



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Análisis de discursos políticos usando procesamiento del lenguaje natural (NLP)

Tesis de Licenciatura en Ciencias de Datos

Manuel Horn

Director: Pablo Balenzuela
Buenos Aires, 2025

ANÁLISIS DE DISCURSOS POLÍTICOS USANDO PROCESAMIENTO DEL LENGUAJE NATURAL (NLP)

Una cuestión fundamental en el comportamiento social colectivo se relaciona con la influencia del discurso público y mediático en la opinión electoral. En este trabajo se propone un enfoque metodológico para examinar cómo características del discurso político se relacionan con manifestaciones medibles de la opinión pública, aplicando una combinación de herramientas que involucran el procesamiento del lenguaje natural y el análisis de series temporales. Se aplicó un análisis de correlación al período correspondiente a la campaña presidencial para las elecciones de Estados Unidos del año 2024, utilizando texto correspondiente al contenido discursivo de los candidatos en carrera. Se compararon las variaciones en las encuestas de intención de voto nacional y estatal con cambios en la relevancia temporal de dichos tópicos nacionales. Los resultados sugieren que la relevancia de ciertos tópicos está correlacionada con las fluctuaciones en la intención de voto a nivel nacional y estatal, esto permite caracterizar a cada estado con los temas que mayor impacto tienen en él y realizar interpretaciones correspondientes. La metodología empleada en este trabajo es de carácter general y puede extenderse fácilmente a otros temas de interés y contextos electorales.

Palabras claves: Procesamiento del lenguaje natural, Aprendizaje no supervisado, Detección de tópicos, Análisis de series temporales, Opinión pública, Política

Índice general

1..	Introducción, antecedentes y objetivo	1
1.1.	Antecedentes	1
1.2.	Objetivo	1
2..	Datos y métodos	3
2.1.	Datos	3
2.1.1.	Extracción y estandarización de datos	4
2.2.	Métodos para detección de tópicos	5
2.2.1.	Text Clustering y Topic Modeling	5
2.2.2.	BERTopic	7
2.3.	Medidas de correlación	9
2.3.1.	Correlación de Pearson	10
2.3.2.	Correlación de Spearman	10
3..	Análisis de discursos	11
3.1.	Representaciones de discursos en el espacio de tópicos y comparación entre ellas	11
3.2.	Armado de series temporales de tópicos	12
3.3.	Descomposición en tópicos: resultados y validación	13
3.3.1.	Resultados	14
3.3.2.	Validación de tópicos	17
4..	Comparación discursos - encuestas	22
4.1.	Representación espacial de discursos por frecuencia y por tópicos	22
4.1.1.	Identificación de estados competitivos (“swing states”)	22
4.1.2.	Análisis de la estrategia de campaña por estado	23
4.1.3.	Representación espacial de tópicos	24
4.2.	Representación temporal y relevancia de los tópicos	31
4.3.	Configuración del análisis de correlación	33
4.3.1.	Definición de períodos de análisis	34
4.3.2.	Metodología de correlación con desplazamiento temporal para series de encuestas	34
4.3.3.	Alcance del análisis estatal	35
4.4.	Correlación entre el contenido de los discursos y las encuestas	36
4.4.1.	Diseño del análisis de correlación	36
4.4.2.	Interpretación de los coeficientes de correlación	36
4.5.	Resultados	37
4.5.1.	Nivel nacional	38
4.5.2.	Nivel estatal	40
5..	Conclusión	46
5.1.	Conclusiones generales	46
5.2.	Conclusiones más específicas y cosas a mejorar o agregar en el futuro	46

Apéndice	49
A.. Listados de Tópicos y Validación	50
A.1. Tópicos de Roll Call	50
A.2. Tópicos generados por BERTopic	51
A.3. Validación de tópicos	51
B.. Tablas de correlaciones	54
B.1. Nivel nacional	54
B.2. Nivel estatal	56
B.2.1. Arizona	56
B.2.2. Georgia	59
B.2.3. Michigan	62
B.2.4. North Carolina	65
B.2.5. Nevada	67
B.2.6. Pennsylvania	70
B.2.7. Wisconsin	73
C.. Test estadístico de permutación	76
C.0.1. Significancia estadística de correlaciones	76

1. INTRODUCCIÓN, ANTECEDENTES Y OBJETIVO

1.1. Antecedentes

Uno de los objetivos fundamentales del grupo de investigación (SoPhy Lab, [1]) es comprender los procesos de formación de opinión pública desde la perspectiva de los sistemas complejos. La opinión pública, entendida como el consenso emergente en una sociedad, resulta de la interacción de múltiples individuos que intercambian información, expresan sus ideas y ajustan sus posturas en función de factores internos y externos. Para abordar este problema, se utilizan herramientas provenientes de la teoría de sistemas complejos y el procesamiento de lenguaje natural (NLP), que permiten analizar patrones en la evolución de la opinión a gran escala. Un caso paradigmático de formación de opinión pública se observa en los períodos electorales, donde los individuos expresan sus preferencias a través de encuestas, debates y redes sociales. En investigaciones previas del grupo, el estudio del impacto de las noticias generadas por los medios de comunicación en este proceso, así como el rol de las redes sociales en su difusión, ocupan un rol preponderante. Se encontró que el flujo de información está condicionado por factores como la polarización política, la dinámica de las redes sociales y la cobertura mediática. Asimismo, las plataformas digitales desempeñan un rol clave en la difusión de información y en la segmentación de la audiencia en comunidades ideológicamente afines [2, 3, 4]. También se ha analizado el impacto de la agenda mediática en la percepción pública y en la intención de voto. En [5] se aplicaron técnicas de modelado de tópicos para analizar la evolución temporal de los temas predominantes en la discusión política y su relación con la opinión pública. Asimismo, [6] exploró el papel del contexto y el sentimiento en la influencia de las noticias en los procesos electorales en EE.UU., encontrando que no solo el contenido de la información, sino también su marco discursivo, juega un papel clave en la persuasión política. Por otro lado, [3] estudió la interacción entre ideología y consumo de noticias en el caso de las elecciones argentinas de 2019, identificando patrones en la forma en que los usuarios comparten información en función de su orientación política. Estas investigaciones proporcionan un marco sólido para el análisis de la formación de opinión en campañas electorales, destacando la importancia de la estructura de la información y su difusión en redes.

1.2. Objetivo

En este contexto, es fundamental estudiar el papel que juegan los discursos presidenciales en la evolución de la intención de voto. A través del análisis de su contenido y su recepción por parte del electorado, se puede obtener una caracterización más precisa de la influencia de los líderes políticos en la configuración del debate público.

El objetivo central de la tesis es identificar los temas que tienen mayor influencia en una campaña electoral presidencial, más específicamente para la última campaña presidencial de Estados Unidos, llevada a cabo en los años 2023 y 2024. El foco de estudio son los discursos, entrevistas, dichos, mensajes y menciones de figuras políticas, de los cuales se pueden extraer tópicos y medir su relevancia en el escenario político.

Se cuenta con el conjunto de discursos correspondientes a ciertas figuras públicas en un

período dado. Se llama informalmente “discurso” a cualquier registro escrito que se tenga de cada candidato; estos registros provienen de transcripciones de conferencias de prensa, ruedas de prensa, entrevistas, comentarios y discursos propios. Con estos datos, se busca separar su contenido en un espacio de tópicos que permita identificar temas centrales para posteriormente construir series temporales que representen la relevancia de cada uno y así comparar estas con las series temporales de intención de voto mediante un análisis de correlaciones. Esto se hace con el objetivo de identificar los temas centrales que más impactaron en la campaña electoral.

También, se busca realizar un análisis exhaustivo de la distribución geográfica y temporal de discursos para cada candidato a presidente. Esto tiene como objetivo ver si existe algún comportamiento o patrón de interés a nivel estatal.

Como fue mencionado anteriormente, el caso de estudio es la última campaña electoral presidencial de 2024 de Estados Unidos, donde los candidatos a presidente eran Donald Trump para el partido republicano, Kamala Harris y, previo al 21 de julio de 2024, Joe Biden para el partido demócrata. Formalmente, el período comprendido en el análisis es desde el 1ero de julio de 2023 hasta el 5 de noviembre de 2024, día de las elecciones.

En resumen, las preguntas centrales de esta tesis son:

- ¿Se puede encontrar una representación del contenido de los discursos en un espacio de tópicos adecuado para el análisis en cuestión? ¿Son estos tópicos relevantes al período electoral? Es decir, ¿Representan las temáticas abordadas por los candidatos en campaña?
- ¿Se puede encontrar una relación entre la presencia de cada candidato con el escenario electoral en cada estado del país?
- ¿Se puede identificar un conjunto de tópicos que definen, o al menos están relacionadas a tendencias de voto en cada estado del país de una manera confiable? ¿De hacerlo, esto puede explicar la naturaleza de cada estado (en qué temática cada estado concentra su atención)? ¿Los resultados son fácilmente interpretables?
- ¿Se puede hacer este mismo análisis a nivel nacional?

La presente tesis se estructura de la siguiente manera. La introducción presenta el problema central, los antecedentes teóricos y el objetivo de la investigación. A continuación, la segunda sección, de datos y métodos, detalla el proceso de recolección de datos y las herramientas utilizadas para su procesamiento. La tercera sección se dedica al análisis de contenido de los discursos, donde se expone el flujo de trabajo para la obtención de tópicos relevantes. Posteriormente, en la cuarta sección se analiza el contexto electoral de Estados Unidos a través de un estudio geográfico y temporal, cuyos resultados se comparan con las series de intención de voto para realizar un análisis de correlaciones y discutir resultados a nivel nacional y estatal. Finalmente, las conclusiones ofrecen una reflexión general sobre el trabajo, discuten sus limitaciones y proponen líneas para futuras investigaciones.

2. DATOS Y MÉTODOS

2.1. Datos

Se contó con tres fuentes principales de datos.

La primera, Roll Call [7], es una de las fuentes más completas de noticias, artículos, archivos y cobertura de las elecciones y política legislativa de Estados Unidos. Cuenta con datos de apariciones públicas de diversos candidatos. Estas comprenden discursos, entrevistas, conferencias de prensa, grupos de prensa, comentarios y Vlogs. Tienen una estructura de transcripción separada por párrafos u oraciones, asociados a su orador, mientras que todo el discurso se asocia al candidato en cuestión:

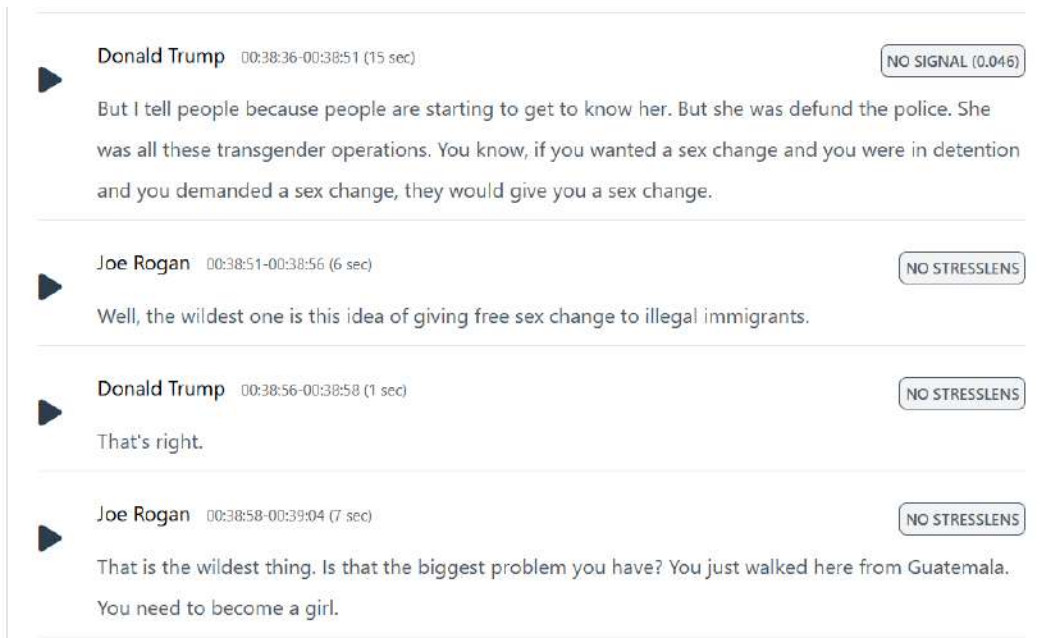


Fig. 2.1: Estructura de un fragmento de entrevista entre Joe Rogan y Donald Trump del 25 de octubre de 2024 (Roll Call, [7]). Este “discurso” se asocia al candidato Donald Trump.

La segunda fuente es RealClearPolling [8], una página con cobertura detallada de encuestas de opinión pública relacionadas con la política. Cuenta con encuestas de intención de voto, imagen pública y seguimiento de gestión (entre otros) para diversas figuras como gobernadores, senadores, legisladores y presidentes, a nivel estatal y nacional. Depende de RealClearPolitics, una organización de periodistas que se encarga de la cobertura extensa de eventos políticos desde el año 2000.

La tercera fuente es 270toWin [9], una plataforma que recolecta resultados electorales históricos de Estados Unidos a nivel estatal. Se autodefine como un sitio no partidista y se dedica a la agregación de encuestas, proyecciones electorales profesionales y contenido noticioso relevante.

2.1.1. Extracción y estandarización de datos

Se construyó una base de datos de discursos a partir del archivo público del sitio web de noticias políticas Roll Call [7]. Para la recolección sistemática de este material, se desarrolló un programa en Python específicamente para esta investigación. Dicho programa fue diseñado para navegar el archivo del sitio y extraer el contenido de interés de los candidatos seleccionados dentro del período temporal definido, guardándolo en un formato estructurado. Para su implementación se utilizaron las librerías *Requests* [10], *BeautifulSoup* [11] y *Selenium* [12].

Se utilizó una estructura estándar para la base de datos de discursos, cuyas variables y descripciones se detallan en la Tabla 2.1.

Tab. 2.1: Estructura de la base de datos de discursos.

Variable	Descripción
ID_único	Identificador numérico único para cada discurso.
candidato	Nombre del candidato que pronunció el discurso.
tipo	Naturaleza del evento (ej. entrevista, conferencia de prensa).
título	Título o encabezado del discurso.
transcripción	Texto completo original del discurso.
fecha	Fecha en que se pronunció el discurso.
ciudad	Ciudad donde se realizó el evento.
estado	Estado o provincia donde se realizó el evento.
tópicos	Temas asociados al discurso, provistos por Roll Call.

En total, la base de datos contiene 1197 discursos de los tres candidatos, abarcando el período de julio de 2023 a noviembre de 2024. Los tópicos, propuestos por Roll Call, están asignados a cada discurso en una estructura jerárquica, donde se dispone de 17 posibles para el primer nivel y 134 para el segundo. Cada discurso puede tener asignado más de un tema por nivel.

Dado que un discurso entero puede tratar diversos tópicos, se decidió trabajar con el párrafo como unidad de texto, ya que normalmente esta unidad se enfoca en un único tema y eso lo hace más manejable en comparación a todo el discurso. Para ello, los discursos fueron separados por párrafos y filtrados para conservar únicamente aquellos con más de 30 caracteres y pronunciados por el candidato asociado al discurso. Este proceso generó una base de datos secundaria, detallada en la Tabla 2.2, que vincula cada párrafo a su discurso original.

Tab. 2.2: Estructura de la base de datos de párrafos.

Variable	Descripción
ID_discurso	ID_único del discurso al que pertenece el párrafo.
transcripción_párrafo	Transcripción del texto del párrafo.

Finalmente, el conjunto de datos de trabajo quedó conformado por 74,239 párrafos, donde cada uno está asociado al discurso donde aparece por su ID_Único.

La segunda fuente de datos usada para este trabajo son las encuestas de opinión pública. Se recolectaron series temporales de intención de voto a nivel nacional y estatal del portal especializado RealClearPolling [8]. Para la obtención de estas series, se desarrolló

un programa en Python que, mediante el análisis de la estructura HTML de las páginas de encuestas, recolectó los datos históricos de intención de voto para los enfrentamientos de interés (Trump-Biden y Trump-Harris).

A partir de esta recolección, se crearon bases de datos para cada estado y para el nivel nacional, indexadas por fecha. Su estructura se detalla en la Tabla 2.3.

Tab. 2.3: Estructura de la base de datos de encuestas diarias.

Variable	Descripción
fecha	Fecha de la medición (índice de la serie temporal).
intencion_voto_trump	Porcentaje de intención de voto para Donald Trump.
intencion_voto_oponente	Porcentaje de intención de voto para el oponente (Biden o Harris).
diferencia	Diferencia porcentual calculada (Trump - Oponente).

Las series temporales reflejan la métrica “RealClearPolitics Poll Average”, un estándar en el análisis político de Estados Unidos que promedia resultados de diversas encuestas de días anteriores a una fecha en particular. Aunque la metodología exacta de este promedio no es pública (en el sentido de que no se conoce la cantidad de días considerados para calcular el promedio para una fecha fija), su uso generalizado lo consolida como un indicador de referencia. Para el enfrentamiento Trump-Biden, se contó con datos hasta el 21 de julio de 2024, fecha del retiro de Biden. Para el caso Trump-Harris, los datos cubren desde el 1 de julio hasta el 5 de noviembre de 2024.

Por último, se obtuvieron resultados electorales para cada estado desde 1996 hasta el 2020 desde el sitio 270toWin [9], estos datos se guardaron en una base de datos por estado como se muestra en la Tabla 2.4. Esto se hizo con el objetivo de entender el contexto electoral y las tendencias de voto en cada estado.

Tab. 2.4: Estructura de la base de datos de resultados electorales históricos para un estado en particular.

Variable	Descripción
año_elección	Año de la elección presidencial (e.g., 1996, 2000, etc.).
resultado_republicano	Porcentaje de votos obtenido por el candidato del partido republicano en ese estado.
resultado_demócrata	Porcentaje de votos obtenido por el candidato del partido demócrata en ese estado.

2.2. Métodos para detección de tópicos

2.2.1. Text Clustering y Topic Modeling

En el ámbito del aprendizaje no supervisado, se trabaja con un conjunto de datos que carece de etiquetas asignadas a priori, disponiéndose únicamente de un conjunto de atributos para cada instancia. El objetivo fundamental consiste en, a partir de dichos atributos, identificar una estructura subyacente que permita explicar el comportamiento inherente a los datos. Un problema característico dentro del aprendizaje no supervisado es el de “*Clustering*” (o agrupamiento), mediante el cual se busca encontrar una segmentación

de los datos en grupos o conglomerados que resulte relevante para su posterior análisis. En este contexto se dispuso de un corpus de texto (un grupo de textos correspondientes a discursos), donde el propósito fue agrupar distintos documentos (párrafos) en clusters (grupos de párrafos) basándose en su contenido semántico, significado y las relaciones inferidas entre ellos. Esta técnica se conoce como “*Text Clustering*” y su objetivo es facilitar la categorización de grandes volúmenes de texto y contribuye a su análisis exploratorio.

Como fue mencionado anteriormente, el uso de párrafos para esta tarea es conveniente ya que la unicidad de tópicos se da en este caso. Un párrafo típicamente hace mención a una idea o tema en particular, mientras que un discurso entero puede hablar de más de un tema. Esto influye en el conjunto de tópicos encontrados en el texto de los discursos; representar un discurso entero con un documento no es una práctica recomendable en el contexto del problema ya que puede generalizar discursos enteros con un sólo tópico y eso no es lo buscado. Por esta razón, se decidió trabajar con los párrafos como documentos.

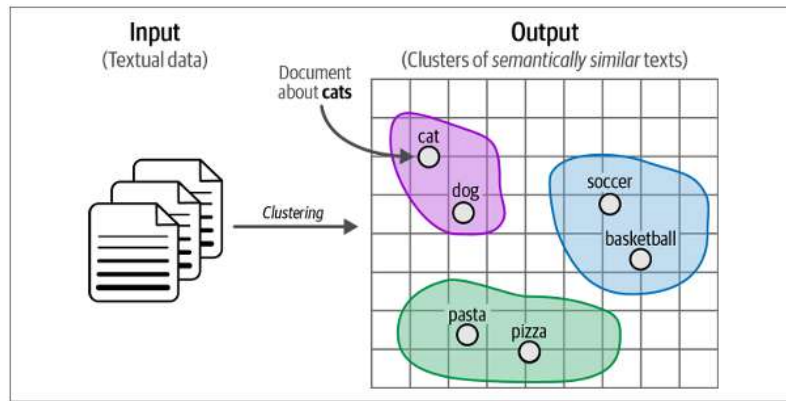


Fig. 2.2: Ilustración conceptual de Text Clustering del libro [13].

Hay interés en la identificación de tópicos abstractos que subyacen en el corpus de texto. Esta tarea se enmarca dentro del área de “*Topic Modeling*” (o modelado de tópicos). El modelado de tópicos busca descubrir las temáticas latentes dentro de una colección de documentos, representando cada tópico como una conjunto de palabras características y asociando cada documento a un tópico en particular. De esta manera, se aspira a que los clusters obtenidos mediante técnicas de agrupamiento puedan ser interpretados o asociados directamente con estos tópicos identificados, donde cada cluster represente un tópico.

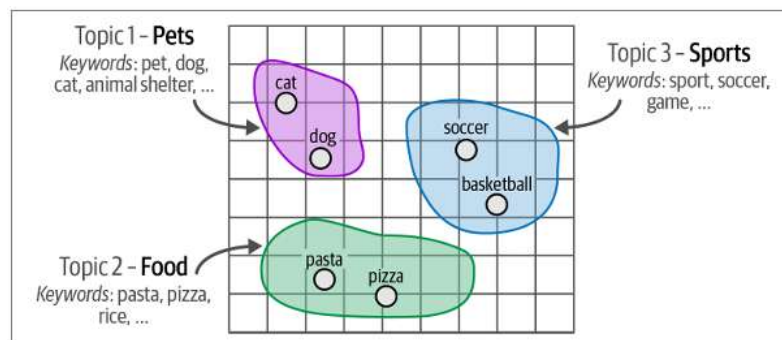


Fig. 2.3: Ilustración conceptual de Topic Modeling del libro [13].

De esta manera, se puede obtener la representación temática de un discurso a través del agrupamiento de sus párrafos correspondientes. Un discurso se puede caracterizar en función de los clusters de tópicos identificados en cada uno de sus párrafos. Esto permite generar para un discurso una distribución en el espacio de tópicos dada por la frecuencia con que un tópico está asociado a sus párrafos.

Mediante la aplicación combinada de técnicas de *Text Clustering* y *Topic Modeling*, el objetivo es identificar tópicos relevantes (es decir, que puedan ser validados) y distintivos (tópicos distinguibles entre sí) para el corpus de texto, derivados de un agrupamiento donde cada cluster idealmente refleje un tópico principal, logrando así una representación efectiva de los discursos.

2.2.2. BERTopic

BERTopic [14, 13] es una técnica que combina elementos de *Text Clustering* y *Topic Modeling*. Su objetivo principal es identificar grupos de documentos (clusters) dentro de un corpus y, a partir de ellos, extraer un conjunto de tópicos.

Para lograr esto, BERTopic aprovecha la arquitectura de los *Transformers*, como BERT (Bidirectional Encoder Representations from Transformers, [15]), para generar representaciones vectoriales (embeddings) para cada documento. Estos embeddings capturan el significado contextual de los textos, permitiendo que palabras con significados distintos en contextos diferentes (polisemia) sean representadas adecuadamente [15].

Una vez que los documentos se transforman en estos vectores densos, BERTopic procede a agruparlos. Una ventaja clave de esta técnica es su flexibilidad en cuanto a la elección de algoritmos en cada paso. Aunque comúnmente se utiliza el algoritmo de reducción de dimensionalidad UMAP y el de clustering HDBSCAN, estos componentes son intercambiables. Esto permite al usuario seleccionar otros algoritmos que se ajusten mejor a las características específicas del problema o de los datos.

Una vez formados los clusters, el siguiente paso es darles una interpretación semántica. Para ello, BERTopic emplea una variante de TF-IDF llamada c-TF-IDF (class-based TF-IDF). El TF-IDF (Term Frequency-Inverse Document Frequency) es una métrica clásica que asigna un peso a cada palabra en un documento, valorando tanto su frecuencia dentro de ese texto como su rareza en el corpus completo [16]. La innovación de c-TF-IDF consiste en tratar a todos los párrafos de un mismo cluster como si fueran un único documento. Luego, aplica la lógica de TF-IDF a nivel de clusters, calcula la importancia de una palabra para ese cluster específico en comparación con su importancia en todos los demás clusters. De este modo, las palabras que son muy frecuentes dentro de un cluster pero raras en el resto, son seleccionadas como las más representativas. Este conjunto de palabras clave constituye la descripción del tópico asociado a dicho cluster.

La estructura de BERTopic es inherentemente modular, permitiendo que cada paso del proceso sea tratado de forma independiente. En la Figura 2.4 se visualiza esta secuencia, que va desde la generación de embeddings hasta la creación de los tópicos.

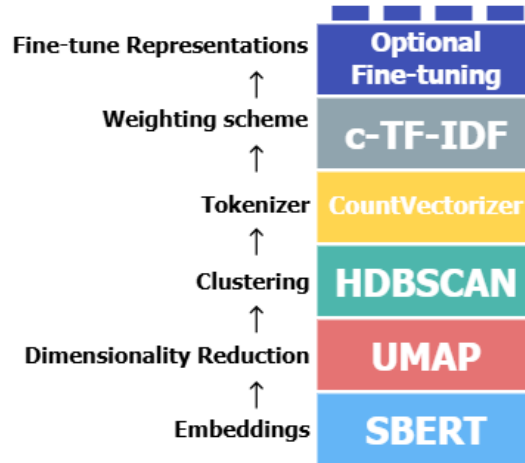


Fig. 2.4: Visualización del estructura modular de BERTopic. Adaptado de [14].

