INTRODUCCIÓN A LA CIENCIA DE DATOS

Informe Tarea 1

Calidad de Datos y Visualización

Análisis exploratorio de los datos, incluyendo la detección de problemas de calidad, la limpieza del texto y la generación de visualizaciones.

Ing. Isabella Buschiazzo

Ing. Lucía Gutierrez

Mayo 2025

Índice

1. Introducción

La recolección de datos es un paso crucial en la generación de conocimiento científico. Sin embargo, es importante tener presente que los datos por sí solos no tienen valor sin una gestión adecuada. La calidad de la información obtenida depende de cómo se gestionan esos datos, y esta gestión incluye diversas dimensiones, entre las que destacan el modelado, la calidad, la integración, así como la evaluación de aspectos éticos, de privacidad y de sesgo. Estas consideraciones son fundamentales para garantizar que los datos sean útiles, precisos y responsables, alineándose con las mejores prácticas en la ciencia y la investigación.

El presente trabajo se centra particularmente en el análisis de la calidad de los datos. Esta dimensión permite determinar qué tipo de información puede extraerse de un conjunto de datos y cuáles son sus limitaciones. Es común encontrar bases de datos que, si bien están disponibles, presentan desorden, incompletitud, inconsistencias o ruido. Dichos problemas pueden originarse en distintas etapas del ciclo de vida de los datos: producción, procesamiento, almacenamiento o utilización. Por lo tanto, mejorar la calidad de los datos requiere identificar la fuente de los problemas y aplicar medidas correctivas adecuadas. Sin embargo, en muchos casos se dispone únicamente de los conjuntos de datos, sin posibilidad de intervenir en su origen. El presente informe se enmarca en dicho contexto, por lo que el objetivo principal del mismo es plantear estrategias que permitan evaluar y mejorar la calidad de los datos disponibles, maximizando así su valor informativo a pesar de las limitaciones.

El análisis de calidad mencionado se realiza para un conjunto de datos que contiene un listado de discursos pronunciados por diferentes candidatos en las elecciones presidenciales de Estados Unidos del año 2020. Los datos recolectados incluyen: el nombre del candidato que pronunció el discurso, la fecha y el lugar junto con el título del discurso, su clasificación y una transcripción del mismo.

A modo de contexto, resulta interesante señalar que el año 2020 estuvo marcado por múltiples acontecimientos de gran relevancia política, económica y social en Estados Unidos. En primer lugar, en el marco de las elecciones presidenciales, hubo un cambio en el mando político del país, resultando electo como nuevo presidente Joe Biden (candidato del Partido Demócrata), y se llevó a cabo un juicio político (*impeachment*) contra el presidente cesante Donald Trump. Ese mismo año, se produjeron masivas y en ocasiones violentas protestas a lo largo del país tras el asesinato de George Floyd, un ciudadano afroamericano, a manos de un oficial de policía en la ciudad de Minneapolis. Por otra parte, se decretó la pandemia de COVID-19, generando una crisis sanitaria sin precedentes a nivel mundial.

Los datos utilizados en este trabajo incluyen la transcripción de 269 discursos (así como su clasificación, fecha y lugar de ocurrencia). A partir de este conjunto, se busca realizar una limpieza preliminar del texto, con el objetivo de extraer información relevante, así como presentar distintas visualizaciones a partir de los discursos de los cinco candidatos con mayor cantidad de intervenciones registradas (asociando esta frecuencia con su popularidad).

Los datos se cargaron en formato Data Frame dentro de un notebook de Jupyter.

2. Desarrollo

2.1. Calidad de datos

En primer lugar, se realizó una inspección visual general del Data Frame cargado con los discursos de los candidatos, con el objetivo de observar su estructura, clasificación y orden. Luego, se examinó la primera columna *speakers* del Data Frame, correspondiente a los nombres de los candidatos presidenciales que pronunciaron discursos durante el año 2020, y se identificaron irregularidades y errores.

Para evaluar la calidad de los datos en la columna *speakers*, se utilizó la función *value_counts* para visualizar los objetos (*inputs*) únicos registrados, así como la cantidad de apariciones de cada uno. De esta manera, se logró reducir los datos a analizar de 269 (cantidad de discursos totales registrados) a 72 (cantidad de nombres únicos de candidatos a presidente que pronunciaron dicho discurso), haciendo mucho más sencilla la inspección visual para detección de anomalías en los datos cargados.

En la Tabla 1 se muestra el resultado. Como se puede observar, figuran los nombres de varios candidatos tanto del Partido Demócrata como del Republicano. Como era de esperarse, los candidatos más relevantes, que aparecieron en mayor cantidad de entrevistas y contaron con mayor cobertura publicitaria y presencia en redes sociales durante la campaña, se encuentran presentes en esta lista, por lo que a priori no se detectan sesgos o preferencias por uno u otro partido político en cuanto a recolección de datos.

Sin embargo, también se detectaron valores faltantes, ingresados al Data Frame como NaN (un total de 3) y una entrada "???" haciendo también referencia a la ausencia del autor de dicho discurso.

Además, se identificaron discursos pronunciados en conjunto por más de un candidato, ya fuera estos en el marco de debates, actos de campaña compartidos o eventos que involucraron a representantes de distintos partidos. En estos casos, el discurso se cargó en el Data Frame como una entrada (*input***)** única bajo el nombre de todos los participantes (separados por comas), en lugar de cargar una entrada por cada participante, con el mismo discurso, lugar, fecha y clasificación pero bajo el nombre de cada candidato (*speaker*) separadamente. También aparecen registros con denominaciones genéricas como “Democratic Candidates” o “Multiple Speakers”, que no permiten identificar cuál o cuáles candidatos los pronunciaron y, por lo tanto, no permiten agruparlos como los demás discursos de los candidatos listados. Esta variación en los criterios de presentación de los nombres de los candidatos constituye una inconsistencia dentro del conjunto de datos, ya que refleja una falta de uniformidad en el modo en que se presentan casos similares.

Todo lo mencionado genera cierta incertidumbre respecto a la validez de estos registros.

Tabla 1: Entradas únicas de la columna “speakers” y la cantidad de repeticiones.

Candidatos	Cantidad de discursos
Joe Biden	71
Donald Trump	53
…	…
Democratic Candidates	8
Multiple Speakers	5
Joe Biden, Kamala Harris	4
…	…
NaN	3
…	…
???	1

Antes de continuar con el análisis de las restantes columnas del Data Frame, se realizó una inspección visual de los datos para detectar posibles errores e inconsistencias y se observó que estos solamente se debían a datos faltantes y algunas inconsistencias.

Por ejemplo, se verificó que en los casos en que el discurso o debate fuera interpretado por más de un candidato, la transcripción incluyera los nombres de estos mismos y no otros, lo mismo en el caso de entrevistas de candidatos individuales.

También se verificó que la fecha que figura en el título de los diferentes discursos coincidiera con la fecha listada en la columna *date* del Data Frame; así como que el tipo de discurso coincidiera con la locación. Por ejemplo, que no aparecieran listadas cadenas de televisión para discursos de campaña presenciales o movilizaciones de votantes; solo para entrevistas o debates, que son instancias más factibles de ocurrir en el estudio de una cadena de televisión.

Al verificar lo anterior, se observó, en la columna *location*, una inconsistencia asociada a la falta de homogeneidad en los criterios de registro: en algunos casos se indicaba la ciudad donde se llevó a cabo el discurso, mientras que en otros se mencionaba una cadena de televisión. Es verdad que un discurso llevado a cabo al aire libre o en algún recinto oficial no es lo mismo que uno realizado en el estudio de una compañía de televisión, pero se podría también incluir la localización de este estudio de grabación. Además, se observó que en el caso de instancias que no se realizaron de forma presencial se contaba con la designación “Virtual” como dato de localización. Este último término es algo menos específico que los demás (cadena de televisión de la entrevista o lugar físico del discurso de campaña o

movilización), pero designa situaciones diferentes a éstas a las que se puede dar un dato de localización más específico. Para resolver esta heterogeneidad, se propone la incorporación de una nueva columna que indique la modalidad del discurso (por ejemplo, presencial, virtual o presencial en estudio), lo cual permitiría establecer una clasificación más clara y dar datos adicionales de localización espacial de los candidatos al momento del discurso.

