1. **Introducción**

En las elecciones presidenciales de Estados Unidos en 2020, celebradas el 3 de noviembre, el candidato Joe Biden resultó electo, con 306 votos electorales frente a 232 que consiguió Donald Trump.

Según artículos de The New York Times, The Wall Street Journal y NPR, los temas de la campaña estuvieron centrados en el COVID-19, la muerte de dos afroamericanos durante manifestaciones, la Ley del Cuidado de Salud a Bajo Precio, el cambio climático, entre otros. [1] [2] [3]

Este trabajo práctico busca realizar un análisis acerca de los discursos de los candidatos estadounidenses en 2020, usando herramientas de ciencia de datos.

1. **Metodología**

**2.1. Datos de partida**

Se utilizó el *data frame* de los discursos dados por los candidatos durante la campaña electoral, este contiene el nombre del candidato, la fecha del discurso, el título, la transcripción completa, la ubicación donde se llevó a cabo y el tipo de evento o discurso.

El trabajo fue realizado en Google Colab, usando Python como lenguaje de programación. Finalmente, el trabajo se cargó en GitLab.

**2.2. Detección de problemas en la calidad de los datos**

Usando la función ‘isnan’ se obtuvieron los datos faltantes. Utilizando la función ‘unique’ se obtienen datos únicos para realizar una inspección visual para encontrar inconsistencias.

**2.3. Discursos por candidatos**

Se listaron todos los candidatos que aparecen al menos una vez en la columna *speaker*. Para aquellos con múltiples candidatos, se utilizó la función ‘split’ para separar los nombres utilizando la coma como separador. Se creó una nueva fila para cada orador manteniendo el resto de la información.

Con el data frame actualizado se contó la cantidad de veces que aparece el orador en la columna speaker. Ordenando y obteniendo el top 5.

Para estos candidatos, se realizó un gráfico de barras con los discursos por mes a lo largo de la campaña, agregando ceros cuando no hay información de discursos del candidato en un mes. En esta instancia se evaluaron los momentos claves de la campaña.

**2.4. Normalización del texto y conteo de palabras**

Para contar las palabras, primero se normalizó el texto: se pasaron todas las letras a minúsculas, se eliminaron las marcas de tiempo, los nombres de los candidatos al comienzo de cada línea, los signos de puntuación y los espacios duplicados. Luego, se listaron las palabras más usadas y se hizo un gráfico de barras adaptado.

Para analizar los discursos, se identificaron los temas principales (como economía, salud, inmigración, etc.) y se listaron palabras clave para cada uno. Posteriormente, se procesaron los textos y se contó cuántas veces aparecía cada palabra clave en cada intervención.

Cuando un discurso era compartido por varios candidatos, se lo contó por separado para cada uno. Con estos datos, se calculó cuántas veces mencionó cada tema cada candidato. Para poder comparar los resultados, esos valores se pasaron a porcentajes sobre el total de menciones de cada candidato, y se mostraron en un mapa de calor. Así se puede ver fácilmente qué temas fueron más importantes para cada figura política.

**2.5. Generación de matriz de menciones entre candidatos**

Este análisis busca identificar con qué frecuencia los candidatos mencionan a sus oponentes en sus discursos. Para ello, se partió de una lista con los nombres de cinco figuras políticas clave, tanto en su forma completa como en la forma más común con la que suelen ser mencionados (por ejemplo, *"Joe Biden"* y *"Biden"*).

Se recorrieron los discursos de cada candidato y, en cada uno, se contabilizó cuántas veces aparecía el nombre de otro candidato. No se contabilizaron las menciones a sí mismos.

Los resultados se almacenaron en una matriz en la que las filas representan a quien habla y las columnas a quien es mencionado. Cada celda muestra la cantidad de veces que un candidato hizo referencia a otro.

1. **Resultados**

**3.1. Datos de partida**

A continuación, se muestra el primer dato del data frame crudo (Tabla 1).

