Entrega Final

Julián Corbo S.

2024-11-14

Table of contents

Consigna **2**

Archivos Entregados **2**

Introducción **3**

Recuperación **3**

Pre-procesamiento **3**

Análisis **4**

Análisis de sentimiento 4

Frecuencia de Ocurrencia 6

Milei 9

Términos Claves 9

Co-ocurrencia 10

Key Word in Context 11

Consigna

Recuperación: *Elegir una fuente de datos de las vistas en el curso (archivos de texto brutos, OCR, web scraping, scrapeo parlamentario, prensa digital, búsquedas de google, transcripción de audio, subtítulos de YouTube, APIs de Google, entre otras) e incluir el proceso de extracción o reconstrucción de la información.*

Pre-procesamiento: *Crear un corpus de datos textuales tabulados, realizar el pre-procesamiento utilizando alguna/s de las herramientas vistas en el curso (manipulación de strings, limpieza o pre/codificación manual)*

Análisis: *Incluir al menos tres técnicas de minería de texto vistas en el curso (frecuencia de ocurrencia, asociación de palabras, contexto de aparición de palabras o frases, diccionarios, análisis de sentimiento, modelado de temas, entre otros posibles)*

Visualización: *Ilustrar el documento con al menos dos visualizaciones que surjan del procesamiento del texto precedente y den cuenta de los resultados del mismo.*

Archivos Entregados

En esta entrega utilicé múltiples recursos que requieren tiempo de procesamiento, por lo que trabajé en etapas agregando cambios sobre distintas bases. Adjunto todas estas con motivo de documentar el proceso pero solo son necesarias para el funcionamiento del script aquellas que están marcadas en negrita.

1. **base_final.Rdata**: Base resultante luego de todos los cambios generados en el preprocesamiento.
2. **casarosada_anotada.Rdata**: Base generada por Udpipe.
3. casarosada_final.Rdata: Base resultante antes de procesarlo con pysentimiento
4. casarosada_final_sentiment.csv: Base resultante luego de procesarlo con pysentimiento
5. casarosada_raw: Base inicial del scraping web
6. sentimiento.ipynb: script en google colab con pysentimiento.

Introducción

En el presente trabajo me propongo realizar un análisis exploratorio de los discursos presidenciales argentinos desde la asunción de Mauricio Macri en 2015 hasta la actualidad. Dado el contexto técnico del curso no pretendo realizar un análisis interpretativo de corte teórico y me limitaré a una exploración de diferentes formas en que se puede explorar la temática y algunos comentarios descriptivos de estos.

El informe se estructurará siguiendo los requerimientos de la consigna, donde se explicará el proceso y las descripciones que se tomaron en el código. A su vez, la sección de análisis será estructurada en base a diferentes preguntas que irán surgiendo.

En términos generales se pretende responder cómo ha ido evolucionando la retórica presidencial argentina en los últimos tres periodos. Dado que la naturaleza discursiva de Javier Milei resulta particularmente llamativa se ahondara en la misma.

Recuperación

Por simplicidad del análisis opté por reducir el universo discursivo de los presidentes a los discursos publicados en la página oficial de la Casa Rosada. Debe tenerse en cuenta que esto implica un gran recorte de los elementos discursivos que hacen a la política contemporánea, pero permite observarlo en su forma más institucionalizada.

La técnica utilizada para la obtención de los datos textuales fue el **scraping** o raspado web. Para esto creé una función que extrae la transcripción (utilizando el paquete *rvest*) del respectivo discurso de un link asignado como argumento. La dificultad consiste en repetir este proceso para los 1390 discursos publicados.

Afortunadamente el comportamiento del url que alberga los discursos era bastante simple: al enlace base de la página (https://www.casarosada.gob.ar/informacion/discursos?start=) se le agregaba una secuencia de múltiplos de 40. Pero implicaba el desafío adicional de extraer cada uno de los hipervínculos que llevaban a los discursos propiamente dichos en cada una de esas páginas.

Lo solucioné realizando una iteración relativamente simple que permita obtener cada uno de estos hipervínculos y extraer el texto con la función que había creado previamente. Además, extraje en este proceso la fecha, título y el presidente y lo guarde en una base de datos llamada *casarosada.*

```
# A tibble: 6 x 4
  titulo                                                presidente fecha discurso
  <chr>                                                 <chr>      <chr> <chr>
1 "Palabras Del Presidente De La Nacion, Javier Milei~ Milei      Miér~ "Buenos~
2 "Palabras Del Presidente De La Nacion, Javier Milei~ Milei      Sába~ "Hola a~
3 "Palabras Del Presidente De La Nacion Javier Milei ~ Milei      Vier~ "Buenas~
4 "Palabras Del Presidente En Jornadas Monetarias Y B~ <NA>       Mart~ "Javier~
5 "Palabras Del Presidente De La Nacion, Javier Milei~ Milei      Sába~ "¡Hola ~
6 "Palabras Del Presidente De La Nacion, Javier Milei~ Milei      Miér~ "Buenos~
```

Pre-procesamiento

Una vez obtenida la base fue necesario realizar un pre procesamiento para obtener un corpus de texto con el que se pueda trabajar.