Esta modularidad ofrece una gran flexibilidad, ya que permite al usuario sustituir o ajustar los algoritmos específicos utilizados en cada etapa (por ejemplo, elegir un modelo de embeddings diferente o un algoritmo de clustering distinto) para adaptar la técnica a las particularidades del conjunto de datos o a los objetivos específicos del análisis.

La elección de este algoritmo se basa en estudios previos que lo comparan con otras técnicas de *Topic Modeling* como la Asignación Latente de Dirichlet (LDA) [17] y la Factorización de Matrices no Negativas (NMF) [18]. La evidencia empírica demuestra que BERTopic obtiene resultados superiores, especialmente en el análisis de textos cortos, donde los métodos tradicionales, basados en la frecuencia de palabras, tienen dificultades para encontrar patrones robustos.

Por ejemplo, en [19] realizaron una evaluación exhaustiva sobre titulares de noticias y concluyeron que BERTopic genera tópicos más coherentes y representativos que LDA y NMF en este tipo de corpus. De manera similar, un análisis comparativo sobre reseñas de clientes confirmó la robustez y el poder de BERTopic para extraer temas significativos donde otros modelos son menos efectivos [20].

Esta ventaja es particularmente relevante para la presente investigación, ya que justifica no solo la elección de BERTopic sobre otras alternativas, sino también la decisión metodológica de segmentar los discursos en párrafos. Al trabajar con estas unidades textuales más breves y temáticamente homogéneas, se aprovecha al máximo la capacidad del modelo para capturar el significado semántico contextual, en lugar de depender únicamente de la estadística de palabras en documentos largos, como así se hace con NMF, por ejemplo.

A continuación se explican brevemente los módulos correspondientes al algoritmo de BERTopic, adaptado al problema en cuestión.

1. Generación de embeddings: El primer paso consistió en vectorizar los párrafos. La elección del modelo de “embeddings”, `all-MiniLM-L6-v2`, se fundamentó en un balance entre eficiencia y rendimiento. Con un tamaño de aproximadamente 80MB, presenta uno de los mejores desempeños para tareas de similitud semántica dentro de su categoría en las evaluaciones comparativas de la librería Sentence-Transformers [21, 22]. Este modelo transforma cada párrafo en un “embedding” de 384 dimensiones, es decir, un vector en

$\mathbb{R}^{384}$ que representa el significado de un párrafo. La intuición de los “embeddings” es que dos párrafos que representen ideas similares tienen que tener una distancia o medida de similitud parecida en el espacio vectorial.

2. Reducción de dimensionalidad y clustering: Para la reducción de dimensionalidad, se empleó UMAP [23] para proyectar los embeddings de 384 a 5 dimensiones. Se mantuvo este algoritmo por ser el método por defecto en la implementación de BERTopic, lo que asegura una integración probada y robusta. Posteriormente, para la tarea de clustering, se optó por HDBSCAN [24]. A diferencia de algoritmos como K-Means, esta elección ofrece dos ventajas cruciales para este trabajo: permite identificar clusters de densidades heterogéneas y, fundamentalmente, no requiere definir la cantidad de clusters a priori.

3. Vectorización y representación de tópicos: La representación de los tópicos se basa en el módulo c-TF-IDF, un componente central de BERTopic. Este módulo utiliza internamente un vectorizador de conteo (basado en la implementación de Scikit-learn [25]) para crear una matriz de frecuencias de palabras para cada cluster. Sobre esta matriz, se aplica la ponderación BM25 [26] para calcular la relevancia de cada palabra dentro de su tópico.

4. Refinamiento de la representación de tópicos: Finalmente, para mejorar la calidad e interpretabilidad de los tópicos, se utilizaron dos módulos de refinamiento integrados en BERTopic.

Primero, el módulo *KeyBERTInspired* [27] reordena las palabras clave de cada tópico para priorizar aquellas que son semánticamente más similares al contenido general del cluster. Esto se hace comparando el vector promedio de embeddings de documentos en un cluster y evaluando la similitud coseno de ese vector contra todas las palabras encontradas por el vectorizador. Luego, se obtiene un ranking en base a las palabras más similares a ese promedio, y por ende las más relevantes.

Segundo, se empleó *Maximal Marginal Relevance* (MMR) [28] para diversificar la lista de palabras clave y reducir la redundancia. Por ejemplo, si un tópico sobre elecciones contiene las palabras “elección”, “electoral” y “comicios”, una alta redundancia podría llevar a que la siguiente palabra sea “votación”. MMR, en cambio, puede promover la inclusión de un término distinto pero relevante como “candidato” o “urna”, enriqueciendo la descripción del tópico.

Más adelante, en la sección 3.3 se detallan los hiperparámetros usados en la implementación de BERTopic.

2.3. Medidas de correlación

En el contexto del problema, las asociaciones entre las series temporales de tópicos y las de encuestas se analizan mediante dos coeficientes de correlación complementarios: el de Pearson [29] y el de Spearman [30]. La elección de ambos permite, por un lado, detectar relaciones lineales estrictas (Pearson) y, por otro, capturar tendencias monótonas más generales que no son necesariamente lineales (Spearman).

2.3.1. Correlación de Pearson

El coeficiente de correlación de Pearson (r) es una medida paramétrica que cuantifica la relación lineal entre dos variables continuas, X e Y . Evalúa la fuerza y dirección de una asociación lineal, asumiendo que las variables están distribuidas normalmente o que, al menos, la relación entre ellas es lineal. Se calcula como la covarianza de las dos variables dividida por el producto de sus desviaciones estándar:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

donde n es el número de observaciones, x_i e y_i son los valores individuales de las variables X e Y respectivamente, y $\bar{x}$ y $\bar{y}$ son las medias muestrales de X e Y . El valor de r varía entre -1 y $+1$, donde $+1$ indica una correlación lineal positiva perfecta, -1 una correlación lineal negativa perfecta, y 0 la ausencia de correlación lineal.

2.3.2. Correlación de Spearman

El coeficiente de correlación de Spearman (ρ o r_s) es una medida no paramétrica de la correlación por rangos. Evalúa qué tan bien puede describirse la relación entre dos variables mediante una función monótona. Una relación es monótona si a medida que una variable aumenta, la otra tiende consistentemente a aumentar o a disminuir, aunque no necesariamente a un ritmo constante, como pasaría si dos variables estuvieran relacionadas linealmente. Para calcularlo, primero se convierten los valores de cada variable a rangos, y luego se aplica la fórmula del coeficiente de correlación de Pearson a estos rangos.

Existen “empates” (ties) cuando múltiples observaciones en una misma variable tienen exactamente el mismo valor, lo que resulta en que comparten el mismo rango; por ejemplo, si en tres días distintos la intención de voto para un candidato fue del 45,2%, a esas tres observaciones se les asignaría el mismo rango. Si no existen empates, la fórmula simplificada es:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

donde $d_i = \text{rg}(x_i) - \text{rg}(y_i)$ es la diferencia entre los rangos de cada par de observaciones (x_i, y_i) . Cuando existen empates, es más preciso utilizar la fórmula de Pearson sobre los rangos asignados.

La inclusión de la correlación de Spearman es fundamental, ya que la relación entre la frecuencia de mención de un tópico en los discursos y la intención de voto no tiene por qué ser estrictamente lineal. Por ejemplo, es posible que un aumento inicial en la discusión sobre un tema tenga un gran impacto en las encuestas, pero que aumentos posteriores tengan un efecto decreciente. Spearman puede capturar estas tendencias monótonas donde Pearson, limitado a la linealidad, podría subestimar o no detectar la asociación. Al igual que el de Pearson, su coeficiente varía entre -1 y $+1$, pero en este caso los valores extremos indican una relación monótona perfecta (positiva o negativa) y un valor de 0 indica que no hay relación entre las variables.

3. ANÁLISIS DE DISCURSOS

3.1. Representaciones de discursos en el espacio de tópicos y comparación entre ellas

Una vez identificado un conjunto de tópicos, el siguiente paso fue cuantificar la presencia de cada tópico en cada discurso. Para ello, se representó cada discurso como un vector de distribución de tópicos, donde cada componente del vector indica el “peso” de un tópico específico dentro de ese discurso.

Este peso se calculó como la frecuencia relativa de los párrafos de un discurso que pertenecen a un tópico determinado. Formalmente, el peso D_i del tópico i en un discurso D se define mediante la Ecuación 3.1:

$$D_i = \frac{\sum_{\text{par} \in \text{párrafos}(D)} \mathbf{1}(\text{par} \in \text{tópico}_i)}{\#(\text{párrafos}(D))} \quad (3.1)$$

Donde $\mathbf{1}(\cdot)$ es la función indicadora, que vale 1 si el párrafo “par” pertenece al tópico i y 0 en caso contrario, y $\#(\text{párrafos}(D))$ es el número total de párrafos en el discurso D . Por ejemplo, si un discurso tiene 20 párrafos en total y 5 de ellos fueron asignados al tópico de “Economía”, el peso de ese tópico para ese discurso sería $5/20 = 0,25$. El conjunto de estos pesos para todos los tópicos forma el vector de distribución del discurso.

En este trabajo se consideran y comparan dos representaciones distintas generadas a partir de este principio: una basada en las etiquetas preexistentes de Roll Call y otra generada con el modelo BERTopic. Si bien ambas respetan la naturaleza multi-etiqueta de los discursos, sus metodologías y características difieren significativamente. Para facilitar la comparación, sus principales atributos se resumen en la Tabla 3.1.

Tab. 3.1: Comparación entre las representaciones de tópicos de Roll Call y BERTopic.

Característica	Representación de Roll Call	Representación de BERTopic
Origen	Etiquetas provistas externamente por el publicador de los datos.	Generadas específicamente para este trabajo a partir del contenido de los discursos.
Metodología	Proceso de asignación desconocido.	Algoritmo transparente y reproducible (BERTopic) basado en similitud semántica.
Unidad de asignación	Asignado al discurso en su totalidad.	Asignado a nivel de párrafo, permitiendo mayor granularidad.
Esquema de pesos	Binario. Un tópico está presente o ausente. Si hay k tópicos, cada uno tiene un peso de $1/k$.	Proporcional. El peso de un tópico es la frecuencia relativa de sus párrafos en el discurso.
Flexibilidad	Fija y no modificable.	Ajustable mediante la configuración de hiperparámetros de BERTopic.

Se desconoce la metodología exacta detrás de la asignación de etiquetas provista por Roll Call. El conjunto completo de los 17 tópicos del primer nivel, junto a su traducción al español, se detalla en la Tabla A.1 del Apéndice. Para esta metodología, se decidió trabajar únicamente con el primer nivel de jerarquía de tópicos, caracterizado por los 17 mencionados. Esta decisión se debe a que clasificar los discursos con el segundo nivel (134 tópicos) no proporciona una cantidad de discursos lo suficientemente alta para el análisis, potencialmente generando series temporales sumamente esparsas (con pocos datos).

La diferencia principal entre las distribuciones de Roll Call y BERTopic se debe a que la distribución de Roll Call asigna igual peso a todos los tópicos presentes en un discurso, mientras que las distribuciones de BERTopic se basan en la frecuencia relativa de los tópicos obtenida a partir del análisis de los párrafos.

A pesar de que la representación de BERTopic ofrece mayor granularidad y transparencia, se decidió analizar ambas distribuciones para evaluar si las conclusiones obtenidas son similares para ambas metodologías.

3.2. Armado de series temporales de tópicos

Fabricar distribuciones en el espacio de tópicos para cada discurso fue el paso previo y necesario para generar series temporales que permitieran capturar la relevancia temporal y espacial de cada tópico.

Como fue mencionado en la Sección 2.1.1, los discursos cuentan con una fecha y ubicación asociados, lo que permite explotarlos geográfica y temporalmente. Con este propósito se construyeron series temporales para cada tópico a nivel nacional y estatal a modo de pre-procesamiento.

$$T_i(t) = \frac{1}{S_i} \sum_{D \in \text{discursos}(t)} D_i \quad (3.2a)$$

$$T_i(t, x) = \frac{1}{S_i(x)} \sum_{D \in \text{discursos}(t, x)} D_i \quad (3.2b)$$

La Ecuación (3.2a) corresponde a la serie de pesos del tópico i al tiempo t , específicamente al día t . La Ecuación (3.2b) corresponde a la serie de pesos del tópico i al día t en el estado x . Ambas series usan la distribución obtenida en la Ecuación (3.1). Las constantes S_i y $S_i(x)$ normalizan ambas series para que sus valores sumen 1 a través del tiempo t , la primera es a nivel nacional y la segunda estatal, por eso depende del estado x . Adicionalmente, los días sin discursos tienen un valor default nulo (0) y se aplicó un suavizado con una media móvil de ventana semanal (7 días, $t \pm 3$).

3.3. Descomposición en tópicos: resultados y validación

Para encontrar un espacio de tópicos de forma no supervisada, se utilizó BERTopic. La configuración final del modelo, resultante del proceso de búsqueda de hiperparámetros que se describe a continuación, se detalla en la Tabla 3.2.

Tab. 3.2: Configuración final del modelo BERTopic.

Componente	Hiperparámetro	Valor
SentenceTransformer	modelo	all-MiniLM-L6-v2
	dimensiones	384
UMAP	n_components	5
HDBSCAN	min_cluster_size	300
	cluster_selection_method	“leaf”
CountVectorizer	max_features	30000
	min_df	0,001
	max_df	0,9
	stop_words	“english”
	n-gram_range	(1, 3)
ClassTfidfTransformer	ponderación_bm25	True
	reduce_frequent_words	True
KeyBERTInspired	habilitado	True
MMR	diversity	0,7

La elección de estos hiperparámetros no fue arbitraria, sino el resultado de una búsqueda manual con el objetivo principal de obtener un número de tópicos del orden de los provistos por Roll Call (~ 17). Esto se hizo con la idea de generar un espacio de tópicos que no fuera ni demasiado granular (una cantidad muy alta de tópicos muy específicos) ni demasiado general (una cantidad baja de tópicos que agrupen temas dispares).

La búsqueda se centró en los dos hiperparámetros de HDBSCAN que más influyen en la estructura y cantidad de los clusters. Los valores explorados en esta búsqueda se muestran en la Tabla 3.3.

Tab. 3.3: Espacio de búsqueda de hiperparámetros de HDBSCAN.

Hiperparámetro	Valores Explorados
cluster_selection_method	{“leaf”, “eom”}
min_cluster_size (para “leaf”)	{200, 220, 240, 260, 280, 300, 320, 340}
min_cluster_size (para “eom”)	{5, 10, 15, 20, 25, 30, 35, 40}

El hiperparámetro “min_cluster_size” define la cantidad mínima de párrafos que un grupo de párrafos debe contener para ser considerado un cluster. Por su parte, “cluster_selection_method” (método de selección de clusters) es crucial porque los algoritmos de clustering jerárquico como HDBSCAN no generan una única partición, sino que construyen una jerarquía de clusters anidados, similar a un árbol. Este método determina cómo se seleccionan los clusters finales a partir de esa estructura jerárquica. Siguiendo la documentación oficial de HDBSCAN [31], se exploraron dos estrategias: “eom” (“Excess of Mass”), que tiende a favorecer clusters grandes y estables dentro de la jerarquía, y “leaf”, que selecciona los nodos hoja del árbol, lo que tiende a producir un mayor número de clusters más pequeños y homogéneos. Se exploraron rangos distintos de `min_cluster_size` para cada estrategia, ya que “leaf” suele requerir un valor más alto para no generar una cantidad excesiva de clusters con pocos párrafos.

3.3.1. Resultados

El proceso de modelado con BERTopic, utilizando la configuración de hiperparámetros descrita, resultó en la identificación de 20 tópicos dentro del corpus de discursos. La Figura 3.1 presenta una visualización de cada uno de estos tópicos mediante nubes de palabras, donde el tamaño de cada término es proporcional a su relevancia según la métrica c-TF-IDF (por ende la relevancia del término dentro del tópico). Para mayor claridad, en la Tabla A.2 del Apéndice se proporciona una lista completa de estos tópicos.



Fig. 3.1: Visualización de los 20 tópicos identificados por *BERTopic*.

Un análisis general de los tópicos revelados en la Figura 3.1 permite agruparlos en varias macro-categorías temáticas que son típicos de una campaña presidencial. A continuación se discute una posible caracterización de los tópicos encontrados, junto a una interpretación de un conjunto de nubes de palabras.

- **Economía y política fiscal:** Tópicos como “Política fiscal / impuestos”, “Comercio y acuerdos económicos”, “Industria, manufactura, tecnología y trabajo” y “Petróleo y energía” están relacionadas al sector productivo y a la economía.
- **Seguridad y política migratoria:** Un eje central de la campaña, reflejado en tópicos como “Métricas de seguridad fronteriza”, “Crítica a la política fronteriza”,

“Inmigración y crimen”, “Fuerzas armadas y seguridad fronteriza” y “Seguridad y orden público”.

- **Asuntos sociales y de salud:** Se identifican temas de gran relevancia social como “Aborto”, “Política educativa”, “Sanidad y programas sociales”, “Asuntos de salud de veteranos” y “Género y deportes”, cubriendo debates desde la política educativa hasta la sanitaria.
- **Política exterior y desastres:** Incluye discusiones sobre “Conflictos internacionales” y la respuesta a eventos como desastres naturales como huracanes y tormentas en “Respuesta a desastres”.
- **Retórica y estrategia política:** Es de particular interés la capacidad del modelo para capturar no solo temas concretos, sino también estilos discursivos. Tópicos como “Retórica patriótica”, “Crítica a ideologías políticas” y “Corrupción legal y política” reflejan más una forma de hablar o un tipo de ataque político (como las acusaciones de corrupción) que un tema de política pública específico. Por otro lado, el tópico “Líderes políticos y comunitarios” logra capturar entidades políticas asociadas a cada partido.

Al observar las nubes de palabras de la Figura 3.1 se puede ver que los tópicos están bien distinguidos entre sí y se pueden hacer diversas interpretaciones. Ciertas nubes de palabras, como “Seguridad y orden público” están claramente influenciadas por eventos o discusiones específicas como el intento de asesinato a Trump ocurrido el 13 de julio de 2024. Los términos “bullet” (bala), “secret service” (servicio secreto), “ear” (oreja) y “blood” (sangre) son característicos de este suceso.

Otro ejemplo de esto ocurre en la nube de palabras “Género y deportes”, donde los términos “transitioned” (del verbo transicionar) y “women sports” (deportes femeninos) dan un indicio de la discusión recurrente impulsada por Trump de atletas hombres que cambian de género y compiten en deportes de mujeres.

La nube de palabras titulada “Respuesta a desastres” hace referencia a los huracanes (ver término “hurricane”) ocurridos en el sureste del país, particularmente el huracán Helena (“Helene”) ocurrido el 26 de septiembre de 2024, que afectó mayormente a los estados de Carolina del Norte, Florida y Georgia. La presencia de términos como “FEMA” (Agencia Federal para el Manejo de Asistencias) o “Starlink” (servicio de internet ofrecido por la empresa *SpaceX*) pueden hacer alusión a la asistencia ofrecida luego de los desastres naturales [32, 33].

Otros temas de discusión distinguibles en las nubes de palabras son los tópicos “Aborto”, con términos como “abortion” (aborto) o “Roe Wade” (histórico fallo de 1973 que garantizó libertades reproductivas a las mujeres [34]); “Política educativa” con términos como “students” (estudiantes), “colleges” (universidades) o “student debt” (deuda estudiantil) y “Conflictos internacionales” con términos relacionados a la guerra en Ucrania (“Ukraine” y “Putin”) el conflicto en el medio oriente entre Israel y Palestina (“hamas”, “hostages”, “Iran”, “jewish”).

Por otro lado, hay nubes de palabras que no muestran un tópico identificable como los mencionados anteriormente. Este es el caso de tópicos como “Retórica patriótica” o “Líderes políticos y comunitarios”. El primer tópico tiene términos relacionados a eslóganes de campaña como “Make America Great Again” o “God bless America” y términos relacionados al apelo al voto como “ballots” (boleta) o “ID” (documento de identidad); estos

términos son genéricos y no parecen presentar una discusión puntual. El segundo tópico está compuesto de figuras políticas diversas, pero no se asocia fácilmente a un suceso en particular.

3.3.2. Validación de tópicos

Para verificar y comparar la validez de los tópicos de Roll Call y BERTopic, se realizó un proceso de validación manual. El objetivo de este proceso fue verificar que las asignaciones de tópicos bajo ambos métodos reflejen el contenido de los discursos y comparar ambas representaciones de tópicos. Este proceso fue llevado a cabo por el autor de esta tesis y consistió en etiquetar una muestra de discursos para luego comparar estas etiquetas “manuales” con las asignaciones originales de cada método, seguido de una comparación entre ambas asignaciones de tópicos originales.

Para asegurar una representación equilibrada, se construyó una muestra de 99 discursos seleccionados aleatoriamente bajo la condición de que cada uno de los 17 tópicos de Roll Call estuviera presente en al menos seis discursos. La distribución final de los tópicos de Roll Call en esta muestra se puede observar en la Figura 3.2.

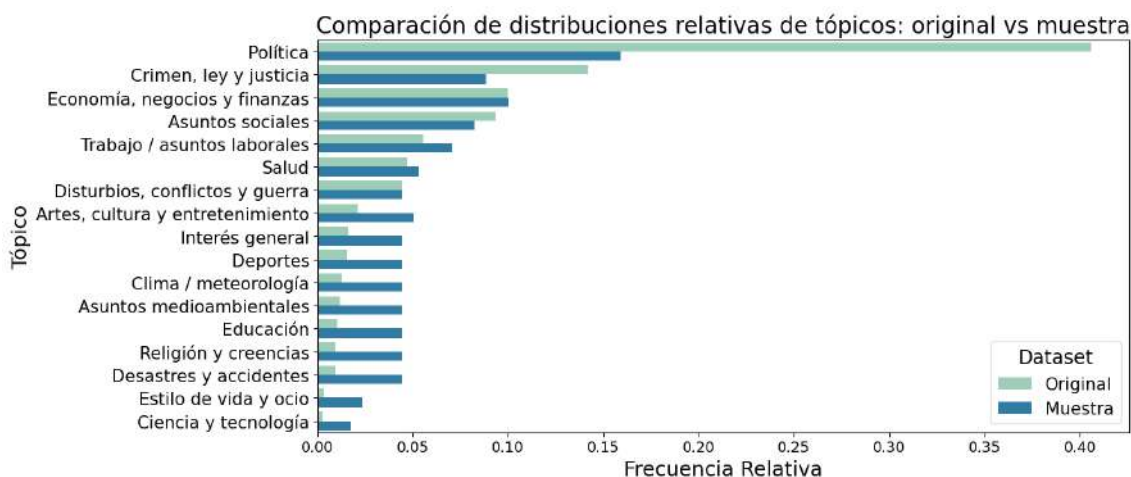


Fig. 3.2: Distribución de tópicos Roll Call en muestra de 99 discursos.

En la Figura 3.2 las barras verdes representan la proporción de discursos de cierto tópico con respecto a la cantidad total de discursos disponibles y las barras azules las mismas proporciones pero con respecto a la muestra de 99 discursos. La distribución de tópicos en la muestra se acerca más a una distribución uniforme que la distribución original de tópicos.

Sobre esta muestra, se diseñaron tres experimentos de validación:

1. **Consistencia interna de Roll Call:** Se compararon las etiquetas originales de Roll Call con un nuevo etiquetado manual realizado por el autor, pero utilizando el mismo conjunto de 17 etiquetas de Roll Call. El objetivo fue medir la consistencia entre las asignaciones de tópicos de este sistema de etiquetas.
2. **Consistencia interna de BERTopic:** De forma análoga, se compararon las etiquetas generadas por BERTopic con un etiquetado manual realizado por el autor, utilizando el conjunto de los 20 tópicos de BERTopic.

3. **Acuerdo entre metodologías (Roll Call vs. BERTopic):** Se compararon las etiquetas originales de Roll Call con las etiquetas asignadas por BERTopic. Para hacerlas comparables, se utilizó un mapeo manual (detallado en la Tabla 3.4) que traduce cada tópico de BERTopic a uno o más tópicos equivalentes de Roll Call. La construcción de dicho mapeo fue una decisión cualitativa de diseño por parte del autor en base a la lectura de la muestra de discursos. Es importante mencionar esto último ya que esto tiene un impacto en la precisión de asociación entre tópicos: una co-ocurrencia alta entre las asignaciones en este caso puede estar influenciada por el propio diseño del mapeo.

Tab. 3.4: Mapeo de tópicos de BERTopic a categorías de Roll Call.