A su vez, esta distinción podría ser útil para identificar fácilmente las relaciones entre el tipo de discurso y el lugar en que fue emitido. Por ejemplo, es probable que los debates y entrevistas se realicen mayoritariamente en estudios de televisión, mientras que los actos de campaña tengan lugar de forma presencial en distintas ciudades.

Una vez analizados estos errores de inconsistencia, se procedió a buscar la cantidad de datos faltantes del conjunto de datos con la función *isna().sum()*. Fuera de la columna *speaker*, no se detectaron entradas erróneas sustituyendo datos faltantes como “???”, por lo que esta función fue suficiente para cuantificar todos los datos no proporcionados. Los resultados de esta búsqueda se presentan en la Tabla 2.

Tabla 2: Cantidad de datos faltantes por columna de datos brindados.

Speakers	Title	Text	Date	Location	Type
3	0	0	0	19	23

Con esto se observa que solamente en las columnas de *location* y *type* hay información que no se completó, por lo que no faltan datos de títulos, transcripciones o fechas de discursos.

Resulta relevante mencionar que, si bien existen datos faltantes en las columnas *location* y *type*, las filas correspondientes no fueron eliminadas, dado que contienen información importante como la transcripción del texto, la fecha y el candidato, que constituyen el foco principal de este trabajo.

2.2. Visualización de cantidad discursos

Tras una primera depuración del conjunto de datos, se seleccionaron los discursos de los 5 candidatos con mayor frecuencia de aparición, con el objetivo de facilitar el tratamiento posterior de la información y otorgar mayor claridad y relevancia a los análisis y visualizaciones subsiguientes.

Los candidatos resultantes fueron: Joe Biden, Donald Trump, Mike Pence, Bernie Sanders y Kamala Harris.

Cabe señalar que, en los casos donde un mismo discurso figuraba bajo la autoría de más de un candidato (con sus nombres separados por comas en la columna *speaker*), dicho discurso fue asignado a ambos. Esto se realizó mediante la combinación de las funciones *str.split(',')* y *explode*. No obstante, aquellos discursos registrados bajo las etiquetas "Multiple Speakers" o

"Democratic Candidates" no pudieron ser asignados a ningún candidato en particular, debido a la falta de información específica.

En la Figura 1 se presenta a modo de *scatter plot* la cantidad de discursos por candidato agrupados mensualmente. Las elecciones presidenciales en Estados Unidos se llevaron a cabo el 3 de noviembre de 2020. Septiembre fue el mes con mayor actividad discursiva, seguido por octubre, coincidiendo con el periodo más intenso de la campaña electoral.

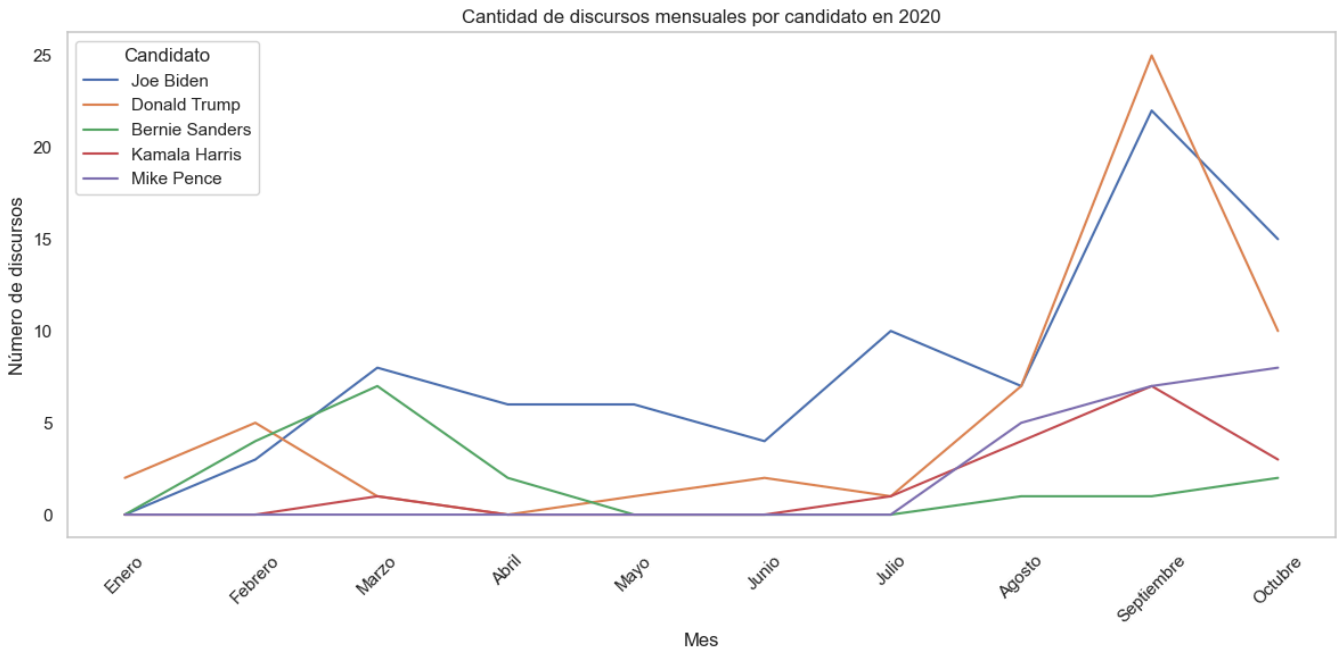

Figura 1: Cantidad de discursos pronunciados por cada candidato a presidente de los Estados Unidos durante el año 2020.

También resulta interesante visualizar individualmente la cantidad y distribución temporal de los discursos de cada candidato, superpuesto al gráfico del total de discursos de los cinco candidatos con más discursos en el período a modo de comparativa.

Esto puede hacerse para todos los (5) candidatos. A modo de ejemplo, en la Figura 2 se presenta a modo de *scatter plot* la comparación para el candidato Joe Biden, los demás gráficos se obtienen de forma análoga.