Primero me interesaba realizar algunos cambios generales a *casarosada* de forma que permitiese un mejor uso de la información disponible: asigné un ID de discurso, cambié el formato de fecha a datetime y completé las celdas *"missing"* de la columna *presidente* con el valor inmediatamente anterior.

Luego, me concentré en las transformaciones necesarias de los datos textuales. Transformé todos los caracteres en minúscula, eliminé caracteres especiales del español, eliminé toda la puntuación y separé el corpus por párrafo (agregándole un ID de párrafo correspondiente).

Cabe destacar que en algunos discursos aparecen otros interlocutores que deberían ser eliminados para el correcto procesamiento del corpus. Sin embargo, dada la dificultad que me suponía y los objetivos de la entrega decidí dejar intactas y asumir despreciables estas intervenciones.

Análisis

Análisis de sentimiento

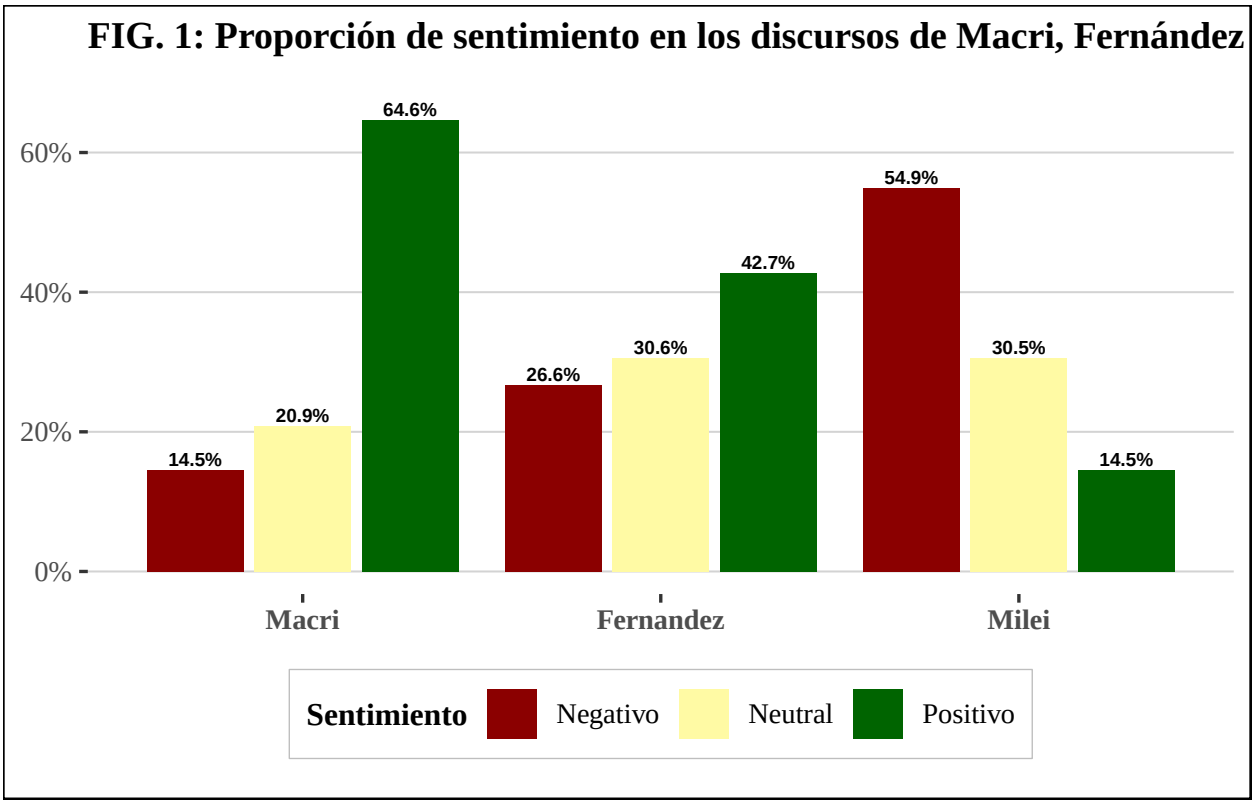
Mi primer interrogante consistía en el tono del discurso. Milei, desde sus comienzos como figura pública, ha sido caracterizado por un tono particularmente disruptivo y negativo. Me interesaba saber si esto se traducía a su comunicación institucional y cómo se comparaba con los predecesores al cargo presidencial.

