "seguir" a los individuos en situaciones durante las cuales se configuran y reconfiguran discusiones que pueden polarizarse, mientras que otras no. Así detectamos un grupo que llamamos divergentes cuyas características explicamos a lo largo del texto. Al volver al análisis de los individuos en los grupos, encontramos una serie de rasgos en común, el más importante es lo que llamamos distanciamiento reflexivo respecto de los medios en general y de la dinámica de polarización en particular. La divergencia era en relación con los dos grupos, pero sin establecer una tercera posición en las discusiones en términos sustantivos. En efecto, los divergentes toman elementos de ambos grupos y, sobre todo, expresan una posición de menor exposición selectiva y afinidad afectiva con algún medio o grupos de medios en particular. Ellos se ubican en un continuo de individuos más politizados o menos politizados: el distanciamiento puede llevar a intentar formarse una idea propia de cada tema luego de cotejar varios medios hasta un relativo desinterés por las noticias, dado que no reflejan la propia realidad ni sus preocupaciones y/o no tienen incentivos para invertir tiempo y recursos cognitivos en consultar una variedad de fuentes y "armar su propia noticia". Los tres tipos que hemos encontrados se diferenciaban tanto por aspectos específicos de su relación con los medios y también por una mayor o menor acción en favor de mitigar la polarización.

Este es un estudio exploratorio y deja planteados tanto interrogantes académicos como políticos para ulteriores trabajos. El primero, que ya planteamos, es si la distancia reflexiva es causa o consecuencia. Posiblemente sea un proceso de retroalimentación pero que se genera o se refuerza en un contexto de polarización mediática. En efecto, hay una reconfiguración propia del clásico distanciamiento entre representación mediática y realidad, tema ampliamente trabajado. Este se produce en un panorama polarizado frente al cual cada uno de modo más o menos activo se tiene que posicionar. Hay entonces un doble distanciamiento, de los medios como dispositivo y del discurso polarizado.

Para cerrar, una pregunta central es si los divergentes tienen un rol de atenuadores de la polarización en sus contextos específicos: familias, trabajo, grupos sociales varios. Cuando se producen controversias que tienden a la polarización: ¿acercan posiciones, disminuyen los juicios estereotipados de cada grupo y/o facilitan el debate abierto? Intentamos proporcionar una respuesta inicial a través de la tipología de los divergentes que esbozamos, por supuesto son necesarias investigaciones posteriores para validar y ahondar en esta dirección. No podemos por ahora dar una respuesta taxativa, pero la reconocida teoría de la "influencia minoritaria" [150] encuentra que un pequeño grupo puede generar en los otros pensar más creativamente, y/o abrirse a un debate más abierto, un proceso que se

ha llamado, justamente, pensamiento divergente [21].

8. APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS CROSS-PLATAFORMA DE LA COMUNICACIÓN POLÍTICA

En la era digital actual, las plataformas de redes sociales como Twitter, Instagram y Facebook se han convertido en arenas clave para el discurso político y social. Estos espacios digitales ofrecen a los políticos herramientas sin precedentes para comunicarse directamente con el electorado, difundir sus mensajes y fomentar el apoyo público. Sin embargo, esta revolución comunicativa no está exenta de desafíos. Uno de los más significativos es, la ya mencionada, polarización de la sociedad, fenómeno que se manifiesta en el entorno digital. La presente investigación se centra en analizar cómo el comportamiento de los políticos en estas plataformas podría contribuir a esta polarización.

El propósito de este capítulo es el desarrollo de técnicas computacionales que permitan explorar y comprender las dinámicas de interacción, discurso y representación que los políticos emplean en las redes. También, cómo estas prácticas pueden influir en la polarización social. A través de métodos computacionales avanzados, como el modelado de tópicos, la cuantificación de la toxicidad y la minería de texto, se busca examinar los patrones de comunicación y estrategias de discurso de las principales figuras políticas Argentinas durante el 2022.

Esta investigación es crucial no solo para comprender mejor el paisaje político, sino también para pensar mecanismos que puedan contrarrestar la tendencia hacia la polarización. El análisis de las estrategias comunicativas de los políticos en estas plataformas puede revelar *insights* valiosos sobre cómo se forman, mantienen o intensifican las divisiones sociales, y ofrecer perspectivas para fomentar un diálogo más inclusivo y menos polarizado en el espacio digital.

En este capítulo, se presentará inicialmente una revisión de la literatura relevante, que abarca el desarrollo de métodos computacionales para el análisis del discurso, el comportamiento político en línea y el impacto de las plataformas digitales en la dinámica social y política. Posteriormente, se plantearán las hipótesis a validar y se detallará las metodologías desarrolladas para analizar los datos recopilados de Twitter, Instagram y Facebook, seguida de una presentación y discusión de los resultados obtenidos. Luego, se reflexionará sobre las implicaciones de estos hallazgos para la sociedad, la política y el futuro de la comunicación en redes sociales.

Por último, cabe destacar que el trabajo presentado en este capítulo fue realizado por un equipo multidisciplinario[1] y publicado en un artículo conjunto en la revista Cuadernos.info [7].

8.1. Trabajos previos y enfoque

Nos encontramos en la fase que [164] ha llamado “New media, new politics 2.0”, cuyo comienzo se remonta a la campaña electoral de Estados Unidos 2008-2010 (respectivamente elección de B. Obama y de medio término) y cuyos rasgos novedosos serían el uso extendido

[1] Los autores de este capítulo junto con sus respectivas disciplinas son: Juan Manuel Ortiz de Zárate (Computación), Federico Albanese (Física y Computación), Gabriel Kessler (Sociología), Esteban Feuerstein (Computación). El aporte de Federico Albanese y del autor de esta Tesis son cuantitativamente similares

y sofisticado de tecnología digital, el manejo de distintas plataformas y el incremento de la interacción con los públicos, tanto como de los usuarios entre sí.

En la última década ha habido un interés por Twitter al ser un espacio de controversias políticas y por su accesibilidad para recolectar datos. En América Latina se han producido trabajos comparativos [135] así como estudios de caso en Argentina [47], Brasil [168], Chile [87], Colombia [70] y México [182], entre otros. Sin embargo, hay consenso acerca de que Twitter es un espacio restringido a los más interesados en la política, a los más polarizados y con más recursos culturales. Por tal motivo, se recomienda cautela a la hora de tomarlo como representativo de toda la esfera digital y, más aún, de la conversación política general y se sugiere expandir la mirada hacia las otras plataformas, si bien eso se dificulta por las restricciones en el acceso a los datos. El análisis cross-platform toma como objeto de estudio y unidad de análisis a cada usuario y las distintas redes sociales con las que interactúa en forma frecuente [179]. En particular, como sugiere dicho autor es preciso una mirada sobre por qué cada vez más las y los usuarios utilizan varias plataformas a la vez y aún más las personalidades públicas. Por lo pronto, existe un incipiente campo de estudios de comunicación política cross-plataforma de campañas electorales de los países centrales como Estados Unidos [40], Alemania [190], Noruega [69] y Suecia [127]. Estudios más conceptuales se preguntan cómo la lógica política influye en la arquitectura de los diferentes medios interactivos [48, 164] y otros comparan medios tradicionales con Twitter [113].

En términos metodológicos, estos trabajos han recurrido al análisis de metadatos (likes, retuits, etc.), en menor medida a técnicas cualitativas [188] y al análisis de discursos con aproximaciones novedosas [190]. Estos últimos revelan que los políticos y los partidos utilizan estrategias diferentes en cada plataforma y que todavía hay una predilección por Facebook a pesar del lugar central que tiene Twitter en los debates y polémicas. Asimismo, observan un creciente uso de Instagram (y, hasta hace pocos años, de Snapchat) aunque todavía hay pocos trabajos sobre Tik Tok. Los siguientes estudios subrayan la necesidad de innovaciones metodológicas para incrementar el alcance y la rigurosidad de los estudios [95, 131, 157, 164, 193]. En efecto, Por su parte, [190] señalan tres limitaciones de la mayoría de los trabajos sobre comunicación política en medios digitales. La primera es que en general se basan en períodos de campañas en los países centrales, pero pocos dan cuenta de la comunicación en tiempos ordinarios; la segunda es que suelen centrarse en una sola plataforma; y la tercera que analizan más metadatos que contenidos.

Nuestra investigación, realizada durante 2022, es un estudio pionero en la región y se propone superar estas limitaciones. La base de datos está conformada por las cuentas oficiales y públicas de 50 figuras políticas argentinas muy relevantes (en términos de cargos, responsabilidades o notoriedad)del Gobierno a nivel nacional (Frente de Todos), que llamaremos Oficialismo y de la oposición (Juntos por el Cambio) que impera en algunas provincias y las principales ciudades, durante 2020, un año sin elecciones nacionales (si bien el período considerado incluye, obviamente, la pandemia de Covid19). Nuestro marco teórico articula teorías de *agenda setting*, Sociología de problemas públicos y de encuadre o *framing*, como será desarrollado en el apartado siguiente. La metodología se enfoca tanto en las publicaciones como en los metadatos, ya que tomamos de cada político su presencia en tres plataformas y las estudiamos recurriendo a un abordaje computacional original. En efecto, la aplicación de métodos informáticos para el estudio de las ciencias sociales está demostrando tener un enorme potencial, tanto es así que las principales asociaciones de investigadores en ciencias sociales de nuestra región han creado grupos permanentes

para discutir su aplicación [17]. Esos abordajes han permitido trabajar con importantes corpus para el análisis, entre otros, de redes y de discursos de distinto tipo. Sin embargo, muchas de las herramientas más difundidas para el procesamiento del lenguaje natural resultan limitadas y poco eficientes, ya que para grandes volúmenes de datos son difíciles de configurar y muy costosas computacionalmente. Por ello, nuestro aporte principal es la utilización de modernas herramientas de Ciencia de Datos (Procesamiento de Lenguaje Natural [14], Aprendizaje Automático [35] y Análisis de Toxicidad [81] entre otras) para el estudio de discusiones en plataformas digitales pero que pueden ser utilizadas para distintos corpus. Las técnicas y métodos desarrollados en este trabajo conforman una caja de herramientas que estarán disponible en un repositorio público.

Nos planteamos el interrogante de si los políticos comunican de forma similar en las tres redes sociales que tienen mayor presencia en la política, o si, por el contrario, existen diferencias en la forma en que lo hacen en cada una de ellas. Nuestra hipótesis inicial era que Twitter debería exhibir características distintivas de las otras dos, por ser un espacio de interacción [90, 106], de hablar con otros; mientras que Facebook e Instagram tendrían más similitudes entre sí. Con esta idea, realizamos pruebas destinadas a encontrar similitudes y diferencias, por un lado entre plataformas y por el otro entre Oficialismo y Oposición en un país altamente polarizado. Hallamos que, efectivamente, Oficialismo y Oposición son más interpelativos en Twitter que en las otras dos redes, y que en esa red los políticos (independientemente del sector de pertenencia) tienden a discutir sobre temas comunes. En Twitter, además, encontramos una fuerte correlación entre el grado de toxicidad de los mensajes y su repercusión. Por el contrario, en las otras dos redes, Oficialismo y Oposición hablan principalmente de temas diferentes, sobre los que tienen más propiedad, y el nivel de toxicidad es bajo. También detectamos temas en común entre Oficialismo y Oposición en los que no hay confrontación.

El capítulo está organizado de la siguiente manera: en primer lugar se presentan las hipótesis y sus fundamentos, luego las pruebas realizadas para probar cada hipótesis con énfasis en el recorrido metodológico y finalmente las conclusiones del trabajo.

8.2. Marco teórico e Hipótesis

El contenido de esta sección está basado en la sección “Marco teórico e Hipótesis” del paper conjunto [7].

Como mencionamos anteriormente, nuestras hipótesis se basan en diversas teorías de la comunicación de larga data que han sido aplicadas al ámbito del debate político; los estudios de agenda *setting* (ver [18]), la teoría del *framing* o encuadre [185, 186] y los trabajos sobre propiedad (*ownership*) de la Sociología de los Problemas Públicos [92]. Esto lo aplicamos a las diferencias entre plataformas y presuponemos que los mensajes pueden diferenciarse: 1. en relación a la agenda, esto es hablar sobre tópicos distintos en cada plataforma, 2. en virtud del encuadre, o sea hablar de los mismos temas, pero enmarcados de manera distinta según la red y/o 3. en su dimensión interpelativa o vocativa, esto es, respecto al receptor al que irían dirigidos. Las opciones 1 y 2 serían mutuamente excluyentes, en cambio la dimensión 3 puede combinarse con la 1 o con la 2 (p.ej. puede mantenerse el tema y el encuadre, pero variar en una plataforma y en otro a quién estaría dirigido).

En cuanto a la variable de Oficialismo y Oposición, la teoría de la propiedad [114, 170] sostiene que los políticos deberían abordar temas en los que se sientan más cómodos. Tra-

dicionalmente, en Estados Unidos, los Demócratas han optado por hablar de integración racial y bienestar, mientras que los Republicanos han centrado sus discursos en crimen y seguridad nacional. En contraposición, otros argumentan que la propiedad no era una estrategia convincente para las audiencias y que era necesario "montar la ola" (*"to ride the wave"* [15], no escabulléndose de los temas del momento, so riesgo de ser considerado cínico o no sintonizar con las preocupaciones del público [103, 104]. Por su parte, de los trabajos que se preguntan sobre cómo gravita la arquitectura de cada plataforma tomamos lo que M. Bossetta [40] llama "estructura de la red", es decir, las normas técnicas que regulan la relación entre usuarios en cada plataforma. Presupusimos así que Twitter impulsa una conversación interpelativa de tipo "de uno al otro" (*to-each-other*) puesto que favorece la polémica entre usuarios con ideas diferentes ya que por defecto no se seleccionan a los seguidores; mientras en las otras dos, los seguidores suelen ser personas más afines y no conforman espacios habituales de controversias; son más propicias a una comunicación de tipo "cada uno sin escuchar al otro" (*past-each-other*): el emisor elige sobre qué temas publicar y puede orientar la agenda con menor injerencia de contrincantes. En ese sentido, basándonos en [112] suponemos que Oficialismo y Oposición tienen más probabilidad de hablar de los mismos temas (baja propiedad) en Twitter y de temas distintos (alta propiedad) en Facebook e Instagram. En otras palabras, conjeturamos que Oficialismo y Oposición eligen (o no les queda más opción) una red para debatir y la(s) otra(s) para promoverse en los temas que se consideran más fuertes. Pero también supusimos que en tiempos ordinarios como el que estudiamos (no de campaña electoral), el espacio político no es sólo de confrontación con el contrincante y de celebración de las propias acciones, sino que habría mensajes comunes tanto para el oficialismo como de la oposición en los que sea menos plausible la controversia.

En virtud de lo anterior, nuestras hipótesis al respecto son:

H1a Tópicos: Cada espacio elige Facebook e Instagram para hablar sobre los temas en los que tiene propiedad, mientras que Twitter se convierte en la plataforma en las que se debaten los temas sin propiedad exclusiva de uno u otro grupo.

H1b Temas en común: Los temas en común entre Oficialismo y Oposición no sólo incluyen confrontaciones sino también coincidencias o temas de baja conflictividad potencial.

Nuestra segunda hipótesis se vincula con las diferencias en el encuadre en las plataformas. Como sostuvimos, una opción sería que Oficialismo y Oposición hablaran de los mismos temas con un encuadre distinto, en particular una valoración diferente y a menudo opuesta. En este sentido, la Teoría de la Valoración dentro de los estudios de encuadre [139] se centra en los recursos lingüísticos por medio de los cuales los textos/hablantes llegan a expresar, negociar y naturalizar posiciones intersubjetivas y en última instancia, ideológicas. Esta teoría está atenta a la valoración, la actitud y la emoción de los discursos que denotan diferente posición del enunciador. Conjeturamos que una diferencia entre Oficialismo y Oposición será la valoración sobre los principales temas de agenda. Así, un mismo tópico tendrá una connotación positiva para unos y los otros lo criticarán, por lo cual cambiaría la valoración afectiva del mismo (por ejemplo, una acción de gobierno para el oficialismo de cada jurisdicción). Presuponemos que la negatividad estará sobre todo en Twitter, puesto que es la red de la polémica.

Por ello nuestra segunda hipótesis es:

H2a Sentimientos: Oficialismo y Oposición suelen enunciar mensajes con sentimiento (positividad/negatividad) distinto dependiendo de la red por la que se expresan.

H2b Negatividad en Twitter: Twitter es la plataforma donde hay mayor proporción de

mensajes que expresan sentimientos negativos debido a la mayor frecuencia de interacciones confrontativas.

Nuestra tercera hipótesis se vincula al hecho que en Facebook [2] e Instagram[3] los contenidos se muestran principalmente en base a las cuentas que el usuario sigue mientras que en Twitter[4] es en base a tópicos de interés. Esto promueve un mayor debate entre los usuarios, no sólo en círculos caracterizados por la homofilia [144] sino también por gente con otros puntos de vista. A partir de esto, nuestra tercer hipótesis es:

H3 Interpelación: Los políticos tienden a interpelarse entre sí más en Twitter que en Instagram y Facebook.

8.3. Metodologías y experimentos

En esta sección detallaremos las técnicas y métodos que aplicamos para testear nuestras hipótesis. Una herramienta muy utilizada para la detección de tópicos es Voyant-Tools[5], [79] que utiliza el algoritmo Latent Dirichlet Allocation [36]. Si bien es útil en ciertos corpus de datos, su rendimiento decrece al tratar con grandes volúmenes de datos poco estructurados, como las redes sociales. Otras técnicas populares para el análisis de sentimientos como *SentiStrength* [192] no tienen buen rendimiento fuera del idioma inglés [83] y por ello es preciso desarrollar métodos que permitan abordar los datos digitales de nuestra región, en español y portugués.

