Título cool

Índice

Índice de cuadros

Índice de figuras

1 Introducción

2 Literatura sobre emotividad y polarización

3 Fuentes de información y preprocesamiento

Esta sección describe las principales características del dataset utilizado y los procedimientos realizados durante la etapa de preprocesamiento. Los datos provienen de tres fuentes de información: 1) textos parlamentarios emitidos desde 1965 a 2022; 2) biografías parlamentarias y 3) votaciones dentro de la cámara de diputados desde 2002 a 2022.

3.1 Textos parlamentarios biblioteca del congreso nacional

Los textos parlamentarios corresponden a todas las transcripciones de intervenciones parlamentarias realizadas en ambas cámaras desde el año 1965 hasta 2022. Debido a que actualmente no existe un set de datos ordenado de los discursos parlamentarios, la información fue obtenida del sitio web de la Biblioteca del Congreso Nacional por medio de técnicas de *webscraping*[1]. De esta manera, fueron obtenidas las transcripciones íntegras, junto a algunos metadatos disponibles, como el título de la intervención, fecha y autores de la misma.

La recolección de información tuvo como resultado un total de 579.663 intervenciones parlamentarias (tabla 1), tanto individuales como grupales. Luego de una edición y selección de intervenciones relevantes, el dataset final quedó conformado por 209.830 textos. Existen dos motivos que explican la reducción en la cantidad de registros. En primer lugar, se seleccionaron aquellas intervenciones en las que participa un solo parlamentario, de modo de asociar claramente un discurso a una persona[2]. El segundo motivo guarda relación con la remoción de ciertas categorías de intervenciones parlamentarias que no son de interés para el presente estudio. Existen 52 categorías de participaciones parlamentarias, muchas de las cuales corresponden a asuntos administrativos o tienen un lenguaje con un fuerte sesgo técnico-jurídico. Dichas intervenciones fueron removidas, puesto que no están asociados al objetivo de este trabajo[3]

Cuadro 1: Total de intervenciones parlamentarias

filtro	cantidad de filas
datos brutos	579.663
datos filtrados	209.830

Tal como muestra la figura 1, existe una ventana de 17 años en la que no se cuenta con información debido al cierre del Congreso Nacional durante la dictadura[4].

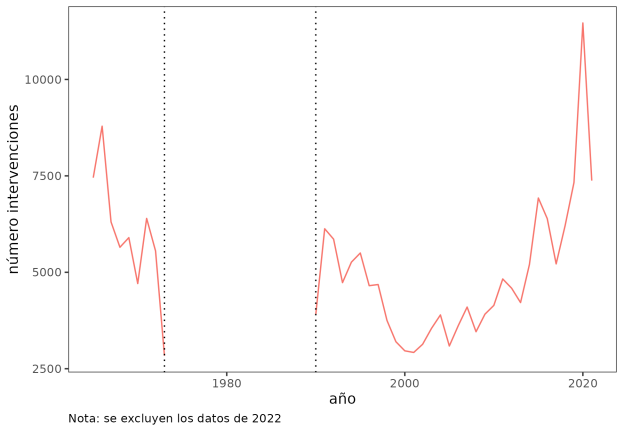
[1]En concreto, se desarrolló un código en R y por medio de una tecnología llamada Selenium se simuló un usuario que navegó a través de todos los discursos parlamentarios durante varias horas.

[2]Es común que una misma intervención esté firmada por dos o más parlamentarios

[3]Para más detalle sobre las categorías de participaciones parlamentarias ver Anexo

[4]No se muestran los datos de 2022 en el gráfico debido al bajo número de intervenciones existen en el momento de la recolección de información. Esta fue realizada durante marzo de 2022, de modo que a dicha fecha solo se registra una pequeña fracción de las intervenciones que usualmente se llevan a cabo durante un año legislativo

Figura 1: Número de intervenciones parlamentarias por año



Con el objeto de convertir los discursos parlamentarios en información estadísticamente relevante, fue necesario llevar a cabo un pre procesamiento de los datos. En primer lugar, se convirtieron todos los textos a minúscula, lo cual facilita una serie de tareas posteriores y reduce el número de palabras únicas. En segundo lugar, se removieron algunos extractos de los textos poco informativos, como los vocativos u otros encabezados similares[5].

En tercer lugar, se dividieron los discursos en párrafos[6], los cuales constituyen la unidad de análisis que da lugar a los resultados de este trabajo. Una vez separados en párrafos, los textos fueron separados (*tokenizados*) en palabras.

En cuarto lugar, se removieron los signos de puntuación y las palabras que en la terminología de NLP se denominan *stopwords*. Estas palabras se caracterizan por ser muy comunes, pues al corresponder a una parte estructural de los idiomas, se utilizan en prácticamente todos los contextos, por ende, para muchas tareas de clasificación de textos no aportan información relevante. Por lo general, las librerías utilizadas para NLP contienen listados de *stopwords*. Estos listados, típicamente, incluyen conjunciones, preposiciones, algunos adverbios y otras partículas.

