Modelos matemáticos para medir la polarización en redes sociales: análisis y ejemplo de aplicación

Mathematical models for measuring polarization in social networks: analysis and application example

Trabajo de Fin de Grado
Curso 2024–2025

Autor
Laura Rodrigo Cañete

Director
Rafael Caballero Roldán

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Matemáticas
Universidad Complutense de Madrid

Modelos matemáticos para medir la polarización en redes sociales: análisis y ejemplo de aplicación

Mathematical models for measuring polarization in social networks: analysis and application example

Trabajo de Fin de Grado en Matemáticas

Autor
Laura Rodrigo Cañete

Director
Rafael Caballero Roldán

Convocatoria: *Junio 2025*

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Matemáticas
Universidad Complutense de Madrid

23 de Junio de 2025

Dedicatoria

A mis padres, que serán los que lean este trabajo con mayor entusiasmo.

Agradecimientos

Quiero agradecer a mi tutor, Rafael Caballero Roldán, por ser un excelente profesor y por haberme ayudado en todo lo que ha podido. Ha confiado en mis decisiones y me ha animado mientras respetaba mi organización y tiempos. Ha animado cada reunión con su sentido del humor y ha convertido este proceso en algo que realmente disfruté y del que aprendí.

Resumen

Modelos matemáticos para medir la polarización en redes sociales: análisis y ejemplo de aplicación

Este trabajo analiza y compara desde una perspectiva matemática cuatro índices de polarización política con el fin de evaluar la división de opiniones en un sistema bipartidista, tomando como caso de estudio las elecciones presidenciales de Estados Unidos. Para ello, se estudian los desarrollos de tres índices conocidos: el Índice de Polarización Gravitatoria, desarrollado recientemente en un contexto similar de análisis de redes sociales, y los modelos clásicos de Foster y Wolfson y de Esteban y Ray, que sientan bases teóricas y se han utilizado ampliamente en la literatura. Además, se propone un nuevo índice original, el Índice Beta de Polarización, diseñado específicamente para adaptarse a las características del estudio de las redes sociales y la clasificación política. A partir de una base de datos compuesta por *tweets* publicados en los periodos de las elecciones estadounidenses de 2016, 2020 y 2024, se diseñan variables que capturan información relevante de las opiniones políticas de los usuarios y se aplican los distintos modelos de polarización con un programa informático. Se comparan los resultados obtenidos para evaluar la efectividad de cada índice para capturar la polarización política.

Palabras clave

Polarización, bipolarización política, modelo, índice, redes sociales, Twitter, elecciones bipartidistas

Abstract

Mathematical models for measuring polarization in social networks: analysis and application example

This work analyzes and compares four political polarization indices from a mathematical perspective to evaluate the division of opinions in a two-party system, using the U.S. presidential elections as a case study. It focuses on three well-known indices: the Gravity Polarization Index, recently developed in a similar context of social media analysis, and the classic models by Foster and Wolfson, and Esteban and Ray, which provide theoretical foundations and have been widely used in the literature. Additionally, a new original index is proposed–the Beta Polarization Index–specifically designed to fit the characteristics of social media and political classification. Using a database of tweets posted during the 2016, 2020, and 2024 U.S. election periods, variables are created to capture relevant information about users' political opinions. The different polarization models are applied using custom software, and the results are compared to evaluate how effectively each index captures political polarization.

Keywords

Polarization, political bipolarization, model, index, social media, Twitter, two-party elections

Índice

Capítulo 1

Introducción

1.1. El Problema de La Polarización

La noción de polarización aparece con frecuencia en los debates públicos para describir situaciones en las que las posiciones ideológicas de los individuos o grupos sociales se muestran fuertemente enfrentadas. Sin embargo, se trata de un concepto cuya definición precisa y cuantificación rigurosa resultan complejas. En este trabajo se aborda un estudio matemático de algunas de las principales medidas de polarización propuestas en la literatura, analizando sus fundamentos teóricos, propiedades formales y potencial de aplicación al análisis de redes sociales.

Desde un enfoque social, la polarización hace referencia al proceso por el cual las opiniones, actitudes o creencias de los individuos dentro de una sociedad tienden a agruparse en extremos opuestos, reduciendo los espacios de consenso y aumentando la confrontación. En el contexto de las redes sociales, este fenómeno se ve amplificado por la dinámica de los algoritmos de recomendación, la creación de cámaras de eco y la facilidad para interactuar selectivamente con contenido afín (Hartmann et al., 2025). Como resultado, los usuarios tienden a exponerse principalmente a discursos que refuerzan sus propias ideas, un fenómeno conocido como *filter bubble*, lo que puede intensificar la separación entre grupos ideológicos y dificultar el diálogo constructivo (Pariser, 2011).

Desde una perspectiva matemática, la polarización se entiende como una propiedad de la distribución de una variable sobre una población, que refleja la concentración de individuos en torno a posiciones alejadas entre sí y escasamente representadas en valores intermedios. A diferencia de la desigualdad, que se centra en la dispersión, la polarización enfatiza la formación de grupos cohesionados y distantes dentro del espacio de opiniones, lo que suele formalizarse mediante medidas que capturan tanto la distancia entre grupos como su masa relativa.

Es fundamental dejar clara desde el inicio la distinción entre polarización y desigualdad, ya que se trata de conceptos diferentes tanto en su fundamento teórico como en las medidas que los cuantifican, tal y como se irá evidenciando a lo largo del trabajo. Una forma de entender esta diferencia es señalar que las medidas clásicas de desigualdad no permiten distinguir bien entre dos situaciones distintas: que todos los individuos tiendan hacia una misma posición intermedia (convergencia

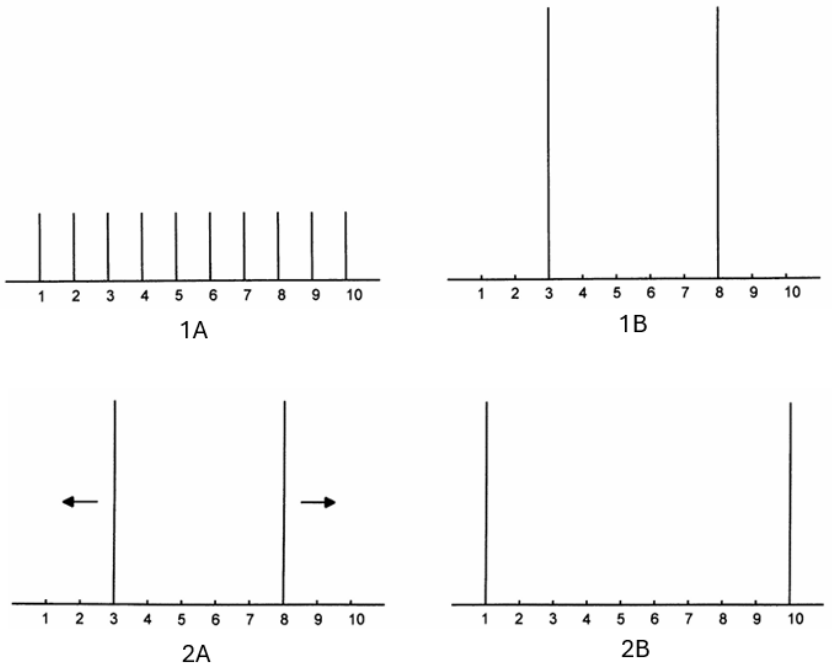

Figura 1.1: Dos ejemplos de distribuciones en evolución que muestran la diferencia entre los conceptos de desigualdad y polarización. Fuente: Esteban y Ray (1994).

global), o que se agrupen en varios bloques separados con opiniones similares dentro de cada grupo (convergencia local o formación de polos). Por ejemplo, si analizamos la evolución de la distribución de riqueza de distintos países, podríamos observar una convergencia interna entre países de bajo crecimiento y otra entre países de alto crecimiento, generando así dos polos claramente diferenciados. Esta estructura implicaría un aumento de polarización, sin embargo, cualquier medida clásica de desigualdad registraría una disminución, ya que se reduciría la dispersión global.

Para ilustrar mejor esta diferencia conceptual, a continuación se presentan dos ejemplos de distribuciones en evolución en la figura 1.1. En ella, supongamos que las distribuciones representadas describen la renta de una población. La altura de cada barra indica la proporción de personas en cada grupo, y los valores del eje horizontal corresponden a niveles de ingreso normalizados. En el primer ejemplo, la distribución inicial (imagen 1A) muestra una situación poco polarizada –ya que cada nivel de renta está igualmente representado– pero muy desigual, pues existen individuos con ingresos desde 1 hasta 10. Esta distribución evoluciona (imagen 1B) hacia otra con menor desigualdad bajo cualquier medida coherente con el orden de Lorenz[1]–solo hay personas con ingresos 3 y 8–, aunque la polarización ha aumentado. En el segundo ejemplo de esta misma figura (imágenes 2A y 2B), partimos de dos grupos diferenciados con ingresos de clase media baja y media alta. Con el tiempo, estos evolucionan hacia una estructura compuesta por pobres y ricos. Para ver cual

[1]La curva de Lorenz es una herramienta gráfica que representa la distribución acumulada de una variable en una población. Cuando una distribución «domina» a otra en el orden de Lorenz, se considera menos desigual: su curva está por encima de la otra en todo punto. Hablaremos de esta noción con más detalle más adelante.

presenta mayor polarización observamos que en términos de homogeneidad interna, ambas distribuciones son similares. Sin embargo, la heterogeneidad entre los grupos aumenta en la imagen 2B, lo que permite concluir que esta configuración es más polarizada. No obstante, en este caso, la desigualdad también ha aumentado. Por tanto, no se afirma que el concepto de polarización entre necesariamente en conflicto con el de desigualdad, sino que ambos capturan aspectos distintos de la estructura social.

1.2. Objetivos

El objetivo general del trabajo es analizar en profundidad y comparar cuatro medidas de polarización: el Índice de Polarización Gravitatoria, de aparición reciente y desarrollado en un contexto similar al nuestro de análisis de redes sociales y clasificación política, en este caso en Venezuela; los modelos de Foster y Wolfson y de Esteban y Ray, que corresponden a dos índices clásicos ampliamente reconocidos en la literatura; y la cuarta es una medida original propuesta en este trabajo, denominada Índice Beta de Polarización y diseñada para adaptarse a las particularidades del caso de estudio.

Además, se estudia la viabilidad y utilidad de dichas medidas en el caso concreto de las elecciones presidenciales bipartidistas en Estados Unidos. Para ello, se toma como base una recopilación de *tweets* que mencionan a los candidatos a las elecciones presidenciales estadounidenses publicados en periodos en torno a las tres últimas elecciones nacionales de 2016, 2020 y 2024. A partir de esta base de datos, se procesan las opiniones políticas de los usuarios y se construyen variables cuantitativas que permiten aplicar y comparar los diferentes modelos de polarización.

Por tanto, los objetivos del trabajo se pueden resumir en:

- Diseñar variables que capturen información relevante de nuestros datos políticos para usarlas como valores de entrada de los modelos de polarización.
- Estudiar y clasificar medidas de polarización desarrolladas en la literatura académica, prestando atención a sus formulaciones matemáticas, propiedades estructurales y analizándolas desde el punto de vista político.
- Diseñar y proponer un nuevo índice de polarización adaptado a las particularidades del problema abordado, combinando elementos de los enfoques clásicos con nuevas ideas extraídas del análisis empírico.
- Aplicar y comparar diferentes modelos de polarización sobre un conjunto real de datos extraídos de Twitter, ahora conocida como X, evaluando su capacidad para capturar la evolución de la opinión pública en un entorno político polarizado.

Aunque ciertamente este trabajo se apoya en una extensa revisión bibliográfica, también contiene múltiples contribuciones originales, que enumeramos a continuación para facilitar su identificación por parte del lector. Tres de los índices de polarización considerados ya existían en la literatura, pero se encontraban presentados

en contextos distintos y de forma dispersa. Aquí se ha realizado un esfuerzo por unificarlos bajo un marco común, homogeneizando su presentación y analizando de manera comparativa sus propiedades desde una perspectiva política. Además, se ha adaptado todo el tratamiento teórico al contexto de análisis de opinión pública en redes sociales, reformulando los índices con nuevas variables diseñadas específicamente para este estudio. Como aportación adicional, se propone un nuevo índice original que extiende los enfoques clásicos al incorporar la entropía individual, capturando así la firmeza o ambigüedad de las opiniones, un aspecto poco tratado por las métricas tradicionales centradas exclusivamente en promedios. Este índice resulta especialmente relevante para el análisis de dinámicas en redes sociales, donde la intensidad de las convicciones es tan significativa como su orientación. Finalmente, otra contribución adicional se basa en la recabación de datos reales y la implementación en *Python* de los cuatro índices para compararlos mediante un caso práctico, lo que permite ilustrar sus diferencias y complementariedades de forma empírica.

1.3. Plan de Trabajo

El plan de trabajo seguido para alcanzar estos objetivos se estructura en las siguientes fases:

1. **Revisión bibliográfica**: se llevó a cabo un estudio de la literatura existente sobre polarización, tanto desde un punto de vista teórico como aplicado. Se estudiaron distintos modelos y se hizo una selección de los más apropiados para desarrollar en el trabajo teniendo en cuenta su relevancia en el campo de estudio y su aplicabilidad al caso de la polarización política que nos ocupa.

2. **Diseño del índice propio y variables**: a partir del conocimiento teórico adquirido, se desarrolló un nuevo índice que se adapta mejor a las características específicas del problema de análisis de opinión política en Twitter. Se diseñan también variables bidimensionales y unidimensionales que reflejan información relevante de los datos y que pueden ser tomadas como entrada de los índices seleccionados.

3. **Aplicación práctica y análisis**: se desarrolló la implementación de las distintas medidas de polarización utilizando herramientas de análisis de datos como la librería *pandas* y el entorno interactivo *Jupyter Notebook*. Se construyó la base de datos de 1 millón de *tweets* reales obtenidos mediante la técnica de *web scraping* con la librería *Selenium* y se aplicaron todas las medidas a muestras de estos datos. Se compararon los resultados obtenidos y se discutieron las diferencias observadas de cada modelo.

   El código se encuentra disponible en el repositorio:

`https://github.com/LauraRodrigoCanete/Aplicacion_Modelos_Polarizacion`

4. **Síntesis e integración**: en esta fase se consolidó el trabajo teórico y empírico desarrollado previamente. Se llevó a cabo la redacción de un análisis detallado de cada medida, incluyendo sus orígenes, formulación, propiedades y

relaciones con conceptos similares. Además, se idearon ejemplos sencillos que acompañaran a cada modelo para favorecer su comprensión. Finalmente, se integraron en este documento los resultados obtenidos del caso práctico con el fin de evaluar el comportamiento de los índices y extraer conclusiones.

Aunque el trabajo se centra en el caso estadounidense y la terminología así lo indica con el uso de los republicanos y demócratas como partidos contrarios, la metodología desarrollada es generalizable a otros contextos de polarización política con dos candidatos, partidos o ideologías enfrentadas.

1.4. Estructura de la Tesis

Este trabajo se estructura en seis capítulos principales, guiando al lector a través de la investigación sobre la polarización política medida con datos de redes sociales. El Capítulo 1, «Introducción», sienta las bases exponiendo el problema de la polarización como motivación de este estudio, los objetivos y el plan de trabajo general del proyecto, además de presentar la organización estructural. El Capítulo 2, «Estado de la Cuestión», proporciona una revisión exhaustiva de la investigación existente, centrándose en la variedad de métodos existentes y desafíos para medir el complejo fenómeno de la polarización en diversos ámbitos (económico, social, religioso...). Pasando a los aspectos prácticos, el Capítulo 3, «Datos Iniciales», se presentan fórmulas para las variables bidimensionales y unidimensionales que capturan información relevante de datos de redes sociales y constituyen los valores de entrada para los modelos de polarización. El Capítulo 4, «Modelos de Polarización» explica el contenido principal, detallando para cada uno de los cuatro índices de medición de polarización política contenidos en el trabajo, incluida la nueva propuesta, su definición, propiedades, consideraciones y ejemplo de aplicación. El Capítulo 5, «Aplicación Práctica de los Modelos», presenta el caso de estudio con datos de Twitter de las elecciones presidenciales de EE. UU. de 2016, 2020 y 2024, incluyendo información sobre los datos evaluados y los resultados de la comparación de los distintos índices. Finalmente, el Capítulo 6, «Conclusiones y Trabajo Futuro», resume las principales conclusiones extraídas de esta investigación y propone diversas vías para futuras investigaciones.

Capítulo 2

Estado de la Cuestión

El fenómeno de la polarización es un concepto presente en campos tan variados como la economía, la ética y la política que ha sido estudiado cada vez más en profundidad desde las últimas décadas, dando lugar a medidas cuantitativas para captarlo. Diversos autores han formalizado este concepto desde la perspectiva matemática, introduciendo índices para medir la intensidad de la polarización en una sociedad. A continuación se presenta un repertorio de las principales contribuciones, abarcando desde medidas tradicionales analíticas de polarización hasta enfoques recientes basados en grafos, conjuntos borrosos y modelos probabilísticos.

2.1. Medidas Analíticas

En cuanto a medidas tradicionales de polarización, nacidas en el ámbito económico, predominan, sin duda, los dos modelos que se estudian con mayor detalle en el presente trabajo: los provenientes de los autores Esteban y Ray junto con Foster y Wolfson. Esteban y Ray (1994) vincularon la polarización con dos nociones clave: identificación (la afinidad de un individuo con quienes son semejantes a él) y alienación (el distanciamiento o antagonismo hacia quienes son distintos). Su índice teórico de polarización parte de una distribución de ingresos dividida en grupos definidos a priori, suponiendo que cada persona siente identificación con los miembros de su mismo grupo y alienación respecto a los de otros grupos. Cabe destacar que la formulación original de Esteban y Ray asumía un número fijo de grupos, pero trabajos posteriores como Esteban et al. (2004) extendieron el índice al caso continuo, que no requiere fijar grupos de antemano. Casi en paralelo, Foster y Wolfson (1992) propusieron una medida de bipolarización enfocada en la desaparición de la clase media. En su trabajo, estos autores definen un índice que considera la distribución de ingresos dividida en dos subconjuntos (por debajo y por encima de la mediana) y cuantifica la polarización en términos de la brecha entre ambas mitades. El índice de Wolfson captura el vaciamiento de la clase media: una mayor concentración de masa en los dos polos extremos conlleva un aumento de la polarización bipartita.

En contextos de conflictividad étnica y política, Montalvo y Reynal-Querol (2005) introdujeron un índice que se define a partir de la distribución de la población entre distintos grupos (por ejemplo, grupos religiosos) y alcanza su máximo cuando la so-

ciedad se divide en dos grupos de igual tamaño. Este índice, similar a una medida de diversidad poblacional, considera únicamente los pesos poblacionales de cada grupo, ignorando diferencias internas como el estatus económico medio de cada grupo. En esencia, trata la polarización como un fenómeno estrictamente composicional. Se le ha criticado por omitir la dimensión socioeconómica, por ejemplo, no distingue si dos grupos étnicos igualmente numerosos ocupan posiciones económicas muy desiguales, lo que podría acentuar la polarización.

Las limitaciones de las medidas previas motivaron extensiones. Apouey (2007) propuso un índice orientado a datos ordinales, aplicándolo a distribuciones de salud autopercibida. Su medida de polarización de la salud es una extensión de los índices tradicionales de ingreso, adaptada para variables ordinales (como niveles de salud) en lugar de cardinales. Apouey retoma la idea de identificar grupos de individuos con niveles similares asumiendo que quienes comparten rango de salud conforman un grupo cohesionado. Sin embargo, la medida de Apouey tampoco incorpora características de la identidad como raza o género, es decir, considera únicamente la distribución ordinal (por ejemplo, cuántos individuos están en cada nivel de salud), pero no si los grupos demográficos se concentran en ciertos niveles. Esto preocupa porque distintas fuentes de polarización pueden superponerse, y las medidas deben evolucionar para capturar esa complejidad.

Una contribución en esa dirección es la de Permanyer y D' Ambrosio (2015), quienes desarrollaron un índice de polarización social multidimensional. Ellos argumentan que ni Reynal-Querol ni Apouey por sí solos reflejan adecuadamente la realidad de muchas sociedades, donde existen grupos con identidades bien definidas y diferencias marcadas en variables socioeconómicas. Por tanto, proponen combinar ambas perspectivas: su índice primero particiona la sociedad en grupos según una característica social (por ejemplo, grupos étnicos) y luego mide qué tan concentrados están dichos grupos en regiones específicas de la distribución de otra variable (por ejemplo, ingreso o salud). De este modo, logran captar situaciones en las que, por ejemplo, un grupo étnico se encuentra mayoritariamente en el estrato económico bajo mientras otro grupo domina el estrato alto –escenario intuitivamente más polarizado-– frente a situaciones donde los grupos étnicos están distribuidos de forma semejante en todos los niveles socioeconómicos. Este enfoque integrador, basado en los principios de identificación y alienación de Esteban y Ray pero aplicado a múltiples dimensiones, permite una visión más rica de la polarización social.

