Reconocimiento y cuantificación de polarización ideológica en redes sociales

Nicolas Alberto Ortiz Aristizabal
Pontificia Universidad Javeriana Cali
Cali, Colombia
nicortiz@javerianacali.edu.co

Junio 2023

Resumen

A lo largo de este trabajo se describen los objetivos y acercamientos de las ciencias de la computación en el análisis de la opinión en redes sociales, se exponen diferentes modelos matemáticos y computacionales que se han implementado para medir y simular la polarización ideológica en redes sociales y posteriormente se expone una metodología para realizar pruebas con datos reales, obtenidos a partir de las interacciones en redes sociales, en los modelos de polarización. Además, se presenta un algoritmo que cuantifica la opinión de mensajes publicados en redes sociales empleando un modelo pre-entrenado de inteligencia artificial, de modo que las opiniones sean representadas en los parámetros de los modelos de polarización. También se implementan diferentes grafos de influencia que definen algunas formas como se pueden representar los intercambios de opinión en redes sociales. Por último, se evalúan las opiniones de un conjunto de datos de Reddit y se generan las simulaciones de polarización empleando el modelo presentado por el grupo Avispa, a partir de las cuales se presentan límites y retroalimentaciones del modelo actual y se proponen nuevos enfoques que pueden favorecer a la simulación de la opinión y la polarización en redes sociales desde las ciencias de la computación.

1. Introducción

Las redes sociales han generado un gran impacto no sólo en el mercado y la publicidad, sino también en campos diversos como la política. Dada la facilidad y los diferentes medios por los cuales compartir y difundir la opinión del individuo, los discursos civiles se han vuelto algo común para la sociedad. Debido a esta realidad, en los últimos años ha nacido una tendencia a estudiar y comprender el comportamiento de las interacciones sociales dentro de las redes sociales y cómo afectan estas a las creencias de cada persona y en general a la opinión pública.

En el presente documento se detalla un proyecto de investigación que tiene como objetivo analizar diferentes estudios sobre uno de los fenómenos psicológicos y sociológicos más famosos en las redes sociales, relacionado a la opinión, las creencias y la polarización ideológica. Además, desarrollar cómo este puede ser estudiado, medido y simulado desde las ciencias de la computación, empleando técnicas y estrategias basadas en modelos matemáticos, inteligencia artificial y modelos probabilísticos.De formas más específicas, el proyecto tiene como objetivo analizar el modelo matemático planteado por el grupo de investigación Avispa, detallando así sus fundamentos teóricos y contrastarlos con estudios similares. Además emplear el mismo para realizar pruebas con datos reales basados en interacciones en redes sociales, con el fin de identificar las características del modelo que favorecen al análisis en cuestión, así como sus limitaciones.

2. Marco Teórico

El fenómeno social de la polarización fue definido rigurosamente por primera vez por los economistas Esteban y Ray [1]. Su medida de polarización es influyente, y es la que adoptamos en este artículo. Li

et al. [2] fueron los primeros en modelar el consenso y la polarización en las redes sociales. Sin embargo, como la mayoría de los otros trabajos, no cuantifican la polarización, sino que se centran en cuándo y bajo qué condiciones una población llega a un consenso. Proskurnikov et al. [3] investigó la formación de consenso o polarización en las redes sociales, pero consideró la polarización de las habitaciones como una falta de consenso, en lugar de un fenómeno en sí mismo.En ciencias sociales, la polarización grupal se refiere a la tendencia natural de los grupos a tomar decisiones más extremas que los individuos. Es decir, que el fenómeno nace del comportamiento de cada individuo, con sus creencias y opiniones, a la hora de interactuar en un grupo con otras personas, donde estas pueden compartir su creencia o estar en desacuerdo con la misma. Además, define cómo a raíz de esta variedad de opiniones compartidas se generan efectos en las propias personas y en la opinión pública, tal que las decisiones grupales pueden ser más radicales que las propias de cada individuo.