Tópico de BERTopic	Categorías de Roll Call
Aborto	Asuntos sociales, Salud
Crítica a la política fronteriza	Política
Métricas de seguridad fronteriza	Crimen, ley y justicia
Respuesta a desastres	Desastres y accidentes
Política educativa	Educación
Género y deportes	Estilo de vida y ocio, Deportes
Sanidad y programas sociales	Salud
Inmigración y crimen	Crimen, ley y justicia
Industria, manufactura, tecnología y trabajo	Trabajo / asuntos laborales, Ciencia y tecnología, Clima / meteorología, Asuntos medioambientales
Conflictos internacionales	Disturbios, conflictos y guerra
Corrupción legal y política	Crimen, ley y justicia
Fuerzas armadas y seguridad fronteriza	Crimen, ley y justicia, Seguridad y orden público
Petróleo y energía	Economía, negocios y finanzas, Asuntos medioambientales
Retórica patriótica	Interés general, Religión y creencias, Política
Crítica a ideologías políticas	Política
Líderes políticos y comunitarios	Política, Crimen, ley y justicia, Artes, cultura y entretenimiento
Seguridad y orden público	Crimen, ley y justicia
Política fiscal / impuestos	Economía, negocios y finanzas
Comercio y acuerdos económicos	Economía, negocios y finanzas, Trabajo / asuntos laborales
Asuntos de salud de veteranos	Salud, Asuntos sociales

Para cada uno de los tres experimentos de validación, se construyó una matriz de confusión global M . El proceso para generarla comienza a nivel de discurso individual. Para cada discurso d , se generó una matriz de co-ocurrencia M^d , donde las filas representan los tópicos de la metodología base (ej. Roll Call original) y las columnas los de la asignación manual (si bien la asignación de BERTopic no es manual en el tercer caso, el mapeo

entre etiquetas de BERTopic y Roll Call lo es, por lo que es adecuado llamarlo asignación manual).

La matriz M^d para un único discurso se rellenó siguiendo un proceso de dos etapas diseñado para capturar primero los acuerdos y luego los desacuerdos:

1. **Etapla 1: identificación de acuerdos (lógica de intersección).** Primero, se identificaron todos los tópicos que aparecían en ambas listas (la base y la manual), sin importar su posición. Por cada tópico i en esta intersección, se sumó 1 a la celda diagonal correspondiente, M_{ii}^d . Este paso se hizo para capturar todos los acuerdos directos entre tópicos.
2. **Etapla 2: comparación posicional de los tópicos restantes.** Una vez registrados los acuerdos, estos se eliminaron de ambas listas, creando dos nuevas listas “residuales” con los tópicos restantes, conservando su orden original. Ambas listas de tópicos fueron ordenadas por relevancia decreciente. Dicha relevancia se define como el orden de aparición de etiquetas para la categoría base (determinado por el orden provisto por Roll Call o la frecuencia relativa ordenada decrecientemente para la asignación de BERTopic) y como el orden de etiquetas para los casos donde hubo una asignación manual de tópicos.
 - Se realizó una comparación posicional sobre estas listas residuales. Para cada posición k de las listas de tópicos, si el tópico base restante era i y el tópico manual restante j , se sumó 1 a la celda M_{ij}^d .
 - Si una de las listas residuales se agotó mientras la otra todavía tenía tópicos, los tópicos sobrantes se asignaban a la categoría “no_match”. Si el tópico sobrante j estaba en la lista manual, se sumó 1 a $M_{(\text{no_match},j)}^d$; si el tópico sobrante i estaba en la lista base, se sumó 1 a $M_{(i,\text{no_match})}^d$.
 - Las posiciones restantes quedaron en 0.

Por ejemplo, si para un discurso la lista base es “[Economía, Seguridad]” y la lista manual “[Economía, Salud, Social]”:

- **Etapla 1:** “Economía” está en ambas listas. Se suma 1 a la celda $M_{(\text{Economía}, \text{Economía})}^d$.
- **Etapla 2:** Las listas residuales son “[Seguridad]” y “[Salud, Social]”. Al comparar posición a posición, “Seguridad” (base) se alinea con “Salud” (manual). Se suma 1 a la celda $M_{(\text{Seguridad}, \text{Salud})}^d$. La lista manual aún tiene el tópico “Social”, que no tiene contraparte. Por lo tanto, se suma 1 a la celda $M_{(\text{no_match}, \text{Social})}^d$.

Finalmente, la matriz de confusión global M para cada experimento se obtuvo sumando elemento a elemento las matrices de todos los 99 discursos de la muestra: $M = \sum_{d \in \text{muestra}} M^d$. Las matrices M resultantes para cada uno de los tres experimentos de validación se presentan de forma visual en la Figura A.1 del Apéndice.

En base a las matrices de confusión (o indistintamente llamadas de co-ocurrencia) se calcularon tres métricas para poder entender el resultado de la validación de tópicos. Primero se calculó la exactitud de clasificación de la muestra de discursos en tópicos como la proporción de casos donde un tópico aparecía en ambas clasificaciones (base y manual), esto consiste en contar la cantidad de casos en la diagonal de la matriz, sobre la cantidad de casos totales de la matriz (sin considerar las fila y columna correspondiente a la categoría

“no_match”). Segundo, se calculó la proporción de casos donde se asignaron más tópicos en la categoría manual que en la base, sobre la cantidad de casos totales de la matriz. Por último, se hizo lo mismo que en el segundo caso pero de modo inverso, se contó la proporción de casos donde había más tópicos en la categoría base que en la manual con respecto a todos los casos de la matriz.

La exactitud sobre tópicos mide qué tanto coinciden ambas asignaciones desde el punto de vista de intersección de tópicos, mientras que las otras dos métricas miden cuánto difieren. En la Tabla 3.5 se pueden observar los valores obtenidos para cada tipo de validación.

Tab. 3.5: Resumen de métricas de validación de etiquetado.

Tipo de validación	Exactitud sobre tópicos (proporción)	Proporción “no_match” (fila)	Proporción “no_match” (columna)
Roll Call vs. Roll Call	0,86	0,09	0,19
BERTopic vs. BERTopic	0,88	0,01	0,42
Roll Call vs. BERTopic	0,51	0,11	0,38

La exactitud es alta para las primeras dos validaciones (0,86 para Roll Call contra Roll Call y 0,88 para BERTopic contra BERTopic), indicando que hay una alta coincidencia entre la clasificación en tópicos por parte del autor con la clasificación externa en el caso de Roll Call y con el algoritmo de BERTopic. Una interpretación de estos resultados sugiere que la construcción de series temporales usando ambos espacios de tópicos es adecuada.

Por otro lado, la exactitud del 51 % en la última validación muestra que las asignaciones no coinciden o separan los discursos en espacios de tópicos distintos. Si bien la cantidad de tópicos de BERTopic y de Roll Call son similares (20 contra 17 respectivamente), la clasificación de párrafos en tópicos de BERTopic es más refinada que la de Roll Call, capturando sucesos o estilos de lenguaje mientras que la clasificación de Roll Call es más genérica.

Este resultado muestra que no es posible elegir un sistema como “mejor” que el otro de forma absoluta; simplemente son diferentes. Esto valida la decisión de realizar el análisis de correlación con los dos espacios de tópicos, sin desestimar ninguno: una vez utilizando las series temporales generadas a partir de los tópicos de Roll Call, y otra vez con las series generadas por BERTopic.

Viendo únicamente la primera validación (Roll Call contra Roll Call), se puede observar que hay menos del 10 % de casos donde la asignación manual superó en cantidad de tópicos a la asignación base, mientras que un valor de 19 % para la categoría de “no_match” por columna sugiere que hay más casos donde la asignación base clasificó con más tópicos que la manual. El mismo comportamiento sucede en las otras dos validaciones, indicando que la asignación manual fue más conservadora, encontrando menos tópicos por discurso que en la asignación base. El caso más pronunciado de esto ocurre en “BERTopic contra BERTopic” (1 % para el caso “no_match” por fila contra un 42 % por columna), sugiriendo que BERTopic encuentra un número mayor de tópicos (posiblemente menos relevantes al texto pero presentes al fin) que una asignación hecha por un humano.

En resumen, la Tabla 3.5 muestra resultados interesantes: tanto Roll Call como BERTopic capturan más tópicos que la asignación hecha por un humano, la exactitud es alta para ambas representaciones y ambos espacios de representación de tópicos tienen una

estructura diferente, ambas útiles para el armado de series temporales.

4. COMPARACIÓN DISCURSOS - ENCUESTAS

En esta sección se muestra un análisis espacial y temporal de los tópicos para tener una comprensión del escenario electoral. Posteriormente, se reportan los resultados de las comparaciones de estos datos con las encuestas a nivel nacional y estatal.

4.1. Representación espacial de discursos por frecuencia y por tópicos

4.1.1. Identificación de estados competitivos (“swing states”)

En el sistema electoral estadounidense, no todos los estados son igualmente disputados. La mayoría se consideran “estados seguros” (safe states), ya que votan consistentemente por el mismo partido en cada elección presidencial. En contraste, existe un pequeño grupo de estados donde el resultado es incierto y el margen de victoria suele ser muy estrecho. Estos son los llamados “swing states” (o “battleground states”), y son cruciales porque su resultado puede decidir la elección.

Aunque no existe una definición formal y única de “swing state”, un criterio comúnmente aceptado es identificar aquellos estados con un margen de victoria muy pequeño en la elección presidencial anterior. Por ejemplo, si se considera una diferencia porcentual de votos menor al 3% en las elecciones de 2020, se obtiene el conjunto de estados que fueron considerados clave para la elección de 2024, como se muestra en la Figura 4.1: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania y Wisconsin.



Fig. 4.1: Mapa de Estados Unidos con los *swing states* resaltados para la elección de 2024.

4.1.2. Análisis de la estrategia de campaña por estado

Para cuantificar la competitividad histórica de los estados, se creó una métrica denominada “swing score”. Para cada estado, se analizaron los resultados electorales históricos desde 1996 hasta 2020, obtenidos del archivo de 270toWin [9]. El “swing score” se calcula como el promedio de la diferencia porcentual entre los votos republicanos y demócratas en un conjunto de elecciones:

$$\text{swing score} = \frac{1}{\# \text{elecciones}} \sum_{e \in \text{elecciones}} (\text{rep \%}(e) - \text{dem \%}(e))$$

Esta métrica varía entre -1 (estado consistentemente demócrata) y $+1$ (estado consistentemente republicano), por lo que los estados más competitivos son aquellos cuyo “swing score” es más cercano a 0.

El objetivo de este análisis es verificar una hipótesis fundamental de la estrategia de campaña: que los candidatos concentran sus esfuerzos y apariciones públicas (discursos) en los estados más competitivos. Para ello, se relacionó la cantidad de discursos registrados en cada estado (entre julio de 2023 y noviembre de 2024) con su “swing score”. Se excluyó el Distrito de Columbia (Washington DC) de este análisis, ya que la alta concentración de discursos ahí se debe a su función como capital del país y no a una estrategia electoral.

La Figura 4.2 muestra esta relación. El eje Y (cantidad de discursos) es fijo, mientras que el eje X (“swing score”) se calcula para diferentes ventanas temporales de elecciones históricas (desde 1996 – 2020 hasta solo 2016 – 2020). Esto permite observar si la estrategia de campaña se alinea mejor con la competitividad reciente o con la histórica.

Relación entre Swing Score y frecuencia de discursos

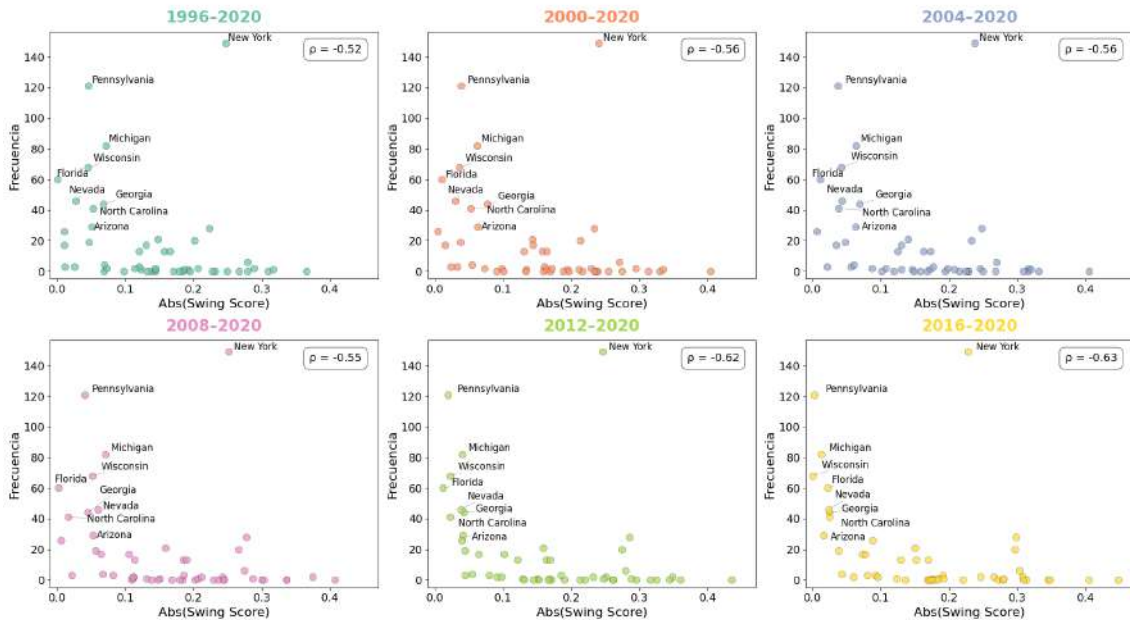


Fig. 4.2: Relación entre la cantidad de discursos por estado (1 de julio 2023 - 5 de noviembre 2024) y los *Swing Scores* calculados para diferentes intervalos de elecciones. Los 9 estados con mayor cantidad de discursos se muestran referenciados en los gráficos.

A partir de los resultados, se observa una clara tendencia: los estados con un “swing score” cercano a cero en valor absoluto concentran la mayor cantidad de discursos. Esta relación es más fuerte cuando se consideran las elecciones más recientes (ej. 2016 – 2020, con una correlación de Spearman de $-0,63$), lo que indica que las campañas ajustan su estrategia basándose en la competitividad actual.

Si se observan los estados con más discursos, el estado de New York, que a pesar de ser un estado seguro demócrata (con un swing score mayor a 0,2 para las últimas dos elecciones), atrae la mayor cantidad de discursos con más de 140, probablemente debido a su importancia como centro mediático y de recaudación de fondos. Por otro lado, el estado de Florida es el otro que no aparece dentro de la lista de swing states definida como aquellos con diferencia porcentual de votos menor al 3 %, sin embargo, presenta un swing score similar a los denominados swing states y una cantidad incluso mayor que algunos de ellos, con alrededor de 60. Esto puede pasar porque la cantidad de votos electorales para los estados de Florida y New York son de 30 y 28, respectivamente, colocándolos en los puestos 3 y 4 de estados con mayor cantidad de votos electorales. Esto es de interés para los candidatos ya que ganar un estado le garantiza todos los votos electorales que ofrece, y una suma mayor o igual a 270 asegura la victoria en la elección [9].

Al observar los primeros 9 estados con mayor cantidad de discursos, los swing states elegidos para la elección del 2024 (mostrados en la Figura 4.1) están incluidos dentro de este conjunto. Asimismo, dichos estados (a excepción de New York) presentan los swing scores más bajos en valor absoluto (calculados para el plazo más reciente, 2016-2020).

Por último, los resultados obtenidos a partir del análisis de resultados históricos indican que hay una mayor cantidad de contenido a analizar en los estados menos decisivos electoralmente, por lo que hay un interés en evaluar los denominados swing states y los estados de Florida y New York.

4.1.3. Representación espacial de tópicos

Con el fin de observar la distribución geográfica de discursos, se obtuvieron mapas para visualizar en qué estado pone el foco cada candidato. En la Figura 4.3 se puede observar cuatro mapas que representan la distribución de discursos en cada estado por candidato y total para todo el período de campaña.

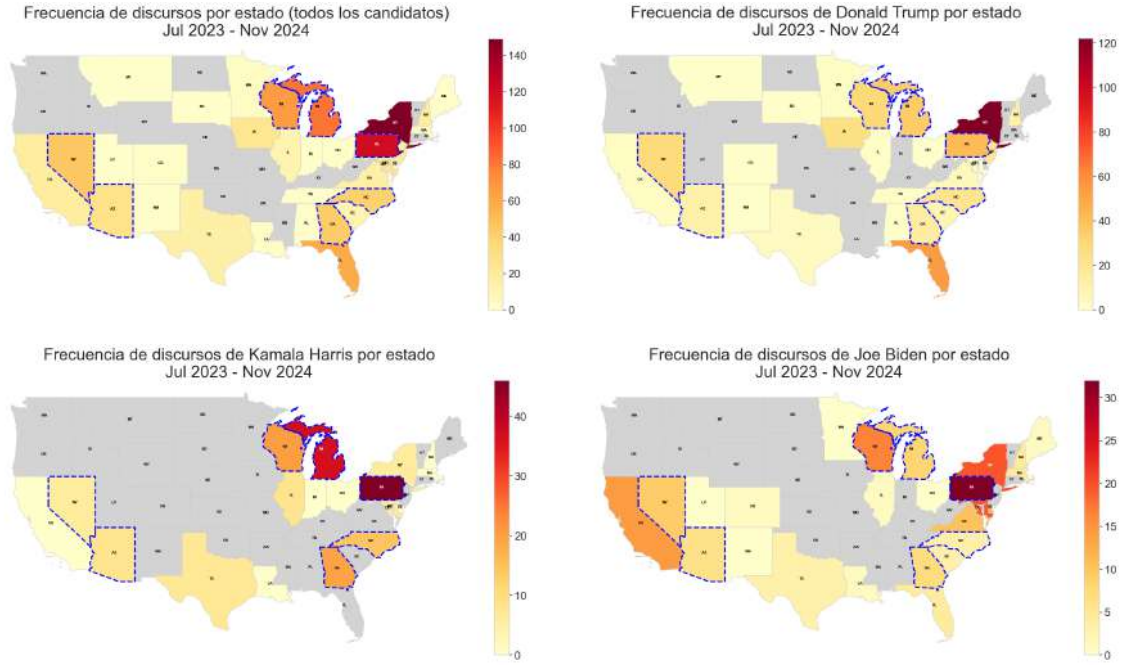


Fig. 4.3: Frecuencias de discursos por estado y candidato (julio 2023 - noviembre 2024). El conjunto de swing states está remarcado en azul.

En conjunto, los tres candidatos juntos ponen foco en el grupo de estados constituido por los swing states, New York y Florida, como se vio en 4.1.2.

Viendo al candidato Trump por separado, se puede observar que tiene una presencia fuerte en los estados de New York y Florida, seguido de los swing states, mientras que Harris se concentra directamente en los swing states del este, con menor presencia en los del oeste. Biden concentra sus discursos en los swing states, pero también pone un énfasis en el estado de California. La presencia de Biden en otros estados fuera del conjunto de swing states, New York y Florida puede deberse al hecho de que fue presidente en el tramo temporal analizado.

Adicionalmente, se construyeron mapas de calor para obtener distribuciones geográficas a partir de los tópicos obtenidos con Roll Call y BERTopic. Esto se hizo únicamente a modo exploratorio, con el fin de dar una interpretación cualitativa de la presencia de los tópicos en los distintos estados.

Para el caso de Roll Call, se construyeron mapas para cada tópico, donde se contabilizó su presencia en los discursos por estado. Los mapas de calor se muestran en la Figura 4.4, donde los estados más oscuros son aquellos con más menciones de cierto tópico.

Para los tópicos obtenidos con BERTopic, se construyeron mapas de calor a nivel estatal a partir de las distribuciones de discursos en tópicos con el fin de analizar la concentración geográfica de cada tema en el discurso de campaña. El proceso comenzó agregando por estado las puntuaciones de probabilidad que BERTopic asigna a cada discurso para cada uno de los 20 tópicos, obteniendo así una "puntuación de atención acumulada" para cada par (estado, tópico). Dado que estas puntuaciones brutas están sesgadas por el número de discursos en cada estado, se aplicó un método de escalado para medir la concentración relativa. Específicamente, se utilizó un escalado por máximo a nivel de tópico. Para cada tópico t , la puntuación de atención de cada estado s se dividió por la puntuación máxima

observada para ese mismo tópico en todo el país, según la Ecuación 4.1:

$$\text{Valor escalado}(s, t) = \frac{\text{Puntuación de Atención}(s, t)}{\max_{s'}(\text{Puntuación de Atención}_{(s', t)})} \quad (4.1)$$

El resultado es un valor entre 0 y 1 que indica la relevancia de un estado para un tema en comparación con el epicentro de esa discusión (el estado con valor 1). Este enfoque permite comparar la estrategia geográfica de cada mensaje de campaña de forma equitativa, revelando dónde se decidió enfocar cada temática.

En la Figura 4.5 se pueden observar los mapas de calor obtenidos con BERTopic.

La elección del enfoque para cada conjunto de mapas fue una consecuencia directa de la estructura de los datos. Para los tópicos de Roll Call, que asignan igual importancia a los temas sin importar su proporción real (por ejemplo un discurso compuesto por un 80 % del tópico “Economía” y 20 % “Género” es tratado como 50/50 en la distribución de probabilidad de tópicos), se utilizó un conteo de frecuencia simple. Este método no solo es el más fiel a la naturaleza de los datos, sino que, al preservar la escala absoluta, permite también comparar la prominencia general entre los distintos temas a lo largo de toda la campaña. Por el contrario, para BERTopic, que ofrece una distribución de probabilidad más ajustada a los párrafos del discurso, se optó por un análisis de probabilidad escalada. De este modo, cada visualización representa con la mayor fidelidad posible la naturaleza de la información que la originó.

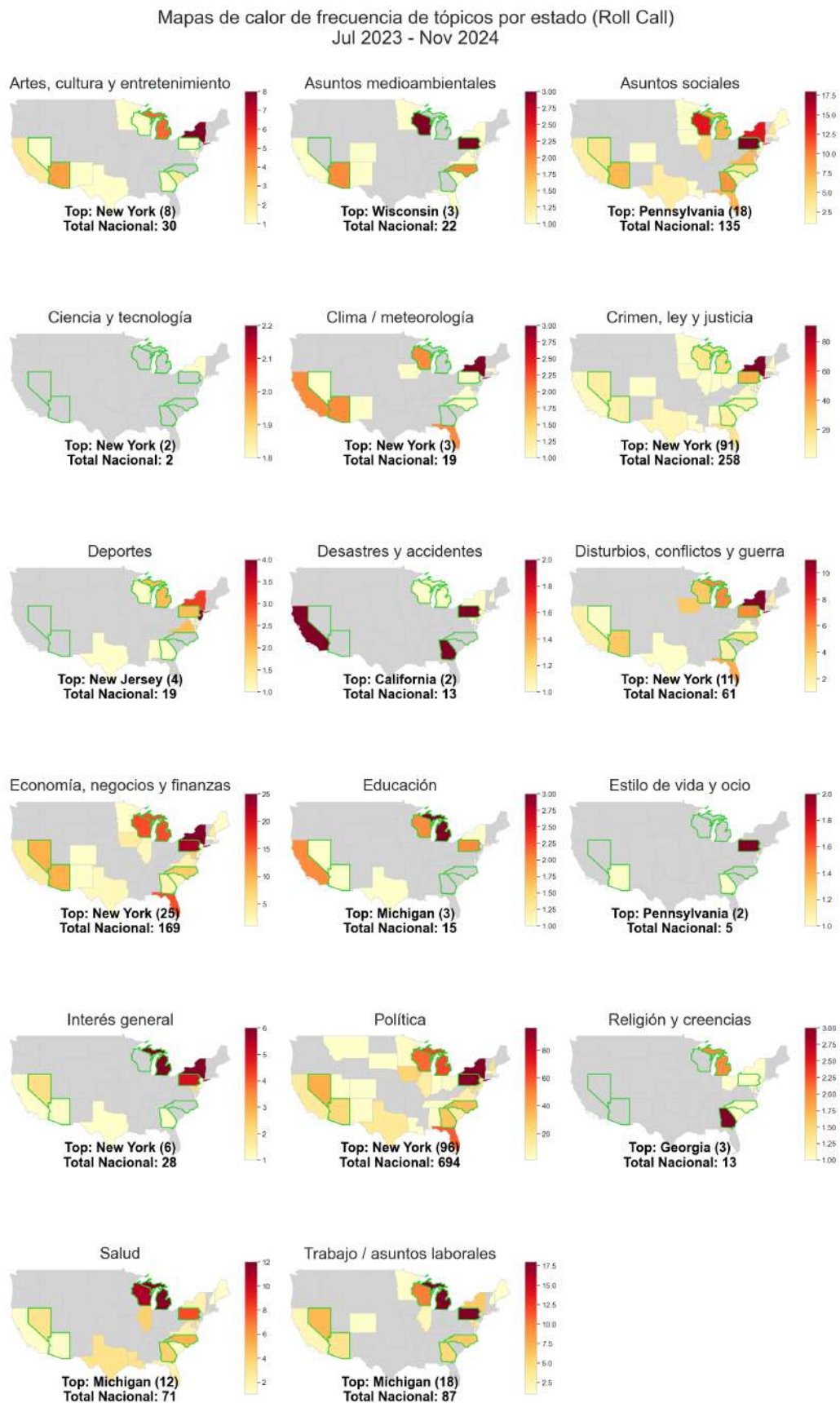


Fig. 4.4: Frecuencia de discursos por tópico (Roll Call; mapas de calor). Los swing states aparecen remarcados en verde.

