



UNIVERSIDAD TÉCNICA
FEDERICO SANTA MARÍA

Departamento de Ingeniería Comercial

**MEMORIA PARA LA OBTENCIÓN DEL GRADO DE
INGENIERÍA COMERCIAL**

**ANÁLISIS DE TENDENCIA POLÍTICA POR MEDIO DE
ANÁLISIS DE TEXTO Y APRENDIZAJE AUTOMATIZADO DE
LA OPINIÓN PÚBLICA. UN CASO DE ESTUDIO DEL
PROCESO CONSTITUCIONAL 2019-2023.**

Benjamín Villalobos Pino, Dr. Fernando Díaz Hurtado, Dr. Pablo Isla Madariaga

Universidad Técnica Federico Santa María

SANTIAGO, MARZO, 2024

Agradezco a Dios por la oportunidad de en ese tiempo, con poco conocimiento, poder tener la fortuna de elegir mi profesión correcta desde un principio. Agradecerle, porque siempre estuvo allí, incluso cuando pensé que no. Por ser el sustento de mis días y el descanso de mis noches.

Agradezco a mi familia. Agradezco a mis padres por sublime esfuerzo realizado solo para que tuviera la oportunidad de estudiar. Gracias por su educación, por entregarme todo lo que estaba en su alcance, para que fuese feliz y fuese la persona que soy hoy. Este título es de ellos, más que mío. A mis hermanas Noemí y Abigail que me han acompañado a lo largo de todo mi proceso y siempre estuvo su consejo

Agradecer a mi novia Nicole, quien estuvo 3 años acompañándome en este proceso y sin ella siempre empujándome esto no sería posible. Gracias por estar siempre para mí, por ser mi refugio, mi mano dercha, mi amiga, mi confidente. Mi día a día sería distinto sin ti. Gracias por ser ese hombro para llorar, mi batería anímica y la inhibidora de malos pensamientos y ansiedad.

Agradecer a mis amigos por tan maravillosa compañía en este último año. A Jorge por su amabilidad y lealtad infinita. Por su ayuda directa en esta tesis para solucionar muchos problemas computacionales. A Constanza y María Olga, mis compañeras de proyecto con quienes mutuamente compartimos alegrías, tristezas, enojos, muchísimo estrés y sobre todo cariño. Sin su apoyo, esto no sería posible. Por último a Santiago, Juan, Bastián y Anaís.

A mi profesor Fernando Díaz, quien me ayudó con todas las herramientas necesarias para realizar esta tesis, tanto teóricamente como empíricamente. Por estar siempre disponible para ayudarme en tareas académicas y personales.

No soy lo que soy y soy lo que no soy. Soy mi pasado, porque me constituye, pero no me determina. Soy mi futuro, porque mi futuro afecta mis decisiones del presente, pero el futuro aún no es. Si no soy lo que soy y soy lo que no soy, entonces soy nada. Pero como soy nada, puedo ser cualquier cosa. Si no soy lo que soy y soy lo que no soy, y mi pasado no me determina y el futuro está abierto, todo el tiempo puede ser algo distinto a lo que he sido.

Índice

1. Problemática	5
1.1. Introducción	5
1.2. Justificación	6
2. Marco Teórico	13
2.1. Aprendizaje Automatizado (<i>Machine Learning</i>)	13
2.2. Procesamiento de Lenguaje Natural (NLP)	14
2.2.1. Corpus	14
2.2.2. Bolsa de Palabras (<i>Bag-of-words</i> y <i>tokens</i>)	16
2.2.3. Matriz de Características del Documento (DFM)	17
2.3. Text Analysis	18
2.3.1. Modelamiento de Tópicos	18
2.3.2. Análisis de Agrupaciones (<i>Clustering</i>)	19
3. Objetivos	26
3.1. Objetivo general	26
3.2. Objetivo específicos	26
3.3. Alcances	26
4. Caso de estudio	27
4.1. Historia del Proceso Constitucional chileno	27
4.2. Metodología de Investigación	29
4.2.1. ¿Cómo se realiza?	29
4.2.2. ¿Cómo se mide opinión pública?	30

4.2.3.	¿Cómo se interpreta, de una forma escalable, las opiniones?	31
4.3.	Procedimiento	32
4.3.1.	Extracción de la Información	32
4.3.2.	Carga, limpieza y transformación de datos	33
4.3.3.	NLP	34
4.3.4.	Modelamiento de Tópicos	35
4.3.5.	<i>Clustering</i>	37
4.3.6.	Análisis cuantitativo de Sentimientos	39
4.4.	Resultados	41
4.4.1.	Nube de Palabras	41
4.4.2.	Términos Relevantes (<i>Keyness</i>)	54
4.4.3.	<i>Topic Modelling</i>	67
4.4.4.	Análisis Cuantitativo de Sentimientos	70
4.4.5.	<i>Clustering</i>	79
5.	Conclusiones	82
5.1.	Conclusiones del caso	82
5.2.	Conclusiones finales	84
6.	Referencias bibliográficas	86

1. Problemática

1.1. Introducción

En la era digital, la opinión pública se manifiesta de manera abrumadora, a través de diversas plataformas en línea. Utilizamos X (o ex-Twitter) para expresar opiniones, instagram para subir contenido social como fotos, memes y videos; Internet para consultar por información. Este fenómeno proporciona una rica fuente de datos que puede ser analizada para entender y prever tendencias políticas, especialmente en contextos electorales y procesos constitucionales. Este estudio se enfoca en la aplicación de técnicas de Aprendizaje Automatizado (Machine Learning) para analizar, justamente estos grandes volúmenes de texto provenientes de diversas fuentes digitales con el objetivo de anticipar y comprender las preferencias políticas en eventos clave, como elecciones presidenciales, votaciones senatoriales, procesos constitucionales y cualquier otro evento político, cuyo voto reside en la ciudadanía.

El análisis de texto mediante *Machine Learning* ha emergido como una herramienta poderosa para extraer información valiosa de grandes conjuntos de datos no estructurados. La capacidad de procesar opiniones expresadas en plataformas digitales ofrece una oportunidad única para comprender la dinámica de la opinión pública en tiempo real. Este enfoque no solo permite capturar la diversidad de perspectivas, sino también anticipar patrones de comportamiento que influyen en la toma de decisiones políticas.

A pesar del crecimiento exponencial de la información disponible en línea, existe una brecha en la capacidad de los métodos tradicionales para analizar y sintetizar opiniones de manera eficiente y precisa. La necesidad de una aproximación más avanzada se vuelve evidente al considerar la rapidez con la que evolucionan las dinámicas políticas y la complejidad de los factores que influyen en la formación de opiniones. En este contexto, surge la pregunta central de este estudio: ¿Cómo podemos

utilizar técnicas de Machine Learning para analizar de manera efectiva la opinión pública expresada en textos digitales y prever tendencias políticas, especialmente en eventos cruciales como elecciones y procesos constitucionales?

1.2. Justificación

Aunque existen muchísimos enfoques de los que es abordable a esta extraña, aunque ya resuelta, situación, política, económica, social, legal, etc., el enfoque más pulcro que puede ser obtenido es a través del análisis de la opinión de la gente. Es la opinión pública la que en sí misma es una medida de satisfacción –o insatisfacción–, que se traduce en demandas sociales. Dahl argumentaba que “una característica clave de la democracia es la capacidad de respuesta continua del gobierno a las preferencias de sus ciudadanos”(Dahl, 1993).

Estas demandas sociales y nuestra hipótesis es que tienen efecto en las políticas públicas. En el estudio "Opinión pública y políticas públicas: la sensibilidad de los ciudadanos y la receptividad de los gobiernos en España" realizado por Ruth Cicuéndez Santamaría (2021) analizaba el índice de demanda de gasto y las asignaciones presupuestarias en la Defensa Nacional, Educación, Salud, Pensiones, entre otras; desde 2005 a 2019, en el cual se demostraba la significancia estadística de este modelo:

$$DG_{p(t)} = \alpha + \beta_1 PR_{p(t)} + \epsilon$$

, tal que $DG_{p(t)}$ es la intensidad de las preferencias de gasto para la política p en el año t , medido por el índice de demanda de gasto, y $PR_{p(t)}$ representa el presupuesto en millones de euros asignado a la política en ese año en la política p en el año anterior $(t - 1)$. (Cicuéndez, 2021)

Los resultados se muestran en la figura 1:

La receptividad de los ciudadanos a las decisiones gubernamentales sobre gasto público (2005-2019). Análisis de autorregresión

	Correlación¹	R	R²	Constante	β_1	beta	t	Sig. t
Defensa	0,283	0,126	0,016	-31,67	0,000851	0,126	0,421	0,68
Seguridad ciudadana	0,115*	0,330	0,109	-18,54	0,001859	0,330	1,160	0,27
Educación	-0,130*	0,666	0,444	141,81	-0,001762	-0,666	-2,964	0,01
Sanidad	-0,118**	0,594	0,353	164,89	-0,001287	-0,594	-2,448	0,03
Pensiones	0,852**	0,838	0,703	9,19	0,000282	0,838	5,325	0,00
Protección al desempleo	-0,144	0,020	0,000	52,84	-2,797E-5	-0,020	-0,670	0,98
Cultura	-0,623*	0,477	0,228	82,13	-0,002824	-0,477	-1,800	0,09
Protección del medioambiente	-0,079	0,189	0,036	79,16	-0,001388	-0,189	-0,638	0,84

Figura 1: Receptividad histórica económica de Chile. Fuente: CLAD

Según la teoría del modelo del termoestado (Wlezien, 1995), si los individuos son sensibles a las decisiones de los responsables públicos, el aumento (o disminución) de la demanda de gasto en una política correlacionará negativamente con el incremento (o reducción) del gasto asignado a esa política. Esto quiere decir que cuando el gasto destinado a la política p aumente, las preferencias ciudadanas se ajustarán, disminuyendo la demanda de recursos; y cuando el gasto se reduzca, aumentará la demanda de recursos.

Sin embargo, nos podemos dar cuenta que: 1) no todas las políticas son estadísticamente significativas, por lo que en sí mismo no podría argumentarse que la opinión pública tiene un efecto total en todas las políticas pública; 2) pareciera ser que los betas son económicamente no significativos, debido a que su valor de beta es bajo. Sin embargo, esta afirmación no es real, debido a que estamos comparando en millones de euros, por lo que una disminución de un millón de euros en salud, por ejemplo, significaría un aumentos del interés por salud evaluada en 1.282 euros; 3) existen factores

que tienen una correlación positiva por lo que, indicaría que un aumento del presupuesto en ella, tiende a aumentar la demanda por esta política, como lo es seguridad y pensiones. Esto porque se teoriza que ambas políticas tienen un efecto más indirecto que los otros y, que por lo tanto, puede que los individuos solo reajusten sus preferencias si las medidas adoptadas son sustantivas y evidentes, o que solo dispongan de información elemental y reaccionen a las decisiones políticas en su conjunto (Wlezien, 1995)

Entonces, el estudio es relativamente funcional en su argumento estadístico al responder si es la política tiene una respuesta en la opinión.

Desde un punto de vista sociopolítico y filosófico, podemos analizar las palabras de Sartori “es en el voto como el ciudadano termina por expresar su propia opinión” (Sartori, 1992)

Lo anterior, basado en que el voto es en sí mismo un acto racional y que está infundado sobre la racionalidad de la opinión pública, ya que los cambios en sus orientaciones responden al devenir de los acontecimientos y a la disponibilidad efectiva de información, hecho que demuestra su capacidad para establecer juicios coherentes a lo largo del tiempo (Del Valle, 2009)

E incluso en sus decisiones irracionales:

“Si bien el hombre de la calle carece de los elementos necesarios para tomar una decisión racional, al menos cuenta con elementos para actuar como si fuera racional: haciendo lo que hacen sus pares (Columbia) o, simplemente, reproduciendo la misma conducta electoral en función de un apego psicológico (Michigan)”(Del Valle, 1996)

Con todo, este es solo una visión de un extenso debate, psicológico y filosófico que lejos de solucionar, más bien ha pretendido generar un hincapié en que, si suponemos cierto que, es la opinión pública, definitoria en cuanto a políticas sociales y en voto, entonces es particularmente atractivo su análisis para la definición de tendencias políticas.

La justificación de esta investigación radica en su potencial para proporcionar a los actores polí-

ticos, analistas y ciudadanos en general, una herramienta más precisa y rápida para comprender las dinámicas políticas en evolución, permitiendo una toma de decisiones más informada y contextualizada, además de tener generar valor agregado al ser una herramienta bastante objetiva comparada con otras como lo son las encuestas, las opiniones profesionales y otros análisis cuantitativos derivados, cuya subjetividad muchas veces está basada en el sesgo de confirmación y representatividad, o simplemente intereses propios.

"El denominado sesgo a mi favor (myside bias), de aquí en más SMF, ha sido caracterizado como la tendencia a considerar y ponderar de manera positiva aquellos argumentos consistentes con las propias opiniones, al mismo tiempo que se ignoran y se evalúan de manera negativa aquellos argumentos que contradicen estas opiniones y creencias previas(Perkins, 2019)

De hecho, es el ejemplo de encuesta, un gran punto de inflexión sobre la tesis de este proyecto. Pensemos en la encuesta política realizada por CADEM. CADEM es una organización privada que realiza diferentes estudios de marketing como Costumer Journey, Testeo de Publicidades, entre otros, dentro de los cuales se encuentran los Estudios de Opinión Pública. Este tipo de encuesta es realizada, a través del servicio llamado Plaza Pública, según reporta su web, a través de la siguiente metodología:

"Nuestro estudio contempla la realización de una encuesta probabilística de 700 casos semanales (con un consolidado mensual que va desde las 2.800 encuestas a 3.500 dependiendo de si el mes tiene 4 o 5 semanas), aplicada en un 100 % a través de teléfonos celulares, utilizando una base de datos propia de Cadem que contiene más de 18 millones de números, tanto de prepago como de post pago, todos ellos obtenidos a través de Random Digit Dialing y consolidados durante los últimos años. Su grupo objetivo son hombres y mujeres mayores de 18 años, habitantes de las 16 regiones del país, en territorios urbanos o rurales. La muestra del estudio es estratificada previamente por región, a partir de las proyecciones poblacionales realizadas por el INE para el año 2023, según el

último Censo disponible (2017). Las comunas, por su parte, son seleccionadas de forma aleatoria. Además, contamos con un consejo consultivo que asesora en la elaboración del cuestionario semanal de Plaza Pública”(Sitio web: CADEM,2024)

Sin embargo, según el estudio realizado por San Martín y Alarcón-Bustamante, 2022., “Dissecting Chilean surveys: The case of missing outcomes”, es justamente la metodología de esta encuesta y la realización de supuestos – lejanos a la realidad –, lo que conllevan a inferencias erróneas sobre los resultados.

Una de las preguntas de su caso de estudio era: “Independiente de su posición política, ¿Usted aprueba o desaprueba la forma como Gabriel Boric está conduciendo su gobierno?” teniendo 4 alternativas votadas: 36 % aprueba la gestión del presidente; 53 % desaprueba la gestión del presidente; 5 % no aprueban ni desaprueban la gestión del presidente; y 6 % no saben o no responde. Empero, es la crítica de esta publicación, el hecho de que por probabilidad, tal aseveración es incorrecta, pues habla de una certeza, más que de una inferencia. De hecho, una de las sugerencias del estudio es la aclaración de los intervalos de confianza, a la hora de declarar un resultado. No obstante, en la práctica Cadem, no lo realizaría (San Martín & Alarcón-Bustamante, 2022).

El estudio demuestra que el intervalo es tan pobre que no genera ninguna información, puesto que el intervalo calculado, por ejemplo para la pregunta anterior, fluctúa entre 0,0693 y 0,9593, lo que debería parafrasearse en la siguiente afirmación: “al menos 6.93 % de las personas encuestadas tienen una imagen positiva o muy positiva del presidente Gabriel Boric, y a lo sumo 95.93 % de las personas encuestadas tienen una imagen positiva o muy positiva del presidente”. La amplitud del rango es de 89 % lo cual es casi desinformativo. Por supuesto que no le conviene a Cadem, hablar de esto, ya que es su negocio (San Martín & Alarcón-Bustamante, 2022).

Pese a que este documento es una crítica clara a Cadem y la metodología de recolección de información, también lo es indirectamente a la medición en sí de la opinión pública que muchas

veces pareciera predecir de mala manera los resultados.

Aún más, es en base a esta publicación, que el Laboratorio Interdisciplinario de Estadística Social (LIES UC)(2022), utilizando los casos de estudios de las encuestas Casem, concluye, frente a las encuestas que:

“En lugar de considerarlas (las encuestas) como representando una sola opinión, hay que verlas como la posibilidad de manifestar la pluralidad de opiniones políticas. Esta pluralidad la hemos representado por medio de un rango de porcentajes posibles, el cual se puede establecer usando los datos proporcionados por la encuesta. Es por ello que, tampoco los resultados de una encuesta, como por ejemplo la CADEM, no deben verse como una predicción de lo que ocurrirá políticamente en el futuro. Es más, lo que se espera en política son acontecimientos, esto es, hechos que interrumpan el (y los) proceso(s) rutinario(s)”(UC, 2022)

En el contexto chileno el proceso constituyente, las opiniones son el aspecto más importante de cara a su futuro, puesto que son estas opiniones las que deciden, ya sea de forma informada o no, si es son las nuevas constituciones, lo que realmente el ciudadano demanda. Ruth Cicuéndez Santamaría (2021) comentaba lo mismo, parafraseando a su vez a diversos autores, afirmando que:

“La gobernanza democrática se fundamenta en involucrar a los ciudadanos en las políticas públicas para asegurar que estas se desarrollan en sintonía con los intereses de la sociedad, la cual exige a la élite política que proporcione alguna respuesta a sus demandas. Desde esta perspectiva, el estudio de la congruencia entre preferencias sociales y políticas públicas resulta fundamental para medir la calidad de un sistema democrático, porque permite evaluar si la actividad gubernamental y los resultados del proceso político se ajustan a los deseos de la población, y si los ciudadanos responden modulando sus demandas en función de tales resultados”(Cicuéndez, 2021)

A partir de aquella importancia, en el éxito de las políticas pública y la democracia per sé, es que la interpretación y la inferencia sobre la verdadera opinión del ciudadano es un elemento crucial para

entender sus demandas sociales. Por tanto, este estudio, por sobre pretender solucionar el problema de Cadem, pretende generar un complemento desde un punto de vista distinto: poder analizar la opinión pública, ya no desde un análisis cuantitativo, sino que a través desde el análisis textual de las opiniones de la gente.

2. Marco Teórico

2.1. Aprendizaje Automatizado (*Machine Learning*)

El Aprendizaje Automático según Franco Alba en "Deep Learning"(2019) es:

"La rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a los ordenadores aprender. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares" (Alba, 2022)

Por tanto, al igual que una regresión, toma la experiencia pasada para tomar decisiones futuras, con la diferencia fundamental de que buscamos que sea capaz de entender reglas o patrones, generalizando y dando como resultados un patrón automático. Este patrón a su vez permitirá realizar clasificación y predicción.

"El Aprendizaje Automático, los humanos suministramos los datos como entrada al igual que las respuestas esperadas de dichos datos con el fin de obtener a la salida las reglas que nos permiten hacer la correspondencia efectiva entre entradas y sus correspondientes salidas" (Alba, 2022)

Un claro ejemplo podría ser Netflix, quien sugiere contenido acorde al entendimiento de patrones de sus usuarios en el consumo de streaming.

Es el Machine Learning parte fundamental de la Inteligencia Artificial, puesto que a partir de aquí derivaría en el entrenamiento de la máquina, para que eventualmente, por sí misma arroje resultados propios u "originales" por medio del *Deep Learning* (fig. 2

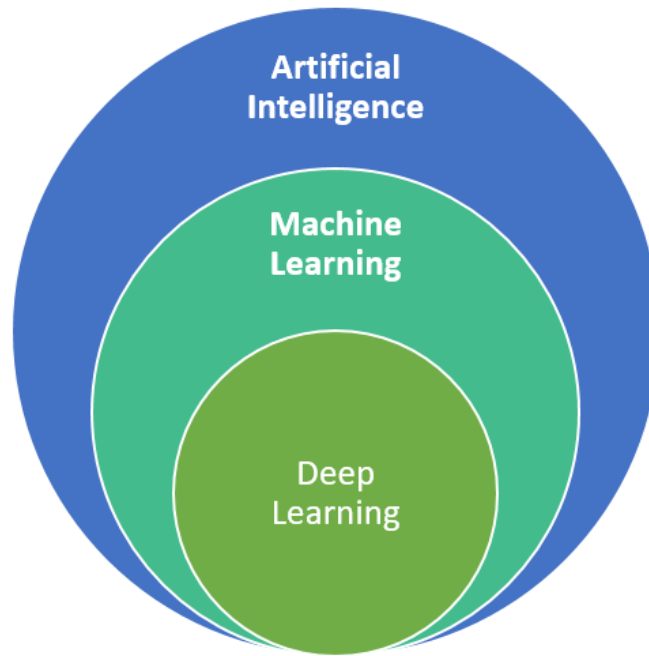


Figura 2: Jerarquía Inteligencia Artificial, *Machine Learning* y *Deep Learning*. Fuente: Master Iesc Angers

2.2. Procesamiento de Lenguaje Natural (NLP)

2.2.1. Corpus

Para el procesamiento del Lenguaje Natural, se debe realizar a limpieza y transformación de los datos, cuyo origen proviene del análisis propio del texto, tanto en su semántica, morfología, sintáxis y significados. A partir de allí, de este análisis teórico se procede a realizar un proceso que determine el *modus operandi* de cualquier texto a la hora de ser analizados.

Es así, como en primera instancia tenemos el corpus que es definido como:

"Un corpus es un conjunto de (fragmentos de) textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en

su totalidad, y reunidos con el propósito de facilitar su estudio científico (cf. Rojo 2014). " (Rojo, 2016)

Según el mismo autor, esta definición nos permite comprender que:

"Los textos deben ser naturales (no artificiales ni creados expresamente para su incorporación al corpus), han de estar en formato electrónico porque esa es la única forma de que podamos recuperar la información que precisamos, tienen que ser representativos de la variedad de la que proceden y, por último, deben permitir su estudio científico (no exclusivamente lingüístico), lo cual suele implicar la adición de información gramatical, léxica y pragmática a la simple secuencia de formas gráficas que constituyen el texto en el sentido más habitual de la palabra" (Rojo, 2016)

Por tanto, es necesario precisar que un corpus no solo requiere del análisis de texto *per sé*, sino que es necesario tener variables que nos indiquen procedencia, época, etc.

Además, son estos textos dentro del corpus, heterogéneos e independientes entre ellos. Su utilización es meramente lingüística, lo cual no obedece al entendimiento contextual de un texto en particular, mas si el entendimiento contextual del conjunto de textos, incluidos fenómenos extra-textuales.

"Un texto se lee de forma continua, línea a línea, mientras que un corpus no se lee de corrido, sino que se estudia, mediante el análisis de las concordancias, el conjunto de casos del fenómeno en cuestión contenidos en él."(Rojo, 2016)

Entonces, es en sí mismo el Corpus es la base de datos utilizada para poder desarrollar el análisis computacional del Análisis de Texto.

En el caso particular de los corpus en la codificación computacional, se debe realizar un proceso similar donde se le da como entrada los textos, llamados "documentos", para luego asignarle las llamadas *docvars* que hacen referencia a las variables extras de los documentos, las que nos permiten justamente conocer el contexto y clasificaciones texto a texto. También podemos asignar nombres a

cada texto por medio de los *docnames*. Si no se le asignan nombre, entonces el algoritmo procede a llamarlas genericamente "*text1*", "*text2*", "*text3*" y así sucesivamente.

2.2.2. Bolsa de Palabras (*Bag-of-words* y *tokens*)

A partir del corpus, como un conjunto de textos es donde se procede a descomponer estos textos en palabras únicas. Este proceso es conocido como "*tokenization*", el cual es el proceso indentifica *tokens*, que a su vez es definido como la unidad básica la cual no será descompuesta en los procesos subsecuentes (Webster & Chunyu, 1992) . De hecho, es en el NLP la entidad "palabra" su unidad más básica. Empero, es la palabra una mezcla más profunda de composiciones que la generan.

Una palabra en sí misma requiere de un lexema y un morfema. El primero es el que por sí solo tiene significado y el segundo es el que acompaña al significado para transformarlo en diferentes subsignificados. Por ejemplo, la palabra gato, se dividiría en el lexema "gat" y en el morfema "o". Si bien en si mismo "gato" hace referencia a una especie, si se cambia el morfema por "a", ahora refleja la especie, más el sexo: hembra. Por tanto, cualquier palabra que se agregue antes, en medio, o después para generar un contexto y una especificación a la palabra, entonces es definido como un morfema. Por supuesto existen direntes tipos de morfema que dan más o menos significados y que por tanto son clasificados de una u otra manera, pero no es relevante para nuestro análisis.