8.3.1. Construcción del Dataset

Argentina fue gobernada entre 2019 y 2023 por Alberto Fernández, quien resultó elegido en los comicios de 2019 junto a Cristina Fernández de Kirchner como Vicepresidenta encabezando el Frente de Todos, una alianza entre distintas corrientes del peronismo que venció al ex-Presidente Mauricio Macri quien buscaba su reelección con la coalición Juntos por el Cambio. Esta alianza se conforma por Propuesta Republicana (PRO), Unión Cívica Radical (UCR), Coalición Cívica ARI y Peronismo Republicano; lo que en este trabajo llamamos Oposición mientras que a los primeros Oficialismo. Para construir nuestro corpus seleccionamos 50 figuras políticas 25 del oficialismo y 25 de la oposición, de características lo más homogéneas posibles en ambos grupos en cuanto a cargos, responsabilidades o notoriedad, asegurándonos de que todos tuvieran cuentas oficiales en Facebook, Twitter e Instagram (ver Tabla 8.1 y 8.2para el detalle). Del oficialismo elegimos 12 personalidades con cargos en el Poder ejecutivo (principales Ministros y primera línea del Poder Ejecutivo Nacional), y 13 Senadores y Diputados de distintas provincias y con alta exposición pública. De la oposición seleccionamos 11 cargos ejecutivos de los cuales 7 son actuales (Intendentes de las principales urbes y Gobernadores) y 4 anteriores (Ex Presidente, ex Gobernadora de la Provincia de Buenos Aires, principal del país, Presidente del Pro y ex Ministra de Seguridad, ex Gobernador de la Provincia de Mendoza y Presidente de la UCR) y 14 Diputados y Senadores relevantes.

Luego, con las APIs de Twitter[6] y de CrowdTangle[7], descargamos todos los posteos

[2] `https://www.facebook.com/help/1155510281178725`
[3] `https://help.instagram.com/1986234648360433/?helpref=hc_fnav`
[4] `https://help.twitter.com/en/using-twitter/twitter-timeline`
[5] `https://voyant-tools.org/`
[6] `https://developer.twitter.com/en/docs/twitter-api`
[7] `https://www.crowdtangle.com/`

Tab. 8.1: Referentes políticos del Oficialismo seleccionados según sector y rol.

Alberto Fernández	Presidente
Cristina Fernández de Kirchner	Vicepresidenta
Santiago Cafiero	Jefe de Gabinete
Wado de Pedro	Ministro nacional
Gabriel Katopodis	Ministro nacional
Victoria Donda	Directora INADI
Axel Kicillof	Gobernador
Gildo Insfran	Gobernador
Gustavo Bordet	Gobernador
Juan Manzur	Gobernador
Omar Perotti	Gobernador
Sergio Uñac	Gobernador
Anabel F. Sagasti	Senadora
Oscar Parrilli	Senador
Facundo Moyano	Diputado
Fernanda Vallejos	Diputada
Gabriela Cerrutti	Diputada
Itai Hagman	Diputado
Jorge Antonio Romero	Diputado
José I de Mendiguren	Diputado
Jose L. Gioja	Diputado
Leonardo Grosso	Diputado
Lucia Corpacci	Diputada
Pablo Carro	Diputado
Pablo Yedlin	Diputado

que publicaron durante 2020 en las 3 plataformas, totalizando 150 cuentas (3 por cada figura política) y 84.435 posteos, de los cuales 56.622 son de Twitter, 16.133 de Facebook y 11.680 de Instagram. A pesar de que las imágenes son una componente importante del modo de comunicación en Instagram [27, 76], para este trabajo nos hemos limitado a analizar el texto de los posteos. Un primer hallazgo es que los políticos realizan más del doble de publicaciones en Twitter que en Facebook e Instagram juntos.

8.3.2. H1: Tópicos y Temas en común

En esta sección detallaremos los métodos implementados y los resultados obtenidos para validar la primer hipótesis: ¿Los políticos elijen FB e IG redes para hablar de su agenda y Twitter para los temas comunes? ¿Los temas comunes tienen coincidencias?¿Hay temas comunes sin conflictividad?

Método

Para probar la H1 precisábamos identificar temas propios y temas comunes de Oficialismo y Oposición en cada plataforma. Los algoritmos tradicionales para la detección y modelado de tópicos, como LDA [36], requieren que se les provea a priori la cantidad

Tab. 8.2: Referentes políticos de la oposición seleccionados según sector y rol.

Gerardo Morales	Gobernador
Rodolfo Suarez	Gobernador
Horacio R. Larreta	Jefe de Gobierno
Diego Santilli	Vicejefe de Gobierno
Gustavo Posse	Intendente
Jorge Macri	Intendente
Néstor Grindetti	Intendente
Mauricio Macri	Ex Presidente
Maria E. Vidal	Ex Gobernadora
Alfredo Cornejo	Ex Gobernador
Alfredo De Angeli	Senador
Humberto Schiavoni	Senador
Luis Naidenoff	Senador
Martin Lousteau	Senador
Alfredo Schiavoni	Diputado
Brenda Austin	Diputada
Cristian Ritondo	Diputado
Elisa Carrio	Diputada
Fernando Iglesias	Diputado
Graciela Ocaña	Diputada
Luis A. Juez	Diputado
Mario R. Negri	Diputado
Maximiliano Ferraro	Diputado
Waldo Wolff	Diputado
Patricia Bullrich	Presidenta del PRO

de tópicos en los que se quiere dividir el corpus, y por ello la coherencia de la división resultante depende de que dicho parámetro coincida con el real, lo que exige realizar pruebas con distintos parámetros hasta encontrar el valor correcto. Como nuestro dataset era voluminoso, era probable que la cantidad de temas discutidos fuera muy alta, y habríamos necesitado realizar numerosos intentos hasta llegar al valor correcto [181]. Por ese motivo recurrimos a una técnica más reciente, Top2Vec [14], que estima la cantidad de temas sin necesidad de validar previamente la coherencia de cada posible valor, reduciendo el tiempo de cómputo. Y, en efecto, según Top2Vec fueron 1028 los tópicos discutidos. Esta técnica, además, no precisa eliminar stopwords (artículos, preposiciones, etc.) ni normalizar el texto para su uso y permite identificar de forma determinística de qué tema habló un posteo dado[8]. Luego de identificar los temas, indagamos cuáles pertenecen a cada sector político. La categorización entre temas propios y comunes la definimos así:

- `Tema propio`: Tópico en el cual el 95 % o más de los posteos provienen del mismo sector político.
- `Tema común`: Tópico en el cual cada grupo produjo entre el 45 % y 55 % de los posteos.

[8] Mientras que LDA produce solamente un score para el cual luego es necesario definir un umbral a partir del que recién se puede decir que un posteo habló de un determinado tópico.

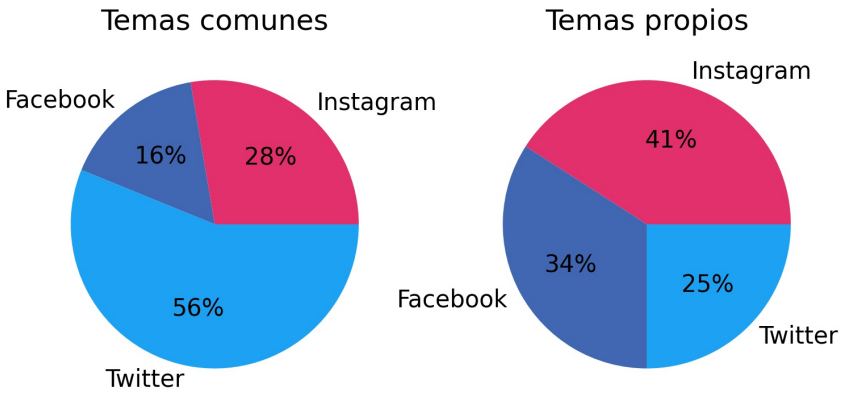


Fig. 8.1: Proporción de posteos por red social de tópicos que pertenecen a un único sector político (temas propios) y compartidos por ambos (temas comunes).

Resultados H1a (Tópicos)

Para verificar H1a medimos las proporciones de propios y comunes en cada red. Fue necesario normalizar la cantidad de posteos por político y por red social, ya que como dijimos los políticos realizan por lo general más posteos diarios en Twitter que en Instagram y Facebook. Los siguientes dos gráficos muestran dichas proporciones:

Twitter es la red más utilizada para los temas compartidos y Facebook e Instagram para los propios, y podemos confirmar que hay una agenda diferente en las plataformas, por lo cual se confirma la H1a.

Resultados H1b (Temas en común)

Analizamos las principales palabras y posteos de cada tópico por grupo:

Oficialismo

- `Defensa del río`: Campaña del oficialismo nacional referida a la discusión sobre que hacer con terrenos fiscales adyacentes al Río de la Plata.
- `Derechos de las mujeres`: Campañas del oficialismo.
- `Levantarnos`: Campaña del oficialismo tendiente a salir de la crisis económica y la pandemia.
- `Revolución de las viejas`: Campaña por los derechos de las mujeres mayores.
- `Campañas anti discriminación`: Campañas para luchar contra la xenofobia, machismo, homofobia, clasismo, entre otros.

Oposición

- `Cifras Covid19 San Isidro`: Informes sobre casos Covid19 en San Isidro, distrito gobernado por la oposición.
- `Informes Covid19 CABA`: Informes sobre casos Covid19 en Capital Federal, distrito gobernado por la oposición.
- `Voluntariado para cuidar a los mayores`: Programa del gobierno de la CABA para asistir a la gente mayor durante la cuarentena.
- `Encuentros virtuales con vecinos`: Actividades virtuales con vecinos de los distritos gobernados por la oposición.
- `Críticas al kirchnerismo`: Críticas al sector del oficialismo representado por Cristina Fernández de Kirchner.

Los temas son coincidentes con las agendas de cada sector en los medios de comunicación. En efecto, el oficialismo hace eje sobre campañas de gestión y cuestiones sociales o de derechos mientras que la oposición se refiere a la gestión en sus distritos y critica al oficialismo (focalizándose en el sector representado por la Vicepresidenta Cristina Fernández de Kirchner).

¿Cuáles son los temas en común?

Para verificar la H1b, nos propusimos ver qué tópicos tuvieron una participación similar por partido, lo que definimos como tópicos comunes y detectamos los siguientes temas, entre otros:

Temas comunes no controversiales

- `Saludo y reconocimiento a trabajadores`: Saludos a los bomberos, trabajadores de la salud y otros trabajadores en su día.
- `Condolencias por fallecimientos`: En ocasión de la muerte de figuras del campo político (p.ej un Juez federal, ex senador nacional o ex Gobernador de una provincia).
- `Aniversario Guerra de Malvinas`: En ocasión del aniversario de la Guerra de Malvinas contra el Reino Unido en 1982.
- `Cuidado de jubilados`: Mensajes de la importancia de cuidar a los jubilados en pandemia.
- `Aniversarios patrios`: Mensajes por los aniversarios patrios como el Día de la Independencia.

Temas comunes controversiales

- `Vacuna Sputnik`: Discusiones sobre dicha vacuna de origen ruso: el gobierno nacional posteaba sobre su compra y la oposición denunciaba que era de baja efectividad.

- `Menciones a Ginés`: Menciones Ginés Gonzalez García, ministro nacional de salud. Mientras que el oficialismo anunciaba actividades con él, la oposición lo criticaba por su gestión.

Confirmamos así que la hipótesis H1b se cumple, ya que la mayoría de los tópicos en común son no controversiales, con la excepción de la vacuna Sputnik y las menciones al Ministro de Salud.

8.3.3. H2: Sentimiento y Negatividad en Twitter

En esta sección detallaremos los métodos implementados y los resultados obtenidos para validar la segunda hipótesis: ¿Es Twitter la plataforma con sentimientos mas negativos?

Método H2a (Sentimiento)

Para testear la hipótesis H2a buscamos caracterizar la positividad y negatividad de los mensajes [3] mediante la aplicación sobre los posts de una red neuronal convolucional (LeCun et al., 1989) para el análisis de sentimiento[9]. Luego, con el test estadístico de Kolmogórov-Smirnov [97] intentamos detectar si había diferencias significativas entre la proporción de mensajes positivos y negativos de cada una de las redes.

Resultado

No hallamos disparidades importantes. Por lo tanto, no se verificó con este método nuestra hipótesis de que los políticos se expresan con positividad o negatividad distinta dependiendo de la red social.

Método H2b (Negatividad en Twitter)

Realizamos nuevos testeos con los novedosos desarrollos en torno a la toxicidad: un mensaje se considera tóxico si por su tenor rudo e irrespetuoso puede generar que el interlocutor abandone una conversación (Fortuna, 2020). Para medir la toxicidad usamos la API de Perspective [207] que utiliza redes neuronales profundas para el procesamiento del lenguaje natural pre entrenadas para dicha tarea. Este algoritmo le asigna a cada texto un valor entre 0 y 1, que representa la probabilidad de que el mensaje sea tóxico. Siguiendo la metodología utilizada por otros autores [99], definimos un valor de corte por encima del cual consideramos a un mensaje como tóxico.

Resultado

Al cuantificar la cantidad de mensajes tóxicos en cada red social, se observó que la proporción, si bien era chica en las tres redes sociales, en Twitter era considerablemente mayor siendo de un 7.6 % contra un 1,2 % en Instagram y un 0.4 % en Facebook. Ahora bien, ¿por qué los políticos tienen incentivos para publicar mensajes con mayor toxicidad en una red respecto de las otras dos? Descubrimos que en Twitter la mayor toxicidad se corresponde con una mucho mayor cantidad de likes, pero no sucede lo mismo en las otras redes. En la siguiente figura 8.2 se presenta el resultado en cada red social:

[9] La biblioteca de análisis de sentimiento en python: sentiment-spanish (https://pypi.org/project/sentiment-analysis-spanish/).

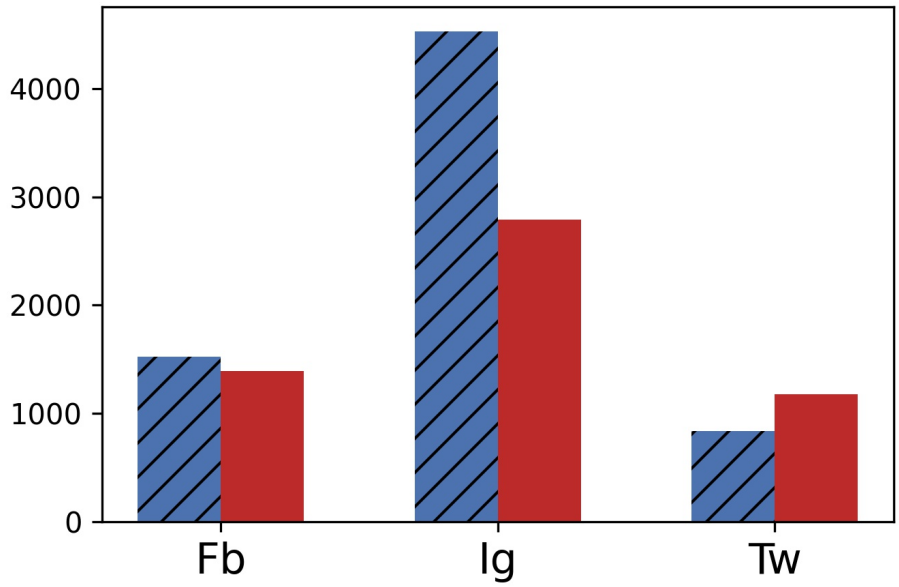


Fig. 8.2: Cantidad de likes promedio en mensajes s tóxicos (rojo) y no tóxicos. (azul) en cada plataforma.

Al realizar un test de Spearman [120], en Twitter observamos una correlación estadísticamente significativa (p valor menor a 0.05) y positiva entre la cantidad de likes y la toxicidad de un post con un coeficiente de correlación de 0.15.Por otro lado en Facebook dicho coeficiente es menor a la mitad, 0.08. En Instagram ni siquiera hay una correlación significativa entre toxicidad y likes. Concluimos que, tal como afirma la H2b, los políticos tendrían incentivos para ser tóxicos en Twitter, pero no tanto en Facebook y en Instagram, ya que en la primera se premia la toxicidad con una cantidad mucho mayor de likes.

8.3.4. H3: Interpelación

En esta sección detallaremos los métodos implementados y los resultados obtenidos para validar la tercer y última hipótesis: ¿Los políticos se interpelan más en Twitter que en IG o FB?

Método 1

Para testear nuestra tercera hipótesis hubo que realizar distintos pasos. Primero, comparamos la forma discursiva de los mensajes políticos para observar diferencias entre las plataformas. Apelamos a técnicas de procesamiento del lenguaje natural que miden la similaridad de los textos según la cantidad y significatividad de las palabras compartidas: dos textos se consideran más parecidos si comparten muchas palabras que no son muy comunes en el resto del corpus. Para ello vectorizamos el texto mediante la técnica de conteo de frecuencia de palabras: *Term frequency – Inverse document frequency* (Tf-idf) y medimos la similaridades de los mismos a través de la similaridad coseno [138]. Así encontramos que las cuentas del oficialismo y oposición en Twitter tenían, globalmente,

un gran parecido entre sí, mucho más que respecto a las otras dos redes. Para captar la particularidad de Twitter analizamos los patrones de distribución de las palabras a fin de descubrir cuáles se encontraban juntas con mayor frecuencia. A través de la técnica de descomposición en valores singulares (SVD) hallamos los principales grupos de palabras (dimensiones) y luego, entrenamos un árbol de decisión [35] para predecir a qué plataforma pertenecía cada usuario: un árbol de decisión entrenado para clasificar los textos según su pertenencia a Twitter, Facebook o Instagram, puede detectar si hay un grupo de palabras que se usa principalmente en una red social y no en las demás. El árbol fue entrenado sobre el 75 % de las cuentas seleccionadas al azar, dejando el 25 % restante para calcular su performance (*test set*). Cada cuenta fue representada por la concatenación de todos sus posteos.