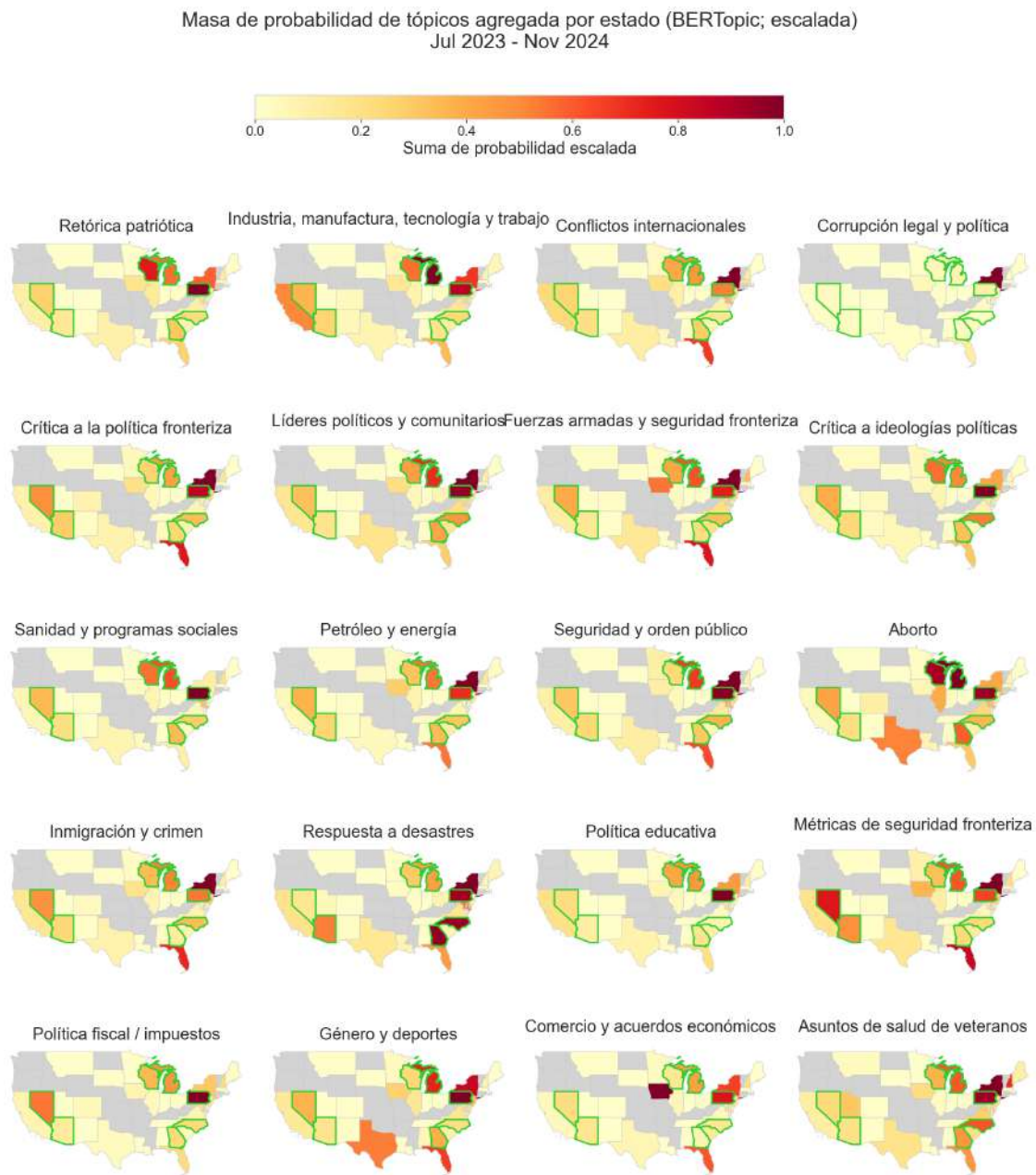


Fig. 4.5: Distribución escalada de tópicos de discursos por estado, (BERTopic, mapas de calor). Los swing states aparecen remarcados en verde.

A partir de las Figuras 4.4 y 4.5 se pueden hacer varias observaciones. Para los tópicos de Roll Call, el tópico “Política” es el más prominente con 694 discursos totales, aunque su presencia está distribuida a lo largo del mapa.

Por otro lado, el tópico “Crimen, ley y justicia” comprende 258 discursos con una presencia abrumadora en el estado de New York, donde se registraron 91 de ellos, una cifra muy superior a la de cualquier otro estado. Esta concentración geográfica coincide directamente con el patrón observado en el mapa del tópico “Corrupción legal y política” de BERTopic.

Este comportamiento atípico en un estado no competitivo (“safe state”) puede atribuirse directamente a factores externos a la estrategia de campaña electoral. Una posible causa puede ser la alta concentración de apariciones públicas y declaraciones de Donald Trump en Nueva York en relación con su juicio penal, conocido como el caso “hush money”, que tuvo lugar en Manhattan durante gran parte del período analizado [35]. Esto es reforzado por el hecho de la presencia de la palabra “indicted” (de “indictment” o “acusación” en español) o “trial” (“juicio”) en la nube de palabras “Corrupción legal y política” de la Figura 3.1, que están relacionadas a los 34 cargos de falsificación de registros comerciales.

Viendo el tópico de Roll Call “Desastres y accidentes”, la presencia de discursos en California o Georgia puede estar relacionada a la presencia de incendios forestales o huracanes, respectivamente. Si se analiza el tópico “Respuesta a desastres” de BERTopic (que puede ser el tópico equivalente en Roll Call) se ve la fuerte presencia de dos estados típicamente afectados por tormentas y huracanes, pero el tópico se ve más distribuido a lo largo del mapa en comparación al de Roll Call.

Como ejemplo adicional de comparación entre ambos espacios de tópicos, se puede observar en ambas categorías que Michigan es un estado fuertemente marcado por dos temas principales: para Roll Call, los tópicos “Salud” y “Trabajo / asuntos laborales” y en BERTopic los tópicos “Aborto” e “Industria, manufactura, tecnología y trabajo” son los que mayor influencia tienen en el estado. Observando la nube de palabras “Industria, manufactura, tecnología y trabajo” en la Figura 3.1, la presencia de términos relacionados a la industria automotriz refuerza la presencia de este tópico en este estado; Michigan es un centro industrial especializado particularmente en este sector [36].

Se pueden hacer estimaciones cualitativas a partir de la visualización de los mapas de calor, pero las estimaciones son meramente interpretativas.

Adicionalmente, con el fin de analizar la agenda (los temas abordados) de cada candidato, se obtuvieron los tópicos más hablados por estado a partir del método mencionado para la Figura 4.5. En la Figura 4.6 se pueden observar los resultados.

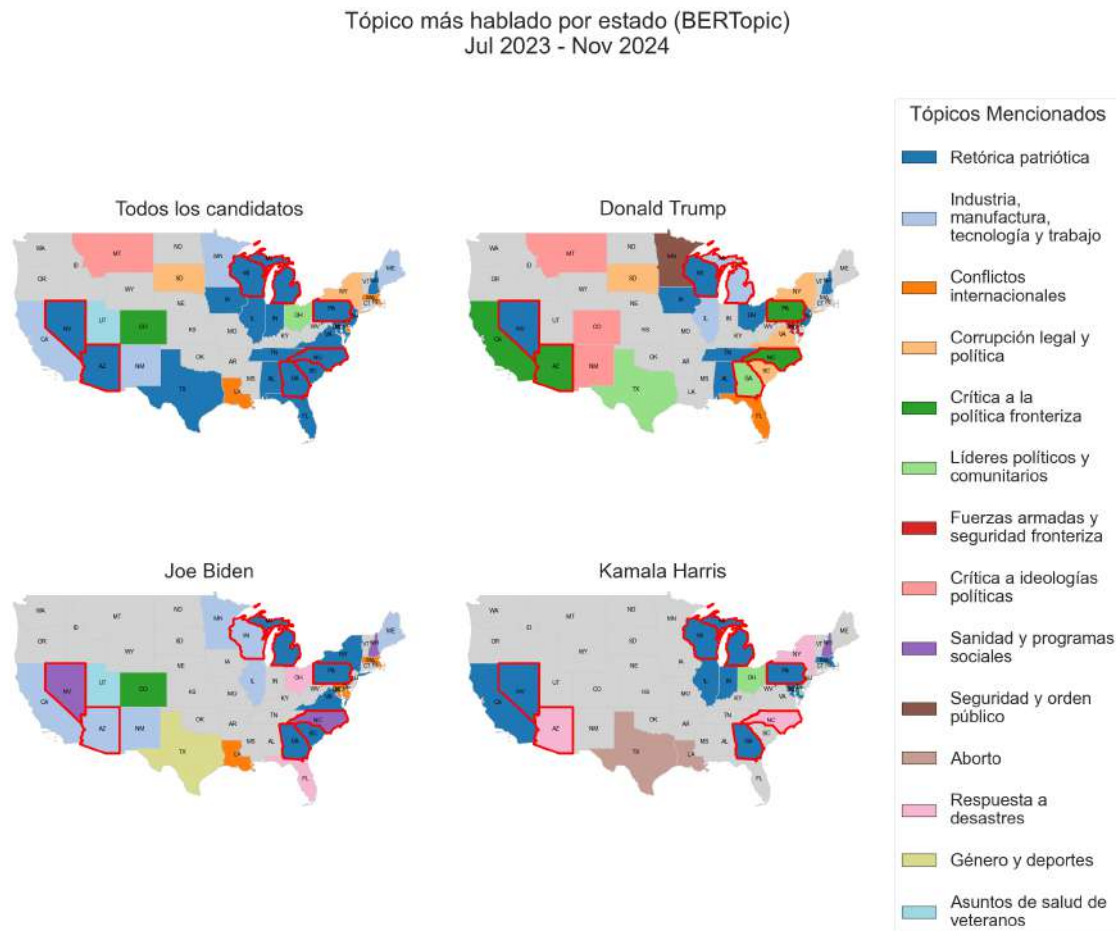


Fig. 4.6: Tópico más hablado por estado, BERTopic. Los swing states aparecen remarcados en rojo.

En la Figura 4.6 se puede ver que el tópico más mencionado, observando únicamente los estados de mayor interés, es el de “Retórica patriótica”. Sin embargo, al ver candidato por candidato, esto cambia.

Para Trump, los temas más importantes dentro de los swing states están relacionados a la política fronteriza o al sector industrial. Una observación importante es la presencia del tópico “Corrupción legal y política” en el estado de New York, reforzando la conjetura de que este tópico está relacionado fuertemente a los juicios de corrupción a este candidato.

Biden discute temas relacionados a la industria en estados de interés como Wisconsin o Arizona, pero también se enfoca en los temas sociales como “Respuesta a desastres” en North Carolina o “Sanidad y programas sociales” en Florida. Esto coincide parcialmente con la agenda de Harris, donde los temas de discusión son mayormente sociales, como “Respuesta a desastres” o “Aborto”.

En resumen, la agenda de Trump está más marcada por la inmigración y la corrupción, mientras que los candidatos demócratas no se enfocan en estos temas y deciden hablar de asuntos sociales.

4.2. Representación temporal y relevancia de los tópicos

El objetivo de esta sección es doble: primero, analizar la distribución temporal de los discursos para justificar la partición del análisis en distintos períodos; y segundo, evaluar la relevancia de los tópicos a lo largo del tiempo para seleccionar los más prominentes.

La Figura 4.7 muestra la frecuencia semanal de discursos para cada candidato. Se observa que tanto Trump como Biden mantienen una actividad constante durante todo el período analizado. En el caso de Kamala Harris, aunque es una figura política activa, los discursos recopilados en este trabajo comienzan en julio de 2024, momento en que asume un rol central en la campaña presidencial (la disponibilidad de discursos de Harris previa a julio de 2024 es escasa, con dos discursos [7]). Es notable el pico de actividad de campaña para Trump y Harris en octubre de 2024, justo antes de las elecciones.

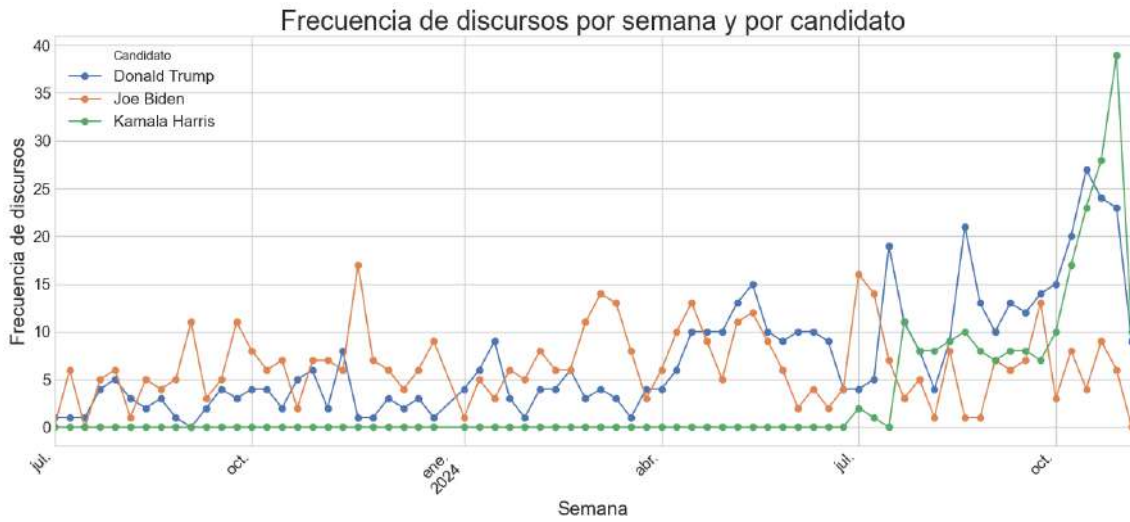


Fig. 4.7: Frecuencia de discursos por semana, por candidato (julio 2023 - noviembre 2024).

Para analizar el contenido de estos discursos en el tiempo, se construyeron series temporales diarias para cada tópico, tal como se definió en la Sección 3.2. La Figura 4.8 visualiza una selección de estas series para ambas representaciones de tópicos (Roll Call y BERTopic).

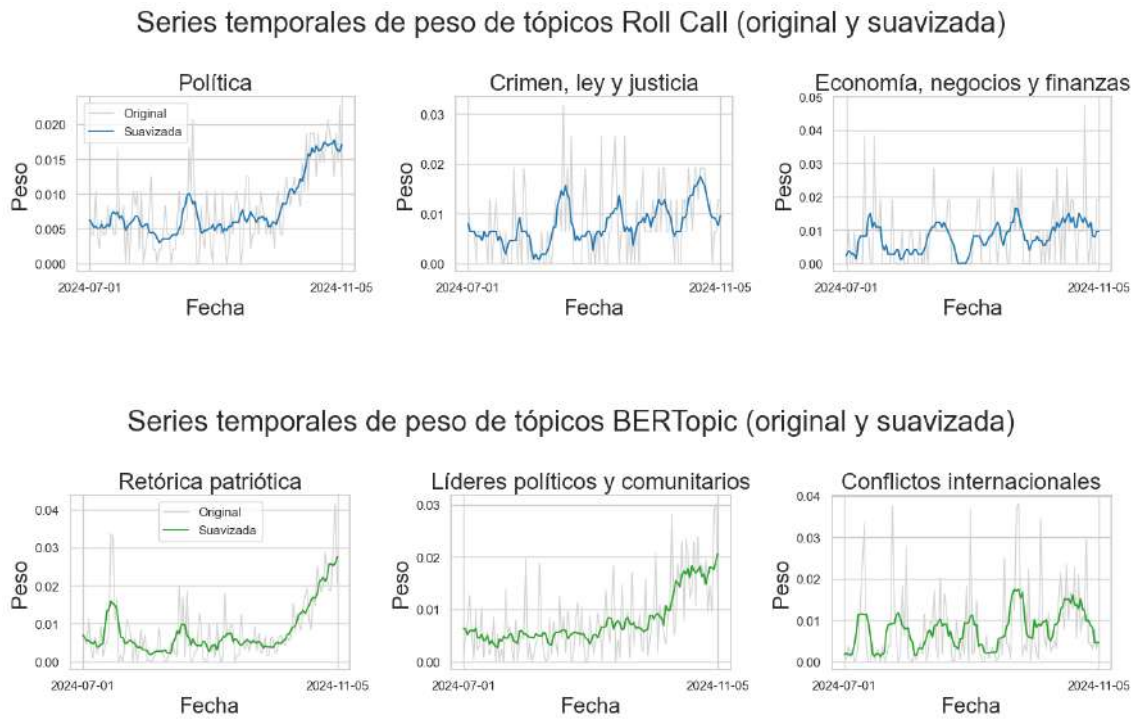


Fig. 4.8: Peso diario de un subconjunto de tópicos entre julio y noviembre de 2024. **Arriba:** Roll Call; **Abajo:** BERTopic. Las fluctuaciones diarias han sido atenuadas mediante una media móvil detallada en la Sección 3.2.

La comparación visual entre los grupos de series de la Figura 4.8 no revelan una diferencia estructural, sin embargo, las distribuciones de discursos de Roll Call son más dispersas (con más valores en cero) que las de BERTopic, por construcción. El método de BERTopic, basado en párrafos, es más propenso a asignar pequeñas probabilidades a múltiples temas, mientras que el etiquetado a nivel de discurso de Roll Call es más binario (el tópico aparece o no en un discurso). La implicancia de esto es que las series temporales de tópicos de BERTopic son densas (tienen pocos ceros), incluso para los tópicos de menor presencia, mientras que las series de Roll Call presentan pocos tópicos suficientemente densos para el análisis de correlaciones.

Las oscilaciones y picos en estas series son de gran interés, ya que reflejan momentos en que la campaña intensificó la comunicación sobre un tema específico, probablemente en respuesta a eventos externos o como parte de una estrategia deliberada. Por ejemplo, la época de huracanes en Estados Unidos ocurre en los meses de septiembre/octubre, por lo que es esperable ver una estacionalidad en series como “Respuesta a desastres”, con picos en estos meses. A modo de ejemplo, en la Figura 4.9 se pueden observar las series de relevancia del tópico “Respuesta a desastres” de BERTopic para el plazo julio-noviembre de 2023 y 2024. Para 2023 se puede observar un pico en septiembre, mientras que en 2024 hay un pico en octubre.

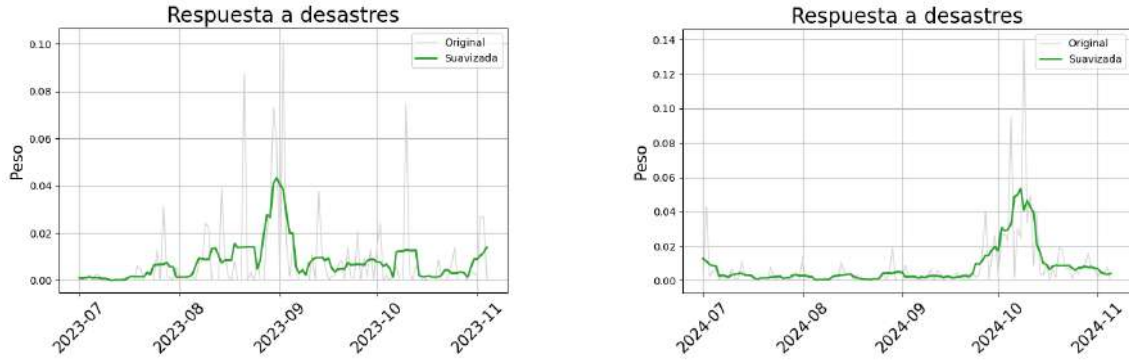


Fig. 4.9: Peso de tópico “Respuesta a desastres” (BERTopic) para julio-noviembre de 2023 (izquierda) y 2024 (derecha).

Por otro lado, los aumentos notables correspondientes a los últimos meses de campaña (como en las series temporales correspondientes a los tópicos “Política” o “Retórica patriótica”) pueden estar relacionados al aumento en la cantidad de discursos visto en la Figura 4.7. Si bien las series están escaladas por una constante de normalización, el aumento en el número de discursos aumenta las chances de que ver una mayor cobertura de los tópicos.

Debido a la alta dispersión de las series de Roll Call, se implementó un criterio de selección para filtrar los tópicos menos frecuentes antes del análisis de correlación. Para cada período definido en la Sección 4.3.1, los tópicos se ordenaron según la cantidad de días con actividad no nula (es decir, con un peso mayor a cero). Luego de aplicar el suavizado se filtraron las series temporales de tópicos con más de 20 valores nulos. Esto permitió seleccionar un subconjunto de los tópicos más consistentemente discutidos y suficientemente densos para el análisis de correlaciones en cada ventana temporal.

Los subconjuntos de tópicos para las primeras dos etapas definidas en la Sección 4.3.1 se mencionan en la Sección B del Apéndice, junto a los resultados del análisis de correlaciones. Para las últimas dos etapas, se consideraron 4 y 7 tópicos de Roll Call para el análisis de correlaciones. Los tópicos correspondientes a la etapa final entre Trump y Biden son “Política”, “Crimen, ley y justicia”, “Asuntos sociales” y “Economía, negocios y finanzas”; los tópicos de la etapa general entre Trump y Harris son “Política”, “Crimen, ley y justicia”, “Economía, negocios y finanzas”, “Asuntos sociales”, “Trabajo / asuntos laborales”, “Salud” y “Disturbios, conflictos y guerra”.

Por otro lado, para BERTopic no se filtraron series, considerando todos los tópicos para el análisis de correlaciones.

4.3. Configuración del análisis de correlación

El objetivo central de esta tesis es cuantificar la relación entre la atención de cada candidato a presidente a un tópico (medida por las series temporales de discursos) y la intención de voto (medida por las encuestas de opinión). Para ello, se diseñó un marco de análisis que considera diferentes períodos temporales y la naturaleza de las series de datos.

4.3.1. Definición de períodos de análisis

La dinámica de una campaña electoral cambia drásticamente con el tiempo. Para capturar estas diferentes fases, el período global de julio de 2023 a noviembre de 2024 se segmentó en cuatro plazos temporales:

- **Etapla temprana (Trump vs. Biden):** 01/07/2023 - 01/11/2023.
- **Etapla media (Trump vs. Biden):** 01/11/2023 - 01/03/2024.
- **Etapla final (Trump vs. Biden):** 01/03/2024 - 21/07/2024, concluyendo con el retiro de Biden.
- **Etapla general (Trump vs. Harris):** 01/07/2024 - 05/11/2024, cubriendo la recta final de la elección.

Esta segmentación en ventanas de aproximadamente cuatro meses (~ 120 días) asegura una cantidad suficiente de puntos de datos para un análisis de correlación robusto, incluso después de aplicar los desplazamientos temporales descritos en la Sección 4.3.2.

En la Sección 4.5 se hace énfasis en los últimos dos períodos, que son los más relevantes simplemente por proximidad a la fecha de elecciones. En el Apéndice (B) se pueden ver los resultados obtenidos para los dos períodos anteriores.

4.3.2. Metodología de correlación con desplazamiento temporal para series de encuestas

El efecto de un discurso sobre la opinión pública puede no necesariamente ser inmediato. Además, como se mencionó en la Sección 2.1.1, la propia métrica de encuestas de RealClearPolling ya incorpora un promedio de días anteriores. Para modelar este retardo, el análisis de correlación se realizó aplicando un desplazamiento temporal (lag) a las series de encuestas.

Específicamente, para cada tópico se calculó la correlación entre su serie temporal y la serie de intención de voto, la cual fue retardada sistemáticamente k días hacia el pasado, con k variando de 7 a 28 días. El valor de k que maximiza la correlación de Spearman se reporta como el “lag” óptimo. La razón para priorizar la correlación de Spearman sobre la de Pearson es que la primera logra capturar relaciones monótonas de todo tipo, mientras que la segunda sólo se limita a encontrar relaciones lineales. Este procedimiento permite identificar si el impacto de un tema en la intención de voto se manifiesta una, dos, o hasta cuatro semanas después de que se discute.

Una consideración importante en el cálculo de correlación entre una serie retardada y otra fija es que la correlación se obtiene de la intersección de ambas series. Es decir, si se considera un lag de k días en un intervalo de tiempo $[1, T]$, sólo las observaciones correspondientes a los tiempos t de la serie fija en el intervalo $[1, T - k]$ se comparan con las observaciones correspondientes a las posiciones $t + k$ de la serie original, del intervalo $[k + 1, T]$. Por lo tanto, se usan k menos datos al calcular la correlación con el uso de lags. Esta fue otra cuestión a considerar en la elección de ventanas temporales discutida en la Sección 4.3.1.

A modo de visualización, en la Figura 4.10 se muestran las series temporales correspondientes a la diferencia de intención de voto (medida en porcentaje) entre Trump y Biden o Harris para todo el período de análisis, a nivel nacional. La diferencia de intención de voto

se mide primero considerando la intención de Trump menos la del candidato demócrata, la diferencia puede variar entre -100 y 100 puntos. Estas series, junto a las de cada estado, son las que se busca correlacionar con la dinámica de los tópicos.