Al abordar este problema desde las ciencias de la computación, se encuentran diferentes estrategias que favorecen al análisis de los datos y el cálculo de los resultados. Técnicas como las referentes a inteligencia artificial, la simulación, los modelos matemáticos y computacionales son algunas de las que suelen relacionarse en trabajos similares a este. Este proyecto, parte de un modelo desarrollado por el grupo de investigación Avispa, el cual se centra en adaptar las teorías sociales y psicológicas afines al reconocimiento de la polarización en las redes sociales, de modo que este se pueda observar como una medida cuantitativa. Más concretamente, el presente proyecto busca enfatizar en las metodologías que permitan transformar los datos presentes en las interacciones de las redes sociales, los cuales son totalmente cualitativos (tales como mensajes y suscripciones), a elementos que sean entendidos por el modelo matemático y computacional planteado por Avispa. Del mismo modo, se piensa enfatizar en estrategias de simulación, probabilidad y predicción basadas en las proyecciones dadas por el mismo modelo o un agente planteado desde la inteligencia artificial.Estas estrategias han sido trabajadas en proyectos similares, como Sirbu et al. [4] el cuál utiliza un modelo algo similar al que se espera analizar en este proyecto, que se actualiza probabilísticamente. En él, se investigan los efectos del sesgo algorítmico en la polarización, contando el número de grupos de opinión, interpretando un solo grupo de opinión como consenso, en lugar de medir directamente la polarización en sí. Leskovec et al. [5] desarrolla redes sociales simuladas y observa la formación de grupos a lo largo del tiempo. Sin embargo, su trabajo no representa una medida formal de polarización.

2.1. Trabajos Relacionados

- Toward a Formal Model for Group Polarization in Social Networks [6]:
  Este trabajo fue realizado por integrantes del grupo de investigación Avispa y es el trabajo principal sobre el que se va a desarrollar la investigación. El trabajo presenta un primer acercamiento a la medición de polarización de grupos en redes sociales, a partir del estudio de diferentes fenómenos sociales y psicológicos. De modo que, plantea un modelo matemático basado en la interacción de agentes en un sistema. Además, se basa en parte de teoría económica para desarrollar la influencia que tiene cada agente en la medición de la polarización.

  Inicialmente en el desarrollo del trabajo se habla de las diferentes causas o interacciones que tienen influencia en la ideología de cada participante del grupo. Cuando 2 personas interactúan entre ellas, dependiendo de algunas variables como su conocimiento del otro y sus propias creencias, pueden tener diferentes tipos de comportamientos, ya sean de reforzar su propia creencia sobre el tema o por el contrario de dudar sobre ella. Posteriormente, se analizan cada uno de los fenómenos y se determinan como agentes, para describir un modelo matemático donde se cuantifica cada agente y se le da un valor de influencia en la medición a cada uno.

  Finalmente, se prueba el modelo, simulando el comportamiento de grupos con parámetros específicos (de extrema diferencia de ideologías, ideologías similares o ideologías normales) con el fin de dar una prueba superficial al modelo matemático y comprobando los comportamientos básicos del mismo. Sin embargo, no se realizan pruebas en conjuntos de datos reales de alguna red social.
- A model of opinion and propagation structure polarization in social media [7]:
  En este artículo se propone un modelo dinámico para el análisis de opinión, fundamentado en el cambio continuo de opiniones de un individuo al interactuar con otros. Los parámetros que se tienen en cuenta para este modelo son la opinión de cada individuo sobre un tema y el grado de conexión

o confianza que se tiene cada individuo con los demás.

Además, se realizan simulaciones sobre la propagación de múltiples opiniones a lo largo de una red social con múltiples individuos y se observa la evolución de los parámetros del modelo. A partir de los resultados de dichas simulaciones, se encontraron patrones de comportamiento de la polarización en grupos y se determinó lo rápido que esta genera un cambio en la opinión de los individuos que participan. Además, se realizaron simulaciones con datos obtenidos en "Twitter" con el fin de desarrollar pruebas con datos reales.

3. Resultados

3.1. Descripción del algoritmo

La metodología que se siguió para la preparación y transformación de los datos obtenidos del conjunto de datos de Reddit [8], de modo que se cuantificara la creencia de los mensajes y se dispusiera un grafo de influencia que permitiera correr las simulaciones, se puede dividir en los siguientes pasos:

1. Extraer los mensajes usando la librería pandas y separarlos en conjuntos de mensajes que se refieran al mismo tema. En el caso de la base de datos empleada en este trabajo, se separaron en base a el hilo o subreddit al que hacen parte.
2. Descargar y preparar el modelo de análisis de sentimiento que se va a emplear y definir las operaciones necesarias para leer y analizar la predicción del mismo. Como se mencionaba anteriormente se decidió usar el modelo pre entrenado de las emociones circumplejo, para ellos se usó la interfaz de huggingface en python por medio de la librería transformers, esta permite importar fácilmente modelos pre entrenados por la comunidad. Además se debe etiquetar correctamente las predicciones del modelo, las cuales son la medición entre 0 y 1 de las 7 emociones presentadas en el capítulo anterior. Con base en los trabajos [7, 9].
3. Clasificar la creencia de los mensajes basado en los resultados del modelo circumplejo [10]. Para esto se siguió la estrategia en la cual se realiza una combinación lineal de las emociones negativas, el cual dará siempre como resultado un valor entre 0 y 1, el cual se tomará como la creencia del agente frente al tema. De modo que para cada uno de los mensajes de un tema en el conjunto de datos, se corre el modelo y se calcula la creencia.
4. Generar el grafo de influencia basado en la estrategia a seguir. Como se describió anteriormente, definir lo que naturalmente sería la confianza o el impacto de la opinión de una persona en otra, es un gran reto a la hora de extraer la información de una red social y significa un sesgo para los datos dentro del modelo. Por ello para este trabajo se optaron por diferentes tipos de estrategias para definir el peso o la influencia entre cada par de agentes.
   - Basado en la inclinación política dada como parte de los datos, donde los agentes que pertenezcan a una misma inclinación política tienen un mayor valor de influencia (0.5) en comparación a los demás (0.1).
   - Basado en una distribución densa y uniforme, donde todos los agentes se relacionan con todos los demás participantes del foro con un mismo valor de influencia (0.1).
   - Basada en la distancia de las creencias entre los agentes, donde todos los usuarios que participan de un mismo foro están relacionados, pero la confianza inicial de unos y los otros depende de la distancia entre sus opiniones. Siendo 0.5 en caso que sean cercanas y 0.05 en caso contrario.
   - Basada en la distancia de los datos, donde únicamente los usuarios con una opinión similar están conectados entre ellos (0.5).
   - Basado en clases según la opinión de los agentes, se clasifican los agentes entre extremo negativo [0, 0.3], neutral (0.3, 0.7) y extremo positivo [0.7, 1]. Todos los agentes que participan del foro están relacionados y la influencia de los nodos que pertenecen al mismo extremo es mayor. Siendo 0.5 para los nodos que pertenecen a las clases extremas y 0.05 para todos los demás.
   - Basado en clases según la opinión de los agentes y clasificado igual que en la estrategia anterior, pero los agentes únicamente se relacionan con aquellos que pertenecen a su misma clase. Siendo 0.4 para los nodos que pertenecen a las clases extremas y 0.1 aquellos que pertenecen a la clase neutral.

Se realizaron múltiples simulaciones con cada uno de los grafos con el fin de optimizar los valores de las influencias para cada caso de forma que se logra evidenciar de forma clara las gráficas de la simulación de cada uno y sus impactos en el desarrollo de las opiniones de los agentes a través del tiempo.

3.2. Evaluación en el modelo

Para la evaluación del modelo desarrollado por el grupo de investigación Avispa se empleó los algoritmos presentados en el artículo [6], para esto se utilizaron los subreddits: "republicans", "Capitalism" y "alltheleft" del conjunto de datos [8]. Para el análisis de los resultados del modelo se realizaron 2 comparativas diferentes, la evolución de la creencia u opinión de todos los agentes a través del tiempo y la evolución de la polarización a través del tiempo para cada uno de los grafos de influencia en cada uno de los temas.

