Tarea 1 - Introducción a la Ciencia de Datos

Bruno Arnuti - Maximiliano Anzibar Fialho - 2025

En esta tarea se trabajo con una base de datos abierta que contiene discuros de los candidatos políticos a las elecciones de Estados Unidos de 2020. El dataset contiene 6 columnas correspondientes a: *speaker* (nombre de quien da el discurso), *title*(título asignado al discurso), *text*(redacción del discurso), *date*(fecha en que se dió el discurso), *location*(locación donde se dió el discurso), y *type* (tipo de discurso, entrevista, discurso de campaña, etc).

1. Cargado y limpieza de datos

Al cargarlo los datos observamos que tenemos un total de 269 entradas (filas). Lo primero que realizamos fue verificar si había datos faltantes, donde comprobamos que teniamos 3 entradas vacías para la columna *speaker*, 18 entradas vacías para la columna *location*, y 21 entradas vacías para la columna *type*. Al eliminar todas las filas que contiene datos vacíos nos quedamos con un total de 237 entradas (no corresponde a la suma de los campos anteriores ya que algunas entradas tienen faltantes tanto en *location* como en *type*).

1.1. Evolución en el tiempo de la cantidad de discursos

Lo siguiente que realizamos fue contar la cantidad de discursos por candidato, y crear un nuevo dataset con los 5 candidatos que tienen más discursos. Estos son **Joe Biden** con 60, **Donald Trump** con 51, **Mike Pence** con 19, **Bernie Sanders** con 12, y **Kamala Harris** con 11. En total este nuevo dataset tienen entonces 153 entradas.

A continuación queremos ver cómo fue la evolución de los discursos a lo largo del tiempo. Para ello realizamos dos análisis. En primer lugar graficamos la cantidad de discursos para cada candidatos en cada fecha. Es decir en todas las fechas donde había una entrada verificamos si uno o más candidatos realizaron un discurso y cuantos realizaron. Los resultados los vemos en la figura 1. Podemos observar que el primer discurso del dataset corresponde a Trump, y que en la mayoría de las fechas los candidatos dan como máximo 1 discurso, excepto sobre los últimos dos meses de campaña donde se vuelve usual que los candidatos den 2 discurso por día (e incluso 4 discursos dados por Biden el 30-09-2020).

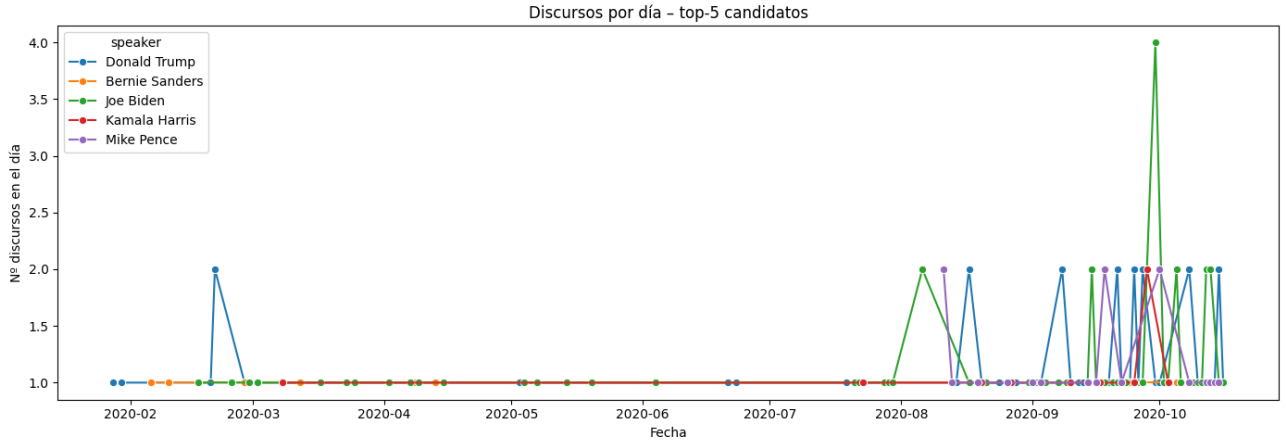

Figura 1: Discursos por fecha

Lo siguiente fue visualizar la cantidad acumulada de discursos por fecha (figura 2). Como podemos ver Joe Biden y Donald Trump son los que sostenidamente dan más discurso y los que también sobre los meses finales aumentan la cantidad de discursos por día (ver la pendiente de las gráficas a partir del mes 8). Otra cosa que podemos notar es que candidatos con menores opciones de ganar las elecciones como Mike Pence comienzan a dar sus discursos más próximo al día de las elecciones, en este caso su primer discurso fue el 11-08-2020; los otros candidatos dieron dus primeros discursos dentro del mes 1, 2 o 3.