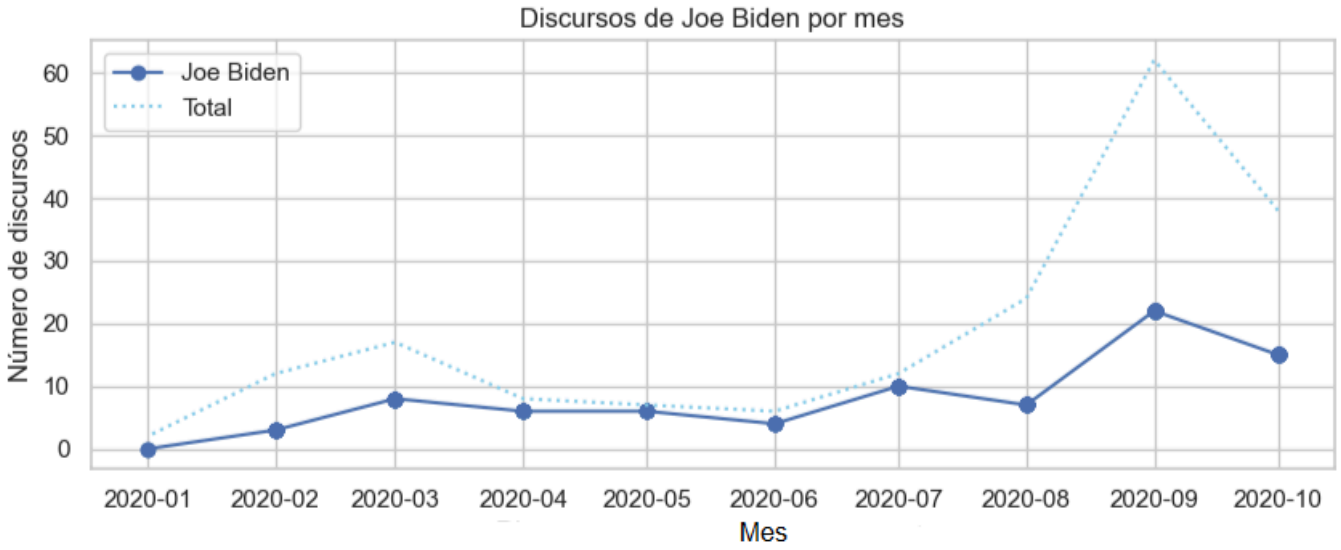

Figura 2: Cantidad de discursos pronunciados por Joe Biden comparado con el total de discursos pronunciados por los 5 candidatos con más discursos durante el año 2020.

2.3. Limpieza del texto

Una de las visualizaciones más relevantes que pueden generarse a partir de este conjunto de datos es aquella que muestra la frecuencia con la que cada candidato menciona determinadas palabras clave, o el conteo de las palabras más utilizadas por cada candidato.

Para llevar adelante este análisis, es necesario normalizar el texto, es decir, convertir todas las palabras a minúsculas y eliminar signos de puntuación como puntos y comas. Con este fin se utilizó la función personalizada *clean_text* para eliminar las primeras palabras hasta el primer \n (normalmente correspondientes a identificadores de la transcripción del discursos, y no a contenido del propio discurso en sí), convertir todo el texto a minúsculas y completar signos de puntuación faltantes. El texto "limpio" de cada discurso se almacenó en una nueva columna *CleanText* en el Data Frame original.

Además de esas modificaciones, se eliminaron algunos puntos que se detectaron por inspección visual para evitar que palabras como "you" y "you." fueran contabilizadas separadamente, utilizando la función personalizada *limpiar_puntos*. También, se hizo manualmente una lista de las contracciones más comunes y utilizadas del idioma inglés con el objetivo de implementar una función que las expandiera (por ejemplo, sustituir "didn't" por "did not"). Si bien "didn't" y "did not" son fonéticamente diferentes, a la hora de contabilizar palabras utilizadas representan el mismo significado ("did" y "not"), por lo que se consideró contabilizarlas junto con sus dos palabras constitutivas y no como una palabra propia. Para esto se utilizó la función personalizada *expand_contractions*.

Finalmente, se utilizó la función *str.split* para separar el texto limpio y homogeneizado en palabras individuales y listarlas. Esto se incluyó en una nueva columna, *WordList,* del Data Frame.

Cabe resaltar que existe otro problema, no necesariamente de calidad de datos, que se encuentra presente en la transcripción del texto de los discursos y que podría afectar al resultado de conteo de palabras. En los casos de entrevistas o debates o incluso algún discurso en que interviene más de una persona, no solo se transcribe lo pronunciado por la persona a quien se atribuye dicho discurso sino también lo pronunciado por el entrevistador, periodista, adversario o participante que le contesta.

Entonces se encuentran textos del estilo: "Donald Trump: …, Speaker 1: …, "Donald Trump": …, "Speaker 1: …". Y esto presenta dos problemas, por un lado el hecho que Donald, Trump, Speaker y 1 son contabilizadas como palabras pronunciadas cuando en realidad no lo fueron y solamente están presentes a modo de organización de la transcripción, y segundo el hecho que todo lo pronunciado por el Speaker 1 será erróneamente atribuido a Donald Trump cuando en realidad fue dicho por otra persona.

Sin embargo, para poder corregir este error se debería conocer todas y cada una de las personas (aparte de los candidatos presidenciales) que participaron de todos y cada uno de los discursos, entrevistas o debates (o incluso tener en cuenta denominaciones genéricas como *Speaker, Interviewer, Person,* etc.) para poder eliminar intervenciones no pertenecientes a los candidatos de todos los debates. Ya que si solo se consideraran algunos (por ejemplo, solo Speaker 1), se corregirían algunos discursos y otros no, generando un sesgo. La forma de considerar a todos los participantes que no son candidatos presidenciales es examinar la totalidad de los discursos uno por uno y, dada su extensión, aún realizándose manualmente existe el riesgo de cometer errores u omitir participantes.

Por lo tanto, se opta por omitir esta corrección y realizar el conteo de palabras con el texto tal cual se obtuvo luego de la anterior limpieza, asumiendo el error en el conteo de palabras como "Donald", "Trump", "Joe" o "Biden" que, además de ser verdaderamente dichas por los propios candidatos, pueden aparecer en la transcripción del texto como elementos de orden o dichos por otro participante (entrevistador, periodista o persona externa).

2.4. Conteo de palabras y visualización

A partir del texto procesado en el ítem anterior, y la columna se construyeron distintas visualizaciones.

En primer lugar, se implementó la función personalizada *total_word_speaker* para visualizar la cantidad de palabras totales pronunciadas por cada candidato. Los resultados se presentan ordenados de mayor a menor en la Tabla 3. En esta se puede observar que las de Donald Trump (candidato que pronunció más palabras) difieren en un orden de magnitud con las de Bernie Sanders (candidato que pronunció menos palabras). Este resultado coincide con la

percepción que se tiene del candidato Donald Trump, persona mediática que genera polémica y que, por lo tanto, se pronuncia con mayor frecuencia.

Tabla 3: Cantidad de palabras pronunciadas por candidato

Candidato	Total de Palabras
Donald Trump	579305
Joe Biden	468916
Mike Pence	121546
Kamala Harris	89725
Bernie Sanders	68693

A continuación, se contaron las palabras más frecuentemente mencionadas por cada candidato, utilizando la función personalizada *rank_words_speaker*. En la Figura 3 se presentan las 10 palabras más mencionadas en general (ordenadas de mayor a menor) y la cantidad de veces que fue mencionada por cada candidato, a modo de gráfico de barras.