Por otra parte, en el campo de la ciencia política surgió una medida distinta enfocada en la polarización ideológica: el índice de Dalton. Dalton (2008) propuso cuantificar la polarización de un sistema multipartidista considerando las posiciones ideológicas de cada partido en el eje izquierda-derecha y ponderándolas según su fuerza electoral. Así, este índice captura la dispersión ideológica del espectro político: vale 0 si todos los partidos se ubican en la misma posición (ausencia total de polarización) y aumenta conforme los partidos se distribuyen más hacia los extremos opuestos del eje ideológico, especialmente si tienen tamaños electorales comparables. En términos matemáticos, equivale esencialmente a una varianza ponderada de las posiciones ideológicas, normalizada para oscilar entre 0 y un máximo cuando se configuran dos bloques en polos opuestos.

En general, para medir la polarización, en particular la política, de forma unidi-

mensional es común encontrar índices que analizan la distribución de preferencias en una escala (por ejemplo, izquierda–derecha) y analizan las varianzas de las posiciones políticas, la diferencia de medias ideológicas entre los votantes de cada partido o calculan cuánto se aleja la distribución de una forma unimodal y centrada. Por ejemplo, la prueba estadística del *dip test* de Hartigan se basa en medir la distancia máxima entre la función de distribución de los datos y la función de distribución unimodal que mejor se ajusta a los datos. Otros índices, como el coeficiente de acuerdo de Van der Eijk (2001) miden el acuerdo comparando la dispersión real de las respuestas respecto a la moda con la dispersión máxima posible. Se basa en la suma de las distancias absolutas individuales a la moda, normalizada para que el coeficiente varíe entre 0 (desacuerdo total) y 1 (acuerdo perfecto).

Aún más reciente es el índice de polarización que proponen Morales et al. (2015), el cual estudiamos con profundidad en este trabajo por su aparente sencillez y por su efectividad probada en reflejar la división existente entre dos bandos políticos. El índice cuantifica cuánto se acerca una distribución al caso de dos picos extremos opuestos de igual tamaño. Su medida está inspirada en el momento dipolar eléctrico en física: análogamente a cómo el momento dipolar aumenta al separar las cargas opuestas de un dipolo, la polarización política aumenta cuanto más distantes estén entre sí dos grupos de opinión de tamaño comparable.

2.2. Medidas de Grafos

Dejando atrás las medidas tradicionales, con el auge de las redes sociales, la polarización política ha empezado a medirse también a través de la estructura de grafos que representan interacciones (amistad, seguimiento, *retweets*, etc.). En estos contextos, la polarización suele manifestarse como una división modular del grafo en dos grandes comunidades con pocos lazos entre sí, a menudo denominadas cámaras de eco[1]. Una herramienta ampliamente utilizada para detectar y cuantificar esta división es la modularidad de la red. La modularidad en Newman (2006) es una medida de cuán bien dividido está un grafo en comunidades: valores altos de modularidad indican que la red puede separarse en grupos con muchas conexiones internas y pocas conexiones entre grupos. Diversos estudios han señalado que la modularidad proporciona una medida conceptualmente clara de la polarización, al revelar tanto el número de grupos relevantes como la intensidad de las divisiones entre ellos. Por ejemplo, aplicada al análisis del Congreso de Estados Unidos en Waugh et al. (2009), la modularidad de la red de votaciones ha servido para indicar la magnitud de la polarización partidista a lo largo del tiempo.

No obstante, es importante matizar que una alta modularidad no siempre equivale a polarización antagonista. Guerra et al. (2013) observaron que la modularidad por sí sola es una métrica incompleta: es posible que incluso en ausencia de fuerte antagonismo ideológico, una red social se divida en comunidades relativamente densas (por ejemplo por intereses comunes, idioma...), produciendo alta modularidad

[1]El fenómeno de la cámara de eco refleja una situación en medios de comunicación y redes sociales en la que los usuarios son presentados con ideas e información que en su mayor parte amplifican y refuerzan sus propias creencias en lugar de exponer ideas contrarias o fuentes diversas.

sin que ello implique polarización política. En consecuencia, los autores proponen analizar no solo la partición en comunidades, sino también las interacciones en la frontera entre comunidades. Por ejemplo, la concentración de nodos altamente conectados en la frontera entre dos comunidades opuestas: en redes altamente polarizadas, los nodos más populares tienden a estar dentro de las comunidades y no en la frontera, reflejando que el debate público ocurre principalmente dentro de cada burbuja ideológica.

Entre las métricas de grafos más recientes destaca una basada en *random walks* o caminos aleatorios. Garimella et al. (2018) propusieron un índice calculado mediante caminos aleatorios que exploran un grafo social. La idea intuitiva es que si una red está fuertemente polarizada en dos comunidades, un camino aleatorio que comience en la comunidad A raramente cruzará a la comunidad B, y viceversa, dado que hay pocos enlaces entre ellas. El índice cuantifica esta dificultad de los caminos para salir de su comunidad de origen. Esta métrica expone el valor de incorporar procesos estocásticos en grafos para medir polarización: en lugar de mirar solo la estructura estática, se evalúa cómo se comporta la difusión simulada de información en la red, lo cual refleja de forma más dinámica las barreras entre comunidades.

2.3. Medidas Difusas y Estocásticas

Otras técnicas de medición modernas se ayudan de materias tan interesantes como los conjuntos borrosos para explicar el fenómeno de la polarización. Las técnicas de conjuntos borrosos o *fuzzy sets* se han introducido para relajar la suposición tradicional de que cada individuo pertenece completamente a un solo grupo. El enfoque borroso permite asignar a cada individuo grados de pertenencia a múltiples grupos, reflejando esta ambigüedad. Bajo esta perspectiva, Guevara et al. (2020) proponen medir la polarización incorporando funciones de pertenencia difusa. Definen un índice donde cada individuo contribuye a la polarización en función de su membresía parcial a los polos en conflicto. Usando operadores de agregación apropiados, evalúan cuánto se superponen o se separan las distribuciones de pertenencia de distintos grupos. Si la mayoría de los individuos tienen afiliaciones mixtas la polarización medida será baja pero si las funciones de pertenencia muestran individuos con alta pertenencia a un grupo y baja al otro, se obtendrá un valor alto de polarización difusa.

Por último, mencionamos la contribución a este campo de distintos modelos probabilísticos. En lugar de enfocarse únicamente en medidas estáticas, varios trabajos recientes emplean modelos dinámicos y estocásticos para entender la polarización como el ya mencionado sobre caminos aleatorios. Otro enfoque representativo es el uso de cadenas de Markov[2] para modelar transiciones de individuos entre estados de opinión. Por ejemplo, en un modelo de opinión pública podemos definir varios estados ideológicos (desde izquierda hasta derecha) y probabilidades de que

[2]Una cadena de Markov es un proceso estocástico en el que el estado futuro depende únicamente del estado actual y no de los anteriores. Es decir, tiene memoria limitada al presente. El proceso avanza de un estado a otro de forma probabilística, siguiendo unas probabilidades de transición que solo dependen del estado actual.

un individuo cambie de un estado a otro mediante interacciones sociales. Böttcher y Gersbach (2020) desarrollaron un marco matemático de evolución política donde los individuos difunden sus ideas a vecinos ideológicamente cercanos, formulando un modelo de cambio político que matemáticamente se describe como una cadena de Markov en el espacio de opiniones. Este tipo de modelo permite simular cómo pequeñas variaciones en la tasa de difusión o en la influencia de actores influyentes pueden llevar a diferentes niveles de polarización a largo plazo, reproduciendo tendencias observadas empíricamente. El resultado es un marco probabilístico donde la polarización puede interpretarse, a partir de la distribución estacionaria de la cadena de Markov. Una distribución estacionaria bimodal equivaldría a alta polarización, mientras que una unimodal centrada indicaría baja polarización.

Dentro de esta categoría se inscribe también el modelo de DeGroot (1974) y sus variantes, utilizados para estudiar la formación de consenso o disenso en grupos. El modelo describe un proceso iterativo en el cual cada individuo actualiza su opinión haciendo un promedio ponderado de la opinión propia con las de sus vecinos. Si se ejecuta indefinidamente, bajo ciertas condiciones todos convergerían a un consenso; pero modificaciones al modelo permiten que se alcancen equilibrios polarizados. Por ejemplo, incorporando individuos inamovibles o grupos desconectados, el modelo puede conducir a múltiples opiniones estables en la población en lugar de una única.

En conclusión, el estudio sobre medidas de polarización muestra un campo creciente y multidisciplinar. Resumiendo lo expuesto en esta sección, todo comienza con las medidas tradicionales, que sentaron las bases axiomatizando la polarización principalmente en términos de distribución socioeconómica, con aportes emblemáticos de Esteban-Ray, Wolfson, y otros. Sobre esa base se han construido nuevos índices que incorporan distintas dimensiones (étnicas, ordinales, ideológicas), como las propuestas de Reynal-Querol, Apouey, Dalton, Permanyer-DÁmbrosio, entre otros, refinando la capacidad descriptiva de los índices. Paralelamente, la irrupción de las redes sociales llevó el análisis de la polarización al terreno de los grafos: ahora se mide cuán separada está la red en comunidades (modularidad) o cuán confinada queda la información en un subgrupo (métricas de caminos aleatorios), revelando la polarización en la esfera comunicacional. Asimismo, se han abierto nuevas fronteras metodológicas mediante la lógica difusa y la modelización estocástica, permitiendo captar grados de pertenencia y dinámicas temporales que los índices estáticos no podían reflejar.

Capítulo 3

Datos Iniciales

Partimos de un conjunto de datos que pueden expresarse en variables bidimensionales y unidimensionales. En algunos casos, será necesario transformar unas en otras en función de las necesidades del análisis. En general, aunque no exclusivamente, los modelos que discutiremos trabajan con variables unidimensionales. Los datos sin tratar de los que disponemos en nuestro caso particular son *tweets* con un sentimiento asociado, que se procesan para derivar las diferentes variables, como discutiremos en el capítulo 5 de la aplicación práctica.

3.1. Variables Bidimensionales

Denotamos por U al conjunto de todos los usuarios de nuestra muestra. Para cada candidato C, definimos el subconjunto $U_C \subseteq U$ como el conjunto de usuarios que han expresado una opinión sobre dicho candidato. Cada usuario $u \in U_C$ tiene asociado un valor $u_C \in [-1, 1]$, calculado como la media de las valoraciones expresadas en sus mensajes referidos a C.[1]

Así, si un usuario u ha publicado m_T mensajes sobre el candidato Trump, con valoraciones $t_i \in \{-1, 0, +1\}$ para cada $i \in \{1, ..., m_T\}$, su valoración agregada será:

$$u_T = \frac{1}{m_T} \sum_{i=1}^{m_T} t_i$$

De forma análoga, si ha publicado m_H mensajes sobre Harris con valoraciones $h_i \in \{-1, 0, +1\}$ para $i \in \{1, ..., m_H\}$, se define:

$$u_H = \frac{1}{m_H} \sum_{i=1}^{m_H} h_i$$

Esto nos permite representar a cada usuario u mediante una variable bidimensional $(u_R, u_D) \in [-1, 1] \times [-1, 1]$, que resume su actitud hacia los candidatos R y

[1] Tanto la media como la moda son opciones válidas; en nuestro caso empleamos la media, aunque la elección depende de la observación empírica. La moda puede ser más robusta ante ruido o valores atípicos.

D, republicano y demócrata respectivamente.

Esta representación presenta dos inconvenientes principales:

1. No todos los usuarios opinan sobre ambos candidatos. Por ello, los conjuntos U_R y U_D pueden diferir en tamaño, y muchos usuarios solo disponen de una de las dos coordenadas. En este contexto, el conjunto de usuarios con opiniones sobre ambos candidatos se denota como $U_{R \cap D} = U_R \cap U_D$.

   Se plantean distintas soluciones:

   - Admitir la ausencia de las opiniones redefiniendo el dominio de la variable bidimensional a $([-1, 1] \cup \{\text{nulo}\}) \times ([-1, 1] \cup \{\text{nulo}\})$, e ignorando los valores nulos en los modelos donde no se puedan utilizar.
   - Imputar los valores faltantes mediante aproximaciones basadas en usuarios similares. Por ejemplo, si un usuario tiene valoración positiva hacia un candidato pero no ha opinado sobre el otro, se puede asignar como valor una media (o moda, si se muestra más precisa) de las opiniones expresadas por usuarios con opiniones similares acerca del partido del que conocemos el sentimiento del usuario.[2]

2. La representación (u_R, u_D) no tiene en cuenta el número de *tweets* que un usuario escribe para opinar sobre cada candidato. Por ejemplo, un usuario que publica un solo *tweet* negativo sobre un candidato y otro que publica mil *tweets* negativos tendrían la misma representación, a pesar de que intuitivamente sabemos que su intensidad de odio podría ser muy distinta.

   Una solución ingenua sería ponderar el valor de cada componente en función del número máximo de mensajes emitidos por todos los usuarios respecto a ese candidato. No obstante, esto podría introducir sesgos por hiperactividad con usuarios que generen una cantidad desproporcionada de *tweets*, además de dificultar la interpretación, ya que al combinar intensidad y volumen en una sola medida se perdería claridad sobre si un valor extremo refleja un sentimiento muy fuerte o simplemente una gran cantidad de *tweets* moderados.

   Otra opción sería incorporar una medida de intensidad o volumen junto a la media, lo que nos lleva a la siguiente sección, en la que transformamos estas variables en medidas unidimensionales que tienen en cuenta para cada usuario el número de *tweets* dedicados comparativamente a cada candidato[3].

3.2. Variables Unidimensionales

A continuación, definimos distintas transformaciones de la variable bidimensional en variables unidimensionales que capturan distintos aspectos del comportamiento

[2]La similitud entre usuarios no requiere igualdad exacta de opiniones. Es recomendable agrupar los usuarios en intervalos de sentimiento (e.g., positivo, neutro, negativo) y calcular la media o moda del grupo correspondiente.

[3]Conviene enfatizar que un mismo *tweet* podría hablar de ambos candidatos y en ese caso contabilizaría para los dos.

político de los usuarios. Estas transformaciones permiten entender mejor los datos y adaptarlos a modelos que requieren una sola dimensión por individuo, facilitando comparaciones agregadas y el uso de ciertos índices de polarización.

3.2.1. Intención de voto (demócrata vs republicano)

Esta variable estima la preferencia política entre dos polos ideológicos opuestos, considerando que una opinión favorable hacia un candidato equivale a una desfavorable hacia su antagonista. Es decir, un *tweet* de apoyo al partido republicano se interpreta como rechazo al partido demócrata, y viceversa.

Se define como:

$$vote\_intention = \frac{u_R \cdot m_R - u_D \cdot m_D}{m_R + m_D}$$

donde $u_R, u_D \in [-1, 1]$ son las valoraciones medias hacia los republicanos y los demócratas respectivamente, y m_R, m_D el número de *tweets* que el usuario ha emitido sobre cada uno. En caso de no existir valoraciones para algún usuario sobre algún candidato se toma la valoración media y el número de *tweets* como 0.

Esta variable toma valores en $[-1, 1]$, donde -1 representa una fuerte preferencia demócrata, y 1 una fuerte preferencia republicana. Para su uso en modelos que requieren valores positivos –como el de Foster y Wolfson, en el que una media o mediana igual a cero haría que el índice no estuviera definido[4]–, usaríamos una versión transformada al intervalo $[0, 1]$:

$$vote\_intention_{[0,1]} = \frac{vote\_intention + 1}{2}$$

3.2.2. Tonalidad del discurso (optimista vs pesimista)

Esta variable mide si el usuario tiende a emitir mensajes cargados afectivamente de forma positiva o negativa, independientemente de hacia quién estén dirigidos. Refleja si el individuo actúa como un simpatizante general (*fan*) o como un opositor sistemático (*hater*).

Se define como:

$$tone\_bias = \frac{u_R \cdot m_R + u_D \cdot m_D}{m_R + m_D}$$

Esta variable también varía en $[-1, 1]$, donde -1 indica una actitud predominantemente negativa, y 1 una actitud mayoritariamente positiva.

3.2.3. Amplitud (neutral vs polarizado)

La amplitud representa la distancia entre las valoraciones de un usuario hacia los dos candidatos, es decir, el grado de polarización interna. Un valor bajo indica

[4]Si después de hacer esta traslación aún así la mediana o la media son 0 se considera un caso degenerado y el índice dará 0 por considerarse que casi toda la población se encuentra en un polo y no existe polarización entre polos.

opiniones similares (ya sean ambas positivas, negativas o neutras), mientras que un valor alto indica una fuerte preferencia por uno de los candidatos en detrimento del otro.

Se define como:

$$spread = \frac{|u_R - u_D|}{2}$$

La variable toma valores en el intervalo $[0, 1]$. Un valor de 0 representa indiferencia o actitud similar hacia ambos candidatos, mientras que 1 indica la máxima polarización posible.

Capítulo 4

Modelos de Polarización

En esta sección se presentan distintos modelos de polarización aplicables a nuestro conjunto de datos. Nos centraremos en índices numéricos, ya que constituyen las herramientas más habituales para el análisis de distribuciones, especialmente en contextos como el de la desigualdad económica o la fragmentación política. Un índice numérico resume el fenómeno de interés en un solo valor, lo que permite inducir un orden completo sobre distintas distribuciones y facilita la comparación entre contextos.

Analizaremos en primer lugar el Índice de Polarización Gravitatoria, seguido del famoso índice propuesto por Foster y Wolfson, el modelo clásico de Esteban y Ray, y finalizaremos con una propuesta propia que pretende capturar de forma simple y efectiva la intensidad bipolar de una sociedad ideológicamente dividida como la americana.

Es importante notar que cada uno de estos modelos puede tomar como entrada cualquiera de las variables definidas en el capítulo anterior, se elegirá una u otra en función de sobre qué aspecto político se desea estudiar la polarización y, a efectos prácticos, modificarla es tan sencillo como cambiar la distribución de entrada del modelo. Es claro que, en general, el aspecto más interesante suele ser la orientación política de los usuarios –en nuestro texto, esta intención de voto es reflejada por la variable *vote_intention*– y al ser la más intuitiva se ha elegido para ejemplificar los tres últimos índices que presentamos. No obstante, la variable bidimensional original también podría ser aplicada a los modelos si cada par (u_R, u_D) se descompone en dos variables unidimensionales. También puede resultar interesante esta variable para analizar, en lugar de la polarización en el espectro republicano-demócrata, la polarización sobre un único candidato en el espectro amor-odio. Por ello, hemos decidido elegirla para ejemplificar el primer índice presentado y así mostrar algo más de variedad.

4.1. Índice de Polarización Gravitatoria

4.1.1. Definición

El Índice de Polarización Gravitatoria (IPG) definido por Morales et al. (2015) considera dos poblaciones que se diferencian entre sí por el signo de su opinión acerca de un candidato político seleccionado: decimos que una población le ama (signo > 0) y la otra le odia (signo < 0). Está inspirado en la física, en particular en el concepto de centro de masa y en el momento dipolar eléctrico (que es proporcional a la separación entre cargas).

En particular, el modelo mide la polarización considerando tanto la diferencia en el tamaño de las poblaciones como la distancia entre sus respectivos centros de gravedad.

Para este modelo, utilizamos la variable bidimensional de sentimiento definida previamente en la sección 3.1. Recordamos que denotamos por U_C al conjunto de usuarios que han expresado una opinión sobre el candidato C y que $u \in U_C$ tiene asignado un valor $u_C \in [-1, 1]$, la opinión promedio del usuario sobre el candidato. En resumen, para cada usuario tenemos la variable (u_A, u_B) donde A y B son candidatos políticos. Dado que este modelo evalúa cada candidato de manera independiente, nos centramos en U_C para un candidato C concreto.

