Tarea 1

Introducción a la Ciencia de Datos

Autores: Juan Pablo de Souza - Jonathan Ramírez
Título del conjunto de datos: *US 2020 Presidential Election Speeches*

Carga y limpieza de Datos

Descripción del Dataset

El presente trabajo se basa en el conjunto de datos *US 2020 Presidential Election Speeches*, disponible públicamente en la plataforma Kaggle. Este dataset recopila transcripciones de discursos pronunciados por diversas figuras políticas estadounidenses durante la campaña presidencial del año 2020.

Las transcripciones fueron obtenidas mediante técnicas de web scraping desde el sitio Rev.com, una plataforma reconocida por su precisión en transcripciones audiovisuales. A su vez, se complementaron manualmente con datos adicionales como el lugar, título y tipo de evento correspondiente a cada discurso.

El dataset contiene un total de **269 filas**, cada una correspondiente a un discurso, distribuidos en las siguientes columnas (Tabla 1):

Columna	**Descripción**
speaker	Nombre del político que pronunció el discurso
title	Título o descripción breve del discurso
text	Transcripción completa del discurso
date	Fecha en la que fue pronunciado el discurso
location	Ciudad y estado donde se dio el discurso
type	Tipo de evento (por ejemplo, campaña, debate, entrevista)

Tabla 1

A continuación, se muestra un fragmento de la tabla, en el cual se puede apreciar cómo es cada entrada del *dataframe*. Ilustración 1

speaker	title	text	date	location	type
David Perdue	Georgia Sen. David Perdue Speech Transcript at Trump Rally: Mocks & Mispronounces Kamala Harris' Nam...	David Perdue: (00:01) How great is it to be back in Macon, Georgia with Donald Trump coming to Macon...	Oct 16, 2020	Macon, Georgia	Campaign Speech
Joe Biden	Joe Biden Southfield, MI Speech on Health Care October 16	Joe Biden: (00:00) Hello, Michigan. Hi, how are you? What's your name? Sam, good to see you, man. Th...	Oct 16, 2020	Southfield ,Michigan	Campaign Speech
Donald Trump	Donald Trump Speech Transcript 'Protecting	President Trump: (00:30) Thank you. What a nice group.	Oct 16, 2020	Fort Myers, Florida	Campaign Speech

Ilustración 1

Inconsistencias detectadas:

Si bien el dataset fue curado manualmente, presenta varias inconsistencias producto del scraping o de la recopilación original:

- En algunos casos, el texto comienza con encabezados redundantes que repiten el nombre del orador, la ubicación o la fecha.
- En varios discursos, especialmente debates, se incluyen múltiples oradores sin una identificación clara en el texto.
- La columna speaker a veces contiene varios nombres separados por comas (ej: "Pete Buttigieg, Amy Klobuchar, O'Rourke"); esto ocurre normalmente en entrevistas o debates.
- Además, se detectaron valores nulos en algunas columnas, como se muestra a continuación (Tabla 2):

Columna	Valores nulos	Porcentaje
speaker	3	1.1%
title	0	0%
text	0	0%
date	0	0%
location	18	6.7%
type	21	7.8%

Tabla 2

Las columnas más afectadas son location y type, mientras que las más relevantes para el análisis (speaker, text, date) están casi completas. En el siguiente histograma se aprecia esta diferencia visualmente (Ilustración 2).

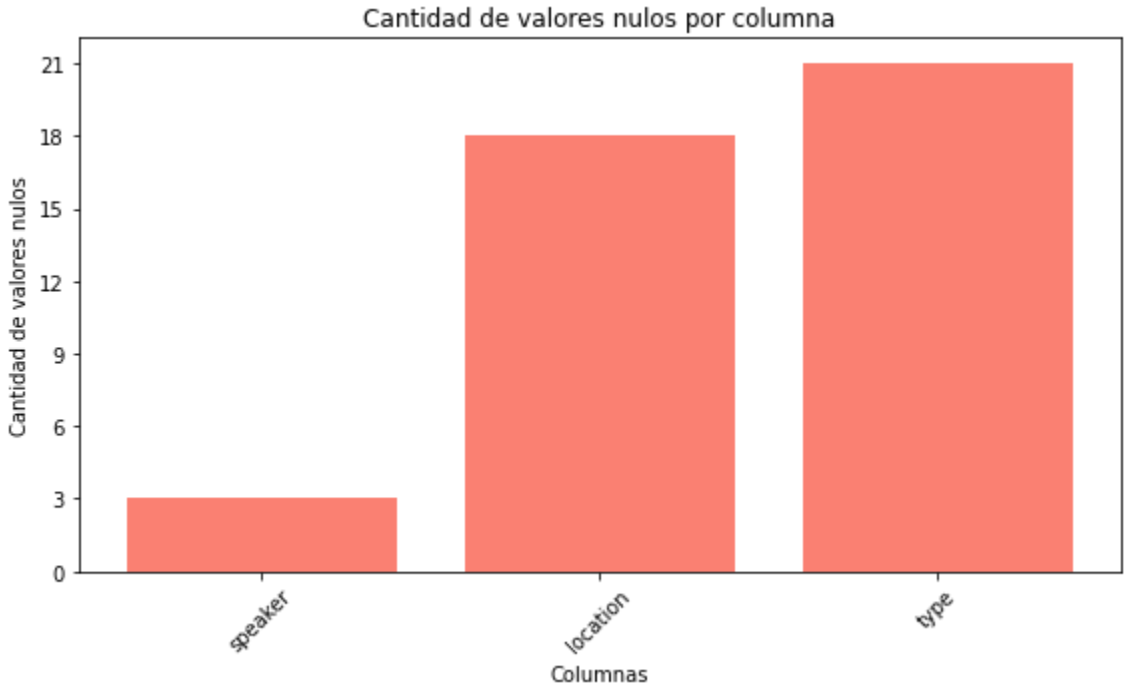

Ilustración 2

Al eliminar los registros con valores nulos en speaker, se pierde solo un 1.1% del total, lo cual es una pérdida aceptable.