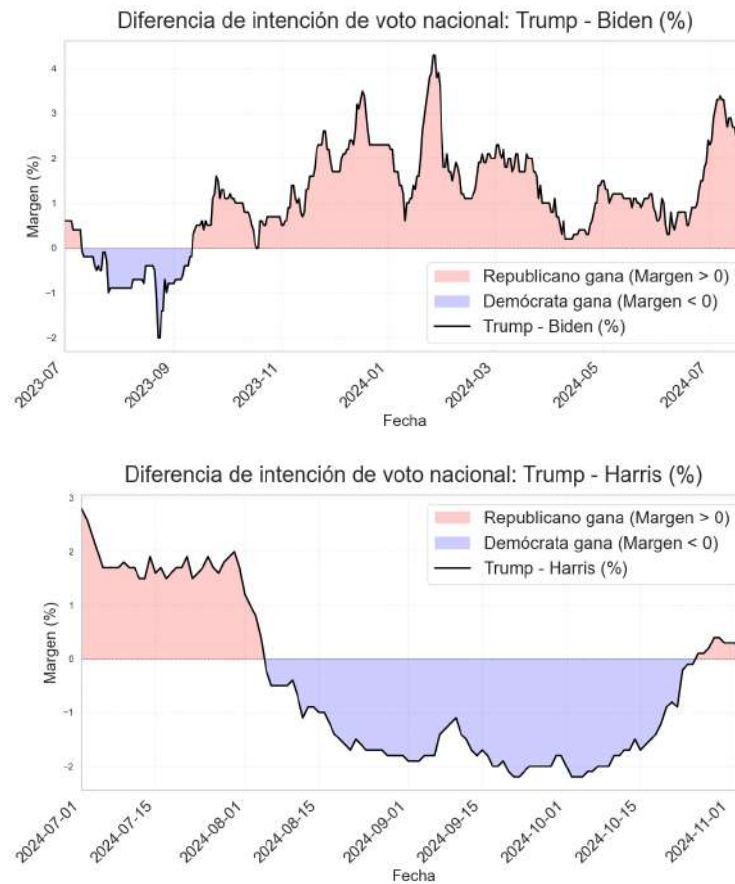


Fig. 4.10: Evolución de la diferencia en intención de voto a nivel nacional (%). **Arriba:** se muestra el margen entre Trump y Biden. **Abajo:** el margen entre Trump y Harris.

En la Figura 4.10 se puede observar que Trump tuvo una ventaja grande sobre Biden a nivel país en casi todo el período. Es razonable asumir que esta fue una de las razones principales de la renuncia de Biden a la carrera presidencial el 21 de julio de 2024. Posteriormente, viendo a Trump contra Harris, la diferencia de intención de voto siguió resultando favorable para Trump hasta que en agosto empezó a resultar favorable para Harris, terminando en un repunte para Trump y una leve ventaja en la recta final. La carrera estuvo pareja para Trump y Harris, donde raramente la diferencia de intención de voto superó los dos puntos.

4.3.3. Alcance del análisis estatal

Además del análisis a nivel nacional, se replicó la metodología de correlación para un conjunto de estados clave. La selección incluyó a los “swing states” identificados previamente (Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania y Wisconsin), a los que se añadieron Florida y New York. La elección de este conjunto de estados fue hecha en base a los resultados obtenidos en la Sección 4.1.2.

4.4. Correlación entre el contenido de los discursos y las encuestas

En esta sección se presenta el núcleo del análisis cuantitativo: la medición de la correlación entre las series temporales de tópicos y las series de intención de voto.

4.4.1. Diseño del análisis de correlación

El análisis se diseñó para medir cómo la discusión de un tema a nivel nacional se relaciona con la opinión pública, tanto a escala nacional como en los “swing states” clave. Para ello, se compararon las series temporales de tópicos nacionales (construidas a partir de todos los discursos, sin importar su ubicación) con dos conjuntos de encuestas: las nacionales y las de cada estado individual. Se optó por utilizar las series de tópicos nacionales para el análisis estatal, en lugar de series estatales individuales, debido a la baja densidad de discursos en los estados, lo que impedía construir series temporales locales suficientemente densas, de acuerdo al criterio de densidad de series discutido en la Sección 4.2.

Como se detalló en la Sección 4.3, para considerar el efecto no inmediato de los discursos en la opinión pública, se aplicó un desplazamiento temporal (lag) y se obtuvieron los lags óptimos (correspondientes a los que maximizaban la correlación de Spearman). Una vez fijado el lag óptimo, se reportaron tanto la correlación de Spearman como la de Pearson. Se utilizó un test de permutación para evaluar la significancia estadística de ambos coeficientes (ver Apéndice C).

4.4.2. Interpretación de los coeficientes de correlación

Una correlación positiva entre la relevancia de un tópico y una serie de diferencia de intención de voto (“Trump - Demócrata”) indica que, a medida que se discute más el tema, el margen de Trump aumenta. Inversamente, una correlación negativa sugiere que la discusión del tema favorece al candidato demócrata.

La evaluación conjunta de los coeficientes de Pearson (r) y Spearman (ρ) en el lag óptimo permite una comprensión más matizada de la naturaleza de la asociación, como se resume en la Tabla 4.1.

Tab. 4.1: Interpretación de la relación entre los coeficientes de Pearson (r) y Spearman (ρ).

	Pearson (r) alto	Pearson (r) bajo
Spearman (ρ) alto	Relación lineal fuerte: la asociación es directa y de ritmo constante. La similitud entre ambos coeficientes confirma la linealidad.	Relación monótona no lineal: hay una tendencia clara, pero afectada por curvatura (ej. rendimientos decrecientes) o por valores atípicos a los que Pearson es sensible.
Spearman (ρ) bajo	Posible valor atípico: la aparente linealidad puede ser una ilusión causada por unos pocos puntos influyentes. Spearman revela la falta de una tendencia general.	Ausencia de relación: no se detecta una asociación clara, ni lineal ni monótona. Los datos se muestran dispersos.

Analizando la Tabla 4.1, un valor alto de correlación de Spearman, independientemente de la magnitud del valor de correlación de Pearson, podría indicar que hay una relación entre la serie de tópicos y de encuestas (lineal si el coeficiente de Pearson es similar o no lineal de no serlo). Esto no es así para el caso del coeficiente de Pearson, que puede ser alto pero un valor de correlación de Spearman bajo indica una falta de relación lineal y posibles valores atípicos influenciando el valor de correlación de Pearson. Esta puede ser una razón para determinar que el coeficiente de Spearman es más robusto y que puede ser interpretado por sí solo, mientras que el coeficiente de Pearson no.

4.5. Resultados

En las siguientes tablas se presentan de las correlaciones más altas y estadísticamente significativas ($p < 0,05$) a nivel nacional y estatal (Sección 4.5.1 y Sección 4.5.2, respectivamente). Los resultados están ordenados decrecientemente según el valor absoluto de la correlación de Spearman y los títulos de cada tópico se muestran en rojo (indicando una correlación positiva, favoreciendo a Trump) o azul (correlación negativa, favoreciendo al candidato demócrata) con el fin de facilitar la interpretación para cada caso.

4.5.1. Nivel nacional

Tab. 4.2: Correlaciones tópico-(Trump - Biden), nación, Marzo - Julio 2024.

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Sanidad y programas sociales	-0,61	-0,54	(L:26, P _S :,002, P _P :,002)
Crítica a la política fronteriza	0,58	0,58	(L:23, P _S :,002, P _P :,002)
Género y deportes	0,58	0,48	(L:12, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	-0,48	-0,43	(L:28, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,47	0,64	(L:28, P _S :,002, P _P :,002)
Política fiscal / impuestos	0,47	0,36	(L:13, P _S :,002, P _P :,002)
Seguridad y orden público	0,45	0,26	(L:27, P _S :,002, P _P :,008)
Respuesta a desastres	-0,41	-0,31	(L:28, P _S :,002, P _P :,002)
Inmigración y crimen	0,39	0,45	(L:28, P _S :,002, P _P :,002)
Política educativa	0,38	0,22	(L:9, P _S :,002, P _P :,026)
Crítica a ideologías políticas	0,36	0,55	(L:7, P _S :,002, P _P :,002)
Retórica patriótica	0,34	0,42	(L:7, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	0,32	0,29	(L:21, P _S :,002, P _P :,006)
Métricas de seguridad fronteriza	0,32	0,39	(L:19, P _S :,002, P _P :,002)
Conflictos internacionales	-0,28	-0,31	(L:28, P _S :,006, P _P :,002)
Corrupción legal y política	-0,26	-0,36	(L:7, P _S :,006, P _P :,002)
Aborto	0,23	0,27	(L:7, P _S :,006, P _P :,004)
Líderes políticos y comunitarios	0,21	0,19	(L:8, P _S :,016, P _P :,044)
<i>Tópicos Roll Call</i>			
Economía, negocios y finanzas	-0,57	-0,59	(L:28, P _S :,002, P _P :,002)
Política	0,30	0,33	(L:7, P _S :,002, P _P :,002)

Tab. 4.3: Correlaciones tópico-(Trump - Harris), nación, Julio - Noviembre 2024.

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Respuesta a desastres	0,57	0,61	(L:28, P _S :,002, P _P :,002)
Petróleo y energía	-0,52	-0,56	(L:7, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	-0,45	-0,43	(L:7, P _S :,002, P _P :,002)
Crítica a la política fronteriza	-0,37	-0,29	(L:28, P _S :,002, P _P :,004)
Industria, manufactura, tecnología y trabajo	-0,36	-0,24	(L:7, P _S :,002, P _P :,008)
Política fiscal / impuestos	0,34	0,39	(L:25, P _S :,002, P _P :,002)
Métricas de seguridad fronteriza	-0,31	-0,26	(L:7, P _S :,002, P _P :,012)
Crítica a ideologías políticas	-0,30	-0,26	(L:13, P _S :,008, P _P :,004)
Retórica patriótica	0,29	0,30	(L:20, P _S :,004, P _P :,002)
Inmigración y crimen	-0,25	-0,26	(L:7, P _S :,010, P _P :,004)
Género y deportes	-0,22	-0,24	(L:25, P _S :,026, P _P :,016)
Corrupción legal y política	0,21	0,18	(L:7, P _S :,022, P _P :,042)
Aborto	-0,20	-0,25	(L:27, P _S :,040, P _P :,016)
<i>Tópicos Roll Call</i>			
Política	0,36	0,25	(L:22, P _S :,002, P _P :,004)
Disturbios, conflictos y guerra	0,33	0,22	(L:27, P _S :,002, P _P :,024)
Crimen, ley y justicia	-0,31	-0,21	(L:7, P _S :,002, P _P :,018)
Trabajo / asuntos laborales	0,30	0,20	(L:27, P _S :,004, P _P :,048)
Asuntos sociales	-0,23	-0,25	(L:7, P _S :,026, P _P :,006)

En la Tabla 4.2 se pueden observar las correlaciones obtenidas para la ventana marzo-julio de 2024 (disputado entre Trump y Biden), mientras que la Tabla 4.3 muestra el periodo julio-noviembre del mismo año (disputado entre Trump y Harris).

Para el período Trump-Biden se puede observar que los tópicos con mayor correlación fueron los relacionados a programas sociales, frontera, deportes y el sector industrial. Para Trump, los tópicos relacionados a frontera y género y deportes están asociados a una mejora en las encuestas, mientras que los programas sociales y el sector industrias favorecieron a Biden. La diferencia entre las correlaciones para el tópico de “Género y deportes” puede indicar que si bien la tendencia general favorecía a Trump, el impacto no fue estrictamente lineal y pudo estar influenciado por picos de atención mediática. Por otro lado, el tópico de “Crítica a la política fronteriza” indica una coincidencia entre las correlaciones de Spearman y Pearson (0,58), donde se puede observar que la relación entre la mención al tópico puede ser lineal.

En términos generales, ninguna correlación supera el 0,65 en valor absoluto. La más alta se obtuvo en el tópico de “Fuerzas armadas y seguridad fronteriza” (Pearson: 0,64), pero este es un caso donde la correlación de Spearman es considerablemente más baja (0,47), indicando que hay una falta de tendencia general fuerte (sea lineal o no).

Comparando con los tópicos de Roll Call, cuya representación de tópicos es más débil, se obtuvo una correlación favorable y lineal para Biden en el tópico “Economía, negocios y finanzas”. Esto puede coincidir con el hecho de que el tópico “Industria, manufactura, tecnología y trabajo” fue favorable para Biden también, aunque no es concluyente ya que

tópicos de BERTopic relacionados a la economía, como “Política fiscal / impuestos” están relacionados a una mejoría para Trump.

Por otro lado, para el período Trump-Harris se pueden observar correlaciones generalmente más bajas, pocas superan 0,5 en valor absoluto. Sin embargo, el tópico “Respuesta a desastres” fue el que mostró mayor correlación para este período, favoreciendo a Trump. Una posible interpretación es la presencia de la temporada de huracanes en septiembre/octubre y la cobertura relacionada a la asistencia por parte de FEMA y Starlink (observada en la nube de palabras de la Figura 3.1) observada en este período. Una observación interesante se encuentra en la Tabla B.1 (período correspondiente a julio-noviembre de 2023), donde se puede ver que este tópico también tuvo las correlaciones más altas (0,68 y 0,60), favoreciendo a Trump. La estacionalidad de la serie puede haber influido en el resultado obtenido, y muestra una consistencia temática conservada para ambos períodos (que cubren los mismos meses pero de años distintos). El segundo tópico con correlación más alta en valor absoluto fue “Petróleo y energía” (con valores de $-0,52$ y $-0,56$), favoreciendo a Harris.

Para los tópicos de Roll Call, los bajos valores de correlación no indican un resultado interesante o fácilmente interpretable. Las series temporales pueden llegar a capturar tópicos de una manera menos confiable debido a la poca suavidad de distribuciones de discursos discutidas en la Tabla 3.1.

El contraste entre los dos períodos es claro. La campaña pasó de estar dominada por la inmigración y los asuntos sociales en la era Trump-Biden, a centrarse en la gestión de crisis (desastres naturales) y la política macroeconómica en la era Trump-Harris.

4.5.2. Nivel estatal

Para analizar correlaciones a nivel estatal y realizar interpretaciones de manera más fácil, se fabricaron gráficos de radar para cada tópico de BERTopic, los tópicos de Roll Call se reportan en la Sección A.

En las Figuras 4.11, 4.12, 4.13, 4.14 y 4.15 se muestran los radares, donde hay un polígono por ventana. Cada vértice de los polígonos muestra la correlación de Spearman obtenida para el tópico en el estado (se prioriza Spearman por su carácter de poder ser interpretado por sí solo, como fue mencionado en la Tabla 4.1) y la extensión de dicho vértice muestra la magnitud de la correlación en valor absoluto. Las tablas completas de correlación para cada estado se muestran en la Sección B del Apéndice, y son analizadas en conjunto con los gráficos de esta sección.

Para facilitar la interpretación, se separaron los tópicos de BERTopic en las cinco categorías mencionadas en la Sección 3.3.1. Estas categorías son: “Economía y política fiscal”, “Seguridad y política migratoria”, “Asuntos sociales y de salud”, “Política exterior y desastres” y “Retórica y estrategia política”.

Para los estados de New York y Florida no se encontraron correlaciones significativas, por lo que el análisis se reduce solamente al conjunto de swing states. A continuación se muestran los gráficos de radar con su análisis interpretativo correspondiente.

Economía y política fiscal

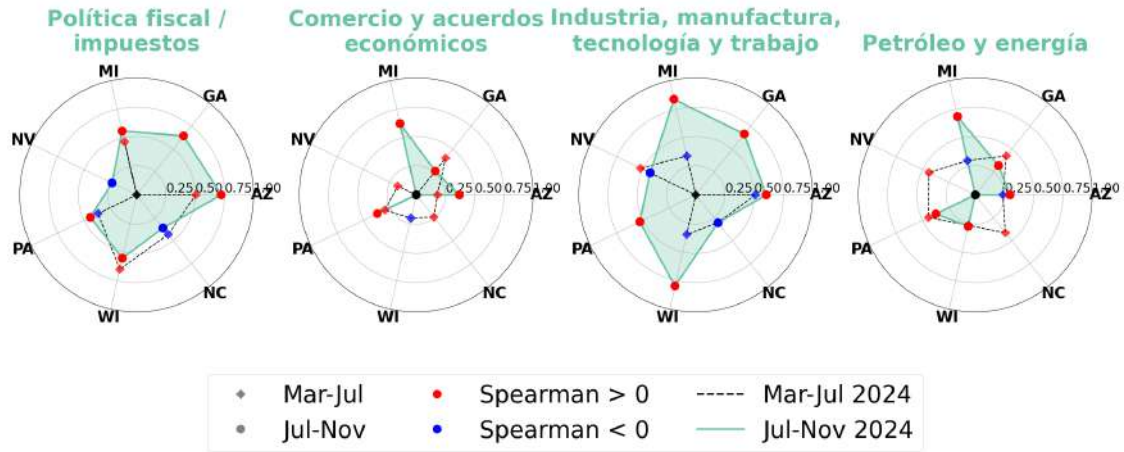
Evolución de tópicos BERTopic, por estado - Categoría: Economía y política fiscal

Fig. 4.11: Evolución de correlaciones por tópico, BERTopic. Categoría: Economía y política fiscal. Períodos marzo - julio y julio - noviembre de 2024.

Viendo la Figura 4.11, se pueden hacer varias apreciaciones. El tópico “Industria, manufactura, tecnología y trabajo” es el que da correlaciones más altas, seguido de “Política fiscal / impuestos”. El estado de Michigan (MI) consistentemente presentó correlaciones altas para todos los tópicos en el tramo de julio - noviembre, a esto se le sumó Wisconsin (WI), que presenta una correlación superior a 0,75 para el tópico de “Industria, manufactura, tecnología y trabajo” y una correlación cercana a 0,6 para “Política fiscal / impuestos”. Haciendo un análisis general, se puede observar que los estados más influenciados por esta categoría de tópicos son Michigan, Wisconsin, Georgia (GA) y Arizona (AZ), mientras que los estados de Pennsylvania (PA) y North Carolina (NC) son influenciados en menor medida. Otra apreciación es que para el tramo correspondiente a Trump-Biden, las encuestas en algunos estados estuvieron relacionadas a un cambio favorable para Biden, como el caso de North Carolina para los tópicos de “Política fiscal / impuestos” e “Industria, manufactura, tecnología y trabajo”, sin embargo, el tramo Trump-Harris estuvo marcado por una clara ventaja para Trump en caso todos los estados.

Si se observan las Figuras 4.5 y 4.6, se puede ver que el tópico correspondiente a “Industria, manufactura, tecnología y trabajo” concentra una distribución fuerte en los estados de Michigan y Wisconsin, sumado al hecho de que este tópico fue el más discutido para Trump en Michigan y para Biden en Wisconsin. Adicionalmente, viendo los términos relacionados a este tópico a partir de la nube de palabras en la Figura 3.1 (“auto”, “cars”, “electric”, “manufacturing”) y dado el contexto industrial de Michigan (donde la *General Motors* contrata a más personas que el propio sector público, [36]), se puede decir que el tópico tiene una gran influencia o relación con las decisiones de los electores. Incluso, la Tabla B.10, correspondiente al mismo estado pero en el período noviembre 2023 - marzo 2024, también posiciona a este tópico como primero, reforzando la influencia del tópico a lo largo del tiempo. Asimismo, evaluando los tópicos de Roll Call para Michigan, se puede ver en la Tabla B.12 muestra correlaciones altas para tópicos como “Trabajo / asuntos laborales” y “Economía, negocios y finanzas”, indicando una consistencia temática y de

signo de correlación (favoreciendo a Trump).

Seguridad y política migratoria

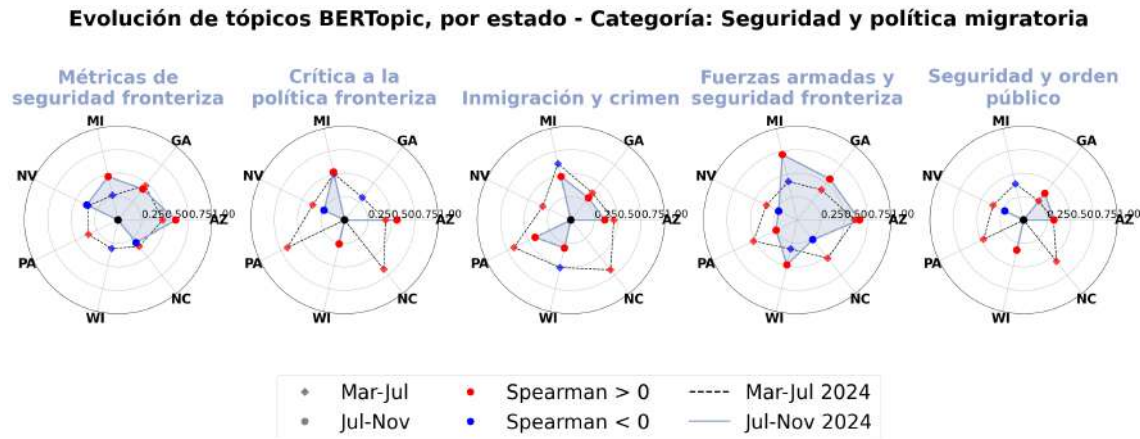


Fig. 4.12: Evolución de correlaciones por tópico, BERTopic. Categoría: Seguridad y política migratoria. Períodos marzo - julio y julio - noviembre de 2024.

Al observar la Figura 4.12 se puede ver que los tópicos relacionados a la política migratoria tuvieron mayor relevancia para el período Trump-Biden. Los tópicos “Fuerzas armadas y seguridad fronteriza” (relacionado a los militares y a la frontera con México), “Inmigración y crimen” (relacionado a prisiones, asilos mentales y terrorismo) y “Crítica a la política fronteriza” (relacionado a la frontera y el cargo conocido como *border czar*) presentaron correlaciones altas para varios estados; particularmente, los últimos dos tópicos mostraron una correlación alta en los estados de Pennsylvania o North Carolina durante el período Trump-Biden mientras que el primer tópico se acentuó más en los estados de Arizona, Michigan y Georgia para el período Trump-Harris. Hay una clara tendencia por parte de este tópico a presentar una mejoría hacia Trump (donde se observan correlaciones positivas y mayores a 0,5 para los casos mencionados).

Una conjetura para la relevancia de temas relacionados a la frontera en el estado de Arizona es que al ser un estado fronterizo con México, la influencia de estos tópicos podría ser mayor. La mejor evidencia de esto es el hecho de que el tópico “Fuerzas armadas y seguridad fronteriza” presentó correlaciones altas (mayor a 0,5) por tres ventanas temporales consecutivas (ver Sección B.2.1).

Por otro lado, el otro swing state cercano a México (en comparación al resto) es Nevada, por lo que la inmigración podría ser un tema interesante ahí. Sin embargo, este estado no tuvo correlaciones fuertes para los tópicos relacionados a esta categoría durante las últimas dos ventanas temporales. El tópico “Métricas de seguridad fronteriza” presentó una correlación alta y favorable a Trump en el período correspondiente a noviembre 2023 - marzo 2024, quedando primero en el orden de tópicos mostrado en la Tabla B.15.

Si bien las interpretaciones pueden seguir una lógica o intuición, estas conclusiones no son inmediatas y requieren un análisis más profundo.

Asuntos sociales y de salud

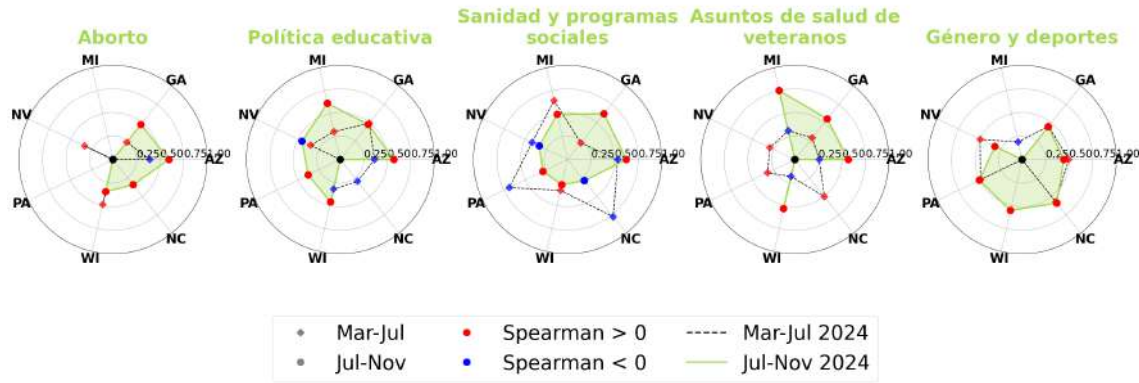
Evolución de tópicos BERTopic, por estado - Categoría: Asuntos sociales y de salud

Fig. 4.13: Evolución de correlaciones por tópico, BERTopic. Categoría: Asuntos sociales y de salud. Períodos marzo - julio y julio - noviembre de 2024.

En la Figura 4.13 se pueden ver las correlaciones obtenidas para los tópicos relacionados a “Asuntos sociales y de salud”.

A primera vista se puede ver que el tópico “Aborto” no fue un tema relevante en comparación al resto, presentando únicamente una correlación superior al 0,5 para el estado de Arizona.

Algo similar ocurre para el tópico de “Política educativa”, donde hay correlaciones superiores a 0,5 únicamente para los estados de Arizona y Michigan. Si bien este tópico está relacionado a la educación superior y a la asistencia con la deuda estudiantil (3.1), no hay indicios de que este tópico haya tenido una relación fuerte a una mejoría para el partido demócrata, que puso en pie planes para perdonar la deuda estudiantil para más de 30 millones de personas [37].

Las correlaciones más altas en valor absoluto se presentaron en los tópicos “Sanidad y programas sociales” y “Asuntos de salud de veteranos”. Para el período Trump-Biden, los asuntos de salud y asistencia social estuvieron relacionados a una mejoría en las encuestas para Biden en los estados de Pennsylvania y North Carolina, mientras que el período Trump-Harris no coincidió con estas tendencias, mostrando mejorías para Trump.