Para comprender la expresión del IPG repasaremos algunos conceptos previos. Definimos la función de probabilidad de encontrar un usuario con un sentimiento promedio x respecto al candidato C entre todos los usuarios con una opinión sobre C como:

$$p(x) = \frac{\#\{u_C = x\}}{\#\{U_C\}}$$

Dado que $u_C \in [-1, 1]$ definimos la función de distribución acumulada (CDF) de la variable X como:

$$F(x) = P(X \leq x) = \int_{-1}^{x} p(t)\, dt.$$

A partir de esta función F, podemos ver las probabilidades acumuladas y expresar las proporciones relativas de usuarios con opiniones negativas y positivas de la siguiente manera:

$$A^- = P(X < 0) = F(0) = \int_{-1}^{0} p(x)\, dx,$$

$$A^+ = P(X > 0) = 1 - F(0) = \int_{0}^{1} p(x)\, dx$$

donde A^- representa la proporción relativa de usuarios con opiniones negativas y A^+ la de usuarios con opiniones positivas. Por ejemplo, si la mitad de los usuarios tiene un valor de opinión $u_C < 0$ y la otra mitad un valor $u_C > 0$, entonces $A^- = A^+ = \frac{1}{2}$. Cabe destacar que estas cantidades solo reflejan la proporción de usuarios en cada grupo y no la intensidad de sus opiniones.

Nótese que no se están teniendo en cuenta a los usuarios con opinión media igual a cero. Esto en el caso continuo no es un problema ya que por ser un punto en el continuo la probabilidad es cero. Sin embargo, en el caso discreto habría que establecer en qué grupo se incluyen estos usuarios. Para evitar sesgos en la distribución de los grupos, proponemos distribuir aleatoriamente los usuarios neutros entre los conjuntos A^+ y A^- de manera proporcional a sus tamaños originales. Por tanto, formalmente, los valores de A^+ y A^- se definirían como:

$$A^+ = P(X > 0) + P(X = 0)\frac{P(X > 0)}{P(X \neq 0)} = P(X > 0)\left[1 + \frac{P(X = 0)}{P(X \neq 0)}\right]$$

$$A^- = P(X < 0)\left[1 + \frac{P(X = 0)}{P(X \neq 0)}\right]$$

Esta asignación garantiza que los conjuntos A^+ y A^- permanezcan disjuntos, preservando la proporción relativa de los grupos y asegurando que la modificación de los centros de gravedad sea lo más equitativa posible. Ahora, una vez aclarada la cuestión de las opiniones neutras, para facilitar la comprensión seguiremos haciendo uso de la formulación inicial de los conjuntos A^+ y A^-.

A partir de estas definiciones, podemos expresar la diferencia normalizada en el tamaño de ambas poblaciones como:

$$\Delta A = A^+ - A^- = P(X > 0) - P(X < 0)$$

Este valor indica qué grupo tiene mayor representación en términos de sentimiento sobre el candidato C.

Sin embargo, para medir la polarización, no solo importa el tamaño relativo de las poblaciones, sino también la intensidad de sus opiniones. Para ello, calculamos los centros de gravedad[1] de la distribución de probabilidad $p(x)$ en los intervalos positivo y negativo. Estos se definen como:

$$gc^- = \frac{\int_{-1}^{0} p(x)x\,dx}{A^-}$$

$$gc^+ = \frac{\int_{0}^{1} p(x)x\,dx}{A^+}$$

La distancia normalizada entre ambos centros de gravedad se define como:

$$d = \frac{|gc^+ - gc^-|}{2}$$

[1]Desde el punto de vista físico, el cálculo realizado en este modelo se corresponde con el término *centro de masa*, el punto en el que se puede considerar que toda la masa de un sistema está concentrada para describir su movimiento en ausencia de fuerzas externas. En contraste, el *centro de gravedad* se define como el punto donde se aplica la resultante de las fuerzas gravitatorias. No obstante, ambos coinciden en campos gravitacionales uniformes, por lo que empleamos este abuso de notación para mantener la coherencia con las fuentes originales.

donde el denominador 2 corresponde al tamaño total del intervalo $[-1, 1]$ en el que pueden situarse los valores de u_C.

Bajo estas definiciones, podemos interpretar que en un caso de polarización extrema (es decir, cuando los sentimientos son muy negativos o muy positivos), los centros de gravedad estarán cerca de ± 1, lo que implica que la distancia definida será $d \approx 1$. En cambio, cuando los sentimientos son más moderados, los valores promedio y los centros de gravedad estarán más cercanos a 0, resultando en una menor distancia d.

A partir de estas definiciones, la expresión final del Índice de Polarización Gravitatoria es:

$$\mu = (1 - |\Delta A|)d$$

Este índice captura tanto la diferencia en tamaño entre los grupos de opinión como la intensidad de sus posiciones, proporcionando una medida cuantitativa de la polarización respecto al candidato C.

4.1.2. Propiedades

El Índice de Polarización Gravitatoria, μ, es una función de dos variables: la diferencia en tamaño entre los grupos de opinión, $|\Delta A|$, y la distancia entre sus centros de gravedad, d. Su comportamiento se muestra en la figura 4.1. En esa gráfica podemos observar que a medida que la diferencia de tamaño aumenta ($|\Delta A| \to 1$), el índice de polarización disminuye, ya que la polarización es mayor cuando los grupos tienen tamaños similares ($|\Delta A| \to 0$). Por otro lado, cuando la distancia entre los centros de gravedad es mayor ($d \to 1$), el índice de polarización aumenta.

Este comportamiento se puede describir a partir de las siguientes propiedades:

- Rango de valores: El índice μ está acotado en el intervalo $[0, 1]$, donde:
  - $\mu = 1$ indica una polarización extrema, es decir, cuando los grupos de opinión tienen tamaños similares ($|\Delta A| \approx 0$) y están completamente separados ($d = 1$).
  - $\mu = 0$ representa una polarización mínima, lo que puede ocurrir en dos casos:
    - Cuando uno de los grupos es significativamente mayor que el otro ($|\Delta A| \approx 1$). Esto refleja que una polarización baja no tiene que significar que la población tiene una opinión neutra. Podría suceder que la opinión media estuviera completamente centrada en un sentimiento extremo, por lo que la proporción de usuarios en el otro sentimiento sería mínima y $|\Delta A| \approx 1$.
    - Cuando las opiniones de ambos grupos son moderadas y se concentran en valores cercanos a 0 ($d \approx 0$).
- Simetría: La métrica es simétrica respecto a la distribución de opiniones, ya que solo depende del tamaño relativo de los grupos y de la distancia entre sus

centros de gravedad, pero no del signo de los sentimientos. Es decir, intercambiar los grupos de opinión negativa y positiva no afecta el valor de μ.

- Invariancia frente a traslaciones en el intervalo: El índice μ solo depende de la distancia relativa entre los centros de gravedad de las opiniones positivas y negativas y del tamaño relativo de los grupos. Si dos distribuciones de opinión tienen la misma separación d y el mismo $|\Delta A|$, entonces μ será el mismo, independientemente de en qué parte del intervalo $[-1, 1]$ se ubiquen los grupos.

  Por ejemplo, si en tres distribuciones con grupos de igual tamaño los sentimientos negativos están concentrados en torno a $-0{,}75$ y los positivos en torno a 0,25, en otra en torno a $-0{,}5$ y 0,5, y en otra en torno a $-0{,}25$ y 0,75, el valor de μ será el mismo.

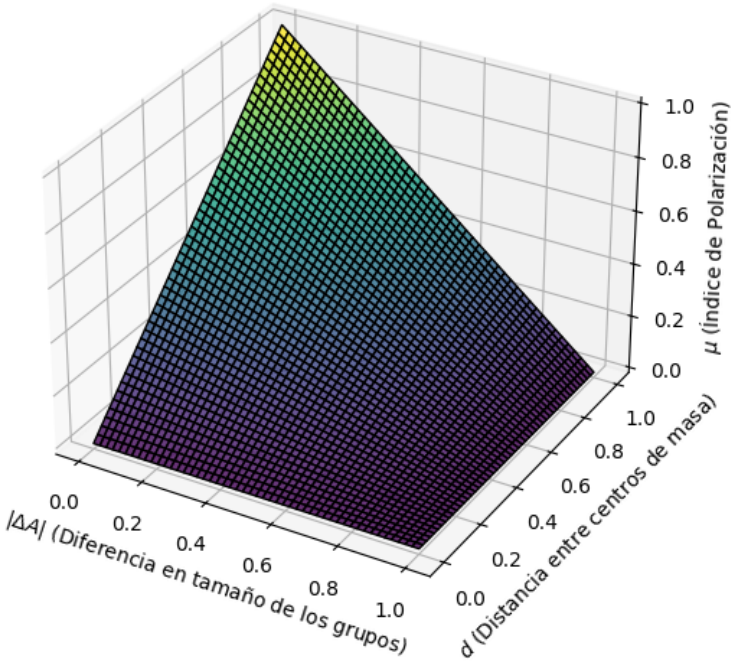

Figura 4.1: Gráfica que muestra el comportamiento del IPG μ en función de la diferencia en el tamaño de los grupos $|\Delta A|$ y de la distancia entre sus centros de gravedad d. Fuente: elaboración propia.

4.1.3. Consideraciones

El Índice de Polarización Gravitatoria permite analizar la opinión de los usuarios sobre cada candidato de manera independiente. Sin embargo, su aplicación requiere considerar algunas particularidades del conjunto de datos. Como se mencionó anteriormente en el estudio de la variable bidimensional, en nuestro caso no todos los usuarios expresan opiniones sobre ambos candidatos por lo que los conjuntos de

usuarios con sentimiento sobre cada candidato, U_A y U_B, pueden tener diferentes tamaños.

Así, solo los usuarios $U_{A\cap B}$ pueden representarse en un espacio bidimensional, lo que permite construir un diagrama de amor-odio como ilustra la figura 4.2. En ella, los ejes muestran el nivel de sentimiento hacia cada uno, desde amor $(+1)$ hasta odio (-1). Los cuadrantes reflejan combinaciones de sentimientos: amor mutuo (superior derecho), odio mutuo (inferior izquierdo), y amor por uno con odio al otro (superior izquierdo e inferior derecho). Las líneas diagonales indican igual intensidad de sentimiento hacia ambos y los ejes que representan el máximo amor hacia cada candidato están representados por corazones azules $(C = +1)$ y por corazones rojos $(T = +1)$.

Este diagrama visualiza simultáneamente la opinión de los usuarios sobre los dos candidatos y proporciona una perspectiva alternativa sobre la polarización, pudiendo también calcularse los centros de gravedad en dos dimensiones. Mientras que el índice μ evalúa la polarización respecto a cada candidato de manera separada, el diagrama de amor-odio se basa únicamente en los usuarios de $U_{A\cap B}$, lo que puede llevar a diferencias en la percepción de la polarización.

Una ventaja clave del diagrama de amor-odio que a la vez ilustra una carencia del IPG es que permite analizar comparativamente si el desacuerdo con un candidato implica necesariamente apoyo al otro. En algunos casos, es posible encontrar una gran cantidad de usuarios con sentimientos negativos hacia ambos candidatos, lo que sugiere un alto nivel de desafección política.

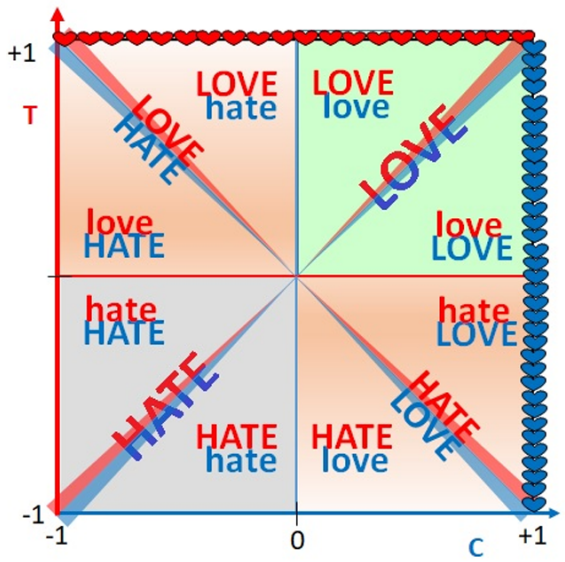

Figura 4.2: Diagrama de amor-odio entre dos candidatos (C y T). Fuente: Losada et al. (2022).

4.1.4. Ejemplo de Aplicación

En el caso particular de una distribución de opiniones con forma gaussiana centrada en cero, el índice de polarización se expresa en términos de la desviación es-

tándar. En este escenario, las opiniones políticas de los usuarios siguen una función de densidad de probabilidad dada por

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}},$$

donde σ es la desviación estándar de la distribución.

El cálculo está detallado en el Apéndice A, se obtiene:

$$\mu = \sigma\sqrt{2/\pi}.$$

Este resultado muestra que en una distribución de opiniones con forma gaussiana centrada en cero, el índice de polarización depende únicamente de la desviación estándar de la distribución. A medida que la dispersión de las opiniones aumenta, el valor de μ crece, reflejando una mayor polarización. Por el contrario, si la desviación estándar es pequeña, las opiniones estarán más concentradas en torno a cero, lo que reduce la polarización medida por el índice.

El hecho de que el valor del IPG dependa solo de la dispersión de la distribución no es exclusivo de la distribución normal, sino que se extiende a cualquier distribución simétrica centrada en cero. Esto es debido a que en este tipo de distribuciones los tamaños de los grupos positivos y negativos son iguales y los centros de gravedad están equidistantes de cero pero en lados opuestos. Por tanto, fórmulas similares podrían derivarse para otras distribuciones de este tipo, como para la distribución uniforme que también se ilustra en el Apéndice A.

4.2. Índice de Foster y Wolfson

4.2.1. Definición

Foster y Wolfson fueron de los primeros autores en proponer un índice de polarización en su artículo de 1992 que se basa en el coeficiente de Gini como indicador de desigualdad. Su medida se utiliza en economía para analizar la distribución de rentas y detectar la desaparición de la clase media. Por tanto, la variable empleada en el modelo originalmente era el ingreso de cada individuo. No obstante, para nuestros datos este modelo se aplica sobre la variable $vote\_intention_{[0,1]}$ de cada individuo i, definida en la sección 3.2.1 como v_i con $i \in \{1, \ldots, n\}$. Al igual que la acumulación de población en valores bajos y altos de ingresos implica polarización por existir gran diferencia entre ricos y pobres, en nuestra variable una gran diferencia entre valores altos y bajos de $vote\_intention_{[0,1]}$ indicará polarización política por la existencia de sentimientos radicales de republicanos y de demócratas y la disminución de sentimientos neutrales representados por valores medios.

El índice de polarización de Foster y Wolfson se define como:

$$P = 2 \cdot \left(\frac{\mu}{m}\right) \cdot (T - G)$$

Donde:

- m y μ son la mediana y la media aritmética respectivamente de los v_i de toda la población
- G es el índice de Gini
- $T = \frac{\mu^+ - \mu^-}{\mu}$, siendo μ^+ y μ^- las medias de los v_i correspondientes a los individuos situados por encima y por debajo de la mediana m, respectivamente[2] Por su definición, T recibe el nombre de desviación de la mediana relativa a la media.
- El factor 2 por el que se multiplica la fórmula simplemente es añadido por los autores para que el rango del índice sea similar al de Gini, según Rodríguez y Salas (2003). No obstante, en algunos desarrollos teóricos es omitido para simplificar los cálculos, para mantener la homogeneidad con otros artículos nosotros también lo omitiremos en las secciones teóricas de Propiedades y Consideraciones.

A continuación, se explican en detalle las componentes de la fórmula y se revisa su interpretación.

Índice de Gini: G

El coeficiente de Gini, propuesto por Corrado Gini en Gini (1912), mide la desigualdad en la distribución de una variable. Su cálculo se basa en la curva de Lorenz, que en nuestro contexto político, compara el porcentaje acumulado de la población ordenada de menor a mayor sentimiento (eje x) con el porcentaje acumulado de sentimiento (eje y).

En un escenario de distribución completamente equitativa, la curva de Lorenz coincide con la diagonal del cuadrado unitario ($y = x$), pues cada segmento de la población aporta proporcionalmente al sentimiento total. Por ejemplo, el 10 % de la población acumularía el 10 % del sentimiento acumulado total de las elecciones. Sin embargo, en una sociedad real, tal y como hemos definido numéricamente los extremos, los sectores más demócratas contribuyen con una fracción menor del *vote_intention* total, mientras que los sectores más republicanos concentran una proporción mucho mayor. Esto hace que la curva de Lorenz siempre se encuentre por debajo de la diagonal.

El coeficiente de Gini mide la desigualdad a partir del área comprendida entre la curva de Lorenz y la diagonal, normalizada respecto al área total del triángulo que forma la diagonal con los ejes (ver figura 4.3). Se expresa como:

$$G = \frac{A}{A + B}$$

donde A es el área entre la curva de Lorenz y la diagonal, y B es el área bajo la curva de Lorenz. Dado que el área total del triángulo es 0.5 (porque el cuadrado

[2]En el caso de una distribución discreta se deben repartir los usuarios con $vote\_intention = m$, a no ser que por su mínimo tamaño se decidan retirar de la muestra. Para que se cumpla la propiedad $\mu = \frac{\mu^+ + \mu^-}{2}$ habrá que asegurarse al redistribuir los neutros de que los grupos por encima y por debajo de la mediana tengan exactamente la misma proporción.

unitario tiene área 1), el coeficiente de Gini también puede interpretarse como el doble del área comprendida entre la curva de Lorenz y la diagonal:

$$G = 2A$$

El coeficiente de Gini varía entre 0 y 1 (o expresado en porcentaje, entre 0 % y 100 %). Un valor de 0 indica igualdad perfecta, es decir, la curva de Lorenz coincide con la diagonal, lo que significa que todos los individuos tienen la misma opinión de preferencia política. Un valor de 1 representa desigualdad absoluta, con una concentración mayoritaria en un extremo ideológico, y no existen posiciones intermedias o moderadas. En este caso la curva de Lorenz iría pegada al lado de abajo del cuadrado por los votantes demócratas radicales y luego pasaría a aumentar drásticamente al acercarse al lado derecho del cuadrado cuando empiecen a sumar los republicanos radicales.

Es importante destacar que el índice de Gini mide desigualdad, no polarización. El índice de Foster y Wolfson, utiliza un enfoque similar al del coeficiente de Gini, pero adaptado para cuantificar la polarización en lugar de la desigualdad.

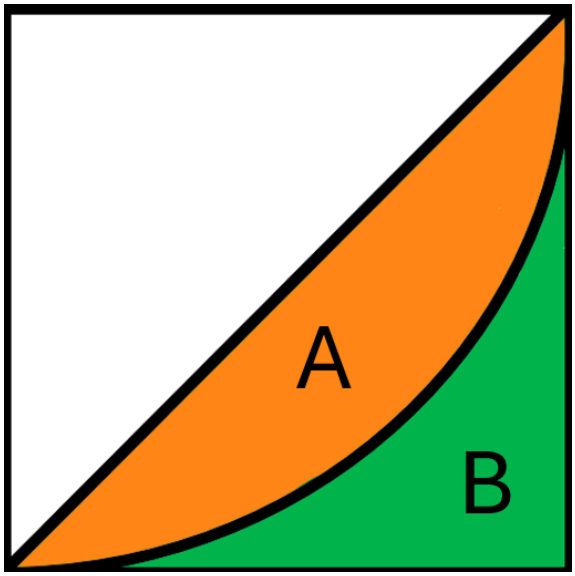

Figura 4.3: Representación del coeficiente de Gini. Si el área entre la línea de perfecta igualdad y la curva de Lorenz es A, y el área por debajo de la curva de Lorenz es B, entonces el coeficiente de Gini es $G = \frac{A}{A+B}$. Fuente: elaboración propia.

Desviación de la mediana relativa a la media: T

El término T se define como:

$$T = \frac{\mu^+ - \mu^-}{\mu}$$

donde μ^+ y μ^- son las medias del *vote_intention* de los individuos situados por encima y por debajo de la mediana, respectivamente, y μ es la media total. T mide la distancia relativa entre ambos grupos en relación con la media. Su interpretación es la siguiente:

- $T < 1$ indica menor polarización, con opiniones más homogéneas.
- $T = 1$ implica una distribución simétrica respecto a la mediana.
- $T > 1$ sugiere una mayor polarización, donde la distancia entre los grupos supera el valor medio, lo que indica una segmentación fuerte entre republicanos y demócratas.