De lo anterior, podremos definir formalmente la palabra como "un item formal (al menos un morfema de largo), cuyo patrón de ocurrencia puede ser descrito en terminos de de una secuencia única ordenada de otros items léxicos que ocurren en su entorno" (Sinclair, 1966)

Además, es importante también tratar las colocaciones de las palabras. Las colocaciones son "construcciones de verbo con complemento preposicional de régimen" (Travalia, 2006). Un ejemplo sería: "break in", cuya combinación verbo *Break* (quebrar o romper) más un adverbio de lugar *In* (dentro), cuyo significado conjunto termina diciendo "Irrumpir". El ejercicio se complica aún más si

consideramos que existen homógrafos, es decir, palabras que se escriben igual pero tienen diferentes significados como lo es *Break*, que siendo verbo se entiende como "Quebrar", pero también como "Frenar".

Para ello, generalmente se utiliza un diccionario que permita identificar este tipo de patrones, donde se pueda filtrar y eliminar estos tokens de forma correcta, quedando solo con tokens que sean realmente significativos.

En el área computacional, existen funciones que realizan esta tarea. Las dos principales son el "Lemmatization" y el "Stemming" cada una con ventajas y desventajas evidentes. La primera busca encontrar el lemma de la palabra, permitiendo estandarizar muchas palabras a una sola raíz léxica. A cambio, requiere de mucho procesamiento para tomar cada token y encontrar su raíz. El segundo, utiliza menos procesamiento, permitiendo un análisis rápido de las palabras, con todo tiende a tener desempeño intermedio debido a que analiza las palabras utilizadas y las trunca, bajo la hipótesis de que si yo tengo la palabra *Love* y *Lover*, entonces basta con cortar la palabra para encontrar su raíz, lo mismo con una basta cantidad de verbos y sustantivos en inglés. Con todo, el español, siendo un lenguaje de conjugación más compleja, no permite el corte de muchas de las palabras, porque tenemos morfemas para definir género como lo es "Oso" y "Osa", por ejemplo. Lo mismo con adjetivos. Esto hace que, el rendimiento de *stemming* en español sea bastante deficiente.

Para este trabajo, no se realiza ninguna debido a que no se cuenta con el procesamiento y la opción eficiente es poco eficaz. Tanto así, que el análisis y sus tokens son más significativos sin cambios.

2.2.3. Matriz de Características del Documento (DFM)

Una vez realizado la tokenización de las palabras, nos queda un último paso y es vectorizar las palabras para poder acceder a funciones que nos permiten estos, como las distancias euclidianas, correlación, frecuencia relativa, entre otros.

Para ello, simplemente se toman todos los tokens de todos los textos y se extraen las palabras únicas de aquí. Esta será nuestra primera dimensión de la matriz. La segunda será la cantidad de textos dentro del corpus –en nuestro caso, 802 textos–. Luego se toma el vector de tokens de cada texto y se le asigna a la matriz el número de frecuencias encontradas en ese texto, para un token específico.

Por ejemplo, digamos que tenemos 2 textos. El primero será la frase: "Somos los mejores del mundo. Los mejores" y el segundo será: "Tengo a los mejores amigos del mundo". Definimos la primera dimensión de la matriz como las palabras únicas de ambos textos. Esto es:

Somos-los-mejores-del-mundo-tengo-a-amigos

Finalmente la Matriz quedaría visualmente algo como el estilo de la tabla 1

	Somos	los	mejores	del	mundo	tengo	a	amigos
texto_1	1	2	2	1	1	0	0	0
texto_2	0	1	1	1	1	1	1	1

Tabla 1: Tabla ejemplo de Matriz y sus dimensiones. Elaboración Propia

2.3. Text Analysis

2.3.1. Modelamiento de Tópicos

El modelamiento de tópicos nos permite encontrar subgrupos dentro de un corpus, con la finalidad de clasificar. Dada la naturaleza de la tarea se dice que el modelamiento de tópicos es un aprendizaje no supervisado, puesto que no busca en sí, ajustar el modelo a una respuesta, más bien, busca entender de las observaciones asignadas, si existen grupos de palabras, las cuales caractericen a un texto.

En ese sentido, se debe procedimentalmente realizar 2 cosas. Primero, generar una cantidad óptima de tópicos que se ajusten a los subgrupos que puedan existir en el texto. Si existen muy pocos tópicos, entonces estamos realizando un mal-ajuste debido a que baja dimensionalidad implica mayor sesgo de los tópicos, debido a que asigna dentro de ella, palabras que no son, y por otra parte, un k muy alto de tópicos puede hacer una especie de sobreajuste, asignando palabras a tópicos que no van a ser relevantes.

Segundo, que cada tópico represente correctamente a los documentos asignados, es decir, si tengo el tópico "Pensiones", debemos asercionarnos que efectivamente esta problemática sea asignada al documento, desde donde es originario. Recordemos que estos modelos solo toman bolsa de palabras y no reciben propiamente tal, una segmentación previa. De hecho justamente la idea es que encuentre por sí solo estas relaciones.

De estos modelos que son llamados "generativos", existen tres principales LDA, CTM y STM, de los cuales abordaremos el primero debido a que este análisis permite la realización de tópicos no correlacionados entre sí, lo que nos permite un mayor inferencia sobre ellos.

2.3.2. Análisis de Agrupaciones (*Clustering*)

También correspondiente a aprendizaje no supervisado, nos permite conocer por medio de distancias semánticas la agrupación y clasificación de los textos. A diferencia con el análisis anterior, esta clasificación utiliza la distancia euclidiana en las matrices de dfm , para poder tender en que tan diferentes son las observaciones. Estas observaciones, pueden provenir de muchas variables explicativas del modelo, por lo que el modelo podría tener más de dos dimensiones lo que nos imposibilitaría ver la distancia entre puntos.

Para poder visualizar la distancia que es multidimensional, debemos de trabajar en un gráfico de dos dimensiones por lo que una de las soluciones para realizar eso, sin perder significancia es realizando

el proceso de selección de variables.

En general, muchas veces en las regresiones, se tiene que una variable es significativa, aunque muchas veces su aporte al modelo, no sea en sí mismo un gran aporte. Si visualizamos el R cuadrado, podría ser que el agregar esta variable, tenga una mejora marginal, lo que no nos conviene.

Por tanto, para estos casos, se utiliza el proceso de selección de variables, eliminando algunas de ellas, que no aportan mucho al modelo. Uno de esos enfoques, es el análisis de componentes principales (PCA, en inglés) quien nos permite realizar una combinación lineal entre los aportes de todas las variables. Es decir genera una dimensión, por medio del promedio ponderado de p-variables.

Se puede realizar tantos componentes principales como variables se tiene, sin embargo, si utilizo la misma cantidad, entonces en realidad, no se está generando ninguna reducción y el modelo sigue siendo el mismo.

Este procedimiento de reducción permite quedarse con las variables más importantes, importancia medida por la cantidad de varianza explicada que crea el componente. A mayor varianza explicada, mayor es la importancia de este componente principal, siendo sus cotas entre 0 y 1.

En el caso particular del clustering, las dos dimensiones de nuestro gráfico, serán precisamente la reducción del modelo con k variables, a tan solo 2 (por medio de componentes principales) y utilizando la distancia semántica podemos visualizar las observaciones sin ningún problema.

Con todo, esto solo nos permite simplificar el modelo, mas no clasificarlo. La clasificación por su parte, se realiza a través de un algoritmo, llamado "K-means" que pretende clasificar por medio de un parametro dado, ese parámetro es el k. El k nos dice la cantidad de clusters (grupos) queremos definir, lo cual en la práctica se define el k óptimo por medio de cross-validation.

Tomando el parametro K crea k-centroides dentro del plano, distribuidos de forma aleatoria, tal y como la figura 3 para $k = 3$.

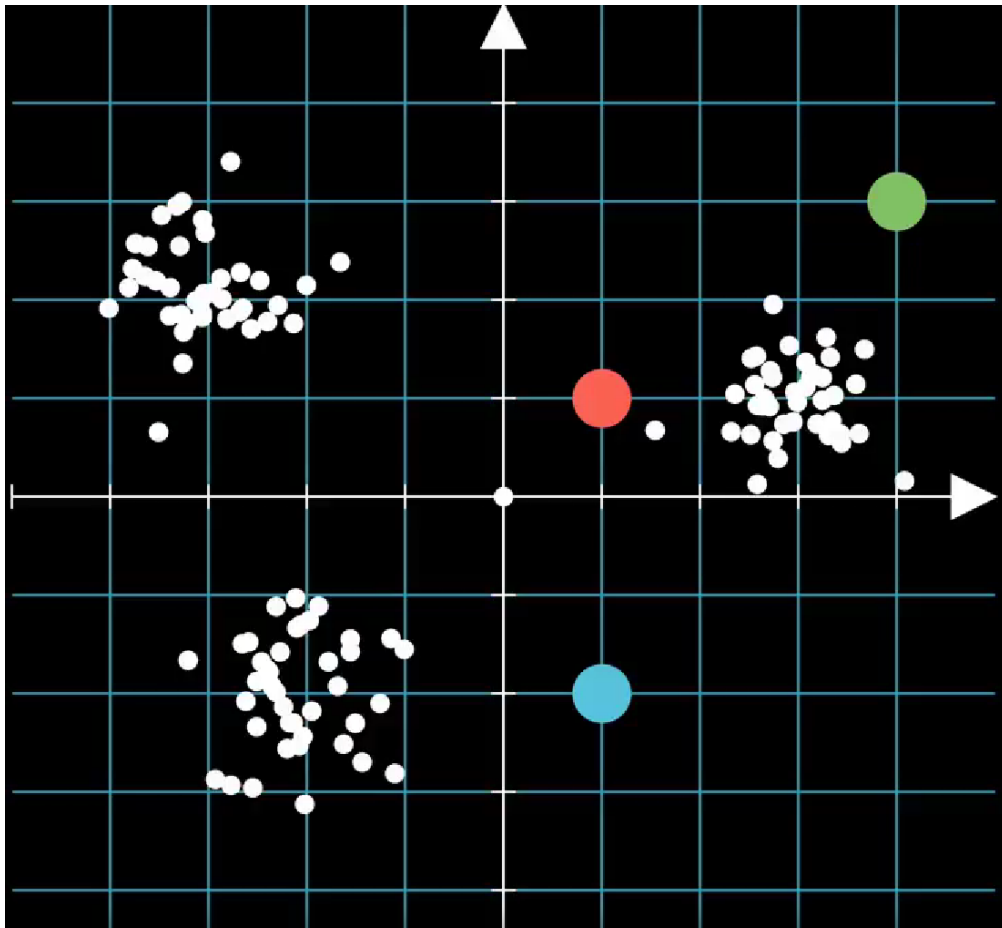


Figura 3: Paso 1: Asignar Centroides Aleatorios. Fuente: Daily Dose of Data Science

Luego, toma un punto cualquiera y calcula la distancia de este punto a cada uno de los centroides

(fig.4

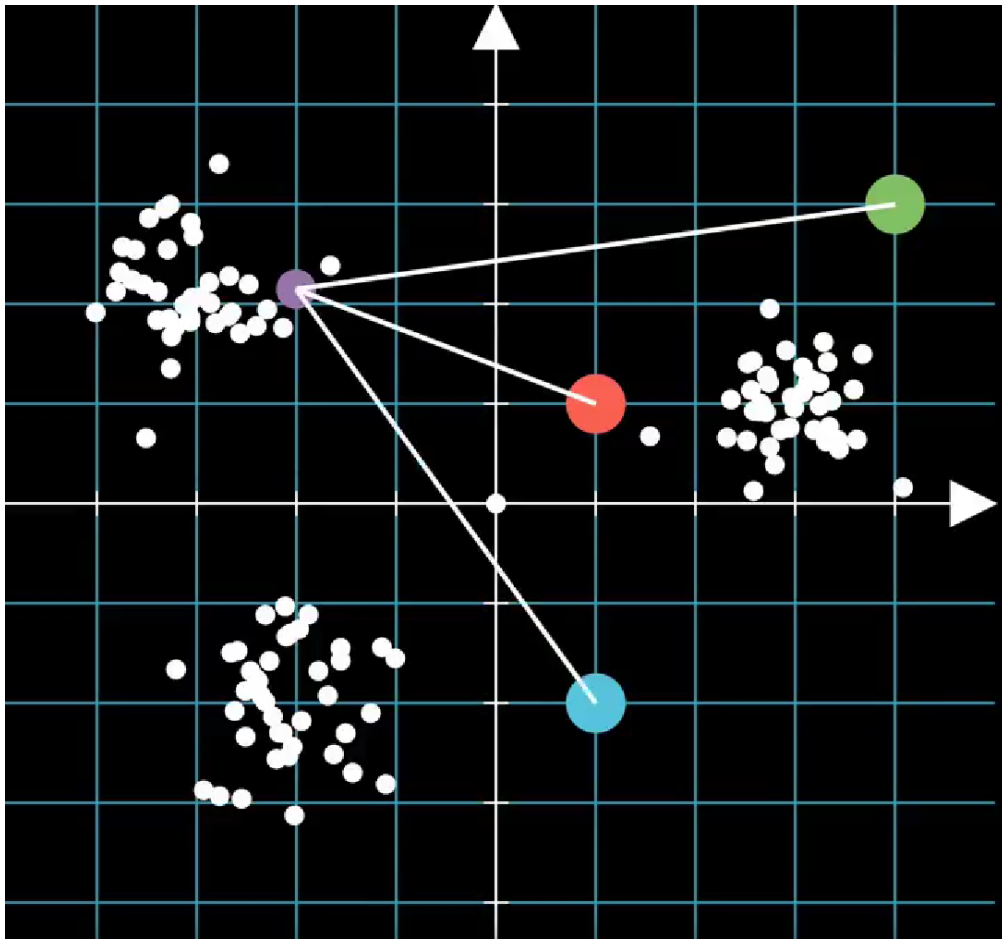


Figura 4: Paso 2: Tomar una observación y calcular la distancia con los centroides. Fuente: Daily Dose of Data Science

Finalmente, se asigna esa observación al centroide más cercano, quedando en este caso de color naranja, como se ven en la figura 5

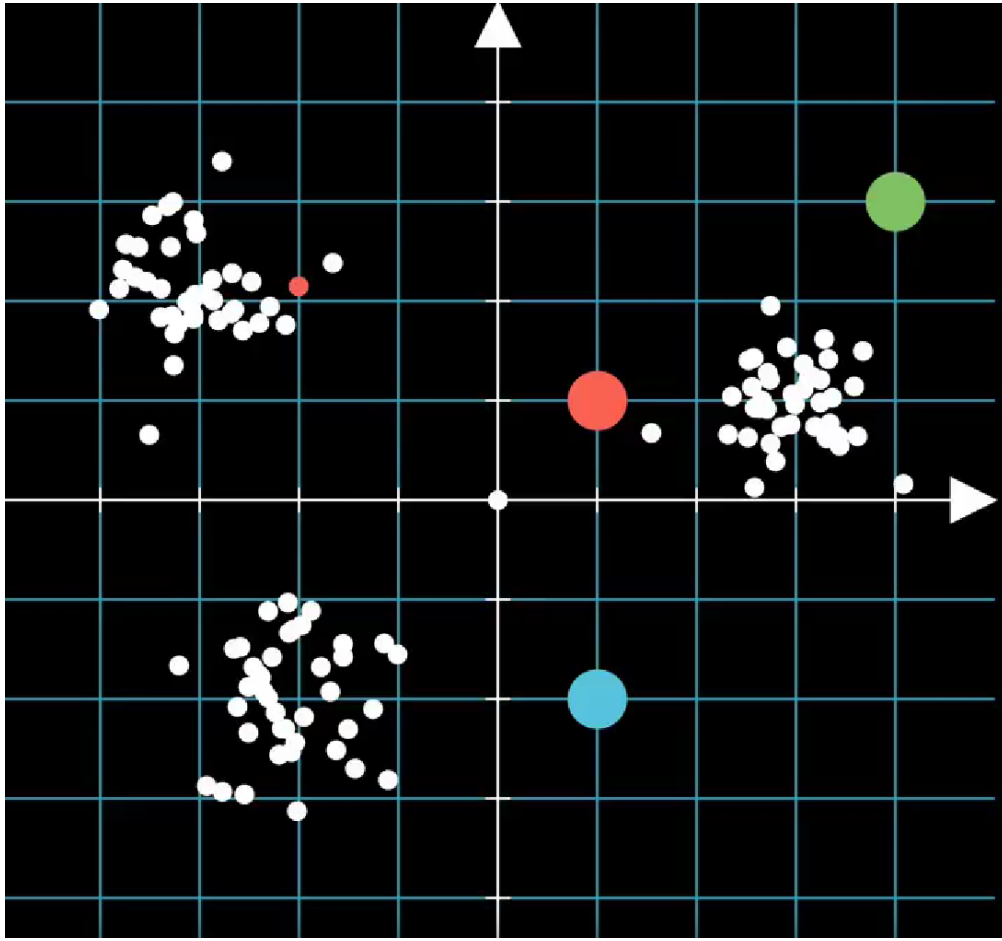


Figura 5: Paso 3: Asignar la observación al centroide más cercano Fuente: Daily Dose of Data Science

Repitiendo el proceso para todas las observaciones, se tiene la figura 6, donde todos los puntos están clasificados en un grupo (*cluster*). Con todo, por supuesto que no es la clasificación óptima, puesto que los centroides asignados aleatoriamente no estaban en la mejor posición.

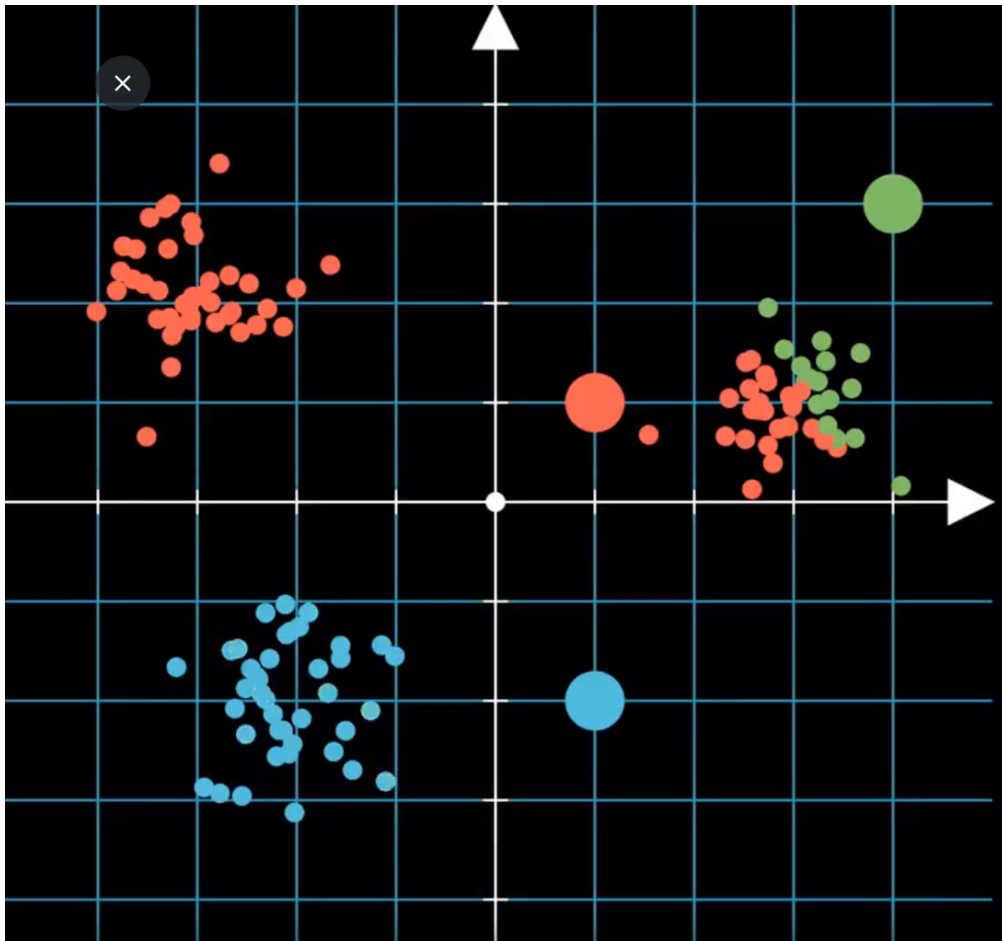


Figura 6: Paso 4: Repetir el proceso para todas las observaciones Fuente: Daily Dose of Data Science

Si repetimos el proceso podemos llegar al resultado que minimiza el error cuadrático medio.

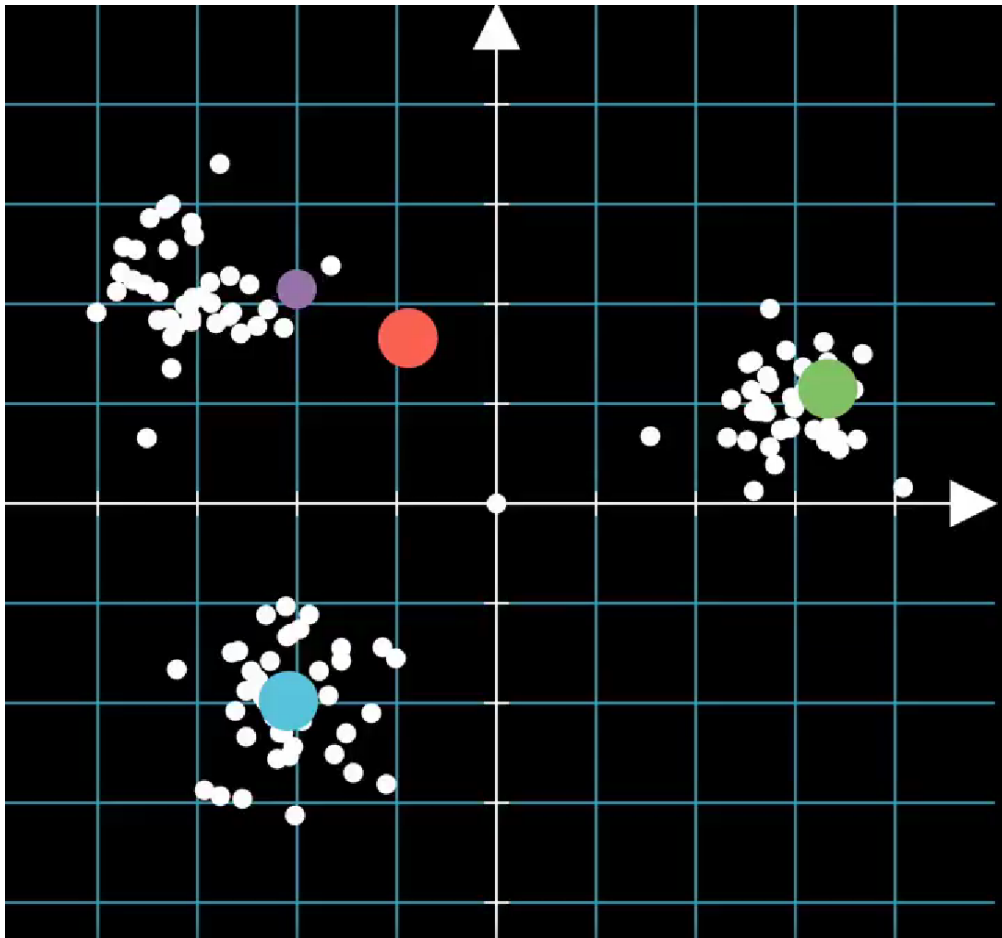


Figura 7: Paso 5: Reposicionar centroides Fuente: Daily Dose of Data Science

Con esto, podemos visualizar un patrón de agrupación que en un análisis preciso, puede darnos información en sentimientos positivos y negativos, por medio de una puntuación del - 5 al 5

3. Objetivos

3.1. Objetivo general

Realizar una nueva forma de análisis para la Comprensión y análisis de tendencias políticas por medio del análisis de texto de la Opinión Pública, utilizando Aprendizaje Automatizado.

3.2. Objetivo específicos

1. Analizar la tendencia política del proceso constituyente para definir un posible voto. Se aprueba o se rechaza la constitución.
2. Establecer el *text analysis*, como una herramienta poderosa para realizar análisis de tendencias políticas
3. Incentivar la creación de nuevos análisis de la opinión pública o mejora de los ya creados.