Para esto utilicé un toolkit de procesamiento de lenguaje natural de Python llamado *pysentimiento* que tiene buenos resultados con el español. La razón por la que previamente separé el corpús en párrafos es que este modelo fue entrenado a base de tweets, por lo que el reconocimiento de tonos en textos largos implicaba un resultado bastante cuestionable en términos de validez metodológica.

Para obtener un sentimiento en el discurso completo le asigné un puntaje de acuerdo al tono por párrafo y sumé cada uno de ellos. De esta forma se obtiene un resultado negativo si el discurso completo fue mayoritariamente negativo y viceversa para el tono positivo.

Desde el punto de vista del tono o sentimiento (Fig. 1) se pueden observar tres formas discursivas bien diferenciadas. En el caso de Fernández, aunque tiene una proporción positiva predominante, hay cierto equilibrio con 26,6%, 30,6% y 42,7% de tono negativo, neutro y positivo, respectivamente. En el caso de Macri, resulta especialmente interesante que más de la mitad de su discurso en todo el mandato fue en una tonalidad positiva y solo un 14.5% puede caracterizarse como negativo. Por último, Milei presenta el caso opuesto al anterior. Aproximadamente la mitad (54.9%) de su discurso hasta la fecha presenta una tonalidad negativa. Le sigue un contenido neutral con una proporcion porcentual del 30.5% y por ultimo un 14.5% del discurso puede considerarse negativo.

Antes de proseguir han de hacerse dos puntualizaciones. Primero se debe tener en cuenta que el sentimiento negativo no se traduce linealmente a agresividad. Por ejemplo, el gobierno de Fernández fue atravesado por una pandemia, por lo que tiene sentido imaginarse altos grados de elementos negativos que comuniquen, por ejemplo, tristeza o lamento hacia la situación. *Pysentimiento* tiene una funcionalidad que permite reconocer *Hate Speech* que podría resultar útil para hacer algunas observaciones mas detalladas al respecto, sin embargo por cuestiones de tiempo de procesamiento no continué por ese camino. Segundo, cabe recordar que Javier Milei no ha terminado su mandato por lo que apenas se está evaluando un año del cargo.



También es posible evaluar la evolución del sentimiento en el tiempo (Fig. 2). En este punto, son llamatorios los picos de positividad que tiene Macri a comienzos del 2016, 2017 y 2018. En la Tabla 1 se muestran los títulos correspondientes a esos tres momentos. De forma simétrica se pueden ver en la Tabla 2 los discursos con tonalidades más negativas que ha presentado Milei. Resulta llamativo que en solo un año de mandato haya alcanzado niveles tan bajos en varias ocasiones.