Al igual que el coeficiente de Gini, T también tiene una interpretación geométrica basada en la curva de Lorenz. Se debe notar que la media total μ se puede expresar como:

$$\mu = \frac{\mu^+ + \mu^-}{2}$$

lo que implica que:

$$\mu^+ = 2\mu - \mu^-$$

Reescribiendo T con la expresión obtenida:

$$T = \frac{\mu^+ - \mu^-}{\mu} = \frac{(2\mu - \mu^-) - \mu^-}{\mu} = \frac{2\mu - 2\mu^-}{\mu} = 2 \cdot \frac{\mu - \mu^-}{\mu}$$

En términos geométricos, se puede interpretar que la expresión:

$$\frac{T}{2} = \frac{\mu - \mu^-}{\mu}$$

es simplemente la distancia vertical entre la curva de Lorenz y la línea de igualdad en $p = 0{,}5$, expresión que denominamos déficit de Lorenz o *Lorenz shortfall*:

$$\frac{T}{2} = 0{,}5 - L(0{,}5)$$

donde $L(0{,}5)$ representa la altura de la curva de Lorenz cuando la proporción acumulada de población es 50 %, tal y como se ilustra en la figura 4.4.

En esta figura visualizamos la interpretación geométrica de T como el doble de la distancia entre la curva de Lorenz y la línea de igualdad en $p = 0{,}5$: T es igual a dos veces el déficit de Lorenz. En la figura se expone un cuadrado unitario con la curva de Lorenz L y la recta tangente a la curva de Lorenz en $p = 0{,}5$. También se muestran las tres medias. μ coindice con $L(1)$ por ser L una CDF y la longitud del intervalo total 1. μ^- es la media de las opiniones de la población por debajo de la mediana, es decir, si todos demócratas opinaran lo mismo trazaríamos una recta desde el punto $(0, 0)$ hasta el punto $(0{,}5, L(0{,}5))$ y donde cortara esa recta con el lado $y = 1$ nos daría una altura de μ^-, aplicando el mismo razonamiento que aplicamos con μ anteriormente. Por tanto, como solo nos estamos quedando con la mitad de esa altura porque queremos medir en $y = 0{,}5$, la longitud señada en el dibujo tiene el valor de $\frac{\mu^-}{2}$. La visualización para $\frac{\mu^+}{2}$ es análoga o simplemente se puede observar que $\frac{\mu^+}{2} = \mu - \frac{\mu^-}{2}$ tal y como se verifica en el dibujo. Así y con todo, pasamos a

calcular la longitud del segmento pintado en rojo como la resta del punto extremo superior (la recta $y = x$ evaluada en $0{,}5 = \frac{\mu}{2}$) menos el punto extremo inferior ($\frac{\mu^-}{2}$):

$$\frac{\mu}{2} - \frac{\mu^-}{2} = \frac{\mu^- + \mu^+}{4} - \frac{mu^-}{2} = \frac{\mu^- - \mu^+}{4} = \frac{T}{2}$$

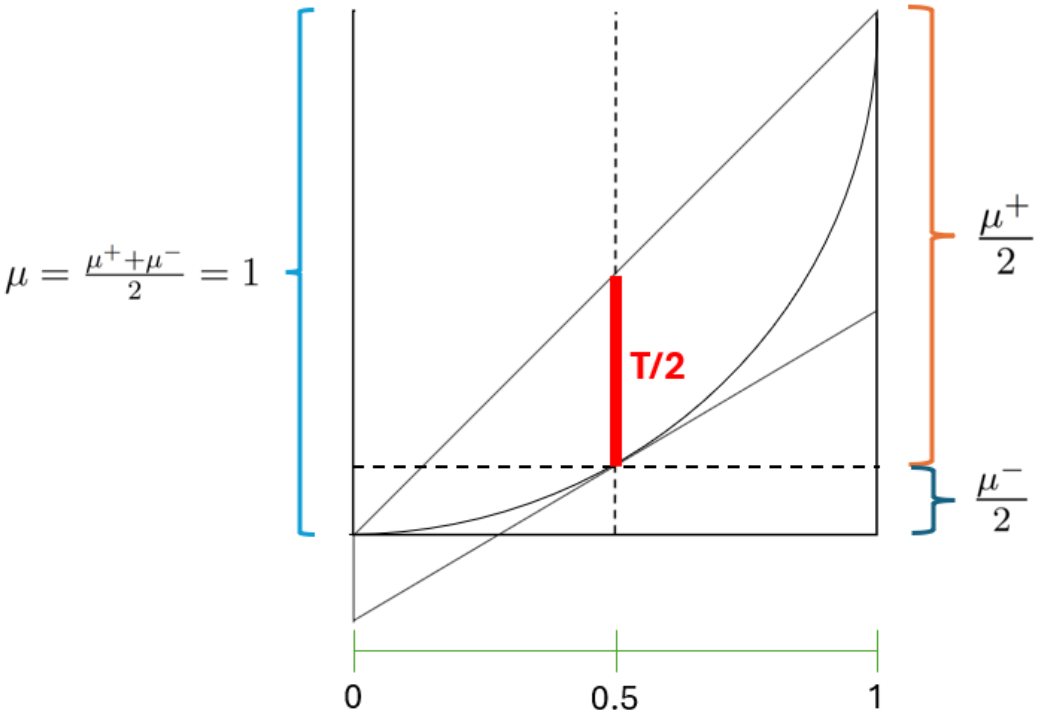

Figura 4.4: Interpretación geométrica de T como el doble de la distancia entre la curva de Lorenz y la línea de igualdad en $p = 0{,}5$. Fuente: elaboración propia.

A partir de esta definición, se obtiene que:

$$T = 1 - 2L(0{,}5)$$

Además, T es equivalente al doble del área del trapecio mostrado en la figura 4.5, ya que a continuación demostraremos que este área es equivalente a la magnitud del déficit de Lorenz.

En la figura, el trapecio está definido por la superficie comprendida entre la recta tangente a la curva de Lorenz en $p = 0{,}5$, la recta de igualdad $y = x$ y los dos ejes verticales en $p = 0$ y $p = 1$. El área está pintada en amarillo y es $0{,}5 - L(0{,}5)$. Por lo tanto, T es el doble del área de este cuadrilátero. A su vez, el índice Gini G se corresponde con el doble del área verde, que es el área entre la diagonal y la curva de Lorenz.

Ahora sí, prosigamos calculando el área entre la tangente y la línea de igualdad desde $p = 0$ a $p = 1$. Sea la ecuación de la recta tangente a la curva de Lorenz L en

$p = 0{,}5$,

$$T(p) = L(0{,}5) + L'(0{,}5)(p - 0{,}5)$$

$$\text{Área} = \int_0^1 (p - T(p))\, dp$$

$$= \int_0^1 [p - (L(0{,}5) + L'(0{,}5)(p - 0{,}5))]\, dp$$

$$= \int_0^1 [p - L(0{,}5) - L'(0{,}5)(p - 0{,}5)]\, dp$$

$$= \int_0^1 p\, dp - \int_0^1 L(0{,}5)\, dp - \int_0^1 L'(0{,}5)(p - 0{,}5)\, dp$$

$$= \frac{1}{2} - L(0{,}5)$$

Así que,

$$T = 2(0{,}5 - L(0{,}5)) = 2 \times \text{Área}$$

Por tanto hemos conseguido expresar T en función de dos características de la curva de Lorenz: el doble de su déficit y el doble del área del trapecio que la encierra.

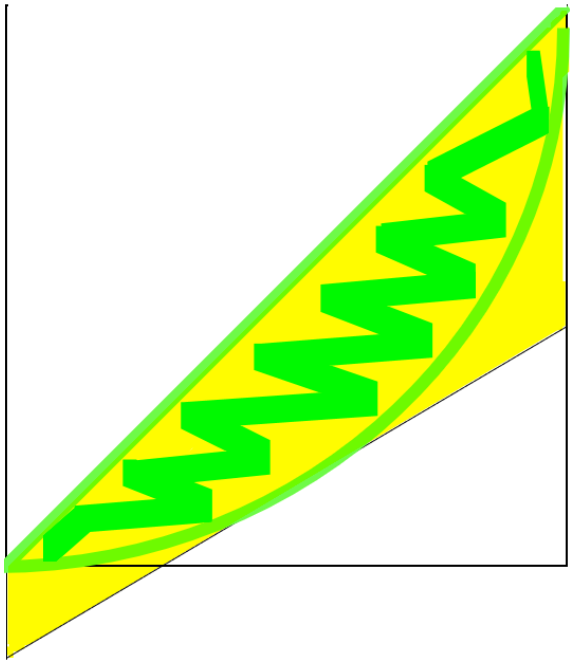

Figura 4.5: Interpretación geométrica de T como el doble del área del trapecio amarillo definido por la curva de Lorenz. Además, el índice de Gini G es el doble del área verde. Fuente: elaboración propia.

A partir de esta construcción, se observa en la figura 4.5 que T es siempre mayor que G cuando este último es distinto de cero.

El factor $\frac{\mu}{m}$

El término $\frac{\mu}{m}$ actúa como un factor de ajuste en el índice de Foster-Wolfson.

La inversa $\frac{m}{\mu}$ se interpreta como la pendiente de la tangente a la curva de Lorenz en el percentil 50 de la población. Para entender esto, recordemos que la derivada de la función de distribución acumulada (CDF) es la función de densidad de probabilidad. La pendiente de la recta tangente a la curva de Lorenz en $p = 0{,}5$ está dada por la función de densidad evaluada en ese punto, que es la mediana. Además se divide entre la media μ porque originalmente el lado del cuadrado en el que se representa la curva tiene longitud μ (ya que la media en una distribución continua es el valor de la CDF dividido entre la longitud del intervalo que es 1, por tanto en este caso la CDF evaluada al final del intervalo, en 1, tiene que tener altura μ). Por lo que dividir por μ permite escalarlo a una escala unitaria.

Al multiplicar por $\frac{\mu}{m}$, se garantiza que la pendiente de la tangente en $L(0{,}5)$ se normaliza a 1, facilitando la comparación de distribuciones con diferentes medias y medianas.

4.2.2. Propiedades

Retornando a la interpretación económica original, cuando la distribución de ingresos es simétrica, se cumple que $\mu = m$, por lo que $\frac{\mu}{m} = 1$, y el índice P se simplifica a:

$$T = G + P$$

Es decir, la desviación relativa respecto a la mediana (T) se puede descomponer como la suma de la desigualdad (G) y polarización (P). En general, la desigualdad y la polarización tienden a aumentar juntas cuando se incrementa la distancia entre los grupos situados por encima y por debajo de la mediana. En este contexto, P mide la parte de T que no se explica por la desigualdad global (G), sino por la concentración en extremos.

Esto suele ocurrir, por ejemplo, cuando se produce una transferencia regresiva a través de la mediana: una redistribución de ingresos desde personas situadas por debajo de la mediana hacia personas por encima de ella. Se denomina regresiva porque aumenta la desigualdad, va en contra de una lógica redistributiva justa donde los que tienen más ayudan a los que tienen menos. Este tipo de transferencia acentúa tanto la desigualdad como la separación entre grupos, incrementando T, G y P.

Sin embargo, desigualdad y polarización no son lo mismo, y pueden moverse en direcciones opuestas. Si T se mantiene constante, una redistribución progresiva (una redistribución de los que están por encima de la mediana a los que están por debajo) dentro de uno de los grupos (solo entre los republicanos o solo entre los demócratas) puede reducir la desigualdad sin modificar o incluso aumentando la distancia entre grupos, y por tanto aumentando la polarización.

4.2.3. Consideraciones

La fórmula del modelo de Foster y Wolfson se encuentra escrita de varias maneras, que dan lugar a interpretaciones interesantes. A continuación se expone su origen y una manera de reescribirla en función de los índices de Gini únicamente.

En el artículo publicado por los autores del índice Foster y Wolfson (2010) se presentan varias curvas de polarización.

La ecuación de la Primera Curva de Polarización se puede encontrar en cualquiera de estas formas:

$$E_F(q) = |\tilde{y}(q) - \tilde{y}(0{,}5)| = \frac{|F^{-1}(q) - F^{-1}(0{,}5)|}{m_F} = \frac{|F^{-1}(q) - m_F|}{m_F}$$

Donde:

- F es la función de distribución acumulativa. Así, si la mediana del *vote_intention* es 0,7, entonces $F(0{,}7) = 0{,}5$, lo que significa que el 50 % de la población tiene preferencia política menor o igual a 0,7.
- F^{-1} es la función cuantil de la distribución, es decir, $F^{-1}(q)$ es el *vote_intention* para el cual la proporción q de la población tiene un *vote_intention* menor o igual.
- $\tilde{y}$ representa el *vote_intention* en el percentil q normalizado por la mediana.
- m_F es la mediana de los *vote_intention*.

Esta curva mide la distancia entre el *vote_intention* del percentil q y la mediana medida en medianas (normalizado). La bipolarización ocurre cuando los *vote_intention* se agrupan en valores alejados de la mediana, lo que se refleja en valores altos de $E_F(q)$. Se muestra un ejemplo de esta curva en la figura 4.6.

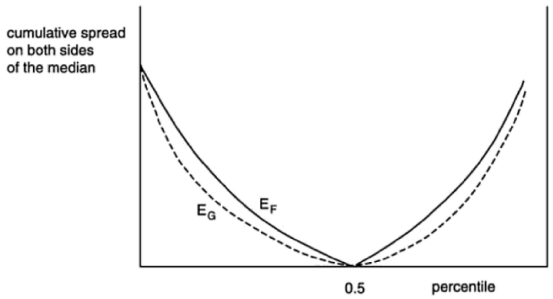

Figura 4.6: Sean F y G dos distribuciones acumulativas de probabilidad. Podemos observar que la curva E_G está debajo de E_F en todo el intervalo, así que hay más bipolarización en la distribución F que en G. Fuente: Nissanov et al. (2013).

La ecuación de la Segunda Curva de Polarización se define en función de la Primera Curva de esta forma:

$$H_F(q) = \left| \int_q^{0{,}5} E_F(p)\, dp \right| = \left| \int_q^{0{,}5} \frac{|F^{-1}(p) - F^{-1}(0{,}5)|}{F^{-1}(0{,}5)}\, dp \right|$$

Esta curva representa el área bajo la Primera Curva de Bipolarización $E_F(q)$, entre los percentiles q y 0.5, es decir, cuantas medianas se tendrían que desplazar los *vote_intention* de los usuarios entre los cuantiles q y 0,5 para alcanzar el valor de la mediana. En otras palabras:

- Si $H_F(q)$ es alto, significa que hay una fuerte acumulación de sentimientos lejos de la mediana, indicando alta bipolarización.
- Si $H_F(q)$ es bajo, implica que los sentimientos están más distribuidos de manera homogénea en torno a la mediana.

Un ejemplo de esta Segunda Curva de Bipolarización se ilustra en la figura 4.7.

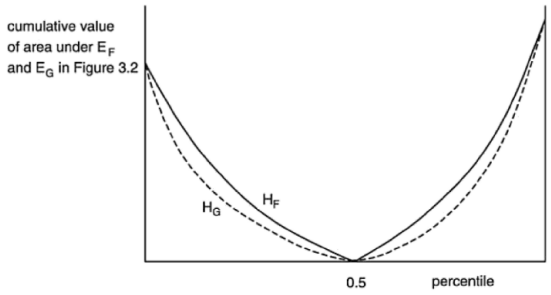

Figura 4.7: Sean F y G dos distribuciones acumulativas de probabilidad. Podemos observar que la curva H_G está debajo de H_F en todo el intervalo, así que hay más bipolarización en la distribución F que en G. Fuente: Nissanov et al. (2013).

Ahora sí, podemos dar la definición original del Índice de Polarización de Foster y Wolfson como el doble del área bajo la Segunda Curva de Bipolarización:

$$P = \int_0^1 2B_F(q)\, dq.$$

Posteriormente en el Apéndice de Foster y Wolfson (2010) se demuestra que este índice dado por curvas se podía definir como la diferencia entre dos áreas, T y G, que es la definición inicial que proporcionamos aquí.

A continuación procedemos a estudiar otra formulación del modelo en función de los índices de Gini únicamente.

Dividimos la población en dos subgrupos: aquellos por encima de la mediana m y aquellos por debajo. La desigualdad total, medida mediante el coeficiente de Gini, se descompone en dos términos:

- Desigualdad entre grupos (G_B): Se obtiene aplicando el coeficiente de Gini a una distribución lineal a trozos, en la cual todos los individuos de un grupo tienen la misma opinión, igual a la media del grupo. Es decir, aquellos por encima de m opinan μ^+ y aquellos por debajo opinan μ^-. Esto equivale a asumir que todos los miembros del grupo tienen la misma intensidad de sentimiento, lo que implica una pendiente constante. G_B mide la desigualdad entre los grupos, la diferencia en intensidad de apoyo entre los dos partidos políticos.

- Desigualdad dentro de los grupos (G_W): Se calcula como un promedio ponderado por población de los niveles de desigualdad dentro de cada grupo. Por ejemplo, dentro de los republicanos puede haber diferencias en la intensidad de apoyo entre los más fanáticos y los menos comprometidos.

La desigualdad total se expresa como la suma de estos dos términos:

$$G = G_B + G_W$$

A partir de esta descomposición, el índice de polarización se define como:

$$P = \frac{\mu}{m}(G_B - G_W)$$

Este índice mide la desigualdad entre los dos subgrupos (G_B), restando la desigualdad interna de cada grupo (G_W), todo ello reescalado por el factor μ/m. Dado que el coeficiente de Gini global es la suma de G_B y G_W, esta descomposición refuerza la diferencia conceptual entre desigualdad y polarización.

- Mayor desigualdad entre los dos subgrupos (G_B grande): aumenta tanto la desigualdad como la polarización.

- Mayor desigualdad dentro de los grupos (G_W grande): incrementa la desigualdad total, pero reduce la polarización, ya que los grupos internamente son más diversos y menos diferenciados entre sí.

La figura 4.8 ilustra esta relación. En la distribución lineal a trozos, si las medias de ambos grupos son similares, la mediana de la curva de Lorenz aumentará, las rectas correspondientes estarán más elevadas y la curva se «aplanará», resultando en un G_B pequeño y menor desigualdad entre grupos. En cambio, si las medias difieren significativamente, el área entre los grupos será mayor, aumentando G_B y reflejando una mayor polarización.

Por otro lado, G_W se puede interpretar como la suma de los coeficientes de Gini dentro de cada grupo, normalizada para cada subpoblación. Si los grupos son internamente más heterogéneos, la desigualdad dentro de ellos aumentará, pero la distancia relativa entre ellos se reducirá, disminuyendo el efecto polarizador.

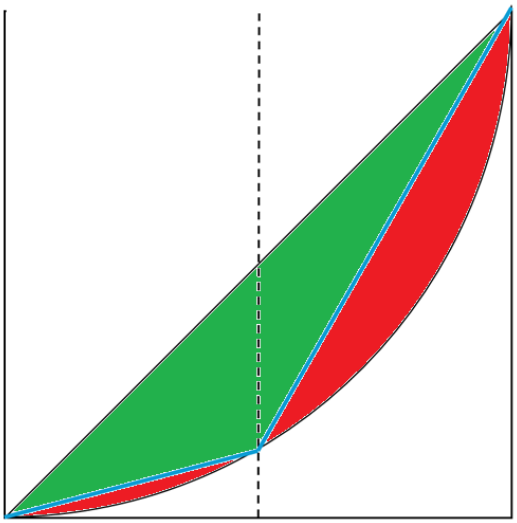

Figura 4.8: En azul se muestra la denominada distribución lineal a trozos, donde cada individuo opina la media de su grupo. El área verde representa $\frac{G_B}{2}$ y el área roja, dado que $G = G_B + G_W$, representa $\frac{G_W}{2}$. Fuente: elaboración propia.

4.2.4. Ejemplo de Aplicación

Para ilustrar el cálculo del índice de Foster y Wolfson, consideremos dos ejemplos con una distribución discreta de la variable v_i, que representa la *vote_intention* de cada individuo, si es necesario haciendo una traslación para que solo pueda tomar valores entre $[0, 1]$.

Consideremos una muestra de $n = 1000$ votantes texanos en 2024 con la siguiente distribución de intención de voto:

- $p_1 = 0{,}50$ (500 votantes apoyan fuertemente a Trump, es decir, $v_i = 1$).
- $p_{0,5} = 0{,}20$ (200 votantes son neutrales o indecisos, es decir, $v_i = 0{,}5$).
- $p_0 = 0{,}30$ (300 votantes apoyan fuertemente a Harris, es decir, $v_i = 0$).