3.3. Alcances

1. El text mining no predice, solo clasifica, por lo que no se pretende predecir de forma exacta, porcentajes de aprobación o rechazo, sino que tendencias políticas, a partir de diferentes métricas de opinión pública que nos podrían indicar su orientación política.
2. No se pretende hablar de otros casos de uso más que el de opinión pública, debido a la limitación misma que emprende el uso de text mining. Esto es: necesidad de una opinión verídica y escrita con sus propias palabras, lo cual en sí mismo es una fortaleza, debido a que es un análisis más cierto que lo que se pueda inferir y encapsular en una encuesta y una debilidad, en cuanto a metodología de obtención de datos.

4. Caso de estudio

4.1. Historia del Proceso Constitucional chileno

En el marco del estallido social en Chile, ocurrido el pasado 19 de octubre de 2019, sucedió lo que a opinión de muchos fue el "despertar de Chile". Un despertar cargado de rabia, sufrimientos e impotencia, ante los 30 años de injusticias que, según la mirada de algunos, vivió el pueblo chileno, desde que Chile volvió a la democracia después de 17 años de régimen militar, culminados en el año 1990. Es de hecho, el año 2019, cuando empieza el declive de la percepción económica según la ciudadanía. Lo que se ve reflejado en la figura 8.

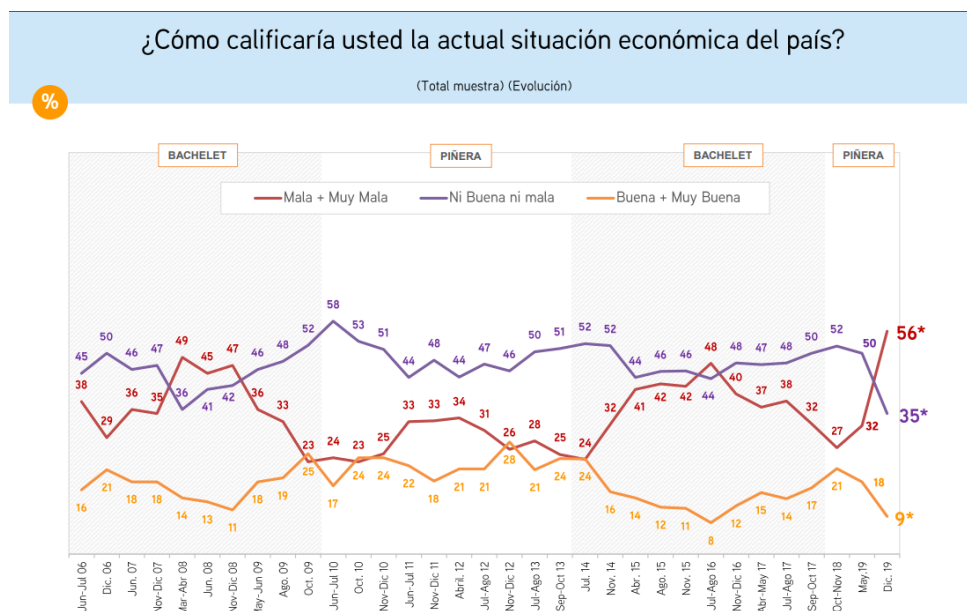


Figura 8: Percepción histórica económica de Chile. Fuente: CEP Chile

Desde las protestas, nació la solución: la creación de una nueva constitución. Una nueva carta magna, desde la cual, el ciudadano pudiese construir un nuevo futuro para Chile

Meses más tarde, Chile desde las masivas marchas, muchas veces pacífica, pero que significó una

revolución, se abrió la posibilidad de eliminar la constitución creada en dictadura, a través de un plebiscito. Este plebiscito aprobado por los chilenos tuvo una amplia mayoría y una participación ciudadana nunca antes vista. Se construía una nueva constitución y no iba a ser realizada por un diputado y/o senador, sino que por los mismos chilenos comunes y corrientes. Personas que supuestamente sabía exactamente el sentir de toda una nación.

No obstante, lejos de toda proyección y desastrosos acuerdos entre los convencionales, se entregó una propuesta a Chile, la cual con la misma fuerza con la que fue aprobada para su redacción, ahora fue rechazada. Los acuerdos políticos, llegaron a la conclusión de que debían construir una nueva constitución esta vez por políticos expertos y a través de un expresión de democracia, ellos redactar la propuesta de constitución. Una segunda propuesta de constitución

Esta segunda propuesta, a su vez tuvo su votación el 17 de diciembre de 2023. Por un lado, estaba la opción de "A favor" quienes votaban por aceptar la nueva propuesta constitucional; y "En contra", para quienes no gustaban de esta nueva propuesta constitucional y que implícitamente aceptaban quedarse con la propuesta actual.

Finalmente, Chile decidió rechazar la segunda propuesta, por lo que aquí se termina la historia del proceso constitucional, al menos, hasta un par de años más.

4.2. Metodología de Investigación

En este caso de estudio, pretendo desde una mirada objetiva predecir cuál sería el resultado del último proceso constitucional, debido a las influencias de todos los procesos culturales en la opinión de la gente sobre el proceso constitucional y por lo tanto, las tendencias políticas, en cuanto a si se aprobaría o no la segunda propuesta de constitución para Chile como un caso hipotético, donde nos situáramos previo a el segundo plebiscito de salida, correspondiente a la segunda propuesta constitucional (17 de diciembre). Actualmente, sabemos que se rechazó esa propuesta, por lo que sabemos que el resultado debería estar orientado hacia allá.

4.2.1. ¿Cómo se realiza?

Para predecir lo anterior, se debe de buscar variables explicativas que predigan en gran parte los resultados aquí obtenidos, sin embargo, se cree que un proceso tan grande como el vivido, es explicado por variados factores tanto políticos, como económicos, sociales y culturales; cuyos efectos individuales son difíciles de cuantificar y no precisan de métricas de aplicación efectivas, factibles y tangibles para mensurar la opinión pública. Es por esto que, que el análisis de estos factores de manera desagregada nos podría conducir a imprecisiones, cuantía de supuestos y sobre todo, resultados insesgados.

Con todo, es importante retomar una discusión previa que realizamos donde nos basaremos en el supuesto de racionalidad del voto, para decir que es el voto una toma de decisión objetiva con la finalidad de considerar maximizar su interés social y el de otros, por medio, de un voto informado y no influenciado por desinformación e influencias externas sobre él. Si así sucediese, entonces hablaríamos de observaciones que generan “ruido” pero que no son relevantes para este análisis. Este supuesto, evita discusiones de cómo las encuestas de opinión no siempre ajusta el voto final en la urna como lo son los precedentes de la Encuesta Cadem sobre los triunfos de algún partido o figura política, o

la discusión socio-psico-política de la veracidad de este.

4.2.2. ¿Cómo se mide opinión pública?

. La opinión pública es actualmente, medida a través de encuestas específicas con preguntas que buscan generar un seguimiento del cambio en las respuestas, como ya vimos anteriormente. Estas preguntas se caracterizan por ser de opción múltiple, y en su mayoría de selección única. Como ventaja, se puede ver que su simplicidad, ayuda en el seguimiento del cambio de la opinión pública así como su inalterabilidad con forme pasan los años, tendencias políticas y cambios sociales. Además, permite evitar ruido, al recibir opiniones escritas que no reflejan necesariamente su pensamiento. Su desventaja es que, como todo tipo de encuesta, no refleja el “verdadero sentir” de la persona en su visión cosmológica e integral, por lo que muchas veces su respuesta se ve restringida a la selección de una respuesta que más represente la respuesta total del encuestado. Sumado a esto una pregunta no es en sí misma, suficiente, sino que deber ser complementada por múltiples preguntas que permitan comprender la totalidad de su opinión, lo cual dependerá de la calidad de pregunta y su intención, la cual puede no funcionar para todos los casos.

Además, este tipo de encuestas podrían tener cierto problema en la recolección de datos, ya sea por cómo se realiza la pregunta al entrevistado, la comprensión de esta pregunta y la coherencia de su respuesta, perdiendo cierto grado de precisión su respuesta. Estos y otro más factores, nos dejan una gran incógnita de cómo podemos obtener en realidad una verdadera opinión de una persona en sus propias palabras, incluyendo sus subjetividades, sesgos de todo tipo, etc. pero que a gran escala representen un fragmento que nos ayuden a realizar todo un prisma de visión sobre qué es lo que los chilenos piensan –realmente– sobre el proceso constitucional.

Es por ello, que decidimos utilizar un escrito que nos permita saber qué es lo que piensa esta persona sobre alguna problemática social. Es allí donde utilizamos las columnas de opinión desde

diarios. Estas columnas se caracterizan por ser realizadas en mayor parte por la misma editorial desde la visión de algún periodista, en el cual se desarrolla una temática, dejando sus críticas al respecto, entablado posibles respuestas y/o soluciones, o encajando nuevas incógnitas sobre la problemática. Es esta una visión editorial de una problemática, una de muchas, las cuales como un todo pueden representar perfectamente la opinión de diferentes sectores políticos, económicos y sociales, sin perder, debido a su característica per sé de los diarios – supuesto, pluralismo –, una mirada más precisa, poco sesgada y representativa de lo que busca la gente.

Es así como, a través de una opinión esbozamos la realidad chilena y su voto en la precisión de sus propias palabras.

4.2.3. ¿Cómo se interpreta, de una forma escalable, las opiniones?

La respuesta es: NLP y Machine Learning. El Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés), nos permitirá con sus herramientas procesar grandes cantidades de textos para trabajarlos de manera más sencilla y con escalabilidad como vimos en la metodología, por medio del *bag-of-words* o los unigramas.

Podemos utilizar el Aprendizaje Automatizado (o Machine Learning, en inglés), donde se procesará estas grandes entradas de datos y aprender de ellas, en el sentido de que sin saber de qué constan los textos, es decir, por medio del aprendizaje no supervisado, realizar clasificación temática de las opiniones sobre la constitución— y/o contexto de la investigación, por sentimientos (positivos o negativos), agrupará en grupos de semejanza lexicosemántica para encontrar relaciones extra-textuales (fuera de lo textual, como el contexto). Todo esto con la finalidad de que podamos utilizar la gran escala de las grandes y complejas opiniones, como un elemento enriquecedor de la investigación, para obtener relaciones que a simple vista no se podrían visualizar y/o que posiblemente requerirían de opiniones expertas para su descubrimiento.

4.3. Procedimiento

Estableciendo previamente el marco metodológico de nuestra investigación, se puede ser capaz de ahondar en los procedimientos que se establecen en el camino y supestos realizados para continuar en el análisis de esta data particular, por encima de lo que se establece en la metodología.

4.3.1. Extracción de la Información

La información la encontramos en el sitio público de <https://espaciopublico.cl/categoria/prensa/>, donde utilizando el método de webscrapping, podemos obtener los diferentes textos de una página web. Es valioso destacar que la fuente de los datos debe ser fidedigna y no sesgada a grupos de interés políticos y/o sociales. Es por eso, que elegimos Espacio Público, debido a que es un organismo independiente que recibe fondos del extranjero, cuya visión es fomentar el estudio público y el desarrollo de nuestras políticas, por lo que es el *website* ideal para obtener nuestra información dura para el análisis.

Desde esta página, retiraremos la información como el título, la fecha de emisión, la temática, el periodico donde fue escrito y el cuerpo (o texto) de la columna. Lo anterior con el objetivo de entablar lo que serán el corpus y las demás variables "docvars".

No obstante, esta extracción no fue suficiente, faltando datos de las diferentes variables como se puede ver en la Tabla 2, debido a que en la mayor parte de ellas, la estructura HTML es diferente. Estas casuísticas las abordaremos en la siguiente sección.

Variables data				
títulos	fechas	medios	tematicas	textos
0	0	17	47	58

Tabla 2: Cantidad de datos faltantes o NA

4.3.2. Carga, limpieza y transformación de datos

Si bien inicialmente extrajimos la información desde la web, el código requirió de algunas modificaciones para la extracciones de algunas casuísticas como comentamos anteriormente. En la mayoría de los casos por diferencias del código HTML. Es valioso comentar, que solo un texto de las 803 observaciones no pudo ser solucionado su problema, el cual fue descartado. Empero es una pérdida mínima. En el caso de las otras variables el problema no es solucionable, puesto que son NA's, donde efectivamente, la página no publica el medio o la temática de la columna de opinión. Los datos faltantes se puede ver en la tabla 3

Variables data				
títulos	fechas	medios	tematicas	textos
0	0	17	47	0

Tabla 3: Cantidad de datos faltantes o NA, después de la carga de datos

En cuánto al proceder de medios y temáticas, no borraremos el registro completo, debido a que puede ser que exista un medio en particular que tenga este modu operandis de no poner el medio o la temática, pero cuyo análisis agregado puede ser interesante. Solo en los análisis –que ya veremos –, donde se segmente por medios o por temáticas, solo allí, borraremos los NA, para no generar ruido.

En segundo lugar, realizamos la transformación de la variable "fechas" a un formato año-més-día, pasando de, por ejemplo, "5 de Octubre de 2019" a "2019-10-05". También se realiza la transformación de la variable "medios" con 2 cambios importantes:

1. Todos los datos venían con "Medio: La Tercera", ""Medio: Diario Financiero", y así sucesivamente, lo que ensucia el dato y por tanto, es retirado, dejando solamente el nombre del medio periodístico.
2. Algunos diarios registraban como diferentes, aunque eran iguales, pero los ponían con mayús-

culas y otros sin, o contracciones como por ejemplo, "Las Últimas Noticias" y "LUN" que son lo mismo. Estandarizamos todos estos registros para que fuesen uno solo.

Por último, creamos una variable más para clasificar las columnas, según el periodo en el que fueron publicados. Esto con la intención de más adelante hacer un análisis intertemporal entre lo que se denomino "periodos" a subgrupos de fechas significativas para las constituciones. Esto es: Periodo previo Plebiscito de entrada (desde 02 de Enero de 2019 a 5 de Octubre de 2020), post-plebiscito (desde la fecha anterior hasta el 4 de Septiembre de 2022); post- rechazo primera constitución (desde la fecha anterior hasta el 17 de Diciembre de 2023) y post-rechazo segunda constitución (desde la fecha anterior y hasta donde se tengan actualizados)

4.3.3. NLP

Se realiza procedimiento de corpus, donde se asigna como *input* todos los textos de la variable "texto" y se nombran como docvars a las demás variables.

Luego, se crean *tokens* con las palabras, creando unigramas, y además realizamos limpieza de puntuación, símbolos, números y conectores, tanto simples (ejemplo: "pero", "además", etc) como compuestos (ejemplo: "sin embargo", "no obstante", etc.). Esto permitirá para el análisis de frecuencias evitar caer en repeticiones de este tipos de palabras como muy frecuentes pero que no tienen relevancia alguna.

Finalmente, utilizamos la herramienta del DFM para establecer un espacio de frecuencia, agrupado por texto, de las diferentes palabras. Es, a partir de aquí que empiezan a salir a la luz los primeros resultados.

Las nubes de palabras, fueron trabajados con un máximo de palabras de 50, aunque dependiendo de la relevancia y también de la resolución del gráfico, fueron agregadas más palabras -o quitado menos palabras-. Además, es muy importante comentar que los resultados no son "puros", puestos

que fueron pasados por un filtro humano, donde se quitaban palabras irrelevantes, asociados a sustantivos, adverbios, adjetivos, etc.

En cuánto al gráfico *Keyness* se realizó el mismo filtro, donde en cada resultado se fue iterando la eliminación de palabras hasta que se logran una consistencia de la significancia de estas palabras.

En cuánto a las tablas de Palabras en Contexto, es importante recalcar que no se realizan desde el dfm, debido a que el contexto necesita de todas las palabras incluyendo conectores, puntuación, etc. Aquí, solo se utiliza el corpus como *input*.

Es importante destacar que, a lo largo de nuestro estudio realizamos una segmentación o filtrado por temática constitucional que si bien, es funcional, no es del todo puro, puesto que la temática es llamada "Corrupción y Democracia", pudiendo caer como significativo, palabras en un contexto aparte de las propuestas constitucionales. Con todo, estos son mínimos para el periodo establecido.

4.3.4. Modelamiento de Tópicos

Para el modelamiento de tópicos se realiza el análisis de k óptimos por perplexity, debido a que, ya que se basa en maximizar la función de verosimilitud, es esta función la que encuentra el mejor parámetro posible, es decir, insesgado y consistente, por lo que, se cree tiene mejor rendimiento que otras medidas.

Para esto se utiliza un proceso de conversión del dfm a dtm, el cuál tiene una estructura más simple de frecuencia de las palabras por sobre el dfm, lo cual permite el calculo de perplexity.

Luego, una vez elegido el k , en mi caso es $k = 4$, puesto que es el k con el que existe una gran disminución eficiente del error cuadrático medio, dado el trade-off sesgo-varianza. Sin embargo, no es el más eficiente que en este caso sería 5 (fig. 9). Esto porque si bien con 5 es óptimo también se realiza un análisis de relevancia de estos y pareciese que el más congruente es con $k = 4$

Teniendo el mejor k , podemos utilizamos la función LDA, que es la que mejor se comporta debido

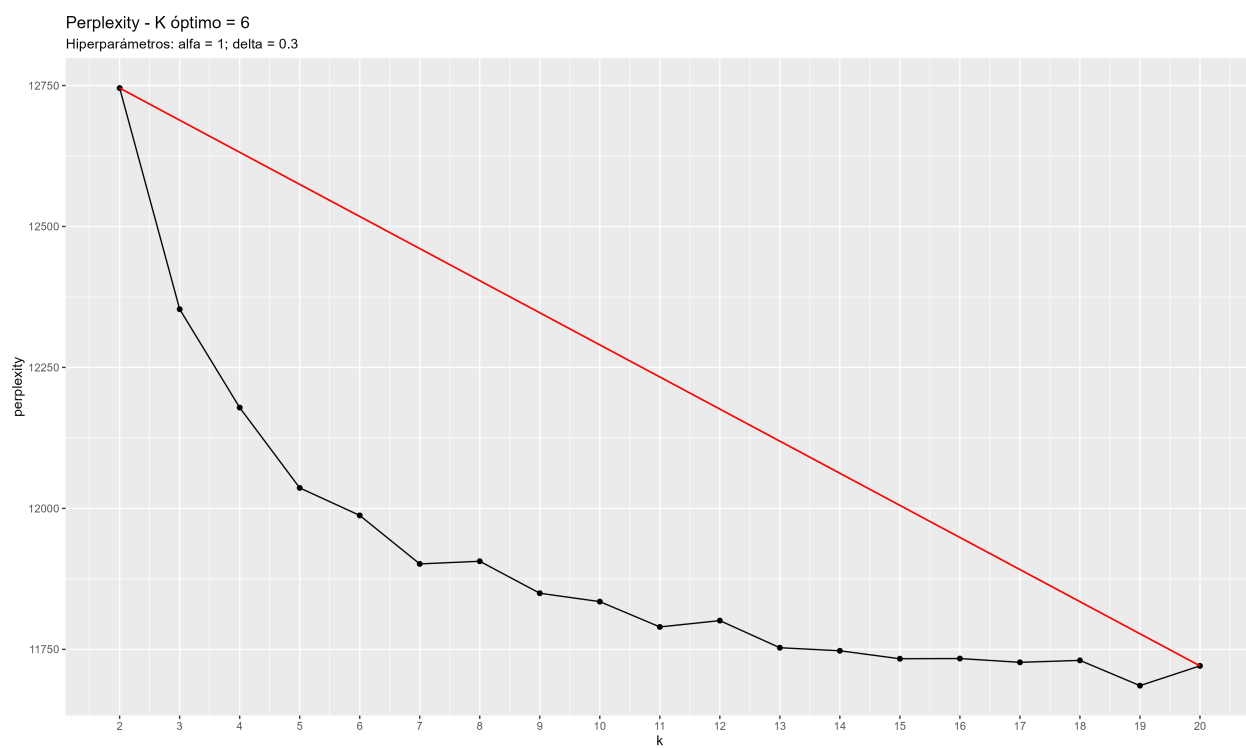


Figura 9: Gráfico de cálculo de perplexidad. Fuente: Elaboración propia

a los supuestos, en este caso, la independencia de tópicos es relevante, pues permite que las palabras entre tópicos no se repitan y tengan una significancia por si sola y no una significancia conjunta. Esta significancia conjunta, es preferida cuando buscamos modelos más complejos y flexibles que predicen mejor, pero cuyas desventaja es la inferencia de estos tópicos. Justamente, acá buscamos la inferencia de estos tópicos para el análisis de los periodos y/o medios.

Una vez aclarado lo anterior, partimos calculando como ya sabemos la distribución Beta de las palabras, asignando la probabilidad *posterior* de que la palabra pertenezca a determinado tópico. Esto permitirá la asignación de palabras. Luego simplemente para simpleza tomamos el top 10 de palabras de cada tópico y vemos su probabilidad de que este bien clasificada esta palabra dentro del tópico.

Finalmente, se hace un procedimiento inverso, que entregue una verdadera conclusión, puesto que si bien la clasificación de las palabras en tópicos es relevante, solo nos indican un orden. Este orden tiene relevancia, cuando se evalua cuales son los tópico preeminentes en los periodos constitucionales o en la opinión de cada medio periodístico. Para ello, se utiliza la matriz gamma que nos indica la prevalencia de tópicos en los documentos, dada la función de distribución gamma. De forma análoga a Beta, ordenamos de mayor a menor, tomando la muestra de los 3 tópicos más relevantes.

4.3.5. *Clustering*

La clusterización se realiza solo sobre los medios, debido a que como comentamos, la clusterización se realiza sobre datos, cuya respuesta no es esperada. Busca asociación entre textos, o conjuntos de textos, porque es un modelo no supervisado. No se puede realizar sobre los periodos, porque son los periodos clasificaciones previas, por lo que ya no sería no-supervisado. Además, no hace sentido realizar clusterización entre cuatro elementos, ni tampoco para buscar alguna relación como tópicos inter-periodos, y que no sabríamos interpretar.

Acá se realiza un proceso clave donde se calcula la matriz distancia de las palabras de los documentos, aprovechandose de que el dfm, nos da en sí mismo , la métrica de *sparsity*, la cual permite saber si 2 textos se parecen, debido a la frecuencia que utilizan las mismas palabras. No obstante, para realizar esta comparación en términos de frecuencia, se debe realizar desde una frecuencia relativa y no absoluta, para asersoriarnos de que se utilizan en la proporción las mismas palabras. Es necesario aclarar que es esta una proporción local, es decir, el uso de la palabra en un documento dividido en el total de frecuencias de palabras de ese documento individual.

Teniendo la distancia, se obtiene una matriz, donde se compara 1 a 1 los documentos de los diferentes medios individuales, siendo la diagonal la comparación de distancia de un medio x, consigo mismo, lo cual da 0. Los otros elementos, nos indica en un número entre 0 y 1 su lejanía semántica. Si es 0, son dos textos idénticos; y si es 1, son 2 textos toalmente diferentes. Tan diferentes que uno ocupa ni una palabra que el otro texto haya utilizado.

Gráficamente, se puede ver en la figura 10

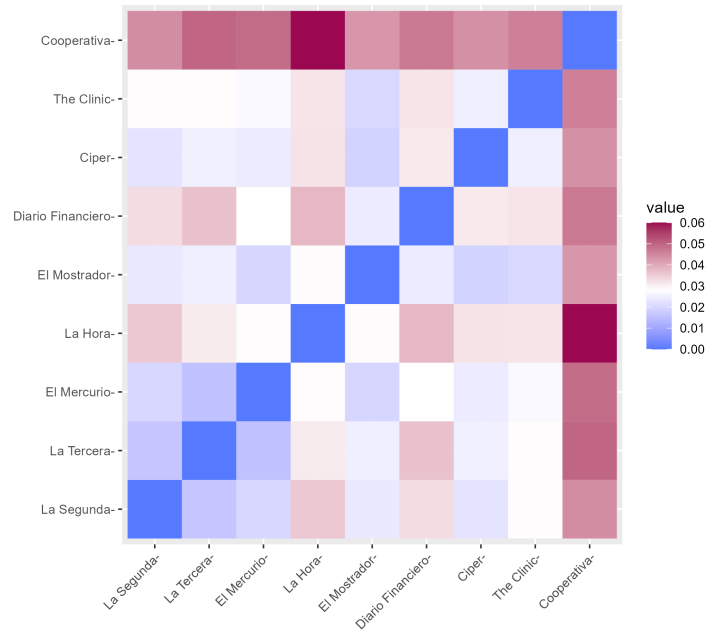


Figura 10: Mapa de calor distancias semánticas. Elaboración propia

Una vez tenidas estas distancias entre documentos, se procede a clasificar acorde, utilizando, además el parametro de $k = 6$ como óptimo de agrupaciones.