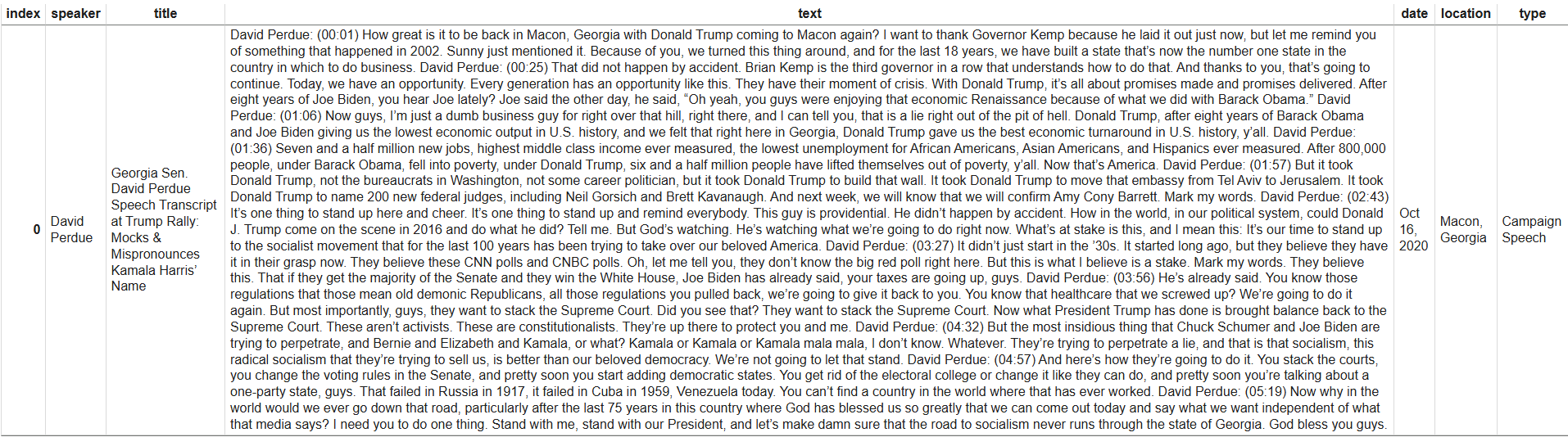


Tabla 1: Primer dato del data frame.

**3.2. Detección de problemas de calidad de los datos**

Existen datos faltantes en algunas columnas (Tabla 2). Esto puede afectar el análisis, en particular cuando la cantidad de datos faltantes es significativa, como en la columna *text*, donde se encuentra la transcripción del discurso.

|  |  |
| --- | --- |
| Columna | Cantidad de datos faltantes |
| speaker | 3 |
| title | 0 |
| text | 65 |
| date | 0 |
| type | 21 |

Tabla 2: Datos faltantes

De la inspección visual se detectaron datos incompletos en la columna speakers, como ‘???’. Asimismo, aparecen datos inconsistentes, donde el código de agrupación de oradores no permite determinar exactamente los candidatos que participaron en el discurso, como ‘Multiple Speakers’, o ‘Democratic Candidates’. Estos errores en la calidad de los datos pueden afectar el conteo de discursos por candidatos realizado más adelante.

En la columna *location*, aparecen nuevamente inconsistencias en los datos. Algunos datos corresponden a cadenas de TV, como ‘Fox News’ o ‘ABC’, pero no queda claro si es un discurso realizado en el canal de forma presencial o virtual. Asimismo, existen locaciones más o menos específicas, como ‘Wisconsin’ y ‘Milwaukee, Wisconsin’, o ‘Darby, Pennsylvania’ y ‘Pennsylvania’. Estas inconsistencias podrían modificar los análisis sobre la campaña, como por ejemplo, el impacto de los discursos presenciales o el interés de realizar un discurso en algún sitio particular por el suceso de algún evento.

**2.3. Discursos por candidatos**

Los candidatos que lideran la cantidad de discursos son los más populares y los más votados (Tabla 3).

|  |  |
| --- | --- |
| Candidatos | Cantidad de discursos |
| Joe Biden | 81 |
| Donald Trump | 54 |
| Mike Pence | 20 |
| Bernie Sanders | 17 |
| Kamala Harris | 16 |

Tabla 3: Cantidad de discursos por candidato

Por otro lado, la campaña se intensifica en los 3 meses previos al 3 de noviembre (Figura 1).

El pico de discursos de Trump y Biden en septiembre coincide con el debate presidencial. Los dos principales candidatos a la vicepresidencia Harris y Pence debatieron en octubre, como se observa en la figura 1. [4]

Joe Biden tiene una presencia más constante en su campaña, manteniendo un compromiso constante. Donald Trump, por otro lado, tiene una estrategia que se concentra sobre el final de la campaña.