Al ordenar los valores, el percentil 50 cae dentro del grupo neutral, por lo que la mediana es $m = 0{,}5$. Para aplicar el índice de Foster y Wolfson, dividimos la muestra en dos grupos: republicanos ($v_i > m$) y demócratas ($v_i < m$). Como el número de votantes indecisos es considerable, no podemos ignorarlo, y como hay 500 republicanos y solo 300 demócratas, asignamos los 200 neutrales completamente al grupo demócrata para igualar la masa de ambos grupos en 0.5 (esto no significa que tomemos los neutros como demócratas y los cálculos no se verán influenciados).

$$\mathcal{R} = 0{,}50, \quad \mathcal{D} = 0{,}30 + 0{,}20 = 0{,}50$$

La media total es:

$$\mu = (1)(0{,}50) + (0{,}5)(0{,}20) + (0)(0{,}30) = 0{,}50 + 0{,}10 = 0{,}60$$

Las medias de cada grupo son:

$$\mu^+ = \frac{(1)(0{,}50)}{0{,}50} = 1, \quad \mu^- = \frac{(0)(0{,}30) + (0{,}5)(0{,}20)}{0{,}50} = \frac{0{,}10}{0{,}50} = 0{,}20$$

Verificamos que tras la reasignación de los neutros se cumple:

$$\mu = \frac{\mu^+ + \mu^-}{2} = \frac{1 + 0{,}20}{2} = 0{,}60$$

El término T es:

$$T = \frac{\mu^+ - \mu^-}{\mu} = \frac{1 - 0{,}20}{0{,}60} = \frac{0{,}80}{0{,}60} = 1.\overline{3}$$

El coeficiente de Gini G se puede calcular con la fórmula discreta como se indica en Cowell (2011):

$$\begin{aligned} G &= \frac{1}{2\mu} \sum_{i,j} p_i p_j |v_i - v_j| \\ &= \frac{1}{2(0{,}60)} \left[(0{,}50)(0{,}20)(0{,}5) + (0{,}50)(0{,}30)(1)\right. \\ &+ (0{,}20)(0{,}50)(0{,}5) + (0{,}20)(0{,}30)(0{,}5) \\ &\left. + (0{,}30)(0{,}50)(1) + (0{,}30)(0{,}20)(0{,}5)\right] \\ &= \frac{1}{1{,}20} \left[0{,}05 + 0{,}15 + 0{,}05 + 0{,}03 + 0{,}15 + 0{,}03\right] \\ &= 0{,}3833 \end{aligned}$$

Finalmente, el índice de Foster y Wolfson es:

$$P = 2 \cdot \left(\frac{\mu}{m}\right) \cdot (T - G) = 2 \cdot \left(\frac{0{,}60}{0{,}5}\right) \cdot (1{,}3333 - 0{,}3833) = 2{,}4 \cdot 0{,}95 = 2{,}28$$

Este valor refleja una polarización significativa con grupos simétricos y bien separados respecto a la mediana, además, nos recuerda que este índice puede presentar valores mayores que 1 a diferencia del IPG estudiado en la sección anterior.

Consideremos ahora otra muestra de $n = 1000$ votantes pensilvanos en 2024 con la siguiente distribución:

- $p_{0,75} \approx 1/3$ (333 votantes apoyan moderadamente a Trump, es decir, $v_i = 0{,}75$).
- $p_{0,5} \approx 1/3$ (334 votantes son neutrales o indecisos, es decir, $v_i = 0{,}5$).
- $p_{0,25} \approx 1/3$ (333 votantes apoyan moderadamente a Harris, es decir, $v_i = 0{,}25$).

La mediana es $m = 0{,}5$. Para aplicar el índice de Foster y Wolfson, dividimos la muestra en dos grupos: republicanos ($v_i > m$) y demócratas ($v_i < m$). Asignamos los 334 votantes neutrales equitativamente: 167 a cada grupo.

$$\mathcal{R} = \frac{333 + 167}{1000} = 0{,}50, \quad \mathcal{D} = \frac{333 + 167}{1000} = 0{,}50$$

La media total es:

$$\mu = (0{,}75)(0{,}333)+(0{,}5)(0{,}333)+(0{,}25)(0{,}333) = 0{,}24975+0{,}1665+0{,}08325 = 0{,}4995 \approx 0{,}5$$

Las medias de cada grupo son:

$$\mu^+ = \frac{(0{,}75)(0{,}333) + (0{,}5)(0{,}167)}{0{,}5} = \frac{0{,}24975 + 0{,}0835}{0{,}5} = \frac{0{,}33325}{0{,}5} = 0{,}6665$$

$$\mu^- = \frac{(0{,}25)(0{,}333) + (0{,}5)(0{,}167)}{0{,}5} = \frac{0{,}08325 + 0{,}0835}{0{,}5} = \frac{0{,}16675}{0{,}5} = 0{,}3335$$

Verificamos:

$$\mu = \frac{\mu^+ + \mu^-}{2} = \frac{0{,}6665 + 0{,}3335}{2} = 0{,}5$$

El término T es:

$$T = \frac{0{,}6665 - 0{,}3335}{0{,}5} = \frac{0{,}333}{0{,}5} = 0{,}666$$

El coeficiente de Gini G se calcula usando los tres posibles valores y asumiendo la simetría al multiplicar por 2 para no tener que calcularlo sobre todos los pares posibles, sino solo cuando $i < j$:

$$\begin{aligned} G &= \frac{1}{2\mu} \sum_{i<j} 2p_i p_j |v_i - v_j| \\ &= \frac{2}{2 \cdot 0{,}5} \cdot \left[(0{,}333)^2(0{,}5) + (0{,}333)^2(0{,}25) + (0{,}333)^2(0{,}25)\right] \\ &= 2 \cdot (0{,}111 \cdot 0{,}5 + 0{,}111 \cdot 0{,}25 + 0{,}111 \cdot 0{,}25) = 2 \cdot (0{,}0555 + 0{,}02775 + 0{,}02775) = 0{,}222 \end{aligned}$$

Finalmente, el índice de Foster y Wolfson:

$$P = 2 \cdot \left(\frac{0{,}5}{0{,}5}\right) \cdot (0{,}666 - 0{,}222) = 2 \cdot 0{,}444 = 0{,}888$$

Este valor refleja una polarización más baja en comparación con el caso anterior, lo cual se corresponde con la intuición por las diferencias entre las distribuciones que evidencia la figura 4.9.

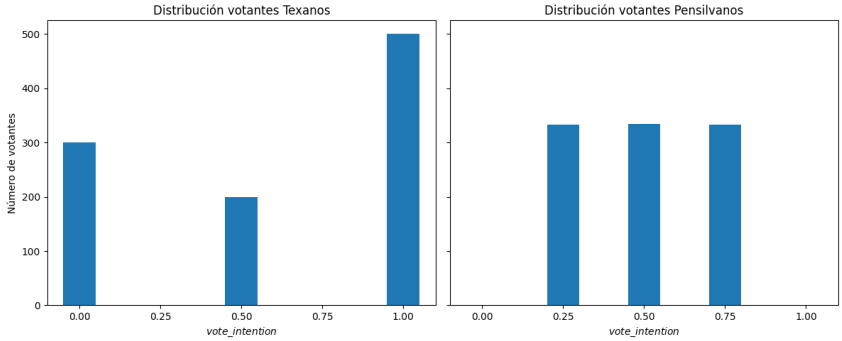

Figura 4.9: Gráficos de barras que muestran las distribuciones discretas de votantes texanos y pensilvanos del ejemplo. Fuente: elaboración propia.

4.3. Índice de Esteban y Ray

4.3.1. Definición

Esteban y Ray, en su artículo de 1994, proponen otro de los primeros modelos formales para medir la polarización dentro del marco de la economía. Su definición se centra en identificar distribuciones polarizadas que cumplan las siguientes características:

- Homogeneidad intra-grupal: los miembros de cada grupo comparten atributos similares.
- Heterogeneidad inter-grupal: los grupos son claramente distintos entre sí.
- Pocos grupos de tamaño significativo: los grupos pequeños o individuos aislados tienen un peso despreciable.

Para formalizar su índice, consideramos $\pi, y \in \mathbb{R}^n$ y una distribución

$$(\pi, y) = (\pi_1, ..., \pi_n; y_1, ..., y_n)$$

donde $\forall i \neq j \ y_i \neq y_j$ y $\forall i \ \pi_i > 0$. La población total está dada por $\sum_{i=1}^{n} \pi_i$. En nuestro caso, los datos provienen de la variable *vote_intention*, con y_i representando la opinión política del individuo i y π_i el tamaño del grupo de los individuos que presentan exactamente el mismo valor de opinión que i.

El modelo se construye sobre las características con las que los autores definen la polarización, consta de dos componentes fundamentales.

En primer lugar, la identificación intra-grupal. El modelo asume que un individuo se siente identificado con quienes comparten su misma opinión política. Se introduce una función de identificación continua $I : \mathbb{R}^+ \to \mathbb{R}^+$ tal que $I(\pi) > 0$ si $\pi > 0$, donde π representa el tamaño del grupo.[3]

[3]El modelo adopta una forma estricta de identificación en la que los individuos solo se sienten vinculados a quienes tienen exactamente el mismo valor de opinión, aunque para que tenga más sentido, con el mismo modelo podemos considerar las opiniones como estimaciones de opiniones

En segundo lugar, el aislamiento inter-grupal. El aislamiento, en textos originales referido como *alienation*, surge de la diferencia entre opiniones. Se define una función continua $a : \mathbb{R}^+ \to \mathbb{R}^+$ con $a(0) = 0$, llamada función de aislamiento. El aislamiento entre los individuos i y j se calcula como $a(\delta(y_i, y_j))$, donde $\delta(y_i, y_j) = |y_i - y_j|$ es la distancia o diferencia de opiniones. Además, a es no decreciente, ya que es razonable pensar que a mayor distancia de opiniones el aislamiento crecerá o al menos se mantendrá constante. Este concepto es simétrico: el aislamiento de i hacia j es el mismo que el de j hacia i.

Finalmente, definimos el antagonismo efectivo que un individuo siente hacia otro como combinación del aislamiento entre los dos individuos con la identificación del grupo del primer individuo (el sujeto). Se define mediante una función continua $T(I, a)$, donde $I = I(\pi)$ y $a = a(\delta(y_i, y_j))$. Se requiere que:

- Cuando el aislamiento sea 0 el antagonismo lo sea también, es decir, $T(I, 0) = 0$ para toda I.

- El antagonismo efectivo para dos individuos crezca cuando el aislamiento entre ellos crezca, es decir, que T sea estrictamente creciente en a cuando $I, a > 0$ [4].

Estos requerimientos permiten que la función identificación tenga efecto o no en la expresión del antagonismo. Por ejemplo, $T(I, a) = a$ es una forma válida si se desea ignorar el efecto de la identificación.

Ahora tenemos las herramientas para presentar el índice de Esteban y Ray, que se obtiene sumando los antagonismos efectivos entre todos los pares de individuos:

$$P(\pi, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i \pi_j \, T\left(I(\pi_i), a\left(\delta(y_i, y_j)\right)\right) \tag{4.1}$$

Aquí, el producto $\pi_i \pi_j$ representa el número de interacciones entre individuos del grupo de i y el grupo de j, mientras que $T(I(\pi_i), a(\delta(y_i, y_j)))$ mide el antagonismo que cualquier individuo del grupo de i siente hacia cualquier individuo del grupo de j. Es irrelevante de qué individuo del grupo se trate puesto que todos tienen la misma opinión. No es necesario excluir del sumatorio el antagonismo efectivo que un individuo siente hacia sí mismo (los términos con $i = j$) ya que en ese caso por los requerimientos exigidos: $\delta = 0$, $a = 0$ y, por tanto, como es lógico $T = 0$. Además, se incluyen tanto el antagonismo de i hacia j como el de j hacia i, que pueden ser distintos debido a los diferentes valores de $I(\pi)$.

En una primera formulación, los autores dejan sin especificar las formas funcionales de I, a y T, permitiendo así que distintas elecciones generen diferentes medidas de polarización.

---

grupales o intervalos. Aunque los autores discuten posibles extensiones que permitan la identificación parcial con individuos con opiniones cercanas pero no iguales, por simplicidad, el modelo se mantiene en esta versión.

[4] Simplemente se exigen valores positivos porque si $a = 0$ entonces $T(I, 0) = 0$ constantemente y si $I = 0$ entonces por lo exigido a a podríamos decir que T es no decreciente pero no podemos asegurar crecimiento estricto. Por último, ni I ni a pueden tomar valores negativos.

4.3.2. Propiedades

A continuación, se proponen una serie de propiedades o axiomas razonables que deben cumplir las funciones implicadas en el modelo. Estos criterios permiten restringir el conjunto de posibles formulaciones a una clase muy específica, de la cual se deriva finalmente el índice concreto de Esteban y Ray. Cada uno de los axiomas se explica intuitivamente a continuación.

- **Axioma 1:** Supongamos que existen dos grupos con opiniones muy cercanas entre sí y ambos son más pequeños que un tercer grupo con opinión más lejana. Si se agrupan los dos pequeños en uno solo, se aumenta la identificación intra-grupal sin modificar la distancia promedio respecto al tercer grupo. Esto debería, por tanto, incrementar la polarización. Ver figura 4.10.

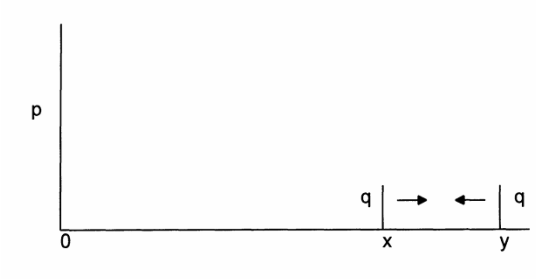

Figura 4.10: Axioma 1. Sean $p, q > 0$, $p > q$, $0 < x < y$. Fijamos $p > 0$ y $x > 0$. Entonces, existen $\varepsilon > 0$ y $\mu > 0$ (posiblemente dependiendo de p y x) tal que si $\delta(x, y) < \varepsilon$ y $q < \mu p$, entonces la unión de las dos masas q en su punto medio, $(x + y)/2$, aumenta la polarización. Fuente: Esteban y Ray (1994).

- **Axioma 2:** Consideremos un grupo situado entre otros dos, más cerca del de menor tamaño. Si solo se permiten pequeños desplazamientos del grupo intermedio, el movimiento que lo acerque al grupo más pequeño y cercano debería incrementar la polarización. Ver figura 4.11.

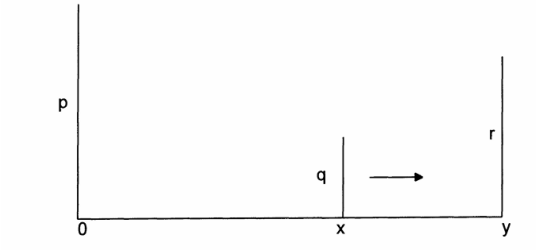

Figura 4.11: Axioma 2. Sean $(p, q, r) > 0$, $p > r$, $x > |y - x|$. Entonces, existe $\varepsilon > 0$ tal que si la masa de población q se mueve a la derecha (hacia r) en una cantidad que no excede ε, la polarización aumenta. Fuente: Esteban y Ray (1994).

- **Axioma 3:** La desaparición de una clase de opinión neutral por absorción en los grupos extremos debe aumentar la polarización. Ver figura 4.12.

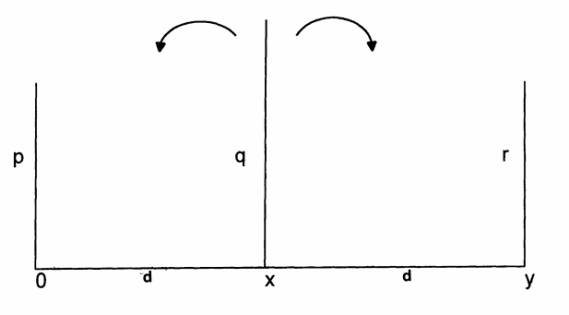

Figura 4.12: Axioma 3. Sean $(p, q) > 0$, $x = y - x = d$. Entonces, cualquier nueva distribución formada al desplazar masa de población de la masa central q equitativamente a las dos masas laterales p, cada una a una distancia d, debe aumentar la polarización. Fuente: Esteban y Ray (1994).

Además, se introduce una propiedad adicional: la de la homotecia. Esta supone que el orden entre dos distribuciones según el índice de polarización no se altera si se escala la población total. Formalmente, si $P(\pi, y) \geq P(\pi', y')$ para dos distribuciones, entonces para todo $\lambda > 0$ se cumple que $P(\lambda\pi, y) \geq P(\lambda\pi', y')$.

Los autores demuestran que una medida de polarización P^* de la familia definida en la ecuación 4.1 satisface los Axiomas 1, 2, 3 y la propiedad de la homotecia si, y solo si, es de la forma:

$$P^*(\pi, y) = K \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{1+\alpha} \pi_j \left| y_i - y_j \right| \tag{4.2}$$

donde $K > 0$ es una constante de normalización (sin impacto en el orden del índice), y $\alpha \in (0, 1{,}6]$[5].

Esta nueva formulación del índice de Esteban y Ray es la que aparece con frecuencia en la literatura. Comparándola con la versión general presentada anteriormente, se observa que:

$$T(I(\pi_i), a(\delta(y_i, y_j))) = K\pi_i^{\alpha}\delta(y_i, y_j)$$

es decir, el antagonismo efectivo T es proporcional a la distancia entre opiniones de los individuos antagónicos δ y a una potencia del tamaño del grupo π con el que el individuo sujeto se identifica[6].

[5]El valor máximo de α se ha estimado mediante análisis numérico a partir de una función auxiliar $f(z, \alpha)$, construida para verificar si el índice cumple los axiomas deseados en distintas configuraciones de población. Analíticamente comprueban que debe verificarse que el límite superior de α esté en el intervalo $(1, 2)$ y con cálculo numérico, evaluando f para muchos valores de z y α, encuentran ese límite superior en aproximadamente 1,6 y se prueba que los axiomas se cumplen si solo si $\alpha \lesssim 1{,}6$, lo que justifica la restricción del parámetro. Véase el artículo original Esteban y Ray (1994) para una discusión más detallada.

[6]La racionalidad detrás de que el antagonismo efectivo T sea proporcional a una potencia del tamaño del grupo con el que el primer individuo se identifica es debido a que T mide el desacuerdo **organizado**. Por ejemplo, en contextos como manifestaciones políticas, el tamaño del grupo con el que se identifica el individuo influye en la visibilidad y fuerza del antagonismo proyectado.

4.3.3. Consideraciones

Conviene señalar que la medida P^* presenta un parecido sorprendente con la conocida medida de desigualdad del coeficiente de Gini (véase la explicación del índice de Gini en el modelo de Foster y Wolfson 4.2.1). De hecho, la fórmula para calcular el coeficiente de Gini sobre distribuciones de probabilidad discreta es:

$$G = \frac{1}{2\mu} \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i \pi_j \left| v_i - v_j \right|$$

lo cual coincide con la medida de Esteban y Ray presentada en 4.2 si se toma $K = \frac{1}{2\mu}$ y $\alpha = 0$. Que en P^* los tamaños grupales se eleven a una potencia mayor que uno genera un comportamiento propio de una medida de polarización y distinto al de una medida de desigualdad. Así, el parámetro α puede interpretarse como el grado de sensibilidad a la polarización del índice: cuanto mayor sea α (dentro de sus límites establecidos) mayor es la desviación respecto a una medida de desigualdad.

Esta diferencia radica de un punto clave: mientras que las medidas de desigualdad no consideran la frecuencia poblacional de todas las categorías, las medidas de polarización sí lo hacen. La formación de grupos con tamaños significativos es esencial para que exista polarización –de hecho es la tercera característica presentada en la definición de este modelo que entienden los autores como necesaria para la existencia de polarización–.

Un ejemplo sencillo lo ilustra bien. Supongamos una sociedad igualitaria compuesta por campesinos. Si se redistribuye un excedente agrícola fijo de forma que una fracción λ de estos campesinos se convierten en campesinos ricos a expensas del resto, ¿para qué valores de λ podríamos decir que se ha generado una diferenciación social significativa?

Valores de λ cercanos a cero o uno difícilmente se interpretarían como una división clara en la población; una población no se considera muy polarizada si casi todos los campesinos tienen lo mismo y solo unos pocos son muy ricos ($\lambda \approx 0$) ni si casi todos son muy ricos y solo unos pocos son campesinos pobres ($\lambda \approx 1$). Sin embargo, menores valores de λ implicarían inequívocamente una mayor desigualdad.

Esto muestra que, al no tener en cuenta cuántas personas hay en cada categoría, las medidas de desigualdad no capturan el fenómeno de la polarización, que se basa precisamente en la existencia de grupos diferenciados con un tamaño significativo dentro de la población.

4.3.4. Ejemplo de Aplicación

Supongamos una población total normalizada fijando $K = 1$. Consideramos una distribución binaria de opiniones políticas en Estados Unidos: una fracción π de la población se identifica como demócrata (y la media de sus opiniones es x), y la fracción restante $1 - \pi$ como republicana (con media de opinión y), siendo $x < y$. Así, hay solo dos grupos ideológicos en disputa, como muestra la figura 4.13.