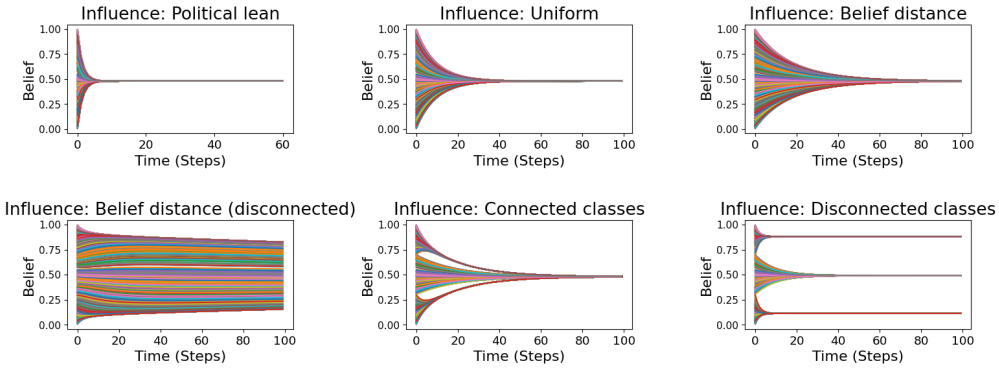

Figura 1: Simulación de opiniones de los agentes en el tema Republicans

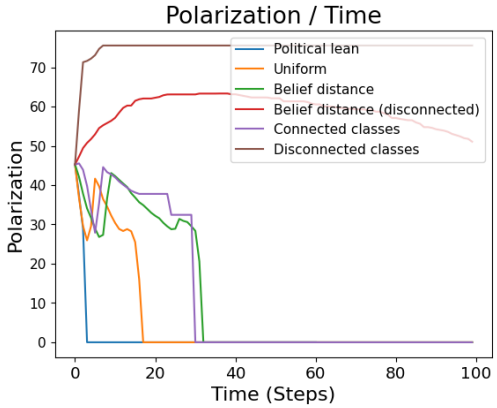

Figura 2: Simulación de polarización de la opinión en el tema Republicans

Inicialmente al comparar los resultados de los diferentes temas o subreddits no es posible identificar características que supongan una variación al modelo o su evolución, más allá de que para el caso de las simulaciones donde la opinión popular converge a un único punto, estas varían el tiempo o pasos que tardan en converger. Además en las gráficas del Capitalismo, es posible ver como una mayor cantidad de agentes que tienen una opinión cercana a 0 u a algún extremo, hace que la convergencia tarde más pasos y el valor al que converga se desvía ligeramente hacia ese extremo. A parte de lo mencionado, los tres temas presentan resultados muy similares, lo que implica que la diversidad que tienen las opiniones del conjunto de datos utilizado es suficiente para que el modelo las interprete de forma similar, sin exponer un resultado diferente entre ellas.

Por otra parte, la comparativa entre los diferentes tipos de influencia tiene grandes resultados en la forma como se simula la polarización a lo largo del tiempo. Para los grados de Inclinación política, uniforme, distancia de opinión y clases conectadas, podemos ver como la opinión converge en un punto cercano a la media 0.5. Entre ellos varía principalmente la cantidad de pasos que toma converger al punto medio, siendo el caso de la inclinación política donde menos tiempo se tarda y la distancia de opinión donde más. A partir de esto podemos evidenciar una relación inversa entre cuán fuertemente conexo sea el grafo de influencia y lo alto que sean las influencias entre los agentes con el tiempo que se tarda en llegar al punto de convergencia.

En el caso de los grafos de influencia desconectados (distancia de opinión desconectada y clases desconectadas) no se llega al punto de convergencia durante la simulación. En el análisis de la distancia de opinión desconectada se presenta una pequeña conversión entre los polos de la creencia para luego formarse muchas clases muy conectadas entre ellas y poco conectadas con las demás, de modo que tienen de converger muy lentamente. Por otro lado el grafo de influencia de clases desconectadas, muestra rápidamente como las 3 clases (Extremo negativo, neutro y extremo positivo) convergen individualmente en la media de cada uno, además debido a que las clases de opinión extrema tienen una mayor influencia, estas convergen antes que la neutral.

Un análisis interesante que destacar es el de las clases conectadas, debido a la estructura del grafo de influencia, donde los agentes que pertenecen a una misma clase tienen mayor influencia entre ellos que con los demás, inicialmente las opiniones de cada clase convergen, para luego converger la opinión de las tres clases en la media de todos los agentes.

Por último se realizaron todas las simulaciones del tema “republicans” empleando la función de actualización del modelo “Avispa” que intenta representar el fenómeno de sesgo de confirmación. Debido a la definición de esta función donde el impacto que tiene una interacción de un agente sobre otro está fuertemente afectada por su propia opinión, causando que las interacciones con agentes con opiniones similares a él tengan más peso en su propia opinión, la opinión de los agentes tarda más en converger, además se puede observar en las gráficas como la curva de convergencia de las opiniones más extremas es menos inclinada que con el método de actualización clásico.