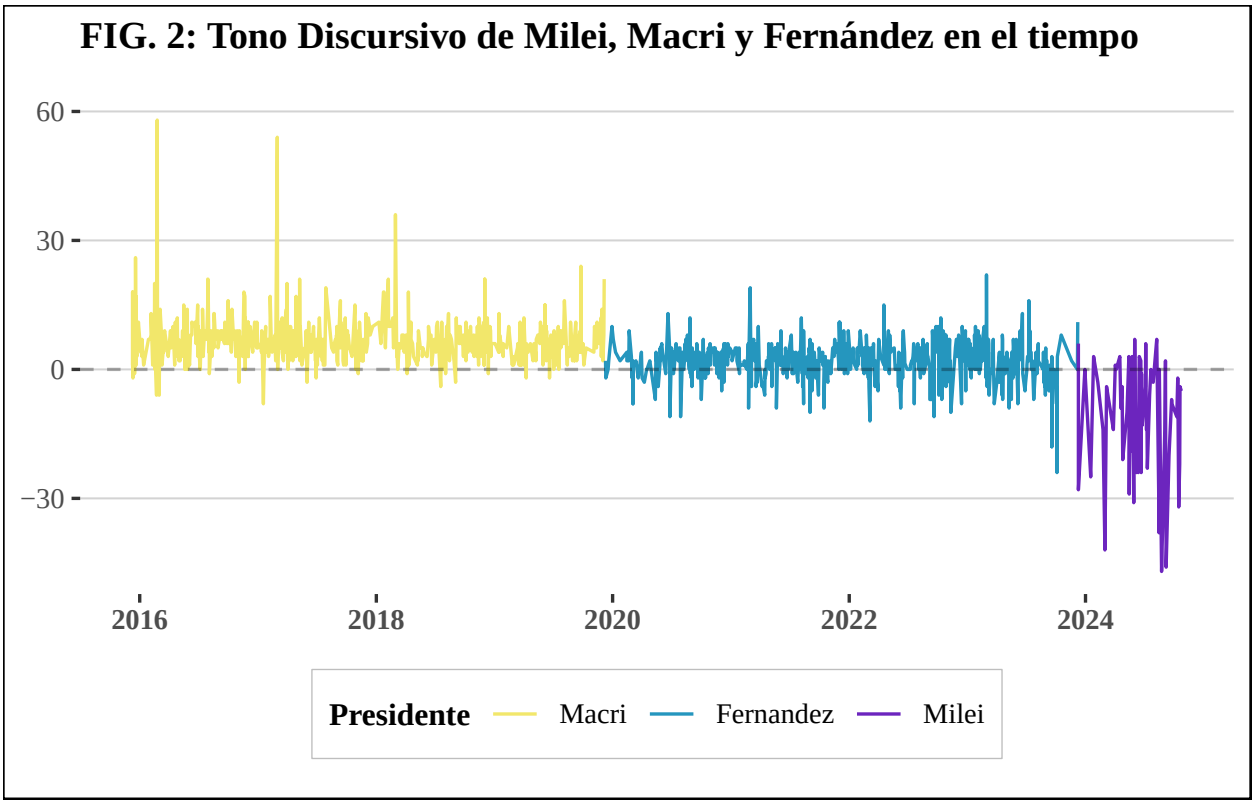


Table 1: Discursos de Macri con mayor positividad.

titulo	fecha
Mensaje Del Presidente Mauricio Macri En La Apertura Del 136° Periodo De Sesiones Ordinarias Del Congreso	2018-03-01
Discurso Del Presidente Mauricio Macri En La Apertura Del 135° Periodo De Sesiones Ordinarias Del Congreso De La Nacion Argentina	2017-03-01
Declaracion Conjunta Del Presidente Mauricio Macri Y Su Par De Francia, Francois Hollande	2016-02-24

Table 2: Discursos de Milei con mayor negatividad.

titulo	fecha
Palabras Del Presidente En Jornadas Monetarias Y Bancarias 2024 ""Deficits Fiscales, Politica Monetaria E Inflacion""	2024-10-15
Discurso Del Presidente Javier Milei En La Convencion Anual Del Iaef En Mendoza	2024-09-06
Palabras Del Presidente Javier Milei En El Acto Por El 140 ° Aniversario De La Bolsa De Comercio De Rosario.	2024-08-23
Palabras Del Presidente De La Nacion En El Congreso De Inversiones Inmobiliarias	2024-08-15
Clase Abierta Del Presidente De La Nacion, Javier Milei, En El Instituto Hoover, De La Universidad De Stanford, En California, Estados Unidos	2024-05-29
Palabras Del Presidente De La Nacion, Javier Milei Al Inaugurar El 142 Periodo De Sesiones Ordinarias De La Asamblea Legislativa, Desde El Congreso De La Nacion	2024-03-01

Frecuencia de Ocurrencia

La segunda interrogante que me surgió fue qué palabras se repetían en el discurso. Además me pareció importante agregar capacidad de análisis logrando identificar qué categoría gramatical corresponde a cada palabra del corpus. Para esta tarea utilicé *Udpipe*[1].

Utilizando el paquete *Tidytext* tokenizé la base *casarosada* y realicé una intersección con la base resultante de *Udpipe* para obtener las categorías de las palabras en una sola base. De esta forma puedo observar fácilmente frecuencia de ocurrencia manteniendo otras variables que puedo incorporar al procesamiento.

Considero que que no tiene mucho valor analítico utilizar frecuencias absolutas, por lo que utilicé el estadístico TF_IDF que permite medir el peso de las palabras dentro de un corpus de texto. Se muestran

[1]Para esto utilice de referencia el procesamiento en: https://www.elinagomez.com/blog/2023-05-4-palabras-presidente/

dos visualizaciones. Uno muestra las palabras con mayor peso cuando se lo compara entre los presidentes (FIG. 3). Con esto pretendo una exploración rápida de algunas temáticas que caracterizaron cada uno de los mandatos.