El índice de polarización propuesto por Esteban y Ray, en este caso, se expresa como:

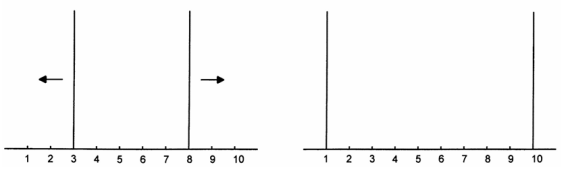

Figura 4.13: Distribución de dos puntos. Fuente: Esteban y Ray (1994).

$$P(\pi, x, y) = \left[\pi^{1+\alpha}(1-\pi) + (1-\pi)^{1+\alpha}\pi\right] \cdot (y-x) = \pi(1-\pi)\left[\pi^{\alpha} + (1-\pi)^{\alpha}\right] \cdot (y-x)$$

La expresión mide la intensidad del antagonismo entre ambos bloques, teniendo en cuenta la identificación intra-grupal y la distancia entre sus opiniones para reflejar el aislamiento inter-grupal.

Es inmediato ver el aumento de polarización que se produce al separar las opiniones: para π fijo, si $y - x$ aumenta (es decir, si las posiciones de los republicanos y demócratas se distancian más), entonces P también crece. Este comportamiento es coherente con la intuición: cuanto mayor es la distancia ideológica entre los grupos, mayor es la polarización.

Estudiamos ahora cómo varía la polarización en función del tamaño relativo de los grupos. Definimos:

$$f(\pi) = \pi^{1+\alpha}(1-\pi) + (1-\pi)^{1+\alpha}\pi$$

Esta función representa la parte del índice que depende de la distribución poblacional. Es simétrica respecto a $\pi = 0{,}5$, puesto que se verifica que $f(\frac{1}{2}+k) = f(\frac{1}{2}-k)$ para cualquier valor de k. Además en el anexo B demostramos que esta función presenta un máximo global en $\pi = 0{,}5$ para $\pi \in [0, 1]$, definido en ese intervalo por ser una fracción del total, y para cualquier f definida con $\alpha \in (0, 1{,}6]$. En la figura 4.14 graficamos f para varios valores de α.

Como se justifica en el anexo B, este resultado solo es válido si $\alpha < 2$, lo que justifica la restricción $\alpha \in (0, 1{,}6]$ discutida anteriormente: no solo garantiza el cumplimiento de los axiomas, sino también la validez estructural de este comportamiento fundamental del índice.

Por tanto, el índice alcanza su valor máximo cuando ambos grupos —republicanos y demócratas— tienen el mismo tamaño, reflejando una situación de máxima confrontación equilibrada: dos bloques simétricos y significativos, claramente diferenciados.

De hecho, la distribución bimodal $(\Pi/2, 0, \ldots, 0, \Pi/2)$, es decir, la distribución que asigna la mitad de la población a la clase de demócratas más extremistas y la otra mitad a la de republicanos más extremistas, es más polarizada que cualquier otra distribución $(\pi_1, \ldots, \pi_n)$ con población total Π, bajo cualquier medida P^* (para la demostración véase el Teorema 2 de Esteban y Ray (1994)).

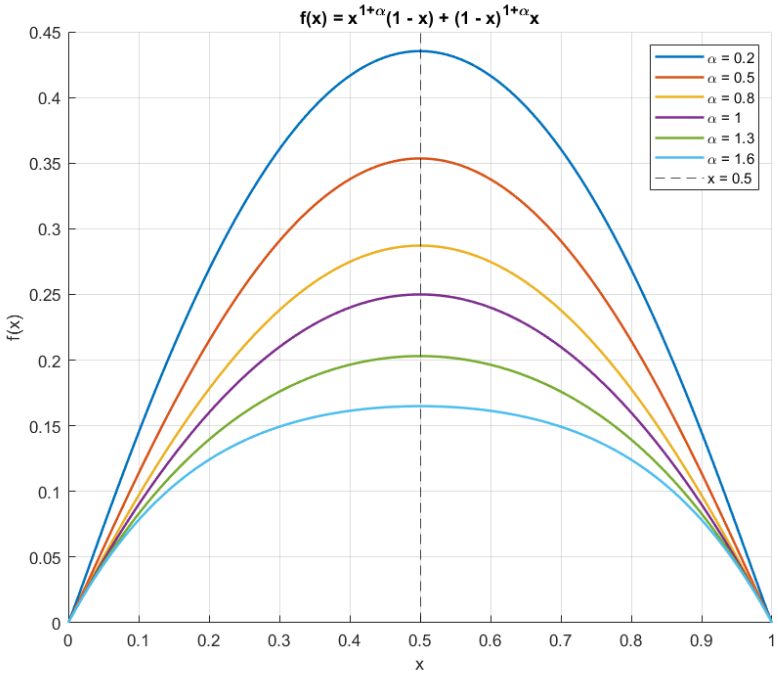

Figura 4.14: Gráficas de $f(x) = x^{1+\alpha}(1-x) + (1-x)^{1+\alpha}x$ para $x \in [0,1]$ y varios valores de α. Fuente: elaboración propia.

4.4. Nuestro modelo: Índice Beta de Polarización

4.4.1. Definición

Sea para cada usuario i la variable *vote_intention* $v_i \in [-1,1]$, donde los valores negativos indican orientación demócrata y los positivos republicana. Definimos los dos polos ideológicos:

$$D = \{i \mid v_i < 0\}, \qquad R = \{i \mid v_i > 0\}.$$

Los usuarios estrictamente neutros ($v_i = 0$) se reparten de forma proporcional entre ambos conjuntos[7]. Además, si alguno de los polos quedara vacío el índice se define como 0, pues no puede haber polarización bipartidista.

Para cada polo tomamos la mediana de la intensidad absoluta:

$$m_D = \text{mediana}\big\{|v_i| : i \in D\big\}, \qquad m_R = \text{mediana}\big\{|v_i| : i \in R\big\}.$$

Elegimos la mediana porque es robusta a valores extremos que podrían inflar la media con sólo unos pocos usuarios muy radicales.

[7]En los infrecuentes casos en los que una minoría de la población se sitúa en los extremos y una mayoría tienen una opinión exactamente neutra, ciertamente tras hacer el reparto proporcional se perdería la interpretabilidad de la situación al mirar los tamaños de L y R, no obstante, esto no influiría en el resultado del índice propuesto porque las medianas de ambos grupos se concentrarían en torno al 0 y se obtendría una polarización baja.

En el campo de la teoría de la información, la entropía de una variable aleatoria cuantifica el nivel promedio de incertidumbre o información asociada con los estados potenciales o resultados posibles de la variable. Etiquetamos cada *tweet* como a favor de los demócratas, neutro o a favor de los republicanos y, para cada usuario i, calculamos las proporciones p_{ik} con $k \in \{\text{dem}, \text{neu}, \text{rep}\}$. La firmeza de la opinión individual se mide con la entropía normalizada entre $\log(3)$ (entropía máxima de una multinomial de 3 clases[8]):

$$h_i \;=\; -\frac{1}{\log 3} \sum_k p_{ik} \, \log p_{ik} \;\in [0,1].$$

Así, un usuario i con $h_i = 0$ tendrá una opinión firme y un usuario con $h_i = 1$ presenta más dudas sobre su orientación política en su contenido publicado. Para evitar ruido, si un usuario tiene un número de *tweets* insignificante se le asigna $h_i = 0$ (opinión considerada clara).

La cohesión interna de cada bloque es la entropía media de sus miembros:

$$H_D = \frac{1}{|D|} \sum_{i \in D} h_i, \qquad H_R = \frac{1}{|R|} \sum_{i \in R} h_i.$$

Sea

$$p = \frac{\text{mín}\{|D|, |R|\}}{|D| + |R|}$$

la fracción del bloque minoritario: toma el valor máximo de 0.5 cuando los dos polos tienen igual tamaño y tiende a 0 cuando uno es muy pequeño.

Con los elementos anteriores definimos el Índice Beta de Polarización:

$$\text{Índice}_\beta \;=\; \big(|m_D| + |m_R|\big) \left[1 - \beta \frac{(H_D + H_R)}{2}\right] p, \qquad \text{Índice}_\beta \in [0,1].$$

- $\beta \in [0,1]$ es un peso que modula la importancia de la cohesión, cuanto mayor sea su valor más castiga el índice la falta de firmeza de opiniones. Por defecto en este texto, si no se especifica lo contrario tomaremos $\beta = 1$.

- $|m_D| + |m_R|$ crece cuanto más alejadas del centro están las posiciones típicas de ambos polos.

- $1 - \beta \frac{(H_D + H_R)}{2}$ refleja la cohesión de opiniones: es máximo cuando las opiniones de los usuarios son firmes ($H_D, H_R \approx 0$).

- p penaliza los escenarios en que un solo polo domina en número.

[8] En la fórmula de la entropía si $p_{ik} = 0$ para alguna k, surgirá una indeterminación de la forma $0 \cdot \log(0)$, para poder mantener la fórmula es común definir el abuso de notación $0 \cdot \log(0) = 0$. Lo más correcto técnicamente sería definir una función f, tal que $f(0) = 0$ y $f(x) = x \log(x)$ para $x \neq 0$, pero no entraremos en estos detalles.

4.4.2. Propiedades

- **Índice**$_\beta \approx 0$ si se verifica alguna de las condiciones:
  1. Uno de los polos es muy reducido ($p \approx 0$)
  2. Ambos polos mantienen posiciones moderadas ($|m_D|, |m_R| \approx 0$)
  3. Las opiniones internas son dispersas ($H_D, H_R \approx 1$)
- **Índice**$_\beta \approx 1$ cuando simultáneamente se verifican:
  1. Los dos polos tienen tamaños parecidos ($p \approx 0{,}5$)
  2. Sus medianas están en los extremos ($|m_D|, |m_R| \approx 1$)
  3. La cohesión interna es alta ($H_D, H_R \approx 0$)

4.4.3. Consideraciones

El Índice Beta de Polarización toma piezas esenciales de las medidas tradicionales definidas por Esteban-Ray y Foster-Wolfson como la distancia a los polos y el equilibrio poblacional. Además amplía la literatura clásica al incorporar la entropía individual que captura la firmeza o ambigüedad de las opiniones, aspecto poco tratado por los índices tradicionales centrados sólo en las opiniones medias. Esto resulta especialmente útil en el estudio de las redes sociales por tratarse de un espacio donde las opiniones pueden ser más fluidas y cambiantes, con usuarios que a menudo expresan puntos de vista contradictorios o ambivalentes, lo que dificulta la clasificación clara en polos ideológicos.

Para cada usuario i, la entropía normalizada h_i se comporta tal que cuando todas sus intervenciones son de signo único, $h_i \approx 0$ y su postura es firme; si reparte *tweets* entre las tres categorías, $h_i \to 1$ y su opinión es volátil. La cohesión de cada bloque se mide con H_D y H_R, las medias de h_i dentro de D y R, respectivamente. El factor $\left[1-\frac{1}{2}(H_D+H_R)\right]$ reduce el índice cuando un polo está compuesto por muchos individuos inseguros: un grupo cuyas opiniones fluctúan es más proclive a cambiar de puntos de vista y ofrece menor capacidad de conflicto por lo que provocarán menor polarización que otro de opiniones estables y permanentes.

Así, se obtiene una medida comprendida entre 0 y 1 que refleja simultáneamente intensidad ideológica, cohesión interna y equilibrio de tamaños, tres componentes que consideramos esenciales para describir la polarización en redes sociales.

4.4.4. Ejemplo de Aplicación

A continuación mostramos tres ejemplos teóricos que ilustran algunos comportamientos del índice con $\beta = 1$.

1. Variación por indecisión de usuarios individualmente. Tomamos $|m_D| = |m_R| = 1$ (posiciones extremas) y un equilibrio perfecto de tamaños ($p = 0{,}5$). Introducimos un parámetro $q \in [0, 0{,}5]$ que representa la probabilidad de que cada usuario publique un *tweet* contrario a su orientación. Además, cada usuario publica

un *tweet* coherente con su polo con probabilidad $1 - q$. Como no se publica ningún *tweet* neutro ahora solo hay dos clases de opinión y por tanto para normalizar se toma el $\log(2)$. La entropía individual normalizada es

$$h(q) = \frac{-(1-q)\log(1-q) - q\log q}{\log 2}, \qquad H_D = H_R = h(q),$$

y el índice se reduce a

$$I(q) = 1 - \frac{h(q) + h(q)}{2} = 1 - h(q).$$

Algunos valores representativos que se obtienen son:

$$I(0) = 1 \qquad I(0{,}1) \approx 0{,}53 \qquad I(0{,}3) \approx 0{,}12 \qquad I(0{,}5) = 0$$

Lo cual ilustra como el índice cae a medida que aumenta la «contaminación» del discurso de los mensajes, llegando a cero cuando los dos tipos de *tweets* son equiprobables.

2. Variación por desigualdad de tamaños de polos. Mantenemos extremos idénticos $|m_D| = |m_R| = 1$ y cohesión total ($H_D = H_R = 0$), pero permitimos que la fracción de usuarios republicanos sea $\lambda \in [0, 1]$.

$$p = \min\{\lambda, 1-\lambda\}, \qquad I(\lambda) = 2p = 2\min\{\lambda, 1-\lambda\}.$$

La curva es una campana isósceles: vale 1 en $\lambda = 0{,}5$ y decrece linealmente hasta 0 cuando un polo acapara toda la población.

3. Variación por proporción de usuarios indecisos. Tomamos $|m_D| = |m_R| = 1$ (posiciones extremas) y un equilibrio perfecto de tamaños ($p = 0{,}5$). Sea α la proporción de usuarios completamente firmes ($h_i = 0$); la clasificación de *tweets* del resto $(1 - \alpha)$ se distribuye equitativamente entre defensores de su polo ($\frac{1}{3}$), del contrario ($\frac{1}{3}$) y neutros ($\frac{1}{3}$), provocando que tengan una entropía individual $h_i = 1$.

$$H_D = H_R = 1 - \alpha, \qquad I(\alpha) = 1 - \frac{1-\alpha+1-\alpha}{2} = \alpha.$$

Así, en el caso límite perfectamente bipartito, el índice coincide exactamente con la fracción de usuarios ideológicamente convencidos, ofreciendo una interpretación directa: $I = 0{,}7$ implica que el 70 % de la población se mantiene inamovible en los extremos.

Capítulo 5

Aplicación Práctica de los Modelos

5.1. Datos Evaluados

Los datos empleados en este estudio corresponden a las tres elecciones presidenciales más recientes en Estados Unidos (2016, 2020 y 2024), y fueron recolectados durante el periodo electoral –desde una semana antes hasta unos días después de la jornada de votación de cada año– a partir de publicaciones en Twitter que mencionaban directamente a los candidatos presidenciales. Para las elecciones de 2016 (Trump vs. Clinton) y 2020 (Trump vs. Biden), los datos fueron proporcionados por el equipo de investigación del profesor Rafael Caballero Roldán, quienes los obtuvieron mediante la API oficial de Twitter como parte de sus propios trabajos académicos. Estos conjuntos incluyen tanto metadatos de los *tweets* como información relevante sobre los usuarios, permitiendo un análisis detallado de la actividad política en la red social durante esos ciclos electorales. En el caso de las elecciones de 2024 (Trump vs. Harris), la recolección de datos fue realizada directamente por los autores de este trabajo. Debido a los cambios recientes en las políticas de acceso a la plataforma –actualmente X– y a la imposibilidad de utilizar su API oficial por razones presupuestarias, se implementó un sistema de *web scraping* desarrollado en Python con herramientas como Selenium y ChromeDriver, lo que permitió extraer contenido textual de los *tweets* publicados por los mismos usuarios que participaron en elecciones anteriores, preservando así el enfoque longitudinal del estudio.

Para la aplicación práctica de este trabajo se ha tomado una muestra representativa de los datos mencionados, compuesta por aproximadamente 50 usuarios aleatorios en cada uno de los tres años analizados, con la condición de que los usuarios seleccionados cada año sumaran un total de al menos 150 *tweets* publicados en el periodo electoral. Esta selección permite observar de forma más controlada la evolución de la polarización en la sociedad a lo largo del tiempo.

Se han etiquetado manualmente los más de 450 *tweets*, con el objetivo de garantizar la mayor fiabilidad posible en los resultados. La clasificación se ha realizado utilizando las etiquetas 1, 0 y -1, que indican respectivamente un sentimiento positivo, neutro o negativo hacia el candidato republicano o demócrata mencionado en el *tweet*. Como cada *tweet* menciona únicamente a un candidato porque así han sido seleccionados, estas etiquetas luego se traducen a «rep» (republicano), «neu»

(neutral) y «dem» (demócrata) fácilmente.

Con el fin de comparar adecuadamente el rendimiento de los modelos propuestos, todos ellos serán aplicados sobre las distribuciones de la variable *vote_ intention* para las tres muestras de usuarios. Sin embargo, en el caso del modelo de Foster y Wolfson, se empleará la versión trasladada $vote\_\ intention_{[0,1]}$ debido a que ni la media ni la mediana de dicho índice pueden asumir valores negativos. Cabe señalar que, se han realizado numerosos estudios dedicados a la comparación entre diferentes índices de polarización y dicha tarea no es sencilla, ya que existen diferencias significativas de escala, traslación y otras propiedades estadísticas. Por ello, en este trabajo nos centraremos principalmente en analizar la evolución relativa de cada índice a lo largo del tiempo, sin entrar en detalles técnicos que exceden el alcance del presente estudio.

5.2. Análisis de Resultados

En la figura 5.1 se presenta en tres histogramas la evolución anual de la distribución de la variable *vote_ intention*, que constituye el objeto de estudio de los índices de polarización.

En las tablas 5.1 y 5.2 se muestran los valores obtenidos en los experimentos, redondeados respectivamente a cuatro cifras decimales y en forma porcentual. Las columnas muestran para cada índice el valor numérico resultante al aplicarlo a la distribución de la variable en cada uno de los tres años, organizados en filas.

El código utilizado para realizar todo el análisis de datos, así como la implementación detallada de los cuatro modelos evaluados, se encuentra documentado y disponible en el repositorio de GitHub:

`https://github.com/LauraRodrigoCanete/Aplicacion_Modelos_Polarizacion`.

Tabla 5.1: Valores de Índices de Polarización por Modelo y Año

Año	IPG	FW	ER ($\alpha = 1{,}5$)	ER ($\alpha = 1$)	ER ($\alpha = 0{,}5$)	IB ($\beta = 1$)	IB ($\beta = 0{,}5$)	IB ($\beta = 0$)
2016	0.6625	8.6930	0.2883	0.4010	0.5639	0.6648	0.7226	0.7805
2020	0.8016	9.2794	0.3283	0.4605	0.6494	0.7662	0.7998	0.8333
2024	0.6987	6.6334	0.3101	0.4300	0.6040	0.7067	0.7423	0.7778

Tabla 5.2: Porcentajes de Índices de Polarización (respecto a 2016) por Modelo y Año

Año	IPG	FW	ER ($\alpha = 1{,}5$)	ER ($\alpha = 1$)	ER ($\alpha = 0{,}5$)	IB ($\beta = 1$)	IB ($\beta = 0{,}5$)	IB ($\beta = 0$)
2016	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
2020	121 %	107 %	114 %	115 %	115 %	115 %	111 %	107 %
2024	105 %	76 %	108 %	107 %	107 %	106 %	103 %	100 %

Volviendo a las gráficas presentadas en la figura de la evolución de la variable *vote_ intention*, es importante destacar que, al tratarse de datos reales procedentes de redes sociales, la forma general de las distribuciones presenta patrones similares entre los tres años, lo cual es esperable dado el contexto de polarización constante que caracteriza a las elecciones presidenciales en Estados Unidos. Por ejemplo, a lo largo de los tres años vemos una agrupación principal de los usuarios en las posiciones extremas de los valores de la variable (-1 y 1) y una menor presencia de valores intermedios o moderados. No obstante, sí es posible identificar algunas diferencias significativas en las distribuciones. Sin contemplar aún los resultados de los índices, y basándonos únicamente en los histogramas de la figura, podemos ver cómo en 2016 la acumulación de valores en los extremos de *vote_ intention* es sustancialmente mayor en -1 que en 1. Esto implica una mayor presencia de usuarios con una intención de voto radical hacia el partido republicano en comparación con sus contrapartes demócratas. Al observar únicamente los valores extremos en los dos años siguientes, se aprecia cómo esta diferencia se va suavizando, aumentando el número de usuarios con una intención de voto clara hacia el partido demócrata.