Finalmente, se seleccionan los sustantivos, adjetivos y verbos mediante un modelo de *spacy*[7] entrenado para hacer POS (*Part of speech*). Mediante esta operación se busca retener aquellas palabras que aportan más significado al contenido de los discursos parlamentarios, lo cual, además, disminuye el tiempo de computación, ya que se elimina una parte importante de las palabras del corpus.

El cuadro 2 muestra un ejemplo de la situación inicial y final de un extracto de una de las intervenciones parlamentarias. Es posible observar lo siguiente: 1) el texto final está en

[5]Una gran cantidad de discursos comienza con el vocativo *señor presidente* o *señora presidenta*. Otro caso muy común se da cuando el presidente o presidenta de la Cámara cede la palabra a un parlamentario, en cuyo caso suele utilizarse la fórmula *el/la diputado/a [nombre] tiene la palabra*

[6]El separador utilizado fue el interlineado

[7]Spacy es una librería de Python ampliamente utilizada para facilitar tareas relacionadas con el procesamiento de lenguaje natural. Spacy contiene modelos para hacer POS, *name entity recognition*, mapeo de palabras a vectores, entre otras herramientas

minúscula, 2) no existen signos de puntuación, 3) varias palabras han sido removidas y 4) el párrafo original está contenido en una lista de palabras.

Cuadro 2: Ejemplo de preprocesamiento

original	final
Lo destaco, porque queremos trabajar en los proyectos de los parlamentarios. Hemos visto lo que se busca con este proyecto, el ministro de Hacienda ya había anticipado que queremos aliviar a las familias en materia crediticia y compartimos el espíritu de lo que se quiere. Y eso es justamente lo que explica que queramos trabajar sobre los diversos proyectos de ley que ustedes han empujado y han sacado adelante.	***['destaco', 'queremos', 'trabajar', 'proyectos', 'parlamentarios', 'visto', 'busca', 'proyecto', 'ministro', 'anticipado', 'queremos', 'aliviar', 'familias', 'materia', 'crediticia', 'compartimos', 'espíritu', 'quiere', 'explica', 'queramos', 'trabajar', 'proyectos', 'ley', 'empujado', 'sacado']***

Para tener una idea general de las características del dataset, el cuadro 3 muestra algunos estadísticos de resumen. Las 209.830 intervenciones, al ser separadas en unidades más pequeñas, dan lugar a un total de 2.649.588 de párrafos, cuya media de palabras es de .

Cuadro 3: Estadísticos de resumen

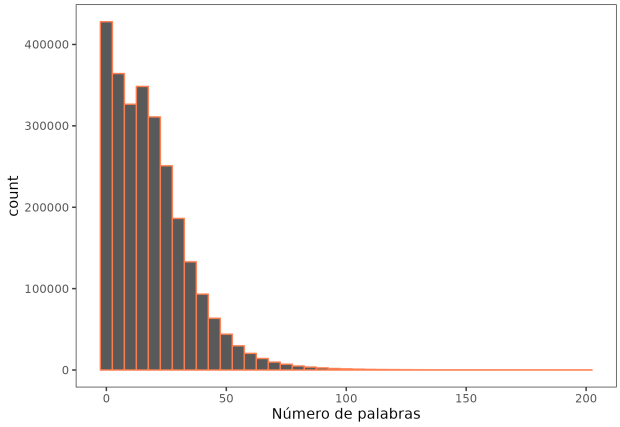
total intervenciones	total párrafos	total palabras	intervención/párrafos	intervención/palabras
209.830	2.649.588	49.065.194	12,63	233,83

Respecto a los párrafos (unidad de análisis), el cuadro 4 y la figura 2 muestran que el número promedio de palabras es aproximadamente 19 y que, en general, los textos no son demasiado extensos, ya que el 50% tiene 15 palabras o menos y el 90% tiene 39 palabras o menos. El hecho de utilizar una unidad de análisis más desagregada que la intervención (párrafo), hace más sencilla la identificación en el texto de características distintivas, las cuales cuales tienden a oscurecerse al trabajar con los textos completos, cuya extensión es significativamente mayor, como se muestra en la tabla 3.

Cuadro 4: Estadísticos de resumen de los párrafos

media	mediana	mínimo	máximo	p90
18.52	15	1	790	39

Figura 2: Histograma del número de palabras por párrafo



3.2 Biografías parlamentarias biblioteca del congreso nacional

Para obtener la historia de militancia política de los parlamentarios, se utilizaron las biografías publicadas en el sitio de la Biblioteca del Congreso Nacional. Al igual que en el caso de las intervenciones, la información fue extraída mediante técnicas de *webscraping*.

Una vez finalizada la extracción de datos, fue posible reconstruir la historia de afiliación política de cada uno de los parlamentarios. A partir de esta información, cada intervención parlamentaria puede ser asociada a una militancia específica. Es relevante constatar que dado que algunos parlamentarios presentan cambios en su militancia, es posible que dos textos enunciados por la misma persona en momentos distintos, estén asociados a partidos políticos diferentes.