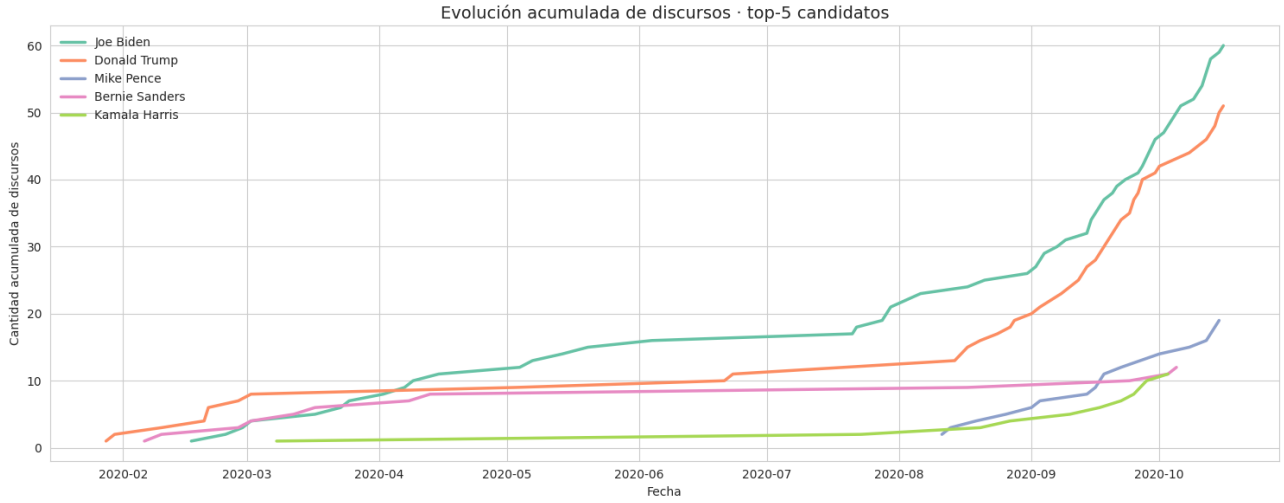

Figura 2: Discursos acumulados por fecha

1.2. Limpieza y acondicionamiento del texto

Previo a trabajar con el texto debemos realizar una depuración de los discursos. Esto consistió en normalizar el texto eliminando puntuaciones indeseadas (puntos, comas, signos de interrogación, etc), pasando todo a minúscula, eliminando formatos de url (como http o www) y eliminando los espacios en blanco. Finalmente convertimos cada discurso a una lista donde cada entrada de la lista es una palabra y agregamos las listas (asociadas a su correspondiente discurso) a una nueva columna del dataset la cual denominamos *WordList*.

2. Conteo de Palabras y Visualizaciones

En esta parte seguimos trabajando con los 5 candidatos con más discursos y lo primero que queremos hacer es encontrar las palabras más frecuentes que menciona cada uno de los candidatos. En la figura 3 y 8 podemos observar que en todos los casos las palabras que más se repiten son conectores, artículos y preposiciones (como "the", .ªnd", "to", .ºf", etc.), conocidas en el mundo de procesamiento de lenguaje como *stop words*. Este tipo de palabras no aporta mucha información sobre el contenido de los discursos e impiden ver términos más significativos desde el punto de vista político, por lo tanto procedemos a eliminar estas palabras para trabajar con los discursos más depurados.

Como se puede observar en los gráficos 4 y 9, las palabras más frecuentes son casi idénticas entre los candidatos y corresponden principalmente a palabras funcionales del idioma (por ejemplo *going* o *us*). Esto dificulta detectar las diferencias reales en el contenido político de cada discurso y oculta términos relevantes. En una etapa posterior del análisis, sería conveniente filtrar estas palabras vacías para permitir que emerjan patrones temáticos más significativos, pero por ahora nos centramos en asegurar que los conteos de frecuencia sean correctos para todas las palabras. Sin embargo, de este análisis podemos detectar que la automención es algo que se repite de forma muy marcada para todos los candidatos.