Discursos con Múltiples Oradores

Se encontraron discursos con múltiples oradores que representan aproximadamente el 5.20% del total. Aunque son interesantes desde una perspectiva contextual, representan un desafío para el análisis textual automatizado, ya que dificultan la asignación de contenido a un solo autor. En la siguiente tabla se observan ejemplos (Tabla 3):

Ejemplos de múltiples oradores	**Tipo de evento**
Lindsey Graham, Jaime Harrison	Debate
Joe Biden, Kamala Harris	Campaign / Entrevista
Kamala Harris, Mike Pence	Debate
Pete Buttigieg, Amy Klobuchar, O'Rourke	Endorsement

Tabla 3

Dado que este tipo de discursos introduce ambigüedad, se decidió eliminarlos para el análisis textual, priorizando la claridad en la autoría.

Análisis de Frecuencia de Discursos por Candidato/a

Se examinó la cantidad total de discursos realizados por cada candidato/a. Este recuento permite identificar a las figuras más activas.

En la siguiente grafica (Ilustración 3) se puede apreciar la cantidad de discursos por los cinco candidatos/as con mayor presencia oratoria.

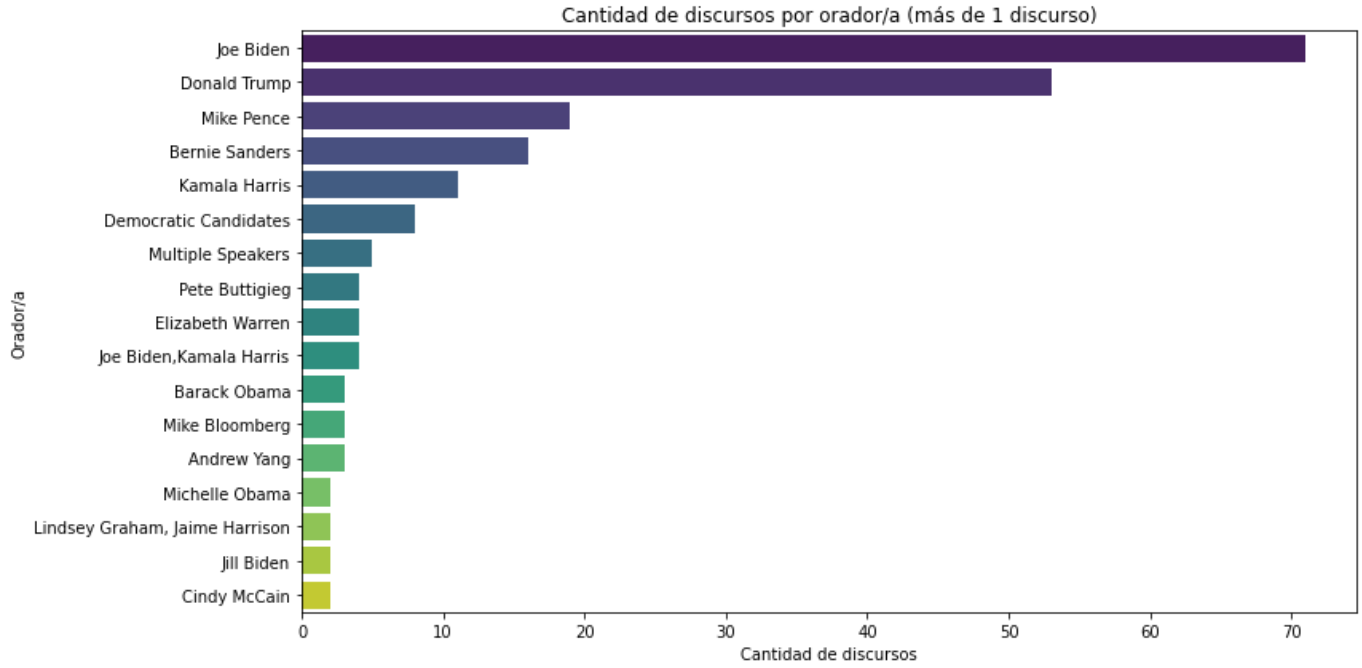

Ilustración 3

A continuación (Tabla 4), presentamos en una tabla los oradores que cuentan con un solo discurso en el dataset.

Chuck Schumer	Nancy Pelosi	Bill Clinton	Alexandria Ocasio-Cortez
Colin Powell	David Perdue	Hillary Clinton	Gavin Newsom
Andrew Cuomo	Cory Booker	John Kasich	Val Demings
Gretchen Whitmer	Kanye West	Amy Klobuchar	Michael Bloomberg
Tulsi Gabbard	Donna Brazile	Sarah Cooper	Rand Paul
Jim Jordan	Nikki Haley	Ivanka Trump	Tom Cotton
Rudy Giuliani	Ben Carson	Mitch McConnell	Chen Guangcheng
Lara Trump	Lou Holtz	Karen Pence	Jack Brewer
Kellyanne Conway	Kayleigh McEnany	Dan Crenshaw	Pam Bondi
Melania Trump	Mike Pompeo	Eric Trump	Nicholas Sandmann
Tiffany Trump	Tim Scott	Kimberly Guilfoyle	Herschel Walker
Donald Trump Jr.			

Tabla 4

Distribución de Discursos: Top 5 Oradores

Se identificaron los cinco oradores con mayor cantidad de discursos, obteniendo un subconjunto de 170 discursos (67.5% del total de discursos filtrados). Esta selección permite concentrarse en los actores políticos más relevantes durante la campaña. A continuación (Ilustración 4), se muestra la distribución de discursos y su porcentaje en una gráfica de torta:

- Joe Biden: 71 discursos
- Donald Trump: 53 discursos
- Mike Pence: 19 discursos
- Bernie Sanders: 16 discursos
- Kamala Harris: 11 discursos