3.3 Votaciones de diputados

La última fuente de información corresponde a las votaciones en sala de los parlamentarios. Para obtener estos datos se utilizó una API (*Application Programming Interface*) dispuesta por la Cámara de Diputados, mediante la cual fue posible extraer todas las votaciones emitidas en la cámara baja desde 2002 en adelante. Es importante mencionar dos limitaciones respecto a esta fuente de información:

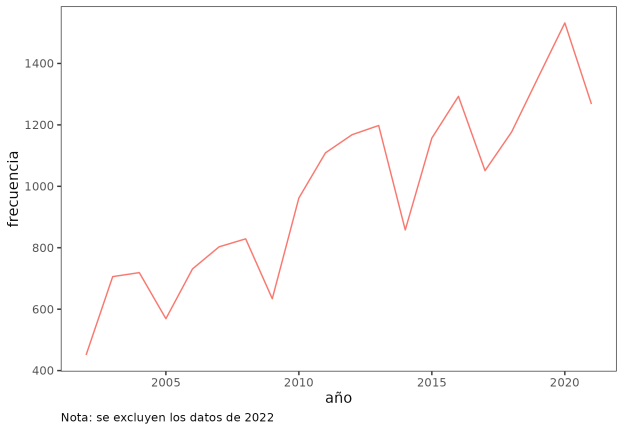
1. Solo fue posible obtener las votaciones para diputados en la ventana de tiempo que va de 2002 a 2022, pues la base de datos dispuesta por la Cámara de Diputados solo contiene datos a partir de dicho año.

2. No se cuenta con datos de votación para senadores. El motivo es que el *web service* del Senado no incluye un método para descargar dichos datos.

Estas brechas de información son parte de las limitaciones del estudio, ya que no es posible descartar que la ausencia de datos más antiguos para diputados y la inexistencia de datos para senadores, esté introduciendo algún sesgo en los resultados.

La descarga desde la API tuvo como resultado un total de 20.271 votaciones, distribuidas a lo largo de aproximadamente 20 años. Estos datos permiten conocer cuál es la situación de todos los diputados en cada una de las votaciones, pudiendo darse 4 posibilidades: *aprobación*, *rechazo*, *abstención* o *dispensado*. La figura 3 muestra una tendencia creciente en el número de votaciones por año a lo largo del tiempo. A partir de estos datos se construyó una medida de posicionamiento político que se describe en el apartado 4.5.

Figura 3: Cantidad de votaciones por año en la Cámara de Diputados



4 Metodología

Esta sección describe los aspectos metodológicos más importantes que se encuentran a la base de los datos presentados en el apartado de resultados. Se entregan las principales características del diccionario utilizado, la metodología de *word embeddings* y cómo es que esta es utilizada para ubicar cada texto en la polaridad cognitiva-afectiva.

4.1 Word embeddings

El procesamiento y análisis de datos de texto comúnmente requiere llevar a cabo alguna operación para convertir el lenguaje humano en una representación numérica que sea legible para un algoritmo. Cualquier procedimiento que permita convertir palabras en vectores numéricos se denomina *word embeddings* (Skansi 2018).

Dentro de las estrategias para construir vectores de palabras, una de las más utilizadas es el modelo *Word2vec*, cuya idea fundamental es que el significado de una palabra depende del contexto en el que esta se encuentre. Siguiendo dicha noción, para aprender vectores de palabras, se entrena una red neuronal utilizando grandes volúmenes de texto, lo cual se puede llevar a cabo mediante dos estrategias alternativas: CBOW (*Continues Bag of Words*) o *skip-gram* (Charu C. 2018). En el modelo CBOW se entrena una red neuronal para que prediga una palabra a partir de su contexto. Al contrario, en el enfoque *skip-gram* se utiliza una palabra para predecir el contexto.

Si se define que el contexto corresponde a dos palabras, en CBOW utilizaremos las dos palabras anteriores y las dos posteriores para predecir una palabra central. A la inversa, bajo la estrategia *skip-gram* se utiliza como entrada la palabra central, para predecir las dos anteriores y dos posteriores.

En términos de arquitectura, los modelos están conformados por una capa de entrada, una capa oculta y una capa de salida. La capa oculta determina la cantidad de dimensiones que tendrán los vectores de palabras. De este modo, si la capa oculta contiene 100 neuronas, el número de dimensiones para representar cada palabra será 100. Cabe señalar que tanto la capa de entrada como la de salida tienen el mismo número de dimensiones, correspondiente a la cantidad de palabras distintas en el corpus utilizado para llevar a cabo el entrenamiento.

Los modelos descritos no se diferencian en lo fundamental de los autocodificadores (*autoencoders*): se busca llevar a cabo un aprendizaje no supervisado (Skansi 2018), lo cual es posible gracias a la disponibilidad de grandes volúmenes de texto. Ahora bien, debido a que el proceso de entrenamiento por lo general es costoso, es común la utilización de modelos desarrollados por personas u organizaciones que cuentan con *hardware* adecuado para este tipo de tareas.