En cuanto al tópico “Asuntos de salud de veteranos, se puede observar que las correlaciones positivas son indicadoras de una mejora en las encuestas para Trump. Esto es de esperar ya que el grupo es predominantemente republicano [38], pero es simplemente una apreciación cualitativa.

Por último, el tópico “Género y deportes” fue otro tema de interés de discusión en la campaña de Trump, oponiéndose fuertemente en contra de la ideología de género. Las correlaciones obtenidas reflejan una mejoría en las encuestas para Trump en todos los estados menos en Michigan, pero los resultados no son indicativos de alguna relación fuerte.

Política exterior y desastres

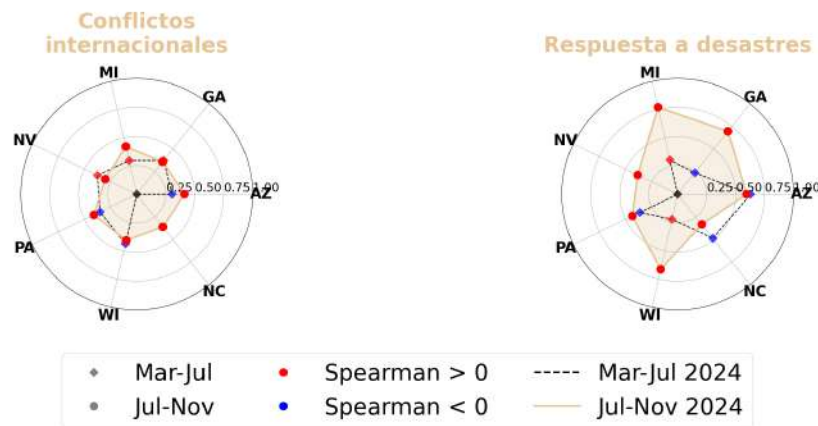
Evolución de tópicos BERTopic, por estado - Categoría: Política exterior y desastres

Fig. 4.14: Evolución de correlaciones por tópico, BERTopic. Categoría: Política exterior y desastres. Períodos marzo - julio y julio - noviembre de 2024.

En la Figura 4.14 se pueden observar los gráficos de radar que explican la evolución de los tópicos relacionados a la “Política exterior y desastres”.

En cuanto al tópico “Conflictos internacionales”, ambos períodos presentan resultados similares, aunque ningún valor de correlación supera 0,5 en valor absoluto. No hay una tendencia para algún estado en particular, esto puede ocurrir ya que este es un tópico relacionado a temas ajenos a la nación y no tiene relevancia local en comparación a otros tópicos. La nube de palabras en la Figura 3.1 muestra para este tópico términos relacionados a la guerra entre Ucrania y Rusia (“Ukraine”, “Putin”, “NATO” (OTAN)), el conflicto entre Israel y Palestina (“jewish”, “Hamas”, “hostages”, “October”, “Iran”) y la guerra en general (“nuclear”, “Middle East”). Si se observan las Tablas correspondientes a las primeras dos ventanas temporales (julio-noviembre 2023 y noviembre 2023 - marzo 2024, B.1 y B.2) el tópico no presenta correlaciones altas a nivel nacional.

Por otro lado, el tópico “Respuesta a desastres” es uno de los más interesantes para evaluar. Los huracanes Idalia, tormentas Lee y Ophelia en 2023 o los huracanes Helene y Milton ocurridos en septiembre y octubre de 2024 pueden ser una causa de los picos vistos en la serie temporal de relevancia de este tópico (Figura 4.9).

Los estados comúnmente afectados por estos eventos son Florida, North Carolina y Georgia, por lo que hay un interés en ver si hay un impacto o relación local del tópico con las encuestas.

Las dos ventanas más interesantes a evaluar para este tópico son las correspondientes a la primera etapa entre Trump y Biden (julio - noviembre de 2023) y la última (entre Trump y Harris, correspondiente al plazo julio - noviembre de 2024). NO hay correlaciones relevantes para estos periodos en North Carolina (ver Sección B.2.4). Para el caso de Georgia, no se encontraron correlaciones significativas en la primera ventana temporal, pero sí para la última (ver Sección B.2.2), donde el tópico fue el más relevante y su correlación indica una mejoría para Trump en las encuestas. Por otro lado, el mismo comportamiento ocurre para esta última ventana en los estados de Michigan, Arizona, Wisconsin y Pennsylvania, pero su interpretación no es inmediata.

Retórica y estrategia política

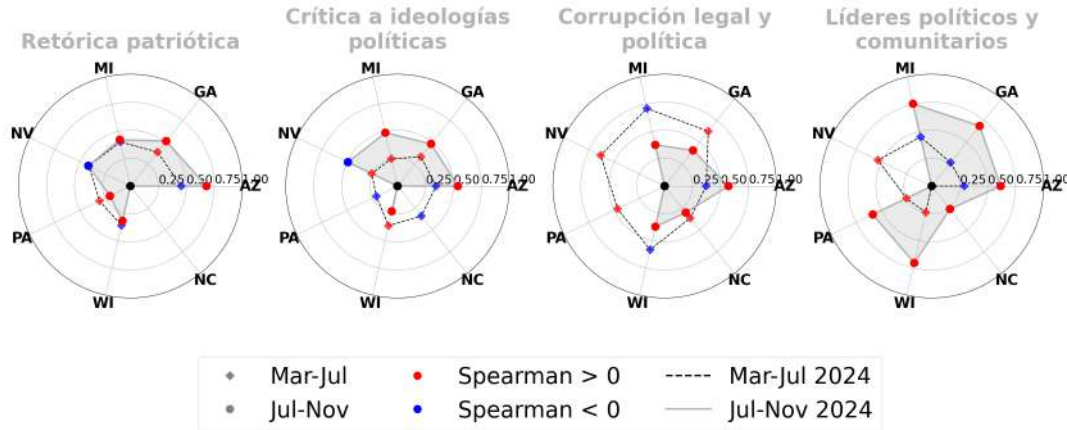
Evolución de tópicos BERTopic, por estado - Categoría: Retórica y estrategia política

Fig. 4.15: Evolución de correlaciones por tópico, BERTopic. Categoría: Retórica y estrategia política. Períodos marzo - julio y julio - noviembre de 2024.

Por último, en la Figura 4.15 se pueden observar las correlaciones obtenidas para los tópicos relacionados a la “Retórica y estrategia política”. Este conjunto de tópicos es el menos interpretable. Las nubes de palabras de estos tópicos (ver Figura 3.1) comprenden términos diversos.

Una consideración a hacer para es que los tópicos como “Retórica patriótica” o “Crítica a ideologías políticas” son parte de estrategias discursivas, mayormente por parte de Trump (ver Figura 3.1).

El tópico “Retórica patriótica” tiene términos como “Make America” (fragmento de eslogan “MAGA (Make America Great Again)”), “ballots” o “America”, que son genéricos. Por otro lado, los picos o subidas en las series temporales de relevancia de este tópico pueden indicar períodos de patriotismo más marcados, por lo que puede ser valioso y puede no ser descartado para el análisis. La relevancia de este tópico se relacionó mejor con las encuestas en Arizona para el último período mientras que las correlaciones fueron más bajas para los otros estados. La tendencia general indica que este tópico favorece a Trump, y esto puede deberse a su carácter discursivo.

El tópico “Líderes políticos y comunitarios” no tuvo correlaciones elevadas para el período entre marzo y julio de 2024, pero para el período entre julio y noviembre de 2024 se puede observar que todos los estados menos North Carolina tuvieron correlaciones entre sus encuestas y la relevancia del tópico mayor a 0,5. Este tópico presenta la mayor dificultad para ser interpretado y necesita más indagación para poder comprenderse mejor.

5. CONCLUSIÓN

5.1. Conclusiones generales

La influencia del discurso público en la opinión electoral ha sido un área de constante investigación, abordada desde múltiples ángulos metodológicos [2, 3, 4, 5, 6]. La creciente disponibilidad de datos textuales provenientes de discursos políticos y la cobertura mediática, junto con los avances en el procesamiento del lenguaje natural (NLP), ofrecen formas para analizar su relación con la opinión pública. En esta tesis, se desarrolló una serie de pasos para investigar las relaciones temporales entre la relevancia de los tópicos en el discurso público y las variaciones en la intención de voto durante la campaña presidencial en los años 2023 y 2024 en Estados Unidos. Específicamente, este trabajo se centró en cuantificar cómo la relevancia de distintos temas, identificados mediante técnicas de modelado de tópicos como BERTopic y clasificaciones predeterminadas externas (Roll Call), se correlaciona con la diferencia en la intención de voto entre Donald Trump y sus oponentes demócratas Joe Biden y Kamala Harris.

Se llevaron a cabo exploraciones del corpus de discursos. Se analizó la distribución geográfica de la frecuencia de los discursos para determinar las regiones y estados donde los principales candidatos concentraron su actividad. Esta distribución se comparó con los resultados electorales históricos de los estados, con el objetivo de verificar si la presencia de discursos se alineaba con aquellos estados considerados de mayor disputa electoral o “swing states”. Adicionalmente, se realizó una clasificación temática del contenido de los discursos para cada candidato, permitiendo identificar los temas predominantes en sus respectivas agendas y ubicaciones donde estos fueron enfatizados. Esta caracterización del panorama discursivo sentó las bases para la posterior evaluación de la correlación entre tópicos específicos y la intención de voto en los estados y a nivel país, buscando patrones temáticos y ofreciendo interpretaciones contextualizadas de las asociaciones encontradas.

Una contribución metodológica clave de este estudio radica en la evaluación rigurosa de estas correlaciones mediante la consideración de desfases temporales (lags) para ver efectos no inmediatos de los tópicos. La comparación sistemática de los resultados de Pearson y Spearman también permitió discernir entre relaciones lineales y monótonas, y evaluar la posible influencia de valores atípicos.

5.2. Conclusiones más específicas y cosas a mejorar o agregar en el futuro

La aplicación de esta metodología permitió extraer las siguientes conclusiones principales:

- **Identificación de tópicos relevantes a nivel nacional y estatal:** Se logró identificar subconjuntos de tópicos (tanto de BERTopic como de RollCall, calculados a partir del corpus nacional de discursos) cuya relevancia temporal mostró correlaciones estadísticamente significativas con la diferencia en la intención de voto. Si bien los tópicos se definieron a nivel nacional, su impacto correlacional se evaluó tanto con encuestas nacionales como con encuestas específicas de estados clave, permitiendo caracterizar cada uno de estos estados por las temáticas nacionales que mayor

asociación presentaron con las fluctuaciones en el sentimiento electoral local. Estos perfiles temáticos ofrecen una visión de las sensibilidades particulares de cada electorado y pueden servir como una base valiosa para futuras investigaciones dentro del marco de SoPhy Lab (DF), orientadas a profundizar en la comprensión de las dinámicas electorales y la influencia temática de los discursos.

- **Correlaciones contextualmente esperables:** En algunos estados, los resultados se alinearon con expectativas contextuales. Por ejemplo, la relevancia del tópico “Respuesta a desastres” mostró correlaciones notables en estados del sudeste del país (una región frecuentemente afectada por eventos climáticos que demandan atención y gestión gubernamental), como el caso de Georgia. Otro ejemplo notable fue el caso del tópico “Industria, manufactura, tecnología y trabajo”, que presentó correlaciones altas en Michigan (estado fuertemente industrial).
- **Asociaciones menos intuitivas y desafíos de interpretación:** En otros estados, se encontraron correlaciones significativas para ciertos tópicos cuya interpretación no resultó inmediata. Esto subraya la complejidad de la opinión pública estatal y sugiere que las dinámicas locales pueden generar asociaciones con temas nacionales que requieren un análisis cualitativo más profundo para comprender plenamente su significado y mecanismo de influencia.
- **Desafíos en la especificidad y densidad de tópicos a nivel estatal:** Un hallazgo metodológico importante fue la dificultad para identificar tópicos suficientemente específicos que se adaptaran perfectamente a las dinámicas particulares de cada estado individual. Los intentos iniciales de generar series temporales de tópicos basadas únicamente en discursos a nivel estatal resultaron en series con una resolución temporal insuficientemente densa para un análisis de correlación robusto. En consecuencia, se optó por usar series de tópicos solamente a nivel nacional, para posteriormente usar con cada serie de encuestas localmente y así usar series más densas.

Si bien este estudio proporciona información valiosa, existen diversos caminos para futuras investigaciones y mejoras que podrían enriquecer la comprensión de la dinámica entre discurso y opinión pública:

- **Resolución y granularidad del modelado de tópicos:** Se podría experimentar con diferentes configuraciones de BERTopic (y otros modelos) para generar representaciones de tópicos con distintas resoluciones. Por un lado, modelos con menos tópicos más genéricos podrían producir series temporales más densas, facilitando el análisis estadístico, aunque a costa de perder especificidad temática. Por otro lado, un mayor número de tópicos más refinados podría capturar mejor matices locales o temas emergentes, pero podría resultar en series temporales más dispersas.
- **Profundización del análisis con sentimiento:** El análisis actual se centra en la relevancia o frecuencia de los tópicos. La incorporación de técnicas de análisis de sentimiento para el contenido de los discursos podría refinar significativamente la comprensión de su impacto. Esto podría aportar una dimensión adicional valiosa para el análisis.
- **Exploración de la causalidad:** Aunque se identificaron correlaciones con lags temporales, la correlación no implica causalidad. Sería valioso aplicar tests estadísticos

más formales, como el test de causalidad de Granger [39], para investigar el carácter predictivo de cada tópico en la intención de voto. Los plazos comprendidos para la realización del trabajo no fueron suficientes para la elaboración de dicho test, pero se podrían explorar causalidades en trabajos futuros.

- **Ampliación de fuentes de datos textuales:** El análisis podría enriquecerse incorporando un corpus de texto más diverso y voluminoso, incluyendo, por ejemplo, publicaciones en redes sociales (Twitter/X, Facebook) o artículos de noticias de un espectro más amplio de medios.
- **Aplicabilidad y extrapolación geográfica:** La metodología desarrollada es inherentemente extrapolable a contextos electorales en otras partes del mundo. Sería de gran interés aplicar un análisis similar para evaluar la dinámica de los tópicos y su relación con la opinión pública en campañas presidenciales en otros países, como por ejemplo Argentina, siempre que se disponga de los datos adecuados de discursos y encuestas.

Apéndice

A. LISTADOS DE TÓPICOS Y VALIDACIÓN

A.1. Tópicos de Roll Call

En la Tabla A.1 se presenta la lista completa de los 17 tópicos de primer nivel provistos por Roll Call, junto con su traducción al español.

Tab. A.1: Tópicos de primer nivel de Roll Call y su traducción.

Tópico original (inglés)	Traducción (español)
Arts, Culture and Entertainment	Artes, cultura y entretenimiento
Crime, Law and Justice	Crimen, ley y justicia
Disaster and Accident	Desastres y accidentes
Economy, Business and Finance	Economía, negocios y finanzas
Education	Educación
Environmental Issue	Asuntos medioambientales
Health	Salud
Human Interest	Interés general
Labor	Trabajo / asuntos laborales
Lifestyle and Leisure	Estilo de vida y ocio
Politics	Política
Religion and Belief	Religión y creencias
Science and Technology	Ciencia y tecnología
Social Issue	Asuntos sociales
Sport	Deporte
Unrest, Conflicts and War	Disturbios, conflictos y guerra
Weather	Clima / meteorología

A.2. Tópicos generados por BERTopic

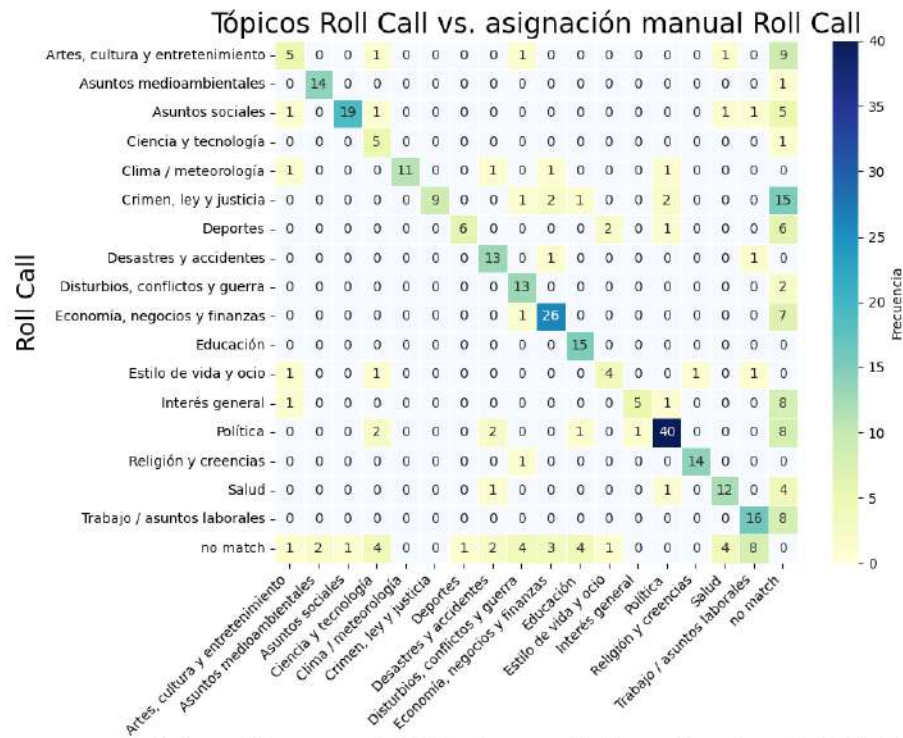
En la Tabla A.2 se presenta la lista de los 20 tópicos identificados mediante BERTopic.

Tab. A.2: Tópicos de BERTopic.

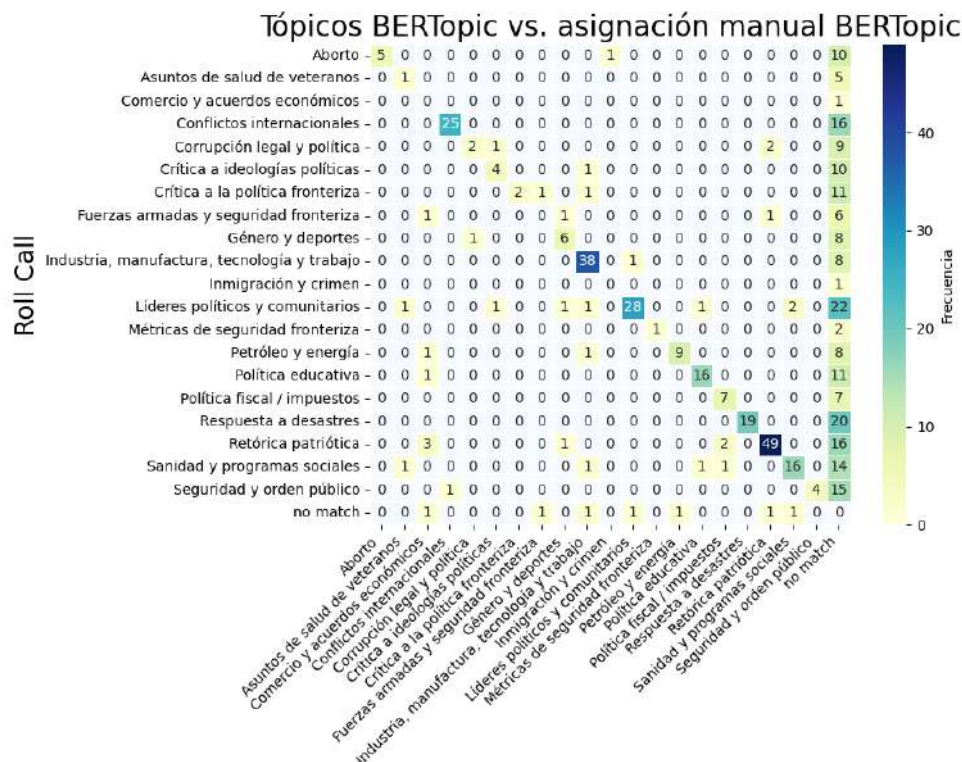
Tópico
Aborto
Métricas de seguridad fronteriza
Crítica a la política fronteriza
Respuesta a desastres
Política educativa
Género y deportes
Sanidad y programas sociales
Inmigración y crimen
Industria, manufactura, tecnología y trabajo
Conflictos internacionales
Corrupción legal y política
Fuerzas armadas y seguridad fronteriza
Petróleo y energía
Retórica patriótica
Líderes políticos y comunitarios
Crítica a ideologías políticas
Seguridad y orden público
Política fiscal / impuestos
Comercio y acuerdos económicos
Asuntos y salud de veteranos

A.3. Validación de tópicos

En la Figura A.1 se presentan las matrices de confusión globales resultantes de los tres experimentos de validación manual descritos en la Sección 3.3.2.



Asignación manual BERTopic, asociada a etiquetas de Roll Cal



Asignación manual BERTopic, asociada a etiquetas de Roll C

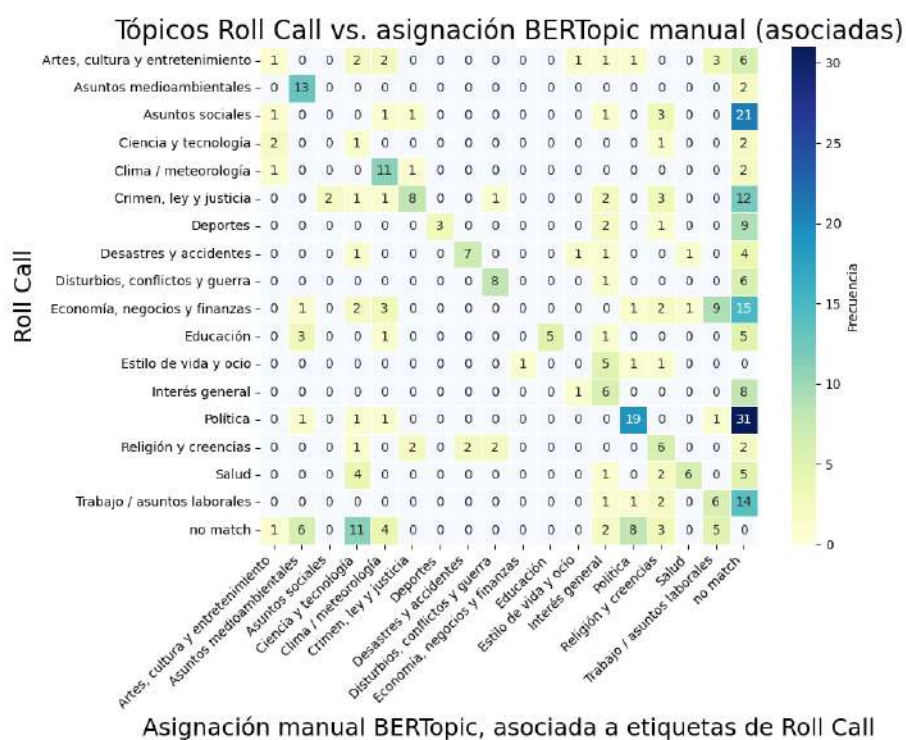


Fig. A.1: Matrices de confusión para los tres experimentos de validación. **Arriba:** validación de consistencia interna de Roll Call (original vs. manual). **Medio:** validación de consistencia interna de BERTopic (original vs. manual). **Abajo:** validación de acuerdo intermetodología (Roll Call vs. BERTopic mapeado).

B. TABLAS DE CORRELACIONES

En este capítulo se muestran tablas de correlación obtenidas para todas las ventanas temporales (para todo el trabajo de tesis). En la Sección 4.5 se priorizó analizar las últimas dos ventanas electorales por proximidad a la fecha de elecciones y para ver los últimos meses de la rivalidad Trump-Biden y Trump-Harris. Se reportan las tablas de correlaciones a modo de material extra y complementario al análisis hecho en la Sección 4.4.

Siguiendo el criterio de filtrado de series temporales mencionado en la Sección 4.2, para las etapas “temprana” y “media” se consideraron 4 y 3 tópicos de Roll Call para el análisis de correlaciones, respectivamente. Los tópicos correspondientes a la etapa temprana entre Trump y Biden son “Política”, “Economía, negocios y finanzas”, “Asuntos sociales” y “Crimen, ley y justicia”; los tópicos de la etapa media son “Política”, “Crimen, ley y justicia” y “Asuntos sociales”.