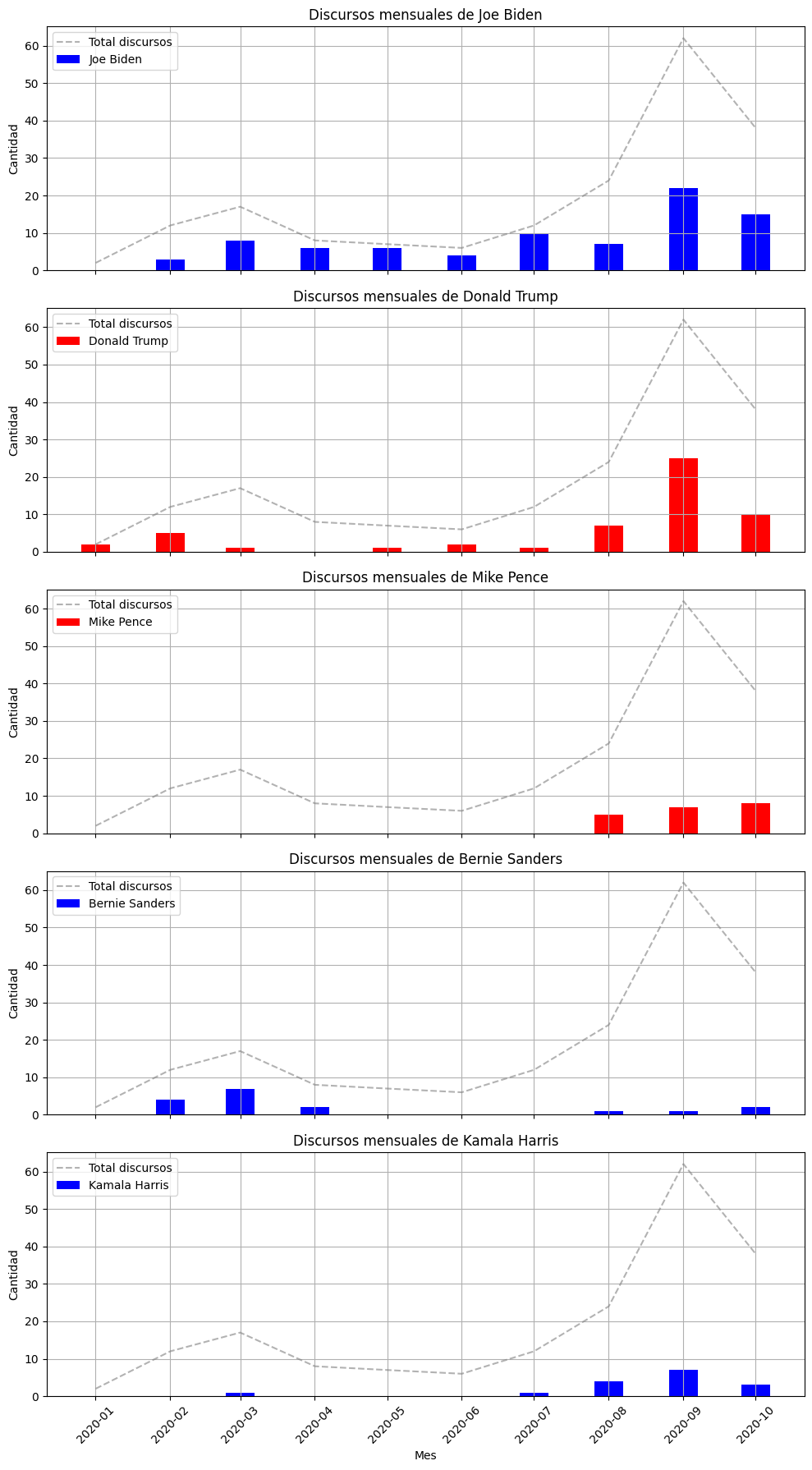


Figura 1: Discursos por candidatos a lo largo del tiempo

**2.4. Normalización del texto y conteo de palabras**

Al tratarse de una transcripción de un discurso hablado, es de esperarse que las palabras más frecuentes sean conjunciones y conectores. Esta tendencia es común entre los cinco candidatos analizados. (Figura 2)