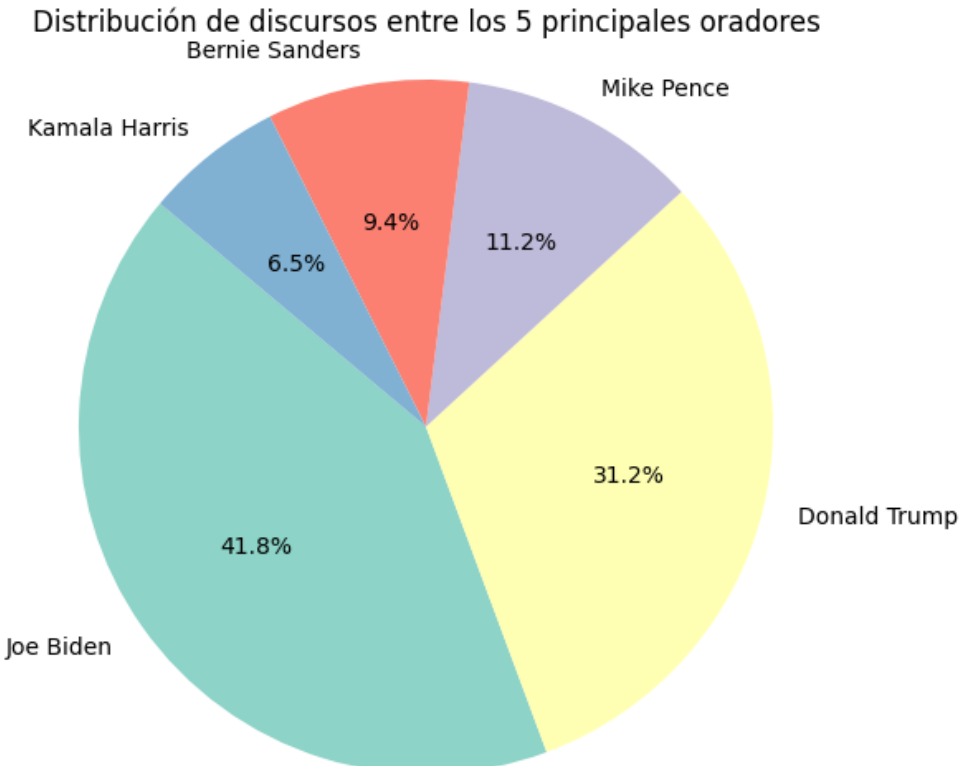

Ilustración 4

Observaciones:

- Joe Biden lidera con 71 discursos, un 25% más que Trump.
- Trump mantiene una presencia sustancial.
- Pence, Sanders y Harris tienen una participación menor.

- Biden y Trump concentran el 72.9% del total de discursos seleccionados.

Esta distribución de discursos tiene sentido considerando ya que, Trump y Biden eran candidatos presidenciales por eso es lógico que hayan realizado más discursos que el resto. Trump, al ser presidente en funciones, no necesitaba tantos discursos. Biden, como oposición, debía tener una presencia comunicacional más fuerte. Pence, Sanders y Harris cumplieron roles secundarios en la campaña.

Análisis contextual de los discursos

En la siguiente gráfica (Ilustración 5), se observa la evolución de la cantidad de discursos de los candidatos a lo largo de las diferentes semanas de campaña.

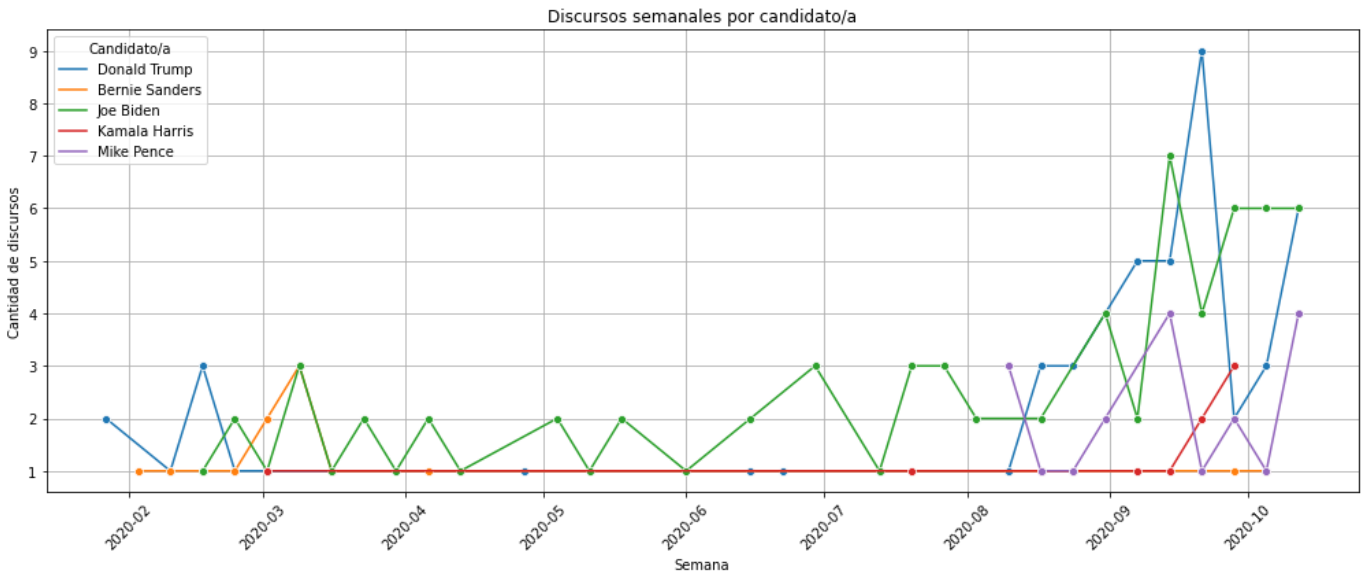

Ilustración 5

Evolución Temporal de los Discursos

Durante las elecciones primarias (enero a marzo de 2020), hay mayor diversidad de oradores, especialmente del partido demócrata.
Desde abril, con la consolidación de Biden como candidato, se incrementa su frecuencia de discursos. Trump, sin competencia interna, mantiene una frecuencia más estable, intensificando recién en los meses previos a la elección.