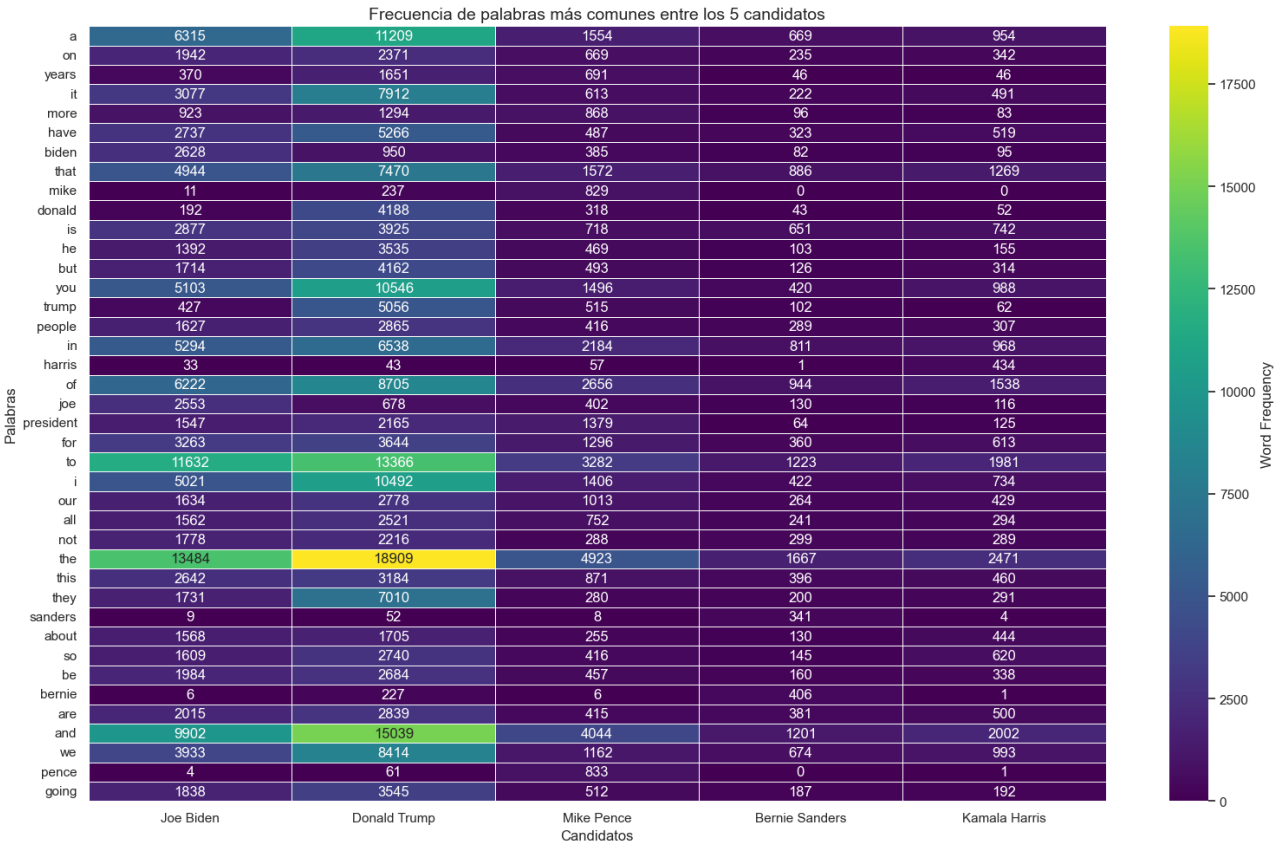

Figura 3: Palabras más frecuentes de cada candidato

Algunos de los análisis que se podrían implementar a futuro son:

- Diferencias entre partidos políticos
  - Agrupar candidatos por afiliación política (Demócratas vs Republicanos) y calcular la frecuencia de palabras agregada para cada partido
  - Crear una visualización de "diferencia de frecuencias"donde el tamaño de la palabra indique cuánto más la usa un partido que otro
  - Implementar análisis de sentimiento para identificar si ciertos temas son abordados con tonos más positivos o negativos según el partido
  - Crear nubes de palabras con códigos de color por partido para identificar visualmente los términos más asociados a cada ideología
- Análisis temporal (fechas)
  - Dividir el período de campaña en fases (primarias, convenciones, debates, semanas finales) y comparar la evolución del vocabulario
  - Crear gráficos de líneas para palabras clave que muestren cómo cambia su frecuencia a lo largo del tiempo
  - Implementar análisis de "trending words"para detectar términos que aparecen súbitamente en respuesta a eventos de campaña
  - Visualizar la evolución del discurso de cada candidato para identificar cambios estratégicos en sus mensajes
- Análisis geográfico (lugares)

Figura 4

- Filtrar discursos por estado o región y comparar las palabras más frecuentes según la ubicación
- Crear mapas de calor geográficos donde el color refleje la frecuencia de términos específicos en cada estado
- Comparar el vocabulario utilizado en estados disputados ("swing states") versus estados tradicionalmente leales a un partido
- Analizar si los candidatos adaptan su mensaje según el contexto regional, económico o demográfico de cada lugar

2.1. Matriz de menciones

Por último calculamos una matriz de menciones, es decir cuantas veces un candidato menciona a otro (figura 5). En el anexo de figuras se muestra también una matriz que contiene la cantidad de menciones para cada candidato en función del total de menciones expresado en porcentaje (figura 10)