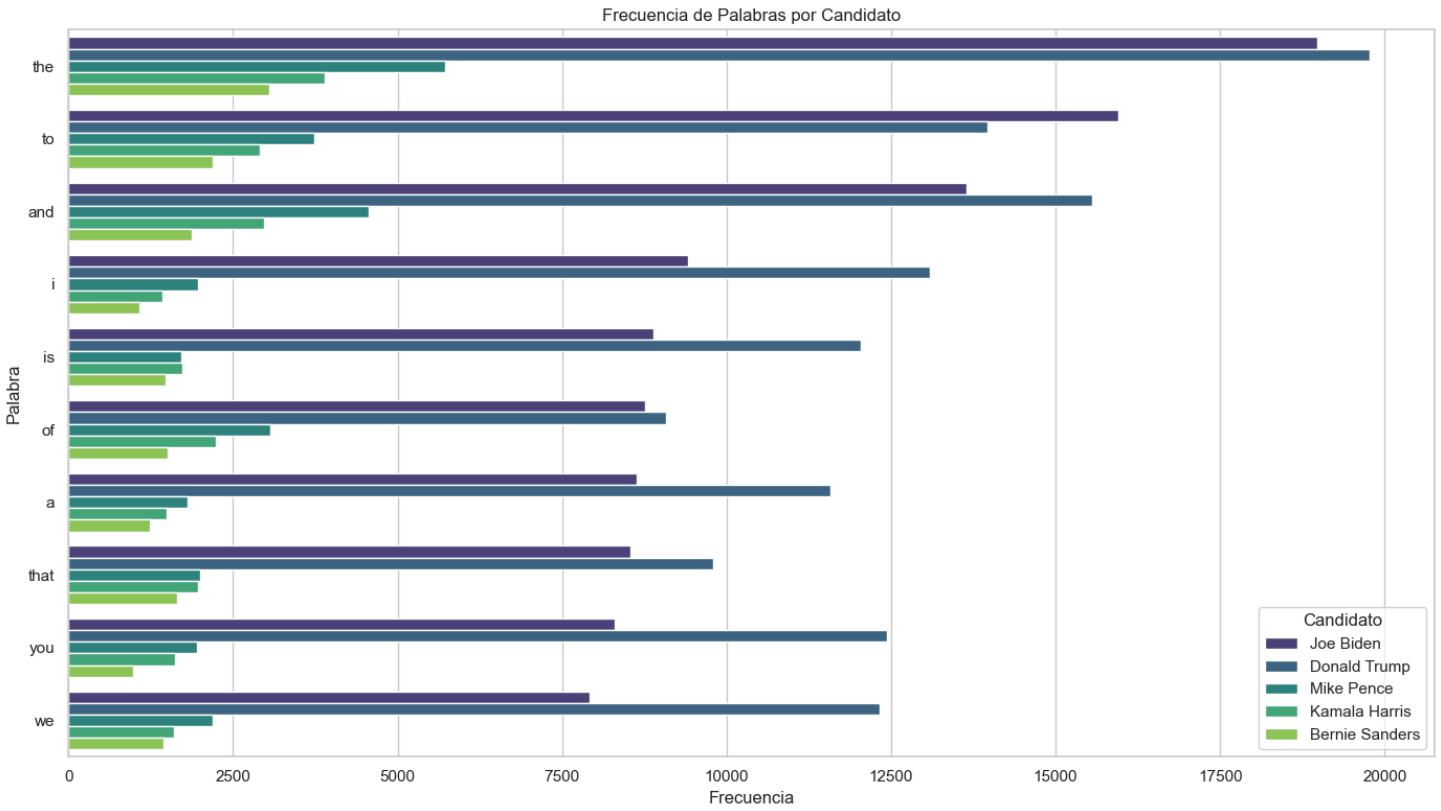

Figura 3: Palabras más frecuentes en el conjunto total de discursos, diferenciadas por candidato.

Cabe resaltar que en esta visualización se cuenta las 10 palabras más mencionadas entre el total de los candidatos para luego contar la cantidad de menciones de cada uno. Pero no necesariamente las 10 palabras más mencionadas globalmente se corresponden con las 10 palabras más mencionadas por cada candidato particular.

La función personalizada *rank_words_speaker* permitió contar las palabras más mencionadas en todos los discursos y ordenarlas según la frecuencia con que fueron utilizadas por cada candidato. A modo de ejemplo se presentan en la Tabla 4 las de Kamala Harris. Un procedimiento análogo se realizó para los demás candidatos, pero se incluye en el informe únicamente este ejemplo con el objetivo de resaltar el hecho de que las palabras más pronunciadas globalmente no tienen porqué coincidir con las más pronunciadas por cada candidato. Esta información se puede observar en la Figura 4, donde se cuentan las palabras más mencionadas por el candidato, en particular. Como se mencionó anteriormente, esta gráfica difiere de la anterior, ya que muestra únicamente las diez palabras más mencionadas por cada candidato. En el caso de la palabra “it”, solo aparece el dato correspondiente a Donald Trump, ya que esta palabra figura entre sus diez más utilizadas. Esto no implica que los demás candidatos no la hayan mencionado, sino que no se encuentra entre sus términos más frecuentes.

Tabla 4: Palabras más utilizadas por Kamala Harris en sus discursos.

Palabra	**Cantidad de menciones**
the	3899
and	2978
to	2911
of	2238
that	1967
is	1729
you	1618
we	1606
a	1488
i	1429

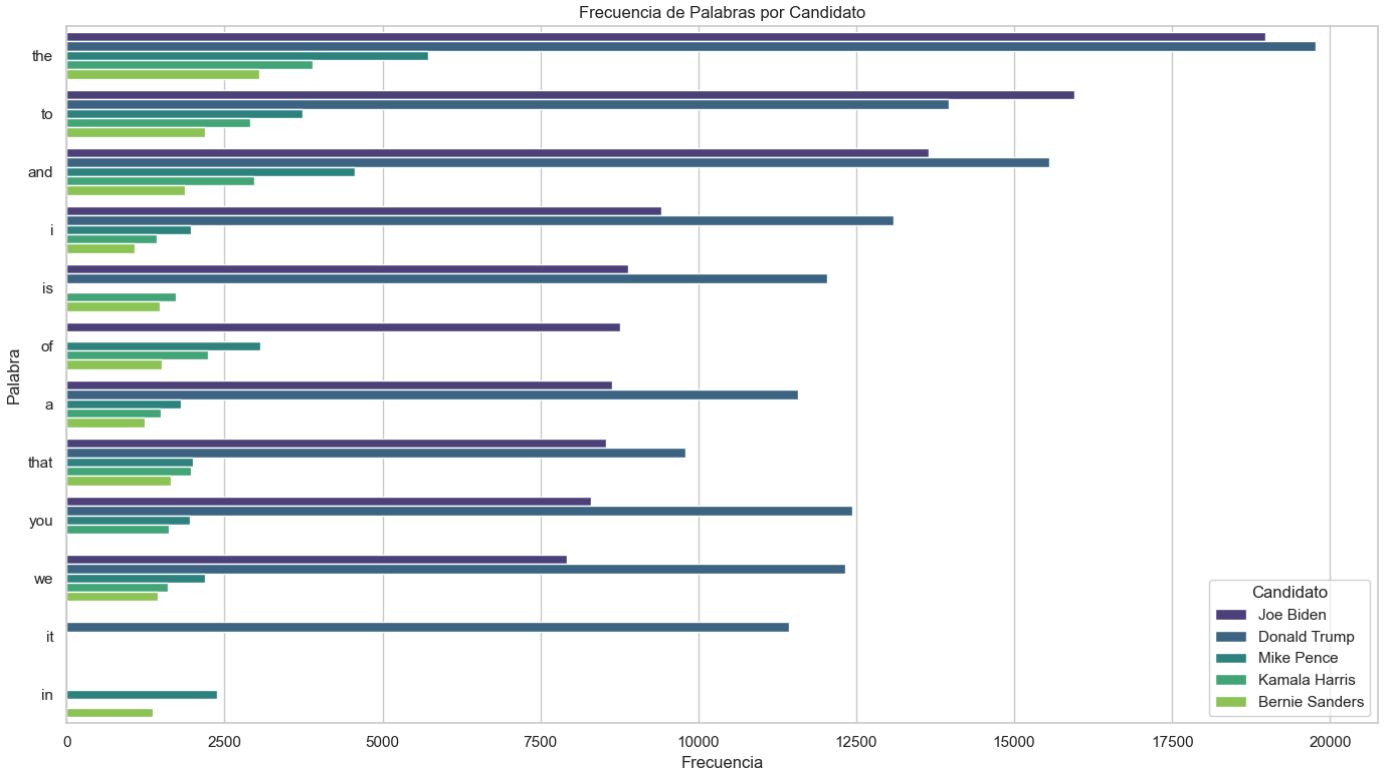

Figura 4: Palabras más frecuentes en los discursos de cada candidato.

Es interesante mencionar que uno de los problemas de visualización que surge al contar palabras de esta manera es que al contar las palabras de todos los discursos en general, se está comparando magnitudes que no son necesariamente equivalentes. Esto se debe a que algunos candidatos mencionaron más una palabra en particular simplemente porque dieron más discursos y pronunciaron más palabras en total, y no necesariamente porque la utilicen con mayor frecuencia que otros candidatos que la pronunciaron menos.