Además, como análisis extra se realiza un dendograma para la comparación de estos cluster, ver-sus este nuevo que permite, no establecer un k manual en el análisis, como discutimos anteriormente. Este dendograma utiliza la misma matriz distancia que el cluster.

4.3.6. Análisis cuantitativo de Sentimientos

Para el análisis de sentimientos, se utiliza un diccionario de emociones en específico llamado "Afinn". Este tiene como característica en particular, no solo clasificar una palabra como "positiva" o "negativa", sino también que tan negativa o positiva puede ser, dependiendo de la naturaleza misma de la palabra, definidas por el marco de trabajo de este centro. Las palabras van desde el -5 siendo

este puntaje una palabras muy negativa, 0 una palabra neutra y 5 una palabra muy positiva. Este tipo de diccionario nos permite, definir dentro del espectro del sentimiento, su intensidad, abarcando análisis cuantitativos que permitan definir sesgos y/o tendencias políticas sobre las constituciones en los periodos constitucionales o la opinión "positiva" o "negativa" que tienen del proceso.

Para ello, a partir de los *tokens*, es decir las palabras sueltas, traducirlas a la clasificación "positiva" o "negativa" y también asignarle puntos.

A partir de allí, a través de diferentes análisis se modifica la información para entregar resultados significativos sobre estos sentimientos del proceso constitucional.

Vale la pena destacar que para que este proceso sea lo más fidedigno posible, se decidió no quitar palabras de ella, para que la clasificación sea más pura, sin intervenciones de por medio.

4.4. Resultados

Utilizando la metodología del caso descrita anteriormente, partimos desde la concepción del NLP. Cuando realizamos el DFM, tenemos una matriz que describe la frecuencia de palabras que aparecen en términos únicos a lo largo de diferentes textos. A partir de aquí podemos realizar un análisis de precisamente estas palabras más frecuentes en los 802 textos a analizar como un todo (corpus).

4.4.1. Nube de Palabras

En primera instancia, podemos ver gráficamente en la figura 11 palabras, a través de una nube la cual muestra las palabras más utilizadas dentro de las columnas de opinión, que como comentamos en el procedimiento, fueron filtradas con un máximo de 50 palabras y se retiraron palabras poco significativas.



Figura 11: Nube de Palabras de todas las columnas de opinión analizadas. Elaboración propia

Las palabras más frecuentes son las más grandes. A medida que se alejan hacia los bordes y pierden tamaño, son palabras menos frecuentes. No obstante, es importante destacar que las palabras menos frecuentes, no son en sí mismas palabras con baja frecuencia. Es tan solo que, en comparación con las demás, no tienen una frecuencia tan alta, sin embargo, siguen siendo palabras recurrentes y significativas, utilizadas dentro de los textos.

Podemos darnos cuenta que las palabras más frecuentes de todos los textos hablan de "reforma", "gobierno", "política", "mujeres", "convención", "crisis", "riesgo", etc. Preeliminarmente estas palabras no concluyen demasiado si consideramos que es un análisis agregado y considera todas las columnas de opinión, en donde podría hablarse de reforma, pensando en la reforma de la constitución, cuando en realidad es la reforma de las pensiones o carbineros. Sin contexto es difícil saber, por lo que, en pos de lo anterior, se realizará dos tipos de análisis: un análisis solo constitucional, que permite saber por las palabras más utilizadas de qué se habla en este periodo sobre la temática, y otro análisis que permita la comprensión de otras opiniones relacionadas a otros ámbitos que sin duda juegan un rol fundamental en la decisión del votante.

Analizando por constituciones exclusivamente, se tiene en la figura 12, nuevas palabras como "violencia", "boric", "vida", entre otras, y que palabras como "salud" y "trabajo", pierden tamaño relativo, a otras palabras que aparecen más. Esto nos dice que por lo menos, se habla menos en las constituciones comparado con el análisis conjunto opinión y constituciones. También, desaparecen palabras como "educación" y "mujeres". Todo esto, por supuesto, en las proporciones de frecuencias de palabras, por tanto, no es que se hable –o se deje de hablar– de algún tópico, es solo que se habla más o menos, en comparación con el resultado anterior ahora que solo se analizan textos exclusivos de constituciones.



Figura 12: Nube de Palabras de columnas de opinión exclusivas de constituciones

Si realizamos un análisis más segmentado, podríamos buscar palabras frecuentes, de textos específicos de nuestro interés como lo son los periodos constitucionales que buscamos analizar. En este caso los 3 periodos constitucionales serán: Pre-Apruebo Plebiscito de Entrada, Post-Apruebo Plebiscito de Entrad (, o Pre-Rechazo Primera Propuesta Constitucional) y Post-Rechazo de Primera Propuesta Constitucional. Lo anterior se ve reflejado en las figura 13, por medio de una nube de palabras, que ahora segmenta por periodo las palabras más frecuentes. La interpretación del tamaño de las palabras se mantiene.

¿Cuánta confianza tiene Ud. en las FF.AA., Carabineros y la PDI?

(Total muestra) (Evolución) (% Mucha + Bastante confianza)

%

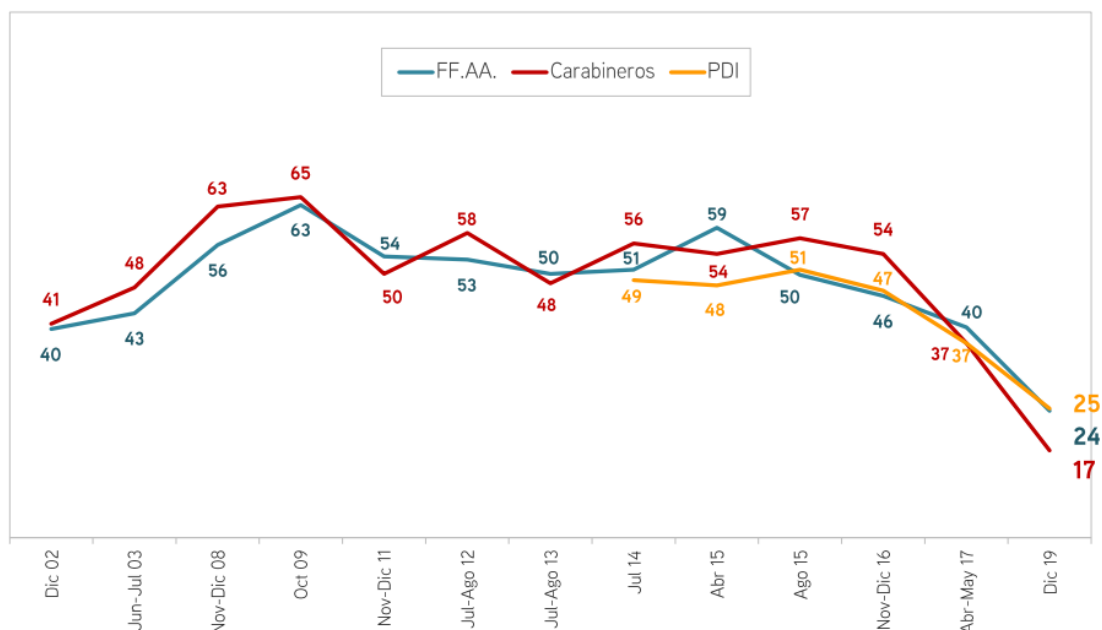
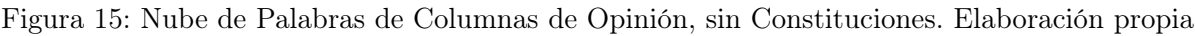


Figura 14: Confianza ciudadana en las Fuerzas Armadas y de orden. Fuente: CEP Chile

Si realizamos el mismo análisis para las opiniones – definiendo opinión, como todo lo que no es propio del proceso constitucional –, podemos ver que las figuras 15 y 16 nos muestran nuevas palabras como "isapres", "crecimiento", "ambiental", "precios", etc. que nos hablan de las problemáticas sociales que en ese periodo se estuvo muy en boga como lo fue la posible eliminación de las isapres y el movimiento "No más AFP" o la preocupación del costo de un nuevo proceso constitucional sumado al ya realizado para la convención constitucional y la redacción de una propuesta que fue rechazada.

Cuando se habla de constitución en el periodo pre-rechazo de constitución, pareciera centrarse más en la discusión política después del triunfo del apruebo.



46

Post-Rechazo Primera Propuesta



Pre-Rechazo Primera Propuesta

Figura 16: Nube de Palabras de Columnas de Opinión, sin Constituciones por periodo. Elaboración propia



Figura 17: Problemáticas sociales prioritarias en el tiempo. Fuente: CEP Chile

Incluso podemos establecer una relación entre palabras por periodos para saber la recurrencia que tienen entre sí las palabras para referirse. En la figura 18, nos muestra un grafo o *network*, donde nos relaciona por medio de estas líneas las palabras relacionadas entre sí para el top 20 de palabras utilizadas en columnas de opinión exclusivas de constitución. A mayor grosor de estas, la relación es más recurrente y por tanto, más estrecha. Por ejemplo, si se habla de "reforma", ésta tienen una relación estrecha con "carabineros", debido a la crisis institucional sufrida por esta última y la demanda social de reformala.

Red de palabras en el periodo Pre-Apruebo Plebiscito de Entrada, solo Constituciones

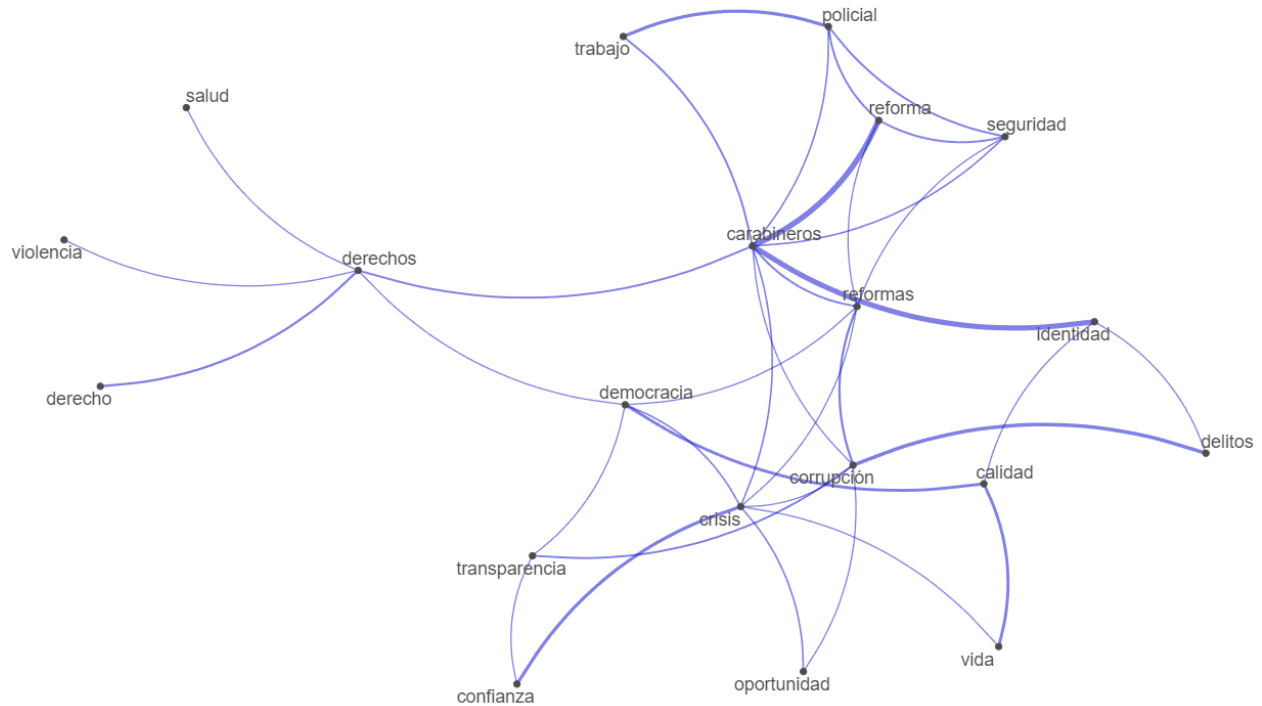


Figura 18: Network de Top 20 palabras Periodo Pre-Apruebo Plebiscito de Entrada. Solo constitución. Elaboración propia

En cuánto, al mismo periodo, pero filtrando por solo opiniones, nos encontramos una estrecha relación entre "crisis" y "climática", así como "calidad" y "educación". Con todo se debe ser claro que no es que cuando se hable de "calidad" necesariamente existe una única relación con "educación", puede que tenga más, sin embargo, solo estamos desarallando la rede de palabras para el top 20 palabras.

Red de palabras en el periodo Pre-Apruebo Plebiscito de Entrada, sin Constituciones

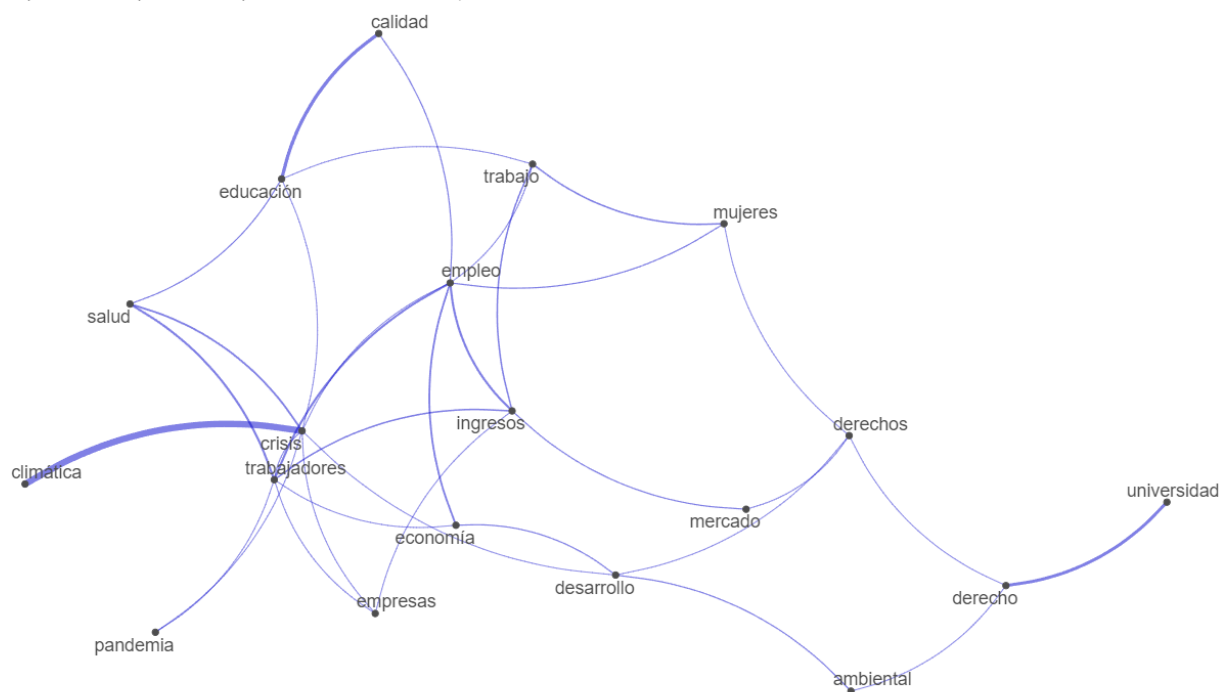


Figura 19: Network de Top 20 palabras Periodo Pre-Apruebo Plebiscito de Entrada. Solo opinión. Elaboración propia

Análogamente, el resto de los periodos. Tenemos que en el periodo post-apruebo, cuando se habla de constituciones, en la figura 20 una estrecha relación entre "convención" y "trabajo" y en general "convención" como nodo de muchas relaciones tales como "confianza", "oportunidad" y "derechos".

Red de palabras en el periodo Post-Apruebo Plebiscito de Entrada, solo Constituciones

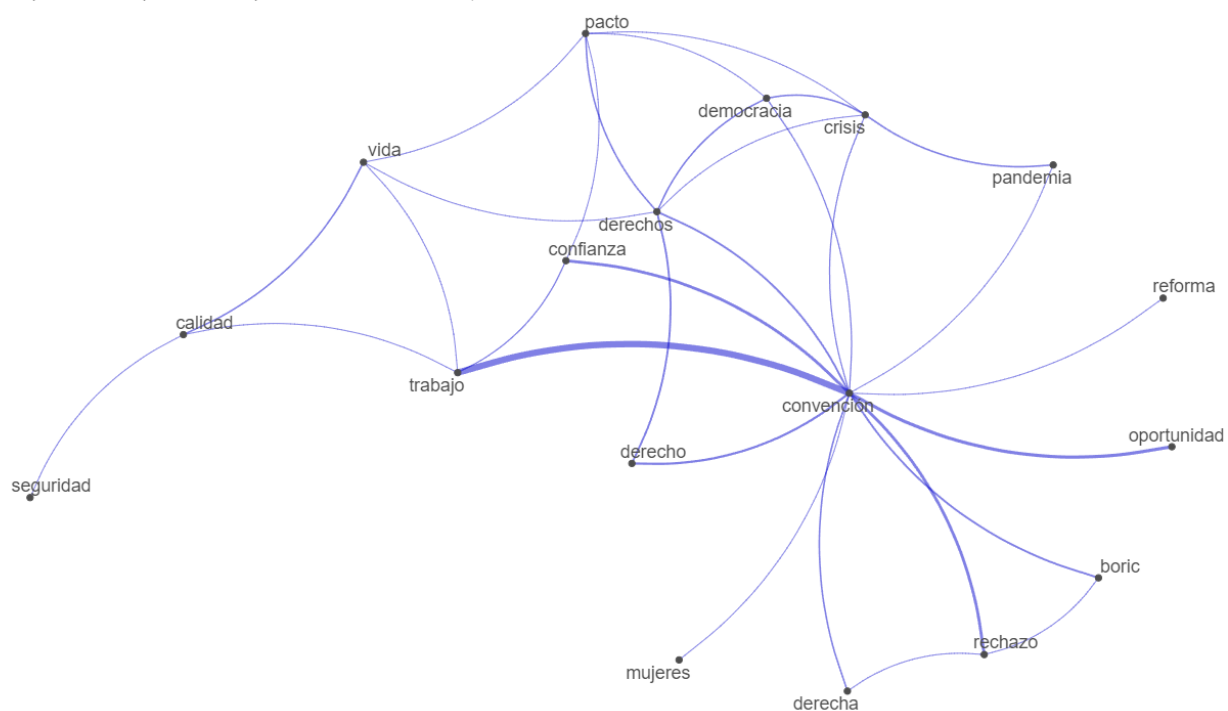


Figura 20: Network de Top 20 palabras Periodo Post-Apruebo Plebiscito de Entrada. Solo constitución. Elaboración propia

Para el caso donde solo se analiza opiniones, también se habla bastante de "trabajo" y "convención", así como de "mujeres" y "laboral" o "empleo" y "trabajadores"

Red de palabras en el periodo Post-Apruebo Plebiscito de Entrada, sin Constituciones

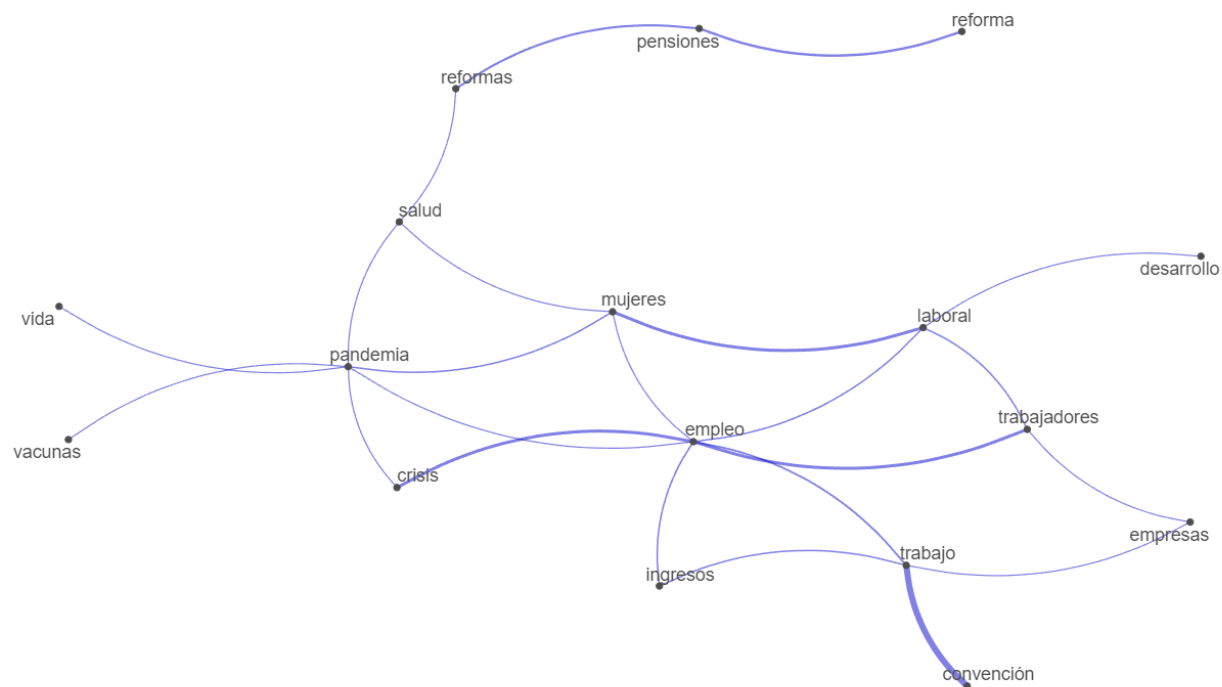


Figura 21: Network de Top 20 palabras Periodo Post-Apruebo Plebiscito de Entrada. Solo opinión. Elaboración propia

Para el periodo post-Rechazo, solo constituciones, las palabras "penal" y "persecución" son estrechas.

Red de palabras en el periodo Post-Rechazo Primera Propuesta Constitucional, solo Constituciones

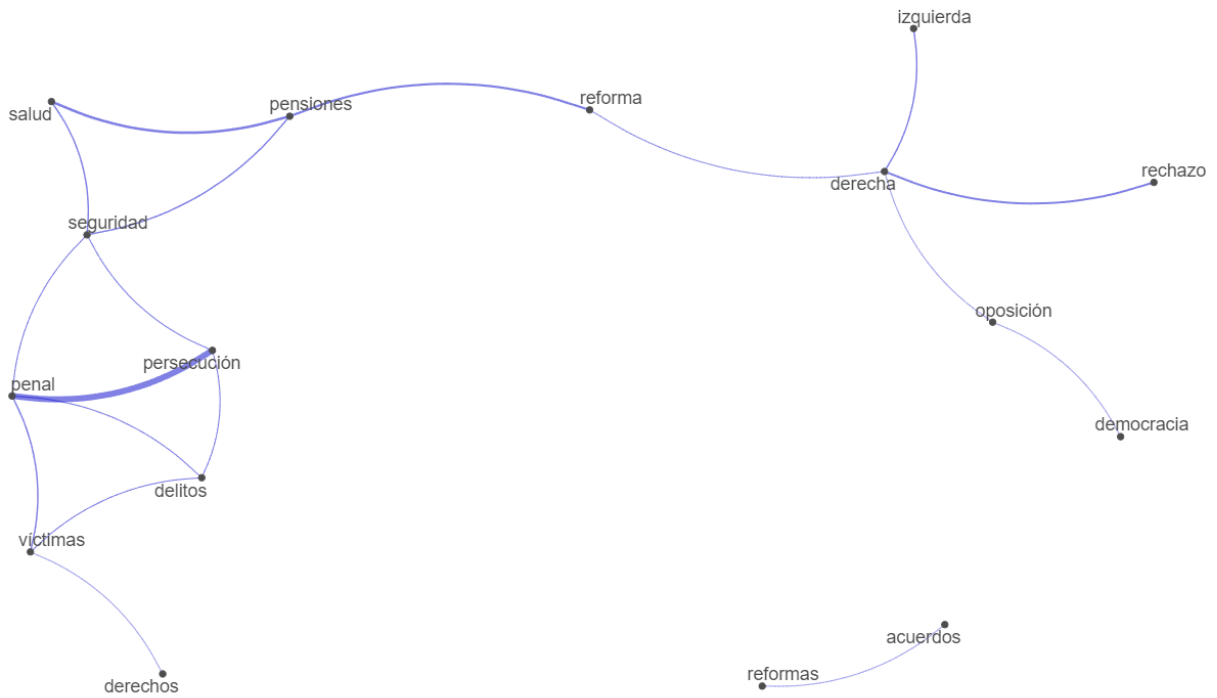


Figura 22: Network de Top 20 palabras Periodo Post-Rechazo Primera Propuesta. Solo constitución. Elaboración propia

Para solo opiniones, tenemos que se suele hablar mucho de "reforma" y "pensiones" de forma recurrente.