3.3. Retroalimentación del modelo

Con base a los resultados obtenidos anteriormente podemos analizar algunas características del modelo presentado por el grupo de investigación Avispa:

- Las opiniones tienden a la media o a converger siempre que el grafo de influencias sea conexo. En otras palabras, siempre que haya algún camino entre cada par de agentes la polarización tiende a desaparecer. Esta conclusión también es presentada por el grupo de investigación en el artículo [11].
- Al representar el grafo de influencia de los datos obtenidos usando diferentes estrategias y evidenciando como únicamente en las que se presentan subgrafos desconectados se mantiene la polarización a través del tiempo, se puede afirmar la existencia de agentes en las redes sociales que o ignoran o no se ven influenciados por las interacciones con otros, incluso al participar de las mismas discusiones o interactuar entre ellos, lo cual corresponde al fenómeno psicológico llamado exposición selectiva. En consecuencia, la interpretación de los datos donde se definen las conexiones o influencia de los agentes con base en la interacción entre ellos no es del todo acertada y deben considerarse nuevas estrategias que permitan representar dicho comportamiento.
- Como la distribución de las creencias obtenidas en los datos reales no presenta una estructura extrema (a pesar de que se pueden identificar polos, también hay opiniones intermedias) el modelo no permite resaltar esas opiniones extremas, y actúa de manera similar a una distribución de opinión uniforme, como la presentada en los datos sintéticos.
- Desde el modelo podríamos concluir que siempre que el grafo de influencia sea conexo, la opinión del sistema debería de converger en algún momento, sin embargo esto parece negar la experiencia empírica que se vive hoy en día frente a los temas evaluados, por lo que se podría investigar nuevos fenómenos o funciones de actualización que permitiera representar o indagar sobre el tipo de comportamiento que evita que la opinión popular pueda converger.
- La representación de los datos en redes sociales tiene limitación a la hora de traducirse al modelo, lo cual limita lo expresivo que este pueda ser o lo cercano a la representación adecuada de las

interacciones en las redes sociales. Como se mencionaba anteriormente, la representación del grafo de influencia o la constancia de las interacciones, no se asemeja a la realidad de las redes sociales. El modelo plantea que para cada instante de tiempo todos los agentes interactúan con todos los demás de forma síncrona, sin embargo una simulación precisa de las interacciones sociales debería de considerar como la comunicación de un único agente modifica la de algunos otros sin influir en su propia opinión, siguiendo el comportamiento de redes sociales como Twitter o Reddit basado en foros o publicaciones.

3.4. Nuevos enfoques

Teniendo en cuenta las características expuestas anteriormente, el análisis realizado sobre los fenómenos que inducen a la polarización y los resultados obtenidos en este trabajo, se proponen los siguientes enfoques que se podrían explorar en busca de favorecer la simulación de la polarización en redes sociales reales del modelo y la representación de los diferentes fenómenos presentes en el intercambio de opiniones.

- Explorar fenómenos psico-sociales diferentes que dificultan el cambio de opinión de agentes que presentan una opinión extrema, de modo que al acercarse a una opinión extrema estos se vean muy poco influenciados por otras opiniones. O incluso influenciados de forma negativa como es el caso del efecto contraproducente. [6]
- Planear diferentes modelos que se alineen con el tipo de estructura de los datos. Grafos fuertemente conexos como los presentes en un foro o débilmente conexos como los presentes en una conversación tienen comportamientos, actualizaciones e influencias diferentes, que podrían ser tratados con mayor precisión si son estructurados de formas específicas para simular su comportamiento.
- Explorar nuevas metodologías de actualización de la opinión a través del tiempo. Como se mencionaba en la sección anterior, las características de las redes sociales basadas en publicaciones o foros son difícilmente representables por un método de actualización donde todos los agentes influyen en aquellos que tienen contacto por cada instante en el tiempo. Explorar métodos basados en la difusión de publicaciones o las interacciones entre grupos de nodos permitirá una mejor representación del comportamiento presente en las interacciones de las redes sociales.
- Con base en los resultados obtenidos en este trabajo, se podría considerar que la actualización dinámica de la influencia presentada por el modelo debería permitir que los agentes eliminen completamente la conexión que tienen unos con otros, de modo que los estados de divergencia que se encuentran en los grafos desconectados puedan ser representados como una consecuencia de la polarización, en lugar de únicamente una causa de ella. Además permitir que el modelo actualice no solo la influencia entre los agentes, sino también la conexión entre ellos, puede permitir explorar la topología del grafo generada durante la simulación del modelo, de modo que se puedan detectar otros fenómenos determinantes de la polarización como son las cámaras de eco.
- Permitir al modelo simular agentes obstinados. Los cuales mantienen una opinión extrema y constante, sin ser influenciados por otros.