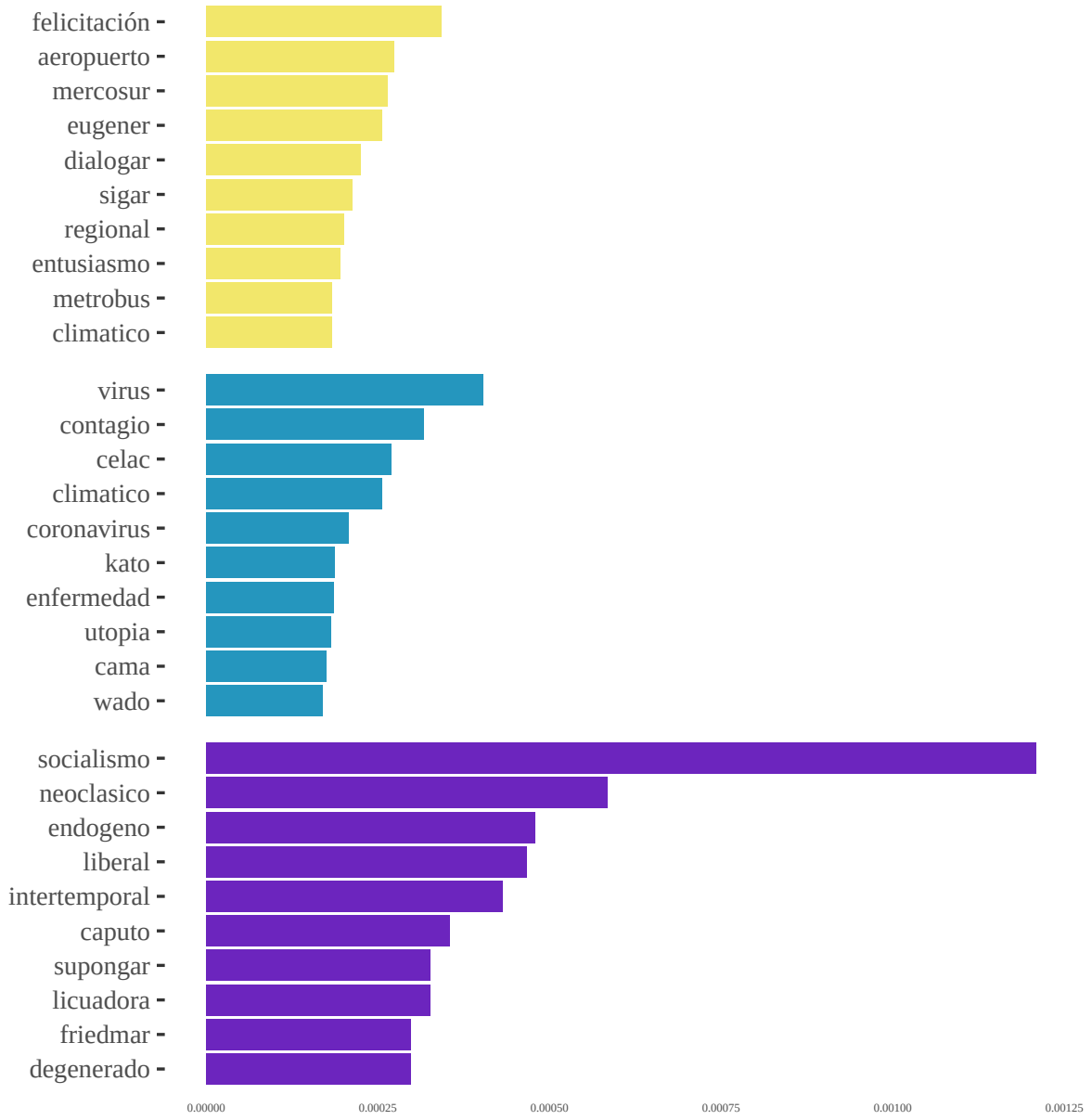
Lógicamente las palabras con mayor ocurrencia de Fernández son aquellas que tienen que ver con el contexto de emergencia sanitaria. También es destacable la presencia de figuras como Wado (de Pedro), sin embargo es esperable que al comparar entre presidentes adquieran relevancia algunas personas específicas a cada uno de los períodos presidenciales . Por otro lado, Macri demuestra, sin una contextualización previa y sin intenciones de ahondar demasiado, en su uso de palabras una retórica que me atrevería a considerar estándar (Diálogo, mercosur, regional, climático).

El caso de Milei, como era de esperarse, difiere ampliamente de los dos casos anteriores. En este sentido, es notorio el componente económico de su discurso[2]. La presencia de "socialismo", "neoclásico", "liberal" y "keynes" resulta bastante ilustrativo de este punto. Además, como en el caso de Fernández, aparecen menciones a varios integrantes de su gobierno (Pettovelo, Caputo, Bausilli).

[2]De hecho, en su momento intenté hacer un topic modeling pero tuve que descartarlo porque no lograba reconocer diferentes temáticas. En su momento pensé que era un problema de la base, sin embargo, me atrevería a decir que no tiene diversidad de temas. Tal vez, a medida que avance en su presidencia diversifica su discurso.