Red de palabras en el periodo Post-Rechazo Primera Propuesta Constitucional, sin Constituciones

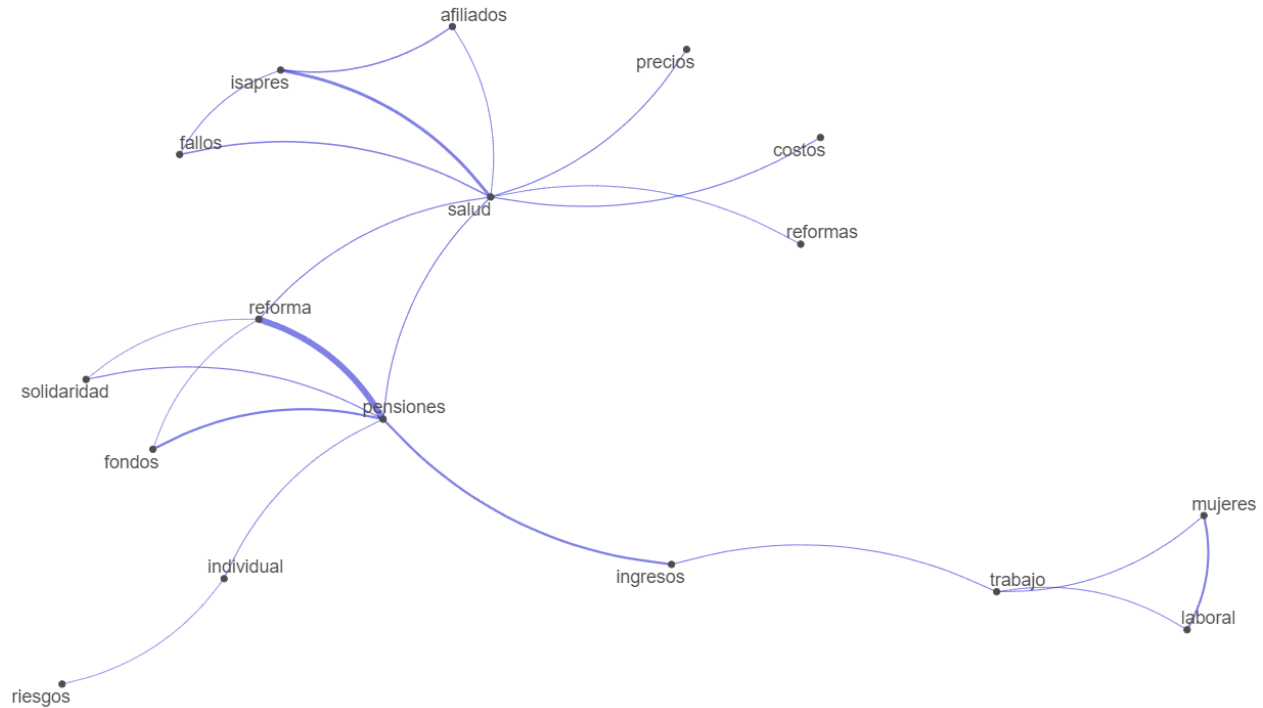


Figura 23: Network de Top 20 palabras Periodo Post-Rechazo Primera Propuesta. Solo opinión. Elaboración propia

4.4.2. Términos Relevantes (*Keyness*)

En paralelo con lo anterior, podemos realizar un análisis no solo de la relevancia en términos de frecuencia como lo apreciamos en la nube de palabras, sino que también, a partir de la relevancia de términos que son característicos de unos textos, dado que en comparativo con otro, este mismo término es mencionado en baja frecuencia. Desde lo anterior, podemos establecer una especie de "relación de autenticidad", si comparamos dos textos, por lo que tiene uno y no tiene el otro, por medio de probabilidades.

En el trasfondo particular del periodo constitucional, podemos justamente complementar lo que mencionábamos anteriormente, y es que por más que existan problemáticas sociales permanentes, estas problemáticas de una forma periódica van tomando mayor o menor relevancia que se visualizan

en, nuestro caso, por medio de los diferentes periodos constitucionales establecidos.

Si realizamos el análisis para el periodo Pre-Apruebo y Post-Apruebo Plebiscito de Entrada, podemos ver en la figura 24, considerando todas las columnas de opinión, que en el periodo pre-apruebo podemos hablar, con propiedad, que se caracterizó, por las crisis de carabineros, los incidentes de vandalismo contra el Metro de Santiago y los movimientos que afectaron al comercio local –pyme en su mayoría– cercano a los lugares de manifestación –principalmente, Plaza Baquedano– .

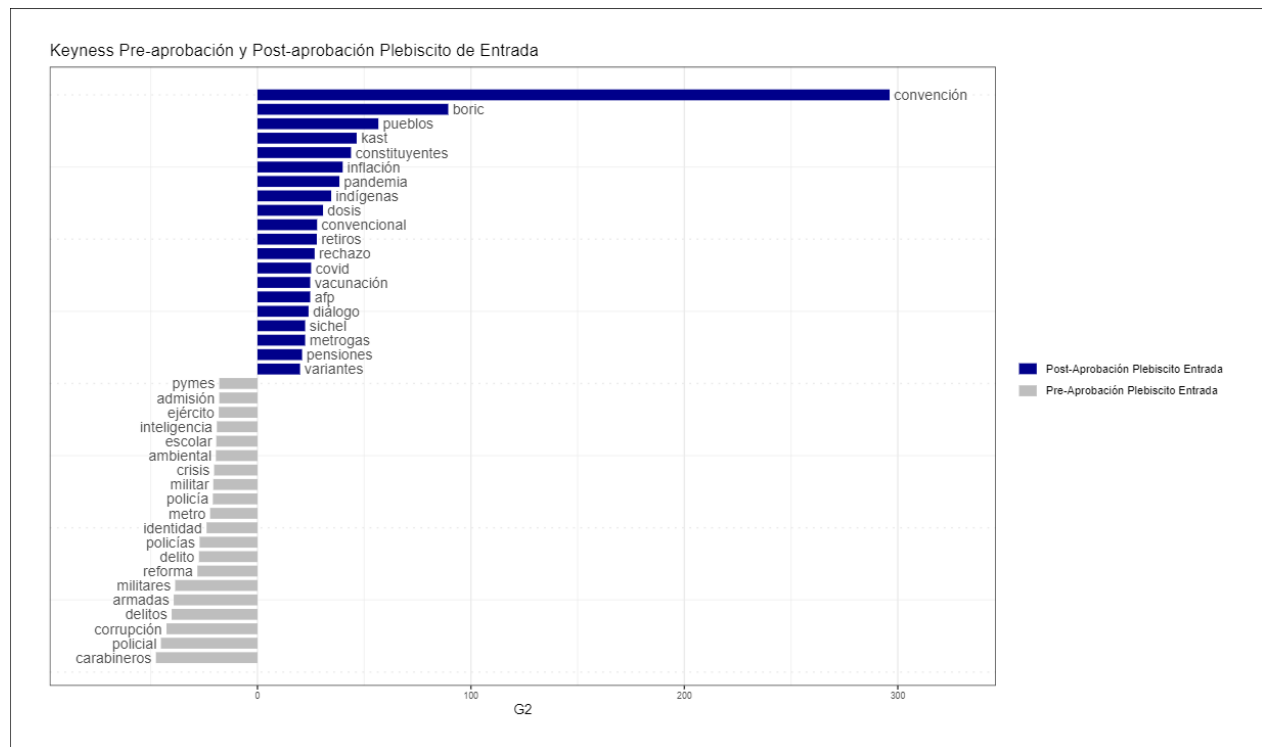


Figura 24: Palabras claves del periodo Post-Apruebo (*target*), en comparación con el periodo Pre-Apruebo. Elaboración propia

Sin embargo, es importante destacar, que realizar estas afirmaciones son un tanto riesgosas, sino tenemos el sustento de su contexto, para verificar si son efectivamente estos, los casos de usos de estos términos analizados. Para ello, nos apoyaremos de las Palabras en Contexto (Keywords in Context, o KWIC), una herramienta que permitirá analizar los contextos en los que se mencionan

estas palabras para evitar mal entendidos y dar mayor sustento a los resultados encontrados por el

Keyness

La herramienta, precisamente, toma la palabra tokenizada y la individualiza en diferentes uso de la misma, dentro del corpus, pudiendo nosotros otorgarle la cantidad de términos de contexto, previo y post a su uso. En nuestro caso, con intención de no ser extensivos, solo se hizo selección de algunas palabras del *Keyness* para dar contexto al periodo y una selección de muestra aleatoria de 5 casos de uso por término, tal como la figura 25

KWIC carabineros		
Contexto Previo	Término	Contexto Posterior
7 puede revisarse con más detalle en el análisis que hice sobre uno de ellos en mi columna de mayo pasado.El retroceso en el discurso gubernamental en materia de reforma a	Carabineros	y el gobierno de la seguridad pública es evidente y preocupante , pero lo que más sorprende es que la autoridad eche pie atrás a lo que se comprometió hace
53 sociales , y también en la casi exclusiva reacción acotada con obstinación al ejercicio de la fuerza . Según datos de la misma encuesta , el 45 % considera que	Carabineros	ha cometido excesos en muchos casos en los que no era necesario , y un 21 % que Carabineros ha usado la fuerza incluso en algunas ocasiones donde no era
27 de los delitos respecto a la situación actual se requiere un fortalecimiento significativo de sus recursos y capacidades . En la contracara , habría que tomar decisiones sobre estructuras de	Carabineros	especializadas existentes en la actualidad en materias de investigación criminal ⁴ . Un tema en particular son los laboratorios de ambas policías . En una definición de especialización real no se
96 sido solo la punta del iceberg . A fines del año 2017 se conoció públicamente la « Operación Huracán » , episodio en el que una unidad de Inteligencia de	Carabineros	de la Araucanía incurrió en la fabricación e implantación masiva de pruebas falsas.Tras una supuesta interceptación de comunicaciones electrónicas encriptadas (en sistemas de mensajería como Telegram y WhatsApp)
38 , al trabajo de Carabineros en el manejo de las protestas y movilización social.Diversos informes emitidos por organizaciones de derechos humanos han entregado una visión muy crítica del trabajo de	Carabineros	dando de cuentas de violaciones graves a derechos como , por ejemplo , la integridad física , la libertad y el derecho a manifestarse libremente en forma pacífica . Más

Figura 25: Palabras en Contexto - Término "Carabineros". Elaboración propia

En la figura 25, se puede apreciar que el término es usado por algo negativo, tanto para hablar de denuncias a actos que atentaban contra derechos humanos, o incluso operaciones, donde se acusa fabricación e implementación de pruebas; como su reforma, por los motivos anteriores.

KWIC corrupción			
	Contexto Previo	Término	Contexto Posterior
86	por qué este caso será uno de los más recordados de los últimos tiempos.Odebrecht es uno de los escándalos de	corrupción	de mayor magnitud y complejidad que recuerde la región . En términos de magnitud , según consigna el repositorio de
139	muchas caras . Se anunciaban « Tiempos mejores » y en algunos aspectos relacionados con la legislación para combatir la	corrupción	, el eslogan puede ser cierto . Luego de varios años en discusión e innumerables llamados de la sociedad civil
23	Transparencia Interhacional y que este año evaluó a 180 países basado en una combinación de encuestas y evaluaciones de la	corrupción	. Las noticias para la región siguen siendo malas con un puntaje promedio de 43 (en una escala de
67	tipo de hechos den información precisa , verídica y comprobable que contribuya a la investigación . En el caso de	corrupción	, se puede reducir la pena en hasta dos grados.La delación compensada a veces es criticada porque puede producir una
75	por un plazo determinado mientras ésta se investiga.La decisión de nuestro máximo tribunal es una buena noticia en materia de	corrupción	, ya que no es posible que no protejamos a aquellos que están dispuestos a denunciar este tipo de hechos

Figura 26: Palabras en Contexto - Término "Corrupción". Elaboración propia

En cuanto a la figura 26, se habla en su mayoría de demandas en cuanto a los aumentos de corrupción y de implementación de medidas preventivas para su disminución.

KWIC metro			
	Contexto Previo	Término	Contexto Posterior
21	el servicio a Maipú-Melipilla y la extensión del Metro de Valparaíso hacia La Calera.Además , debiera corregir la « Ley de Espejo » , empujando grandes proyectos regionales como el	Metro	a Concepción o un tren entre Puerto Montt y Puerto Varas . Por último , la ministra deberá enfrentarse a los ruidosos taxistas para regular de una vez por todas
15	del Transantiago como un hecho aislado , ya que viene a complementar varios cambios que se han implementado en los últimos años . El más relevante , fue situar a	Metro	como el soporte del sistema de transporte público , impulsando una ambiciosa extensión de líneas.Esta era la idea original , pero fue distorsionada por los técnicos que diseñaron el Transantiago
9	recientemente aumentado su tarifa . Una acción política de protesta , sin embargo el el gobierno lo interpretó como un problema de orden y seguridad , llenando las estaciones de	metro	con policía . Ello aumentó los niveles de tensión y conflicto , los cuales escalaron a tal nivel que el gobierno decidió cerrar toda la red de metro el viernes
20	destacado ingeniero Franco Basso . Un segundo objetivo es avanzar con los trenes , priorizando el puente ferroviario sobre el Biobío , el servicio a Maipú-Melipilla y la extensión del	Metro	de Valparaíso hacia La Calera.Además , debiera corregir la « Ley de Espejo » , empujando grandes proyectos regionales como el Metro a Concepción o un tren entre Puerto Montt
16	de transporte público , impulsando una ambiciosa extensión de líneas.Esta era la idea original , pero fue distorsionada por los técnicos que diseñaron el Transantiago final y que pretendían reemplazar	Metro	por 300 kilómetros de corredores segregados de buses de alto costo y poco beneficio . Además se eliminaron millones de transbordos inútiles , se modificaron cientos de recorridos para aumentar

Figura 27: Palabras en Contexto - Término "Metro". Elaboración propia

Acá en la muestra, lejos de toda proyección, su uso está más orientado, a las críticas del sistema de transporte en general, en lo que se menciona a TranSantiago como un proyecto mal ejecutado y posibles extensiones de Líneas. En tan solo una observación se habla de lo sucedido el 19 de Octubre con el comienzo de las evasiones masivas y el Estallido Social.

KWIC pymes		
	Contexto Previo	Término Contexto Posterior
3	autoridad como para los trabajadores . Establecerlo requiere la acción colectiva y la colaboración de la gran empresa , la extensión a proveedores y , finalmente , hacia todas las	pymes . Los montos envueltos no son pequeños , pero son mucho menores que el costo de la pérdida de producto y empleo que nos aguarda si no ponemos a la
9	la reducida fracción de pymes que están en el régimen semiintegrado . ¿ Son sociedades de inversión ? ¿ Basta con ampliar el acceso al régimen que existe para las	pymes para resolver su supuesta carga desmedida ? Asimismo , no sabemos quiénes son los contribuyentes en el tramo exento que pagan impuestos por no recibir crédito completo por las utilidades
4	considera transferencias importantes para todos los trabajadores , no solo para quienes tienen acceso al subsidio de cesantía , y que incluye un programa importante para proveer crédito a las	pymes . Dejando , además , abierta la posibilidad de ir ajustando estos programas e incorporar nuevas iniciativas de acuerdo con cómo evolucione la crisis.El caso del ministro de Salud ,
6	de poner en riesgo el propio sistema financiero.Apoyo a empresas El Gobierno contempla la postergación del pago de impuestos y el adelanto de devoluciones , medidas destinadas principalmente a las	pymes . A ellas se deben agregar apoyos adicionales a través de la provisión de garantías de crédito , profundizando su extensión . Por ejemplo , pueden simplificarse los requisitos y
12	importante de pymes que tributan en el sistema semiintegrado . Pero no se necesita reintegrar a todos para lograr este alivio : basta con que se les permita a estas	pymes pasar al régimen especial que hoy existe para ellas , uno en que tributan sobre la base de su flujo de caja y que ya está integrado . Las pymes

Figura 28: Palabras en Contexto - Término "Pymes". Elaboración propia

El contexto de las "pymes", está envuelto en el periodo de la pandemia, donde debido al confinamiento de las personas en sus casas por el Covid-19. El sector más afectado fueron las Pymes quienes no podían pagar un sueldo, si no podían abrir, y así mismo sus trabajadores no podían en un periodo difícil no recibir sus salarios. Las demandas eran la flexibilización del régimen tributario, la opción a crédito con aval del Estado para su supervivencia y el subsidio del sueldo para trabajadores, cuya fuente de empleo no podían pagarle o debían de despedirle.

KWIC reforma		
	Contexto Previo	Término Contexto Posterior
89	, en la reforma constitucional de 1970 , en virtud de la cual se requería la autorización del Senado solo si el viaje era superior a quince días . Esa	reforma , sin embargo , dejó en evidencia la forma y modo en el cual la tradición constitucional chilena entendía la función presidencial como una actividad continua y permanente . Algunos
20	reflexión distinta a la que ha surgido a partir de la segunda mitad del año 2019 y que ha llevado a un consenso importante sobre la necesidad de realizar una	reforma profunda o estructural a Carabineros y no solo un proceso de modernización ¹ . Como ya señalaba , el proyecto es más bien tributario de esta última visión (es de
25	a Carabineros , y atendiendo a los comentarios previos , no me parece razonable que se avance en este proyecto sin antes decidir cuáles son los pasos prioritarios para la	reforma y cómo este proyecto debiera coordinarse con ella . No solamente es un problema de temporalidad , sino centralmente de contenidos . Avanzar en este proyecto sin que contenga lo
78	deberíamos analizar es la manera de proteger los derechos de los trabajadores , permitiendo a la vez la innovación de las empresas . Un asunto que , combinado con la	reforma del sistema de pensiones , de la salud y la propia reforma tributaria , forma parte de los cuatro ejes estructurales que definirán nuestra sociedad para los próximos años.Así las
152	el patronazgo de la política sobre la administración pública , garantizar la inversión adecuada de los fondos públicos , fiscalizar que los funcionarios cumplan la ley y , tras la	reforma constitucional de 1943 , controlar directamente el presidencialismo . La derecha no debe olvidar que en el pasado no sólo hizo uso y abuso sistemático de la Contraloría denunciando al

Figura 29: Palabras en Contexto - Término "Reforma". Elaboración propia

Finalmente, en la figura 29, se habla de la reforma en tres contextos: reforma del sistema de pensiones, reforma institucional de carabineros, y finalmente reforma constitucional. Todo esto con un trasfondo, donde incluso se quería eliminar las AFP's (movimiento "No más AFP") y empezar desde cero una nueva policía.

Si analizamos el periodo Post-Apruebo de la figura 24, podemos darnos cuenta que sus palabras características obedecen a demandas y problemáticas sociales fuera de la constitución que se dieron en este periodo, como la conversación a posteriori sobre la elección de constituyentes, los escaños para los pueblos originarios, los nombres de tres candidatos a la presidencia: Boric, Kast y Sichel, de los cuales más adelante fue electo el primero de ellos. Así mismo de los populares retiros previsionales de las AFP, como el latente Covid-19 y sus nuevas variantes.

Si se busca dar aún mayor contexto a estas situaciones, se puede utilizar nuevamente el KWIC para complementar esta información.

En la figura 30, se puede apreciar, que efectivamente en la muestra el contexto de Convención es para referirse precisamente a la figura de la Convención Constitucional, quien era le encargada de redactar la nueva propuesta constitucional para los chilenos. Vale la pena destacar, que estos 5 ejemplos hablan sobre el proceso administrativo y las discusiones políticas que se pueden generan desde diferentes perspectivas críticas sobre el desarrollo de la mirada pública, con respecto, tanto a sus personajes políticos como el desarrollo propio de la propuesta.

KWIC convención		
Contexto Previo	Término	Contexto Posterior
326 vale la pena . La probabilidad de que nuestro voto decida el resultado , ya sea entre el Apruebo y el Rechazo , o entre la Convención Constitucional y la	Convención	Mixta , es tan pequeña , que el más mínimo costo asociado a votar será mayor que el beneficio esperado de ser el voto decisivo . ¿ Para qué votar
245 , como herramientas complementarias para lograr cambios . Sin embargo , nada de eso sirve si no somos capaces de percibir el momento en que nos encontramos.De ahí que la	Convención	Constituyente debiese comprender rápidamente que , mientras las facciones internas disputan protagonismos superficiales , el tiempo avanza y el pacto social que deben proponer para la Constitución es precisamente el
156 una información falsa o tergiversada puede propagarse masivamente afectando la imagen de la Convención o de algunos convencionales . Un 48 % del total de la muestra confía en la	Convención	Constitucional , sin embargo , entre quienes se informan por redes sociales la confianza sube a más de 55 % y entre quienes lo hacen primordialmente mediante prensa escrita ,
228 tiempos necesarios para la entrada en vigencia de las nuevas instituciones que determine el texto , o el plazo para que las instituciones existentes incorporen estos cambios.A medida que la	Convención	aborde el trabajo sustantivo de nuevas normas constitucionales , la discusión pública sobre la adecuación del cuerpo normativo a estas disposiciones puede tener un efecto informativo en el esclarecimiento de
292 , después de mantenerlos ausentes de debates televisivos y entrevistas a página completa en los diarios . Pero esos 17 escaños serán clave en la creación de alianzas en una	convención	donde nadie tiene derecho a veto de un tercio.Esta fue una elección donde todos estábamos mirando la apelación que hizo la derecha al voto estratégico , pero que nos sorprendió

Figura 30: Palabras en Contexto - Término "Convención". Elaboración propia

En el caso particular del término "inflación" de la figura 31, se ve la preocupación de la ciudadanía, con respecto al proceso inflacionario que comenzó a vivir el mundo y Chile, junto con una baja en el crecimiento económico que podía hacer caer en una crisis recesiva económica que después de una pandemia, podría ser extremadamente dañina.

KWIC inflación		
Contexto Previo	Término	Contexto Posterior
6 ¿ Chile ha logrado construir una economía en la que los shocks inflacionarios persisten menos que en el pasado y en la que los agentes económicos confían en que la	inflación	será de 3 % en el mediano plazo . Es precisamente eso lo que se arriesga : que empresas y personas comiencen a usar con más intensidad mecanismos para protegerse
27 dejando este gobierno con los precios de la electricidad y del transporte colectivo congelados y las gasolinas subsidiadas . Son , justamente , estos precios los que más reflejarían la	inflación	internacional.El ajuste es inevitable . El presupuesto aprobado normalizará la política fiscal . El Banco Central ha subido la tasa de interés y , si quiere que la inflación eventualmente
29 suavizar la convergencia ? No es una tarea fácil , las herramientas son toscas . Al menos el Banco Central puede amortiguar el ajuste , en la medida que la	inflación	lo permita.Es fundamental intentar que la eventual recesión sea breve y no una ' de verdad ' (cualquiera sea su definición) . Y que , después del ajuste
24 la inflación estuviera cerca de la meta , el Banco Central podría decir " esperemos , es un shock de precios por una vez y contractivo " . Con la	inflación	disparada no tiene ese lujo . La tasa de interés deberá ser aún más alta.A esto se agrega la incertidumbre de un nuevo gobierno . Aunque el presidente entrante ha
35 La decisión correcta sería subir gradualmente las tasas para evitar alzas posteriores más costosas . ¿ Qué ocurre , sin embargo , si llegamos a fin de año con menor	inflación	y alto desempleo ? Algunos podrían proponer hacer una acusación por subir tasas cuando el desempleo es aún 10 % y la inflación está controlada . Que la inflación no

Figura 31: Palabras en Contexto - Término "Inflación". Elaboración propia

Si nos referimos al término "pueblos", se da en un contexto sumamente nuevo para Chile y

es el involucramiento de una convención constituyente, cuyos escaños, tenían un mínimo de cupo representativo de pueblos originarios. En particular de los 5 ejemplos, todos estaban a favor de estos escaños reservados para los representantes indígenas.