4. Conclusiones

Luego de haber explorado la teoría detrás de la polarización ideológica como un fenómeno de los intercambios de opiniones entre individuos, los diferentes fenómenos psico-sociales ampliamente estudiados que promueven a la aparición de estos polos extremos de opinión del mismo modo como los factores externos a la conversación que se presentan con el fin de alterar la opinión popular frente a un tema y que se han visto fuertemente impulsados con el intercambio masivo de opiniones presente en las redes sociales. Además de analizar cada uno de estos factores y como las ciencias de la computación y el análisis de datos puede favorecer al estudio, simulación y reconocimiento de este fenómeno en las redes sociales, así como los acercamientos que se han realizado al tema y los retos presentes en el mismo. Se analizaron los retos que tienen que afrontar los modelos de polarización, tanto para representar la información presente en las redes sociales, como para simular de forma precisa el comportamiento de la opinión de múltiples individuos al interactuar entre sí de forma precisa, teniendo en cuenta los múltiples fenómenos conocidos que afectan de forma, en ocasiones, contraintuitiva el impacto que tiene una opinión sobre otra.

Sin embargo, los diferentes modelos que se han realizado hasta el momento permiten evidenciar algunos comportamientos de la polarización en redes sociales que vivimos día a día. Así como explorar y sacar algunas conclusiones sobre el impacto que pueden estar teniendo los fenómenos sociales en la

comunicación dentro de las redes sociales, además de evidenciar algunas tendencias y perspectivas gracias al análisis computacional, como es las redes topológicas o las cámaras de eco y como estas se hacen presentes y evolucionan fuertemente relacionadas a la polarización. Se presentó además una metodología y algoritmo para la recolección, análisis, clasificación y visualización de las redes de opinión a partir de las interacciones en redes sociales que siguen un formato tipo foro. El cual puede contribuir a las pruebas de los modelos de polarización con datos reales, así como la integración futura de dichos modelos para lograr aplicaciones como la medición de polarización o la detección de factores externos que promueven la aparición de este fenómeno.

Además a partir de las pruebas realizadas en el modelo de polarización presentado por el grupo avispa con conjunto de datos obtenido de la red social Reddit. Se logró evidenciar el comportamiento del modelo en casos reales e identificar retos en la representación de estos, como son la actualización de las conexiones dentro de la red por medio de la difusión de publicaciones en lugar de la influencia de todos los agentes simultáneamente, la representación de personas con opiniones extremistas e inflexibles o la inclusión de factores externos, como bots o agentes obstinados. También se logró evidenciar características presentadas por dicho trabajo, como la relación entre la aparición de polarización en el modelo y el grado de conexión del grafo de influencia. Por último basado en la investigación realizada sobre la polarización ideológica, diferentes acercamientos al tema que se han desarrollado en los últimos años y las conclusiones obtenidas de los resultados del modelo con datos reales, se presentaron diferentes propuestas que pueden permitir a futuros modelos de polarización una mejor representación de las interacciones presentes en las redes sociales, los fenómenos que son causa/consecuencia de la polarización y la representación de estructuras dentro de influencia que promueven la aparición de este fenómeno social.

Referencias

[1] J. Esteban and D. Ray, "The measurement of polarization," *Econometrica*, 1994.

[2] S. A. S. A. Z. Q. Li, L., "Consensus, polarization and clustering of opinions in social networks.," *IEEE Journal on Selected Areas in Communications*, 2013.

[3] M. A. C. M. Proskurnikov, A.V., "Opinion dynamics in social networks with hostile camps: Consensus vs. polarization," *IEEE Transactions on Automatic Control*, 2016.

[4] P. D. G. F. K. J. Sirbu, A., "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model," *PLOS ONE*, 2019.

[5] G. Y. Gargiulo, F., "The role of homophily in the emergence of opinion controversies," *preprint arXiv*, 2016.

[6] F. V. M. Alvim, S. Knight, "Toward a formal model for group polarization in social networks," *Lecture Notes in Computer Science series*, vol. 11760, 1960.

[7] T. M. Hafzh A. Prasetya, "A model of opinion and propagation structure polarization in social media," *Computational social networks. Springer*, 2020.

[8] N. Gajare, "Liberals vs conservatives on reddit [13000 posts]." `https://www.kaggle.com/datasets/neelgajare/liberals-vs-conservatives-on-reddit-13000-posts`, 2022.

[9] F. S. D. O. S. o. Movilizatorio, Ford Foundation, "Xenofobia y polarizacion para lograr inclusion y cohesion en colombia," 2020.

[10] J. Hartmann, "Emotion english distilroberta-base." `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

[11] S. K. S. Q. Mario S. Alvim, Bernardo Amorim and F. Valencia, "Polarization and belief convergence of agents in strongly-connected influence graphs," 2020.