Para corregir este problema, se pueden normalizar los datos. Dado que esta gráfica simplemente incluye conectores funcionales, verbos o pronombres sin valor semántico, no se realizó esta corrección, pero sí es una consideración que se tuvo en cuenta en gráficos que se presentarán más adelante en este informe que cuentan palabras de mayor interés.

También se incluye en la Figura 5 una visualización en forma de gráfica de barras de las 10 palabras más utilizadas en este caso por Mike Pence. Nuevamente se observa que no necesariamente coinciden con las 10 más mencionadas globalmente por Kamala Harris.

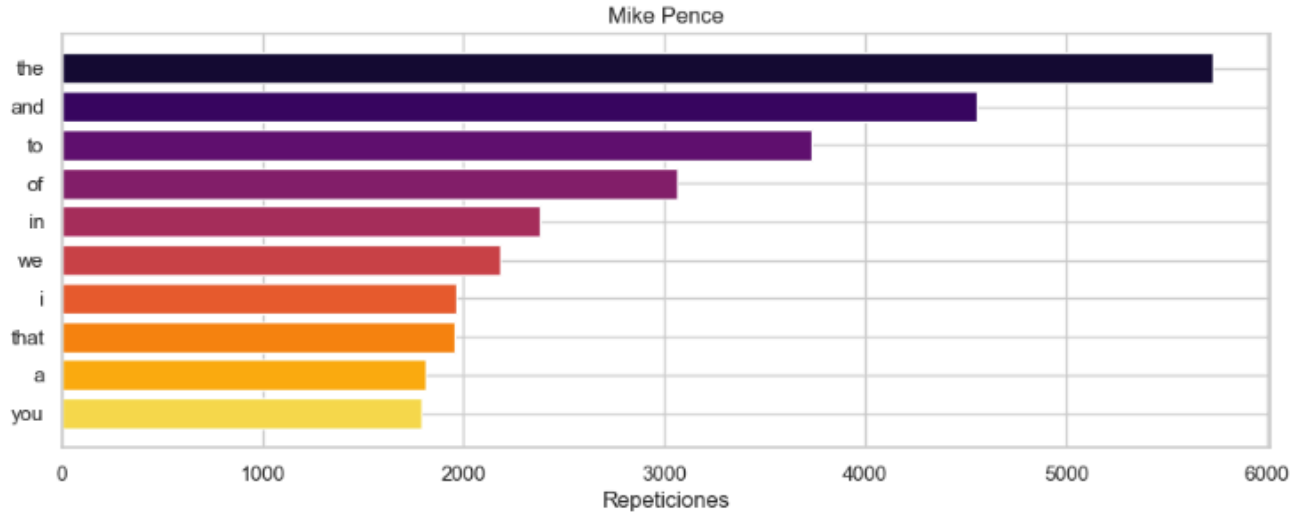

Figura 5: Palabras más utilizadas por Mike Pence

No obstante, como se mencionaba respecto a la Figura 3 y 4 (y también es aplicable para la Figura 5), este enfoque inicial de conteo de palabras sin ningún criterio resultó limitado, dado que las palabras más frecuentes son generalmente términos funcionales como “a”, “the” o “to”, comunes a todos los discursos y sin contenido semántico relevante.

El análisis de frecuencia puede resultar de mayor utilidad o interés al enfocarse en términos sustantivos vinculados a asuntos de interés público para los Estados Unidos, como la salud, el desempleo o la inmigración. Para evaluar esto, se estudió la cantidad de veces que cada candidato mencionó determinados términos clave asociados a estos temas. Para esto se definió una lista de palabras clave seleccionadas manualmente de acuerdo a consideraciones personales de lo que debería formar parte de la discusión política en el marco de una campaña electoral: "democracy", "covid", "economy", "education", "security", "unemployment", "rights", "healthcare", "racial", "immigration", "wall", "climate change", "poverty", "war". Se implementó la función personalizada *count_ keywords_per_speaker* para realizar el conteo de las mismas entre los discursos.

Finalmente, en este caso sí se normalizaron los datos para comparar frecuencia de menciones por cada candidato y no solo cantidad de menciones totales, para evaluar a qué asuntos de interés público cada candidato da más importancia y dedica más tiempo en sus discursos. En la Figura 6 se grafica la cantidad de repeticiones de estas palabras clave seleccionadas por cada 1000 palabras pronunciadas por cada candidato.

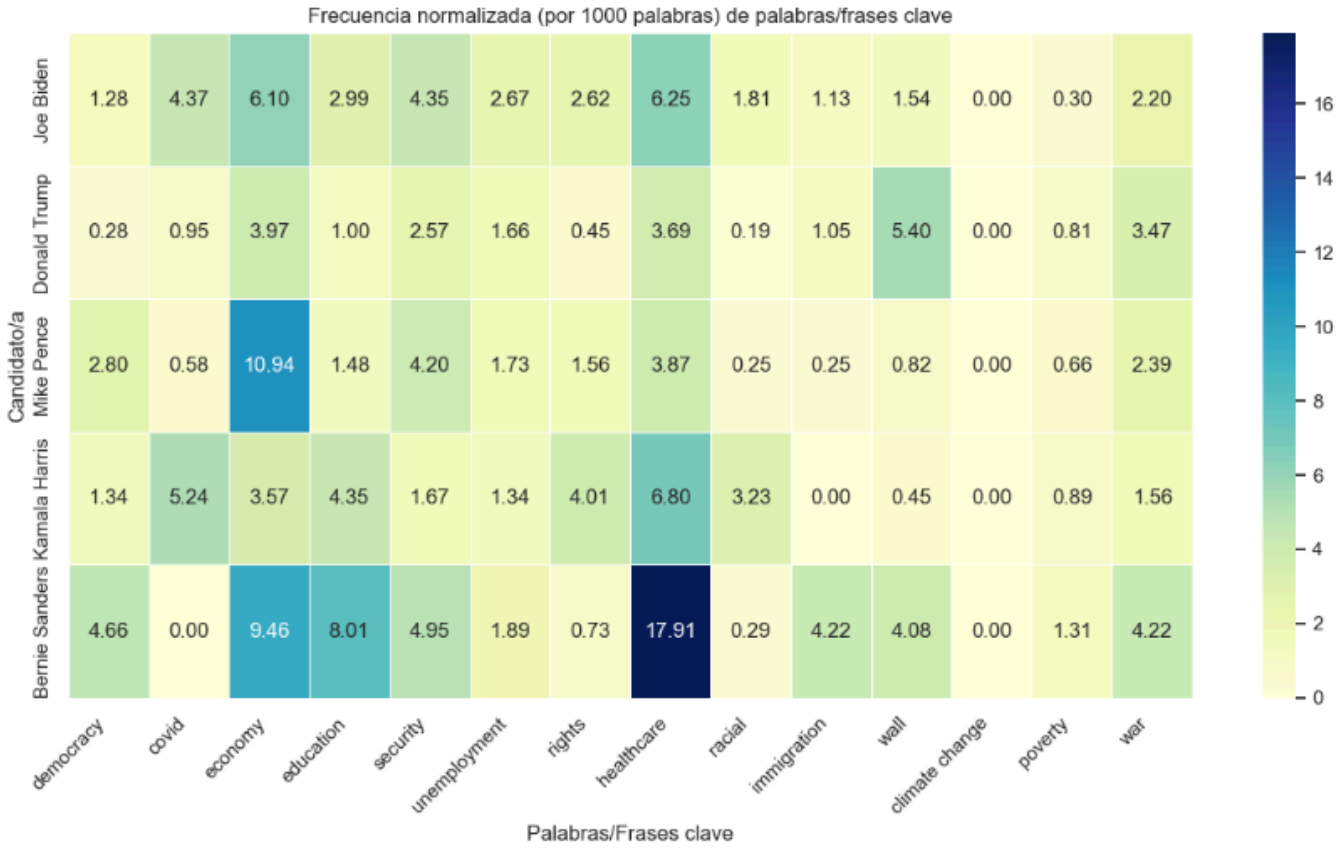

Figura 6: Conteo de palabras clave mencionadas por candidato

Es interesante observar la relevancia que el candidato Bernie Sanders otorga al asunto de la salud pública y el protagonismo que le da en sus discursos, lo cual es coherente con su orientación Demócrata. Sanders es el presidente que presenta un protagonista más marcado (dentro de los términos considerados, por supuesto), seguido por Mike Pence y su enfoque en la economía, también coherente dada su orientación Republicana y el propio interés popular en el tema (es de esperar que enfaticen en aquellos asuntos de mayor preocupación popular).