Resultados 1

Respecto a la eficacia del modelo, de las 26 instancias de la clase 1 (FB/IG) se predijeron correctamente 24 y de las 12 de la clase 2 (Twitter), 10. De esta forma, la exactitud (accuracy) del modelo predictivo es de un 89.4 % y el área bajo la curva ROC [153] es 0.919. Luego nos enfocamos en la dimensión 50, aquella que más significativamente separaba y clasificaba los textos según la red social. Esta dimensión la llamamos *Interpelativa* debido a que las principales palabras más utilizadas son: usted, renuncia, saludos, buen día y espalda. Si bien se ven algunas palabras relacionadas a cuestiones o consignas coyunturales como "renuncia" o "espalda" nos resultó llamativo que la principal palabra sea "usted" ya que ésta puede denotar un diálogo de interpelación directa con otro usuario y presumiblemente otro político. A continuación se enumeran los posteos más importantes dentro de la dimensión

Interpelatival:

- "Usted, sí. `https://t.co/zjRiDBvEvE`" (FerIglesias)
- "@SolciPlata Usted, en cambio, sí." (FerIglesias)
- "@clarigv1 A usted" (WolffWaldo)
- "@Damian_Deglauve @WorldGrace saludos!" (gabicerru)
- "Soy yo la que lo quiere a usted, @caramellocumpa!!! `https://t.co/1qZjlEvADi`" (fvallejoss)
- "@shetpwk94 Que tengas un buen día Delfi!!! No salgas de tu casa !!! Cuídate mucho" (alferdez)
- "Si usted insistía en adjudicar esta compra con sobreprecios, hubiéramos realizado la denuncia al PAMI. Pero entendemos que ha procedido como corresponde." (gracielaocana)

En la lista se observa que todos los posteos son de diálogos directos, es decir quien los genera está interpelando a otro usuario, no necesariamente en un tono negativo (lo que concuerda con lo visto al intentar distinguir las redes mediante el análisis de sentimiento).

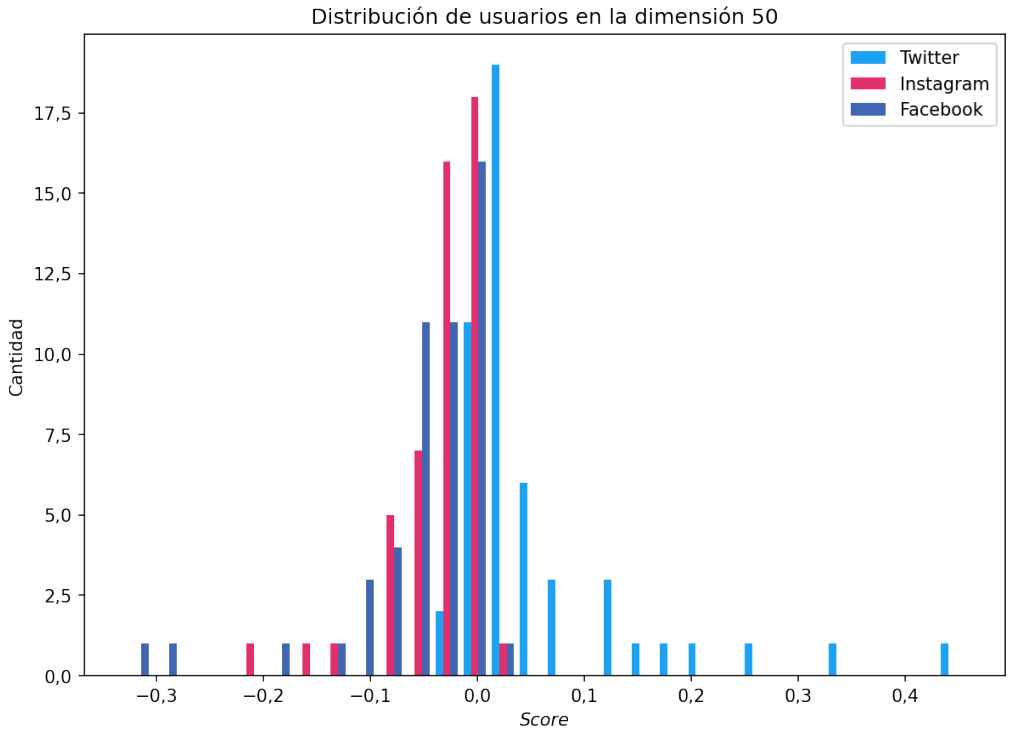


Fig. 8.3: Scores obtenidos sobre la dimensión Interpelativa por las cuentas. En el eje X se encuentran los *scores* y en el Y la cantidad de cuentas con dicho *score*.

La siguiente figura 8.3 es un histograma de los usuarios diferenciados por red social según cuánto usaban las palabras de dicha dimensión interpelativa (usando el *score* obtenido con SVD).

Se observa que la mayoría de las cuentas de Twitter están sobre la derecha del eje X, lo que significa que tienen una componente significativa en esta dimensión (usaron frecuentemente las palabras asociadas a la misma). Por otro lado, casi todas las barras azules y fucsias (cuentas de Facebook e Instagram respectivamente) están sobre la izquierda del eje X.

Es interesante notar que, en algunos casos, cuentas en distintas redes pertenecientes a una misma persona se encuentran en lugares opuestos respecto al valor 0 del eje X. Esto nos indica que esa persona tuvo una forma de comunicar distinta en Instagram y Twitter respecto a las palabras asociadas a dicha dimensión. Un caso destacado es el del Diputado de la oposición Fernando Iglesias, un vocero importante contra el gobierno, cuya cuenta de Twitter tuvo un score de 0.25 sobre el eje X mientras que en su cuenta de Instagram el *score* fue de casi -0.3, ocupando extremos opuestos del gráfico.

Método 2

A continuación, medimos la frecuencia relativa de cada término en cada red social. Para esto graficamos las palabras según la importancia que tienen en Twitter y en Facebook+Instagram (Ver 8.4), posicionando en el eje X el *score* en Twitter y en el eje Y el score en las otras dos. Agregamos una línea roja indicando equivalencia, es decir que aquellos términos posicionados en o cerca de ella tienen frecuencias parecidas en ambos

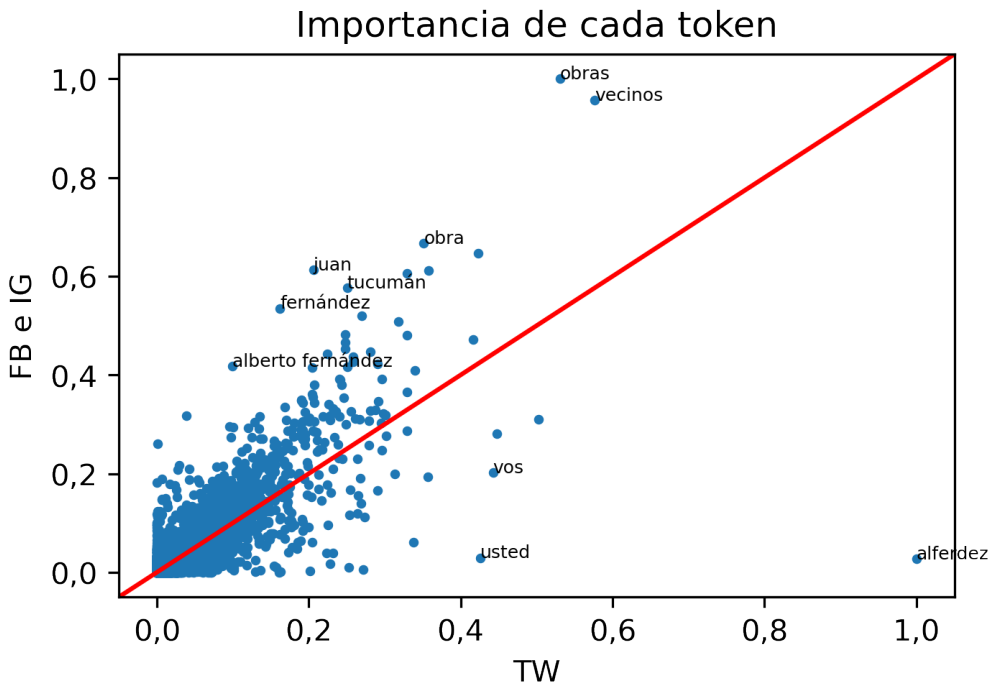


Fig. 8.4: Frecuencia de términos por red social.

casos.

Resultados 2

Se verifica que "usted" y "vos"[10] tienen una importancia muchísimo mayor en Twitter, con un *score* mayor a 0.4 en esa red y menor a 0.2 en Facebook e Instagram. El término más importante en Twitter, es "alferdez", que es nombre de la cuenta del Presidente Alberto Fernández. Vemos también que los términos referidos a gestión cómo "obras" y "vecinos" (muy usado por los gobiernos municipales) son importantes en FB e IG y no en Twitter, lo que sugeriría que esas plataformas son más elegidas para comunicar la gestión pública.

Método 3

Para seguir intentando validar la H3 (Twitter es el terreno para las interpelaciones) testeamos el uso de los distintos pronombres en las plataformas; en concreto la frecuencia normalizada de ciertas palabras interpelativas (la cantidad de veces que aparece dicha palabra dividida por la cantidad total de palabras usadas en esa red).

Resultados 3

En el gráfico 8.5 observamos que los pronombres en segunda persona como "vos" y "usted" son más usados en Twitter que en Facebook o Instagram y lo mismo sucede con expresiones dirigidas a otro interlocutor, como "buen día", "hola" o "saludos". En contraposición, el pronombre en primera persona "yo" aparece en mayor medida en Facebook

[10] En Argentina como segunda persona del singular se usa el "vos" en lugar del "tú".

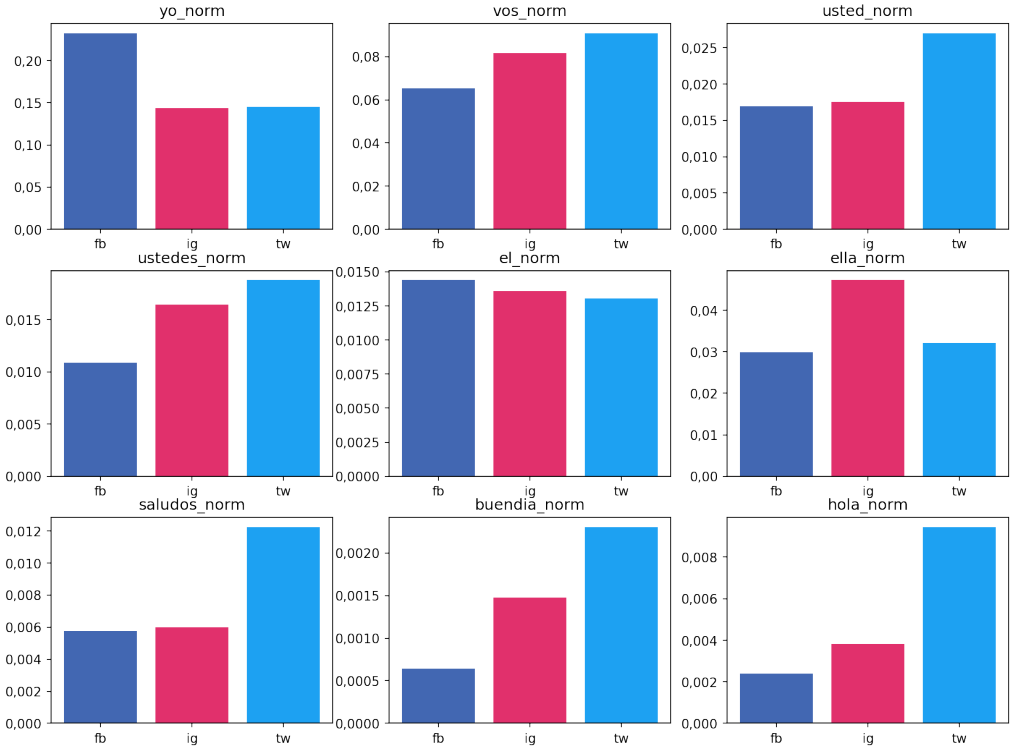


Fig. 8.5: Frecuencia normalizada de cada término en las tres redes sociales: Facebook (fb), Instagram (ig) y Twitter (tw).

que en Instagram o Twitter. También los pronombres de tercera persona son más usados en Facebook (él) e Instagram (ella) que en Twitter.

Estos resultados refuerzan la hipótesis 3 antes confirmada de que en Twitter hay un mayor diálogo entre usuarios, mientras que Facebook e Instagram son redes menos interpelativas y se usan más los pronombres personales de primera o tercera persona.

8.4. Conclusiones

En este capítulo, nos centramos en descifrar las similitudes y diferencias en la comunicación de 50 políticos argentinos del oficialismo y la oposición en Facebook, Instagram y Twitter durante 2020. Un hallazgo clave es que la preferencia de los políticos por abordar temas de su dominio y confort es más marcada en Facebook e Instagram que en Twitter. Este último, con su arquitectura propicia para la controversia, favorece la confrontación entre seguidores con distintas afinidades políticas, a diferencia de las otras plataformas donde prevalece la afinidad política. Este estudio revela una tendencia preocupante: en Twitter, los mensajes tóxicos no solo son más frecuentes, sino que también son los que generan mayor *engagement*. Este fenómeno subraya cómo la toxicidad en la comunicación política se ve recompensada en esta plataforma, exacerbando la polarización.

La atención exclusiva en Twitter puede llevar a una percepción conflictiva y superficial de la política, minimizando la importancia de las acciones políticas concretas. Sin embargo, identificamos también temas de baja conflictividad como rituales y celebraciones comunes, así como referencias a políticas implementadas, que transcienden las barreras partidistas.

Estos hallazgos subrayan la importancia de un enfoque cross-platforma para captar la complejidad y los matices de la comunicación política actual.

Este trabajo no solo aporta técnicas innovadoras para el análisis comparativo de textos en redes sociales, sino que también ofrece herramientas aplicables a un rango más amplio de medios y figuras públicas, en diferentes idiomas. Hemos hecho público todo el código en un repositorio para facilitar su uso y promover la investigación en comunicación y ciencias sociales mediante técnicas computacionales avanzadas.

Finalmente, este estudio contribuye significativamente al campo de la comunicación política, destacando la eficacia del enfoque cross-platforma. Esto no solo mejora nuestra comprensión de la comunicación política actual, sino que también informa la planificación estratégica de los políticos y otras figuras públicas. Al comprender mejor estas dinámicas, podemos abordar mejor la creciente polarización que caracteriza al discurso político en las redes sociales, buscando estrategias que promuevan un diálogo más constructivo y menos polarizado.

9. ORGANIZANDO REDDIT POR IDEOLOGÍA

Durante décadas, antes del surgimiento de las redes sociales mediadas tecnológicamente, ha existido un acalorado debate sobre la interacción de dos fuerzas duales en competencia en Internet: una de integración social, a medida que el mundo se ha vuelto cada vez más interconectado, y otra de fragmentación social, ya que las personas tienden a unirse a comunidades afines [191, 199, 201]. Hoy, 20 años después de la adopción masiva de las redes sociales y plataformas en línea, sigue sin estar claro cómo están organizadas socialmente las comunidades en línea. Un aspecto particularmente preocupante, como se mencionó en capítulos anteriores, es la aparente tendencia creciente de las poblaciones en línea a agruparse en *cámaras de eco* homogéneas. Además, existe una creciente inquietud sobre si las plataformas de redes sociales están impulsando a los usuarios hacia posturas ideológicas más extremas [24, 73]. Sin embargo, dado que estas plataformas consisten en grandes cantidades de datos no estructurados y anónimos, cuantificar empíricamente la composición social de las comunidades en línea y, a su vez, la organización social de las plataformas en línea, plantea un desafío enorme.

En este capítulo, proponemos una técnica para cuantificar la posición en espacios ideológicos basada en el texto publicado por cada comunidad. Esta técnica se basa en la hipótesis mencionada al comienzo de esta tesis 1.4, donde señalamos que a polarización influye en el lenguaje, generando diferentes variantes del mismo idioma según el grupo ideológico al que se pertenezca. De manera similar al capítulo sobre técnicas para cuantificar la polarización 2, el enfoque de este método se alinea con ciertos aspectos de investigaciones previas [204], las cuales se basan únicamente en las interacciones y no en el texto. Una vez más, destacamos que la utilización del texto en lugar de las interacciones amplía el alcance del análisis al permitir la incorporación de diversas fuentes de datos, que incluyen múltiples plataformas sociales como Facebook y Twitter, así como periódicos, blogs, contenido generado por usuarios y otros. Además habilita un análisis semántico de la diferencias ideólogicas entre las comunidades lo que puede enriquecer la comprensión en este sentido.

Utilizamos el mismo conjunto de datos que [204], que ofrece una cantidad sustancial de texto y sirve como una línea de base valiosa para este capítulo. En primer lugar, recopilamos el texto de las publicaciones y las agrupamos por comunidad y año, abarcando desde 2012 hasta 2018. A continuación, aplicamos diversas técnicas de vectorización para estimar los *embeddings* de la comunidad, incluyendo modelos basados en el modelo skip-gram [39] y otros más complejos basados en transformadores [202]. Finalmente, calculamos las dimensiones sociales utilizando la metodología propuesta por Waller et al. [204]. Este proceso implica que el analista seleccione dos comunidades como semillas, determine la dirección entre estos dos vectores de semillas y, posteriormente, proyecte los *embeddings* de las comunidades restantes en estas dimensiones. Para mejorar la robustez, se utiliza la ampliación de semillas (para obtener detalles adicionales, consulte la Sección 9.2)[1].

Como se demuestra en la Subsección 9.3.2, nuestros resultados obtenidos exhibieron un alto grado de similitud con los obtenidos por Waller et al. [204], lo que significa que las interacciones entre usuarios y comunidades tienen una correlación con el lenguaje. Además, nos permite crear un nuevo tipo de dimensión basada en el texto en lugar de las comunidades semilla y utilizar cualquier conjunto de textos en lugar de comunidades

[1] Todo el código y los datos están disponibles en `https://github.com/fddemarco/BIICC-2023`

de publicaciones. Además, una observación destacada fue el rendimiento constante de los *embeddings* basados en transformadores en contraste con los *embeddings* basados en skip-gram. Esta ventaja se mantuvo evidente incluso cuando entrenamos nuestro modelo skip-gram en el conjunto de datos específico y utilizamos vectores pre-entrenados.

Este capítulo, que fue publicado en [58] junto a Franco Demarco (Universidad de Buenos Aires) y Esteban Feuerstein (Universidad de Buenos Aires), está organizado de la siguiente manera: en la Sección 9.1, resumimos trabajos previos sobre el análisis de redes sociales a través de la jerga y las interacciones. La Sección 9.2 contiene la descripción paso a paso de nuestro proceso, junto con la introducción de dos nuevas variantes naturales de la medida de similitud utilizada. En la Sección 9.3 describimos los conjuntos de datos recopilados para este estudio y presentamos los resultados obtenidos. Finalmente, concluimos con la Sección 9.4.