B.1. Nivel nacional

Tab. B.1: Correlaciones tópico - (Trump - Biden) a nivel nacional (Julio - Noviembre 2023).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Respuesta a desastres	0,68	0,60	(L:28, P _S :,002, P _P :,002)
Política fiscal / impuestos	0,58	0,58	(L:21, P _S :,002, P _P :,002)
Política educativa	0,58	0,53	(L:18, P _S :,002, P _P :,002)
Seguridad y orden público	0,54	0,54	(L:21, P _S :,002, P _P :,002)
Inmigración y crimen	0,48	0,39	(L:14, P _S :,002, P _P :,002)
Crítica a ideologías políticas	0,45	0,39	(L:20, P _S :,002, P _P :,002)
Aborto	0,43	0,48	(L:13, P _S :,002, P _P :,002)
Corrupción legal y política	0,42	0,34	(L:7, P _S :,002, P _P :,002)
Crítica a la política fronteriza	0,39	0,37	(L:17, P _S :,004, P _P :,002)
Conflictos internacionales	0,38	0,42	(L:7, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	0,34	0,32	(L:24, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	-0,33	-0,32	(L:28, P _S :,002, P _P :,004)
Fuerzas armadas y seguridad fronteriza	0,31	0,33	(L:7, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	0,27	0,33	(L:7, P _S :,006, P _P :,002)
Retórica patriótica	0,27	0,27	(L:21, P _S :,012, P _P :,006)
Sanidad y programas sociales	0,25	0,25	(L:25, P _S :,010, P _P :,008)
Género y deportes	0,23	0,20	(L:12, P _S :,036, P _P :,034)
<i>Tópicos Roll Call</i>			
Política	0,46	0,56	(L:7, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,44	0,46	(L:25, P _S :,002, P _P :,002)
Asuntos sociales	0,28	0,23	(L:22, P _S :,010, P _P :,018)
Economía, negocios y finanzas	-0,20	-0,22	(L:16, P _S :,030, P _P :,026)

Tab. B.2: Correlaciones tópico - (Trump - Biden) a nivel nacional (Noviembre 2023 - Marzo 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Sanidad y programas sociales	-0,56	-0,59	(L:27, P _S :,002, P _P :,002)
Corrupción legal y política	0,55	0,72	(L:15, P _S :,002, P _P :,002)
Métricas de seguridad fronteriza	-0,53	-0,50	(L:28, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	0,49	0,43	(L:15, P _S :,002, P _P :,002)
Inmigración y crimen	-0,49	-0,46	(L:25, P _S :,002, P _P :,002)
Aborto	-0,49	-0,52	(L:24, P _S :,002, P _P :,002)
Respuesta a desastres	0,48	0,49	(L:16, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,47	-0,39	(L:24, P _S :,002, P _P :,002)
Crítica a ideologías políticas	-0,46	-0,29	(L:23, P _S :,002, P _P :,004)
Crítica a la política fronteriza	0,46	0,44	(L:16, P _S :,002, P _P :,002)
Petróleo y energía	-0,45	-0,47	(L:28, P _S :,002, P _P :,002)
Conflictos internacionales	0,40	0,23	(L:12, P _S :,002, P _P :,020)
Líderes políticos y comunitarios	-0,39	-0,38	(L:20, P _S :,002, P _P :,002)
Política fiscal / impuestos	-0,36	-0,47	(L:25, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	-0,22	-0,23	(L:7, P _S :,030, P _P :,010)
Género y deportes	-0,21	-0,22	(L:20, P _S :,042, P _P :,036)
<i>Tópicos Roll Call</i>			
Economía, negocios y finanzas	-0,58	-0,61	(L:24, P _S :,002, P _P :,002)
Política	-0,55	-0,59	(L:26, P _S :,002, P _P :,002)
Crimen, ley y justicia	-0,49	-0,47	(L:25, P _S :,002, P _P :,002)

B.2. Nivel estatal

B.2.1. Arizona

No se encontraron correlaciones significativas para la primera ventana temporal.

Tab. B.3: Correlaciones t3pico - (Trump - Biden), Arizona (Noviembre 2023 - Marzo 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
M3tricas de seguridad fronteriza	-0,69	-0,60	(L:28, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,53	-0,47	(L:26, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	-0,50	-0,57	(L:28, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	-0,48	-0,22	(L:14, P _S :,002, P _P :,014)
Pol3tica fiscal / impuestos	-0,46	-0,41	(L:17, P _S :,002, P _P :,004)
Aborto	-0,40	-0,38	(L:13, P _S :,002, P _P :,002)
Corrupci3n legal y pol3tica	-0,38	-0,45	(L:28, P _S :,002, P _P :,004)
Ret3rica patri3tica	-0,36	-0,24	(L:16, P _S :,002, P _P :,020)
Industria, manufactura, tecnolog3a y trabajo	0,34	0,23	(L:28, P _S :,006, P _P :,030)
Respuesta a desastres	0,31	0,32	(L:14, P _S :,004, P _P :,004)
Seguridad y orden p3blico	-0,30	-0,25	(L:17, P _S :,004, P _P :,008)
Asuntos de salud de veteranos	-0,29	-0,20	(L:7, P _S :,002, P _P :,040)
L3deres pol3ticos y comunitarios	0,21	0,22	(L:28, P _S :,042, P _P :,010)
G3nero y deportes	-0,20	-0,27	(L:8, P _S :,048, P _P :,006)
<i>T3picos Roll Call</i>			
Pol3tica	-0,50	-0,43	(L:26, P _S :,002, P _P :,002)
Crimen, ley y justicia	-0,39	-0,40	(L:27, P _S :,002, P _P :,002)

Tab. B.4: Correlaciones t3pico - (Trump - Biden), Arizona (Marzo - Julio 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Respuesta a desastres	-0,63	-0,45	(L:25, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,61	0,55	(L:23, P _S :,002, P _P :,002)
Sanidad y programas sociales	-0,53	-0,42	(L:25, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,51	-0,44	(L:26, P _S :,002, P _P :,002)
Pol3tica fiscal / impuestos	0,51	0,53	(L:8, P _S :,002, P _P :,002)
G3nero y deportes	0,49	0,38	(L:12, P _S :,002, P _P :,002)
M3tricas de seguridad fronteriza	0,48	0,49	(L:20, P _S :,002, P _P :,002)
Inmigraci3n y crimen	0,46	0,42	(L:28, P _S :,002, P _P :,002)
Ret3rica patri3tica	-0,45	-0,41	(L:26, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	0,44	0,43	(L:22, P _S :,002, P _P :,002)
Aborto	-0,39	-0,33	(L:28, P _S :,002, P _P :,002)
Corrupci3n legal y pol3tica	-0,37	-0,35	(L:7, P _S :,002, P _P :,002)
Pol3tica educativa	-0,36	-0,30	(L:28, P _S :,002, P _P :,008)
Cr3tica a ideolog3as pol3ticas	-0,34	-0,26	(L:27, P _S :,002, P _P :,008)
Seguridad y orden p3blico	0,32	0,33	(L:26, P _S :,002, P _P :,002)
Conflictos internacionales	-0,30	-0,23	(L:26, P _S :,004, P _P :,016)
L3deres pol3ticos y comunitarios	-0,29	-0,31	(L:24, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	-0,26	-0,40	(L:7, P _S :,002, P _P :,002)
Petr3leo y energ3a	-0,24	-0,31	(L:28, P _S :,010, P _P :,002)
Comercio y acuerdos econ3micos	0,18	0,18	(L:7, P _S :,036, P _P :,038)
<i>T3picos Roll Call</i>			
Econom3a, negocios y finanzas	-0,47	-0,48	(L:28, P _S :,002, P _P :,002)
Crimen, ley y justicia	-0,31	-0,41	(L:25, P _S :,002, P _P :,002)
Asuntos sociales	0,31	0,30	(L:11, P _S :,002, P _P :,002)
Pol3tica	-0,25	-0,26	(L:24, P _S :,012, P _P :,008)

Tab. B.5: Correlaciones tópico - (Trump - Harris), Arizona (Julio - Noviembre 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Política fiscal / impuestos	0,72	0,73	(L:19, P _S :,002, P _P :,002)
Retórica patriótica	0,68	0,73	(L:16, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,66	0,71	(L:16, P _S :,002, P _P :,002)
Sanidad y programas sociales	0,62	0,66	(L:20, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	0,62	0,69	(L:15, P _S :,002, P _P :,002)
Métricas de seguridad fronteriza	0,61	0,51	(L:14, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	0,60	0,55	(L:24, P _S :,002, P _P :,002)
Aborto	0,60	0,59	(L:9, P _S :,002, P _P :,002)
Respuesta a desastres	0,59	0,62	(L:24, P _S :,002, P _P :,002)
Política educativa	0,57	0,53	(L:14, P _S :,002, P _P :,002)
Corrupción legal y política	0,57	0,62	(L:16, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,57	0,68	(L:24, P _S :,002, P _P :,002)
Crítica a la política fronteriza	0,56	0,57	(L:16, P _S :,002, P _P :,002)
Crítica a ideologías políticas	0,54	0,57	(L:15, P _S :,002, P _P :,002)
Género y deportes	0,44	0,60	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	0,41	0,40	(L:7, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	0,37	0,39	(L:19, P _S :,004, P _P :,004)
Inmigración y crimen	0,36	0,46	(L:19, P _S :,004, P _P :,002)
Seguridad y orden público	0,32	0,34	(L:12, P _S :,004, P _P :,002)
Petróleo y energía	0,30	0,35	(L:19, P _S :,004, P _P :,004)
<i>Tópicos Roll Call</i>			
Política	0,62	0,69	(L:16, P _S :,002, P _P :,002)
Asuntos sociales	0,51	0,62	(L:24, P _S :,002, P _P :,002)
Trabajo / asuntos laborales	0,49	0,44	(L:22, P _S :,002, P _P :,002)
Disturbios, conflictos y guerra	0,47	0,43	(L:7, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,45	0,48	(L:22, P _S :,002, P _P :,002)
Salud	0,38	0,35	(L:14, P _S :,002, P _P :,004)
Economía, negocios y finanzas	0,34	0,36	(L:20, P _S :,006, P _P :,002)

B.2.2. Georgia

No se encontraron correlaciones significativas para la primera ventana temporal.

Tab. B.6: Correlaciones t3pico - (Trump - Biden), Georgia (Noviembre 2023 - Marzo 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
M3tricas de seguridad fronteriza	0,74	0,67	(L:13, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,58	0,42	(L:27, P _S :,002, P _P :,002)
Ret3rica patri3tica	0,51	0,32	(L:13, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,50	-0,48	(L:28, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	0,47	0,37	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	-0,46	-0,50	(L:23, P _S :,002, P _P :,002)
Pol3tica fiscal / impuestos	0,42	0,38	(L:7, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	0,42	0,36	(L:11, P _S :,002, P _P :,002)
Respuesta a desastres	-0,32	-0,23	(L:27, P _S :,004, P _P :,038)
Pol3tica educativa	-0,28	-0,42	(L:28, P _S :,006, P _P :,002)
Corrupci3n legal y pol3tica	0,24	0,25	(L:28, P _S :,028, P _P :,020)
Aborto	0,24	0,22	(L:12, P _S :,010, P _P :,020)
<i>T3picos Roll Call</i>			
Pol3tica	0,54	0,56	(L:13, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,44	0,50	(L:8, P _S :,002, P _P :,002)

Tab. B.7: Correlaciones t3pico - (Trump - Biden), Georgia (Marzo - Julio 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Corrupci3n legal y pol3tica	0,63	0,56	(L:28, P _S :,002, P _P :,002)
Pol3tica educativa	0,49	0,51	(L:14, P _S :,002, P _P :,002)
M3tricas de seguridad fronteriza	0,47	0,44	(L:14, P _S :,002, P _P :,002)
G3nero y deportes	0,45	0,42	(L:19, P _S :,002, P _P :,002)
Petr3leo y energ3a	0,43	0,36	(L:16, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,41	0,21	(L:20, P _S :,002, P _P :,014)
Comercio y acuerdos econ3micos	0,40	0,47	(L:13, P _S :,002, P _P :,002)
Ret3rica patri3tica	0,39	0,38	(L:19, P _S :,002, P _P :,002)
Conflictos internacionales	0,37	0,29	(L:14, P _S :,002, P _P :,002)
Inmigraci3n y crimen	0,37	0,30	(L:14, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	0,34	0,38	(L:19, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	-0,31	-0,31	(L:7, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,30	0,29	(L:15, P _S :,002, P _P :,004)
L3deres pol3ticos y comunitarios	-0,27	-0,35	(L:7, P _S :,008, P _P :,002)
Seguridad y orden p3blico	0,26	0,23	(L:19, P _S :,004, P _P :,016)
Respuesta a desastres	-0,24	-0,27	(L:7, P _S :,002, P _P :,002)
Aborto	0,24	0,30	(L:22, P _S :,018, P _P :,006)
Sanidad y programas sociales	0,23	0,36	(L:7, P _S :,012, P _P :,002)
<i>T3picos Roll Call</i>			
Crimen, ley y justicia	0,54	0,59	(L:28, P _S :,002, P _P :,002)
Pol3tica	0,52	0,54	(L:18, P _S :,002, P _P :,002)
Asuntos sociales	-0,37	-0,30	(L:22, P _S :,002, P _P :,004)
Econom3a, negocios y finanzas	0,35	0,47	(L:24, P _S :,002, P _P :,002)

Tab. B.8: Correlaciones tópico - (Trump - Harris), Georgia (Julio - Noviembre 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Respuesta a desastres	0,70	0,62	(L:24, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	0,69	0,71	(L:12, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	0,66	0,61	(L:26, P _S :,002, P _P :,002)
Política fiscal / impuestos	0,64	0,68	(L:14, P _S :,002, P _P :,002)
Sanidad y programas sociales	0,62	0,68	(L:11, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,56	0,51	(L:26, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,55	0,61	(L:26, P _S :,002, P _P :,002)
Retórica patriótica	0,51	0,55	(L:10, P _S :,002, P _P :,002)
Crítica a ideologías políticas	0,48	0,47	(L:7, P _S :,002, P _P :,002)
Política educativa	0,48	0,43	(L:8, P _S :,002, P _P :,002)
Aborto	0,48	0,53	(L:7, P _S :,002, P _P :,002)
Género y deportes	0,44	0,40	(L:7, P _S :,002, P _P :,002)
Métricas de seguridad fronteriza	0,42	0,28	(L:7, P _S :,002, P _P :,004)
Corrupción legal y política	0,41	0,43	(L:16, P _S :,002, P _P :,002)
Seguridad y orden público	0,36	0,42	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	0,36	0,37	(L:7, P _S :,002, P _P :,002)
Petróleo y energía	0,32	0,33	(L:9, P _S :,004, P _P :,006)
Inmigración y crimen	0,30	0,26	(L:7, P _S :,010, P _P :,020)
Comercio y acuerdos económicos	0,26	0,25	(L:10, P _S :,006, P _P :,026)
<i>Tópicos Roll Call</i>			
Política	0,56	0,59	(L:8, P _S :,002, P _P :,002)
Salud	0,55	0,53	(L:11, P _S :,002, P _P :,002)
Trabajo / asuntos laborales	0,51	0,55	(L:24, P _S :,002, P _P :,002)
Disturbios, conflictos y guerra	0,50	0,41	(L:7, P _S :,002, P _P :,002)
Economía, negocios y finanzas	0,48	0,50	(L:9, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,29	0,28	(L:10, P _S :,004, P _P :,008)
Asuntos sociales	0,29	0,31	(L:26, P _S :,006, P _P :,008)

B.2.3. Michigan

Tab. B.9: Correlaciones t3pico - (Trump - Biden), Michigan (Julio - Noviembre 2023).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Ret3rica patri3tica	-0,58	-0,56	(L:9, P _S :,002, P _P :,002)
Pol3tica educativa	-0,51	-0,47	(L:7, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	-0,47	-0,41	(L:7, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,41	-0,37	(L:10, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	-0,37	-0,32	(L:7, P _S :,002, P _P :,004)
Pol3tica fiscal / impuestos	-0,35	-0,28	(L:7, P _S :,004, P _P :,014)
Aborto	-0,31	-0,32	(L:7, P _S :,010, P _P :,006)
G3nero y deportes	0,29	0,28	(L:7, P _S :,010, P _P :,012)
Asuntos de salud de veteranos	-0,29	-0,23	(L:10, P _S :,008, P _P :,044)
Corrupci3n legal y pol3tica	0,27	0,45	(L:10, P _S :,026, P _P :,002)
Conflictos internacionales	0,26	0,30	(L:10, P _S :,020, P _P :,010)

Tab. B.10: Correlaciones t3pico - (Trump - Biden), Michigan (Noviembre 2023 - Marzo 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Industria, manufactura, tecnolog3a y trabajo	-0,78	-0,72	(L:23, P _S :,002, P _P :,002)
Pol3tica educativa	-0,64	-0,69	(L:16, P _S :,002, P _P :,002)
L3deres pol3ticos y comunitarios	-0,58	-0,62	(L:18, P _S :,002, P _P :,002)
G3nero y deportes	-0,56	-0,61	(L:19, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	-0,56	-0,81	(L:26, P _S :,002, P _P :,002)
Pol3tica fiscal / impuestos	-0,56	-0,26	(L:28, P _S :,002, P _P :,012)
Ret3rica patri3tica	-0,54	-0,32	(L:28, P _S :,002, P _P :,002)
Respuesta a desastres	-0,53	-0,34	(L:21, P _S :,002, P _P :,002)
Seguridad y orden p3blico	-0,52	-0,45	(L:18, P _S :,002, P _P :,002)
Petr3leo y energ3a	-0,49	-0,62	(L:28, P _S :,002, P _P :,002)
M3tricas de seguridad fronteriza	0,40	0,30	(L:7, P _S :,002, P _P :,002)
Inmigraci3n y crimen	-0,39	-0,40	(L:18, P _S :,002, P _P :,002)
Comercio y acuerdos econ3micos	-0,31	-0,56	(L:28, P _S :,008, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,27	-0,41	(L:28, P _S :,010, P _P :,004)
Cr3tica a ideolog3as pol3ticas	-0,24	-0,51	(L:26, P _S :,020, P _P :,002)
<i>T3picos Roll Call</i>			
Econom3a, negocios y finanzas	-0,41	-0,52	(L:24, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,31	0,37	(L:7, P _S :,002, P _P :,002)
Pol3tica	-0,27	-0,27	(L:28, P _S :,010, P _P :,004)

Tab. B.11: Correlaciones tópico - (Trump - Biden), Michigan (Marzo - Julio 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Corrupción legal y política	-0,71	-0,65	(L:25, P _S :.002, P _P :.002)
Sanidad y programas sociales	0,64	0,68	(L:17, P _S :.002, P _P :.002)
Inmigración y crimen	-0,62	-0,54	(L:8, P _S :.002, P _P :.002)
Crítica a la política fronteriza	-0,50	-0,54	(L:7, P _S :.002, P _P :.002)
Política fiscal / impuestos	0,47	0,33	(L:10, P _S :.002, P _P :.004)
Líderes políticos y comunitarios	-0,45	-0,43	(L:25, P _S :.002, P _P :.002)
Fuerzas armadas y seguridad fronteriza	-0,42	-0,40	(L:7, P _S :.002, P _P :.002)
Retórica patriótica	-0,41	-0,36	(L:28, P _S :.002, P _P :.002)
Seguridad y orden público	-0,40	-0,39	(L:17, P _S :.002, P _P :.002)
Industria, manufactura, tecnología y trabajo	-0,34	-0,34	(L:26, P _S :.002, P _P :.002)
Asuntos de salud de veteranos	-0,31	-0,32	(L:21, P _S :.006, P _P :.002)
Respuesta a desastres	0,30	0,27	(L:15, P _S :.002, P _P :.006)
Política educativa	0,30	0,26	(L:12, P _S :.002, P _P :.010)
Petróleo y energía	-0,30	-0,33	(L:24, P _S :.004, P _P :.002)
Conflictos internacionales	0,30	0,27	(L:7, P _S :.002, P _P :.004)
Métricas de seguridad fronteriza	-0,27	-0,21	(L:7, P _S :.004, P _P :.010)
Crítica a ideologías políticas	0,25	0,37	(L:18, P _S :.010, P _P :.002)
Género y deportes	-0,19	-0,18	(L:7, P _S :.020, P _P :.040)
<i>Tópicos Roll Call</i>			
Crimen, ley y justicia	-0,72	-0,64	(L:28, P _S :.002, P _P :.002)
Economía, negocios y finanzas	0,35	0,19	(L:7, P _S :.002, P _P :.038)
Política	-0,35	-0,28	(L:25, P _S :.002, P _P :.002)

Tab. B.12: Correlaciones t pico - (Trump - Harris), Michigan (Julio - Noviembre 2024).

T�pico	Spearman	Pearson	Detalles
<i>T�picos BERTopic</i>			
Industria, manufactura, tecnolog�a y trabajo	0,84	0,69	(L:13, P _S :,002, P _P :,002)
Respuesta a desastres	0,77	0,80	(L:7, P _S :,002, P _P :,002)
L�deres pol�ticos y comunitarios	0,75	0,65	(L:25, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,75	0,53	(L:16, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,72	0,57	(L:11, P _S :,002, P _P :,002)
Petr�leo y energ�a	0,69	0,65	(L:21, P _S :,002, P _P :,002)
Comercio y acuerdos econ�micos	0,62	0,44	(L:19, P _S :,002, P _P :,002)
Pol�tica educativa	0,61	0,52	(L:24, P _S :,002, P _P :,002)
Pol�tica fiscal / impuestos	0,56	0,46	(L:7, P _S :,002, P _P :,002)
Cr�tica a la pol�tica fronteriza	0,53	0,38	(L:7, P _S :,002, P _P :,002)
Sanidad y programas sociales	0,49	0,42	(L:25, P _S :,002, P _P :,002)
Cr�tica a ideolog�as pol�ticas	0,49	0,41	(L:24, P _S :,002, P _P :,002)
Inmigraci�n y crimen	0,48	0,39	(L:14, P _S :,002, P _P :,002)
M�tricas de seguridad fronteriza	0,48	0,27	(L:23, P _S :,002, P _P :,022)
Ret�rica patri�tica	0,42	0,32	(L:7, P _S :,002, P _P :,004)
Conflictos internacionales	0,42	0,41	(L:13, P _S :,002, P _P :,002)
Corrupci�n legal y pol�tica	0,38	0,40	(L:7, P _S :,002, P _P :,002)
<i>T�picos Roll Call</i>			
Trabajo / asuntos laborales	0,69	0,74	(L:14, P _S :,002, P _P :,002)
Asuntos sociales	0,68	0,59	(L:16, P _S :,002, P _P :,002)
Econom�a, negocios y finanzas	0,65	0,66	(L:25, P _S :,002, P _P :,002)
Pol�tica	0,59	0,41	(L:13, P _S :,002, P _P :,002)
Disturbios, conflictos y guerra	0,56	0,49	(L:12, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,51	0,37	(L:24, P _S :,002, P _P :,002)
Salud	0,50	0,46	(L:25, P _S :,002, P _P :,002)

B.2.4. North Carolina

No se encontraron correlaciones significativas para ambas ventanas temporales.

Tab. B.13: Correlaciones t3pico - (Trump - Biden), North Carolina (Marzo - Julio 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Sanidad y programas sociales	-0,78	-0,70	(L:21, P _S :,002, P _P :,002)
Inmigraci3n y crimen	0,68	0,69	(L:28, P _S :,002, P _P :,002)
Cr3tica a la pol3tica fronteriza	0,67	0,66	(L:23, P _S :,002, P _P :,002)
G3nero y deportes	0,57	0,46	(L:7, P _S :,002, P _P :,002)
Seguridad y orden p3blico	0,56	0,54	(L:28, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,52	0,46	(L:10, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,50	0,33	(L:23, P _S :,002, P _P :,002)
Respuesta a desastres	-0,49	-0,32	(L:16, P _S :,002, P _P :,002)
Pol3tica fiscal / impuestos	-0,43	-0,44	(L:28, P _S :,002, P _P :,002)
Petr3leo y energ3a	0,41	0,18	(L:7, P _S :,002, P _P :,022)
Corrupci3n legal y pol3tica	0,37	0,28	(L:27, P _S :,002, P _P :,004)
M3tricas de seguridad fronteriza	0,36	0,41	(L:23, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	-0,34	-0,36	(L:20, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,30	-0,21	(L:11, P _S :,002, P _P :,030)
Pol3tica educativa	-0,29	-0,27	(L:28, P _S :,002, P _P :,006)
Comercio y acuerdos econ3micos	0,24	0,26	(L:15, P _S :,008, P _P :,004)
<i>T3picos Roll Call</i>			
Econom3a, negocios y finanzas	-0,47	-0,40	(L:28, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,25	0,22	(L:27, P _S :,008, P _P :,032)

Tab. B.14: Correlaciones t3pico - (Trump - Harris), North Carolina (Julio - Noviembre 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
G3nero y deportes	0,59	0,70	(L:18, P _S :,002, P _P :,002)
Pol3tica fiscal / impuestos	-0,36	-0,33	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	0,36	0,27	(L:24, P _S :,004, P _P :,016)
Aborto	0,34	0,39	(L:15, P _S :,008, P _P :,002)
Respuesta a desastres	0,33	0,46	(L:23, P _S :,004, P _P :,002)
M3tricas de seguridad fronteriza	-0,31	-0,34	(L:7, P _S :,006, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,31	-0,35	(L:9, P _S :,002, P _P :,002)
Corrupci3n legal y pol3tica	0,30	0,35	(L:19, P _S :,004, P _P :,002)
Sanidad y programas sociales	-0,29	-0,33	(L:9, P _S :,006, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,27	-0,34	(L:7, P _S :,008, P _P :,002)
L3deres pol3ticos y comunitarios	0,26	0,29	(L:19, P _S :,008, P _P :,012)
<i>T3picos Roll Call</i>			
Asuntos sociales	-0,47	-0,42	(L:13, P _S :,002, P _P :,002)
Disturbios, conflictos y guerra	0,42	0,41	(L:18, P _S :,002, P _P :,002)
Trabajo / asuntos laborales	0,29	0,24	(L:21, P _S :,016, P _P :,038)
Econom3a, negocios y finanzas	-0,26	-0,39	(L:8, P _S :,012, P _P :,002)
Salud	0,24	0,29	(L:23, P _S :,042, P _P :,018)

B.2.5. Nevada

No se encontraron correlaciones significativas para la primera ventana temporal.