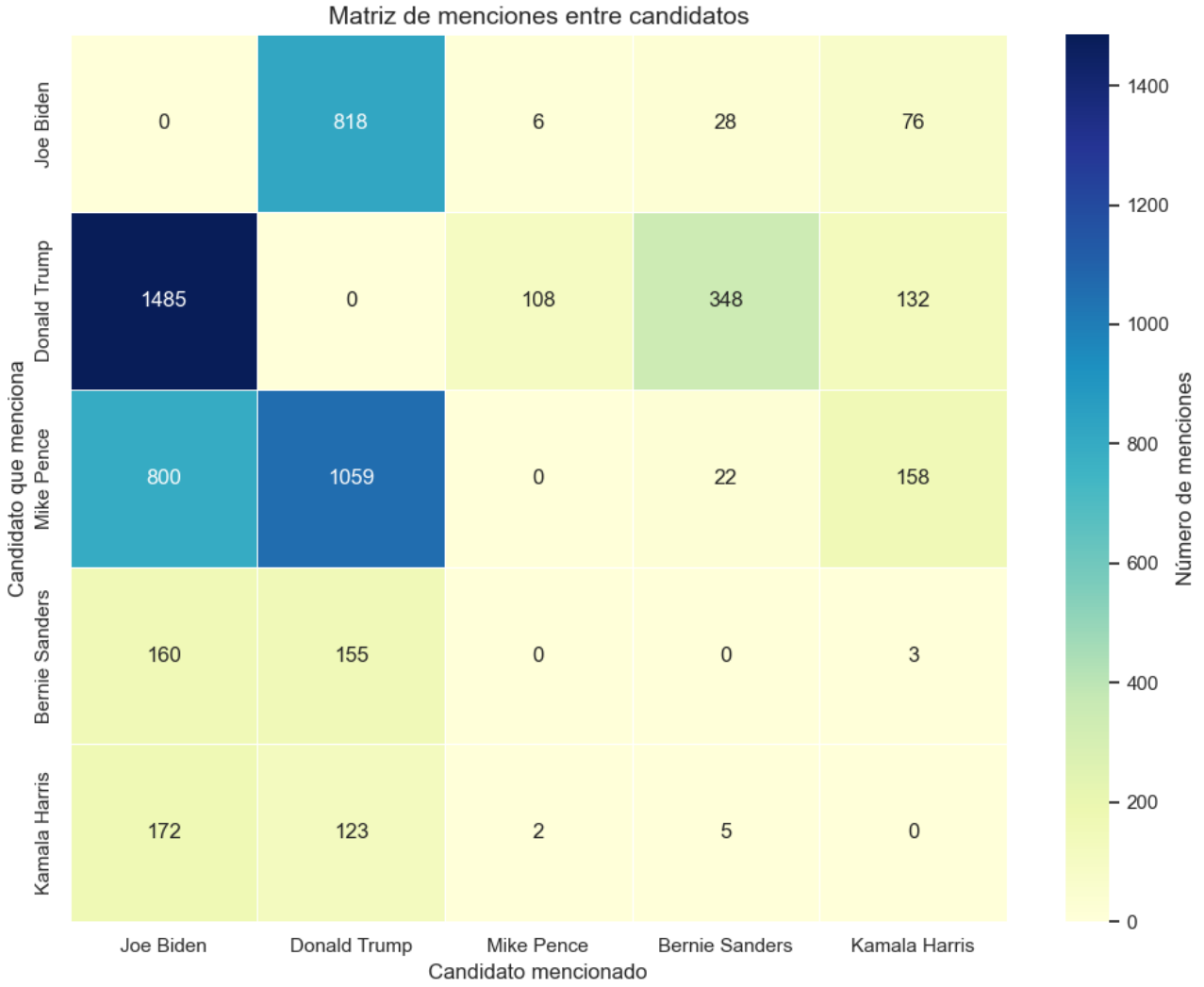

Figura 5

En el discurso político, los candidatos raramente son mencionados con un único nombre. Por ejemplo, Joe Biden puede ser referido como "Biden", "Joe Biden", "Vice President Biden.º "Former Vice President". Para capturar esta diversidad, definimos múltiples variantes para cada candidato:

candidate_name_variants = "Joe Biden": ["biden", "joe biden", "vice president biden", "vp biden", "former vice president"], "Donald Trump": ["trump", "donald trump", "president trump", "donald j trump"], ...otros candidatos

Esto nos permite detectar menciones independientemente de cómo un candidato sea referido, aumentando considerablemente la precisión del análisis. Para asegurar que solo contamos palabras completas (y no partes de otras palabras), implementamos una función que utiliza expresiones regulares con delimitadores de palabras (ver código).

Los delimitadores son importantes porque aseguran que solo contemos palabras completas. Por

ejemplo, con este enfoque, "trumpeting"no se contará como una mención de "trump".

Luego implementamos una visualización de grafo para observar las menciones entre los distintos candidatos (figura 6). En este caso al ser un grafo bidireccional (es decir que A menciona a B pero B también menciona a A), grafo resultante puede ser difícil de visualizar ambas direcciones de mención, ya que las aristas suelen superponerse. En este caso agregamos el peso en formato numérico en cada arista y como el grosor de la arista. Una solución que encontramos para mejorar la visualización es hacer 5 grafos distintos, uno para cada candidato (ver figura 7). En este caso podemos ver con más detalle y de forma ordenada cuantas veces cada candidato menciona al resto.