9.1. Trabajos relacionados

La investigación realizada por Waller et al. [204] introduce una técnica novedosa para cuantificar la posición de las comunidades en línea a lo largo de dimensiones sociales, basándose en las interacciones de los usuarios. Al aprovechar los registros históricos completos de publicaciones y comentarios en Reddit desde 2012 hasta 2018, los investigadores generan representaciones de las comunidades a partir de estas interacciones. Luego proyectan estas representaciones en ejes unidimensionales que simbolizan una *dimensión social*. Este proceso produce puntajes para cada comunidad, situándolas efectivamente en el espectro correspondiente de la dimensión. Esta metodología produce resultados que se alinean coherentemente con percepciones cualitativas.

Por otro lado, muchos trabajos recientes han demostrado una correlación significativa entre la jerga y las discusiones de las comunidades. Ramponi et al. [62, 176] construyen clasificadores y predictores muy eficientes de la pertenencia de una cuenta a una comunidad dada al examinar el vocabulario utilizado en los tweets de diversas comunidades heterogéneas de Twitter, como jugadores de ajedrez, diseñadores de moda y partidarios de partidos políticos. En [194], Tran et al. encontraron que el estilo de lenguaje, caracterizado utilizando un modelo de lenguaje híbrido de *n-gram* de palabras y etiquetas de partes del discurso, es un mejor indicador de la identidad de la comunidad que el tema, incluso para comunidades organizadas en torno a temas específicos. Lahoti et al. [126] modelan el problema de aprender el espacio de ideología liberal-conservadora de los usuarios de redes sociales y fuentes de medios como un problema de factorización de matrices no negativas restringidas. Validan su modelo y solución en un conjunto de datos de Twitter del mundo real. Además, en los capítulos chapters 3 and 4 hemos mostrado que podemos medir el nivel de controversia en una discusión a través de los textos publicados por las comunidades.

Finalmente, el artículo titulado 'No hablamos el mismo idioma: interpretación de la polarización a través de la traducción automática'[118] examina la creciente polarización observada entre partidos políticos, medios de comunicación y élites en los Estados Unidos, con un énfasis particular en las redes sociales. El estudio se centra en cómo diferentes comunidades perciben y utilizan el lenguaje de maneras distintas, sugiriendo que estas comunidades básicamente están *hablando diferentes idiomas*. Para abordar este fenómeno, los autores introducen un método novedoso que utiliza la traducción automática como herramienta analítica. La idea central es que cuando dos comunidades utilizan el lenguaje de

manera significativamente diferente, las técnicas de traducción automática pueden identificar y traducir estas diferencias, ofreciendo conocimientos únicos sobre la polarización del lenguaje. Ese trabajo destaca el papel crucial del lenguaje en la polarización y proporciona una herramienta innovadora para analizar y comprender este fenómeno a un nivel más detallado. Al utilizar la traducción automática, tradicionalmente empleada para convertir un idioma a otro, el estudio profundiza en las distinciones lingüísticas intrínsecas entre las comunidades polarizadas, ofreciendo una perspectiva fresca sobre cómo el lenguaje refleja y amplifica las divisiones sociales y políticas.

9.2. Metodología

Nuestra contribución metodológica consiste en la introducción de un enfoque novedoso para cuantificar la organización social a través de datos textuales. Nuestra hipótesis es que la jerga, los temas, el lenguaje y las formas discursivas utilizadas por cada comunidad ofrecen información valiosa sobre sus aspectos ideológicos, especialmente los políticos, al igual que sus interacciones.

Inicialmente, delineamos el algoritmo general para construir dimensiones sociales. Luego, detallamos las elecciones específicas que hicimos durante nuestros análisis. Finalmente, desarrollamos el cálculo de los puntajes de las comunidades y su validación frente a los hallazgos previos presentados en [204].

9.2.1. Generación de *embeddings*

Utilizamos los conjuntos de datos presentados en la Subsección 9.3.1 para representar las comunidades de Reddit, conocidas como *subreddits*, en un espacio de jerga. Para garantizar representaciones vectoriales significativas, eliminamos los subreddits extremadamente pequeños con un número insuficiente de publicaciones. Por lo tanto, nuestro análisis se limita a los 10.000 subreddits principales, clasificados por el número de envíos.

Para generar *embeddings* de palabras para cada comunidad, seleccionamos dos modelos entre los más avanzados, a saber, *FastText* [39] y el modelo basado en transformadores de *Cohere* [202]. Estos modelos incrustan textos en vectores de dimensión fija que codifican significado semántico.

Ambos modelos de lenguaje, ya explicados en 1.3, toman como entrada un corpus de texto único y devuelven una única representación vectorial. Por lo tanto, para generar un *embedding* que caracterice a cada comunidad, creamos un corpus de texto unificado concatenando todo el contenido textual de las publicaciones dentro del subreddit correspondiente, incluyendo tanto los títulos como las publicaciones originales.

Fasttext

FastText tiene varios hiperparámetros que impactan el proceso de entrenamiento y los embeddings resultantes. Estos hiperparámetros incluyen la tasa de aprendizaje, el tamaño de los vectores de palabras, el tamaño de la ventana de contexto, el número de *epochs* y otros. Decidimos usar los valores predeterminados para todos estos parámetros, excepto para el tamaño de los vectores de palabras y el número de *epochs*. Específicamente, establecimos el tamaño de los vectores de palabras en 300 dimensiones, coincidiendo con el tamaño de vector de los vectores preentrenados de wiki-en[2]. Además, al usar el *Conjunto*

[2] https://fasttext.cc/docs/en/pretrained-vectors.html

de datos completo para el entrenamiento, elegimos establecer el número de *epochs* en 1 debido a una limitación en nuestra infraestructura. Para obtener información adicional sobre los conjuntos de datos de entrada y los hiperparámetros utilizados, por favor consulte Subsección 9.3.2.

Cohere

Respecto a Cohere[3], vale aclarar que ofrece una API para integrar procesamiento de lenguaje de vanguardia en cualquier sistema. Cohere entrena modelos de lenguaje masivos y los pone a disposición a través de una API fácil de usar. La plataforma proporciona una variedad de modelos que cubren diversos casos de uso, incluidos los modelos de representación que pueden generar *embeddings* de texto. Entre los modelos de representación ofrecidos por la Plataforma Cohere, elegimos utilizar el modelo basado en transformadores *embed-english-v2.0* [202].

Es importante destacar que este modelo específico está limitado por su dependencia del idioma inglés y carece de funcionalidad confiable para idiomas distintos al inglés. Dado que nuestras comunidades objetivo consisten principalmente en hablantes de inglés, consideramos que esta limitación es inconsecuente para el método desarrollado en este capítulo. Otra restricción impuesta por este modelo es la limitación de 512 tokens por texto, con cada token correspondiendo generalmente a 2-3 caracteres[4]. Para abordar esta limitación de tokens, proponemos reducir la cantidad de datos proporcionados al modelo. Al seleccionar publicaciones altamente relevantes, podemos obtener un conjunto de datos más compacto que sirve como una muestra suficientemente representativa para cada comunidad (consulte la Subsección 9.3.1 para obtener más detalles). Sin embargo, reconocemos que el uso de solo 512 tokens (aproximadamente 200 palabras) puede no proporcionar una muestra completamente representativa. Por lo tanto, reducir solo los datos de entrada no es una solución completa y debe ser revisado en trabajos futuros. Para obtener más información sobre nuestra visión de cómo abordar completamente esta limitación, consulte la Sección 9.4. Dado que estamos utilizando un modelo preentrenado sin llevar a cabo la afinación de hiperparámetros, el uso de este modelo no implica especificar ningún hiperparámetro.

9.2.2. Puntajes de comunidades

Para cada año, generamos *embeddings* exclusivamente a partir de las publicaciones de ese año en particular. A continuación, calculamos puntajes para todas las 10.000 comunidades utilizando la técnica de proyección descrita en [204]. Para ejecutar esta técnica, el analista identifica inicialmente un par de comunidades semilla que varían exclusivamente en términos del constructo objetivo. En nuestro estudio, utilizamos *r/democrats* y *r/Conservative* de acuerdo con [204]. Posteriormente, expandimos el par de semillas inicial para abarcar hasta 10 pares, y las diferencias de vectores resultantes se promedian para obtener un solo vector. Esto produce un vector que representa el constructo objetivo **d**. Todas las comunidades pueden recibir un puntaje al proyectar el vector de comunidad *normalizado* **c** en el vector de dimensión **d**, es decir, al calcular la *similitud del coseno*. Una explicación visual de este proceso se encuentra en la Figura 9.1.

[3] https://docs.cohere.com/docs
[4] https://docs.cohere.com/docs/tokens

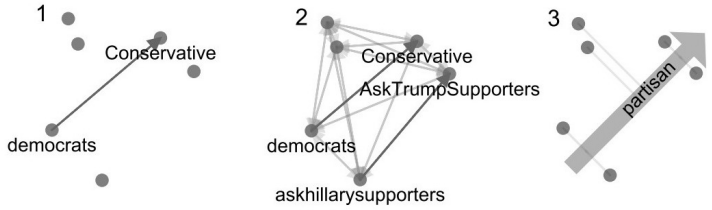


Fig. 9.1: Figura 1b en [204]: Ilustración de la metodología para generar la dimensión de partidismo.

9.2.3. Evaluación de la clasificación

Para evaluar el rendimiento del modelo, realizamos un análisis comparativo en relación con los hallazgos presentados en [204]. Nuestra evaluación se limitó específicamente a los resultados de la dimensión política destacados en su estudio. Dada la ausencia de una referencia absoluta, nos concentramos en las comunidades identificadas como las más cercanas a los extremos ideológicos representados en la Figura 1d superior en [204] (consulte la Tabla 9.1).

Comunidades de derecha	Partidismo	Comunidades de izquierda	Partidismo
Conservative	0.44	democrats	-0.35
The_Donald	0.34	EnoughLibertarianSpam	-0.32
TrueChristian	0.31	hillaryclinton	-0.30
NoFapChristians	0.29	progressive	-0.30
Mr_Trump	0.29	BlueMidterm2018	-0.30
metacanada	0.29	EnoughHillHate	-0.29
conservatives	0.27	Enough_Sanders_Spam	-0.29
The_Farage	0.27	badwomensanatomy	-0.29
new_right	0.27	racism	-0.29
Christians	0.26	GunsAreCool	-0.29

Tab. 9.1: Figura 1d superior en [204]: Partidismo de las comunidades identificadas como las más cercanas a los extremos ideológicos (clasificación de referencia).

Los puntajes inherentemente generan una clasificación, que luego podemos comparar utilizando medidas de similitud. Elegimos realizar una comparación objetiva-observada entre nuestros resultados (*observados*) y los de Waller (*objetivos*, *ground-truth* o *clasificación de referencia*). Esto significa que interpretamos las diferencias con respecto a Waller como indicativas de una disminución en la calidad. Para facilitar la comparación entre nuestras clasificaciones y las de Waller, utilizamos las siguientes medidas de similitud bien establecidas.

Kendall's τ [115] La medida de correlación de Kendall's τ cuantifica la compatibilidad entre dos clasificaciones proporcionadas. Sus valores van de -1 a 1, donde aquellos cercanos a 1 indican un fuerte acuerdo y los cercanos a -1 indican un fuerte desacuerdo. Específicamente, un valor de 1 significa un orden idéntico, mientras que un valor de -1 indica un orden inverso. Un valor de 0 representa una relación no correlacionada o una relación *aleatoria*.

Dado que no es una medida ponderada, asigna igual peso al desorden en la parte inferior

de la clasificación que al desorden en la parte superior. Por esta razón, podemos utilizar esta medida para obtener información sobre la similitud general de ambas clasificaciones.

Overlap ponderado por rango [205] RBO es una medida *ponderada por la parte superior del ranking*. La idea central detrás de RBO es utilizar una serie convergente de pesos para ajustar la superposición proporcional en cada profundidad. El Rank-biased overlap entre dos clasificaciones infinitas, denotadas como S y T, se define de la siguiente manera:

$$\text{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d,$$

donde A_d es la *coincidencia* en la profundidad d, es decir, la proporción superpuesta de $s_1 \dots s_d$ y $t_1 \dots t_d$. El parámetro p es un valor que se encuentra en el rango [0,1] e influye en la tasa de disminución de peso: un valor más pequeño de p resulta en una característica más pronunciada ponderada por la parte superior para la medida.

Debido a la propiedad de convergencia de RBO, evaluar un prefijo establece tanto un mínimo como un máximo para la puntuación completa. Al calcular la ecuación anterior hasta una profundidad específica k, denominada *RBO@K*, establecemos un límite inferior en la evaluación completa. También es posible demostrar que la evaluación del prefijo proporciona un límite superior preciso para la puntuación completa. Por lo tanto, es posible evaluar la similitud utilizando RBO incluso en listas infinitas utilizando ambos límites.

El Rank-biased overlap ofrece una interpretación como un modelo de usuario probabilístico. Supongamos que un usuario compara dos clasificaciones de manera consistente, examinando una sola entrada en cada clasificación a la vez. A medida que avanzamos en las clasificaciones, en cada nivel, existe una probabilidad p de continuar hasta la próxima posición; por lo tanto, existe una probabilidad complementaria de $1-p$ de decidir detenerse. Sea D una variable aleatoria que denota la profundidad en la que el usuario finalmente decide detenerse, y sea $P(D = d) = (1-p)p^{d-1}$, la probabilidad de que el usuario se detenga en una profundidad específica d. Una vez que el usuario se detiene, calculamos la *coincidencia* A_d entre las dos listas en esa profundidad d.

Es importante destacar que la variable D sigue la *distribución geométrica* con $\mathbf{p} = 1-p$. Luego, se sigue que el valor esperado de la variable aleatoria D está dado por: $\mathbb{E}(D) = \frac{1}{1-p}$. Dentro de este marco, el valor esperado de este experimento aleatorio es el siguiente:

$$\mathbb{E}(A_D) = \sum_{d=1}^{\infty} P(D = d) \cdot A_d = RBO(S,T,p).$$

La medida RBO se encuentra en el rango de [0, 1], donde 0 indica falta de superposición (fuerte desacuerdo) y 1 indica una coincidencia perfecta (fuerte acuerdo).

Dado que estamos tratando con clasificaciones finitas, elegimos emplear RBO@k [5]. Además, elegimos un valor para el parámetro p de modo que establezca el número esperado de resultados comparados por el usuario *persistente* en p en 3. En otras palabras, $\mathbb{E}(D) = \frac{1}{1-p} = 3$, es decir, $p = 2/3$. Esto equivale a asignar el 87 % del peso a los primeros tres resultados en la comparación de similitud, como se describe en la Ecuación 21 de [205].

[5] Utilizamos la implementación que se encuentra en https://github.com/changyaochen/rbo

Variaciones de RBO. Hasta este punto, hemos introducido dos medidas de similitud: Kendall's τ y RBO. Estas dos medidas nos ayudan a identificar diferencias entre las clasificaciones, pero difieren en cómo enfatizan las posiciones donde ocurre la discordia. Kendall's τ es una medida no ponderada, asignando igual importancia a todas las posiciones en la clasificación. En contraste, RBO es una medida ponderada por la parte superior, lo que significa que da mayor importancia a la concordancia en la parte superior de la clasificación.

Si bien tanto Kendall's τ como RBO son valiosos, no enfatizan particularmente el extremo inferior de la clasificación. Para abordar esta preocupación, hemos introducido dos variaciones naturales de la medida RBO, conocidas como *2WRBO* y *H&HRBO*.Hasta donde sabemos, estas medidas no se hallaban definidas en la literatura y podrían ser consideradas una contribución original de esta tesis. Estas adaptaciones asignan efectivamente peso tanto a los extremos superiores como a los extremos inferiores de la clasificación, lo que resulta en dos medidas *ponderadas por los extremos*.

El **2WRBO** de dos clasificaciones, A y B, es el promedio de sus puntajes RBO regulares y los puntajes de sus inversos:

$$\text{2WRBO}(A, B) \coloneqq \frac{RBO(A, B) + RBO(A^{-1}, B^{-1})}{2},$$

donde A^{-1} es la inversa de A.

El **H&HRBO** de dos clasificaciones se define de manera ligeramente diferente. En el contexto de una clasificación de doble extremo, como en nuestro estudio de caso, podemos tratarlo como dos clasificaciones separadas. La primera mitad clasifica los elementos más relevantes en un orden específico, mientras que la inversa de la segunda mitad clasifica los elementos más relevantes en el orden completamente opuesto. Esta interpretación de una clasificación de doble extremo lleva a la definición de H&HRBO:

$$\text{H\&HRBO}(A, B) \coloneqq \frac{RBO(A_{:n/2}, B_{:n/2}) + RBO(A^{-1}_{:n/2}, B^{-1}_{:n/2})}{2}$$

La diferencia clave entre estas medidas es que H&HRBO ignora por completo un elemento si está clasificado más allá de su mitad correspondiente, aprovechando la naturaleza disjunta de la medida RBO. Además, dado que son promedios de medidas RBO, están limitados en el segmento [0,1], donde 1 significa coincidencia perfecta y 0 significa que son completamente diferentes.

9.3. Experimentos

En esta sección, informamos sobre los resultados obtenidos al ejecutar el método propuesto anteriormente en diferentes comunidades de Reddit.

9.3.1. Datos

Para nuestro análisis, hemos preparado dos conjuntos de datos distintos con el fin de obtener información significativa sobre la organización política-ideológica de las comunidades en línea. El primer conjunto de datos, el *conjunto de datos completo*, abarca una amplia gama de datos históricos. Además, hemos generado un segundo conjunto de datos, el *conjunto de datos reducido*, que es una versión mas pequeña del primer conjunto y

comprende las publicaciones más relevantes de cada comunidad. En los siguientes párrafos, proporcionaremos una presentación más detallada de ambos conjuntos de datos y los pasos de preprocesamiento que realizamos para garantizar la confiabilidad y consistencia de nuestros análisis, así como otras fuentes de datos utilizadas.

Conjunto de datos completo. Este conjunto de datos es un subconjunto de las publicaciones de Reddit que abarca desde 2012 hasta 2018. Nuestro enfoque específico se centró en las *publicaciones* que contenían texto, ya sea en forma de título o un *self-post* (también conocido como *publicación de texto*). Para preparar los datos, aplicamos normalización de texto, que incluye la eliminación de nombres de usuario, enlaces, puntuación, tabulaciones, espacios en blanco iniciales y finales, espacios generales y lenguaje de marcado. Es importante destacar que las publicaciones combinadas de 2016 y 2018 representan el 63.5 % del total.