En el marco de este trabajo se utilizaron los vectores entrenados por (Perez y Cañete 2019) del Departamento de Ciencias de la Computación de la Universidad de Chile. Los autores utilizan el algoritmo *FastText* (Bojanowski et al. 2016) sobre un corpus en español llamado *Spanish*

Unannotated Corpora (SUC)[8]. *FastText* recoge la idea de que es posible capturar el significado de las palabras a partir de sus contextos, sin embargo, se diferencia de *Word2Vec* en el hecho de que el texto no es dividido en palabras, sino en conjuntos de caracteres más pequeños. El significado se construye en este caso a partir de cadenas de caracteres que componen las palabras. Ello hace posible, entre otras cosas, obtener vectores para cualquier palabra, independiente de que estas hayan estado o no presentes en el corpus de entrenamiento.

Pérez y Cañete (2019) ponen a disposición varios modelos, cuya diferencia principal dice relación con el número de dimensiones que tienen los vectores. El más pequeño está conformado por vectores de 10 dimensiones, mientras que el más grande, por vectores de 300 dimensiones. Con el objeto de facilitar el procesamiento de datos, en esta investigación se utiliza un modelo de 100 dimensiones. Cabe señalar que si bien los vectores de 300 dimensiones debiesen reflejar de mejor manera el significado de las palabras, el modelo de 100 dimensiones ofrece resultados satisfactorios a un costo de procesamiento significativamente menor.

Los vectores de palabras construidos mediante *FastText* y *Word2Vec* han demostrado ser capaces de capturar el significado de las palabras. Así, palabras que aparecen en contextos similares, estarán cerca en el espacio proyectado, lo cual implica que es posible llevar a cabo operaciones algebraicas y agrupar palabras según la dirección en la que apunten los vectores. Por ejemplo, si buscamos los vectores más cercanos a *rojo* (mediante similitud coseno u otra medida de distancia), utilizando el modelo de 100 dimensiones, se observa que el resultado corresponde a otros colores.

```
colores =  wordvectors.most_similar(positive=['rojo'],  topn = 5)
```

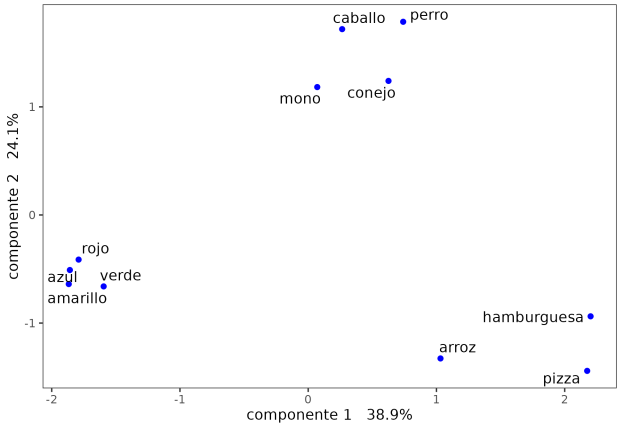
Cuadro 5: Palabras más cercanas a rojo

palabra	similitud
amarillo	0.906
azul	0.903
blanco	0.866
negro	0.853
anaranjado	0.843

La idea de que las palabras cercanas tienen un correlato en el espacio se puede expresar de manera gráfica mediante un ejercicio de reducción de dimensionalidad. La figura 4 corresponde a las dos primeras componentes de un Análisis de Componentes Principales (PCA). Se puede observar que al proyectar los vectores en este nuevo espacio de dos dimensiones, las posiciones de las palabras generan agrupaciones conceptuales. De hecho, podemos observar tres grupos claramente definidos: animales, colores y comidas.

[8]Corpus construido a partir de una gran cantidad de fuentes. El dataset está conformado por 300 millones de líneas. Para mayores detalles sobre el dataset, ver https://github.com/josecannete/spanish-corpora

Figura 4: Agrupación de palabras en un espacio bidimensional



Los vectores también permiten construir analogías del tipo a es a b como x es a y y establecer operaciones como la siguiente:

$$reina \approx rey - hombre + mujer \tag{1}$$

La ecuación 1 es una manera de representar algebráicamente la relación *hombre es a rey, como mujer es a reina*. Mediante alguna medida de distancia (usualmente, similitud coseno) se busca el vector más cercano a *rey* y a *mujer* y que, al mismo tiempo, se aleje del vector *hombre*. La tabla 6 muestra que el vector más parecido, efectivamente, corresponde a *reina*, seguido por *princesa* y otras palabras que podrían ajustarse a la analogía. En ese sentido, los vectores permiten construir relaciones semánticas complejas y, por ende, son útiles para representar el lenguaje humano.