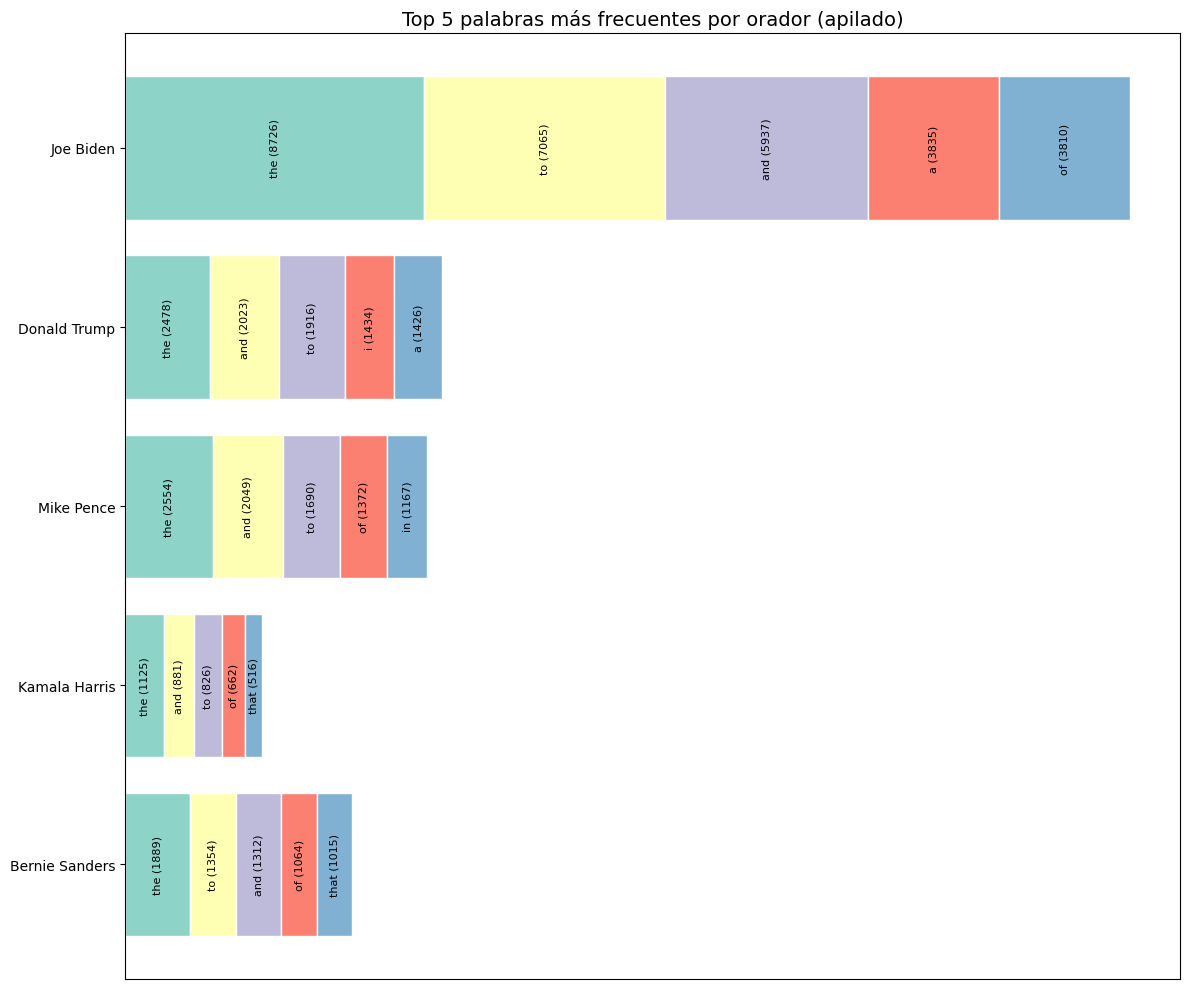


Figura 2: Cantidad de palabras por candidato

Debido a que Joe Biden tiene mayor cantidad de discursos, es esperable que tenga mayor cantidad de palabras. Sin embargo, la diferencia en la cantidad de palabras no se explica únicamente por este factor. En primer lugar, un 65% de los discursos de Donald Trump (el segundo candidato con más discursos) no han sido transcritos, lo que genera una representación incompleta de su participación, especialmente si se compara con Joe Biden, quien solo tiene un 14% de discursos sin transcribir. Por otro lado, los discursos transcritos de Joe Biden tienden a ser más extensos que los de Trump, lo cual se traduce en una mayor cantidad de conjunciones. En el caso de los otros candidatos, la cantidad de discursos es considerablemente menor, por lo que los factores mencionados anteriormente no tienen un impacto significativo en los resultados. (Tabla 4)