Impacto del COVID-19

Desde marzo de 2020, la pandemia influye fuertemente en el contenido y formato de los discursos:

- Biden adopta un tono empático en sus discursos, centrado en la salud pública y la unidad nacional.
- Trump minimiza el impacto de la pandemia, haciendo foco en la recuperación económica.
- Se reducen los eventos presenciales, aumentando los discursos virtuales.

Protestas sociales y raciales

Luego del asesinato de George Floyd en mayo de 2020, los discursos reflejan posiciones opuestas:

- Biden promueve reformas y apela a la unidad
- Trump refuerza el mensaje de “ley y orden”

Conclusiones del análisis contextual:

La distribución de discursos es desigual: Biden tiene un 25% más de discursos que Trump y ambos superan ampliamente al resto de los oradores.

Además, el contexto político —pandemia y protestas— influyó fuertemente en el contenido y la estrategia discursiva.

Biden tuvo una estrategia más constante en el tiempo, intensificando su actividad en septiembre, mientras que Trump comenzó a aumentar su presencia oratoria recién en agosto.

Proceso de limpieza de datos

Para mejorar la calidad del dataset se aplicaron los siguientes pasos.

- Eliminación de entradas con valores nulos en speaker.

Después de este proceso, se conservaron 252 discursos, lo que implicó una pérdida del 6.3% del total original de datos. Esta pérdida es razonable, considerando la mejora en la calidad del dataset.

Otros procesos de limpieza aplicados:

- Conversión de todos los textos a minúsculas
- Eliminación de signos de puntuación y caracteres especiales
- Remoción de encabezados redundantes en la columna text
- Verificación de que date sea del tipo datetime

Estos pasos permiten que el análisis de los discursos sea lo más claro y preciso posible. Más adelante, se aplicarán procesos adicionales que serán explicados oportunamente.

Conteo de palabras y Visualizaciones

Estrategia para el Análisis de Discursos

El objetivo central de esta sección es analizar las ideas predominantes de cada candidato/a mediante el conteo de palabras más frecuentes, partiendo de la premisa de que los términos repetidos con mayor frecuencia reflejan la esencia del mensaje que cada orador intenta transmitir.

Sin embargo, antes de realizar este análisis, se detectó un desequilibrio significativo en la cantidad de discursos disponibles por cada candidato/a. En total, se cuenta con 170 discursos distribuidos entre los cinco oradores principales, pero con una fuerte disparidad: Joe Biden posee 71 discursos, mientras que Kamala Harris solo 11. Esta diferencia podría distorsionar el análisis, ya que un mayor volumen de discursos aumenta naturalmente la aparición de determinadas palabras, sin necesariamente reflejar su importancia relativa.

Para abordar este problema, se plantearon tres posibles estrategias:

1. **Mantener el dataset original sin modificaciones**
   Esta opción conserva todos los discursos disponibles, lo cual permite capturar completamente el estilo y los temas predominantes de cada candidato. Sin embargo, favorece a quienes tienen más discursos, ya que se analizará un volumen mayor de texto para ellos.
2. **Equilibrar la cantidad de discursos entre candidatos**
   Se propuso tomar la misma cantidad de discursos por cada orador (11 por cada uno en este caso, limitados por el orador de menos intervenciones). Esta estrategia mejora la equidad en la comparación, ya que garantiza que todas las frecuencias de palabras se basen en volúmenes de texto equivalentes. No obstante, implica eliminar más de 100 discursos, lo cual puede conducir a una pérdida considerable de información valiosa, sobre todo en el caso de los candidatos con más intervenciones.
3. **Nivelar parcialmente: igualar Biden y Trump, mantener los demás**
   Esta opción busca un equilibrio intermedio. Se reduce la cantidad de discursos de Joe Biden a 53, igualando los de Trump, pero se mantiene sin cambios al resto. De este modo, se logra reducir el sesgo de volumen entre los dos candidatos con mayor presencia, pero sin sacrificar tanta información como en la estrategia anterior.

La estrategia que consideramos más adecuada para realizar el conteo de palabras y las tareas del análisis fue igualar la cantidad de discursos entre Biden y Trump, manteniendo sin cambios al resto de los candidatos. Esta opción permite reducir el sesgo generado por la disparidad en el número de discursos entre los dos principales oradores, mejorando la comparabilidad directa entre ellos, que fueron los protagonistas centrales de la campaña. Al mismo tiempo, se evita una pérdida excesiva de información como en la estrategia que iguala a todos los candidatos, ya que solo se descartan 18 discursos de Biden en lugar de más de 100. Así, se logra un equilibrio entre representatividad, equidad en la comparación y preservación del contenido original del dataset. (Ilustración 6)

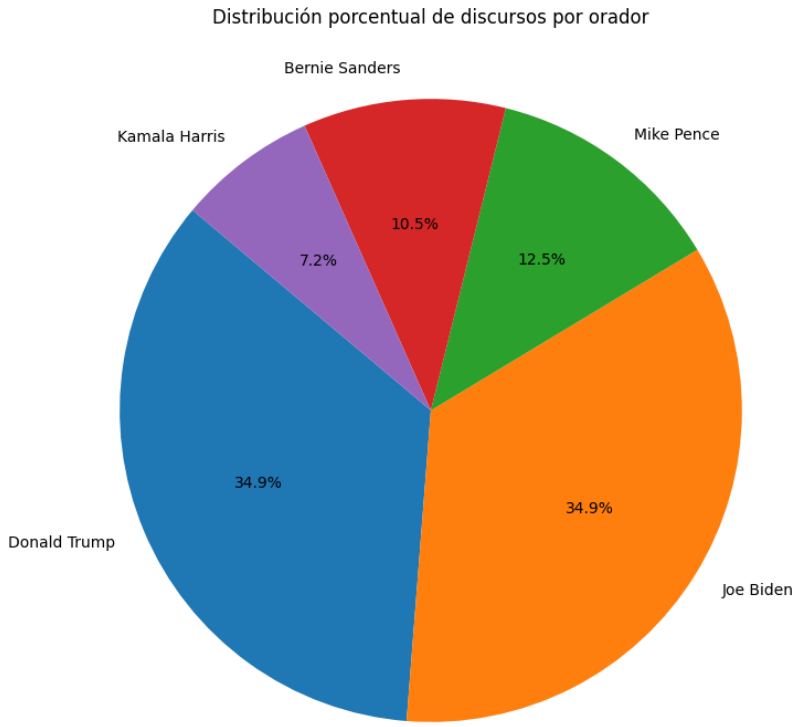

Ilustración 6

Visualización de las Palabras Más Frecuentes por Candidato/a

Para identificar las ideas principales de cada uno de los cinco candidatos/as con más discursos, se realizó un análisis de frecuencia de palabras. Antes de generar la visualización, se aplicaron procesos de limpieza al texto: eliminación de stopwords, lematización, y filtrado de caracteres no alfabéticos, asegurando así que las palabras seleccionadas reflejen contenido relevante y no términos vacíos. La siguiente gráfica