Si bien este análisis ya brinda información más valiosa que el simple conteo de palabras más reptiles sin ningún criterio realizado anteriormente, una estrategia sin duda más robusta podría consistir en ampliar el conjunto de términos representativos de cada categoría temática. Por ejemplo, para analizar el interés por el tema de la economía del país se puede considerar los siguientes términos afines: “economic”, “growth”, “inflation”, “gdp”, “jobs”, “employment”, “unemployment”, “wages”, “recession” y “recovery”.

Esto permite cuantificar aquellas menciones al mismo asunto hechas con términos no exactamente iguales a los considerados originalmente como palabras clave. El resultado se presenta en la Figura 7 y se puede observar algunos cambios en las abundancias relativas respecto a los resultados expuestos en la Figura 6 anterior.

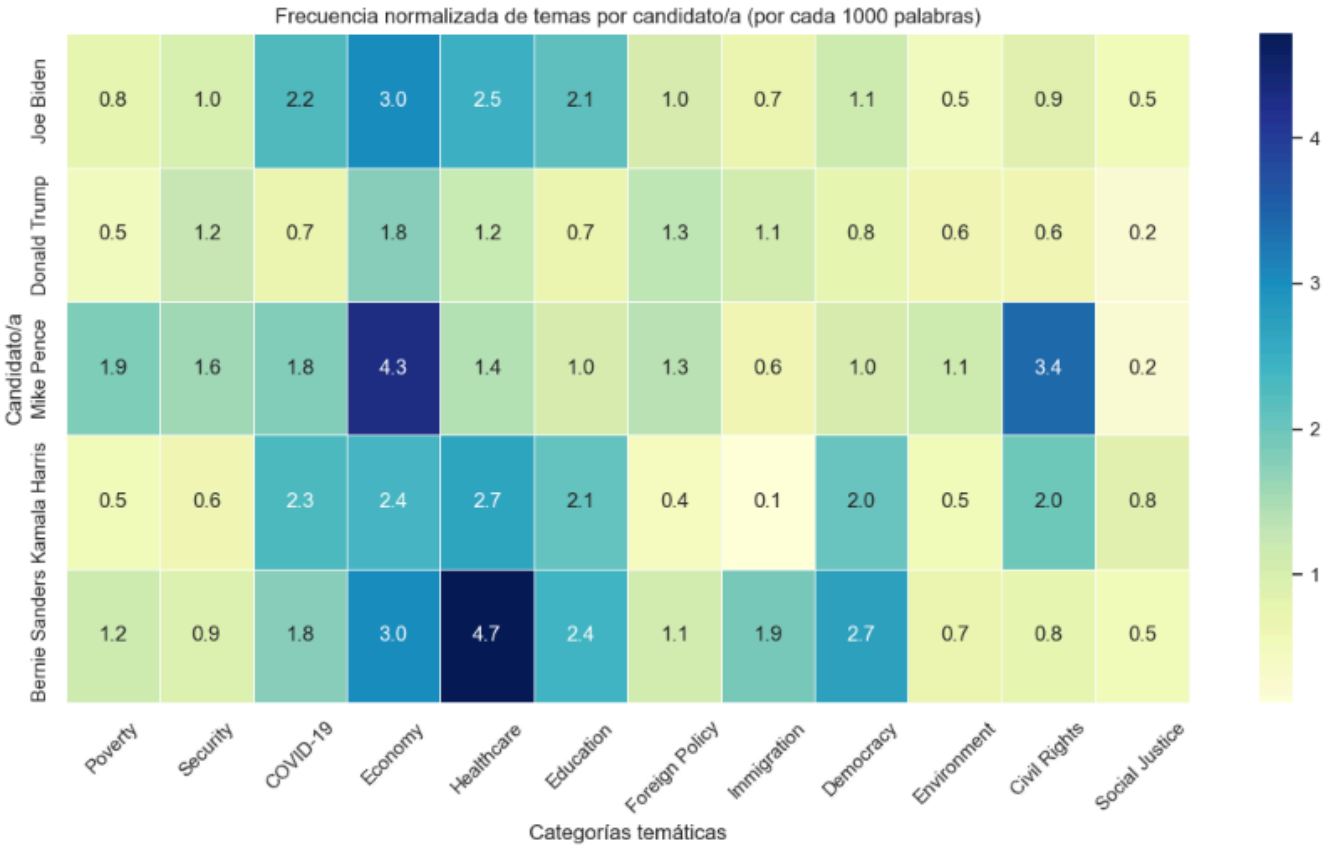

Figura 7: Conteo de categorías clave abordadas por candidato

Si bien los temas más frecuentes parecen ser los mismos para todos los candidatos con este nuevo análisis, las abundancias se vieron ligeramente modificadas debido a la consideración de diferentes acepciones para el mismo tema.

Como este se pueden realizar otros análisis semejantes, agrupando la información disponible de acuerdo con otros criterios más allá de los temas de interés público. Un análisis interesante sería, por ejemplo, analizar la distribución en el tiempo de las menciones de ciertos términos, tal vez concentradas cerca de la fecha de ocurrencia de algún hecho.

El 3 de marzo de 2020, la Organización Mundial de la Salud (OMS) declaró oficialmente al COVID-19 como una pandemia global. Con el objetivo de analizar la importancia que cada candidato le otorgó a este tema a lo largo del año 2020, se identificaron en sus discursos los siguientes términos relacionados con la pandemia: “covid”, “coronavirus”, “pandemic”, “quarantine”, “vaccine”, “virus”, “mask” y “contagion”.

Para hacer comparables los resultados entre candidatos y a lo largo del tiempo, la frecuencia de mención fue normalizada según el número de discursos pronunciados por cada uno. Los resultados se presentan en la Figura 8, donde se observa que la mayoría de los candidatos comenzaron a referirse al tema a partir de marzo. En particular, Joe Biden mantuvo un número constante y elevado de menciones por discurso, mientras que Donald Trump —a

pesar de ser presidente en ese momento— mostró picos de menciones aislados y, en general, una frecuencia inferior a la del candidato anteriormente mencionado.