|  |  |  |  |
| --- | --- | --- | --- |
| Candidatos | Discursos no transcritos | Discursos vacíos (%) | Promedio de palabras por discurso |
| Donald Trump | 35 | 65 | 1229 |
| Joe Biden | 14 | 17 | 2323 |
| Kamala Harris | 2 | 13 | 1465 |
| Bernie Sanders | 1 | 6 | 2357 |
| Mike Pence | 1 | 5 | 2512 |

Tabla 4: Características de los discursos por candidatos

Para realizar un análisis más enriquecedor acerca del enfoque que cada candidato da a sus discursos, se seleccionaron distintos ejes temáticos con palabras claves relacionadas a los mismos. Los ejes temáticos y sus palabras claves se muestran a continuación:

**COVID:** covid, coronavirus, pandemic, vaccine, vaccination, mask, quarantine, lockdown, health crisis.

**Economía:** economy, jobs, employment, tax, unemployment, business, market.

**Justicia racial:** race, racism, racial, justice, equality, black lives matter, blm, discrimination, civil rights, protest.

**Salud:** healthcare, insurance, obamacare, medicare, hospital, doctor, patient, public option.

**Cambio climático:** climate, environment, global warming, pollution, green energy, renewable, sustainability.

**Inmigración:** immigration, border, wall, deportation, visa, migrant, undocumented.

**Seguridad pública:** police, crime, violence, gun, shooting, safety.

**Política exterior:** china, russia, iran, north korea, allies, nato, trade war, treaty.

**Corte Suprema:** supreme court, justice, judge, constitution.

**Derechos civiles:** equal rights, lgbtq, women rights, freedom.

**Educación:** education, school, student, college, university, tuition, remote learning.

En la distribución porcentual de los temas mencionados por los candidatos se observa que la economía es el tema central de la campaña, lo cual tiene sentido considerando el peso que tiene Estados Unidos en el mercado global y el impacto que tienen las decisiones económicas del país. (Figura 3)