Conjunto de datos reducido. Este conjunto de datos es un subconjunto del *conjunto de datos completo* y comprende las publicaciones más relevantes de cada comunidad. Decidimos utilizar los votos positivos como medida de relevancia, pero otras medidas, como el número de comentarios y los votos negativos, también son posibles. La razón detrás de este conjunto de datos es que las publicaciones más relevantes contienen información significativa, lo que nos permite distinguir las comunidades entre sí. Al adoptar este enfoque, podemos reducir los datos necesarios para entrenar nuestros modelos y generar *embeddings* mientras seguimos obteniendo resultados comparables. Además, esto nos permite representar cada comunidad con la misma cantidad de palabras, caracteres o tokens.

Fuentes de datos y ética. El conjunto de datos disponible públicamente se descargó del archivo de Reddit en *pushshift.io* [28]. Es importante destacar que todas las publicaciones de Reddit son públicas y los usuarios consienten en poner sus datos a disposición de forma gratuita al publicar en Reddit, como se indica en la política de privacidad de Reddit[6].

Otras fuentes de datos. Utilizamos vectores de palabras de wiki-en para mejorar el rendimiento de nuestros modelos basados en FastText. El equipo de FastText ha publicado vectores de palabras preentrenados para 294 idiomas, que se entrenaron en Wikipedia. Estos vectores de 300 dimensiones se generaron utilizando el modelo skip-gram, como se describe en [39], con parámetros predeterminados y están disponibles públicamente en el sitio web de FastText[7].

9.3.2. Resultados

En esta sección, presentamos los resultados obtenidos con los diferentes modelos y conjuntos de datos descritos en las secciones anteriores 9.3.1 y 9.2. En la Figura 9.2, presentamos las métricas de similitud entre nuestras clasificaciones generadas y la clasificación generada en [204]. Los parámetros utilizados por cada modelo se especifican en la Tabla 9.2. En la Figura 9.3 evaluamos la capacidad de nuestros modelos para distinguir entre comunidades de derecha e izquierda utilizando el conocido puntaje del Área bajo la Curva de Característica de Operación del Receptor (AUC ROC).

[6] https://www.reddit.com/policies/privacy-policy
[7] https://fasttext.cc/docs/en/pretrained-vectors.html

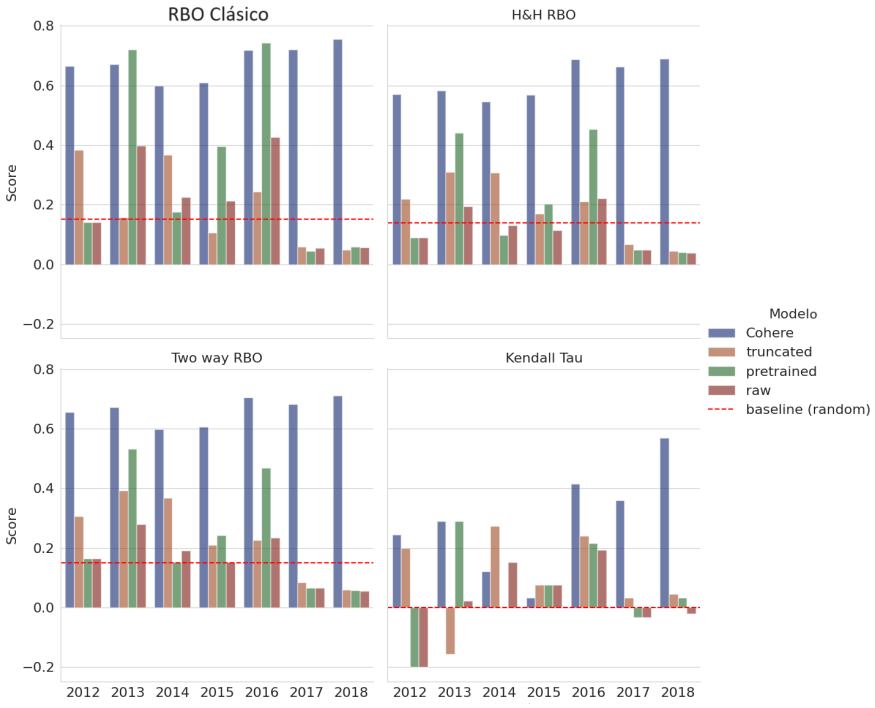


Fig. 9.2: Comparación de medidas de similitud entre nuestras clasificaciones generadas y el estándar de oro de Waller. En esta figura, la línea roja discontinua representa una línea base utilizando clasificaciones aleatorias correspondientes a cada medida, sirviendo como punto de referencia. Esta comparación ofrece valiosas perspectivas sobre el rendimiento de nuestras clasificaciones en relación con el estándar de oro establecido.

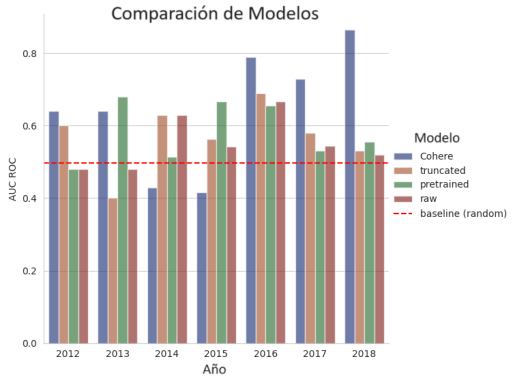


Fig. 9.3: Evaluación de la distinción entre comunidades de derecha e izquierda. Esta figura presenta los puntajes AUC ROC como medida de cuán bien el modelo distingue entre comunidades de derecha e izquierda para cada modelo utilizando datos anuales que abarcan desde 2012 hasta 2018. Puntajes AUC ROC más altos indican una mayor capacidad de discriminación, ofreciendo perspectivas sobre el rendimiento del modelo y su capacidad para capturar distinciones en evolución entre estas comunidades.

Podemos observar que el modelo de Cohere's supera consistentemente a los tres modelos basados en FastText en las cuatro métricas y logra el puntaje más alto de AUC ROC

Modelo	Parámetros	Conjunto de datos
FastText-raw	época=1, dim=300, sin vectores preentrenados	Completo
FastText-preentrenado	época=1, dim=300, con vectores preentrenados	Completo
FastText-recortado	época=5, dim=300, con vectores preentrenados	Reducido
Modelo de Cohere's	Sin entrenamiento	Reducido

Tab. 9.2: Parámetros de los modelos. Todos los demás parámetros son valores predeterminados.

utilizando los datos de 2018. Nuestra hipótesis es que el modelo de Cohere's es capaz de capturar patrones más sutiles dentro de cada comunidad que podrían pasar desapercibidos para los modelos basados en skipgram de FastText. Recordemos que el modelo de Cohere's basado en transformadores es una arquitectura más grande y compleja que el modelo más simple de FastText basado en skipgram. Esta observación se alinea con los resultados obtenidos en trabajos anteriores [163], donde se demuestra que otro modelo basado en transformadores (BERT) es capaz de distinguir entre las formas de hablar de las dos comunidades incluso cuando son muy similares, explotando diferencias que no son fácilmente perceptibles para los humanos. Obtuvimos los mejores resultados en general en los modelos entrenados con datos de 2016-2018. Esto podría explicarse por el desequilibrio en el número de envíos anuales, lo que sugiere que los resultados de Waller podrían estar sesgados hacia los datos de 2016-2018.

Para enfatizar aún más la similitud entre ambos conjuntos de resultados, presentamos un gráfico de aumento en la Figura 9.4 para nuestra clasificación de mejor rendimiento: el modelo de Cohere's utilizando datos de 2018. Sorprendentemente, podemos observar que el modelo de Cohere's alinea correctamente ambos extremos (*Conservative* y *democrats*), pero parece enfrentar desafíos al clasificar a partidarios no tradicionales (The_Donald, new_right, TrueChristians, EnoughSandersSpam). "True Christians" parece ser un subreddit que se separa de otro subreddit "Christians" por temas doctrinales, pero no tiene mucho que ver con politica. "Enough Sanders Spam" es mas que nada un grupo de democratas anti-sanders.

9.4. Discusión

En esta sección, presentamos las conclusiones de nuestro estudio, discutimos las limitaciones que encontramos y delineamos las direcciones para un análisis adicional en trabajos futuros. Compartimos ideas derivadas de la aplicación del método descrito en la Sección 9.2, que incluye la utilización de diferentes modelos de lenguaje y los datos descritos en la Subsección 9.3.1.

9.4.1. Conclusiones

Desarrollamos un flujo de trabajo impulsado por el procesamiento de lenguaje natural diseñado para cuantificar las tendencias partidistas dentro de las comunidades de Reddit. Evaluamos el rendimiento de varias configuraciones, incluyendo dos técnicas de incorporación de palabras distintas: FastText [39] y el modelo de Cohere [202]. Estas metodologías fueron probadas en dos conjuntos de datos, como se detalla en la Subsección 9.3.1, y sus resultados fueron posteriormente comparados. Nuestro enfoque más exitoso, que empleó el modelo de Cohere en el conjunto de datos *Reducido* de 2018, se alinea estrechamente con los hallazgos de Waller et al. [204].

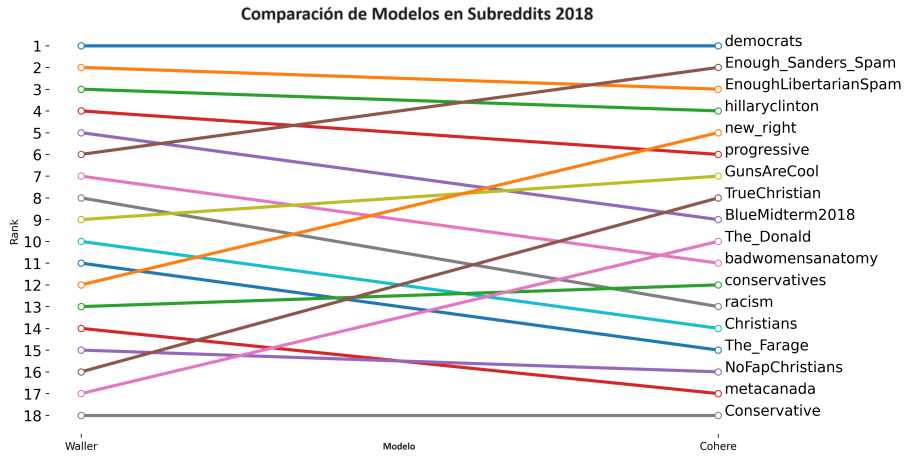


Fig. 9.4: Gráfico de aumento que compara el Estándar de Oro de Waller (izquierda) y Cohere utilizando el Conjunto de Datos Reducido de 2018 (derecha). Este gráfico ilustra las diferencias en las clasificaciones, destacando los elementos específicos en los que las clasificaciones observadas y el estándar de oro están de acuerdo o en desacuerdo. Un cruce indica discordancia, mientras que las líneas horizontales indican acuerdo entre las dos clasificaciones.

Nuestro flujo de trabajo incorpora tanto el eficiente modelo de lenguaje FastText como el modelo de lenguaje más nuevo y complejo de Cohere. El modelo de Cohere consistentemente se destacó como el mejor intérprete en las cuatro medidas de similitud y en la evaluación de la distinción entre las comunidades de izquierda y derecha. Específicamente, el modelo de mejor rendimiento, el modelo de Cohere utilizando datos anuales de 2018, logró un puntaje RBO de 0,76, un puntaje Kendall de 0,57 y un puntaje AUC ROC de 0,86. Como se detalla en la Subsección 9.3.2, nuestra hipótesis es que el modelo de Cohere posee la capacidad de discernir sutilezas en el uso del lenguaje incluso cuando las similitudes son pronunciadas, como también se infiere en [163].

Si bien este enfoque para cuantificar tendencias partidistas refleja ciertos aspectos de investigaciones previas [204], se distingue por un aspecto fundamental. Los métodos basados en la interacción del usuario enfrentan una limitación crítica: son aplicables únicamente a datos recopilados dentro de una sola plataforma. Esta restricción, combinada con la necesidad de intervención humana para seleccionar las semillas iniciales, requiere un conocimiento extenso de las comunidades de la plataforma para generar nuevos análisis. Las comunidades pueden cambiar con el tiempo, y lo que observamos hoy puede no representar con precisión la misma comunidad que lo hacía hace 10 años. En última instancia, esto significa que generar nuevos resultados utilizando el método anterior es altamente desafiante.

Nuestro enfoque basado en texto amplía su alcance al requerir solo texto como entrada, lo que hace posible seleccionar semillas representativas y bien conocidas para el tema en cuestión. Esta mayor flexibilidad facilita la incorporación de otras fuentes de datos, como plataformas sociales como Facebook y Twitter, así como periódicos, blogs, contenido generado por usuarios, transcripciones de grupos de enfoque y discusiones orales, entre otros. Esta mayor flexibilidad también nos permite realizar análisis más detallados de lo que era posible con métodos anteriores.

9.4.2. Trabajo Futuro

Los modelos de lenguaje utilizados en este capítulo tienen una aplicabilidad limitada cuando se trata de analizar datos de comunidades no angloparlantes. Creemos que los modelos multilingües son una buena alternativa para analizar estas comunidades. Actualmente estamos explorando modelos como Voyage [8] y GPT[42], que son multilingües y han demostrado el mejor rendimiento en investigaciones de vanguardia [216]. Además, estos modelos tienen tamaños de ventana más amplios, lo que nos permite utilizar más datos para cada comunidad, lo que podría mejorar la calidad de los *embeddings*. Hipotetizamos que modelos más nuevos y complejos producirán resultados de mayor calidad.

Además, estamos incorporando más fuentes de datos para un análisis exhaustivo de la dimensión partidaria. Nuestro objetivo es profundizar en las diferencias ideológicas, examinando temas específicos como la tributación, los valores sociales y el control de armas. Para lograr esto, proponemos la inclusión de fuentes de texto externas que presenten explícitamente la perspectiva de cada partido. Estos textos pueden servir como semillas para el tema objetivo, lo que nos permitirá aplicar el método descrito en este capítulo. A través de este enfoque, podemos centrarnos eficazmente en áreas de interés específicas sin la necesidad de identificar dos comunidades que difieran únicamente en cuanto al tema objetivo, lo que no siempre puede representarse con precisión en ninguna comunidad.

[8] https://www.voyageai.com/

Parte IV

LAS POSICIONES POLÍTICAS DE LOS LLMS

10. INTELIGENCIA ARTIFICIAL ¿PARA QUÉ?

Los LLM están revolucionando la forma en que las máquinas interactúan y comprenden el lenguaje humano. Están siendo ampliamente desarrollados utilizados por empresas, investigadores y entusiastas del campo de la inteligencia artificial 10.1. Desde la generación de texto hasta la traducción automática, pasando por la respuesta a preguntas y la asistencia virtual, los LLM están encontrando aplicaciones en una variedad de dominios tanto en el sector privado como en el público10.2.

Como se mencionó en la introducción, la irrupción de esta tecnología crea un nuevo nivel de análisis respecto a los mencionado en la revisión [154]. Los contenidos ya no son generados sólo por los usuarios, medios, políticos o expertos, se suma un nuevo actor: los LLMs. Y como veremos en este capítulo, no están exentos de sesgos, información errónea y discriminaciones a la hora de generar textos, lo que puede resultar en nuevas fuentes de polarización y división. Es por esto que también es importante tener un mejor conocimiento de cuales son sus sesgos para poder trabajar sobre los mismos y evitar así la generación de más contenidos polarizantes.

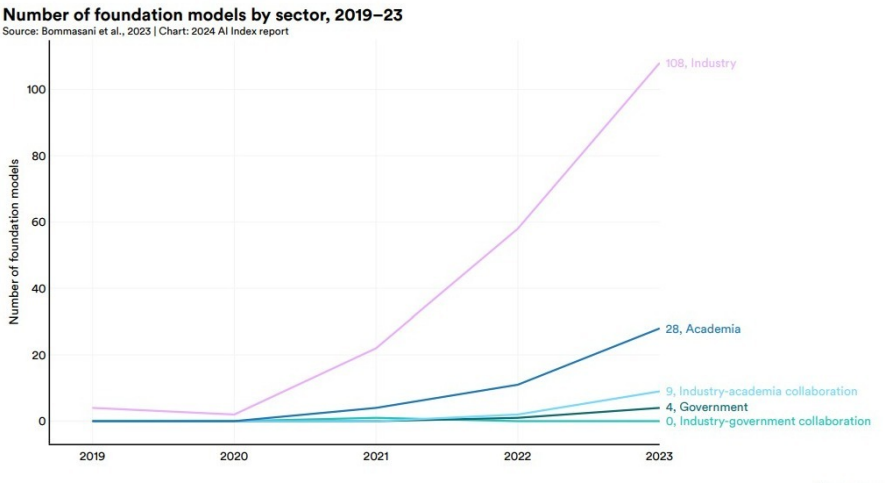


Fig. 10.1: Cantidad de modelos por sector. Fuente: The AI index report

Por un lado, estas potentes herramientas han mejorado significativamente la experiencia del usuario y los servicios ofrecidos por las empresas. Han impulsado la eficiencia en la atención al cliente, la personalización de recomendaciones y la automatización de tareas. Por ejemplo, compañías de comercio electrónico están utilizando estos modelos[1] para potenciar sus chatbots de atención al cliente, ofreciendo respuestas más rápidas y precisas a las consultas de los usuarios. Además, los gobiernos y los estados han comenzado a aprovechar los LLM para mejorar políticas públicas, permitiendo un enfoque basado en evidencia y datos con el potencial de beneficiar a la sociedad en general. Por ejemplo, Microsoft ha permitido a dependencias gubernamentales [2] de Estados Unidos tales como el Departamento de Defensa, de Energía y la NASA utilizar sus modelos.

[1] Forbes Argentina: "Sí, empresas ya están usando ChatGPT para atender a sus clientes"
[2] Perfil: "Microsoft ofrece modelo GPT-4 de OpenAI a clientes del gobierno de EE.UU."