Cuadro 6: Ejemplo de analogía con Word Embeddings. 5 palabras más cercanas a la analogía

palabra	similitud coseno
reina	0.763
princesa	0.665
consorte	0.665
sibila	0.653
isabel	0.650

Nota: Se calcula similitud coseno entre el vector resultante de la ecuación 1 y todos los demás vectores. La tabla presenta las 5 palabras más cercanas a dicho vector

Es importante mencionar que una estrategia alternativa a la de *word embeddings* es utilizar un listado de palabras previamente clasificadas e identificar cada aparición de estas en los textos. Una vez hecho lo anterior, es posible construir una medida sintética para cada

documento mediante alguna operación de agregación, como suma simple, suma ponderada u otro procedimiento similar. Existen al menos dos grandes ventajas de utilizar el enfoque de *word embeddings* en lugar de estrategias que busquen simplemente la presencia o ausencia de palabras en un texto.

1. No se requiere un *match* exacto de palabras, ya que es posible trabajar con la noción de distancia en un espacio vectorial. A modo de ejemplo, si un diccionario contiene la palabra *rabia* y no la palabra *ira* y se intenta clasificar el texto *los políticos a veces sienten ira*, el enfoque de *word embeddings* será capaz de detectar que la palabra *ira* apunta hacia una dirección cercana a *rabia*, asignando un puntaje conforme a alguna medida de distancia. Al contrario, una estrategia que considere únicamente la presencia de una palabra, no podrá asignar puntaje.

2. No se requiere establecer *a priori* el puntaje de cada palabra del diccionario. Inevitablemente, al utilizar diccionarios surge la pregunta sobre la intensidad de una palabra respecto a algún concepto. Por ejemplo, ¿las palabras amistad y amor deberían tener el mismo puntaje de afectividad? Existen diccionarios, como AFINN (Nielsen 2011), SentiWordNet o VADER (Hutto y Gilbert 2014), que establecen puntajes en la polaridad negativo-positivo, utilizando una metodología basada en jueces. Esto hace surgir preguntas respecto al modo en que el puntaje fue asignado: ¿Cuántos jueces deben votar? ¿Qué palabras deben seleccionarse? ¿Qué escala se utilizará?, etc. Otra estrategia posible es que todas las palabras tengan puntaje igual a 1, de modo de evaluar simplemente la presencia o ausencia de las mismas en un texto, como se hace en el diccionario Bing (Hu y Liu 2004), sin embargo, asignar el mismo puntaje no resuelve el problema, pues ello también es una ponderación (todas las palabras tienen la misma ponderación).

El enfoque de *word embeddings* no requiere lidiar con este tipo de decisiones, ya que el vector que representa una palabra contiene su significado. En ese sentido, si el entrenamiento funcionó y los vectores efectivamente dan cuenta del significado de las palabras, entonces, no es necesario tomar decisiones respecto a la asignación de ponderaciones. En términos empíricos, Gennaro y Ash (2021) entregan evidencia de que una estrategia basada en *word embeddings* genera mejores resultados que una estrategia basada un *match* exacto de palabras, para el análisis de textos parlamentarios en EEUU.

4.2 Diccionario LIWC

La estrategia para construir los polos cognitivo y emotivo comienza con un diccionario llamado LIWC (*Linguistic Inquiry and Word Count*). Este diccionario clasifica una gran cantidad de palabras en una serie de dimensiones. Su construcción ha sido validada por psicólogos del lenguaje (Pennebaker et al. 2015) y presenta una serie de propiedades psicométricas que lo hacen confiable para fines estadísticos. Dentro de las dimensiones del diccionario existe una relacionada con procesos psicológicos, la cual a su vez contiene

las subdimensiones de procesos cognitivos y procesos afectivos. El primer paso, entonces, consiste en seleccionar todas las palabras que están etiquetadas en estas 2 subdimensiones.

Siguiendo la metodología propuesta por Gennaro y Ash (2021), se lleva a cabo una selección de palabras en dos pasos. En primer lugar, se extraen los sustantivos comunes, adjetivos y verbos, por medio de una técnica de etiquetado llamada POS (*part of speech*). Con ello, se busca retener aquellas palabras que aportan mayor significado a la clasificación en la polaridad cognitivo-afectivo. Al llevar a cabo dicho filtro, la cantidad inicial de palabras en el polo afectivo cae de 1.586 a 1.390 y de 1.656 a 1.468, en el polo afectivo (tabla 7).

Cuadro 7: Total de intervenciones parlamentarias

polo	conteo inicial	conteo POS	conteo final
afectivo	1.586	1.390	278
cognitivo	1.656	1.468	294

Fuente: Elaboración propia con datos de LIWC

El segundo paso en la selección de palabras consiste en remover aquellas que estén menos correlacionadas con cada una de las polaridades. Ambos polos contienen una gran cantidad de palabras y es posible que algunas no estén fuertemente correlacionadas con los polos que se pretende medir. De hecho, de acuerdo a la metodología de LIWC es posible que una misma palabra se encuentre etiquetada tanto en el polo cognitivo como emotivo. En ese sentido, es deseable eliminar aquellas palabras que introduzcan ruido y/o que no faciliten una correcta discriminación entre los polos.