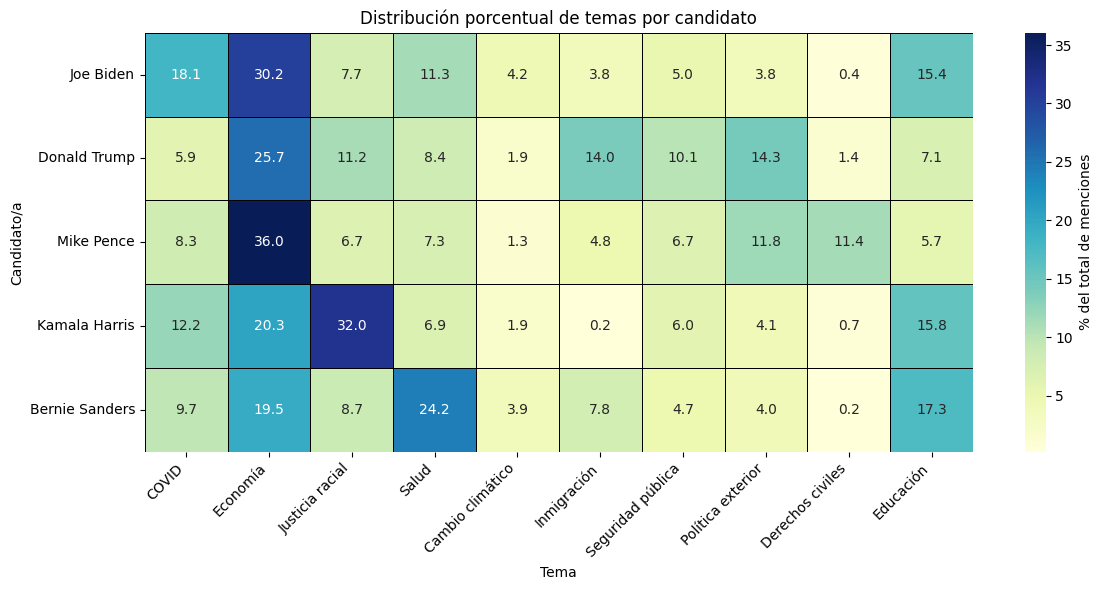


Figura 3: Distribución de temas por candidato

Los discursos de Joe Biden se centraron en la economía, la pandemia de COVID-19 y la educación. Kamala Harris abordó temas como la justicia social, la economía y la educación. Por su parte, Bernie Sanders hizo énfasis en la salud, la economía y la educación.

En cuanto a los republicanos, Donald Trump enfocó sus intervenciones en la economía, la política exterior y la inmigración, mientras que Mike Pence destacó la economía, la política exterior y los derechos civiles.

Los candidatos demócratas coinciden en su enfoque en la economía y la educación, temas alineados con su énfasis en la equidad social, mientras que los republicanos muestran un interés común en la economía y la política exterior, en línea con su visión de libre mercado y seguridad nacional.

A pesar de tener una gran industria, el cambio climático y la agenda ambientalista no parece ser de gran importancia para la sociedad estadounidense.

Es importante destacar que esta representación está basada en el conteo de palabras clave asociadas a cada eje temático. Esto introduce un sesgo en el análisis: la selección de términos puede dejar fuera sinónimos o formas indirectas de mencionar los temas. Además, no se considera el tono con el que se habla de cada asunto (si es positivo, negativo o neutro), ni la duración total de los discursos.

**3.5. Generación de matriz de menciones entre candidatos**

En la Tabla 5 se muestra el conteo de menciones de un candidato a otro. Las filas representan el candidato y las columnas a quien es mencionado. Por ejemplo, Joe Biden mencionó a Donald Trump 420 veces. Cabe destacar que Trump es el candidato más mencionado por todos los políticos, lo cual puede deberse a su personalidad provocadora.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Joe Biden | Donald Trump | Bernie Sanders | Kamala Harris | Mike Pence |
| Joe Biden | 0 | 420 | 8 | 23 | 3 |
| Donald Trump | 110 | 0 | 7 | 4 | 2 |
| Bernie Sanders | 61 | 127 | 0 | 1 | 0 |
| Kamala Harris | 35 | 36 | 0 | 0 | 4 |
| Mike Pence | 249 | 312 | 5 | 41 | 0 |

Tabla 5: Matriz de menciones entre candidatos

1. **Reflexiones finales**

Este trabajo muestra cómo la ciencia de datos puede ayudarnos a entender mejor qué están diciendo los políticos y hacia dónde apuntan sus discursos. Con herramientas simples, como contar palabras clave y visualizar los resultados, se puede tener una idea bastante clara de los temas más mencionados y las prioridades que se repiten.

Pero también es importante tener en cuenta que los datos no hablan por sí solos: todo análisis está influido por las decisiones que tomamos, como qué palabras elegir o cómo agrupar los temas. Eso quiere decir que, según cómo armemos el análisis, podemos llegar a conclusiones distintas.

1. **Bibliografía**

[1] Armour, S., & Siddiqui, S. (2020, September 25). *Supreme Court vacancy makes health care leading issue in 2020 elections*. The Wall Street Journal.<https://www.wsj.com/politics/policy/supreme-court-vacancy-makes-health-care-leading-issue-in-2020-elections-11600948807>

[2] Bouie, J. (2020, June 3). *Trump or Biden? George Floyd has shown us the answer*. The New York Times.<https://www.nytimes.com/2020/06/03/opinion/george-floyd-trump-biden.html>

[3] Hersher, R. (2020, October 22). *There’s a lot at stake for the climate in the 2020 election*. NPR.<https://www.npr.org/2020/10/22/925468716/theres-a-lot-at-stake-for-the-climate-in-the-2020-election>

[4] Wikipedia contributors. (s.f.). *Debates presidenciales de Estados Unidos de 2020*. Wikipedia.<https://es.wikipedia.org/wiki/Debates_presidenciales_de_Estados_Unidos_de_2020>