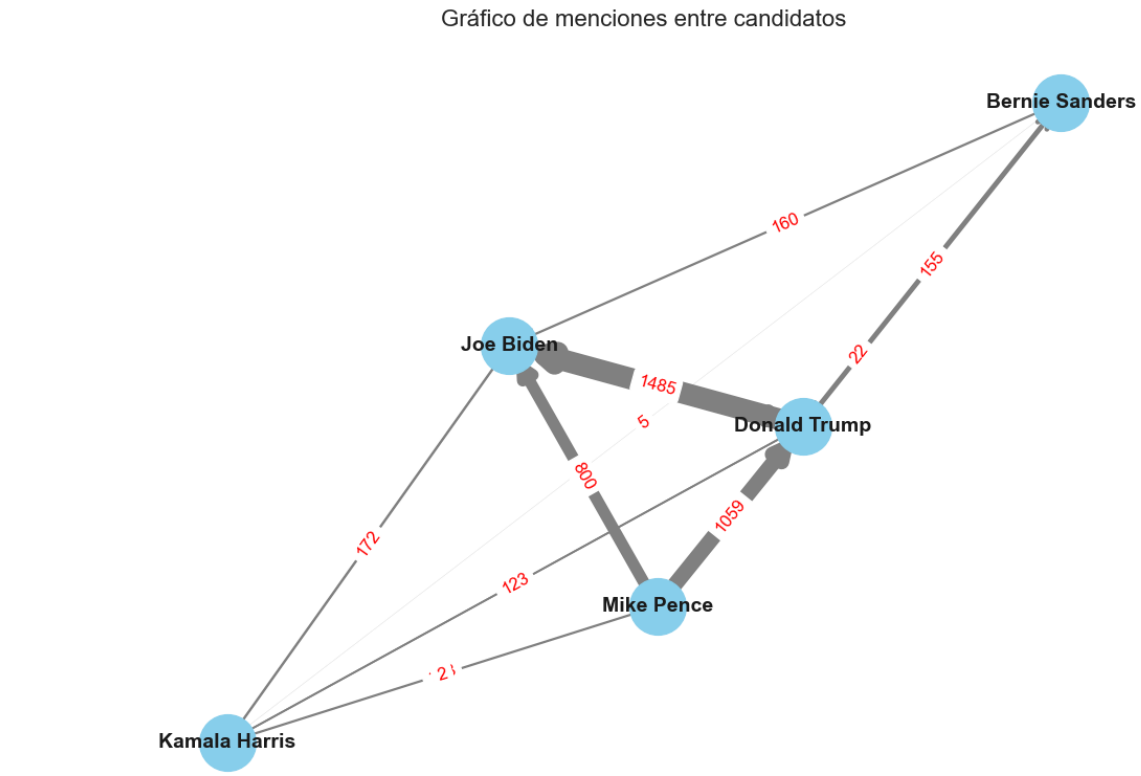

Figura 6

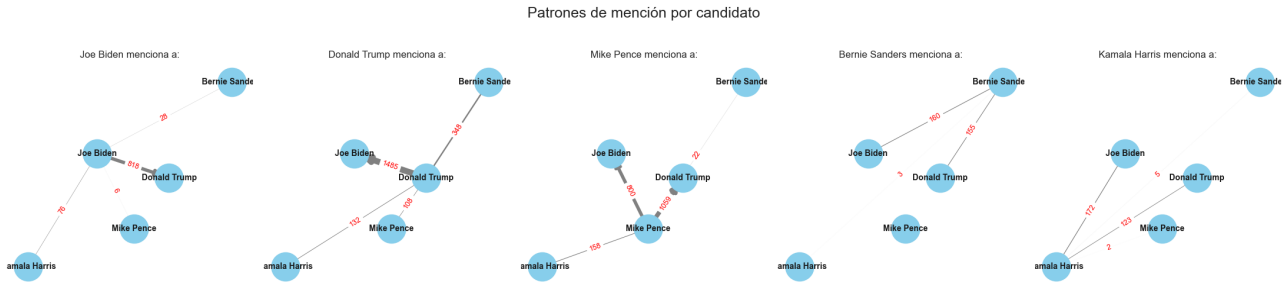

Figura 7

2.2. Preguntas adicionales para análisis futuro

Para cerrar proponemos algunas preguntas que puedan servir como disparadoras para futuros análisis:

1. **¿Cómo evolucionó el lenguaje y los temas de los candidatos a lo largo de la campaña?**

Los discursos políticos suelen adaptarse a eventos externos, resultados de encuestas y estrategias cambiantes. Analizar la evolución temporal del lenguaje nos permitiría entender cómo los candidatos ajustan sus mensajes.

El enfoque metodológico sería dividir los discursos en intervalos temporales significativos (mensual, por etapas de campaña) Para cada intervalo, extraer los temas principales usando técnicas como modelado de tópicos (LDA) Analizar cambios en la frecuencia de palabras clave específicas a lo largo del tiempo Comparar cómo cambia el tono emocional de los discursos mediante análisis de sentimiento Este análisis revelaría, por ejemplo, si Trump cambió su retórica después de eventos clave o si Biden modificó sus temas prioritarios en diferentes estados.