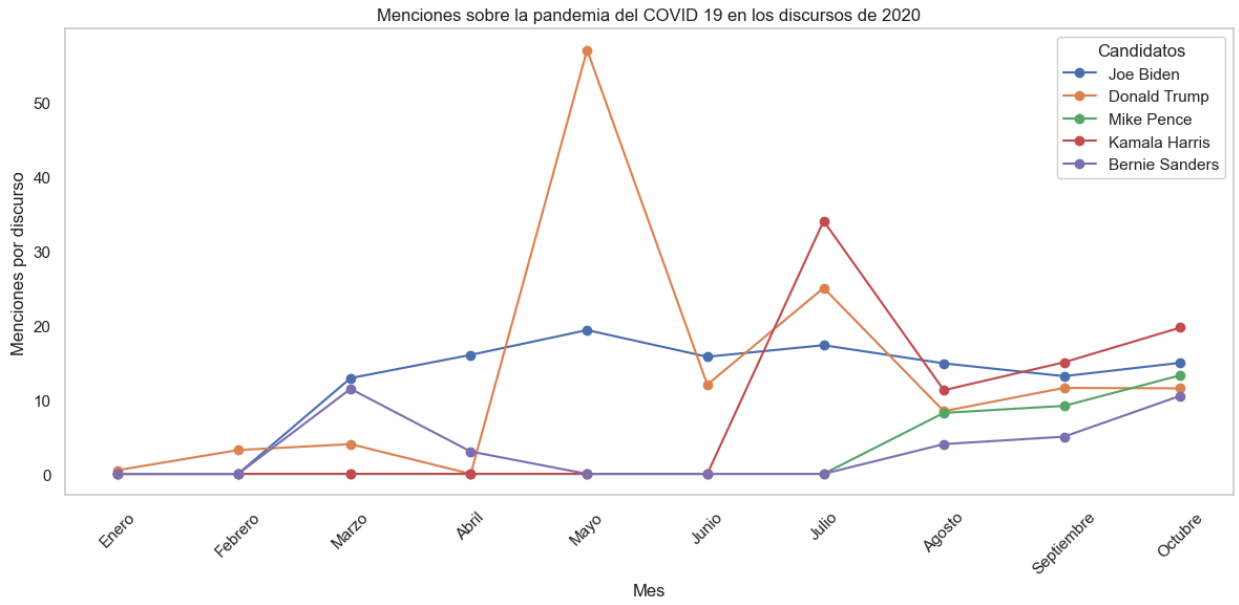

Figura 8: Menciones sobre la pandemia del COVID 19 en los discursos de los candidatos.

Otro hecho que marcó la política y la situación social de Estados Unidos fue el asesinato de George Floyd, ocurrido el 25 de mayo de 2020, cuando el oficial de policía Derek Chauvin presionó su rodilla contra su cuello durante más de nueve minutos durante un arresto. De forma análoga al análisis anterior, se realizó una búsqueda de términos relacionados con este acontecimiento en los discursos y se normalizó la frecuencia de mención. Los resultados se presentan en la Figura 9.

Como era esperable, las menciones comienzan en mayo, mes en que ocurrió el hecho, y alcanzan su punto máximo en junio, destacándose Donald Trump —entonces presidente— como el candidato con más referencias al tema. Llama la atención que Bernie Sanders no haya realizado ninguna mención, lo cual podría atribuirse a un sesgo en la selección de los términos utilizados para el análisis

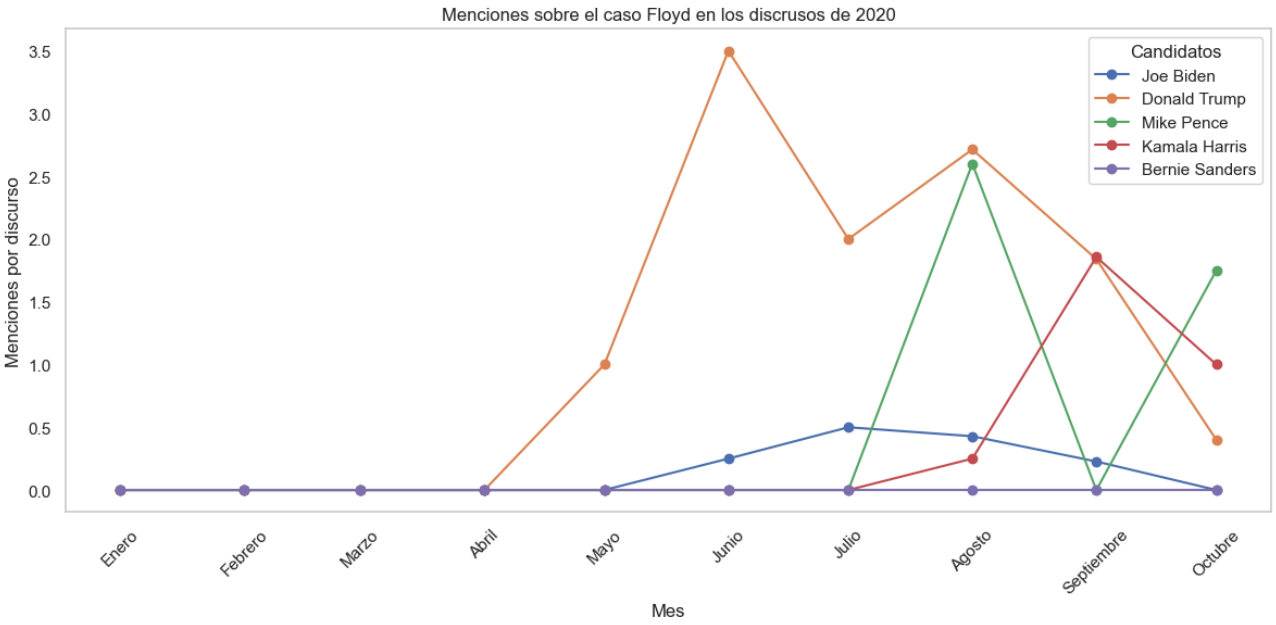

Figura 9: Menciones sobre el caso Floyd en los discursos de los candidatos.

Por otra parte, el análisis puede extenderse a nivel partidario. En lugar de evaluar individualmente a cada candidato, es posible agrupar los candidatos de acuerdo con su partido político y contabilizar las menciones de ciertos términos por bloque partidario. Esto permitiría identificar diferencias ideológicas o programáticas entre los principales partidos políticos estadounidenses de forma más ordenada.

2.5. Menciones cruzadas

Otro tema de interés a analizar puede ser la cantidad de menciones de cada candidato a los demás (dentro de los 5 candidatos que más discursos tienen).

Para eso se contabilizó la cantidad de veces que cada candidato mencionó el nombre completo, solo el nombre o solo el apellido (evitando contabilizar tres veces menciones a Nombre Apellido como “Nombre Apellido”, “Nombre” y “Apellido) de otro candidato. El resultado se presenta en la Figura 10 en formato de *heat map*.

Para este análisis se consideraron nulas las auto-menciones. Si bien es posible que, por ejemplo, “Donald Trump” haga referencia a la “Administración Trump”, en cuyo caso se mencionaría una vez a sí mismo, no se considera relevante para este análisis. Además, al omitir las menciones propias dentro del discurso, se evita el sesgo generado por palabras utilizadas en el ordenamiento del texto, como en los patrones: 'Donald Trump: ...', 'Speaker 1: ...', o similares, donde podrían contabilizarse erróneamente menciones del candidato a sí mismo.

Por otro lado, no se eliminaron del análisis las palabras correspondientes a intervenciones de personas externas (entrevistadores, periodistas u otros), lo que introduce un posible error asociado a la mención de un candidato por alguien que no es el orador principal. Esto se hizo para evitar el sesgo que se podría introducir al no considerar absolutamente todos los hablantes externos posibles.

Igualmente, es importante tener en cuenta que esta consideración puede generar ciertas desviaciones en los resultados presentados en la Figura 10.