Fig. 3: Comparación de Palabras Relevantes entre Discursos Presidenciales

Palabras con mayor peso (TF IDF) entre los discursos de Macri, Fernández y Milei

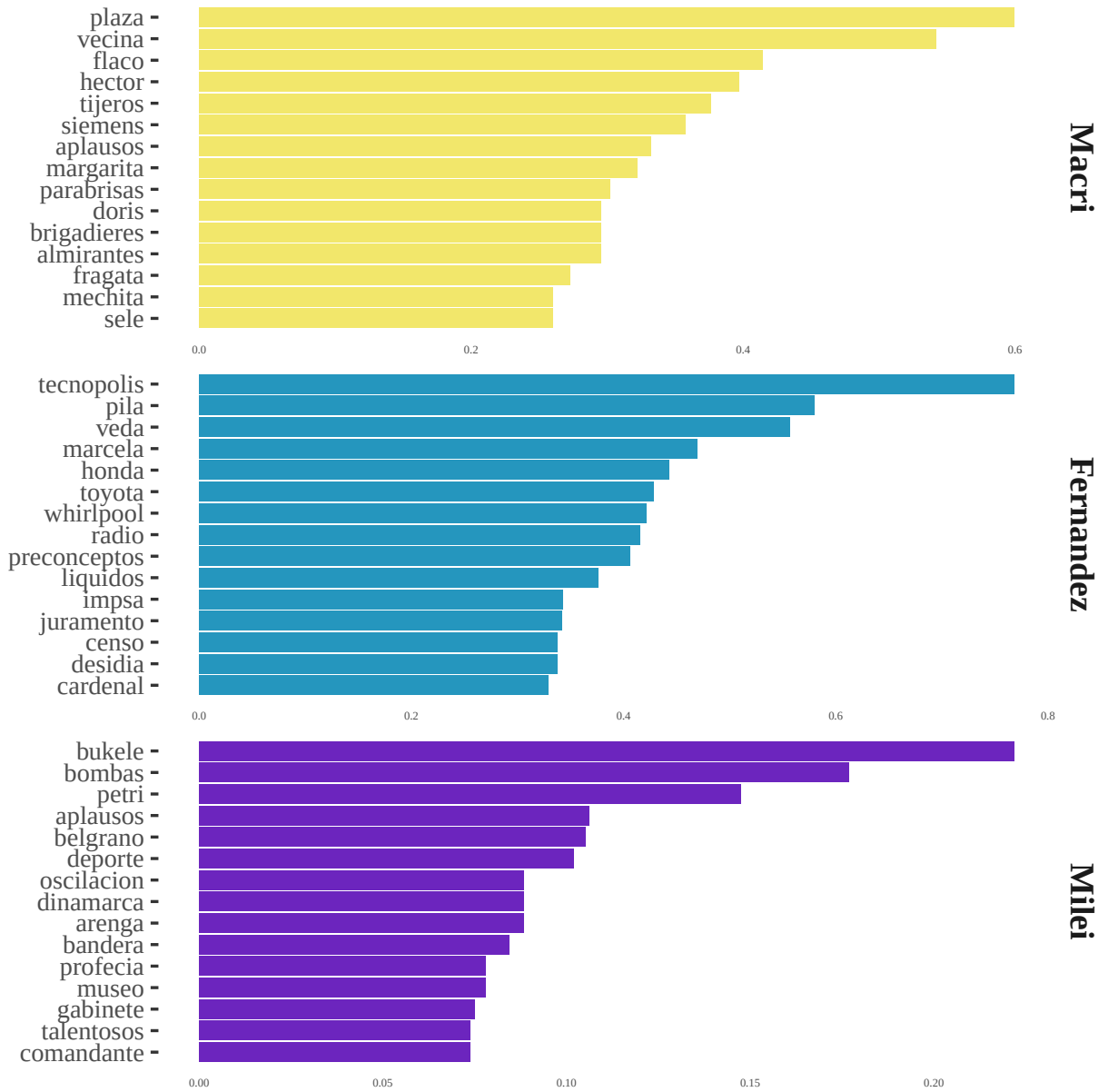


La segunda visualización (FIG. 4) ilustra el peso de las palabras en los discursos de cada uno de los presidentes. Para este punto también decidí reducir la base a sustantivos, adjetivos y nombres propios. No parecería haber nuevas observaciones relevantes mas allá de una expansión temática de lo mencionado anteriormente. Para el caso de Fernández resulta curioso que "tecnópolis" tenga más peso que otros conceptos que refieren a la situación pandémica.

En la retórica de Milei el peso que tienen los conceptos de corte económico parecen perder relevancia y aparecen menciones a personas como Bukele y Petri y símbolos nacionales como la bandera y Belgrano.