2. **¿Qué diferencias existen en el discurso de los candidatos según la ubicación geográfica?**

   Los candidatos adaptan sus mensajes según las preocupaciones locales y la composición demográfica de cada región. Esta pregunta nos permitiría comprender cómo personalizan su comunicación.

   El enfoque metodológico sería agrupar discursos por estado o región (noreste, sur, medio oeste, etc.) Identificar términos distintivos para cada región mediante análisis de frecuencia comparativa Analizar si hay temas específicos que solo se abordan en ciertas regiones Examinar las menciones de industrias o sectores económicos relevantes para cada estado Este análisis podría mostrar, por ejemplo, que en estados con alta producción agrícola se habla más de subsidios y aranceles, mientras que en zonas urbanas se enfatiza vivienda o transporte.

3. **¿Qué estrategias retóricas caracterizan a cada candidato y cómo se relacionan con su éxito electoral?**

   Más allá de los temas abordados, la forma en que los candidatos comunican sus ideas puede ser tan importante como el contenido mismo. Esta pregunta explora patrones lingüísticos distintivos.

   El enfoque metodológico sería nalizar la complejidad lingüística (longitud de oraciones, diversidad de vocabulario) Identificar patrones de uso de la primera persona (yo, nosotros) versus tercera persona. Examinar el uso de lenguaje emocional o cargado versus lenguaje neutro. Estudiar la frecuencia de ciertos recursos retóricos: preguntas, llamados a la acción, anécdotas Correlacionar estos patrones con la respuesta del público (encuestas posteriores a discursos) Este análisis podría revelar, por ejemplo, que algunos candidatos utilizan lenguaje más sencillo y directo, mientras que otros emplean un vocabulario más sofisticado, y cómo estas diferencias impactan en su capacidad para conectar con distintos segmentos del electorado.