Tab. B.15: Correlaciones tópico - (Trump - Biden), Nevada (Noviembre 2023 - Marzo 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Métricas de seguridad fronteriza	0,64	0,50	(L:7, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	-0,60	-0,44	(L:23, P _S :,002, P _P :,002)
Respuesta a desastres	-0,43	-0,27	(L:26, P _S :,002, P _P :,008)
Fuerzas armadas y seguridad fronteriza	0,37	0,36	(L:28, P _S :,002, P _P :,002)
Crítica a ideologías políticas	0,35	0,43	(L:16, P _S :,002, P _P :,002)
Conflictos internacionales	-0,33	-0,28	(L:24, P _S :,002, P _P :,012)
Política educativa	-0,32	-0,40	(L:28, P _S :,004, P _P :,002)
Crítica a la política fronteriza	0,30	0,24	(L:7, P _S :,002, P _P :,014)
Corrupción legal y política	0,28	0,19	(L:7, P _S :,004, P _P :,044)
Líderes políticos y comunitarios	-0,28	-0,28	(L:10, P _S :,006, P _P :,008)
Género y deportes	-0,23	-0,25	(L:28, P _S :,020, P _P :,022)
Aborto	0,23	0,26	(L:24, P _S :,022, P _P :,008)
Petróleo y energía	-0,22	-0,21	(L:26, P _S :,036, P _P :,032)
Comercio y acuerdos económicos	-0,22	-0,24	(L:14, P _S :,030, P _P :,016)
Asuntos de salud de veteranos	0,19	-0,19	(L:8, P _S :,044, P _P :,040)
<i>Tópicos Roll Call</i>			
Crimen, ley y justicia	0,48	0,53	(L:9, P _S :,002, P _P :,002)
Política	0,42	0,42	(L:12, P _S :,002, P _P :,002)
Economía, negocios y finanzas	-0,31	-0,25	(L:14, P _S :,002, P _P :,012)

Tab. B.16: Correlaciones t pico - (Trump - Biden), Nevada (Marzo - Julio 2024).

T�pico	Spearman	Pearson	Detalles
<i>T�picos BERTopic</i>			
Corrupci�n legal y pol�tica	0,63	0,51	(L:17, P _S :,002, P _P :,002)
L�deres pol�ticos y comunitarios	0,53	0,49	(L:28, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog�a y trabajo	0,52	0,52	(L:27, P _S :,002, P _P :,002)
G�nero y deportes	0,49	0,51	(L:7, P _S :,002, P _P :,002)
Petr�leo y energ�a	0,44	0,38	(L:7, P _S :,002, P _P :,002)
Sanidad y programas sociales	-0,42	-0,55	(L:18, P _S :,002, P _P :,002)
Ret�rica patri�tica	0,41	0,37	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	0,38	0,32	(L:7, P _S :,002, P _P :,002)
Cr�tica a la pol�tica fronteriza	0,37	0,39	(L:17, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,36	0,32	(L:7, P _S :,002, P _P :,002)
Seguridad y orden p�blico	0,36	0,38	(L:13, P _S :,002, P _P :,002)
M�tricas de seguridad fronteriza	-0,35	-0,41	(L:26, P _S :,002, P _P :,002)
Pol�tica educativa	0,35	0,49	(L:7, P _S :,002, P _P :,002)
Aborto	0,33	0,26	(L:7, P _S :,002, P _P :,006)
Inmigraci�n y crimen	0,33	0,27	(L:13, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,29	0,24	(L:7, P _S :,004, P _P :,012)
Cr�tica a ideolog�as pol�ticas	0,26	0,23	(L:7, P _S :,002, P _P :,008)
Comercio y acuerdos econ�micos	0,18	0,20	(L:13, P _S :,044, P _P :,010)
<i>T�picos Roll Call</i>			
Pol�tica	0,64	0,62	(L:7, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,60	0,55	(L:12, P _S :,002, P _P :,002)
Econom�a, negocios y finanzas	0,30	0,32	(L:9, P _S :,002, P _P :,002)
Asuntos sociales	-0,20	-0,23	(L:22, P _S :,022, P _P :,020)

Tab. B.17: Correlaciones t3pico - (Trump - Harris), Nevada (Julio - Noviembre 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Cr3tica a ideolog3as pol3ticas	-0,49	-0,52	(L:19, P _S :,002, P _P :,002)
Pol3tica educativa	-0,45	-0,45	(L:16, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,43	-0,40	(L:9, P _S :,002, P _P :,002)
Ret3rica patri3tica	-0,42	-0,26	(L:17, P _S :,002, P _P :,022)
Respuesta a desastres	0,38	0,39	(L:20, P _S :,002, P _P :,002)
M3tricas de seguridad fronteriza	-0,37	-0,32	(L:16, P _S :,002, P _P :,008)
Sanidad y programas sociales	-0,33	-0,30	(L:20, P _S :,012, P _P :,006)
G3nero y deportes	0,32	0,51	(L:10, P _S :,004, P _P :,002)
Conflictos internacionales	0,30	0,32	(L:20, P _S :,010, P _P :,006)
Cr3tica a la pol3tica fronteriza	-0,24	-0,25	(L:19, P _S :,032, P _P :,022)
Pol3tica fiscal / impuestos	-0,23	-0,22	(L:10, P _S :,022, P _P :,030)
Seguridad y orden p3blico	-0,22	-0,29	(L:19, P _S :,038, P _P :,012)
Fuerzas armadas y seguridad fronteriza	-0,22	-0,20	(L:11, P _S :,038, P _P :,038)
<i>T3picos Roll Call</i>			
Disturbios, conflictos y guerra	0,39	0,35	(L:20, P _S :,002, P _P :,002)
Trabajo / asuntos laborales	0,28	0,28	(L:20, P _S :,016, P _P :,024)
Crimen, ley y justicia	-0,28	-0,32	(L:17, P _S :,018, P _P :,004)

B.2.6. Pennsylvania

No se encontraron correlaciones significativas para la primera ventana temporal.

Tab. B.18: Correlaciones tópico - (Trump - Biden), Pennsylvania (Noviembre 2023 - Marzo 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Industria, manufactura, tecnología y trabajo	0,75	0,64	(L:27, P _S :,002, P _P :,002)
Respuesta a desastres	0,59	0,42	(L:27, P _S :,002, P _P :,002)
Métricas de seguridad fronteriza	-0,58	-0,62	(L:12, P _S :,002, P _P :,002)
Política educativa	0,50	0,52	(L:28, P _S :,002, P _P :,002)
Género y deportes	0,48	0,50	(L:28, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	0,45	0,45	(L:28, P _S :,002, P _P :,002)
Seguridad y orden público	0,43	0,37	(L:28, P _S :,002, P _P :,002)
Conflictos internacionales	0,41	0,42	(L:25, P _S :,002, P _P :,002)
Petróleo y energía	0,38	0,36	(L:27, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	0,37	0,34	(L:28, P _S :,002, P _P :,006)
Retórica patriótica	0,34	0,36	(L:28, P _S :,006, P _P :,002)
Corrupción legal y política	-0,29	-0,32	(L:14, P _S :,014, P _P :,002)
Inmigración y crimen	0,27	0,20	(L:25, P _S :,008, P _P :,048)
Crítica a ideologías políticas	-0,27	-0,26	(L:16, P _S :,006, P _P :,012)
Asuntos de salud de veteranos	0,25	0,42	(L:28, P _S :,028, P _P :,002)
Crítica a la política fronteriza	-0,22	-0,37	(L:7, P _S :,024, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,20	-0,22	(L:20, P _S :,038, P _P :,028)
<i>Tópicos Roll Call</i>			
Crimen, ley y justicia	-0,53	-0,54	(L:10, P _S :,002, P _P :,002)
Economía, negocios y finanzas	0,45	0,45	(L:27, P _S :,002, P _P :,002)
Política	-0,44	-0,45	(L:12, P _S :,002, P _P :,002)

Tab. B.19: Correlaciones tópico - (Trump - Biden), Pennsylvania (Marzo - Julio 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Sanidad y programas sociales	-0,69	-0,65	(L:20, P _S :.002, P _P :.002)
Crítica a la política fronteriza	0,68	0,63	(L:22, P _S :.002, P _P :.002)
Inmigración y crimen	0,67	0,51	(L:27, P _S :.002, P _P :.002)
Fuerzas armadas y seguridad fronteriza	0,52	0,41	(L:22, P _S :.002, P _P :.002)
Género y deportes	0,49	0,40	(L:12, P _S :.002, P _P :.002)
Seguridad y orden público	0,47	0,36	(L:26, P _S :.002, P _P :.002)
Corrupción legal y política	0,47	0,26	(L:27, P _S :.002, P _P :.006)
Petróleo y energía	0,44	0,32	(L:27, P _S :.002, P _P :.006)
Política fiscal / impuestos	-0,37	-0,38	(L:28, P _S :.002, P _P :.002)
Respuesta a desastres	-0,36	-0,40	(L:20, P _S :.002, P _P :.002)
Métricas de seguridad fronteriza	0,35	0,29	(L:20, P _S :.002, P _P :.010)
Conflictos internacionales	-0,35	-0,39	(L:13, P _S :.002, P _P :.002)
Asuntos de salud de veteranos	0,32	0,20	(L:27, P _S :.002, P _P :.030)
Retórica patriótica	0,31	0,50	(L:7, P _S :.002, P _P :.002)
Comercio y acuerdos económicos	0,30	0,25	(L:27, P _S :.004, P _P :.018)
Líderes políticos y comunitarios	0,25	0,35	(L:8, P _S :.006, P _P :.002)
Crítica a ideologías políticas	-0,21	-0,34	(L:25, P _S :.028, P _P :.004)
<i>Tópicos Roll Call</i>			
Economía, negocios y finanzas	-0,40	-0,37	(L:26, P _S :.002, P _P :.002)
Crimen, ley y justicia	0,36	0,23	(L:27, P _S :.002, P _P :.038)
Política	0,25	0,36	(L:7, P _S :.008, P _P :.002)

Tab. B.20: Correlaciones tópico - (Trump - Harris), Pennsylvania (Julio - Noviembre 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Líderes políticos y comunitarios	0,59	0,48	(L:28, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	0,53	0,53	(L:26, P _S :,002, P _P :,002)
Género y deportes	0,51	0,50	(L:8, P _S :,002, P _P :,002)
Política fiscal / impuestos	0,44	0,47	(L:24, P _S :,002, P _P :,002)
Respuesta a desastres	0,43	0,40	(L:25, P _S :,002, P _P :,004)
Inmigración y crimen	0,42	0,33	(L:24, P _S :,002, P _P :,002)
Conflictos internacionales	0,41	0,44	(L:26, P _S :,002, P _P :,002)
Política educativa	0,38	0,23	(L:28, P _S :,002, P _P :,032)
Petróleo y energía	0,37	0,40	(L:25, P _S :,004, P _P :,002)
Comercio y acuerdos económicos	0,37	0,28	(L:24, P _S :,002, P _P :,008)
Sanidad y programas sociales	0,29	0,33	(L:28, P _S :,010, P _P :,006)
Fuerzas armadas y seguridad fronteriza	0,25	0,21	(L:20, P _S :,020, P _P :,048)
Retórica patriótica	0,20	0,36	(L:18, P _S :,042, P _P :,004)
<i>Tópicos Roll Call</i>			
Trabajo / asuntos laborales	0,59	0,58	(L:26, P _S :,002, P _P :,002)
Disturbios, conflictos y guerra	0,54	0,52	(L:27, P _S :,002, P _P :,002)
Política	0,50	0,40	(L:25, P _S :,002, P _P :,004)
Economía, negocios y finanzas	0,43	0,54	(L:26, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,41	0,37	(L:26, P _S :,002, P _P :,004)
Salud	0,36	0,37	(L:28, P _S :,002, P _P :,002)
Asuntos sociales	0,35	0,38	(L:24, P _S :,008, P _P :,004)

B.2.7. Wisconsin

No se encontraron correlaciones significativas para la primera ventana temporal.

Tab. B.21: Correlaciones tópico - (Trump - Biden), Wisconsin (Noviembre 2023 - Marzo 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Métricas de seguridad fronteriza	0,60	0,51	(L:28, P _S :,002, P _P :,002)
Industria, manufactura, tecnología y trabajo	-0,56	-0,54	(L:8, P _S :,002, P _P :,002)
Comercio y acuerdos económicos	-0,44	-0,44	(L:8, P _S :,002, P _P :,002)
Sanidad y programas sociales	0,39	0,25	(L:22, P _S :,002, P _P :,020)
Respuesta a desastres	-0,39	-0,35	(L:15, P _S :,002, P _P :,004)
Crítica a ideologías políticas	0,37	0,28	(L:16, P _S :,002, P _P :,008)
Política educativa	-0,35	-0,63	(L:20, P _S :,004, P _P :,002)
Aborto	0,33	0,31	(L:18, P _S :,004, P _P :,004)
Seguridad y orden público	-0,32	-0,35	(L:8, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	-0,31	-0,40	(L:9, P _S :,002, P _P :,002)
Petróleo y energía	-0,30	-0,35	(L:8, P _S :,004, P _P :,002)
Conflictos internacionales	0,26	0,44	(L:7, P _S :,012, P _P :,002)
Política fiscal / impuestos	0,26	0,31	(L:19, P _S :,016, P _P :,004)
Crítica a la política fronteriza	0,24	0,27	(L:28, P _S :,008, P _P :,014)
Género y deportes	-0,23	-0,45	(L:20, P _S :,020, P _P :,002)
Retórica patriótica	-0,21	-0,31	(L:8, P _S :,030, P _P :,002)
<i>Tópicos Roll Call</i>			
Crimen, ley y justicia	0,50	0,54	(L:10, P _S :,002, P _P :,002)
Política	0,35	0,31	(L:13, P _S :,002, P _P :,004)
Economía, negocios y finanzas	-0,33	-0,42	(L:27, P _S :,004, P _P :,002)

Tab. B.22: Correlaciones t3pico - (Trump - Biden), Wisconsin (Marzo - Julio 2024).

T3pico	Spearman	Pearson	Detalles
<i>T3picos BERTopic</i>			
Pol3tica fiscal / impuestos	0,65	0,65	(L:9, P _S :,002, P _P :,002)
Corrupci3n legal y pol3tica	-0,58	-0,60	(L:26, P _S :,002, P _P :,002)
Inmigraci3n y crimen	-0,52	-0,42	(L:9, P _S :,002, P _P :,002)
Aborto	0,49	0,43	(L:7, P _S :,002, P _P :,002)
Conflictos internacionales	-0,44	-0,41	(L:28, P _S :,002, P _P :,002)
Cr3tica a ideolog3as pol3ticas	0,36	0,46	(L:7, P _S :,002, P _P :,002)
Ret3rica patri3tica	-0,36	-0,29	(L:28, P _S :,002, P _P :,002)
Industria, manufactura, tecnolog3a y trabajo	-0,35	-0,39	(L:28, P _S :,002, P _P :,002)
Sanidad y programas sociales	0,34	0,17	(L:7, P _S :,002, P _P :,042)
Pol3tica educativa	-0,32	-0,23	(L:28, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	-0,32	-0,25	(L:12, P _S :,002, P _P :,002)
M3tricas de seguridad fronteriza	-0,31	-0,29	(L:9, P _S :,002, P _P :,002)
Petr3leo y energ3a	-0,27	-0,22	(L:26, P _S :,010, P _P :,002)
L3deres pol3ticos y comunitarios	0,24	0,30	(L:8, P _S :,004, P _P :,002)
Respuesta a desastres	0,22	0,23	(L:7, P _S :,012, P _P :,014)
Comercio y acuerdos econ3micos	-0,20	-0,24	(L:11, P _S :,018, P _P :,002)
Asuntos de salud de veteranos	-0,18	-0,24	(L:25, P _S :,048, P _P :,006)
<i>T3picos Roll Call</i>			
Crimen, ley y justicia	-0,70	-0,64	(L:26, P _S :,002, P _P :,002)
Econom3a, negocios y finanzas	0,57	0,51	(L:7, P _S :,002, P _P :,002)
Pol3tica	-0,49	-0,43	(L:26, P _S :,002, P _P :,002)
Asuntos sociales	0,30	0,32	(L:10, P _S :,002, P _P :,002)

Tab. B.23: Correlaciones tópico - (Trump - Harris), Wisconsin (Julio - Noviembre 2024).

Tópico	Spearman	Pearson	Detalles
<i>Tópicos BERTopic</i>			
Industria, manufactura, tecnología y trabajo	0,80	0,77	(L:27, P _S :,002, P _P :,002)
Líderes políticos y comunitarios	0,70	0,66	(L:27, P _S :,002, P _P :,002)
Respuesta a desastres	0,67	0,67	(L:21, P _S :,002, P _P :,002)
Género y deportes	0,55	0,51	(L:7, P _S :,002, P _P :,002)
Política fiscal / impuestos	0,55	0,50	(L:27, P _S :,002, P _P :,002)
Asuntos de salud de veteranos	0,53	0,40	(L:27, P _S :,002, P _P :,002)
Fuerzas armadas y seguridad fronteriza	0,49	0,36	(L:27, P _S :,002, P _P :,002)
Política educativa	0,46	0,30	(L:27, P _S :,002, P _P :,010)
Conflictos internacionales	0,41	0,35	(L:7, P _S :,002, P _P :,002)
Corrupción legal y política	0,37	0,46	(L:7, P _S :,002, P _P :,002)
Aborto	0,35	0,39	(L:7, P _S :,002, P _P :,002)
Seguridad y orden público	0,33	0,30	(L:7, P _S :,008, P _P :,004)
Retórica patriótica	0,32	0,37	(L:9, P _S :,002, P _P :,002)
Inmigración y crimen	0,31	0,24	(L:7, P _S :,006, P _P :,006)
Petróleo y energía	0,27	0,31	(L:27, P _S :,012, P _P :,006)
Sanidad y programas sociales	0,27	0,29	(L:15, P _S :,014, P _P :,012)
Crítica a la política fronteriza	0,26	0,27	(L:7, P _S :,006, P _P :,010)
Crítica a ideologías políticas	0,23	0,24	(L:7, P _S :,028, P _P :,026)
<i>Tópicos Roll Call</i>			
Disturbios, conflictos y guerra	0,58	0,45	(L:13, P _S :,002, P _P :,002)
Trabajo / asuntos laborales	0,55	0,57	(L:21, P _S :,002, P _P :,002)
Política	0,54	0,52	(L:9, P _S :,002, P _P :,002)
Salud	0,40	0,40	(L:9, P _S :,002, P _P :,002)
Crimen, ley y justicia	0,37	0,28	(L:26, P _S :,004, P _P :,022)
Asuntos sociales	0,32	0,32	(L:23, P _S :,004, P _P :,008)
Economía, negocios y finanzas	0,28	0,30	(L:27, P _S :,016, P _P :,010)

C. TEST ESTADÍSTICO DE PERMUTACIÓN

C.0.1. Significancia estadística de correlaciones

Para evaluar la significancia estadística de los coeficientes de correlación de Pearson (r) y de Rango de Spearman (ρ) entre la relevancia temporal de los tópicos y las series de intención de voto, se emplearon tests de permutación no paramétricos [40]. Este método es robusto frente a violaciones de los supuestos distribucionales de los tests analíticos tradicionales y es particularmente adecuado para tamaños de muestra moderados, como los encontrados en las ventanas temporales, que tienen un orden de 120 observaciones (ventanas de cuatro meses).

El principio fundamental del test de permutación se basa en la hipótesis nula (H_0) de que no existe una asociación real entre las dos variables comparadas. Si H_0 es cierta, cualquier correlación observada es puramente aleatoria. El test construye empíricamente la distribución del coeficiente de correlación bajo la hipótesis nula mediante los siguientes pasos:

1. **Cálculo del coeficiente observado:** para un par de series (X, Y) de N observaciones alineadas (donde X representa la serie de relevancia del tópico y Y la serie de intención de voto, ajustada por un lag específico), se calcula el coeficiente de correlación original, c_{obs} (siendo c ya sea r de Pearson o ρ de Spearman).
2. **Generación de la distribución nula:** se realizaron M iteraciones (remuestreos). En cada iteración i :
 - Se mantuvo fija la serie X .
 - Los valores de la serie Y fueron permutados aleatoriamente para crear una nueva serie $Y'_{perm,i}$, rompiendo así cualquier asociación original con X .
 - Se calculó el coeficiente de correlación $c_{perm,i}$ entre X y $Y'_{perm,i}$.

Se utilizaron $M = 999$ permutaciones para cada test.

3. **Cálculo del p-valor (dos colas):** el p-valor se estimó como la proporción de coeficientes permutados cuya magnitud fue igual o mayor a la magnitud del coeficiente observado:

$$p = \frac{(\text{Número de veces que } |c_{perm,i}| \geq |c_{obs}|) + 1}{M + 1}$$

Este cálculo proporciona la probabilidad de observar una correlación tan fuerte o más fuerte que la original si la hipótesis nula fuera verdadera.

Este procedimiento se aplicó de forma independiente para obtener los p-valores tanto para las correlaciones de Pearson como para las de Spearman reportadas en los análisis.

Bibliografía

- [1] SoPhy Lab. Sophy: Social physics at df – uba. <https://sophy.df.uba.ar/>, 2019.
- [2] S. M. del Pozo, S. Pinto, M. Serafino, F. Moss, T. Cicchini, H. A. Makse, and P. Balenzuela. The hidden dimension of information diffusion: A latent space representation of social media news sharing behavior. *Scientific Reports*, 15:516, 2025.
- [3] S. M. del Pozo, S. Pinto, M. Serafino, L. Garcia, H. A. Makse, and P. Balenzuela. Analyzing user ideologies and shared news during the 2019 argentinian elections. *EPJ Data Science*, 13:54, 2024.
- [4] T. Cicchini, S. M. del Pozo, E. Tagliazucchi, and P. Balenzuela. News-sharing on twitter reveals emergent fragmentation of media agenda and persistent polarization. *EPJ Data Science*, 11:48, 2022.
- [5] S. Pinto, F. Albanese, C. O. Dorso, and P. Balenzuela. Quantifying time-dependent media agenda and public opinion by topic modeling. *Physica A*, 524:614–624, 2019.
- [6] Federico Albanese, Sebastián Pinto, Viktoriya Semeshenko, and Pablo Balenzuela. Analyzing mass media influence using natural language processing and time series analysis. *Journal of Physics: Complexity*, 1(2):025005, 2020.
- [7] FiscalNote. Roll call. <https://rollcall.com/>, 2025.
- [8] RealClearPolitics. Realclearpolling. <https://www.realclearpolling.com/>, 2025.
- [9] 270towin. <https://www.270towin.com/>, 2025.
- [10] Kenneth Reitz. Requests: Http for humans. <https://pypi.org/project/requests/>, 2023.
- [11] Leonard Richardson. Beautiful soup. <https://www.crummy.com/software/BeautifulSoup/>, 2024.
- [12] The Selenium Authors. Selenium webdriver. <https://www.selenium.dev/>, 2024.
- [13] Jay Alamar and Maarten Grootendorst. *Advanced Techniques for Large Language Models*, chapter 5. O’Reilly, 2024.
- [14] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
- [16] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.
- [17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

-
- [18] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [19] O. Babalola, B. Ojokoh, and O. Boyinbode. Comprehensive evaluation of lda, nmf, and bertopic’s performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2(2):268–289, 2024.
- [20] Anusuya Krishnan. Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis, 2023.
- [21] SBERT.net. Modelos pre-entrenados — sentence-transformers (documentación). https://www.sbert.net/docs/sentence_transformer/pretrained_models.html, 2024.
- [22] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
- [23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [24] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, 7819:160–172, 2013.
- [25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011.
- [26] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In Donna K. Harman, editor, *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. National Institute of Standards and Technology (NIST), 1995.
- [27] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert, 2020.
- [28] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
- [29] Karl Pearson. Notes on the history of correlation. *Biometrika*, 1920.
- [30] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 1904.
- [31] HDBSCAN development team. Selección de hiperparámetros para hdbscan — documentación de hdbscan. https://hdbscan.readthedocs.io/en/latest/parameter_selection.html, 2023.

-
- [32] Project HOPE Staff. “Cronología visual de huracanes: Respuesta de emergencia del proyecto HOPE en 2024”. <https://www.projecthope.org/news-stories/story/hurricane-visual-timeline/>, 2024.
- [33] Samantha Mathewson. “SpaceX ofrece servicio de internet Starlink gratuito para las personas afectadas por los huracanes Helene y Milton.”. <https://www.space.com/spacex-starlink-free-2024-hurricane-helene-milton>, 2024.
- [34] J. E. Luebering. Roe v. wade. <https://www.britannica.com/event/Roe-v-Wade>, 2024.
- [35] Ballotpedia. Seguimiento de procesos judiciales de donald trump, 2023-2025. https://ballotpedia.org/Donald_Trump_indictments,_2023-2025, 2025.
- [36] IBISWorld. Perfil de economía de michigan. <https://www.ibisworld.com/united-states/economic-profiles/michigan/>, 2024.
- [37] The White House. “El presidente Joe Biden presenta nuevos planes para aliviar la deuda estudiantil de más de 30 millones de estadounidenses bajo la administración Biden-Harris.”. <https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2024/04/08/president-joe-biden-outlines-new-plans-to-deliver-student-debt-relief-to-over-30-million-americans-under-the-biden-harris-administration/>, 2024.
- [38] Carroll Doherty, Alec Tyson, and Hannah Hartig. “Los veteranos militares siguen siendo un grupo republicano, apoyando a Trump sobre Harris por un amplio margen.”. <https://pewrsr.ch/3Y7DeL6>, 2024.
- [39] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 1969.
- [40] Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, New York, third edition, 2005.