(Ilustración 7) muestra las 10 palabras más frecuentes empleadas por cada orador, lo que permite comparar sus enfoques discursivos y prioridades temáticas.

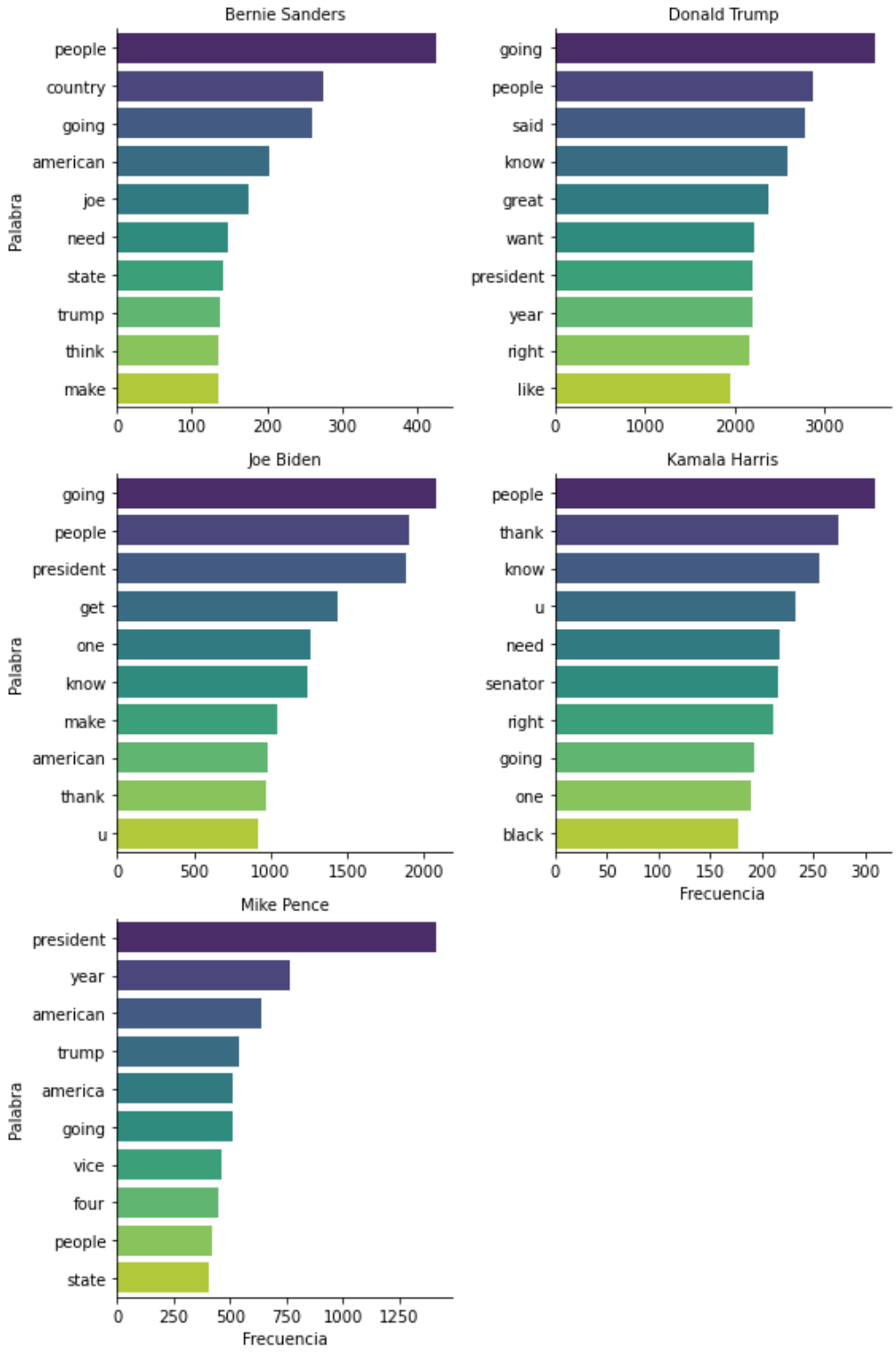

Ilustración 7

Aunque las diez palabras mas utilizadas en cada discurso no permiten sacar conclusiones definitivas, a simple vista se pueden observar las siguientes ideas:

- Bernie Sanders (integrante del partido demócrata) habla demasiado sobre Trump, ya que aparece entre las diez primeras palabras. también utiliza verbos como Necesitar y hacer, los cuales remarcan que el gobierno actual está en falta. También aparece la palabra “going” que se utiliza normalmente para hablar de futuro. Se puede suponer que como en toda campaña electoral se prometen cosas a cambiar, mejorar para el país, las personas.
- Del mismo modo que Bernie, los otros integrantes del partido demócrata (Biden y Kamala) utilizan los verbos necesitar, hacer, saber, de forma que se podrían tratar de críticas al gobierno actual de ese entonces.
- Por otro lado, Trump (presidente en funciones de ese momento) dice palabras como dijeron, genial, cierto mostrando otro enfoque en sus discursos, no tanto como la critica de la oposición, sino remarcando las cosas que se hicieron en el gobierno.
- Por último, Mike Pence, vicepresidente de Trump, habla de su rol, ya que aparece la palabra “vice” y tiene el mismo estilo de discursos que Trump.