KWIC pueblos			
	Contexto Previo	Término	Contexto Posterior
33	como durante la jornada eleccionaria surgieron denuncias de electores indígenas , y también de las y los vocales de mesa : falta de papeletas verdes con candidaturas para los 10	Pueblos	que participaron , desconocimiento de parte de electores / as respecto de su incorporación en el padrón electoral , falta de proactividad en vocales de mesa para ofrecer la alternativa
12	constitucional de la descentralización el poder , una que ha estado presente desde los inicios de la República y que ha provocado permanentes frustraciones . Y el pacto con los	pueblos	originarios , uno que desde el retorno a la democracia ha tratado de canalizar mecanismos para corregir las injusticias provocadas a las comunidades indígenas , especialmente desde finales del siglo
57	. Será primera vez en la historia de Chile en que votaremos por las mujeres y hombres que escribirán la nueva Constitución , en paridad y con representantes de los	pueblos	indígenas . También será la primera vez que lo haremos en dos días consecutivos.Frente a la gran relevancia de esta elección en sí , la forma parece un detalle ,
2	, muchas veces con argumentos más bien efectistas . Tengo una visión fruto de años de trabajo en el desarrollo de economía intercultural indígena en Chile y con los diez	pueblos	indígenas del Cauca Colombiano.Soy partidario de la plurinacionalidad como reconocimiento a los pueblos originarios , parte de un esfuerzo para cicatrizar las heridas y deudas históricas que dejó la mal
37	en sus diferentes ámbitos , siendo la administración pública uno de ellos . En diferentes casos se ha evidenciado lo complejo que resulta concretar diferentes derechos que atañen a estos	Pueblos	y que no son coherentes con lo establecido en el Convenio 169 , como se ha dado cuenta en diferentes experiencias para el caso chileno . A este análisis se

Figura 32: Palabras en Contexto - Término "Pueblos". Elaboración propia

En el caso del término "rechazo" de la figura 33 su uso es diverso. Por un lado, es la referencia a la campaña del Rechazo –quienes no querían aprobar la nueva propuesta de constitución – y su aparente crecimiento; y por otro, para referirse al rechazo que realizó la ciudadanía a que la propuesta constitucional fuese redactada por diputados y/o políticos, sino por ciudadanos comunes.

KWIC rechazo		
	Contexto Previo	Término Contexto Posterior
35	tiempos actuales . La otra posibilidad es que se ponga a prueba la teoría de quienes sostienen que ni los desequilibrios macroeconómicos ni el maximalismo adolescente acarrearán grandes consecuencias.Alarma y	rechazo ha causado entre muchos arquitectos chilenos , en ámbitos profesionales y académicos , un pretendido decreto del Ministerio de Vivienda y Urbanismo que modifica la ordenanza vigente sobre el acondicionamiento
27	ha facilitado enormemente la labor de quienes están por el rechazo . Los convencionales debieran tomar conciencia de cómo , sin proponérselo , algunos han contribuido a que crezca el	rechazo en las encuestas y pensarlo dos veces antes de presentar al pleno normas que terminarán siendo testimoniales.El 4 de julio la Convención concluye su labor y comienza una campaña que
25	otras que son impopulares en parte importante del electorado . Existe un problema adicional , sin embargo , que también se debe considerar.Pareciera que la decisión de votar apruebo o	rechazo en el plebiscito de salida no solo dependerá de las normas que se incluyan en el texto definitivo , sino también del proceso que llevó a su aprobación . Se
24	varias normas ya aprobadas que no son esenciales para lograr los objetivos principales de la nueva Constitución y que pueden restar muchos votos.La propuesta anterior supone que el incremento del	rechazo se debe a que las normas que está aprobando el pleno incluyen varias que son desprolijas y otras que son impopulares en parte importante del electorado . Existe un problema
37	creen en el discurso del mérito individual , pero que desconfían profundamente de los actores políticos tradicionales.En vez de identificarse con algún partido o candidato , pareciera que votan en	rechazo a las ofertas tradicionales . Este tipo de movimientos no es nuevo en política . Lo vimos en Italia con el surgimiento de Beppe Grillo y el movimiento de las

Figura 33: Palabras en Contexto - Término "Rechazo". Elaboración propia

Finalmente, con respecto a la palabra "retiros" de la figura 34, los usos convergen al contexto de los retiros previsionales como método de financiamiento para los chilenos, debido a la pandemia y el desempleo causado por la anterior mencionada. Con todo, la perspectiva es la de que los retiros son una respuesta social del descontento social, con respecto al sistema.

KWIC retiros		
	Contexto Previo	Término Contexto Posterior
7	ahora , sin elecciones a la vista y una economía recalentada , esa obsesión se vuelve el reflejo de una política pobre en proyectos de futuro.El argumento de que los	retiros afectan las pensiones futuras ya no genera mayor impacto . La realidad es que la burbuja de que el sistema de AFP podía proveer pensiones dignas se rompió hace rato
18	irrelevante dado que la inmensa mayoría de los afiliados no se ha jubilado aún.La pérdida de legitimidad del sistema de AFP es un primer elemento para explicar por qué los	retiros no sucedieron antes . Esta pérdida se cristaliza cuando comienzan las jubilaciones masivas bajo el nuevo sistema , hace una década . Las tasas de reemplazo (pensión como fracción
8	La realidad es que la burbuja de que el sistema de AFP podía proveer pensiones dignas se rompió hace rato . Y si algo quedaba de esa ilusión , los	retiros terminaron matándola para quienes menos ahorros tenían . Además , el argumento de su efecto en la inflación es , al menos , difícil de vender.Varias cosas se combinan en
23	mayor participación de mercado que ni siquiera la acepta en forma presencial . Todo esto precisamente en el momento en que la demanda por efectivo ha crecido tanto -por los	retiros de las AFP – que el Banco Central ha debido importar billetes por vía aérea , aumentando aún más el costo del efectivo . ¿Cuál podría ser la razón
1	de beneficios , con tope mínimo y máximo en pesos , hace progresivo al sistema . Como los recursos son comunes , no existen las herencias desde el FCS ni	retiros al jubilar.Este diseño no es casual : es fruto del aprendizaje de los 20 años de operación previa del sistema de pensiones . Por un lado , se agregó un

Figura 34: Palabras en Contexto - Término "Retiros". Elaboración propia

Análogamente para los periodos Post-Apruebo Plebiscito de Entrada y Post-Rechazo Plebiscito

de Entrada, podemos ver, en la figura 35, que ahora algunos términos han cambiado, ahora que hemos cambiado la comparación. Entran al análisis palabras como "televisión", "desconfianza" y "esperanza"

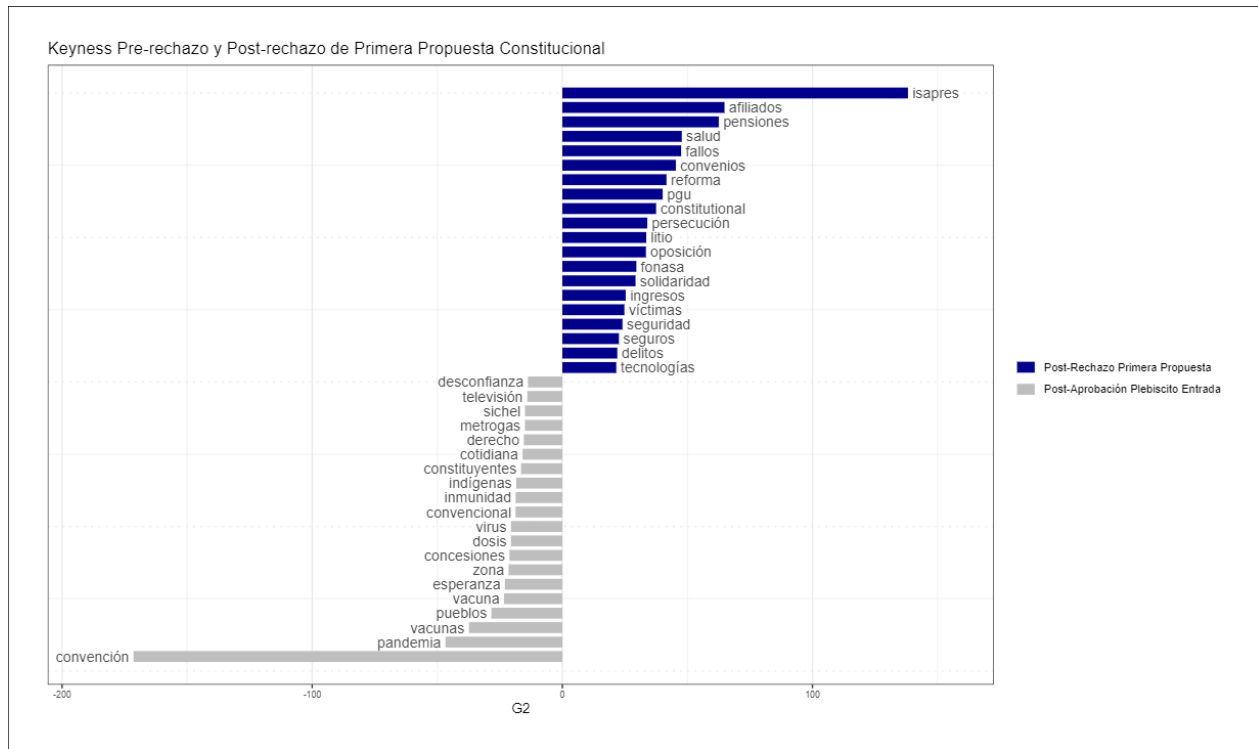


Figura 35: Palabras claves del periodo Post-Rechazo Primera Propuesta (*target*), en comparación con el periodo Post-Apruebo. Elaboración propia

Si realizamos un análisis de contexto de las palabras mencionadas anteriormente podemos ver que en la figura 36, la palabra "desconfianza" se utiliza para hablar de la incertidumbre que existía del proceso y la desconfianza de si efectivamente el proceso desembocaría en un resultado aceptable o en un fracaso.

KWIC desconfianza		
	Contexto Previo	Término Contexto Posterior
7	' poco realistas como si fueran pujas de una subasta entre buenos y malos . Cuando ese remate de promesas se produce en la Constitución , la frustración y la	desconfianza están garantizadas.Esto es significativo . Ninguno de los últimos gobiernos pudo , simultáneamente , entusiasmar a la ciudadanía y generar confianza de parte de los mercados . De acuerdo con
8	el proceso constituyente se han mantenido estables ; la esperanza y la alegría siguen siendo las que lideran este camino , aunque les siguen de cerca la incertidumbre y la	desconfianza . Nuevamente , las personas evidencian una alta preferencia por que convencionales privilegien los acuerdos por sobre sus propias posturas , aunque aquello implique renunciar a ellas . Y si
25	eso es importante partir por convencerse (y comunicar) que las cifras son lo que son y no lo que uno quisiera que fueran.Existe el temor de que la	desconfianza que está instalada entre nosotros frustra el proceso constituyente.La desconfianza influye en la forma en que evaluamos y discutimos casi todas las propuestas que vienen de quienes piensan distinto .
37	quienes quieran ejercerlo como personas con Covid positivo , postradas o privadas de libertad que no han perdido este derecho . Segundo : que no decaiga la participación por la	desconfianza en las instituciones cuyas autoridades habrá que elegir . Tercero : que el entusiasmo que se vivió el domingo se traduzca en vías efectivas de abrir la participación a grupos
1	En la sexta y última medición , realizada entre el 12 y el 25 de julio , la esperanza fue superada en promedio por la incertidumbre y alcanzada por la	desconfianza . Pareciera ser que el momento nacional reúne un poco de todo : esperanza , desconfianza y falta de certezas , y cuesta separar cuánto de esto es resultado del

Figura 36: Palabras en Contexto - Término "Desconfianza". Elaboración propia

En La palabra "esperanza" de la figura 37, nos relata otro sentir de la ciudadanía que se tenía sobre el proceso constituyente. Esperanza que en caso de que no cumpla la expectativas, terminará con un posible voto en contra, lo que finalmente pasó.

KWIC esperanza		
	Contexto Previo	Término Contexto Posterior
33	las emociones es lo que permite que los engranajes , en vez de chirriar , pongan el barco en movimiento.En Chile hay desconfianza , pero paradójicamente coexiste con una gran	esperanza puesta en el proceso constituyente (¿ sabiduría de las emociones ?) . Todo es difícil y las cosas pueden salir mal . Y sí , hay procesos que
31	a las élites . Parafraseando a Mudde , el populismo sería una respuesta democrática antiliberal a un liberalismo que ha sido antidemocrático.Entonces , quienes miran con agrado , o incluso	esperanza , a los líderes populistas , no son simplemente personas encandiladas por la farándula o las luces . Si están dispuestos a seguir a un político que hace política desde
34	su democracia.Como muchos otros chilenos que vivieron su exilio en España , Italia y otras partes del sur de Europa , para mi familia este ejemplo era una luz de	esperanza dentro de un túnel terriblemente oscuro . Quizás hubiera arena de playa debajo de los adoquines , quizás algún día se abrieran otra vez las grandes alamedas . Casi 50
44	destacar tres . En primer lugar , y pese al engegucimiento de cierta elite , la conversación en estos diálogos se realiza sobre el deseo de un futuro visto con	esperanza , la creencia de que Chile puede estar mejor que lo que está hoy . En la misma línea , en una destacada entrevista a Kathya Araujo realizada en este
18	chilenos respecto al proceso constituyente . En primer lugar , los resultados son consistentes con las emociones que los encuestados expresan , en general , hacia el proceso constituyente :	esperanza (57 %) , pero también incertidumbre (37 %) . Esto se expresa que la población muestra , de manera mayoritaria , una alta expectativa de cambios

Figura 37: Palabras en Contexto - Término "Esperanza". Elaboración propia

Finalmente, se habla de la televisión, en particular, debido a la desconfianza que existía por este medio de información. Se alegaba que ocultaban la verdad o simplemente desinformaban con una

supuesta intención oculta de manipular la opinión de lo que sucedía. La mejor información provenía de las redes sociales. Con todo, se tuvo que tener mucho cuidado con las *fake news*, debido a que se estableció las rr.ss. como primera fuente de información "confiable".

KWIC televisión		
	Contexto Previo	Término Contexto Posterior
8	su respeto y colaboración constructiva . Además , Kast decidió asistir al hotel donde se encontraba Gabriel Boric y tuvieron una pequeña reunión . Luego , todos pudimos ver por	televisión abierta una videollamada entre el Presidente en ejercicio y quién asumirá el liderazgo del Ejecutivo en marzo de 2022 . Fue una conversación formal , republicana y con altura .
10	prácticas de la política tradicional . Porque pagar lo que probablemente sean sueldos abultados a asesores de campaña , que además presidían partidos de la coalición , y confesar en	televisión que se emitieron boletas ideológicamente falsas se parece demasiado a la vieja política que se iba a reemplazar . ¿ Benefician los problemas de Boric a la candidatura de Provoste
3	la desconfianza se destacan entre quienes se informan por diarios o revistas en papel ; y , finalmente , la incertidumbre y desconfianza prevalece entre quienes se informan por la	televisión . Al considerar la reciente renovación de la mesa directiva de la Convención y el comienzo de la segunda etapa , más enfocada en la discusión propia de las normas
14	a su vez , el actual ministro de Defensa comunicó su respaldo al reproche de un general de Brigada del Ejército a un segmento de humor de un canal de	televisión . ¿ Qué tiene en común todo lo anterior ? Cada uno de estos hechos da cuenta de señales en la dirección equivocada , donde no prima la condena a
12	posible identificar que los grupos más jóvenes se informaban preferentemente a través de redes sociales , mientras que los de mayor edad lo hacían por medio de una combinación entre	televisión y portales de internet.Medios de información sobre el proceso constituyente:Al profundizar en los discursos de las personas sobre el rol de los medios de comunicación en la entrega de información

Figura 38: Palabras en Contexto - Término "Televisión". Elaboración propia

Finalmente, si analizamos el periodo constitucional Post-Rechazo de la Primera propuesta, se puede analizar la palabra "Isapres" y "pensiones" en las figuras 39, 40 y 41, respectivamente, convergen a la conclusión de que aún rechazada la propuesta las problemáticas sociales siguen latentes más que nunca, pero no pareciera considerarse en esta ocasión que la solución provenga de una constitución.

KWIC isapres		
	Contexto Previo	Término Contexto Posterior
10	la Corte Suprema , el mandato de ' asegurar la sostenibilidad del sistema ' era un desafío mayor . El primero , de noviembre de 2022 , obliga a las	isapres a restituir las ganancias excesivas que habrían obtenido entre 2020 y 2022 , por no haber considerado en el precio de los planes la tabla única de factores emitida por
60	, como la salud . La pandemia demostró la necesidad de un Estado activo y con capacidad preferente en la provisión de salud , y la crisis autoinfligida de las	isapres nos muestra lo que pasa cuando los prestadores privados funcionan con criterios de negocios . La propuesta aprobada en general en la Comisión de Expertos vuelve a consagrar esta división
47	indican que las isapres siguen siendo sumamente impopulares para la ciudadanía.La solución a la crisis de las isapres requiere que se cumpla el fallo , dando viabilidad financiera a las	isapres y , al mismo tiempo , que se introduzcan reformas largamente postergadas . Viabilidad financiera sin reformas no es políticamente factible ni socialmente deseable y , por otro lado ,
29	, la opción más eficiente.Hay mucho que mejorar en los sistemas de pensiones e isapre . También en el seguro de cesantía , en Fonasa y en la interacción entre	isapres y Fonasa . Lo mismo vale para otras prestaciones de seguridad social . Es imposible predecir en qué dirección evolucionarán las mejoras en cada caso . Dependerá de lo que
59	a que las personas de mayor riesgo y menores ingresos se concentren en Fonasa . También son síntomas de las falencias del sistema la significativa judicialización de controversias con las	isapres y un creciente sentimiento de desprotección en los afiliados . En la práctica , cuando los espacios para compartir riesgos y solidaridad se constriñen , la libertad de elección resulta

Figura 39: Palabras en Contexto - Término "Isapres". Elaboración propia

KWIC pensiones		
	Contexto Previo	Término Contexto Posterior
105	Congreso llegue a un acuerdo en una reforma de pensiones . Este dato nos indica que la gente no está poco informada porque no le interese modificar el sistema de	pensiones , sino porque no se la están explicando bien . Por otra parte , la multiplicidad de visiones indica que no será fácil construir un nuevo contrato social en pensiones
57	y , entre ellos , cerca de 200 mil pensionados . En total se retiraron \$ 53 mil millones de dólares , equivalentes a 25 % de los fondos de	pensiones antes de los retiros . El impacto sobre el sistema de pensiones resultó sustantivo y se estima que los efectos son mayores en mujeres y jóvenes y en los afiliados
131	, incluyendo rentas generales , aseguramiento social y pagos de las familias . El envejecimiento de la población presionará fuertemente el gasto en cuidados , pero lo mismo ocurrirá con	pensiones y salud . Por ello , es imprescindible considerar que ante recursos que pueden expandirse en forma gradual y limitada , las definiciones respecto de un sistema repercutirán en las
37	elección , aunque sí hay competencia , competencia por la cancha , en lugar de competencia en la cancha . Los diseños del seguro de cesantía , del sistema de	pensiones de vejez y de los seguros privados de salud ilustran bien la diversidad de opciones que se emplean en Chile para proveer la seguridad social.La provisión del seguro de cesantía
132	las posibilidades y la velocidad de desarrollo de otros . Esta disyuntiva debiera estar presente en las definiciones de política pública que se adopten en la discusión de reformas como	pensiones y saludUno de los dilemas centrales de la política es como transformar las palabras en acción , el deseo en faena y el ideal en una realidad posible . La

Figura 40: Palabras en Contexto - Término "Pensiones". Elaboración propia

KWIC seguridad		
	Contexto Previo	Término Contexto Posterior
16	última década , sino también observar de manera comparada nuestro gasto en relación con otros países y si este es concordante con nuestros niveles de ingreso.En nuestro reciente informe sobre	seguridad pública , buscamos responder algunas de estas interrogantes , tomando como punto de partida la información presupuestaria que existe tanto en Chile como en diversos países en el mundo .
22	en materias que corresponde dejar a la política ordinaria . Por ejemplo , la regulación del financiamiento y provisión de ciertos derechos sociales (libertad de elección en salud y	seguridad social , propiedad de los fondos de pensiones , y normas sobre financiamiento y currículum de la educación escolar , entre otros) , la definición de actividades asociadas a
7	de construir más cárceles , lo que tampoco es , valga la pena decir , una política fácil de concretar.No parece posible seguir con este frenesí legislativo en materia de	seguridad sin hacernos cargo de una vez por todas , de manera seria y con perspectiva de Estado y largo plazo , del enorme elefante que hay en la cristalería .
19	sistemas de justicia es proporcionalmente alto en relación con la proporción del gasto en seguridad que países europeos destinan a sus policías.Esta revisión busca poner en perspectiva el gasto de	seguridad en Chile . Más allá del monto concreto que tenga el aumento en próximos ejercicios presupuestarios , hacemos un llamado a que la determinación del presupuesto esté acompañada de una
63	vividos durante los últimos años y los daños cometidos por sus autores . ¿ Es eso una justificación para tomar una decisión tan compleja como retirarse de los acuerdos de	seguridad pública ? No , en mi opinión eso es una impostura . Los delitos cometidos por las personas indultadas se dieron dentro de un contexto , y es eso mismo

Figura 41: Palabras en Contexto - Término "Seguridad". Elaboración propia

4.4.3. *Topic Modelling*

Pareciese que ya encontramos un contexto bastante potente de lo sucedido y como la opinión pública a evolucionado con el tiempo. Con todo, quedan muchas interrogantes, debido a que si bien este análisis cuantitativo es relevante, necesita de un sustento mayor. Es por esto que, buscaremos por medio del *Topic Modelling* (modelo de clasificación no supervisado) y ya hablando de *Machine Learning*, intentar encontrar tópicos a lo largo de estos periodos constitucionales en los cuales podamos entablar una conclusión, ¿qué es lo preeminente de un periodo?

Para ello, se utiliza el LDA para clasificar un texto, esperando encontrar estos patrones, pero no patrones definidos, sino agrupaciones de texto que tienen entre ellos alguna relación aparente. De hecho, nosotros solo facilitamos el texto, sin dar ninguna clasificación previa.

Como ya explicamos en el procedimiento se elije previamente por medio de *perplexity*, el k óptimo de tópicos, que para este caso de estudio será 4. Siendo así, procedemos de manera "rústica" a asignar nombre a esos tópicos de manera que tengan no solo un sentido matemático, sino que también político y es allí donde nos muestra la figura 42, como los 4 grandes tópicos se basan en la

Crisis de carabineros, Discusión Post-Apruebo, Pensiones y Salud; y Demandas estallido social

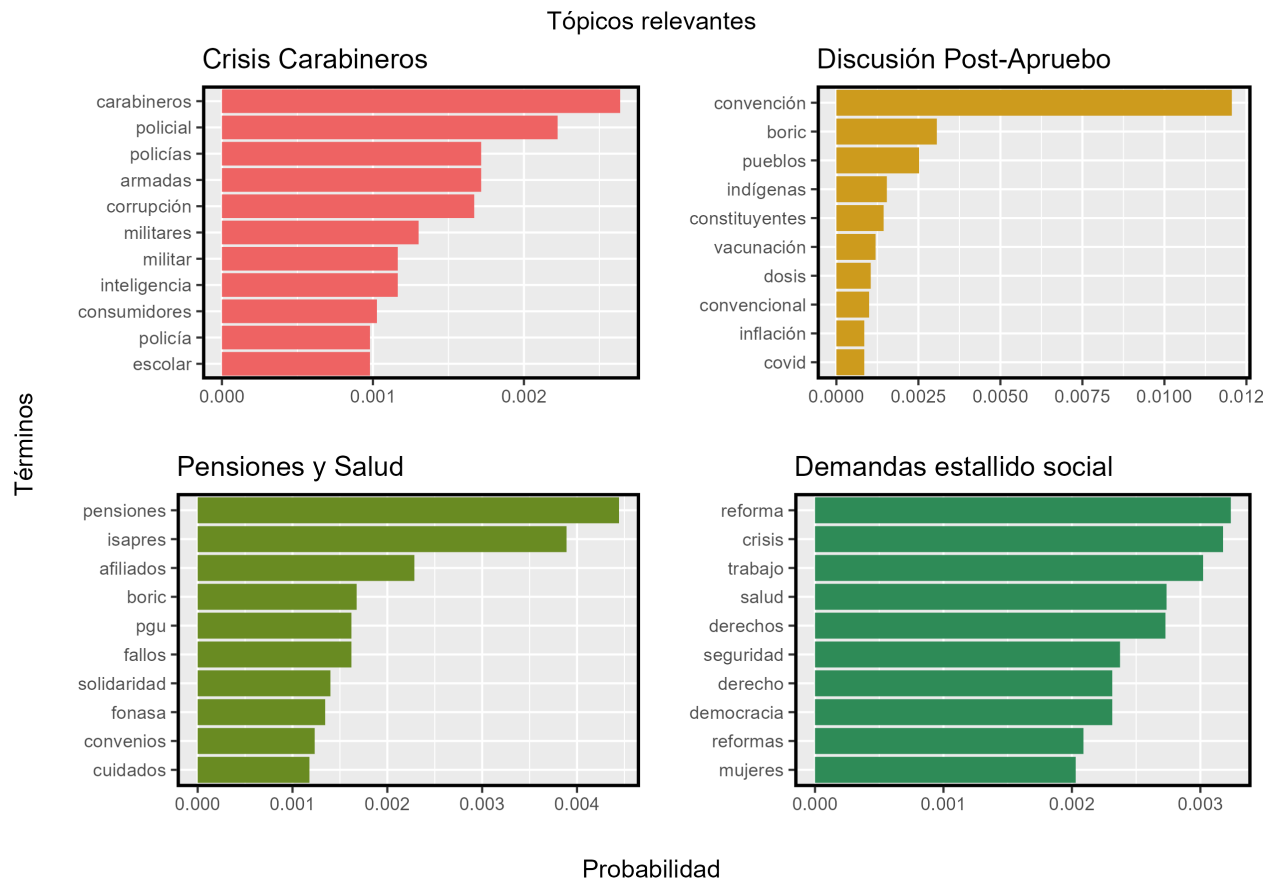


Figura 42: Modelamiento de Tópicos, $k = 4$. Elaboración propia

Como puede ver, estos tópicos no son en sí mismo, una problemática social en sí que aqueja a la sociedad y que se debería ver reflejada aquí, mas bien es una caracterización de grandes conversaciones que se tuvieron a lo largo de la opinión pública. Es decir, si queremos saber qué sucedió en el proceso constitucional, entonces tenemos que remontarnos a explicar estos 4 tópicos, cuyo contenido refleja el proceso constitucional y la evolución de la opinión pública.