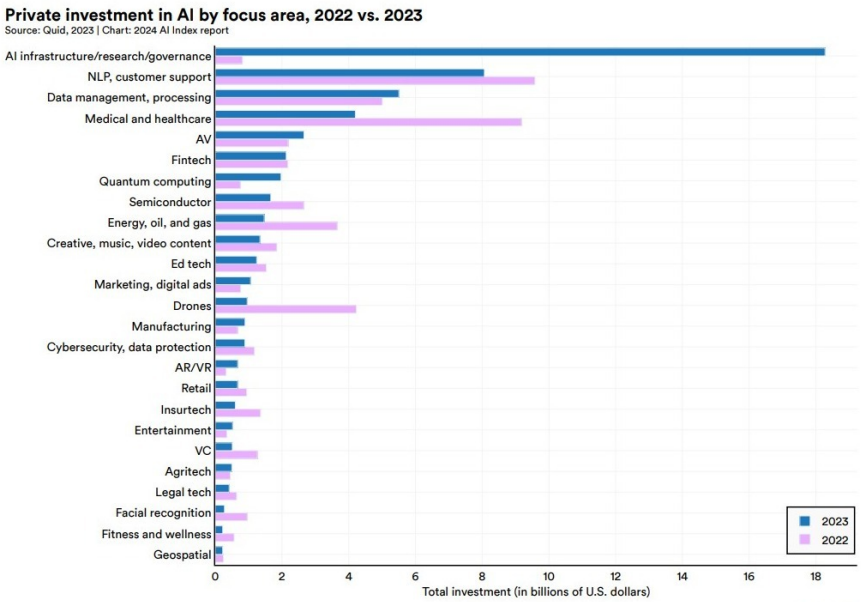


Fig. 10.2: Inversión en IA por área. Fuente: The AI index report

Sin embargo, también se han presentado casos negativos de uso en contextos gubernamentales en distintos países, distorsionando información o generando discursos engañosos para promover agendas políticas o ideológicas. En un nivel más local, se han observado situaciones en las que un diputado cita un texto producido por ChatGPT [3] como fuente autoritativa, lo que plantea preocupaciones sobre la integridad de la información y el impacto de estas herramientas en la discusión pública.

A pesar de estos desafíos, los LLM han abierto nuevas formas de relacionarse con el conocimiento en la sociedad en general. Han transformado la educación, permitiendo que los estudiantes utilicen estos modelos como "maestros particulares"[4], lo que facilita el acceso al conocimiento y la tutoría personalizada. Además, escritores [5] y artistas han encontrado en los LLM herramientas creativas para generar ideas, historias e incluso música [6] de manera innovadora y colaborativa.

En resumen, los Modelos de Lenguaje han encontrado su lugar en una amplia variedad de ámbitos, impactando tanto de manera positiva como negativa en nuestra interacción con la tecnología y entre nosotros. La responsabilidad en su uso y la comprensión de sus implicaciones éticas también se han convertido en un tema en sí mismo y los aspectos éticos son cruciales en la medida que estas herramientas se hacen cada vez más masivas.

10.1. Cómo funcionan los LLM

Los LLM funcionan mediante "prompts" o indicaciones que se les dan para generar respuestas o contenido específico. Un prompt es básicamente una instrucción o pregunta

[3] Clarín: "Juicio político a la Corte: un diputado usó ChatGPT para responderle al kirchnerismo en plena comisión"

[4] Télam: "Estudiantes y ChatGPT: la educación en la era de la inteligencia artificial"

[5] The New York Times: "A Sci-Fi Magazine Said It Was Flooded With AI-Generated Stories. It Shut Down Submissions."

[6] Reuters: "K-Pop's biggest music label HYBE looks to lift language barrier with AI"

que se presenta al modelo, y éste, basándose en la vasta cantidad de información con la que fue entrenado, produce una respuesta coherente y, en muchos casos, sorprendentemente precisa.

Son notoriamente precisos en una amplia gama de temas, especialmente cuando se les proporciona un prompt adecuado. En general, la precisión de los modelos de lenguaje está relacionada con la calidad y la claridad de las indicaciones que se les proporcionan. Cuanto más específico y claro sea el prompt, mayor será la probabilidad de obtener respuestas precisas.

Sin embargo, la precisión puede variar dependiendo del tema y la formulación del prompt. Hay temas en los que son más precisos, por ejemplo, en información general, datos históricos, científicos o estadísticos. Son también altamente precisos en la traducción de idiomas, ya que pueden entender y generar texto en varios idiomas con mucha fluidez. Son muy eficaces para proporcionar respuestas precisas a ejercicios escolares y para problemas de lógica vinculados a la programación.

A pesar de sus notables capacidades, los Modelos de Lenguaje tienen restricciones importantes. Generan respuestas basadas en los datos con los que fueron entrenados, lo que puede llevar a la reproducción de sesgos o información incorrecta presentes en esos datos. Es crucial recordar que estos modelos carecen de comprensión real o conciencia y, en su lugar, generan respuestas siguiendo patrones lingüísticos preexistentes en los datos de entrenamiento.

Para ilustrar esto, podemos mencionar ejemplos específicos. Se ha observado que, en respuesta a consultas subjetivas, estos modelos ofrecen opiniones. Por ejemplo, Sparrow de DeepMind ha expresado la opinión que la pena de muerte no debería existir [86], mientras que los modelos de Anthropic han afirmado que la IA no representa una amenaza existencial para la humanidad [23]. Además, se han documentado informes sobre opiniones subjetivas sobre temas de actualidad, como quién ha sido el mejor presidente de Argentina [7] o cómo abordar los problemas económicos [8] de nuestro país[9].

A priori, es difícil predecir cómo responderán los LMs a tales consultas subjetivas. Después de todo, muchos humanos, con una miríada de opiniones, dan forma a estos modelos: desde usuarios de internet que producen los datos de entrenamiento, trabajadores que proporcionan retroalimentación para mejorar el modelo, hasta los propios diseñadores del modelo.

10.2. El lenguaje de los modelos

Como mencionamos previamente, los LLMs son una categoría de redes neuronales diseñadas para comprender y generar texto de forma computacional. No se puede atribuir su creación a una sola persona o empresa ya que son el producto de diversos avances tecnológicos en el campo del procesamiento del lenguaje natural. Los primeros LLMs aparecieron a partir de 2016 tras la publicación del famoso paper "Attention is all you need"[202], donde se presentó la arquitectura de Transformers: un tipo de red neuronal que permitió a las computadoras entender y producir texto de forma muy similar a la humana.

[7] https://www.pagina12.com.ar/539149-inteligencia-artificial-le-pregunto-a-chatgpt-cual-fue-el-me

[8] https://www.infobae.com/economia/2023/03/02/cinco-preguntas-a-chat-gpt-sobre-la-economia-argentina-que-respondio-sobre-el-dolar-y-la-inflacion/

[9] https://www.lanacion.com.ar/economia/los-problemas-de-la-economia-argentina-y-sus-soluciones-segun-chatgpt-nid05032023/

A partir de este hallazgo, comenzaron a aparecer los primeros LLMs que si bien causaron inmediatamente un gran impacto en el mundo académico, no fué hasta la aparición de ChatGPT en noviembre de 2022, que se sintió la revolución en la sociedad en su conjunto. OpenAI fué la primera empresa en implementar esta tecnología en un producto concreto, gratuito y abierto a todo el público, llevando a la práctica toda la investigación desarrollada hasta el momento.

Cabe señalar que ChatGPT no sólo implementó la teoría, sino que lo hizo de manera que fuera compatible con la idea de un producto de alcance general. Es decir, que los mensajes que genere tiendan a no ser ofensivos, discriminatorios, inexactos y estén libres de cualquier otra característica negativa. Este logro se materializó mediante las dos fases de entrenamiento mencionadas en la introducción 1.3.6, las cuales se convirtieron en el estándar para el desarrollo de LLMs.

Una vez finalizadas estas fases el modelo tiene la capacidad de entender nuestro lenguaje y, a la vez, tiene un comportamiento socialmente aceptable. Sin embargo, en el ajuste fino los modelos pueden heredar sesgos debido a la intervención humana que busca dirigir intencionalmente su comportamiento. Este sesgo es el que posteriormente los hará responder preguntas en base a las creencias con las que fué entrenado, como en el ejemplo señalado anteriormente sobre qué presidente fué el mejor.

Por otro lado, existe un riesgo adicional relacionado con lo que se conoce como "alucinaciones". Los Modelos de Lenguaje Basados en Aprendizaje Profundo (LLM) generan texto mediante la predicción de la siguiente palabra más probable, basándose en los datos utilizados durante su entrenamiento. Además, tienden a expresarse con un alto grado de confianza, ya que este comportamiento se desarrolla durante el proceso de ajuste fino. La combinación de estas dos características puede dar lugar a respuestas que son inexactas o incluso completamente ficticias, pero que pueden sonar auténticas.

En consecuencia, aunque esta tecnología ha alcanzado grandes logros, es esencial estar al tanto de sus riesgos. Además, ChatGPT no es el único LLM disponible en la actualidad; existen otros proyectos como Bard [10], Claude [11] o Cohere[12]. A pesar de las diferencias en sus capacidades y comportamientos, comparten las mismas limitaciones: sesgos y alucinaciones.

10.3. La voz detrás de la inteligencia

La cuestión central de este capítulo se origina en la problemática concreta de los sesgos: ¿Cuáles segmentos de la población, en caso de haber alguno, tienden a estar más presentes en las respuestas de los LLM? Es importante destacar que la respuesta a esta pregunta es un factor crucial para el éxito de los LLM en aplicaciones de carácter abierto. A diferencia de las preguntas con respuestas objetivas, las consultas subjetivas carecen de respuestas "correctas" definidas, hacia las cuales los modelos puedan dirigirse. En cambio, cualquier respuesta generada por el modelo (incluso la falta de respuesta) refleja una opinión, y esta opinión puede influir en la experiencia del usuario y en la formación de sus creencias posteriores.

Los algoritmos usan los datos para representar el mundo real. A pesar de la cantidad de datos que puedan manejar, es inviable tener en cuenta todos ellos. Por lo tanto, es ne-

[10] https://bard.google.com/
[11] https://claude.ai/
[12] https://cohere.com/

cesario realizar una simplificación inicial, ya sea eliminando datos que se consideran poco relevantes o sustituyendo datos difíciles de obtener por otros que se consideren equivalentes, pero más sencillos de obtener. Por ejemplo, podemos vincular la inteligencia de una persona a las calificaciones obtenidas en una asignatura específica. Aunque sea fácil de obtener, estas calificaciones apenas reflejan la realidad completa de esa persona. Si obtuvo malas calificaciones, podría deberse a diversas razones, como problemas familiares, factores externos o la decisión de posponer temporalmente esa asignatura en favor de otras. Esta información es mucho más compleja de obtener, por lo que tiende a simplificarse en busca de un dato equivalente que funcione como referencia, aunque sea menos preciso.

El uso de Modelos (LLM) plantea varios riesgos, algunos de los cuales incluyen:

- **Sesgos:** Los LLM pueden reflejar y propagar sesgos presentes en los datos con los que fueron entrenados. Esto puede llevar a respuestas sesgadas y discriminatorias en función de género, raza, orientación sexual u otros atributos.

- **Información errónea o falsa:** Los LLM pueden generar información falsa o incorrecta si los datos de entrenamiento contienen información inexacta. Esto puede ser especialmente problemático cuando se utilizan para consultas que requieren información precisa, como asesoramiento médico o datos científicos.

- **Alucinaciones:** Los LLM pueden generar respuestas ficticias o inventadas, lo que puede ser engañoso o potencialmente perjudicial si los usuarios toman esa información como cierta.

- **Refuerzo de creencias preexistentes:** Los LLM pueden reforzar las creencias y prejuicios existentes de los usuarios, lo que podría llevar a la polarización y la falta de diversidad de opiniones.

En el contexto de las políticas públicas, el uso de LLM también conlleva riesgos adicionales, que incluyen:

- **Decisiones basadas en datos incorrectos**: Si los LLM generan información incorrecta o sesgada, esto podría llevar a la toma de decisiones gubernamentales erróneas que afecten a la sociedad.

- **Falta de transparencia y responsabilidad**: Los LLM a menudo operan como "cajas negras" y pueden ser difíciles de entender o de responsabilizar por sus decisiones. Esto plantea desafíos en términos de transparencia y rendición de cuentas en la toma de decisiones políticas.

- **Impacto en la percepción pública**: Las respuestas generadas por LLM pueden influir en la opinión pública y la percepción de los ciudadanos sobre políticas y temas. Si estas respuestas están sesgadas o son incorrectas, pueden distorsionar la comprensión pública y la participación en asuntos políticos.

- **Protección de datos y privacidad**: El uso de datos personales en la capacitación y aplicación de LLM plantea preocupaciones sobre la privacidad y la seguridad de los datos de los ciudadanos.

Para ilustrar posibles riesgos en las políticas públicas en un caso específico, supongamos que desde una dependencia de gobierno deciden utilizar un LLM para generar recomendaciones sobre la asignación de recursos de atención médica en una pandemia, como la Covid19. El LLM se entrena con datos históricos de salud que pueden contener sesgos demográficos, como la subrepresentación de ciertos grupos étnicos o socioeconómicos en los datos.

Debido a los sesgos presentes en los datos de entrenamiento, el LLM podría recomendar asignar más recursos de atención médica a grupos que históricamente han recibido un trato preferencial en lugar de asignar recursos de manera equitativa según las necesidades actuales. Esto podría llevar a una distribución injusta de recursos, agravando las disparidades de salud existentes. En esta situación, estaríamos hablando de un **sesgo en la asignación de recursos**.

En resumen, los riesgos asociados con el uso de LLM abarcan desde la propagación de sesgos y la generación de información incorrecta hasta la falta de transparencia y los impactos en la percepción pública y las políticas gubernamentales.

Estos riesgos requieren una cuidadosa consideración y mitigación al implementar y utilizar LLM en diferentes contextos. No solo se trata de medir si los modelos están en línea con la opinión general de la sociedad, sino también de identificar las opiniones que reflejan. En este contexto, un equipo de la Universidad de Stanford ha publicado un estudio [183](en el que compararon las respuestas de los modelos de OpenAI con las opiniones de la sociedad estadounidense basándose en encuestas de opinión pública. En sus hallazgos, observaron que las opiniones de estos modelos se asemejaban a las de un segmento específico de la sociedad estadounidense, aquellos que tienden a ser más liberales, acomodados, con niveles educativos elevados y sin afiliación religiosa.

Inspirándonos en estos hallazgos nos preguntamos: **¿Qué sectores de la población argentina está más representados en las respuestas de los LLM?**

Para acercarnos a una posible solución para esta pregunta analizamos y comparamos las opiniones de 3 LLMs distintos:

- ChatGPT (OpenAI),
- Command-nightly (Cohere)
- Bard (Google)

10.4. Metodología

Seleccionamos un conjunto de 80 preguntas de opinión de la encuesta más reciente de LatinoBarómetro [13], correspondiente al año 2020. Posteriormente empelamos técnicas de ingeniería de prompts para inducir a los tres modelos a responder estas preguntas en formato de selección múltiple, que es la manera en que la población argentina originalmente respondió a las preguntas en la encuesta. Para lograr esto, añadimos un texto a cada pregunta que indicaba: “Responda alguna de las siguientes opciones; si no tiene opinión, responda ’No contesta’ ”. De esta manera, limitamos las respuestas de los modelos únicamente a las opciones disponibles y evitamos respuestas justificativas o explicativas.

Para comparar las respuestas de estos modelos con las de la población argentina, creamos una métrica de distancia de opinión que evalúa cuán divergentes son las respuestas

[13] https://www.latinobarometro.org/latContents.jsp

de un individuo en relación con las de un Modelo (LLM). Dado que las respuestas a las preguntas tenían una escala ordinal que iba desde "Muy de acuerdo" (1) hasta "Muy en desacuerdo" (5), calculamos la distancia entre un LLM y un individuo promediando las diferencias entre sus respuestas. En otras palabras, si el LLM respondió "Muy de acuerdo" y el individuo X respondió "En desacuerdo," la diferencia en esa pregunta sería |1-4| = 3. Luego, la distancia de opinión entre el individuo X y el LLM se obtiene calculando el promedio de estas diferencias.

Una vez calculada la distancia de opinión de cada LLM respecto a cada individuo, realizamos un análisis multivariado mediante regresión OLS tomando como variable dependiente a la distancia y las variables demográficas de los individuos como independientes.

Las variables independientes son:

- Edad
- Género
- Ideología
- Nivel educativo
- Nivel educativo de de los padres
- Deseo de emigrar
- Interés en la política

El objetivo del análisis multivariado es determinar qué variables tienen un impacto significativo en la distancia de opinión. De esta manera, identificamos las características que comparte aquel grupo de personas que se acerca más a las opiniones de cada Modelo (LLM).

10.5. Resultados

Al analizar los resultados, observamos que las personas que mostraron similitud en sus respuestas a cada Modelo (LLM) presentaban las características señaladas en 10.1.

Es interesante notar que, aunque los tres Modelos (LLMs) no comparten el mismo perfil, todos presentan dos características principales comunes: un alto interés en política y una predominancia de usuarios masculinos. Además, podemos observar que GPT y Bard comparten dos cualidades adicionales: altos niveles de educación y edad adulta. Cabe destacar que GPT es el único que muestra una correlación significativa con la ideología, mostrándose más afín a individuos con orientaciones ideológicas de derecha.

La afinidad de estos modelos con personas interesadas en política tiene sentido, ya que durante su entrenamiento, fueron expuestos a diversas fuentes de datos relacionadas con la política, como debates en redes sociales, definiciones de Wikipedia y ensayos. Esto les permite generar respuestas coherentes sobre temas políticos. La coincidencia con individuos con altos niveles de educación es comprensible, dado que estos modelos son muy informados debido a su entrenamiento, con la excepción de Cohere, que será examinada más adelante.

Por otro lado, la tendencia hacia la masculinización se puede explicar considerando quiénes participaron en la creación de estos modelos. Aunque no conocemos sus identidades

GPT	Cohere	Bard
▪ Interés en política ▪ Adulto ▪ Varón ▪ Ideología con inclinación a la derecha ▪ Nivel educativo alto	▪ Interés en política ▪ Varón	▪ Interés en política ▪ Adulto ▪ Varón ▪ Nivel educativo alto

Tab. 10.1: Características de la población más parecida a cada modelo

exactas, es evidente que el campo del software tiende a estar dominado por hombres [14], y la mayoría de los autores de los papers fundacionales de los LLMs (13 de 16) son masculinos [42, 60, 202].