Conteo de Palabras por Candidato

Para poder lograr un acertado conteo de palabras de cada candidato, se propuso dividir los textos de cada discurso en listas de palabras de modo de evitar los valores nulos y que al dividir de otra forma queden varias palabras juntas. Otro cuidado que se tuvo que realizar es que la columna speaker este correctamente formateada, ya que un formato erróneo podría afectar el proceso de agrupación de palabras. De esta forma, una vez solucionado los potenciales problemas de los datos, se podría calcular la cantidad total de palabras por cada candidato y ordenar a ellos de mayor cantidad a menor.

En la siguiente grafica (Ilustración 8) se puede observar la cantidad de palabras que utilizo cada candidato en todos sus discursos. Vale aclarar que para este análisis se utilizaron los 170 discursos de los 5 candidatos con más discursos.

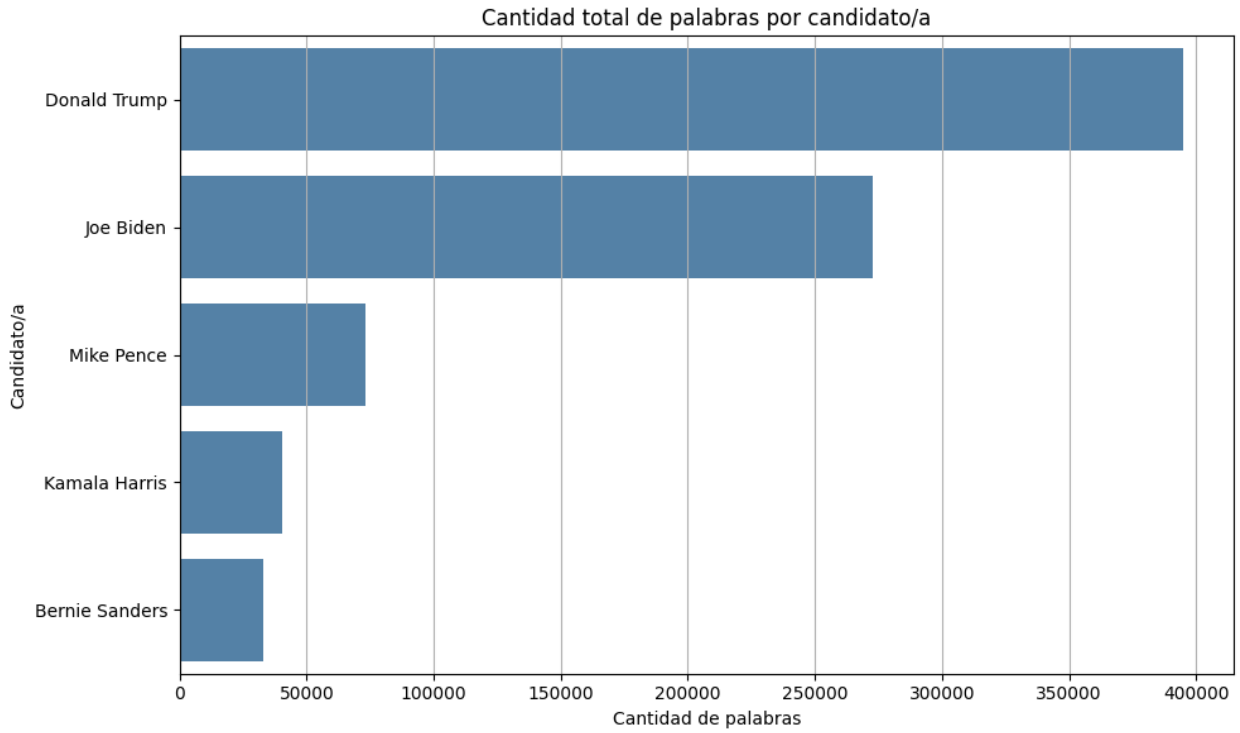

Ilustración 8

A simple vista se puede apreciar que, si bien Trump no es el candidato con mas discursos, es el candidato que utilizo mas palabra, esto nos lleva a pensar que los discursos de Trump fueron potencialmente mas largos que los discursos de los demás candidatos.

También, se puede ver que los discursos de Bernie, Kamala y Mike son insignificantes tanto en cantidad de discursos como en duración, en comparación con los de Trump o Biden. Esto nos termina de confirmar que ellos tres fueron apoyo en las campañas de Trum y Biden respectivamente

Análisis de Menciones entre Candidatos

En esta parte del informe, el objetivo es, basado en los cinco candidatos con mas discursos, observar cuanto se nombran los unos a los otros y cuanto se nombran a ellos mismos. Para eso, primero se construyo la siguiente matriz de correlación. Donde cada entrada de la matriz (i,j) corresponde a la cantidad de veces que el candidato i de la fila, nombro al candidato j de la columna correspondiente. En la diagonal (Ilustración 9) se puede observar la cantidad de veces que cada candidato se nombró a sí mismo.