Fig. 4: Palabras Destacadas en Discursos Presidenciales

Palabras con mayor peso (TF IDF) en los discursos de Macri, Fernández y Milei

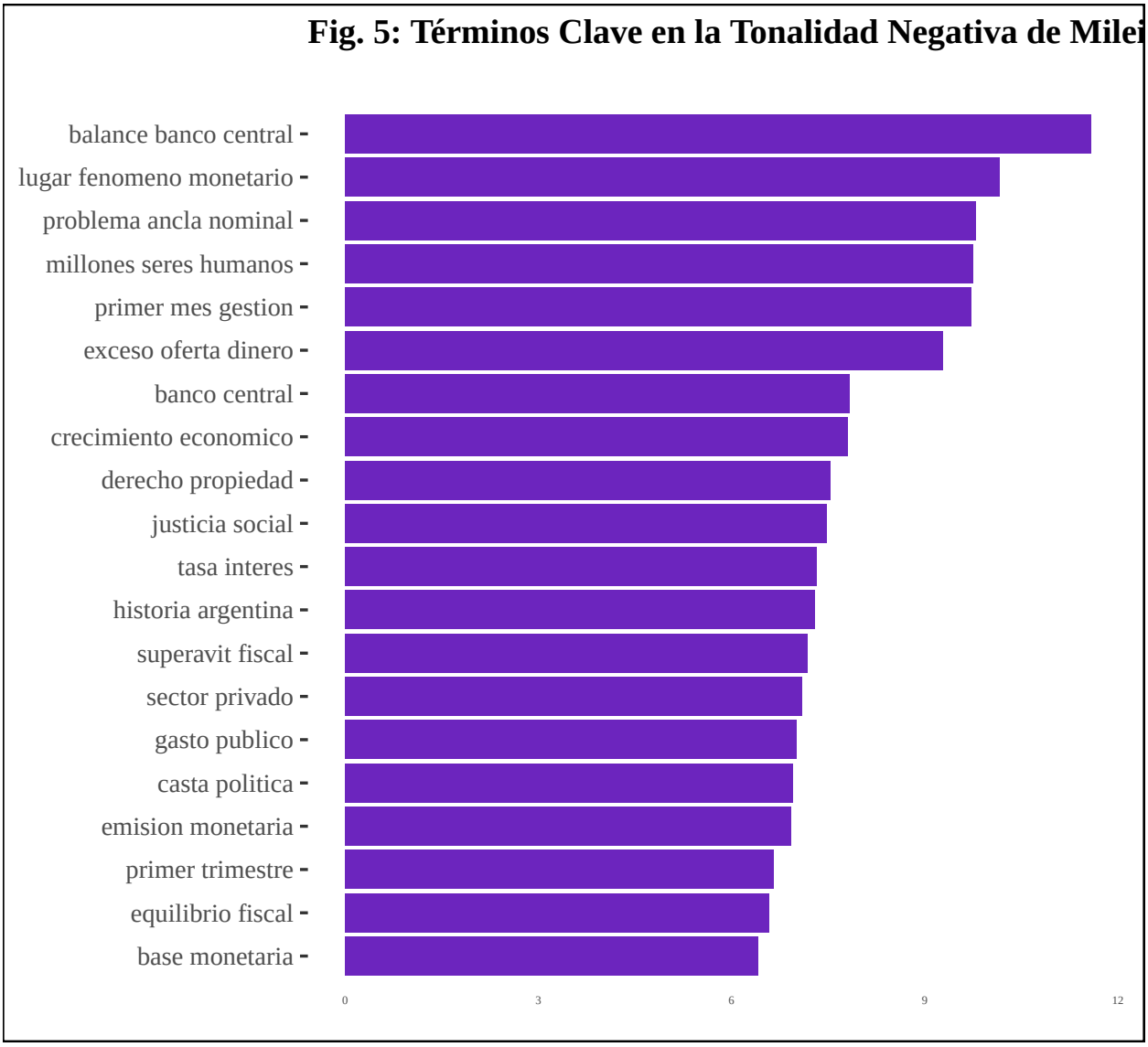


Milei

Términos Claves

A continuación quise ahondar en el caso de Milei y, específicamente, en el componente que identifiqué como negativo en su discurso. Me propuse ahondar aún más que en palabras individuales, por lo que decidí usar la función *keywords_rake* de *udpipe* que utiliza un *"algoritmo de extracción de palabras clave que intenta determinar frases relevantes en un cuerpo de texto mediante el análisis de la frecuencia de aparición de las*

palabras y su coocurrencia con otras palabras en el texto" [3].