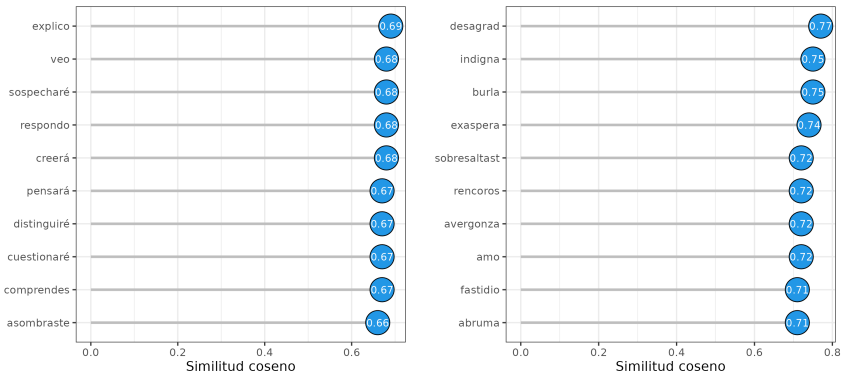
Para generar un set de palabras final para cada polaridad, se utilizan los vectores descritos en el apartado 4.1, es decir, cada una de las palabras es *mapeada* a un vector de 100 dimensiones, lo cual genera una matriz de 1.390X100 para el polo afectivo y de 1.468X100 para el polo cognitivo. Una vez finalizado dicho procedimiento, se realizan los siguientes pasos:

1. Se calcula el centroide de cada una de las matrices
2. Se calcula la similitud coseno de cada uno de los vectores con su respectivo centroide
3. Se ordenan las palabras de menor a mayor similitud
4. Se conserva el 20% de palabras en cada polaridad[9]

El listado final de palabras, luego de aplicar los pasos anteriores, es de 278 en el polo afectivo y 294, en el cognitivo (tabla 7). El objetivo de remover palabras dice relación con la necesidad de construir medidas de afectividad y cognición consistentes en si mismas, y que permitan discriminar correctamente entre discursos de una polaridad u otra. La figura 5 muestra (a través del tamaño) las palabras del diccionario que más se acercan al centroide de cada polo, es decir, aquellas palabras que mejor dan cuenta de la dimensión cognitiva y afectiva.

[9]Para determinar este porcentaje se consideró la cercanía resultante entre los vectores cognitivo y afectivo y se intentó maximizar la distancia entre ambos. Dado que las polaridades se utilizan para discriminar entre distintos tipos de textos, es deseable que los vectores no se acerquen demasiado. Para revisar los valores obtenidos a partir de diferentes porcentajes de palabras retenido, ver Anexo.

Figura 5: Palabras del diccionario más representativas de cada polaridad



Con el objetivo de validar que los vectores efectivamente estén midiendo afectividad y cognición, es importante observar cuáles son las palabras del corpus (conjunto de intervenciones políticas) que más se acercan a cada una de las polaridades. Para ello, se calculó la similitud coseno entre cada una de las 193.205[10] palabras distintas del corpus que no están dentro del diccionario y los vectores que representan a los polos cognitivo y afectivo. Las figuras 6a y 6b muestran (a través de su tamaño) cuán cerca se encuentra una palabra de cada una de las polaridades. Se observa que, efectivamente, los vectores construidos para cada una de las polaridades dan cuenta de afectividad y cognición, ya que mientras en el panel izquierdo (polaridad afectiva) palabras como *rencoroso*, *atormentado* y *enfado* muestran predominancia, en el panel derecho resaltan verbos como *decir*, *demostrar* y *preguntar*.

Cabe mencionar que las palabras del polo afectivo presentan un sesgo hacia emociones tradicionalmente consideradas como negativas. El motivo de ello es que LIWC (diccionario utilizado) tiene un sesgo hacia palabras de este tipo, lo que implica que la construcción del vector de afectividad esté sesgado hacia ese tipo de emociones, cuestión que debe tenerse en consideración al momento de analizar los resultados. Con el objeto de descartar que el polo afectivo esté capturando únicamente emociones negativas, se llevaron a cabo algunas pruebas con palabras usualmente consideradas positivas como *amor*, *alegría* o *risa*. Este ejercicio arrojó como resultado una asociación más fuerte con el vector emotivo que con el cognitivo, lo que da cuenta de que si bien existe un sesgo hacia emociones negativas, el instrumento es capaz de dar cuenta también de emociones positivas[11].

[10]Este número corresponde al total de palabras luego de haber aplicado un procedimiento de *tokenización*, mediante el cual se eliminan *stoptwords* y se seleccionan solo los adjetivos, sustantivos y verbos

[11]Para más detalles sobre estas pruebas, ver el anexo.

Figura 6: Nubes de palabras

(a) Polo afectivo *(b)* Polo cognitivo

Nota: Cada nube contiene las 200 palabras con mayor similitud coseno respecto a los vectores cognitivo y afectivo. El tamaño de las palabras se pondera de acuerdo al valor de la similitud coseno.