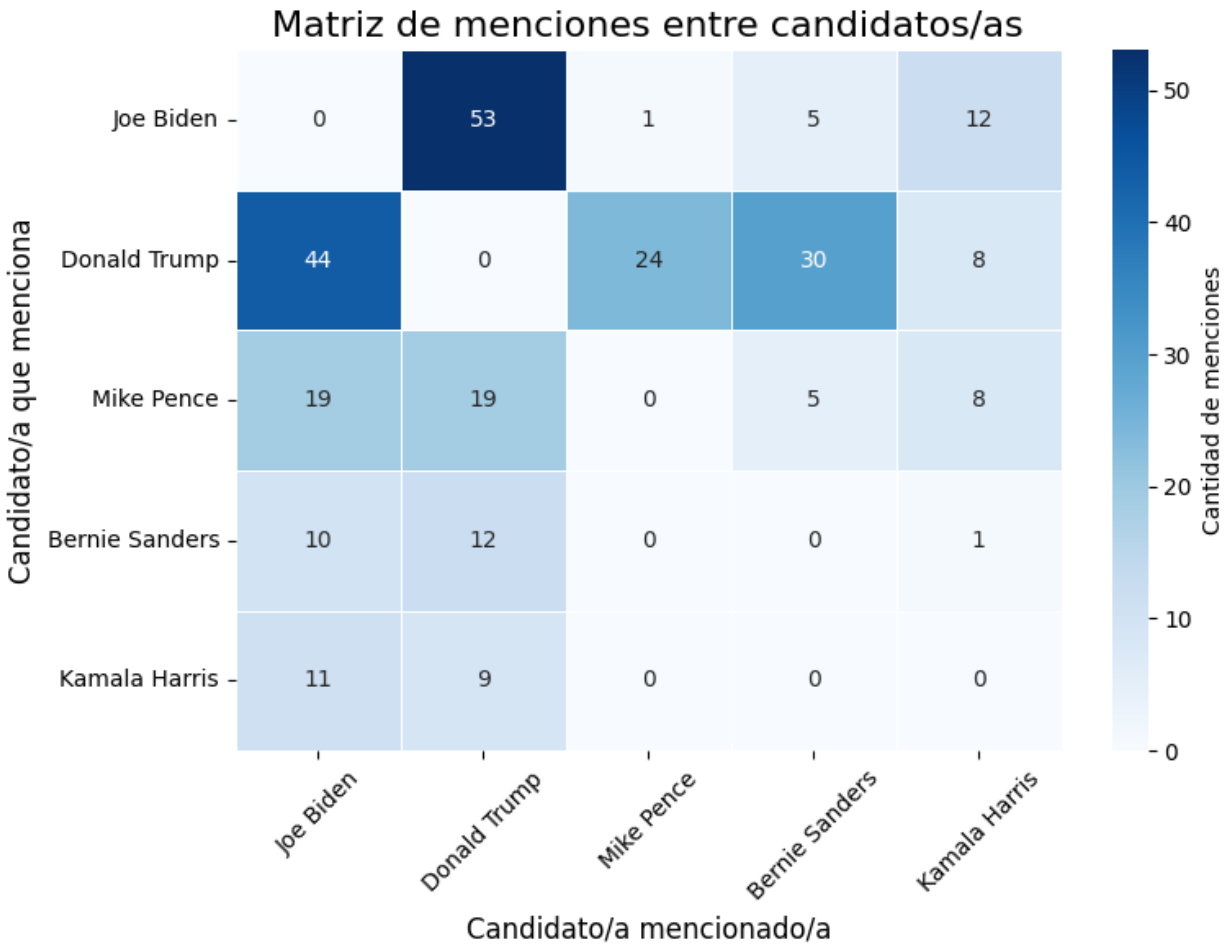

Ilustración 9

En la matriz de menciones de candidatos se puede observar que claramente Trump y Biden se nombraron mutuamente, ya que eran los candidatos a la presidencia. También, se observa que todos los otros oradores nombraron mayoritariamente a Trump y Biden y no se nombraron entre ellos.

Adicionalmente se puede ver que Trump menciono a todos los otros oradores, aunque se enfocó más en los partícipes de la oposición que en su vicepresidente de aquel momento Mike.

Sin embargo, Biden se dedicó a nombrar a Trump (su contrincante en las elecciones) y a su compañera de fórmula Kamala, dejando a los otros al margen de sus discursos.

Otra forma de poder observar la información anterior mas gráficamente, es mediante un grafo dirigido. Los grafos, son estructuras de datos que se utilizan para representar relaciones entre esos datos. Se componen de vértices (los datos) y aristas que conectan los vértices.

Para representar de esta forma las menciones entre oradores, se optó por construir un grafo dirigido, donde cada nodo representa a un candidato y las aristas indican las menciones entre ellos. El peso de cada arista se corresponde con la cantidad de menciones realizadas entre los candidatos. Esto permite visualizar no solo quién menciona a quién, sino también la intensidad de esas menciones.

A continuación (Ilustración 10), se muestra el grafo dirigido basado en la matriz de menciones:

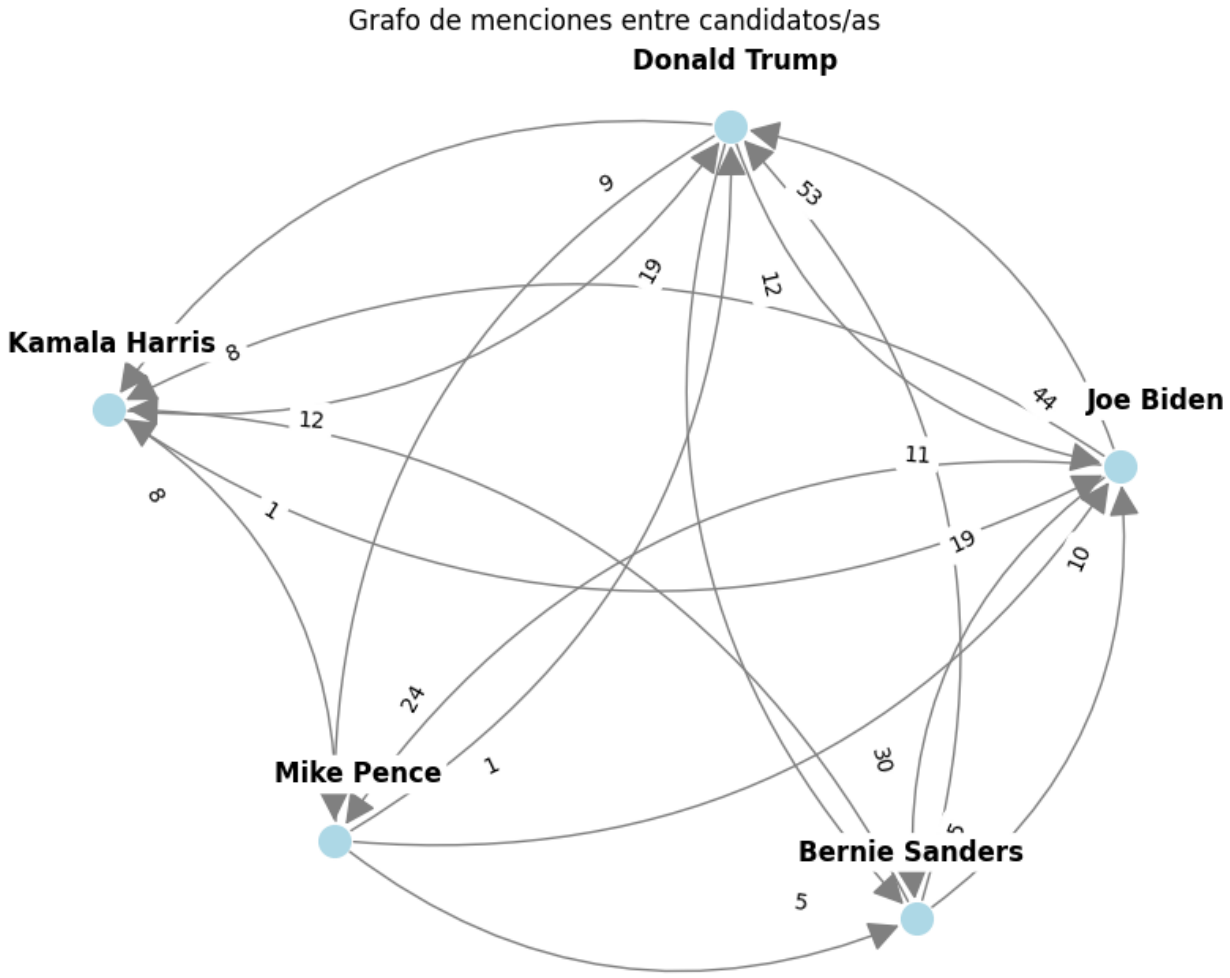

Ilustración 10

Este grafo es útil para entender las dinámicas entre los candidatos, ya que las flechas entre candidato y candidato significan “candidato x nombra a candidato y” según para donde apunta cada flecha. Este grafo permite identificar cuáles son los candidatos más mencionados o quiénes tienden a interactuar más entre sí a través de menciones directas.

Preguntas de análisis de los discursos

A lo largo de la presente tarea, se nos ocurrieron diferentes preguntas que se podrían responder con este juego de datos, debido a la escasez de tiempo y la profundidad de la materia, no se pudo analizar a fondo todo lo que se podría hacer con el dataset.

Algunas de las preguntas que se pudieran responder con estos datos son:

1. ¿Cómo ha cambiado el tono de los discursos de cada candidato a lo largo de la campaña, especialmente durante eventos clave como la pandemia de COVID-19 o las protestas sociales?

- Análisis de Sentimiento: Podríamos utilizar herramientas como TextBlob o VADER para analizar si el tono de los discursos de los candidatos varió con el tiempo, evaluando si se tornaron más negativos o positivos durante momentos críticos de la campaña, como cuando comenzó la pandemia o después del asesinato de George Floyd.
- Análisis Temporal: Se podría dividir la campaña en distintas etapas (antes y después de la pandemia, por ejemplo) para ver cómo cambia el tono de los discursos en relación con los eventos que van ocurriendo.
- Comparación entre Candidatos: También podríamos comparar cómo los candidatos Biden y Trump, por ejemplo, usan el tono de forma diferente en los mismos eventos, lo que podría reflejar sus estrategias de comunicación.

2. ¿Los discursos de los candidatos muestran signos de polarización? ¿Cómo se relacionan las menciones que hacen entre sí con este fenómeno?

- Menciones Directas: Usando la matriz de menciones y el grafo, podemos investigar si los candidatos tienden a mencionar más a sus rivales de forma directa. Si es así, podríamos ver si esto está relacionado con un tono más agresivo o polarizado.
- Tópicos y Temas: Con el análisis de tópicos (por ejemplo, LDA), podríamos ver si ciertos candidatos tienden a centrarse en temas más polarizados, como la inmigración o el control de armas, y cómo eso se refleja en sus menciones hacia otros candidatos.

3. ¿Cuáles son los principales temas de cada candidato y cómo varían a lo largo de la campaña?

- Podríamos realizar un análisis de las palabras más frecuentes en los discursos de cada candidato en diferentes momentos de la campaña. Esto nos ayudaría a ver cómo sus prioridades y temas de campaña cambiaron con el tiempo.

- Utilizando herramientas como LDA, se podrían extraer los temas principales de los discursos de cada candidato, observando cómo estos temas cambian durante las distintas fases de la campaña (por ejemplo, primarias, convenciones, y periodo electoral).

Conclusión

El conjunto de datos proporciona una visión integral de las estrategias discursivas empleadas durante la campaña presidencial de EE. UU. en 2020, permitiendo identificar patrones clave en la frecuencia, longitud y contenido de los discursos. A través de un proceso meticuloso de limpieza de datos, fue posible extraer información relevante que abre la puerta a un análisis más profundo de las tácticas discursivas utilizadas por los candidatos.

Sin embargo, este análisis presenta algunas limitaciones que deben ser tenidas en cuenta. En primer lugar, la disparidad en la cantidad de discursos pronunciados por cada candidato podría sesgar los resultados, dado que los candidatos con más discursos proporcionan una mayor cantidad de datos, lo que podría llevar a interpretaciones desproporcionadas en su favor. Además, la presencia de entradas con múltiples oradores y la falta de etiquetas partidarias explícitas dificultan una segmentación clara y precisa de los discursos según su ideología o afiliación política, lo cual sería clave para un análisis más detallado.

Este análisis, aunque útil, solo sienta las bases para investigaciones más profundas sobre el discurso político. En etapas futuras, sería valioso explorar no solo la estructura y los patrones discursivos, sino también aspectos más complejos como el tono del discurso, las prioridades temáticas de los candidatos, el grado de polarización y la personalización del mensaje. Incorporar técnicas avanzadas de análisis de datos, como el análisis de sentimiento, la detección de entidades nombradas o el análisis semántico, permitiría una interpretación más matizada de los discursos, ayudando a captar las intenciones subyacentes y los enfoques estratégicos empleados por cada candidato.