Nuevamente, aparece en la Figura 5 una centralidad en la economía , pero nos permite obtener un poco más de contexto. En este línea, aparece una relevancia importante en el banco central, el gasto y la propiedad. Además hay una reiterada referencia a la temporalidad: "historia argentina", "primer trimestre" y "primer mes de gestión".

Co-ocurrencia

Llegando al final de mi práctica quise explorar en las posibilidades de análisis de la co-ocurrencia. Esto resultó un desafío porque quería mantener bigramas para mejorar las posibilidades de análisis: en los primeros intentos me generaba una correlacion de casi 1.0 entre banco y central, por lo que me propuse agregar algunos de estos conceptos como tokens. Esto derivó en un segundo problema: ahora el token "banco central" tenía una correlación de 1.0 con "banco". Luego de varios intentos logré generar una base tokenizada que tome en

[3] https://pypi.org/project/rake-nltk/

cuenta estos factores. Ahora la co-ocurrencia más alta de “banco central” es “inflación” con una correlacion de 0.817.

Podemos hacer lo mismo por ejemplo para “gobierno anterior” como se ve en la Tabla 3.

Table 3: Co-ocurrencia con el Término ”Gobierno Anterior”

feature1	feature2	correlation
trimestre cayo	gobierno anterior	0.7964968
equivalente	gobierno anterior	0.7879647
etapa	gobierno anterior	0.7719295
anterior hacia	gobierno anterior	0.7676369
argumentos	gobierno anterior	0.7676369
hacia deficit	gobierno anterior	0.7676369
periodistas ensobrados	gobierno anterior	0.7676369
terminos positivos	gobierno anterior	0.7676369
banco central	gobierno anterior	0.7545666
sentido	gobierno anterior	0.7439739

Es claro que existe en el discurso de Milei una fuerte asociación entre “gobierno anterior” y la situación económica. También aparece banco central y, curiosamente, periodistas ensobrados con una correlación considerable.

Key Word in Context

Por último, me pareció importante expandir brevemente en el contenido propiamente dicho a través de la técnica denominada *Keyword in context (KWIC)* que permite observar un fragmento de texto del corpus que contiene algún concepto de interés en el que se quiera ahondar de forma agregada. Seleccioné dos conceptos: “periodistas ensobrados” (Tabla 4) que se obtuvo en la sección anterior y “casta política” (Tabla 5) , que se ha convertido casi en un eslógan asociado a su persona.

Table 4: Palabra clave en contexto (KWIC): ’periodistas ensobrados’

fecha	contexto
2024-09-06	van meses vayan y vean mis archivos no los editados vean los completos no los que los sucios de los periodistas ensobrados no todos una gran parte editan para deformar la informacion y para mentir entonces vayan y vean los videos originales
2024-09-02	objetivos por si a alguien le interesa ir a las fuentes y dejar de depender de la informacion inventada por periodistas ensobrados en ese mismo riesgo pais que sube cada vez que los degenerados fiscales del congreso pasan leyes impagables sin decir
2024-06-12	sea la casta no solo son los politicos ladrones sino que tambien son los empresarios prebendarios son los medios y periodistas ensobrados pero tambien estan los profesionales que son complices de estas aberraciones entonces otra cosa que se nos exigio fue que

Es posible confirmar algunos comentarios respecto a la forma disruptiva de Milei que, claramente, no ha perdido intensidad durante su presidencia. Además, se puede ver una doble intención. Primero de generar desconfianza ante una figura enemiga representada por el sistema político y los medios de comunicación. Y segundo, una glorificación de su persona y lo que representa como cabeza de Estado; “El primer libertario” presidente de la Argentina.

Table 5: Palabra clave en contexto (KWIC): 'casta politica'

fecha	words
2024-08-23	a dejar pasar ni un milimetro a los degenerados fiscales que quieren arruinar este pais por tener beneficios para la casta politica
2024-08-23	bajando la inflacion despues voy en la parte monetaria voy a entrar en los detalles evidentemente la basura de la casta politica quiere rompernos el equilibrio fiscal saben por que porque si sale bien se les termino el curro entonces es importante
2024-09-06	no tiene esa consigna es decir por lo tanto no tienen el respaldo que aquellos billetes tenian es decir la casta politica nos volvio a estafar pero nos dijeron que esa vez o esta vez va a ser diferente y no fue
2024-09-06	millones de argentinos que no les alcanza para comer ese es el desastre que lo generaron lo generaron toda la casta politica no nosotros que vinimos a arreglar todos estos despioles y les paso un dato mas porque dicen no la caida
2024-09-06	estamos bajando la inflacion sino que ademas tampoco cayo tanto la actividad yo entiendo esto es muy doloroso para la casta politica les estamos demostrando que son unos inutiles que no sirven para nada que lo unico que hacen es dano y

#