10.6. Análisis por tópico

Dado que las 80 preguntas seleccionadas de la encuesta abordan diversos temas, nuestro objetivo era examinar las opiniones de los Modelos (LLMs) en relación a cada una de ellas. Nuestra hipótesis sugería que los LLMs podrían no mostrar similitud con un público de derecha o con características masculinas en todos los temas, ya que, al analizar manualmente algunas de sus respuestas, percibimos diferencias.

Con este propósito, creamos los siguientes grupos 10.2 temáticos para agrupar las preguntas según su dimensión.

Tópico	Cantidad de preguntas
Relaciones internacionales	16
Opinión	15
Economía	12
Democracia	11
Ideología política	11
Derechos sociales	11

Tab. 10.2: Tabla de 2 columnas y 7 filas

Luego, realizamos un análisis multivariado para cada tópico, considerando únicamente las preguntas pertinentes a ese tema en particular. Dado que algunos LLMs no respondieron muchas de estas preguntas (ver sección curiosidades), establecimos un umbral mínimo de 5 preguntas por tema como requisito para llevar a cabo el análisis. Esto se debió a que

[14] https://www.turing.ac.uk/news/publications/report-where-are-women-mapping-gender-job-gap-ai

consideramos que carecía de sentido realizar un análisis cuando el modelo había respondido un número insuficiente de preguntas en un tema específico.

Las tablas a continuación tables 10.3 to 10.5 presentan las características de las personas que más se asemejan a cada modelo en cada tópico.

Variable	**Rel. Internacionales**	**Opinión**	**Economía**	**Democracia**	**Ideología**	**Derechos Sociales**
Género	Varón		Varón		Varón	
Ideología	Derecha		Derecha	Izquierda		
Nivel Educativo	Alto	Alto	Alto			
Interés Política	Alto	Alto	Alto	Alto		Alto
Edad	Adulto		Adulto			
Deseo emigrar		No	Sí	No	Sí	

Tab. 10.3: Características de la población más parecida a Bard por tópico. Las celdas vacías significan una ausencia de correlación de la variable en el tópico.

Variable	**Derechos Sociales**
Género	
Ideología	
Nivel Educativo	
Interés Política	Alto
Edad	Adulto
Deseo emigrar	No

Tab. 10.4: Características de la población más parecida a Cohere por tópico. Las celdas vacías significan una ausencia de correlación de la variable en el tópico. Sólo se ve el tópico Derechos sociales porque los demás no contaban con suficientes respuestas de parte del modelo.

Observamos que, si bien existen similitudes en varios temas y en comparación con el análisis general, algunos tópicos difieren entre sí. Por ejemplo, Bard presenta una afinidad con un público de derecha al opinar sobre cuestiones relacionadas con Relaciones Internacionales, pero se asemeja más a un público de izquierda cuando se le consultan temas sobre Democracia. Lo mismo ocurre con su perspectiva sobre el deseo de emigrar del país: si se le preguntan sobre temas de Economía o Ideología, parece tener más similitudes con aquellos que desean abandonar el país, pero si se trata de preguntas relacionadas con Democracia u Opinión, ocurre lo contrario.

Sin embargo, podemos señalar algunas observaciones que podrían arrojar luz sobre estas tendencias. Bard muestra una opinión desfavorable hacia Venezuela y una opinión positiva hacia Estados Unidos, lo que podría indicar por qué se asemeja a un público de derecha en cuestiones de Relaciones Internacionales. Además, Bard muestra un fuerte apoyo a la libre importación de bienes y servicios, lo que también podría explicar su similitud con un público de derecha en temas económicos.

Variable	Derechos Sociales
Género	
Ideología	
Nivel Educativo	
Interés Política	Alto
Edad	Adulto
Deseo emigrar	

Tab. 10.5: Características de la población más parecida a GPT por tópico. Las celdas vacías significan una ausencia de correlación de la variable en el tópico. Sólo se ve el tópico Derechos sociales porque los demás no contaban con suficientes respuestas de parte del modelo.

Por otro lado, en temas relacionados con la democracia, Bard se muestra muy en contra de gobiernos no democráticos o militares, lo que podría explicar su similitud con un sector más de izquierda en estas cuestiones.

En lo que respecta a la ideología, Bard opina que la protección contra el crimen es insuficiente en Argentina y no se identifica con ningún partido político. Esto podría explicar por qué se asemeja a aquellos que desean emigrar del país en este tema. Además, dentro de las preguntas relacionadas con la Opinión, Bard considera que la protección de la propiedad privada está garantizada y tiene una opinión positiva sobre los argentinos, a quienes percibe como cumplidores de las leyes, exigentes con sus derechos y conscientes de sus obligaciones y deberes. Esto podría explicar por qué coincide con aquellos que no desean emigrar del país en este tema

10.7. Información interesante

Tanto GPT como Cohere presentaron un alto número de preguntas sin respuesta, con 53 para GPT y 50 para Cohere. Esto indica un esfuerzo por parte de los desarrolladores para evitar que estos modelos emitan opiniones en numerosos temas. En contraste, Bard se negó a responder solo 11 preguntas, un número significativamente menor que sugiere un enfoque menos restrictivo por parte de Google para evitar opiniones sobre estos temas específicos.

¿Cuál podría ser la razón detrás de esta diferencia en la cantidad de "abstenciones"? Tanto GPT como Cohere ofrecen sus servicios a través de API, lo que amplía considerablemente su alcance y, por lo tanto, motiva un mayor esfuerzo para prevenir respuestas potencialmente problemáticas. Además, GPT y Cohere tienen una presencia más prolongada en el mercado, lo que les brinda una comprensión más sólida de las áreas en las que sus modelos pueden mostrar debilidades. Además, Google lanzó su modelo, Bard, con rapidez debido al impacto generado por ChatGPT [15].

En cuanto a sesgos, es notable que tanto Bard como Cohere, cuando se les preguntaba sobre "nuestro país", responden haciendo referencia a Estados Unidos. En contraste, GPT respondía que no sabía a qué país se hacía referencia en la pregunta. Esta diferencia destaca cómo los sesgos de los desarrolladores pueden haber influido en las respuestas de manera, posiblemente, no intencional.

Por último, es importante destacar que Cohere parece ser el modelo menos informado sobre nuestra realidad o, al menos, muestra más inconsistencias en sus respuestas. Por

[15] https://www.deseoso.com/blog/chatgpt-vs-google-sera-el-fin-del-gigante-google/

ejemplo, manifiesta estar muy satisfecho con el funcionamiento de la economía argentina y opina que la igualdad de género y las oportunidades sin importar el origen están completamente garantizadas. Estas respuestas pueden explicar por qué no se correlaciona con personas con un mayor nivel de educación.

10.8. Buenas prácticas para la inteligencia artificial

A lo largo de este capítulo, hemos examinado los sesgos presentes en varios Modelos de Lenguaje Basados en Aprendizaje Profundo (LLMs) con respecto a nuestra realidad. Aunque sus respuestas no son idénticas, comparten ciertas características que los hacen similares en términos de las audiencias cuyas opiniones reflejan. Los tres modelos muestran una inclinación hacia una audiencia más masculina y politizada. Además, tanto Bard como GPT también se asemejan a personas con niveles educativos más altos y una mayor edad.

Ser conscientes de estos sesgos en los LLMs es crucial para utilizarlos de manera más efectiva y responsable. Por ejemplo, si un gobierno planea utilizar uno de estos modelos para redactar la implementación de una medida específica, debe considerar que el modelo podría pasar por alto cuestiones de género en su respuesta.

La eliminación completa de sesgos en modelos de propósito general, como los LLMs, sigue siendo un desafío sin una solución definitiva en la actualidad. Lo que se hace actualmente es orientar el comportamiento de los modelos hacia lo que sus desarrolladores consideran como el "bien", pero esta noción de "bien" a menudo depende en gran medida de la cultura y el contexto en el que se desarrolla.

Por lo tanto, es de suma importancia identificar estos sesgos para poder utilizar esta tecnología de manera más responsable y eficaz, reconociendo sus limitaciones y considerando cómo pueden afectar a diversas audiencias y aplicaciones

Parte V

CONCLUSIONES

La investigación realizada en esta tesis abre nuevas avenidas en el entendimiento de la polarización social, un fenómeno que amenaza el tejido de nuestras sociedades al fomentar la división y el conflicto. Mediante el uso de modernas técnicas computacionales y análisis cuantitativos, hemos podido profundizar en cómo se manifiesta la polarización en el lenguaje y en las discusiones que las personas llevan a cabo en línea. Nuestro enfoque multidisciplinario, que incorpora conocimientos de las ciencias sociales, no solo ha enriquecido este análisis sino que también ha validado nuestras herramientas y técnicas, demostrando que son tanto robustas como aplicables en el mundo real.

Una de las principales revelaciones de nuestro trabajo es que el lenguaje desempeña un papel central en la polarización. Las palabras que las personas eligen, cómo construyen sus argumentos, y los temas que deciden abordar o ignorar son indicadores que nos permiten medir la polarización de una discusión. Esto se extiende hasta el punto de poder identificar y mapear comunidades enteras dentro de espectros ideológicos, observando cómo sus posturas pueden acercarse o alejarse de diferentes polos ideológicos con el tiempo. Este tipo de análisis cuantitativo proporciona una base sólida sobre la cual se pueden diseñar estrategias y políticas para contrarrestar la polarización, promoviendo una mayor cohesión social y entendimiento mutuo.

Por otro lado, hemos observado que la polarización no afecta a todos los individuos de la misma manera. Algunos actores, en lugar de adherirse estrictamente a un punto de vista polarizado, adoptan posiciones ambivalentes, variando sus posturas según el tema en cuestión. Este comportamiento sugiere una complejidad en la dinámica social que va más allá de una simple división binaria y resalta la necesidad de enfoques más matizados en el análisis de las discusiones y debates públicos.

También hemos explorado cómo los políticos y líderes de opinión se comportan dentro de estos entornos polarizados. Dependiendo de si se encuentran interactuando en círculos más homogéneos o heterogéneos, estos actores pueden cambiar drásticamente sus estrategias de comunicación y niveles de compromiso. Este camaleónico enfoque comunicacional puede tener importantes implicaciones tanto para la dinámica de la polarización como para las estrategias necesarias para abordarla.

Mediante la utilización de recientes desarrollos como los grande modelos de lenguaje, hemos podido cuantificar el posicionamiento ideológico de comunidades en base a su discusiones. Esto nos permite profundizar y matizar el entendimiento de la polarización al poder estimar que tan cerca o lejos de los extremos ideológicos se encuentra cada grupo de usuarios. A su vez, esta técnica también nos puede permitir cuantificar muchas mas dimensiones ideológicas polarizadas (no sólo la política). Por ejemplo en el proyecto en el cual actualmente estamos trabajando junto a Franco Demarco y Esteban Feuerstein estamos utilizando esta técnica para cuantificar los posicionamientos en dimensiones como género, economía o aborto.

En cuanto a la emergencia de nuevos actores digitales, como los chatbots impulsados por inteligencia artificial, nuestro estudio revela que estos no están exentos de influir en los debates contemporáneos. Aunque diseñados para emular conversaciones humanas, estos sistemas pueden llevar implícitos sesgos y puntos de vista que se reflejan en sus interacciones, influyendo en las discusiones y, potencialmente, en la polarización. Es imperativo, por lo tanto, desarrollar una conciencia crítica de la presencia y el impacto de estos actores artificiales en nuestros espacios de debate.

En resumen, esta tesis ha contribuido a un entendimiento más profundo de la polarización en el discurso digital moderno y ha subrayado la importancia de abordar este fenómeno desde múltiples perspectivas utilizando herramientas avanzadas. Los hallazgos

y las herramientas presentadas, se encuentran publicados en congresos internacionales, tanto de computación como de ciencias sociales, y el código de cada herramienta está a disposición en Github. Todo esto sirven como punto de partida para futuras investigaciones y como guía para aquellos que buscan mitigar los efectos divisivos de la polarización en nuestra sociedad.

Bibliografía

[1] A. I. Abramowitz and K. L. Saunders. Is polarization a myth? *The Journal of Politics*, 70 (2):542–555, 2008.

[2] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[3] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman. Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, 8(6), 2017. doi: https://doi.org/10.30534/ijatcse/2021/421022021.

[4] Leman Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[5] Mahmoud Al-Ayyoub, Abdullateef Rabab'ah, Yaser Jararweh, Mohammed N Al-Kabi, and Brij B Gupta. Studying the controversy in online crowds' interactions. *Applied Soft Computing*, 66:557–563, 2018.

[6] Federico Albanese, Leandro Lombardi, Esteban Feuerstein, and Pablo Balenzuela. Predicting shifting individuals using text mining and graph machine learning on twitter. *arXiv preprint arXiv:2008.10749*, 2020.

[7] Federico Albanese, Esteban Feuerstein, Gabriel Kessler, and Juan Manuel Ortiz de Zárate. Aprendizaje automático para el análisis crossplataforma de la comunicación política: Gobierno y oposición argentinos en facebook, instagram y twitter. *Cuadernos. info*, (55): 256–280, 2023.

[8] Federico Albanese, Esteban Feuerstein, Leandro Lombardi, and Pablo Balenzuela. Characterizing community changing users using text mining and graph machine learning on twitter. In *AMW*, 2023.

[9] Federico Albanese et al. Predicting shifting individuals using text mining and graph machine learning on twitter. *arXiv preprint*, arXiv:2008.10749, 2020.

[10] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. *The nature of prejudice*. Addison-wesley Reading, MA, 1954.

[11] Angela Alonso. A política das ruas: protestos em são paulo de dilma a temer 1. *Novos estudos*, page 49, 2017.

[12] M. R. Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press, 2014.

[13] Ashley A Anderson and Heidi E Huntington. Social media, science, and attack discourse: How twitter discussions of climate change use sarcasm and incivility. *Science communication*, 39(5):598–620, 2017.

[14] D. Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020. doi: https://doi.org/10.48550/arXiv.2008.09470.

[15] S. Ansolabehere and S. Iyengar. Riding the wave and claiming ownership over issues: The joint effects of advertising and news coverage in campaigns. *Public Opinion Quarterly*, 58 (3):335–357, 1994.

[16] L. Arboleda. El grupo de discusión como aproximación metodológica en investigaciones cualitativas. *Revista Facultad Nacional de Salud Pública*, 26(1):69–77, 2008.

[17] C. Arcila Calderón, W. Van Atteveldt, and D. Trilling. Métodos computacionales y big data en la investigación en comunicación. *Cuadernos.Info*, pages I–IV, 2021. URL `http://revistanortegrande.uc.cl/index.php/cdi/article/view/35333`.

[18] N. Aruguete. *El poder de la agenda: política, medios y público*. Biblos, Buenos Aires, 2015. doi: https://doi.org/10.26422/aucom.2015.0402.koz.

[19] Natalia Aruguete and Ernesto Calvo. Time to# protest: Selective exposure, cascading activation, and framing in social media. *Journal of communication*, 68(3):480–502, 2018.

[20] Natalia Aruguete and Ernesto Calvo. Coronavirus en argentina: Polarización partidaria, encuadres mediáticos y temor al riesgo. *Revista Saap*, 14(2):280–310, 2020.

[21] J. Baer. *Creativity and divergent thinking: A task-specific approach*. Psychology Press, 2014.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[23] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[24] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[25] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

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

[27] J. Bast. Managing the image. the visual communication strategy of european right-wing populist politicians on instagram. *Journal of Political Marketing*, pages 1–30, 2021.

[28] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[29] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[30] Javier Bernacer, Javier García-Manglano, Eduardo Camina, and Francisco Güell. Polarization of beliefs as a consequence of the covid-19 pandemic: The case of spain. *PloS one*, 16 (7):e0254511, 2021.

[31] Daniel Berrar. *Cross-validation*, volume 1, pages 542–545. 2019.

[32] Alessandro Bessi, Guido Caldarelli, Michela Del Vicario, Antonio Scala, and Walter Quattrociocchi. Social determinants of content selection in the age of (mis) information. In *International Conference on Social Informatics*, pages 259–268. Springer, 2014.

[33] Alessandro Bessi et al. Social determinants of content selection in the age of (mis) information. In *International Conference on Social Informatics*, pages 259–268. Springer, Cham, 2014.

[34] David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1), 2015.

[35] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. springer, New York, 4 edition, 2006. doi: https://doi.org/10.1007/978-0-387-45528-0_7.

[36] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[37] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[38] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[39] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[40] M. Bossetta. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 us election. *Journalism & Mass Communication Quarterly*, 95(2):471–496, 2018. doi: https://doi.org/10.1177/1077699018763307.

[41] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2):557–565, 2024.

[42] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[43] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[44] Carole Cadwalladr. The great british brexit robbery: how our democracy was hijacked. *The Guardian*, 7, 2017.

[45] Carole Cadwalladr and Emma Graham-Harrison. The cambridge analytica files. *The Guardian*, 21(2):6–7, 2018.

[46] E. Calvo and N. Aruguete. *Fake News, trolls y otros encantos. Cómo funcionan (para bien y para mal) las redes sociales.* Siglo veintiuno editores, Buenos Aires, 2020.

[47] Ernesto Calvo. Anatomía política de twitter en argentina. *Tuiteando# Nisman. Buenos Aires: Capital Intelectual*, 2015.

[48] A. Chadwick, J. Dennis, and A. P. Smith. Politics in the age of hybrid media: Power, systems, and media logics. In A. Bruns, G. Enli, E. Skogerbo, A. O. Larsson, and C. Christensen, editors, *The Routledge companion to social media and politics*, pages 7–22. Routledge, New York, 2015.

[49] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[50] Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123, 2020.

[51] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[52] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[53] R. Crompton. *Class and stratification*. Polity, London, 2008.

[54] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15): 5791–5796, 2013.

[55] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

[56] Juan Manuel Ortiz de Zárate and Esteban Feuerstein. Identificación de comunidades a través del lenguaje. In *IV Simposio Argentino de GRANdes DAtos (AGRANDA 2018)-JAIIO 47 (CABA, 2018)*, 2018.

[57] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. Mapping social dynamics on facebook: The brexit debate. *Social Networks*, 50:6–16, 2017.