Sin embargo, no solo los usuarios en los extremos juegan un papel relevante, sino que también es necesario contemplar la distribución en su conjunto. Por ello, proseguimos con el análisis de la polarización a través de los índices. A partir de una primera inspección visual de las tablas, se observa que, con excepción del índice de Foster y Wolfson, los demás índices presentan una misma tendencia: la mayor polarización se registra en 2020, seguida por 2024, siendo 2016 el año con menor polarización. En contraste, el índice FW muestra una mayor polarización en 2016 que en 2024, posiblemente debido a que en la distribución de 2024 la media y la mediana tienen el mismo valor, logrando un cociente del término del índice $\frac{\mu}{m} = 1$ mientras que en 2016 la distribución está ligeramente sesgada a la derecha o positivamente, causando un aumento del término $\frac{\mu}{m} = 3{,}26$ y aumentando ligeramente la polarización –no aumenta en un factor tan alto ya que este es solo uno de los términos que juegan un papel en el valor final del índice–. Además el modelo de FW no está acotado, lo que dificulta la interpretación directa de sus valores absolutos.

En cuanto a la variación de los índices entre los distintos años, los que muestran una mayor sensibilidad –es decir, una mayor diferencia porcentual entre sus valores mínimos y máximos– son el índice FW y el índice IPG. En el otro extremo, el índice que muestra una menor sensibilidad resulta ser IB para $\beta = 0$ que muestra solo una amplitud del 7 % entre sus valores mínimo y máximo.

Por otro lado, en términos de la magnitud de la polarización, después del modelo Foster-Wolfson, el Índice Beta presenta los valores absolutos de polarización más altos mientras que los Índices de Esteban-Ray tienden a mostrar los valores absolutos de polarización más bajos. No obstante, como ya hemos comentado al principio, por la diferencia entre escalas de estos índices nos abstendremos de profundizar en estas comparaciones absolutas.

Asimismo, se identifican algunas diferencias entre los índices que incluyen hiperparámetros. En el caso del índice ER, apenas se observan variaciones porcentuales al modificar el parámetro α, el cual representa el grado de sensibilidad del índice a la polarización. En cambio, si miramos los valores absolutos, a medida que α aumenta (de 0,5 a 1,0 y a 1,5), el índice de polarización absoluto disminuye para cualquier año

dado. Esto es debido a que α potencia las proporciones de los grupos (π^α y $(1-\pi)^\alpha$). Como estas proporciones son números entre 0 y 1, elevarlas a una potencia mayor las hace más pequeñas, reduciendo así el valor total del índice. Conceptualmente, α actúa como un factor de sensibilidad al aislamiento –un α más alto hace que el índice sea más exigente, otorgando valores de polarización más bajos para cualquier distribución dada–. Para que el índice muestre un valor alto de polarización cuando α es alto, la distancia entre los grupos $(y - x)$ debe ser muy grande y/o la división de la población en grupos opuestos ($\pi(1-\pi)$ se maximiza en $\pi = 0{,}5$) debe ser muy clara para superar esta atenuación.

En el caso de nuestro Índice Beta, en términos de porcentajes, al disminuir el valor de β, se reduce también la variabilidad de la polarización reportada, es decir, la sensibilidad del índice entre distinguir entre picos y valles disminuye. Dado que el parámetro β controla el peso asignado a la cohesión –penalizando la falta de firmeza en las opiniones–, estos resultados indican que el índice es capaz de captar y reflejar adecuadamente los matices presentes en los datos. En efecto, un análisis más detallado revela que un 27 %, 17 % y 20 % de los usuarios en los años 2016, 2020 y 2024 respectivamente, han sido asignados al menos dos etiquetas («dem», «rep» o «neu») distintas en sus *tweets*, lo que pone de manifiesto una cierta ambivalencia o evolución en sus posturas políticas. Esta variabilidad constituye una fuente de información valiosa que el índice puede incorporar. De hecho, cuando se fija $\beta = 1$, es decir, se otorga al índice la máxima capacidad de penalizar la falta de cohesión, los resultados obtenidos en porcentaje se aproximan notablemente a los del índice ER, mostrando como ambos índices son capaces de llegar a conclusiones similares con el uso de distinta información. Además, si nos fijamos en los valores absolutos vemos que a medida que β aumenta (de 0 a 0,5 y a 1), el índice de polarización absoluto disminuye para cualquier año dado. Esto se debe precisamente a que se le da al índice la oportunidad de que considere a esos usuarios que hemos mencionado que tienen opiniones más inestables o variables y que contribuyen a «relajar» a los polos y por tanto disminuyen la polarización globalmente. Esto demuestra que, mediante el uso de técnicas adaptadas a las particularidades del entorno digital –como lo son las redes sociales, caracterizadas por una alta densidad de mensajes o publicaciones–, es posible capturar con buena precisión las dinámicas subyacentes a la polarización política.

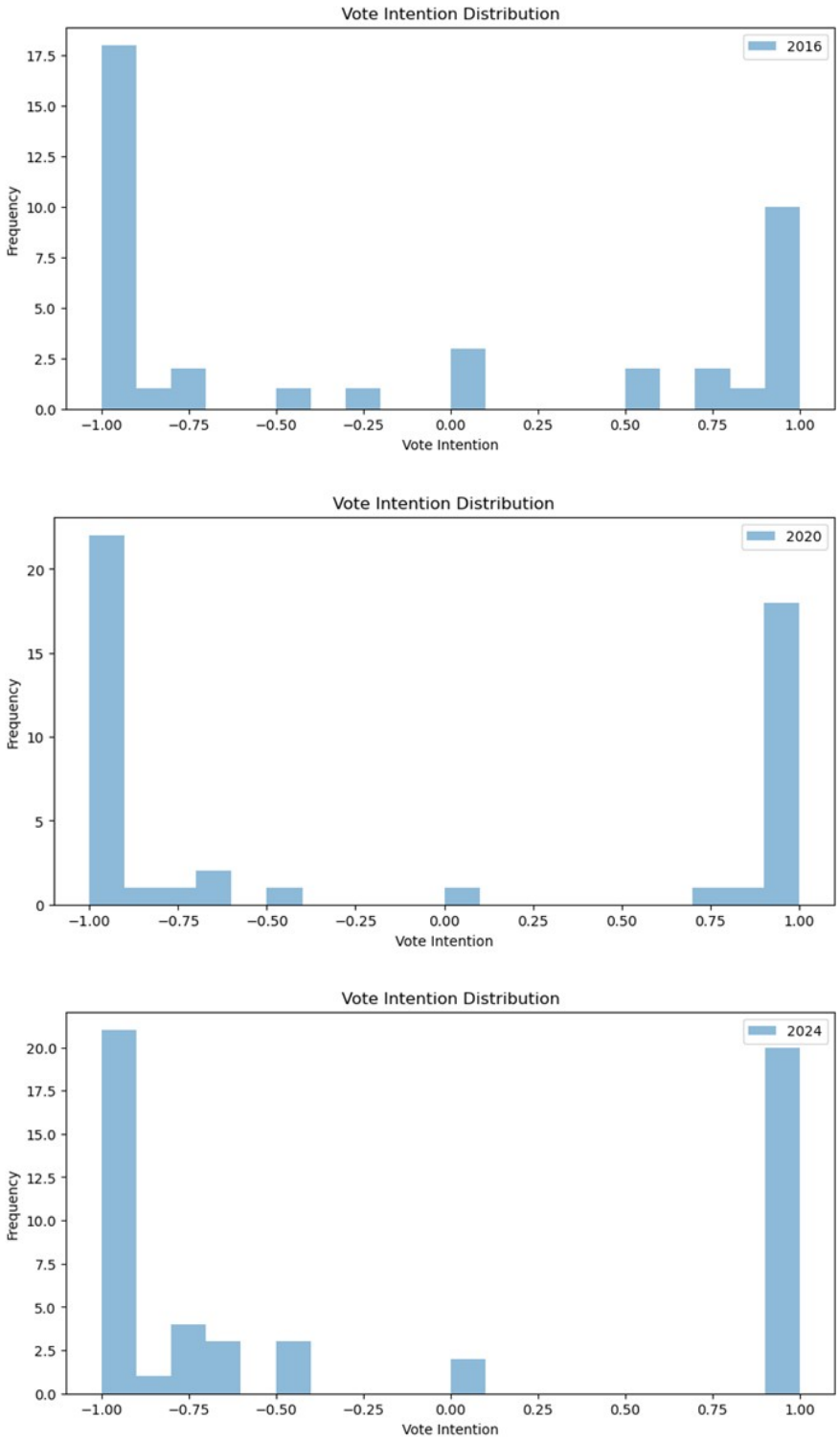

Figura 5.1: Gráficas que muestran la evolución de la distribución de la variable *vote_intention* sobre las muestras tomadas en 2016, 2020 y 2024. Fuente: elaboración propia.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

En el presente trabajo hemos investigado en profundidad tres índices representativos del estudio matemático de la medición de la polarización, además de formular el nuestro propio. Para ello hemos diseñado variables que funcionan como la entrada de estos modelos y que capturan información de las distribuciones de los datos. Además, hemos ilustrado los modelos con ejemplos y hemos terminado con el análisis de un caso práctico que daba pie a una breve comparación entre los índices.

En primer lugar, presentamos el índice que aquí hemos bautizado como Índice de Polarización Gravitatoria, ya que en su definición original en Morales et al. (2015) es nombrado como Índice de Polarización, un término poco específico para este texto. Este índice es reciente e intuitivo a la vez que eficiente. Es una función de la diferencia de los tamaños de grupos de opiniones opuestas y de la distancia entre sus centros de gravedad. Presenta buenas propiedades, como estar acotado en el intervalo $[0, 1]$ y ser simétrico e invariante a traslaciones respecto al intervalo de la distribución de opiniones que analiza. Además, sus fórmulas resultan especialmente directas para el caso de distribuciones gaussianas.

A continuación, se estudia el desarrollo del Índice de Foster y Wolfson, uno de los primeros en ser publicados y de los más reconocidos. Este índice depende principalmente del coeficiente de Gini G y de la desviación de la mediana relativa a la media T. Gracias a este índice estudiamos también en detalle el coeficiente de Gini, que –a pesar de medir desigualdad y no polarización directamente– tiene unas interpretaciones y unas ideas muy relevantes, y nos sirve para comprender como el Índice de Foster-Wolfson utiliza un enfoque muy similar a él. Interpretamos también gráficamente y de dos formas distintas el significado de la desviación de la mediana relativa a la media T, interpretaciones basadas en la curva de Lorenz, expresando T como el doble del llamado déficit de Lorenz y como el doble del área del trapecio que encierra a dicha curva. Así estas dos medidas G y T quedan estrechamente relacionadas. Por su relevancia histórica, también estudiamos la formulación original de este modelo, como el doble del área bajo la Segunda Curva de Bipolarización.

Seguidamente pasamos a otro de los índices más influyentes y fundamentales que se han propuesto en este área, el de Esteban y Ray. Este índice introduce una serie de

características como la homogeneidad intra-grupal y la heterogeneidad inter-grupal que deben de cumplir las distribuciones polarizadas y que en su índice se reflejan con las funciones de identificación y aislamiento. Con estas ideas se obtiene una formulación general del índice pero, tras proponerse una serie de axiomas, esta formulación queda explicitada casi completamente a excepción de un parámetro α (y una constante de normalización sin impacto alguno), que resulta ser un hiperparámetro de la medida. Resulta interesante como esta formulación final se asemeja en verdad a la del índice de Gini, salvo por este exponente dado por α, esto arroja luz sobre las diferencias entre la medición de la desigualdad y la polarización, y como la formación de grupos con tamaños significativos es esencial para que exista polarización.

Finalmente, y con ganas de formar nosotros también parte de esta larga lista de autores que han contribuido con sus ideas a la medición de un fenómeno de tal relevancia actualmente, nos animamos a plantear una propuesta de índice de polarización que puede tener buenos resultados especialmente en el contexto de las redes sociales. Nuestro índice, llamado Índice Beta por el hiperparámetro de tal nombre, tiene en cuenta elementos fundamentales que han quedado evidenciados por otros autores –como las medianas y los tamaños de los bloques–, pero además incorpora un concepto nuevo en este tema, como lo es la entropía de cada usuario sobre la distribución de sus opiniones en los *tweets*. En el índice se consideran más globalmente las medias de estas entropías individuales por grupos de opiniones. La intuición viene de que la entropía se comporta tal que cuando todas las intervenciones de un usuario son de un mismo carácter político, su postura es firme, y genera polarización y enfrentamiento directo con sus opuestos. No obstante, si este mismo usuario resulta publicar *tweets* con una opinión más repartida o fluctuante, aparenta tener una opinión más volátil o incluso ser más tolerante y más proclive a cambiar su punto de vista, por lo que constituye una menor amenaza de polarización.

Terminamos el trabajo con una aplicación práctica de todos los modelos a un conjunto de datos real de mensajes publicados en Twitter de votantes de las elecciones americanas. Observamos que, en general, los índices aplicados reflejan la mínima polarización en 2016, con un aumento de la polarización en las elecciones posteriores del 2020, seguida por una ligera disminución en 2024. Además, el análisis de los distintos índices de polarización revela que los más sensibles a la variación entre años son el índice de Foster-Wolfson y el Índice de Polarización Gravitatorio, mientras que el Índice Beta con $\beta = 0$ muestra la menor amplitud de cambio. En términos de magnitud absoluta, el Índice de Foster-Wolfson presenta los valores más altos y el de Esteban-Ray con $\alpha = 1{,}5$ los más bajos, aunque se evita comparar directamente debido a las diferencias de escala. En cuanto a los índices con hiperparámetros, se observa que el parámetro α del índice ER reduce los valores absolutos al aumentar, al hacer el índice más exigente con la distribución de los grupos. Por su parte, el Índice Beta refleja una mayor capacidad para capturar la variabilidad interna de las opiniones: al aumentar β, disminuyen los valores absolutos de polarización, ya que el índice penaliza la falta de cohesión que observamos empíricamente en los datos. Así, vemos que incorporar la entropía ofrece una manera alternativa de reflejar la diversidad y fluctuación de las posturas, proporcionando una visión más detallada y adaptada de la polarización en plataformas digitales.

6.2. Trabajo Futuro

En cuanto a trabajos futuros de investigación, a pesar de los avances en la medición de la polarización en tan poco tiempo, existen varias líneas de investigación que consideramos que pueden enriquecer este campo. Primero, aunque se han desarrollado enfoques basados en redes sociales, aún queda mucho por hacer en términos de dinámica temporal: la mayoría de los modelos actuales miden la polarización en un instante dado, pero pocos abordan cómo la polarización evoluciona a lo largo del tiempo, especialmente ante eventos políticos de gran magnitud, como elecciones o crisis sociales. Explorar modelos dinámicos que integren tanto la información temporal como la evolución de la estructura de un grafo puede proporcionar una visión más precisa de cómo se construyen y refuerzan las burbujas ideológicas en tiempo real.

Integrar este enfoque con otras variables, como el comportamiento de las audiencias en las redes sociales o el análisis de interacciones emocionales en debates, podría proporcionar una medición más rica. En este sentido, un área por explorar podría ser desarrollar métodos que midan la polarización desde una perspectiva relacional, considerando las interacciones entre usuarios y no solo sus características estáticas.

Por otro lado, la aplicación de herramientas de inteligencia artificial y *big data* a la medición de la polarización social y política en plataformas digitales está en una etapa temprana, por lo que aún queda mucho por investigar en cuanto a la precisión y robustez de estos enfoques.

La interacción entre diferentes tipos de polarización (por ejemplo, entre polarización política y social) es algo poco explorado, pero que podría indicar cómo los fenómenos de polarización se refuerzan mutuamente.

Finalmente, queda por investigar cómo estas mediciones de polarización pueden aplicarse a políticas públicas y estrategias de prevención de radicalización, de modo que los enfoques cuantitativos sean útiles en la intervención social y en el diseño de políticas que busquen mitigar la polarización extrema.

Introduction

6.3. The Problem of Polarization

The concept of polarization often appears in public discussions to describe situations where ideological positions of individuals or social groups are strongly opposed. However, precisely defining and rigorously measuring polarization is complex. This work presents a mathematical study of several key polarization measures proposed in the literature, analyzing their theoretical foundations, formal properties, and potential applications in social network analysis.

From a social perspective, polarization refers to the process by which opinions, attitudes, or beliefs within a society tend to cluster at opposing extremes, reducing consensus and increasing confrontation. On social media, this effect is amplified by recommendation algorithms, echo chambers, and the ability to selectively interact with like-minded content (Hartmann et al., 2025). As a result, users are mostly exposed to content that reinforces their existing views, a phenomenon known as the *filter bubble*, which can deepen ideological divides and complicate constructive dialogue (Pariser, 2011).

Mathematically, polarization is seen as a property of the distribution of a variable across a population, reflecting the concentration of individuals at distant positions with few people in the middle. Unlike inequality, which focuses on overall dispersion, polarization emphasizes the formation of distinct and cohesive groups with opposing views. This is usually captured through measures that account for both the distance between groups and their relative sizes.

It is important to distinguish from the beginning polarization from inequality, as they are conceptually and mathematically different. Classical inequality measures cannot distinguish between global convergence (everyone moving to a central position) and local convergence (formation of separate opinion clusters). For instance, in the evolution of income distribution across countries, we may observe two clearly defined poles: low-growth and high-growth countries. While this would indicate greater polarization, traditional inequality measures would show a decrease in inequality due to reduced global dispersion.

Figure 6.1 shows two illustrative examples. Suppose the bars represent income levels and the height of each bar the proportion of people in that group. In the first example (1A to 1B), the initial distribution is unequal but not polarized. It evolves

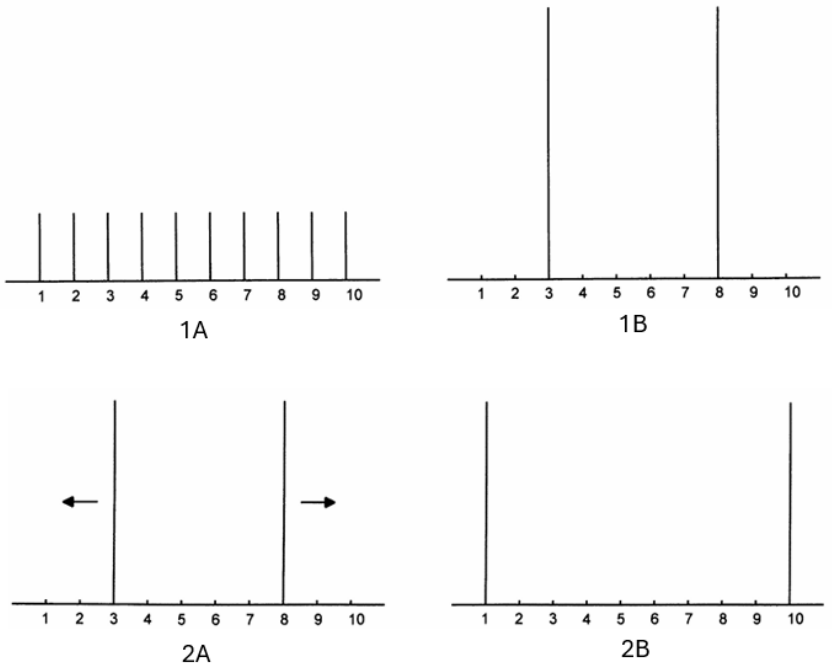

Figura 6.1: Two evolving distributions illustrating the difference between inequality and polarization. Source: Esteban y Ray (1994).

into a distribution with less inequality–only two income levels (3 and 8)–but greater polarization. In the second example (2A to 2B), two groups shift from middle-income positions to poor and rich. Internal homogeneity remains, but inter-group difference increases, leading to higher polarization. In this case, both inequality and polarization increase. Thus, while polarization and inequality are not mutually exclusive, they capture different aspects of social structure.

6.4. Objectives

The main goal of this work is to analyze and compare four polarization measures: the Gravitational Polarization Index, recently developed in a context similar to ours (political classification in Venezuela); two classic models, Foster-Wolfson and Esteban-Ray, and a new original index proposed in this work, the Beta Polarization Index, designed to fit the particularities of our case study.

We also explore the applicability of these measures to the case of bipartisan U.S. presidential elections. The analysis is based on a dataset of tweets mentioning candidates in the 2016, 2020, and 2024 elections. From this dataset, we extract political opinions and construct quantitative variables to apply and compare the polarization models.

The specific objectives are:

- Design variables that capture relevant information from political data to use as inputs for the polarization models.