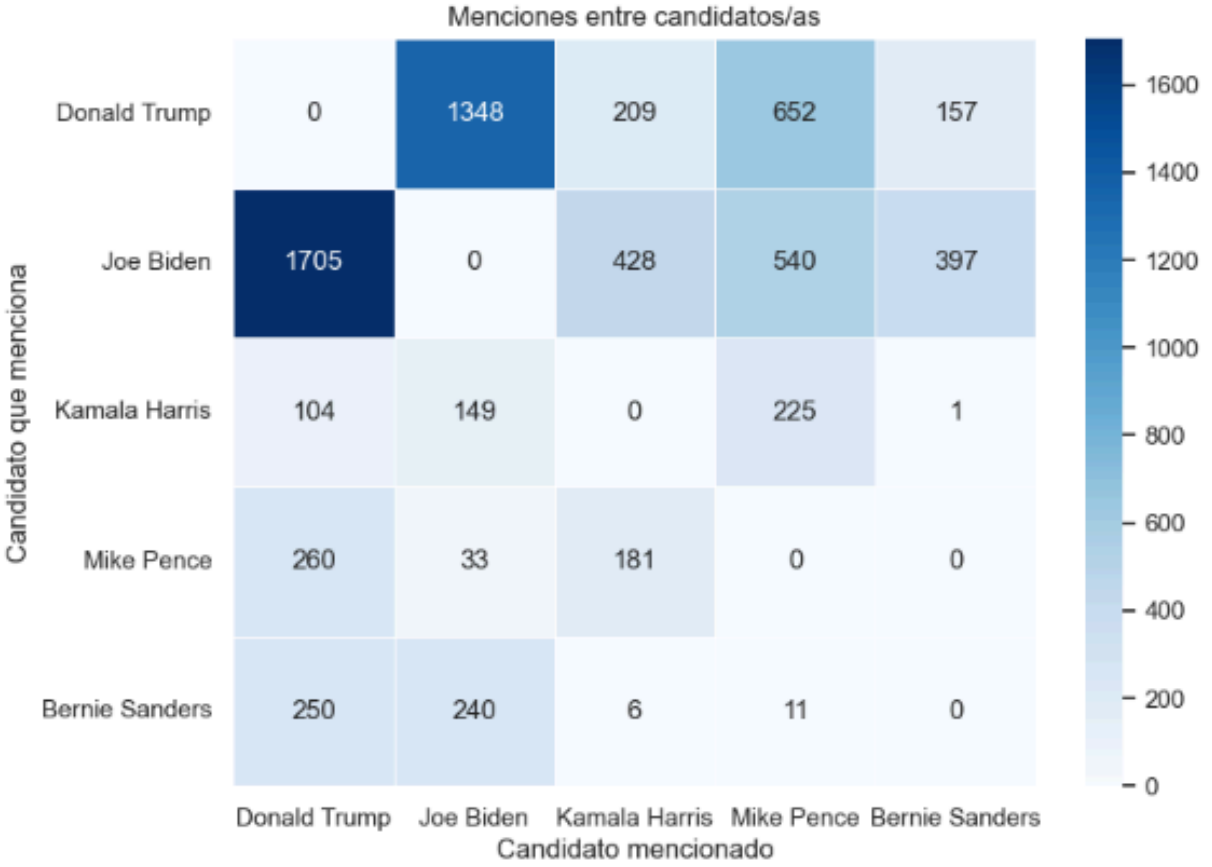

Figura 10: Menciones de unos candidatos a otros (ignorando auto-menciones) en discursos.

2.6. Preguntas a responder

En línea con el análisis anterior, algunas preguntas que surgen y pueden ser respondidas mediante el análisis del set de datos anteriormente presentado (y que van en la misma línea que algunas estrategias ya implementadas) son:

1. ¿Qué temas de interés público dominan los discursos de cada candidato?
   Esto fue lo que se pretendió evaluar con los *heat maps* presentados anteriormente.

2. ¿Cómo varía la frecuencia de ciertos temas a lo largo del tiempo?
   Por ejemplo, ¿cuándo comienza a aumentar el uso de términos asociados al COVID-19 o al asesinato de George Floyd?
3. ¿Qué candidato/a menciona más la palabra “justicia” o “economía”?
   Podría llevarse a cabo un análisis que hiciera énfasis en ciertos valores simbólicos importantes ya sea para el entrevistador o para un público específico al que se quiere apuntar una campaña electoral.
4. ¿Qué tan diverso es el vocabulario utilizado por cada candidato?
   Mediante métricas como riqueza léxica (número de palabras únicas / número total de palabras) o entropía (mide cuán uniformemente se usan las palabras en un texto).
   Útil si fuera de interés un análisis del estilo de oratoria de cada candidato. Por ejemplo, para evaluar por qué hay más personas que concuerdan o se oponen a las ideologías planteadas por algún candidato; o si alguno suena más convincente.
5. ¿Hay correlación entre ciertos temas?
   Por ejemplo, ¿los discursos que hablan de inmigración también tienden a hablar de seguridad o de desempleo?
6. ¿Cómo varía la longitud de los discursos según el orador? ¿Hay alguno que repita la misma idea una y otra vez dentro del discurso? ¿Quién es el más sintético?
7. ¿Se observa un cambio de tono o enfoque en discursos de campaña en comparación con entrevistas o debates?
8. ¿Hay algún entrevistador o cadena de televisión que presente un sesgo o mayor enfoque e interés hacia ciertos temas? ¿Todos hablan de los mismos?
9. ¿Qué temas se intensifican en los discursos más cercanos a las elecciones ?
   Por ejemplo en el mes de septiembre que es el de mayor cantidad de discursos o en el mes de octubre que es el que es inmediatamente anterior a las elecciones.
10. ¿Hay diferencias de estilo entre hombres y mujeres candidatos?
    Puede referir al uso de ciertos términos o conectores con mayor frecuencia, o también a menciones a ciertos temas de interés público (como cuidado, familia, comunidad).
11. ¿Qué candidato es más mencionado por otros candidatos?
12. ¿A qué cierto candidato menciona más otro candidato en particular?
13. Se puede evaluar la estrategia discursiva de Joe Biden, presidente electo del 2020, según su frecuencia, su oratoria, su léxico, el lugar donde los realiza y los temas en los que centra sus declaraciones.

3. Conclusión

El trabajo realizado para el presente informe permitió llevar a cabo un abordaje preliminar a la limpieza y transformación de un conjunto de datos compuesto por discursos políticos de presidentes de los Estados Unidos en el marco de las elecciones presidenciales del año 2020. Esto se realizó con el objetivo de habilitar su análisis cuantitativo, permitiendo comparar la intervención o relevancia de cada participante, así como los temas o palabras más empleadas por cada uno.

A partir de la detección de errores de inconsistencia y datos faltantes así como la estandarización del formato del campo que contenía las transcripciones de los discursos, fue posible construir representaciones estructuradas que facilitaron la exploración temática del contenido.

Se realizó un conteo general de palabras sin ningún criterio a priori y, en base a las conclusiones obtenidas, se definieron categorías semánticas relevantes (como salud, economía, educación, seguridad, entre otras) y se agruparon palabras clave bajo cada una de ellas para obtener información de mayor significado.

Esto permitió calcular y visualizar la frecuencia relativa con la que cada candidato abordó distintos temas. Las visualizaciones resultantes (por ejemplo, *heat maps*) ofrecieron una mirada clara sobre las prioridades discursivas de cada orador y permitieron detectar patrones interesantes.

Además del análisis temático, se exploró la dinámica de las menciones cruzadas entre candidatos. Este enfoque reveló qué figuras políticas fueron más mencionadas por sus pares. Incluso podría extenderse el análisis a la evaluación de en qué contextos se mencionó a otro candidato, permitiendo observar interacciones discursivas y estrategias de confrontación o alusión y, por lo tanto, aportando una dimensión relacional al análisis del lenguaje utilizado en campaña.

En síntesis, el trabajo mostró cómo una adecuada preparación del conjunto de datos, combinada con herramientas básicas de análisis de texto, permite extraer información significativa y relevante desde una perspectiva tanto descriptiva como comparativa.

4. Bibliografía

Diario Las Américas. (2020, diciembre 30). *Sucesos que marcaron el 2020 en EEUU*. https://www.diariolasamericas.com/eeuu/sucesos-que-marcaron-el-2020-eeuu-n4213289

Diapositivas del curso *Introducción a la Ciencia de Datos 2025*. Facultad de Ingeniería. UdelaR.