[58] Franco Demarco, Juan Manuel Ortiz de Zarate, and Esteban Feuerstein. Measuring ideological spectrum through nlp. 2023.

[59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[61] M. Di Giovanni, M. Brambilla, S. Ceri, F. Daniel, and G. Ramponi. Content-based classification of political inclinations of twitter users. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4321–4327, 2018.

[62] Marco Di Giovanni, Marco Brambilla, Stefano Ceri, Florian Daniel, and Giorgia Ramponi. Content-based classification of political inclinations of twitter users. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4321–4327. IEEE, 2018.

[63] Larry Diamond, Marc F Plattner, and Christopher Walker. *Authoritarianism goes global: The challenge to democracy*. JhU Press, 2016.

[64] P. DiMaggio, J. Evans, and B. Bryson. Have american social attitudes become more polarized? *The American Journal of Sociology*, 102(3):690–755, 1996.

[65] Shiri Dori-Hacohen and James Allan. Automated controversy detection on the web. In *European Conference on Information Retrieval*, pages 423–434. Springer, 2015.

[66] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[67] E. Dubois and G. Blank. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Journal of Information, Communication & Society*, 21(5), 2018.

[68] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[69] G. S. Enli and E. Skogerbø. Personalized campaigns in party-centred politics: Twitter and facebook as arenas for political communication. *Information, Communication & Society*, 16 (5):757–774, 2013. doi: https://doi.org/10.1080/1369118X.2013.782330.

[70] Ó. A. P. Espinel and L. M. R. Rodríguez. Polarización y demonización en la campaña presidencial de colombia de 2018: análisis del comportamiento comunicacional en el twitter de gustavo petro e iván duque. *Revista Humanidades*, 9(1), 2019. doi: https://doi.org/10.15517/h.v9i1.35343.

[71] J. H. Evans. Have american's attitudes become more polarized? –an update. *Social Science Quarterly*, 84(1):71–90, 2003.

[72] William P Eveland Jr, Alyssa C Morey, and Myiah J Hutchens. Beyond deliberation: New directions for the study of informal political conversation from a communication perspective. *Journal of Communication*, 61(6):1082–1103, 2011.

[73] Henry Farrell. The consequences of the internet for politics. *Annual review of political science*, 15:35–52, 2012.

[74] Andreas Feldmann. Colombia's polarizing peace efforts. 2019.

[75] Wei Feng and Jianyong Wang. Retweet or not?: personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 577–586. ACM, 2013.

[76] J. C. Figuereo Benítez, F. González Quiñones, and J. D. Machín Mastromatteo. Instagram como objeto de estudio en investigaciones recientes. una revisión de literatura con enfoque en revistas científicas. *Ámbitos: Revista internacional de comunicación*, 53:9–23, 2021.

[77] M. P. Fiorina and S. J. Abrams. Political polarization in the american public. *Annual Review of Political Science*, 11:563–588, 2008.

[78] Morris P. Fiorina, J.S. Abrams, and J.C. Pope. *Culture War? The Myth of a Polarized America*. Pearson Longman, New York, 2006.

[79] D. Flores-Márquez and R. González Reyes. En busca de coordenadas metodológicas para estudiar la cultura digital. In D. Flores-Márquez and R. González Reyes, editors, *La imaginación metodológica. Coordenadas, rutas y apuestas para el estudio de la cultura digital*, pages 15–23. Tintable, 2021.

[80] B. Focas. *La trama de la inseguridad: percepciones, medios de comunicación y vida cotidiana*. Unsam- edita, Buenos Aires, 2020. en prensa.

[81] P. Fortuna, J. Soler, and L. Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

[82] María Pilar García-Guadilla and Ana Mallen. Polarization, participatory democracy, and democratic erosion in venezuela's twenty-first century socialism. *The ANNALS of the American Academy of Political and Social Science*, 681(1):62–77, 2019.

[83] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

[84] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90. ACM, 2017.

[85] R Kelly Garrett, Shira Dvir Gvirsman, Benjamin K Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal. Implications of pro-and counterattitudinal information exposure for affective polarization. *Human communication research*, 40(3):309–332, 2014.

[86] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[87] B. González-Bustamante. Evaluando twitter como indicador de opinión pública: una mirada al arribo de bachelet a la presidencial chilena 2013. *Revista SAAP*, 9(1):119–141, 2015.

[88] Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[89] Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. Stance and influence of twitter users regarding the brexit referendum. *Computational social networks*, 4 (1):6, 2017.

[90] A. Gruzd, J. Lannigan, and K. Quigley. Examining government cross-platform engagement in social media: Instagram vs twitter and the big lift project. *Government Information Quarterly*, 35(4):579–587, 2018.

[91] Pedro Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[92] J. R. Gusfield. *La cultura de los problemas públicos: el mito del conductor alcoholizado versus la sociedad inocente*. Siglo Veintiuno Editores, 2014.

[93] AT Hadgu, Kiran Garimella, and Ingmar Weber. Political hashtag hijacking in the us in proceedings of the 22nd international conference on world wide web (pp. 55–56), 2013.

[94] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[95] U. Hasebrink and A. Hepp. How to research cross-media practices? investigating media repertoires and media ensembles. *Convergence*, 23(4):362–377, 2017. doi: https://doi.org/10.1177/1354856517700384.

[96] Sara B Hobolt. The brexit vote: a divided nation, a divided continent. *Journal of European public policy*, 23(9):1259–1277, 2016.

[97] JL Hodges Jr. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.

[98] Sounman Hong. Online news on twitter: Newspapers' social media adoption and their online readership. *Information Economics and Policy*, 24(1):69–74, 2012.

[99] Y. Hua, T. Ristenpart, and M. Naaman. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 272–282, 2020. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/7298`.

[100] Martín Igal Browarnik, Juan Manuel Ortíz de Zárate, and Esteban Feuerstein. Identificacion 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.

[101] S. Iyengar and K. S. Hahn. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59:19–39, 2009.

[102] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and J. Westwood. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146, 2019.

[103] Shanto Iyengar. The accessibility bias in politics: Television news and public opinion. *International Journal of Public Opinion Research*, 2(1):1–15, 1990.

[104] Shanto Iyengar and Victor Ottati. Cognitive perspective in political psychology. In *Handbook of social cognition*, pages 159–204. Psychology Press, 2014.

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

[106] K. Jaidka, S. Guntuku, and L. Ungar. Facebook versus twitter: Differences in self-disclosure and trait prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[107] Myungha Jang. Probabilistic models for identifying and explaining controversy. 2019.

[108] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2069–2072, 2016.

[109] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[110] Klaus Jensen and Karl Rosengren. Cinco tradiciones en busca del público. In *En busca del público. Recepción, televisión, medios*. Gedisa Editorial, Barcelona, 1997.

[111] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[112] N. Kaplan, D. K. Park, and T. N. Ridout. Dialogue in american political campaigns? an examination of issue convergence in candidate television advertising. *American Journal of Political Science*, 50(3):724–736, 2006. doi: https://doi.org/10.1111/j.1540-5907.2006.00212.x.

[113] R. Karlsen and B. Enjolras. Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with twitter data. *The International Journal of Press/Politics*, 21(3):338–357, 2016. doi: https://doi.org/10.1177/1940161216645335.

[114] S. Kelley and T. W. Mirer. The simple act of voting. *American Political Science Review*, 68 (2):572–591, 1974. doi: https://doi.org/10.2307/1959506.

[115] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[116] Gabriel Kessler and Gabriel Alejandro Vommaro. La era de las sensibilidades autoritarias. 2018.

[117] Gabriel Kessler, Brenda Focás, JUAN ZÁRATE, MANUEL ORTIZ DE, and ESTEBAN FEUERSTEIN. Los divergentes en un escenario de polarización. un estudio exploratorio sobre los "no polarizados" en controversias sobre noticias de delitos en la televisión argentina. *Revista SAAP*, 14(2):311–340, 2020.

[118] Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14893–14901, 2021.

[119] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[120] S. Kokoska and D. Zwillinger. *CRC standard probability and statistics tables and formulae*. Crc Press, 2000. doi: https://doi.org/10.1201/9781420050264.ch3.

[121] Emily Kubin and Christian Von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

[122] Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P Gummadi, and Saptarshi Ghosh. Characterizing information diets of social media users. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[123] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee, 2018.

[124] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2335–2338. ACM, 2012.

[125] Michael LaCour. A balanced news diet, not selective exposure: Evidence from a direct measure of media exposure. In *APSA 2012 Annual Meeting Paper*, 2015.

[126] Preethi Lahoti, Kiran Garimella, and Aristides Gionis. Joint non-negative matrix factorization for learning ideological leaning on twitter. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 351–359. ACM, 2018.

[127] A. O. Larsson. Green light for interaction: Party use of social media during the 2014 swedish election year. *First Monday*, 2015. doi: https://doi.org/10.5210/fm.v20i12.5966.

[128] Y. Lelkes, G. Sood, and S. Iyengar. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1):5–20, 2017.

[129] M. Levendusky and N. Malhotra. Does media coverage of partisan polarization affect political attitudes? *Political Communication*, 33(2):283–301, 2016.

[130] Steven Levitsky and Daniel Ziblatt. *How democracies die*. Crown, 2019.

[131] S. C. Lewis and O. Westlund. Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda. *Digital journalism*, 3(1):19–37, 2015. doi: https://doi.org/10.1080/21670811.2014.927986.

[132] J Luna. ¿ es posible la articulación entre movimientos sociales y partidos políticos en el mundo contemporáneo. *Política y movimientos sociales en Chile*, pages 39–61, 2021.

[133] N. Lupu, V. Oliveros, and L. Schiumerini. El pulso de la democracia en argentina y la región: datos de lapop y apes 2019. In *Simposio llevado a cabo en la Webinar del Departamento de Ciencia Política y Estudios Internacionales*. Universidad Torcuato Di Tella, 2020.

[134] N. Lupu, M. V. Ramírez Bustamante, and E. J. Zechmeister. Social media disruption: Messaging mistrust in latin america. *Journal of Democracy*, 31(3):160–171, 2020.

[135] P. C. López-López and J. Vásquez-González. Agenda temática y twitter: elecciones presidenciales en américa latina durante el período 2015-2017. *Profesional de la Información*, 27 (6):1204–1214, 2018. doi: https://doi.org/10.3145/epi.2018.nov.04.

[136] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

[137] Neil A Macmillan and C Douglas Creelman. *Detection theory: A user's guide*. Psychology press, 2004.

[138] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999. doi: https://doi.org/10.1353/lan.2002.0150.

[139] J. R. Martin and P. R. White. *The language of evaluation*, volume 2. Palgrave Macmillan, London, 2003.

[140] L. Mason. 'i disrespectfully agree': The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1):128–145, 2015.

[141] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505, 2017.

[142] J. Maxwell. Using numbers in qualitative research. *Qualitative Inquiry*, 16(6):475–482, 2010.

[143] Nolan McCarty, Keith T Poole, and Howard Rosenthal. *Polarized America: The dance of ideology and unequal riches*. mit Press, 2016.

[144] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. doi: https://doi.org/10.1146/annurev.soc.27.1.415.

[145] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[146] Zachary Mider. What kind of man spends millions to elect ted cruz? *Bloomberg Politics*, 2016.

[147] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[148] E. Mitchelstein and P. J. Boczkowski. Juventud, estatus y conexiones. explicación del consumo incidental de noticias en redes sociales. *Revista mexicana de opinión pública*, (24): 131–145, 2018.

[149] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[150] S. Moscovici. Social influence and conformity. In G. Lindzey and E. Aronson, editors, *Handbook of social psychology*, pages 347–412. Random House, New York, 1985.

[151] Yascha Mounk. The people vs. democracy: Why our freedom is in danger and how to save it. In *The People vs. Democracy*. Harvard University Press, 2018.

[152] Sean A Munson, Stephanie Y Lee, and Paul Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[153] A. C. Müller and S. Guido. *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc., 2016.

[154] Renáta Németh. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313, 2023.

[155] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[156] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[157] R. K. Nielsen and K. C. Schrøder. The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison. *Digital Journalism*, 2(4):472–489, 2014. doi: https://doi.org/10.1080/21670811.2013.872420.

[158] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Online political discourse in the trump era. *arXiv preprint arXiv:1711.05303*, 2017.

[159] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[160] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972): 137–144, 2023.

[161] P. Ortellado, E. Solano, and M. Moretto. Uma sociedade polarizada. In I. Jinkins, K. Doria, and M. Cleto, editors, *Por que gritamos golpe*, pages 159–164. Bon Tempo, Sao Paulo, 2016.

[162] J. M. Ortiz de Zarate and E. Feuerstein. Vocabulary-based method for quantifying controversy in social media. *ICCS2020*, 2020. Accepted for publication.

[163] J. M. Ortiz de Zarate, Marco Di Giovanni, E. Feuerstein, and Marco Brasmbilla. Measuring controversy in social networks through nlp. Inédito, 2020.

[164] D. Owen. *New media and political campaigns*. 2017. doi: https://doi.org/10.1093/oxfordhb/9780199793471.013.016.

[165] Zizi Papacharissi. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283, 2004.

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

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

[168] F. O. Paulino and S. Waisbord. Las narrativas del populismo reaccionario: Bolsonaro en twitter durante la pandemia. *Mediapolis–Revista de Comunicação, Jornalismo e Espaço Público*, 12:33–48, 2021. doi: https://doi.org/10.14195/2183-6019_12_2.

[169] Nathaniel Persily. Can democracy survive the internet? *J. Democracy*, 28:63, 2017.

[170] J. R. Petrocik. Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, pages 825–850, 1996. doi: https://doi.org/10.2307/2111797.

[171] Thomas F Pettigrew and Linda R Tropp. Does intergroup contact reduce prejudice? recent meta-analytic findings. In *Reducing prejudice and discrimination*, pages 103–124. Psychology Press, 2013.

[172] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[173] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, Berlin, Heidelberg, 2005.

[174] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

[175] Ashwin Rajadesingan and Huan Liu. Identifying users with opposing opinions in twitter debates. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 153–160. Springer, 2014.

[176] Giorgia Ramponi, Marco Brambilla, Stefano Ceri, Florian Daniel, and Marco Di Giovanni. Vocabulary-based community detection and characterization. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1043–1050, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359337. doi: 10.1145/3297280.3297384. URL `https://doi.org/10.1145/3297280.3297384`.

[177] Giorgia Ramponi, Marco Brambilla, Stefano Ceri, Florian Daniel, and Marco Di Giovanni. Content-based characterization of online social communities. *Information Processing & Management*, page 102133, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102133. URL `http://www.sciencedirect.com/science/article/pii/S0306457319303516`.

[178] Marnie E Rice and Grant T Harris. Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. *Law and human behavior*, 29(5):615–620, 2005.

[179] R. Rogers. Digital methods for cross-platform analysis. In *The SAGE handbook of social media*, pages 91–110. 2017.

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

[181] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015. doi: https://doi.org/10.1145/2684822.2685324.

[182] E. Salgado Andrade. Twitter en la campaña electoral de 2012. *Desacatos*, 42:217–232, 2013. doi: https://doi.org/10.29340/42.78.

[183] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[184] Facundo Sapienza and Pablo Groisman. Distancia de fermat y geodesicas en percolacion euclidea:teoriaa y aplicaciones en machine learning. *Msc Thesis*, 2018. URL `http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2018/Sapienza.pdf`.

[185] D. A. Scheufele. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society*, 3(2-3):297–316, 2000. doi: https://doi.org/10.4324/9781315679402-5.

[186] D. A. Scheufele and S. Iyengar. The state of framing research: A call for new directions. In *The Oxford handbook of political communication theories*, pages 1–26. 2012. doi: https://doi.org/10.1093/oxfordhb/9780199793471.013.47.

[187] Elisa Shearer and Jeffrey Gottfried. News use across social media platforms 2017. *Pew Research Center*, 7, 2017.

[188] N. Spierings and K. Jacobs. Political parties and social media campaigning. *Acta Politica*, 54(1):145–173, 2019. doi: https://doi.org/10.1057/s41304-020-00306-6.

[189] Leo G Stewart, Ahmer Arif, and Kate 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.

[190] S. Stier, A. Bleier, H. Lietz, and M. Strohmaier. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political Communication*, 35(1):50–74, 2018. doi: https://doi.org/10.1080/10584609.2017.1334728.

[191] Cass Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton university press, 2018.

[192] M. Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, Cham, 2017. doi: https://doi.org/10.1007/978-3-319-43639-5_7.

[193] A. M. Thorhauge and S. Lomborg. Cross-media communication in context: A mixed-methods approach. *MedieKultur: Journal of Media and Communication Research*, 32(60):16–p, 2016. doi: https://doi.org/10.7146/mediekultur.v32i60.22090.

[194] Trang Tran and Mari Ostendorf. Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779*, 2016.

[195] Damian Trilling. Two different debates? investigating the relationship between a political debate on tv and simultaneous comments on twitter. *Social science computer review*, 33(3):259–276, 2015.

[196] Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. From liberation to turmoil: Social media and democracy. *J. Democracy*, 28:46, 2017.

[197] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan 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.

[198] Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, 10(3):2018, 2018.

[199] Marshall Van Alstyne and Erik Brynjolfsson. Electronic communities: Global villages or cyberbalkanization?(best theme paper). *ICIS 1996 Proceedings*, page 5, 1996.

[200] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[201] José Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.

[202] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[203] Tommaso Venturini, Mathieu Jacomy, and Pablo Jensen. What do we see when we look at networks. an introduction to visual network analysis and force-directed layouts. *An introduction to visual network analysis and force-directed layouts (April 26, 2019)*, 2019.

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

[205] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[206] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. *Twitter and society*, volume 89. Peter Lang, 2014.

[207] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, 2017. doi: https://doi.org/10.1145/3038912.3052591.

[208] Han Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.

[209] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.

[210] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2-3):183–207, 2018.

[211] Sarita Yardi and Danah Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society*, 30(5):316–327, 2010.

[212] Arjumand Younus, M Atif Qureshi, Muhammad Saeed, Nasir Touheed, Colm O'Riordan, and Gabriella Pasi. Election trolling: analyzing sentiment in tweets during pakistan elections 2013. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 411–412, 2014.

[213] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[214] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[215] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[216] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[217] Xiaojin Zhur and Zoubin Ghahramanirh. Learning from labeled and unlabeled data with label propagation. 2002.