3. Anexo Figuras

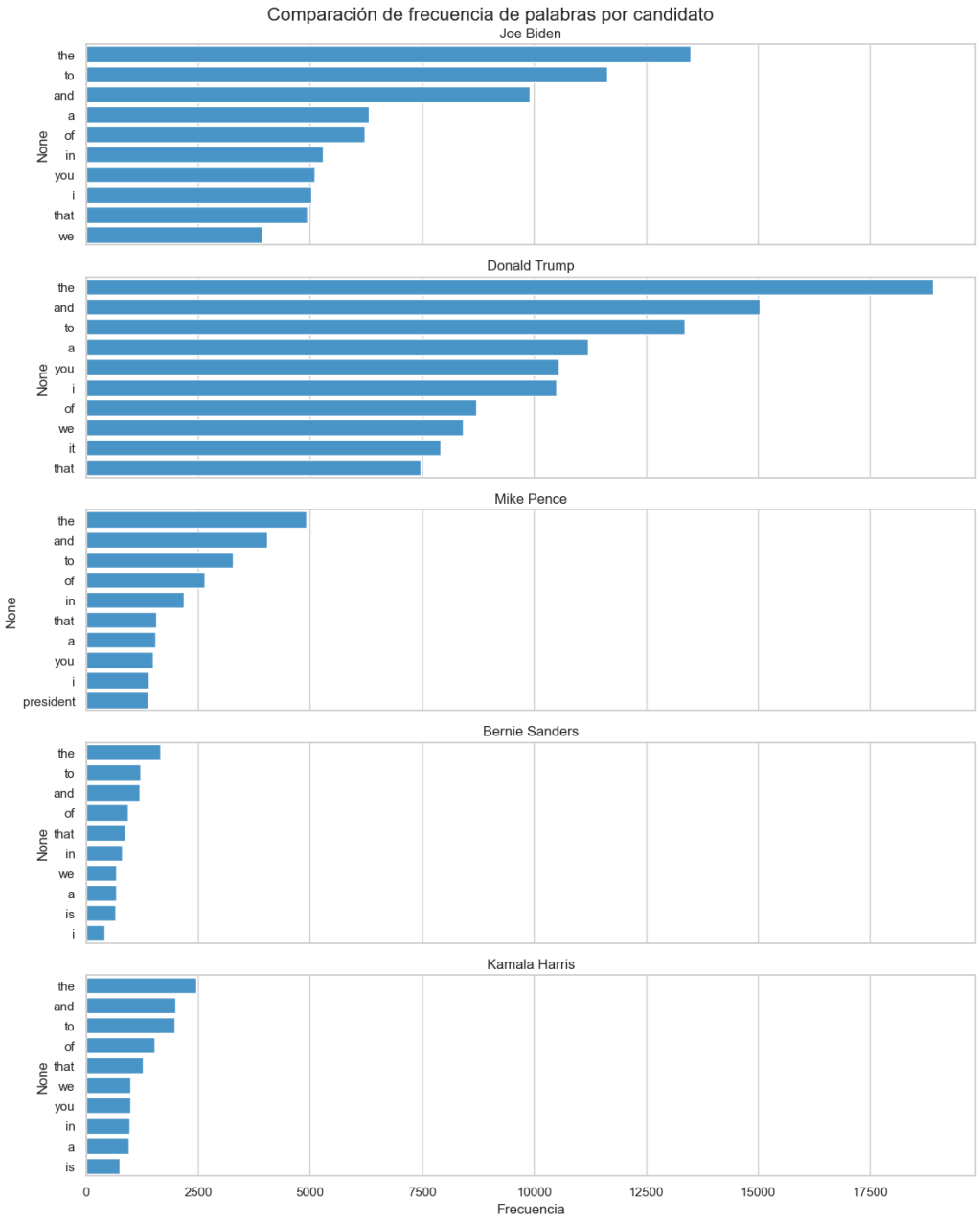

Figura 8

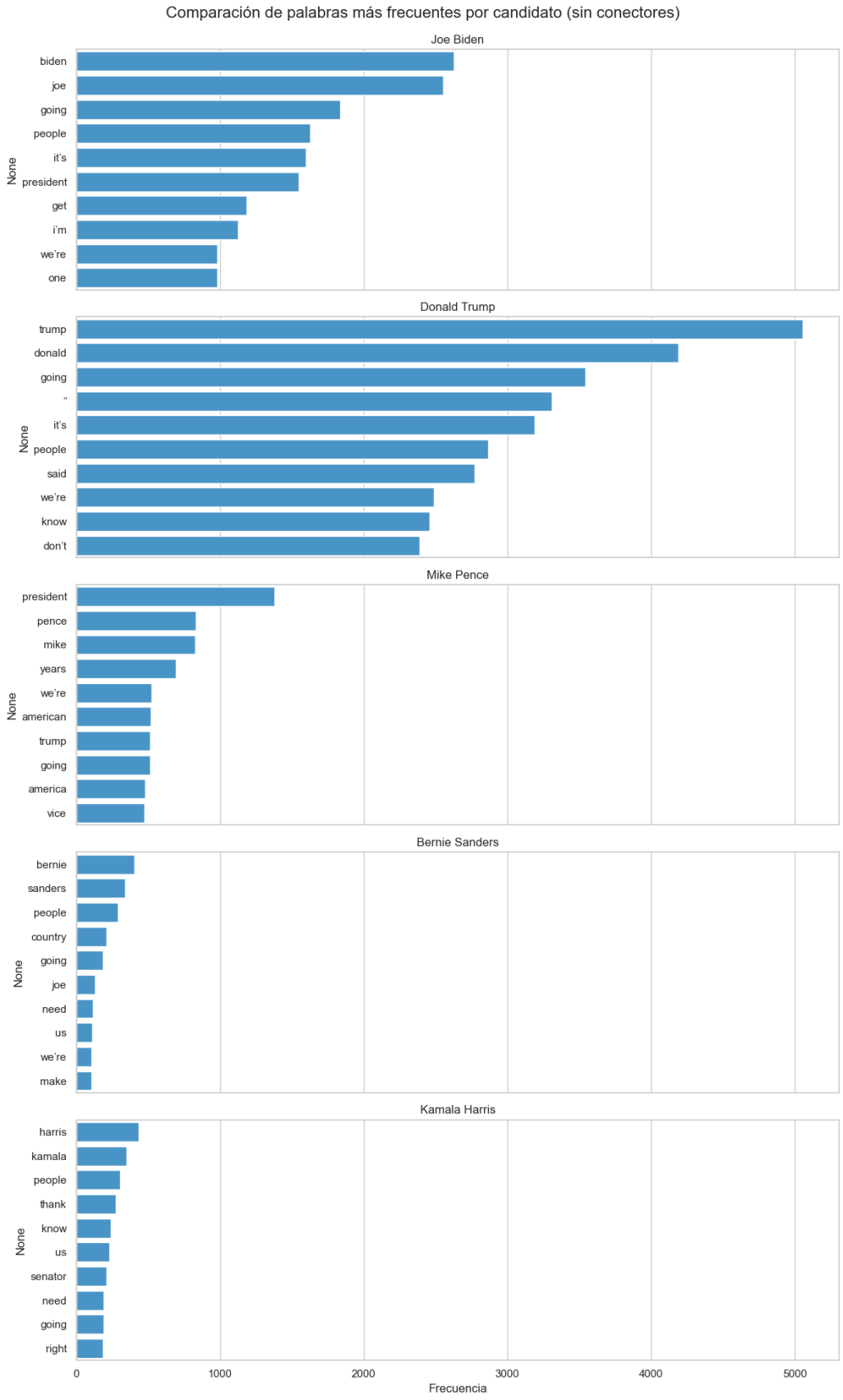
Comparación de palabras más frecuentes por candidato (sin conectores)
Joe Biden
biden
joe
going
people
it's
president
get
i'm
we're
one
Donald Trump
trump
donald
going
"
it's
people
said
we're
know
don't
Mike Pence
president
pence
mike
years
we're
american
trump
going
america
vice
Bernie Sanders
bernie
sanders
people
country
going
joe
need
us
we're
make
Kamala Harris
harris
kamala
people
thank
know
us
senator
need
going
right
None
0
1000
2000
3000
4000
5000
Frecuencia