- Study and classify polarization measures from the academic literature, focusing on their mathematical definitions, structural properties, and political implications.
- Design and propose a new polarization index tailored to our specific problem, combining classical ideas with new elements from empirical analysis.
- Apply and compare different polarization models using real Twitter data to assess their ability to reflect public opinion trends in a polarized political environment.

While based on extensive literature, this work includes several original contributions. Three of the indices already exist but were scattered across different contexts. We bring them together under a common framework, adapting their theoretical formulation to the context of public opinion in social media. Additionally, we propose a new index that incorporates individual entropy to capture opinion firmness or ambiguity–an aspect often overlooked by traditional metrics–. This is particularly relevant in social media analysis, where conviction strength is as important as ideological orientation. Another key contribution is the collection of real data and implementation of the four indices in *Python*, enabling a practical comparison and empirical insight into their behavior.

6.5. Work Plan

The project followed the steps below:

1. **Literature Review**: A study of theoretical and applied polarization literature. Key models were selected based on relevance and suitability for the political context.
2. **Index and Variable Design**: Using insights from the literature, we created a new index and designed one-dimensional and two-dimensional variables that reflect the data's key features and serve as inputs to the models.
3. **Practical Implementation and Analysis**: All polarization measures were implemented using *pandas* and *Jupyter Notebook*. A dataset of 1 million real tweets was collected using *Selenium* for web scraping. The models were applied to samples, and the results compared and analyzed.

   The code is available at:

`https://github.com/LauraRodrigoCanete/Aplicacion_Modelos_Polarizacion`

4. **Synthesis and Integration**: We wrote detailed analyses of each index, including their origins, definitions, and properties, with simple examples for better understanding. We integrated the empirical results to evaluate model performance and draw conclusions.

Although the analysis focuses on the U.S., the models are general and can be applied to other polarized political contexts involving two candidates, parties, or ideologies.

6.6. Structure of the Thesis

This work is divided into six main chapters, guiding the reader through the study of political polarization using social media data. Chapter 1, *Introduction*, introduces the polarization problem, project objectives, methodology, and structure. Chapter 2, *State of the Art*, reviews existing literature and methods for measuring polarization in different contexts. Chapter 3, *Initial Data*, defines the one- and two-dimensional variables used as inputs. Chapter 4, *Polarization Models*, presents the four indices, including the new proposal, detailing their definitions, properties, and examples. Chapter 5, *Practical Application*, analyzes U.S. Twitter data from the 2016, 2020, and 2024 elections and compares the indices' performance. Finally, Chapter 6, *Conclusions and Future Work*, summarizes the main findings and outlines directions for future research.

Conclusions and Future Work

6.7. Conclusions

In this work, we have explored in depth three key indices used in the mathematical study of polarization, and we have also proposed our own. To do this, we designed input variables for the models that capture the structure of the data distributions. We illustrated each model with examples and concluded with a practical case study that allowed for a brief comparison between the indices.

First, we presented the index we named the Gravitational Polarization Index, originally introduced in Morales et al. (2015) simply as the Polarization Index. We chose a more specific name to better fit this text. This recent and intuitive index is based on the difference in size between opposing opinion groups and the distance between their centers of gravity. It has desirable properties, such as being bounded in $[0, 1]$, symmetric, and invariant under translations within the opinion distribution interval. It also has particularly simple formulas in the case of Gaussian distributions.

Next, we analyzed the Foster-Wolfson Index, one of the first and most well-known polarization measures. This index is mainly based on the Gini coefficient G and the relative median deviation from the mean T. This led us to study the Gini coefficient in depth, as it plays a key role despite measuring inequality rather than polarization directly. We provided two interpretations of T using the Lorenz curve: as twice the Lorenz deficit and as twice the area of the trapezoid under the curve. These insights revealed a strong relationship between G and T. We also reviewed the original formulation of the index as twice the area under the Second Bipolarization Curve.

We then studied the influential Esteban-Ray Index, which introduces key concepts such as intra-group homogeneity and inter-group heterogeneity, represented through identification and alienation functions. A general formulation of the index was proposed and then refined using a set of axioms, resulting in a nearly closed formula except for a parameter α (and a normalization constant), which acts as a hyperparameter. Interestingly, the final formula closely resembles the Gini index, except for the α exponent. This highlights the key difference between inequality and polarization: polarization requires the formation of significantly sized groups.

Finally, aiming to contribute our own ideas to this field, we proposed a new index

called the Beta Index, named after its hyperparameter. This index is especially suitable for social media contexts. It includes common elements used by other indices—such as medians and group sizes—but also introduces a novel component: the entropy of each user's opinion distribution in their tweets. The index considers the average entropy within each opinion group. The intuition is that if a user consistently tweets from one political perspective, their stance is firm and contributes to polarization. In contrast, users who tweet with more varied opinions appear more flexible or tolerant, posing less risk of polarization.

We concluded the study by applying all models to real Twitter data from U.S. election voters. We found that polarization was lowest in 2016, increased in 2020, and slightly declined in 2024. Among the indices, the Foster-Wolfson and Gravitational indices were most sensitive to changes over time, while the Beta Index with $\beta = 0$ showed the least variation. In terms of absolute values, the Foster-Wolfson Index presented the highest values, and the Esteban-Ray Index with $\alpha = 1.5$ the lowest, though comparisons are limited by differences in scale. Regarding hyperparameters, increasing α in the ER index lowers the absolute values, making it more demanding on group structures. Similarly, increasing β in the Beta Index decreases polarization scores, as it penalizes internal incoherence within opinion groups. This shows how entropy helps reflect opinion diversity and fluctuation, offering a more detailed view of polarization in digital spaces.

6.8. Future Work

Although significant progress has been made in measuring polarization, there are still many directions worth exploring. First, most current models measure polarization at a specific point in time. Few address how polarization evolves, especially around major political events like elections or crises. Developing dynamic models that combine temporal information and evolving graph structures could give better insights into how ideological bubbles form and persist in real time.

Combining this with other data–such as audience behavior or emotional dynamics in debates–could enrich the analysis. One promising direction is to create relational models that focus on interactions between users, not just their static features.

Furthermore, the application of AI and big data tools to polarization measurement is still in early stages. Much remains to be done to improve the accuracy and robustness of these approaches.

Another underexplored topic is how different forms of polarization (e.g., political vs. social) interact and potentially reinforce each other.

Finally, there is a need to explore how polarization metrics can inform public policy and radicalization prevention strategies. The goal is to ensure that quantitative methods have practical value in designing interventions and policies aimed at reducing extreme polarization.

Bibliografía

Andrews, L. C. *Special functions of mathematics for engineers*, vol. 49. Spie Press, 1998.

Apouey, B. Measuring health polarization with self-assessed health data. *Health Economics*, vol. 16(9), páginas 875–894, 2007.

Böttcher, L. y Gersbach, H. The great divide: drivers of polarization in the us public. *EPJ data science*, vol. 9(1), páginas 1–13, 2020.

Cowell, F. A. *Measuring inequality*. Oxford University Press, 2011.

Dalton, R. J. The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative political studies*, vol. 41(7), páginas 899–920, 2008.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical association*, vol. 69(345), páginas 118–121, 1974.

Van der Eijk, C. Measuring agreement in ordered rating scales. *Quality and Quantity*, vol. 35, páginas 325–341, 2001.

Esteban, J., Ray, D. y Duclos, J.-Y. Polarization: concepts, measurement, estimation. *Econometrica*, vol. 72(6), páginas 1737–1772, 2004.

Esteban, J.-M. y Ray, D. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, páginas 819–851, 1994.

Foster, J. E. y Wolfson, M. C. Polarization and the decline of the middle class: Canada and the us mimeo. *Vanderbilt University*, vol. 31, 1992.

Foster, J. E. y Wolfson, M. C. Polarization and the decline of the middle class: Canada and the us. *The Journal of Economic Inequality*, vol. 8, páginas 247–273, 2010.

Garimella, K., Morales, G. D. F., Gionis, A. y Mathioudakis, M. Quantifying controversy on social media. *ACM Transactions on Social Computing*, vol. 1(1), páginas 1–27, 2018.

Gini, C. Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini, Bologna, Italy*, vol. 156, 1912.

Guerra, P., Meira Jr, W., Cardie, C. y Kleinberg, R. A measure of polarization on social media networks based on community boundaries. En *Proceedings of the international AAAI conference on web and social media*, vol. 7, páginas 215–224. 2013.

Guevara, J. A., Gómez, D., Robles, J. M. y Montero, J. Measuring polarization: A fuzzy set theoretical approach. En *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, páginas 510–522. Springer, 2020.

Hartmann, D., Wang, S. M., Pohlmann, L. y Berendt, B. A systematic review of echo chamber research: Comparative analysis of conceptualizations, operationalizations, and varying outcomes. *Journal of Computational Social Science*, vol. 8(2), página 52, 2025.

Losada, J. C., Robles, J. M., Benito, R. M. y Caballero, R. Love and hate during political campaigns in social networks. En *Complex Networks & Their Applications X: Volume 2, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10*, páginas 66–77. Springer, 2022.

Montalvo, J. G. y Reynal-Querol, M. Ethnic polarization, potential conflict, and civil wars. *American economic review*, vol. 95(3), páginas 796–816, 2005.

Morales, A. J., Borondo, J., Losada, J. C. y Benito, R. M. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25(3), 2015.

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

Nissanov, Z., Poggi, A. y Silber, J. Measuring bi-polarization and polarization: a survey. En *The Measurement of Individual Well-Being and Group Inequalities*, páginas 49–87. Routledge, 2013.

Pariser, E. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

Permanyer, I. y D' Ambrosio, C. Measuring social polarization with ordinal and categorical data. *Journal of Public Economic Theory*, vol. 17(3), páginas 311–327, 2015.

Rodríguez, J. G. y Salas, R. Extended bi-polarization and inequality measures. En *Inequality, Welfare and Poverty: Theory and Measurement*, páginas 69–83. Emerald Group Publishing Limited, 2003.

Waugh, A. S., Pei, L., Fowler, J. H., Mucha, P. J. y Porter, M. A. Party polarization in congress: A network science approach. *arXiv preprint arXiv:0907.3509*, 2009.

Apéndice A

Cálculo del Índice de Polarización Gravitatoria en una Distribución Normal

En este escenario, las opiniones de los usuarios siguen una función de densidad de probabilidad dada por

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

donde σ es la desviación estándar de la distribución. Ha de señalarse que para que las opiniones de la población estén en el intervalo $[-1, 1]$, y tenga sentido hablar del IPG tal y como lo hemos definido con sus límites de integración, la normal debe tener una desviación estándar lo suficientemente pequeña como para que solo una insignificante parte de las opiniones queden fuera de los límites del intervalo que empleamos. De no ser este el caso y tratarse de una distribución normal con una σ grande podemos aproximar la distribución normal por una distribución uniforme continua en el intervalo $[-1, 1]$. Gracias a esta aproximación garantizamos que se preserva la forma aplanada de la normal pero manteniendo todas las opiniones dentro del intervalo. A continuación se detallarán los cálculos para ambos casos.

Para calcular la distancia entre los centros de gravedad de las opiniones positivas y negativas, se evalúan las siguientes integrales:

$$gc^- = \frac{\int_{-1}^{0} xp(x)\,dx}{\int_{-1}^{0} p(x)\,dx}$$

$$gc^+ = \frac{\int_{0}^{1} xp(x)\,dx}{\int_{0}^{1} p(x)\,dx}$$

Es bien conocido en el campo de la estadística que la función de distribución acumulada de la distribución normal se puede expresar en términos de la función error de esta forma:

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

Con esto y tomando la media de la distribución $\mu = 0$ para nuestro caso, desarrollamos la expresión del denominador de gc^-:

$$\begin{aligned}
A^- &= \int_{-1}^{0} p(x)\, dx \\
&= \int_{-\infty}^{0} p(x)\, dx - \int_{-\infty}^{-1} p(x)\, dx \\
&= F(0) - F(-1) \\
&= \frac{1}{2}\left((1 + \operatorname{erf}(0)) - \left(1 + \operatorname{erf}\left(\frac{-1}{\sqrt{2}}\right)\right)\right) \\
&= -\frac{1}{2}\operatorname{erf}\left(\frac{-1}{\sqrt{2}}\right) \\
&= \frac{1}{2}\operatorname{erf}\left(\frac{1}{\sqrt{2}}\right)
\end{aligned}$$

Para el numerador resolvemos aplicando el cambio de variable $u = \frac{-x^2}{2\sigma^2}$:

$$\begin{aligned}
\int_{-1}^{0} xp(x)\, dx &= \int_{-1}^{0} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-1}^{0} x e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= -\frac{\sigma}{\sqrt{2\pi}} \int_{\frac{-1}{2\sigma^2}}^{0} e^u\, du \\
&= -\frac{\sigma}{\sqrt{2\pi}} e^u \Big|_{\frac{-1}{2\sigma^2}}^{0} \\
&= -\frac{\sigma}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \Big|_{-1}^{0} \\
&= -\frac{\sigma}{\sqrt{2\pi}} \left(1 - e^{\frac{-1}{2\sigma^2}}\right)
\end{aligned}$$

Por lo tanto, combinando ambos términos, obtenemos el centro de gravedad negativo:

$$gc^- = \frac{-\frac{\sigma}{\sqrt{2\pi}}\left(1 - e^{-1/(2\sigma^2)}\right)}{\frac{1}{2}\operatorname{erf}\left(\frac{1}{\sigma\sqrt{2}}\right)} = -\sigma\sqrt{\frac{2}{\pi}} \cdot \frac{1 - e^{-1/(2\sigma^2)}}{\operatorname{erf}\left(\frac{1}{\sigma\sqrt{2}}\right)} \tag{A.1}$$

Un cálculo análogo para el centro de gravedad positivo nos da:

$$gc^+ = -gc^-.$$

Finalmente, la distancia entre los centros de gravedad es

$$d = \frac{|gc^+ - gc^-|}{2} = |gc^-|$$

Dado que la distribución es simétrica respecto al origen, el tamaño relativo de los grupos es idéntico, es decir, $|\Delta A| = 0$. Sustituyendo en la ecuación del índice de polarización, se obtiene:

$$\mu = d = |gc^-|$$

Ahora para simplificar la expresión de $|gc^-|$ recordamos lo que se mencionaba al principio del apéndice sobre el tamaño de σ.

Si σ tiene un valor pequeño podemos continuar nuestros cálculos y emplear la aproximación que se menciona a continuación.

Cuando el argumento de la función de error es grande, ésta se aproxima a 1 como ilustra la Figura A.1:

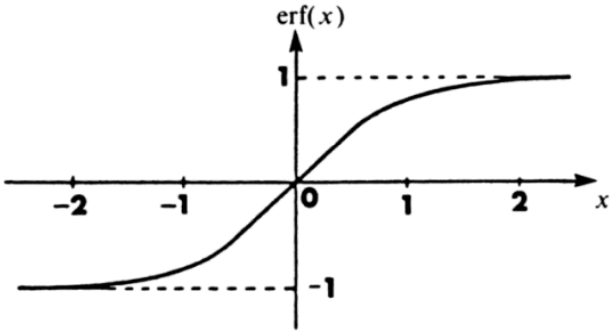

Figura A.1: La función error. Fuente: Andrews (1998).

$$\text{erf}(x) \approx 1, \quad \text{para } x \to \infty.$$

Por tanto,

$$\text{erf}\left(\frac{1}{\sigma\sqrt{2}}\right) \approx 1, \quad \text{para } \sigma \to 0.$$

Además,

$$\lim_{\sigma \to 0} e^{-\frac{1}{2\sigma^2}} = 0$$

Finalmente substituyendo esto en la fórmula A.1 obtenemos el resultado mencionado en el capítulo:

$$\mu = d = |gc^-| = \sigma\sqrt{2/\pi}$$

En concreto, ahora podemos observar que para que $\mu \in [-1, 1]$ y la mayoría de opiniones queden dentro del intervalo la condición de que σ sea pequeño se traduce en $\sigma\sqrt{\frac{2}{\pi}} < 1 \iff \sigma < \sqrt{\frac{\pi}{2}} \approx 1,2533$.

En el caso de que σ tenga un valor muy grande procedemos a aproximar la distribución normal por la uniforme, $U[-1,1]$. En este caso, calculamos los parámetros del índice de polarización gravitatoria.

Dado que la distribución es simétrica, la diferencia de tamaños de los grupos es:

$$\Delta A = 0$$

Para calcular el centro de gravedad del grupo negativo, evaluamos:

$$gc^- = \frac{\int_{-1}^{0} xp(x)\,dx}{\int_{-1}^{0} p(x)\,dx}$$

Dado que en una distribución uniforme $p(x) = \frac{1}{2}$ en el intervalo $[-1,1]$, el numerador se resuelve como:

$$\int_{-1}^{0} 0{,}5x\,dx = \frac{1}{2}\int_{-1}^{0} x\,dx = \frac{1}{2}\left[\frac{x^2}{2}\right]_{-1}^{0} = \frac{1}{2}\left(0 - \frac{1}{2}\right) = -\frac{1}{4}$$

El denominador es simplemente:

$$\int_{-1}^{0} \frac{1}{2}\,dx = \frac{1}{2}(0 - (-1)) = \frac{1}{2}$$

Por lo tanto:

$$gc^- = \frac{-1/4}{1/2} = -\frac{1}{2}.$$

De manera análoga, para el centro de gravedad positivo:

$$gc^+ = \frac{1}{2}$$

La distancia entre los centros de gravedad es:

$$d = \frac{|gc^+ - gc^-|}{2} = \frac{1}{2}$$

Finalmente,

$$\mu = (1 - |\Delta A|)d = \frac{1}{2}$$

Este resultado muestra que en una distribución uniforme en $[-1,1]$, el índice de polarización gravitatoria alcanza un valor medio, lo que refleja una considerable dispersión de opiniones en el intervalo considerado.

Apéndice B

Demostración de que $\pi = 0{,}5$ es el máximo global de la función $f(\pi)$

Sea la función definida como:

$$f(\pi) = \pi^{1+\alpha}(1-\pi) + (1-\pi)^{1+\alpha}\pi$$

para $\pi \in [0,1]$ y $\alpha \in (0,1{,}6]$. Queremos demostrar que $\pi = 0{,}5$ es su único máximo global en ese intervalo.

Calculamos la derivada de f:

$$f'(\pi) = (1+\alpha)\pi^{\alpha}(1-\pi) - \pi^{1+\alpha} - (1+\alpha)(1-\pi)^{\alpha}\pi + (1-\pi)^{1+\alpha}$$

En $\pi = 0{,}5$, todas las potencias simétricas coinciden:

$$f'(0{,}5) = (1+\alpha)(0{,}5)^{\alpha}(0{,}5) - (0{,}5)^{1+\alpha} - (1+\alpha)(0{,}5)^{\alpha}(0{,}5) + (0{,}5)^{1+\alpha} = 0$$

Verificamos que la segunda derivada en $\pi = 0{,}5$ es negativa, por lo que $\pi = 0{,}5$ es un máximo local:

$$f''(\pi) = (1+\alpha)(\alpha x^{1+\alpha}(1-x) + \alpha(1-x)^{1+\alpha}x - 2x^{\alpha} - 1(1-x)^{\alpha})$$
$$f''(0{,}5) = 0{,}5^{\alpha}(\alpha - 2)$$

Vemos que para que $f''(0{,}5) < 0$ y se verifique la condición de máximo debe garantizarse $0{,}5^{\alpha}(\alpha-2) < 0 \iff \alpha - 2 < 0 \iff \alpha < 2$ que se cumple porque tomábamos $\alpha \in (0,1{,}6]$.

Para ver ahora que es un máximo global, podemos reescribir la función como:

$$f(\pi) = \pi(1-\pi)\left[\pi^{\alpha} + (1-\pi)^{\alpha}\right]$$

Observamos que para $\pi \in [0,1]$:

$$\pi(1-\pi) \leq 0{,}25 \quad \text{con igualdad sólo en } \pi = 0{,}5$$
$$\pi^{\alpha} + (1-\pi)^{\alpha} \leq 2(0{,}5)^{\alpha} \quad \text{con igualdad sólo en } \pi = 0{,}5$$

Por tanto con eso hemos demostrado la globalidad, ya que:

$$f(\pi) = \pi(1-\pi)\left[\pi^{\alpha} + (1-\pi)^{\alpha}\right] < 0{,}25 \cdot 2(0{,}5)^{\alpha} = (0{,}5)^{1+\alpha} = f(0{,}5) \quad \text{para todo } \pi \neq 0{,}5$$

Concluyendo, para todo $\alpha \in (0,1{,}6]$ y para todo $\pi \in [0,1]$, la función $f(\pi)$ alcanza su único máximo global en $\pi = 0{,}5$.