4.3 Identificación de la polaridad

Para convertir en vectores cada uno de los párrafos que componen las intervenciones parlamentarias, se implementan los procedimientos descritos en el apartado 4.1. En primer lugar, se busca un vector para cada una de las palabras que están dentro de un texto. Luego, para generar un indicador agregado de cada texto, se calcula el centroide de todas las palabras que lo componen. De esta manera, sin importar la cantidad de palabras contenidas en un texto, su representación final será siempre un vector de 100 dimensiones, que funciona como "resumen" del texto original.

Una vez que los más de 2 millones de párrafos son *mapeados* a su respectivo vector, es posible llevar a cabo todo tipo de operaciones algebraicas con ellos. Para identificar la polaridad de cada párrafo, se utiliza la metodología propuesta por (Gennaro y Ash 2021). La idea de fondo es que un texto puede contener simultáneamente emotividad y cognición. Ello implica que la medida utilizada debe dar cuenta de dicha dualidad y generar un valor sintético considerando ambas dimensiones. El indicador utilizado para medir emocionalidad de un texto es el siguiente:

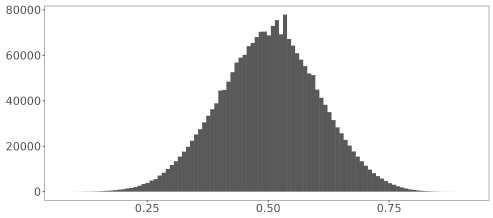
$$Y_i = \frac{sim(d_i, A) + b}{sim(d_i, C) + b} \tag{2}$$

Donde A representa al vector del polo afectivo y C, al vector cognitivo. La expresión $sim(v, w) = (v \cdot w)/(||v||\ ||w||)$ corresponde a la similitud coseno entre los vectores v y w. El término b se introduce para suavizar posibles *outliers* y puede ser cualquier número positivo pequeño. Respecto a la interpretación, un incremento en Y_i corresponde a un movimiento hacia la polaridad afectiva. Cuando Y_i toma valor 1 significa que el texto es neutro en la polaridad afectivo-cognitivo.

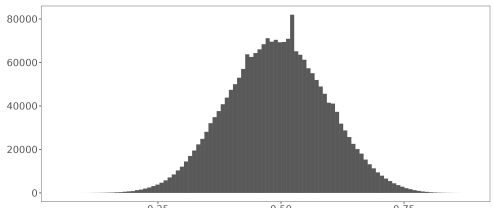
Los gráficos de la figura 7 muestran las distribuciones de Y_i, $sim(d_i, A)$ y $sim(d_i, C)$. En el gráfico del panel 7c se puede observar que el indicador sintético se mueve aproximadamente entre 0.8 y 1.2, con una distribución levemente sesgada hacia la derecha, es decir, hacia el

polo afectivo (valores mayores a 1 indican afectividad). Por su parte, los indicadores parciales de afectividad y cognición se encuentran centrados en 0.5 y se mueven entre 0 y 1.

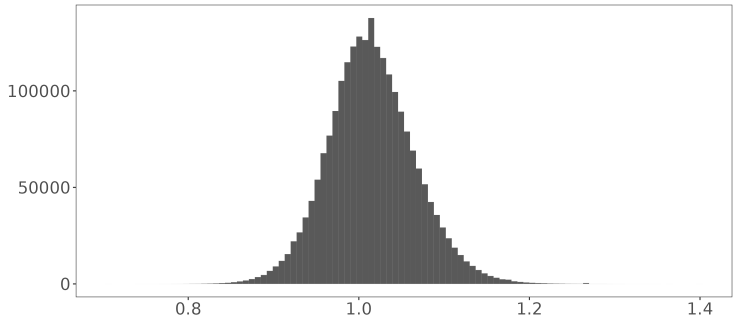
Figura 7: Indicadores de afectividad y cognición



(a) Similitud coseno polo afectivo



(b) Similitud coseno polo cognitivo



(c) Indicador sintético de emocionalidad

Nota: blabla

Con el objeto de entregar evidencia de que el indicador propuesto funciona, un ejercicio posible es inspeccionar visualmente cómo son los textos que presentan puntajes elevados en el polo afectivo y cognitivo, respectivamente. En los cuadros 8 y 9 se muestran las 15 frases con mayor puntaje en el polo cognitivo y afectivo. Una lectura rápida muestra que, efectivamente, el indicador está dando cuenta de la polaridad que se pretende medir. En el caso del cuadro 8 se observan verbos como analizar, fijar, buscar, modificar, decidir, que de alguna manera se asocian a actividades con un fuerte componente cognitivo. Por su parte, la tabla 8 contiene palabras como sensación, inseguridad, brutal, desorden, anarquía, etc, es decir, palabras que apuntan hacia una dimensión afectiva.