Aún más, si quisiéramos caracterizar lo anterior, podemos establecer quienes son los tópicos relevantes para cada periodo – en nuestro caso, se seleccionó 2 –, podemos ver en la figura 43 que efectivamente para el periodo pre apruebo, lo más significativo de este proceso fue las demandas

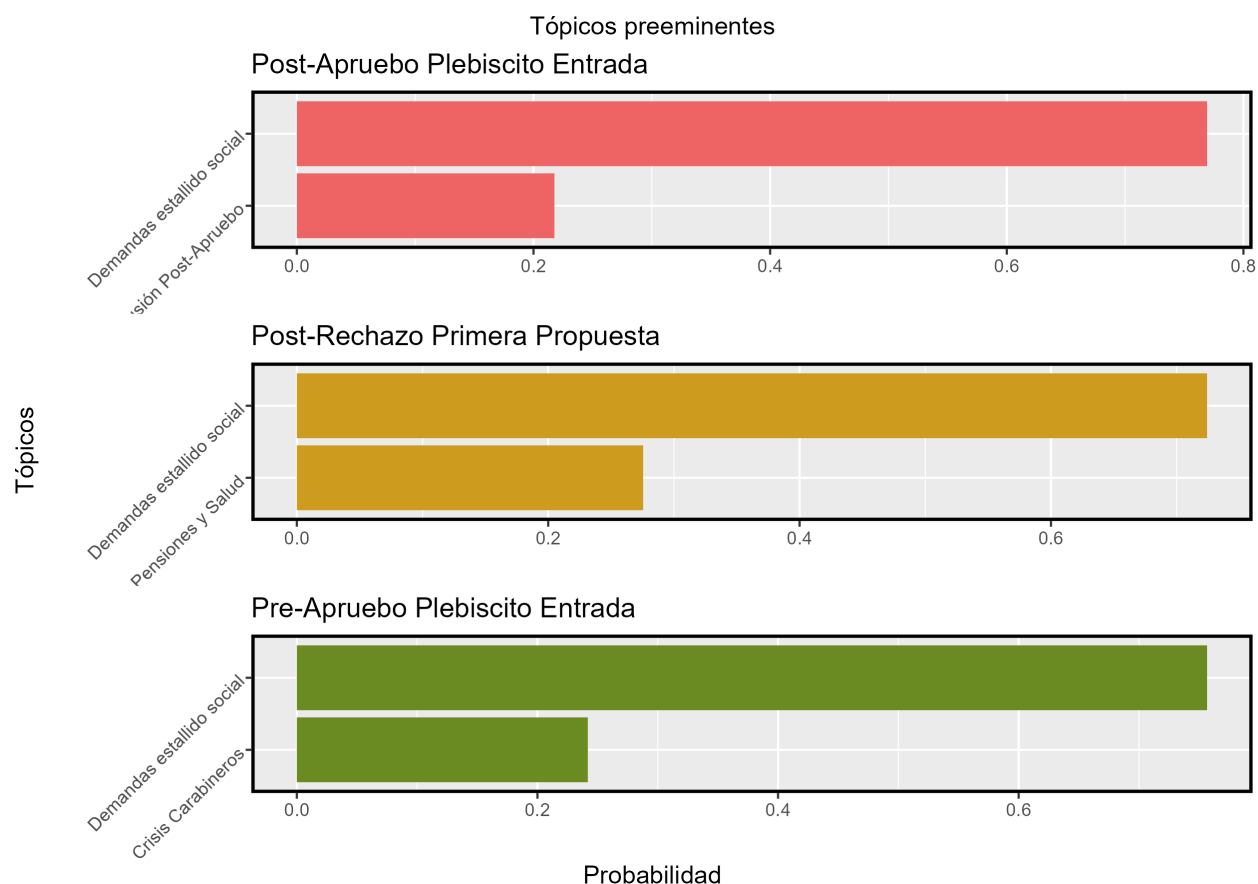


Figura 43: Tópicos Preeminentes ($k = 2$) por periodo constitucional. Elaboración propia

sociales que iniciaron el proceso constituyente y la crisis intitucional de carabineros. Si hablamos del proceso Post-Apruebo, entonces las demandas siguen vigentes más que nunca, pero ahora sigue la discusión si es que serán estas satisfechas por la propuesta constitucional y toda la incertidumbre detrás de ello. Finalmente, si hablamos del Post-Rechazo de propuesta, se caracteriza por la discusión más que nunca de pensiones y salud, esta vez como un tema cercano, donde se están viviendo las consecuencias y no solo un descontento. Además se sigue hablando de las mismas demandas sociales, pero no de solucionarlo por medio de una segunda propuesta.

De hecho, es la segunda propuesta constitucional una decisión política, "pensando en el ciudadano", más no del ciudadano *per sé* como lo fue la primera propuesta.

4.4.4. Análisis Cuantitativo de Sentimientos

Conociendo el contexto del proceso, el sentir de la población y su opinión con respecto, tanto a la constitución como otras variables que afectan su opinión. Sin embargo, nos falta ese *je ne sais quoi*, para definir un voto. No se puede definir un voto, solamente definiendo descriptivamente los hechos sucedidos, sino que debemos saber cómo afectó este contexto en la percepción de calidad y éxito de la constitución para que ésta pueda ser aceptada. Si se percibe mal la constitución, entonces lo más probable es que se vote "en contra" y viceversa. Por tanto, si la ciudadanía percibe el proceso, como en una especie promedio, tanto de contexto social como constitucional exclusivamente, como negativo, la propuesta de constitución se debería rechazar.

Son estos sentimientos positivos como negativos en las columnas de opinión, los que serán analizados, con tal de estimar este posible voto.

Este análisis lo hemos denominado "análisis cuantitativo de sentimiento", lo que nos permitirá no solo estimar de forma dicotómica "Positivo" o "Negativo", sino que también el grado de positividad o negatividad, según corresponda.

Utilizando el procedimiento descrito, en las páginas anteriores podemos clasificar cada palabra usada a lo largo del texto, de forma que en primera instancia podemos encontrar la figura 44, en la cual se calcula la proporción de apariciones de términos positivos en contraposición con los términos negativos. Para ello, se realiza la proporción de los puntajes de palabras segmentadas por periodo constitucional.

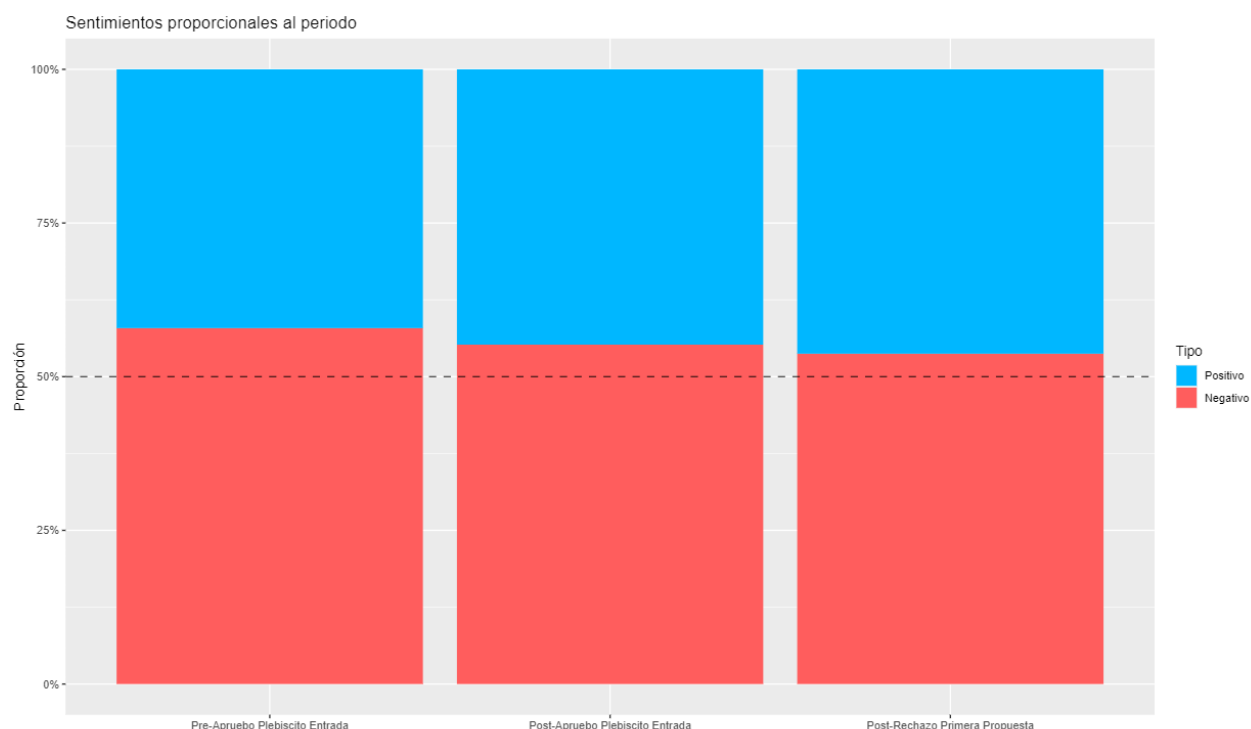


Figura 44: Enter Caption

Podemos ver claramente, que las proporciones no son significativamente altas más bien están dentro de un rango que pudiese llamarse neutro, pero la tendencia clara es el sentimiento negativo de la constitución. De hecho la proporción del periodo previo al rechazo, marca aproximadamente 55,22 % muy cercano a la votación real que rechazo la segunda propuesta constitucional con 55,7 %. Con todo, este pronóstico se condice que por ejemplo, el periodo Pre-Apruebo del Plebiscito de Entrada, como una especie de canalización, el sentimiento negativo, se traduciría en el porcentaje de aprobación de la Entrada, lo cual claramente no es (57,9 % sentimiento negativo versus 78,28 % de aprobación para el plebiscito de entrada).

Si analizamos más de cerca el proceso, fijándonos en la opinión exclusivamente sobre las propuestas constitucionales, podemos ver en la figura 45 que, interesantemente, aumentó levemente la proporción de sentimientos. Parciera que si hablamos de la opinión pública de forma agregada,

entonces podemos encontrar que los sentimientos no son tan negativos como cuando se habla de la opinión pública en cuánto al proceso constitucional.

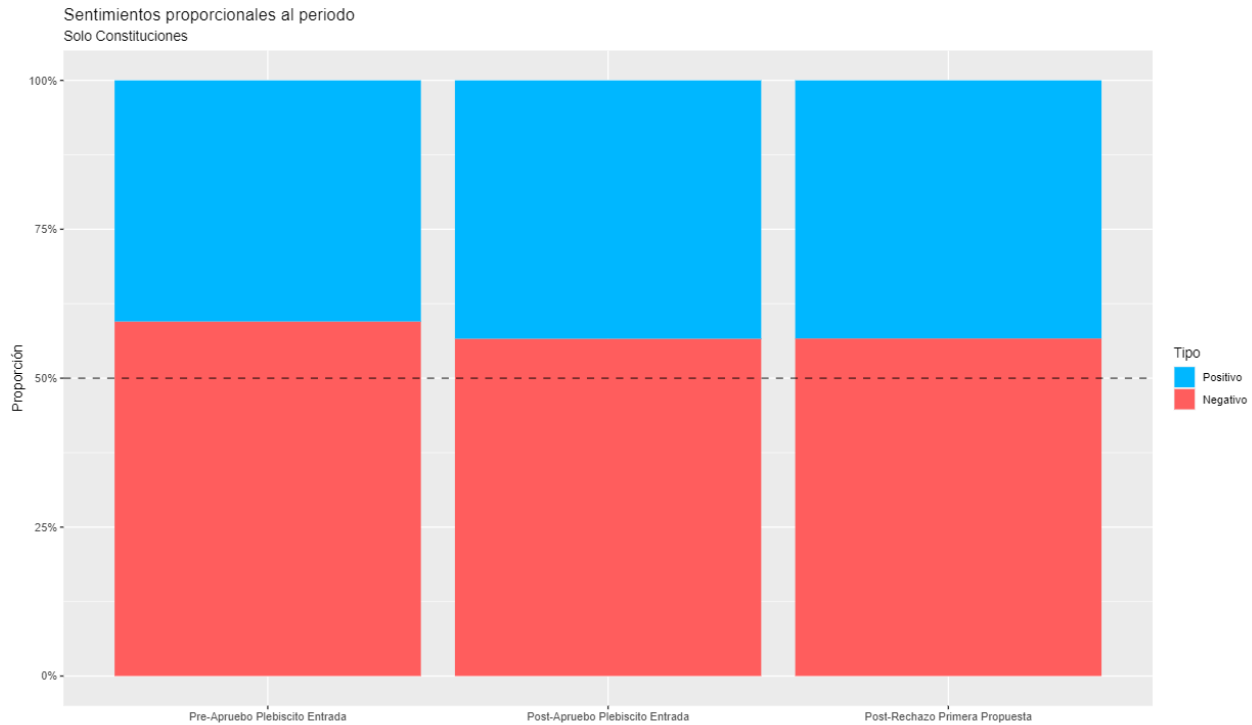


Figura 45: Enter Caption

Por otro lado, podríamos analizar los sentimientos dentro de un evolutivo en el tiempo, en el cual se pueda apreciar los cambios de percepción de esta constitución por sobre una métrica absoluta.

En la figura 46, se aprecia un cambio en el periodo pre-aprueba del plebiscito de entrada, desde una mirada más neutro, con fluctuaciones entre 0 y -0.2 para a finales del año 2019 tener una baja importante y acentuarse aún más para mediados del 2020. De allí hubo una recuperación a niveles positivos, luego del triunfo del aprueba. Con todo existió bajas importantes al llegar al año 2022. Desde allí hubo una recuperación importante que se mantuvo relativamente el resto del periodo, llegando al post-rechazo donde hubo un aumento del positivismo en este periodo y una baja considerable justo antes de las votaciones para el plebiscito de la segunda propuesta consitucional.

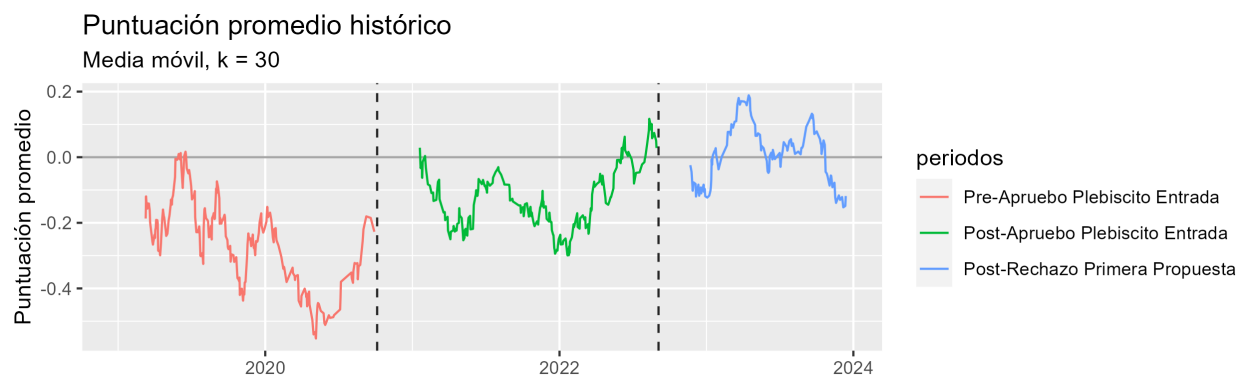


Figura 46: Enter Caption

Aún más, si utilizamos regresión local para el suavizamiento de la series, para enfocarnos mayormente en las tendencias, podemos ver lo siguiente:

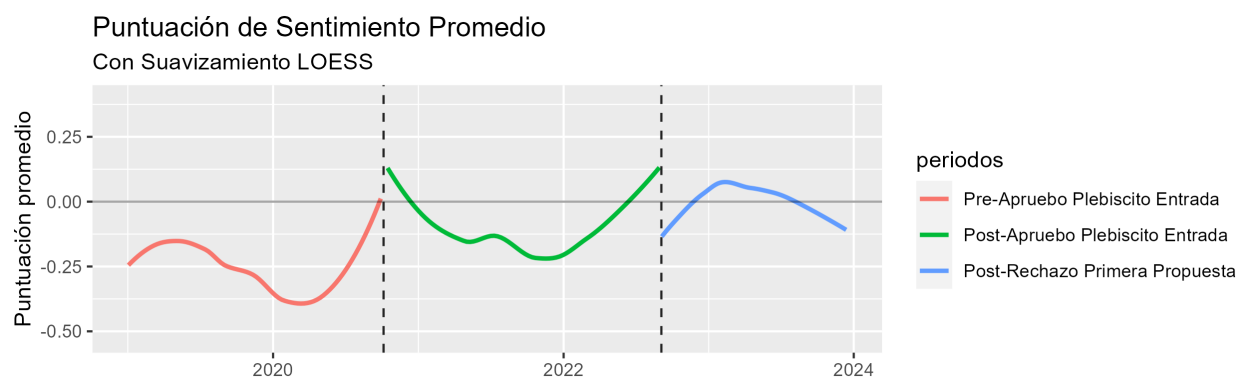


Figura 47: Enter Caption

Pero, si analizamos desde el punto de visto solo constitucional en la figura 48, podemos ver una forma similar a las tendencias del análisis agregado, pero con un nivel mayor (es más negativo), pudiendose recalcar que esta vez el periodo post-rechazo primera propeusta constitucional, tiene una caída más notoria con respecto al agregado, previa votación de la segunda propuesta constitucional (fig. 49).

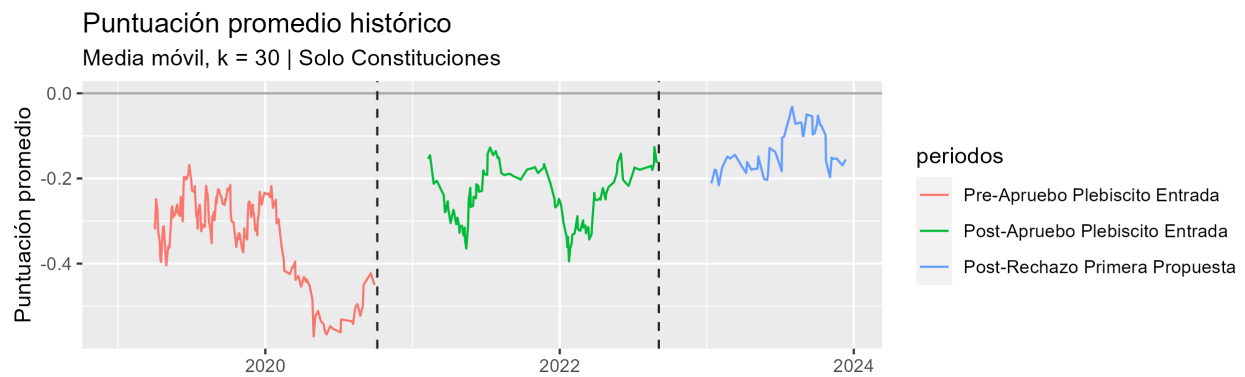


Figura 48: Enter Caption

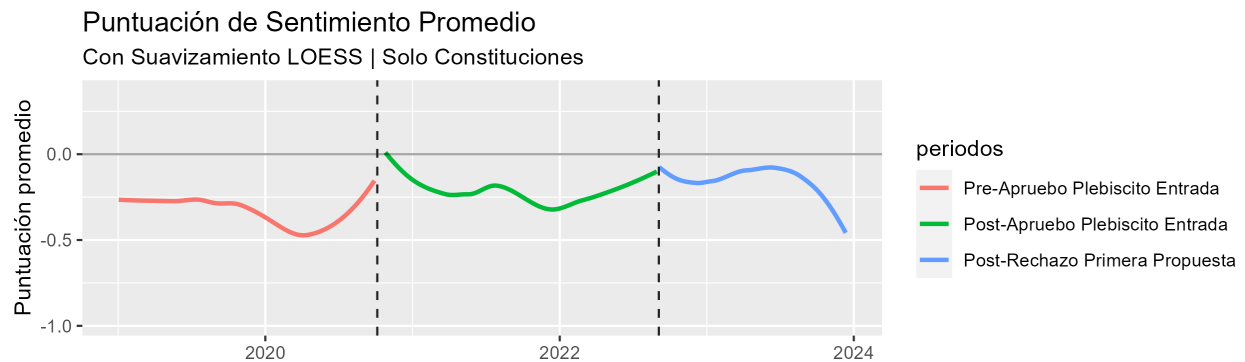


Figura 49: Enter Caption

No obstante, es importante destacar la naturaleza de los datos y es que al tratar con columnas de opinión de editoriales de medios periodísticos, éstas, si bien emiten una opinión política, esta no puede ser extremadamente subjetiva, ni refleja de forma fidedigna el pensamiento del autor de la columna. Esta es revisada y filtrada por una editorial para que los pensamientos se unan más a la línea –política– que se desea.

Es por esto que si analizamos los sentimientos de forma desagregada por cada uno de los diferentes medios periodísticos, podríamos ver si es en efecto esta hipótesis correcta.

En la figura 50, podemos ver que de forma agregada, tomando solo 7 de todos los medios selec-

cionados (debido a que solo aportaban menos de 10 observaciones), 5 de los 7 medios tienen una opinión negativa a lo largo del periodo constitucional, destacando La Hora, cuya proporción es la más negativa con poco más del 60 %.

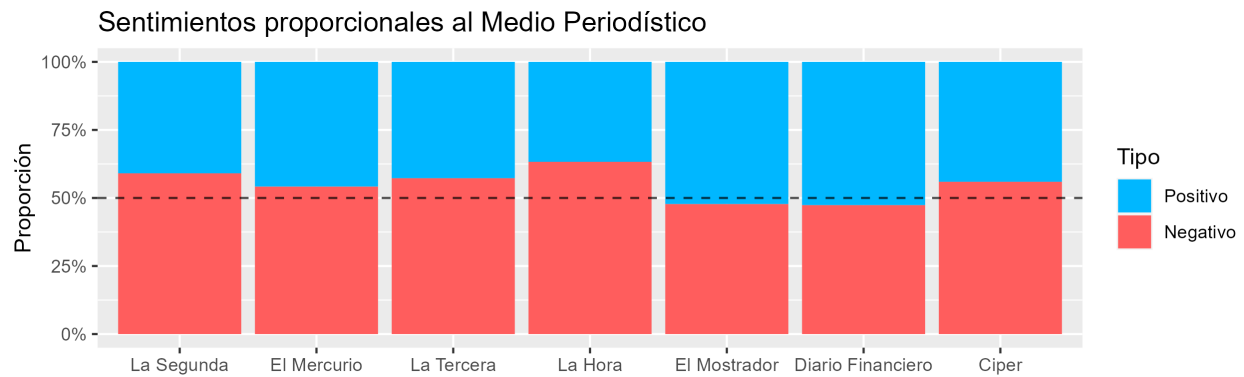


Figura 50: Enter Caption

A modo descriptivo igualmente, podemos visualizar en la figura 51, que los tres diarios con más observaciones tienen su mediana, entorno al -1 con un rango intercuartil entre -2 y 2. Comportándose sus datos de manera idéntica. Es natural, igualmente encontrar que no pasen los máximos ni mínimos del 5 y -5, respectivamente, debido a que es el puntaje máximo asignado por el diccionario. Para el caso de Ciper, su distribución es similar, sin lograr los máximos ni mínimos que si tienen los otros diarios. Esto nos permite también afirmar que, dado que Ciper es un medio independiente de los holdings de "La Segunda", "El Mercurio", "La Tercera" y "La Hora", por lo que la opinión de los diarios no viene, dada necesariamente por grupos de interés económicos. Finalmente, "El Mostrador" y "Diario Financiero", mantienen una mediana por sobre 0, argumentando una posible tendencia política diferente de las anteriores.