Figura 9

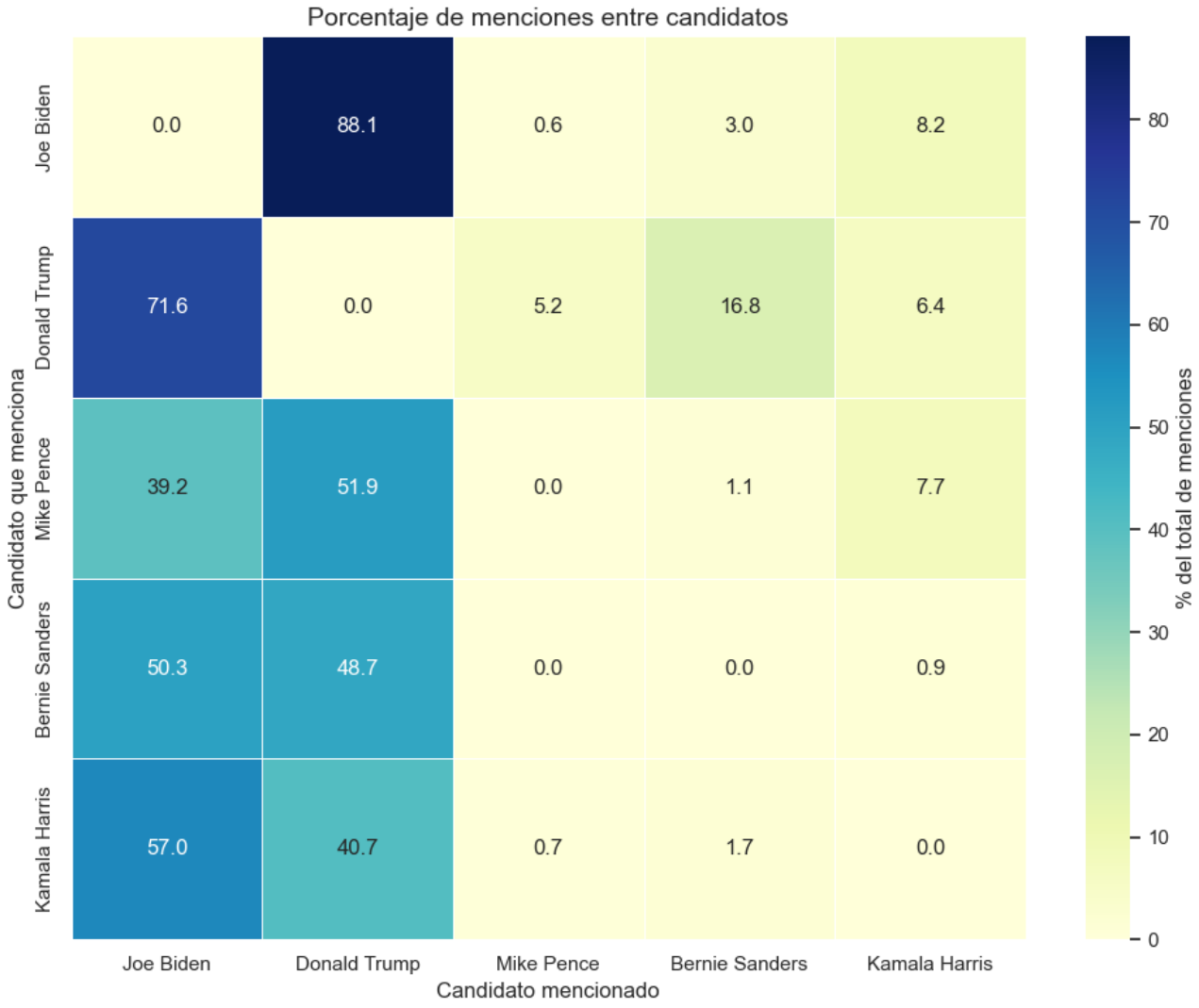
Porcentaje de menciones entre candidatos
Candidato que menciona
Joe Biden
Donald Trump
Mike Pence
Bernie Sanders
Kamala Harris
0.0
88.1
0.6
3.0
8.2
71.6
0.0
5.2
16.8
6.4
39.2
51.9
0.0
1.1
7.7
50.3
48.7
0.0
0.0
0.9
57.0
40.7
0.7
1.7
0.0
Joe Biden
Donald Trump
Mike Pence
Bernie Sanders
Kamala Harris
Candidato mencionado
% del total de menciones
80
70
60
50
40
30
20
10
0

Figura 10