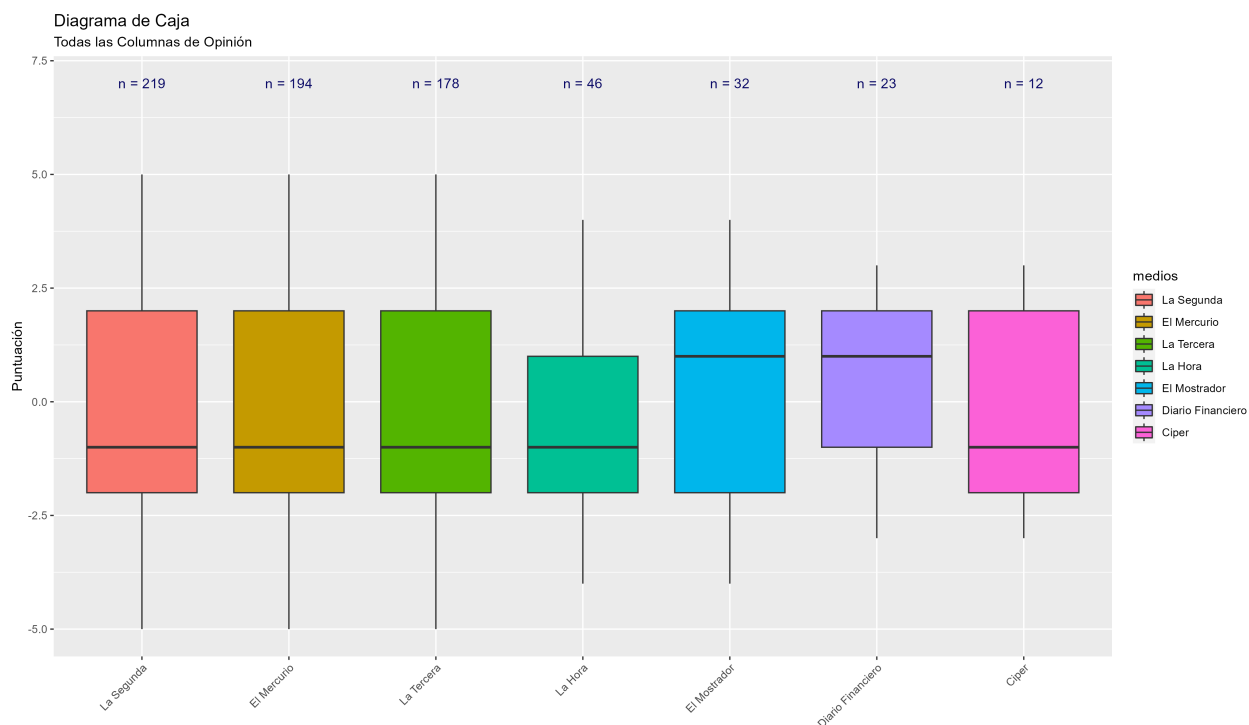


Figura 51: Enter Caption

Las figuras 52 y 53 nos muestran la media móvil para medios y la tendencia suavizada. En estos, solo están presentes 3 de los 7 seleccionados, debido a que al realizarse una media de 30 observaciones, los demás medios restantes no contaban con trayectoria suficiente, como para ser significativas.

"El Mercurio" y "La Segunda", siendo del mismo holding, opinan bastante diferente del periodo constitucional, mostrando en el caso del segundo una baja sostenida a lo largo de todo el periodo pre-apruebo plebiscito de entrada. A partir de allí un alza sostenida a lo largo de los procesos restantes, hasta llegara un sentimiento promedio positivo hacia el día de la votación, mientras que "El Mercurio", casi de manera cíclica siempre, cercano a las hitos relevantes (plebiscitos), muestra una tendencia negativa y luego repunta a niveles positivos por el resto del periodo. Hacia el final muestra una tendencia neutral en las votaciones.

Finalmente, "La Tercera", muestra una tendencia similar a la del Mercurio con ciclos más pro-

nunciados con una mayor variación de sentimientos a lo largo del proceso, pero con una tendencia a
acentuar su percepción negativa días antes de la votación.



Figura 52: Enter Caption

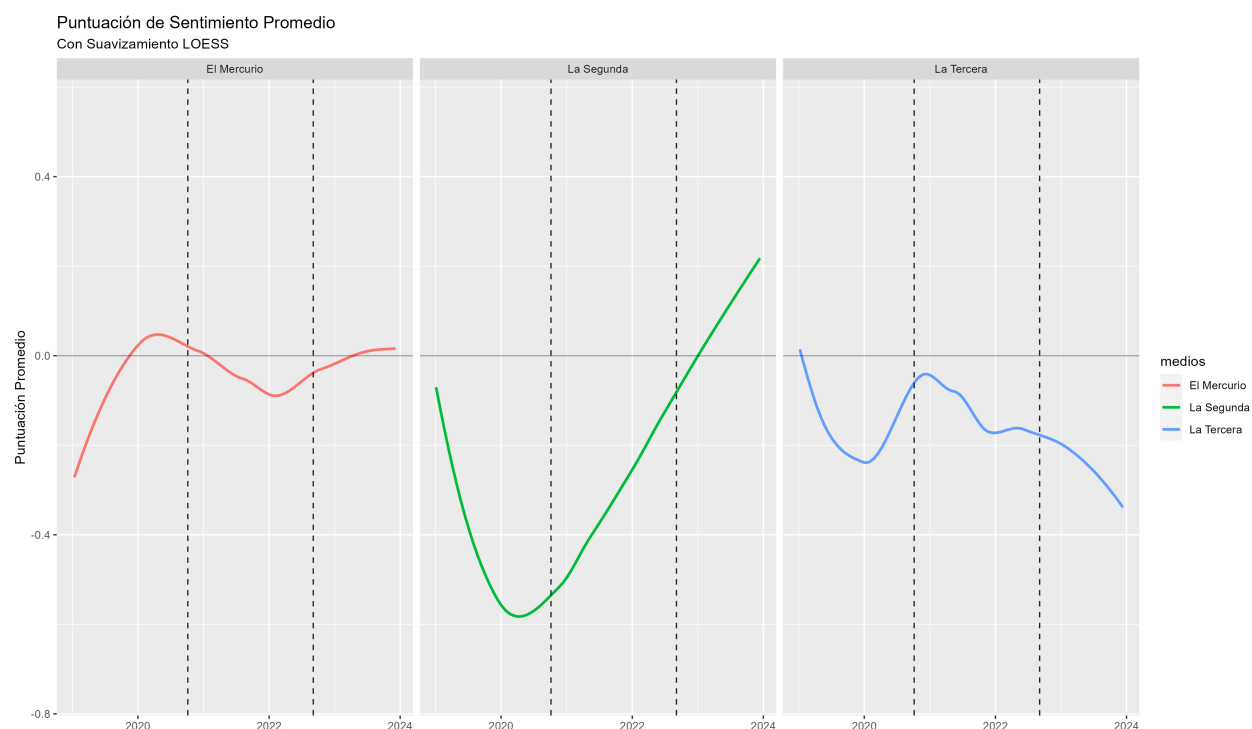


Figura 53: Enter Caption

De forma análoga, si realizamos un filtro por solo las columnas de opinión que hablen de constituciones, tenemos que, en las figuras 54 y 55. La tendencia de "El Mercurio" cambia a una tendencia más claramente negativa, en cuanto al proceso constituyente se habla y tiende a tener una pequeña disminución de negatividad en el final. "La Segunda" mantiene su tendencia y nivel de negatividad con repunte a lo largo del proceso, al igual que "La Tercera", solo que con menos variabilidad.



Figura 54: Enter Caption

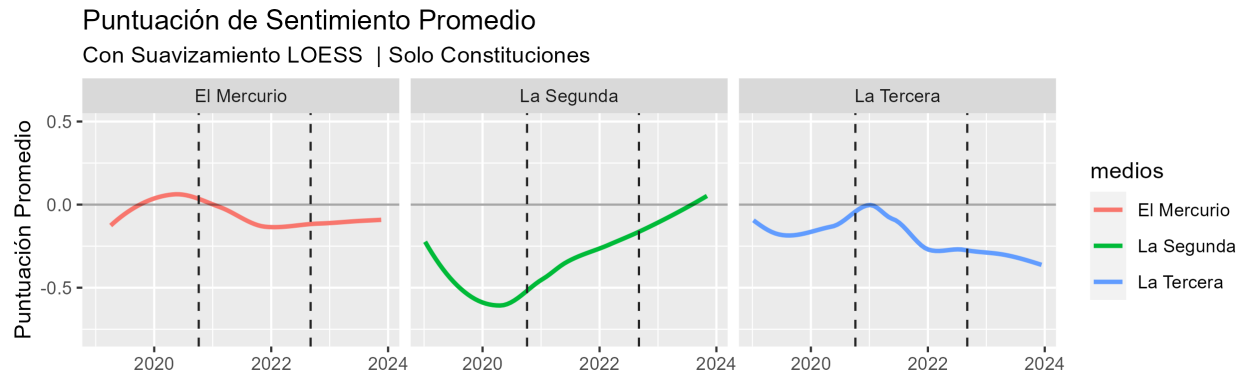


Figura 55: Enter Caption

4.4.5. *Clustering*

Como sabemos el Clustering nos podría ayudar justamente a vislumbrar estas diferencias que existiesen entre medios periodísticos, a través de la distancia semántica entre las palabras. Es así, donde no solo a través de su percepción podríamos hablar de algún sesgo político, sino que también, a través de las distancias entre sus opiniones.

Como mencionamos en el procedimiento, nuestro k óptimo será igual a 4, lo cual se ve reflejado en la figura 56, donde podemos ver la misma tendencia de los sentimientos visto en la figura 51, donde

"La Segunda", "La Tercera" y "El Mercurio" tenían una similitud notable. Sorprendentemente, Ciper es similar a El Mostrador semánticamente hablando, incluso aún cuando pareciera que sus inclinaciones políticas eran diferentes y finalmente, como casos apartes, tenemos a Diario Financiero y La Hora como dos tipos de medios con una redacción y semántica bastante diferentes de las demás. Con todo es importante destacar que esta clasificación no es tan significativa dado que la suma de las dos dimensiones, suma tan solo 57,3% cuando en general los componentes principales para ser considerado bueno, requiere de un 80 a 90 %.

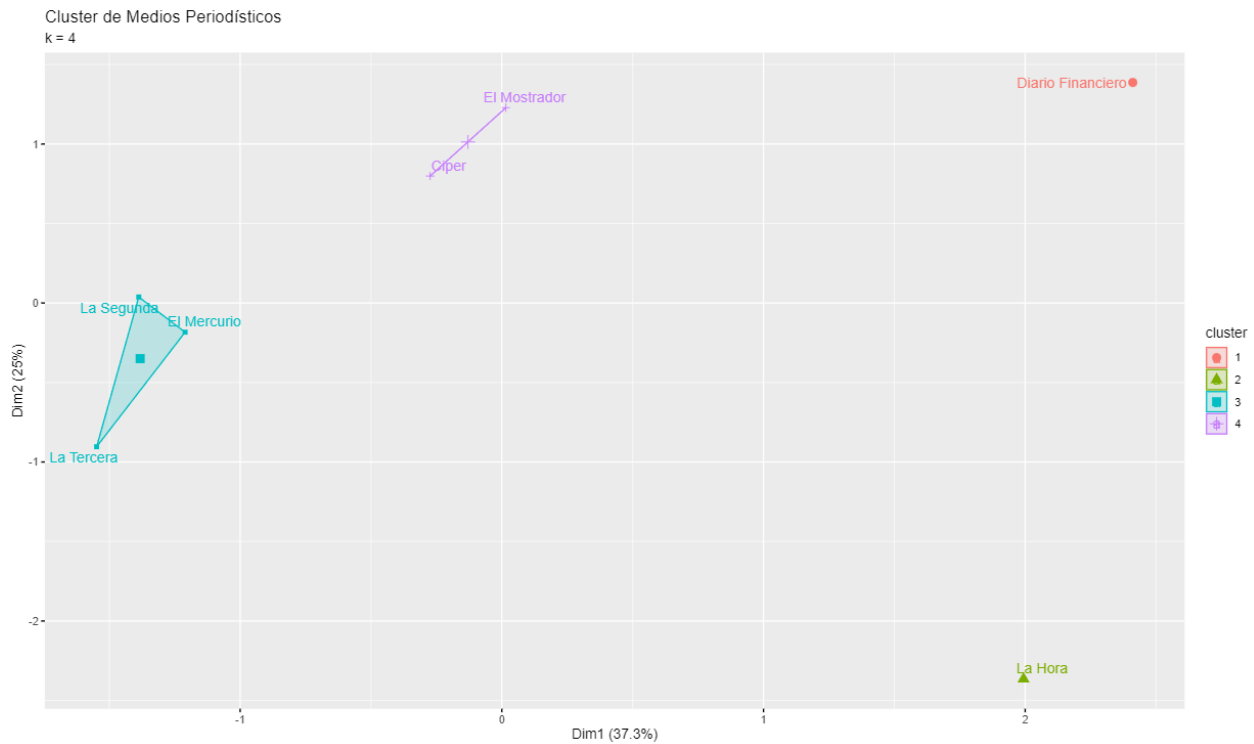


Figura 56: Enter Caption

Además como doble comprobación, podemos darnos cuenta que el *clustering* jerárquico nos proporciona los mismos grupos.

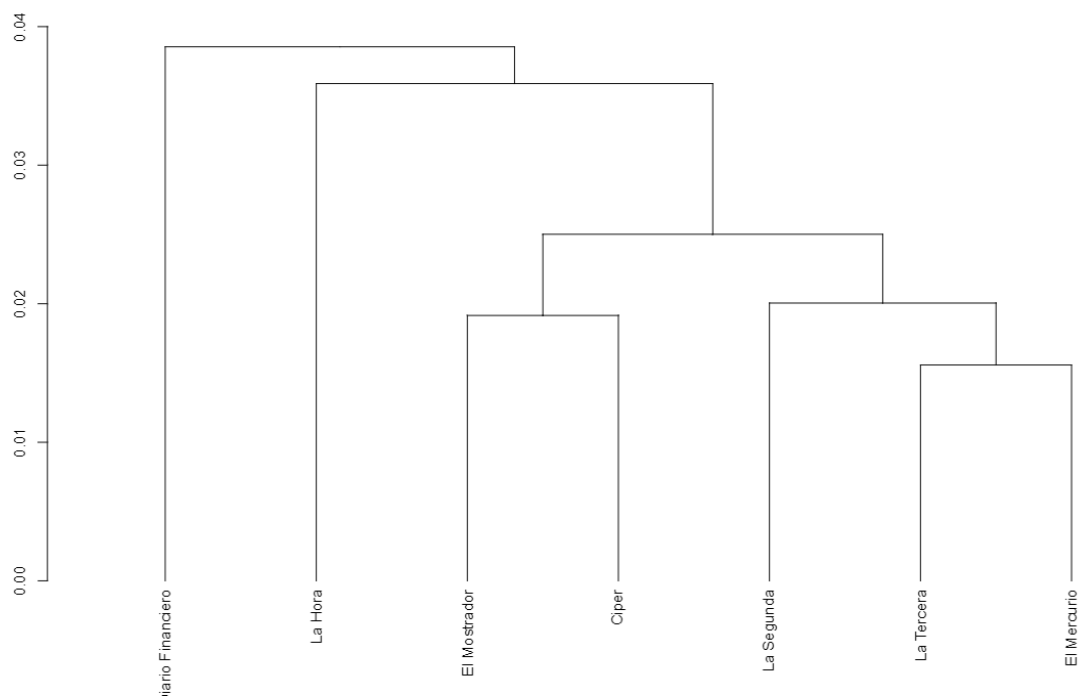


Figura 57: Enter Caption

5. Conclusiones

5.1. Conclusiones del caso

En resumen, la utilización del NLP nos permitió poner en contexto y comprobar situaciones que fueron afectando la opinión pública. Dimos un contexto por cada situación a lo largo de cada periodo constitucional, pudiendo comprobar hechos puntuales que caracterizaron a cada uno de los periodos. Increíblemente, en un análisis a escala de la frecuencia de palabras nos permitió conocer y discutir procesos, aún en la complejidad de estos procesos, de una manera más simple, permitiendonos entregar resultados sin sesgo político. En ese sentido, perfectamente este análisis podría ser utilizado por otras disciplinas para comparar, o mejor un análisis interdisciplinario con el objetivo de generar una mayor riqueza en el análisis.

Por medio del *topic modelling* pudimos entablar una manera probabilística y matemática acertada de comprobar los análisis de NLP por medio de clasificación no supervisada. Por tanto, desde dos perspectivas diferentes se pudo llegar a las mismas conclusiones, estableciendo un nivel de confianza mayor en los resultados.

Finalmente, por medio del análisis de sentimiento definir la denominada opinión pública, por medio de la percepción positiva y negativa, tanto de los procesos constitucionales como procesos sociopolíticas ocurridos en el periodo, encontrando niveles de significancia diferentes para cada uno. También segmentar por medios periodísticos acentuando en pequeñas tendencias políticas de sus editoriales, aportando también un caso de estudio, para posibles extensiones sobre este tema, debido a que actualmente, existen muy pocas publicaciones sobre esto, en el contexto chileno y de los medios periodístico aquí utilizados.

Inclusive, pudimos encontrar por medio del *clustering* relaciones entre medios periodísticos para identificar posibles sesgos periodísticos de nuestro análisis, dandonos cuenta que de las observaciones

obtenidas, los 3 medios periodísticos más relevantes (con mayor cantidad de observaciones), tienden a agruparse, lo que nos indica que, sumado a la distribución similar de los sentimientos, que estos medios podrían tener un sesgo político asociado, y que en general en el universo de medios periodístico existe una heterogeneidad de expresión sobre las noticias que informan.

Este análisis, con todo, lejos de ser perfecto, tiene sus defectos y es que si bien su procesamiento permite conclusiones sencillas, puede ser esta justamente también su debilidad. Esto, debido a la ultra simplificación de procesos sociopolíticos, en general, ultra complejos que requieren de muchos puntos de vistas diferentes, lo que puede llevarnos a sesgos. Es importante destacar, que si bien se realiza un análisis de gran escala, muchas veces pasan desapercibidos ciertos micro procesos, como lo pudo ser la gran caída de la "lista del pueblo", movimiento político con mayor escaños utilizados de la convención que se caracterizaba por ser los canalizadores de las demandas sociales. Eran ellos, ciudadanos comunes y corrientes que supuestamente "sabían mejor que nunca lo que quería el chileno". Muchos de ellos, participantes activos de las marchas del estallido social como Giovanna Grandón (conocida como "tía Pikachu") o Rodrigo Rojas Vade ("pelao Vade"), fueron justamente la mayor fuente de críticas, por su actuar dudoso. Sobre todo este último quien ya siendo electo como constituyente, tuvo que renunciar, debido a información revelada de que su supuesta condición médica era en realidad falsa y que fue utilizada como móvil para ser reconocido y electo. A partir de allí, la lista del pueblo fue desmoronándose y con ello, creciendo la incertidumbre sobre el proceso. Podría ser incluso este evento, un canón que repercutiría para siempre en la percepción ciudadana en la confianza del proceso constitucional, empero no pudo ser percibido por este análisis.

No obstante, a favor de este, podríamos decir que si este proceso de la lista del pueblo, no fue analizado, se podría deber a que si bien este proceso es relevante, dentro de la escala del análisis, no lo es tanto. No es este un problema tan transversal de algún periodo característico, sino que de un pequeño momento. Eso, sumado a que tal vez un análisis no tan agregado, sino que identificando

procesos más pequeños que los establecidos en esta publicación ayudarían a individualizar este tipo de acontecimientos. Además, podríamos argumentar, que si bien no se logra identificar, si puede ser canalizado por la opinión pública, por medio del análisis de sentimiento, como forma agregada que ponderaba todos los sentimientos, incluso en las pocas observaciones que pudiesen existir sobre la estafa de Rojas Vade.

Por otro lado, se pudo entablar una posición política, en la cual se marca un precedente al intentar cuantificar por medio de columnas de opinión (parcialmente, al menos) la percepción pública sobre los procesos constitucionales, utilizando precisamente opiniones y no encuestas, dejando la puerta abierta a mejores análisis que puedan analizarse con datos potencialmente más precisos, como lo podría ser otras fuentes de opinión fidedignas y verídicas extraídas directamente del ciudadano común; y por otro lado, que es la opinión pública una buena fuente de información, que permite entablar una predicción de tendencia política. En este caso, es la predicción acertada del voto "En Contra", dada la contextualización del pensamiento política existente y la percepción a nivel de la carga sentimental que se realiza de forma subjetiva dentro de la opinión sobre un tema.

Sin embargo, lejos de afirmar que es el Machine Learning, sea un reemplazo de encuestas de opinión u otras metodologías de medición de la misma; o la forma definitiva de predicción de tendencias políticas, queda como tarea comprobar la precisión de esta a lo largo de otros análisis similares a nuestro caso de estudios, así como de posibles extensiones que se puedan realizar para la mejora sustancial de esta clase de estudios.

5.2. Conclusiones finales

Es la opinión pública por lo tanto, una serie de factores, la cual se ve reflejada en el sentir de la población. Son, por ende, las demandas sociales periódicas que van y vienen tomando fuerza en mayor o menor medida, dado ciertos contextos socio-políticos y económicos, lo que crea una evolución

del pensamiento del ciudadano y a la larga, en una posición política, y es, precisamente, esta posición política, la que determina en sí misma el voto final del ciudadano.

En efecto, es por medio del Análisis de Sentimiento, que se puede canalizar estas posiciones políticas de una forma más simple, con tal de predecir un voto. Predicción que nos da como resultado

Finalmente se debe velar por el uso ético de estos análisis, debido a la sensibilidad de sus resultados, al ser modificados por el ojo humano.

- Realizar una nueva forma de análisis para la Comprensión y análisis de tendencias políticas por medio del análisis de texto de la Opinión Pública, utilizando Aprendizaje Automatizado.

6. Referencias bibliográficas

Referencias

- Alba, F. (2022). Deep Learning. *Universidad de Sevilla*, 1-3.
- Cicuéndez, R. (2021). Opinión pública y políticas públicas: la sensibilidad de los ciudadanos y la receptividad de los gobiernos en España. *Revista del CLAD Reforma y Democracia*, (80), 109-148.
- Dahl, R. (1993). La poliarquía: participación y oposición. *Social Policy*, 35(2), 227-237.
- Del Valle, H. (1996). El debate sobre las propiedades de la Opinión Pública en Estados Unidos. *Sociedad*, 145-159.
- Del Valle, H. (2009). Opinión pública y comportamiento electoral: De las opiniones al voto. *American Journal of Political Science*.
- Perkins, D. (2019). Learning to reason: The influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues. *Judgment and Decision Making*, 14(6), 624.
- Rojo, G. (2016). Los corpus textuales del español. *Enciclopedia lingüística hispánica*, 285-296.
- San Martín, E., & Alarcón-Bustamante, E. (2022). Dissecting Chilean Surveys: The case of Missing Outcomes. *Chilean Journal of Statistics*, 13(1), 17-46.
- Sartori, G. (1992). *Elementos de teoría política*.
- Sinclair, J. (1966). Beginning the study of lexis.
- Travalia, C. (2006). LAS COLOCACIONES GRAMATICALES EN ESPAÑOL. *Anuario de Estudios Filológicos*, 279-293.
- UC, L. (2022). Datos Faltantes: ¿Imputarlos o expresar la incerteza inducida por ellos?

Webster, J., & Chunyu, K. (1992). TOKENIZATION AS THE INITIAL PHASE IN NLP. *City Polytechnic of Hong Kong*.

Wlezien, C. (1995). The Public as Thermostat: Dynamics of Preferences for Spending. *American Journal of Political Science*, 39(4), 981-1000.