Cuadro 8: 15 frases más cognitivas

text
- después fijaremos la fecha exacta.
- más adelante analizaré los otros artículos del proyecto.
- los numerales de que consta son los siguientes.
- eso es lo que tenemos y lo que buscamos modificar.
- aquí efectuamos una modificación que señalaré más adelante.
- aquí dijimos revisemos esto veamos qué ocurre.
- porque cambiamos la ubicación de ese artículo.
- ¡si no los tenemos ahora ni los tendremos mañana si no llegamos a acuerdo.
- entonces votamos y establecemos un plazo para las.
- además tenemos proyectos aprobados.
- que los pocos instrumentos que tenemos los utilizaremos para llegar a un acuerdo con el gobierno eso haremos.
- corresponden a las indicaciones números 233 y 234.
- eso lo discutiremos cuando llegue el texto respectivo.
- tenemos que analizar cómo lo hacemos para adelante.
- vamos a decidir si sacamos o no de la tabla el proyecto.

Cuadro 9: 15 frases más afectivas

text
- desocupación de los jóvenes un 30 por ciento de la población juvenil.
- la sensación de inseguridad en la población.
- sea una indolencia brutal total.
- sembrando un clima de inquietud de inseguridad de violencia.
- entonces más fragmentación más inestabilidad más desgobierno.
- para alentarlos al desorden y a la anarquía.
- a con esfuerzo físico excesivo.
- en consecuencia con el mismo ánimo solidario reflejo mi preocupación por aquello.
- acá hay lluvias en demasía y sequías excesivas hay falta de agua en el verano y exceso en el invierno.
- f la campaña del terror desatada por los latifundistas.
- sin el ánimo de disminuir la importancia de la iniciativa.
- además ésa es una inquietud de numerosos sectores de nuestra ciudadanía.
- esta situación ha contribuido a la exacerbación y recrudecimiento de otros males socialmente nefastos creando un clima de inseguridad desconfianza y desesperanza.
- ¡eso y no el boicot es lo que está provocando la escasez de alimentos.
- cierta hilaridad es manifestación de nerviosismo.

Para aportar más información respecto al funcionamiento del indicador, las siguientes figuras muestran un resumen de los 5.000 textos con mayor puntaje cognitivo y 5.000 con mayor puntaje afectivo. El ejercicio consiste en calcular la frecuencia de las palabra de cada uno de los conjuntos de datos (luego de haber removido las *stopwords*) y graficar dicha información mediante nubes de palabras. Así, palabras de mayor tamaño reflejan una alta frecuencia y viceversa. Se puede observar que mientras en el polo cognitivo resaltan palabras como proyecto, artículo, o indicación, en el polo afectivo se observan palabras como manifestaciones, aplausos y violencia.

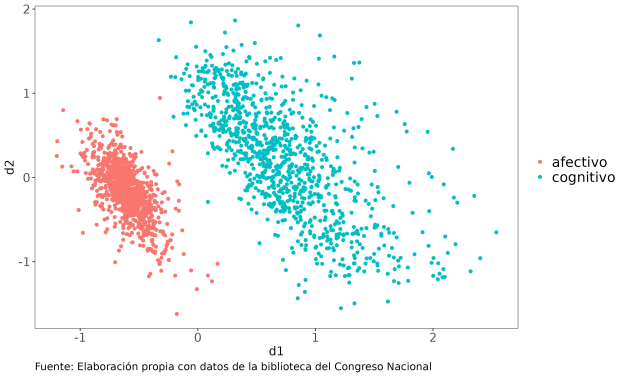
Figura 8: 5.000 frases más cognitivas y afectivas

(a) Palabras polo cognitivo *(b)* Palabras polo afectivo

Nota: Fueron removidas las stoptwords para mejorar la visualización. Cada nube contiene un máximo de 150 palabras

Un último ejercicio para evaluar el indicador consiste en comprobar si la agrupación de palabras mostrada en las nubes de palabras (figura 8) tiene un correlato en términos espaciales. Si el indicador utilizado realmente refleja dos polaridades, debería ser capaz de discriminar entre diferentes tipos de contenido textual y generar agrupaciones. En ese sentido, es posible seleccionar la representación vectorial construida mediante *word embeddings* (100 dimensiones) de los los textos con mayor puntaje cognitivo y afectivo (1.000 por polaridad), y proyectar ese espacio de 100 dimensiones en uno de 2, mediante PCA. La figura 9 da cuenta de que pese a que se ha reducido de 100 a 2 dimensiones, la información conservada es capaz de generar una agrupación de textos coherente, ya que efectivamente los textos con contenido afectivo y cognitivo ocupan espacios que no se superponen.

Figura 9: Proyección en dos dimensiones de los mil primeros textos cognitivos y afectivos



Fuente: Elaboración propia con datos de la biblioteca del Congreso Nacional

4.4 Identificación de tópicos

La identificación de tópicos se realizó mediante un modelo basado en ELECTRA [12], al cual se le aplicó un procedimiento de *fine-tuning* para la tarea específica de detectar tópicos. Esta arquitectura (Clark et al. 2020) está compuesta por dos redes: red generadora y red

[12]El modelo original fue bautizado como ELECTRA y SELECTRA corresponde a su versión en español.

discriminadora. La primera es entrenada para predecir una palabra a partir de su contexto, mientras que la segunda (red discriminadora) recibe un entrenamiento para discriminar si una palabra corresponde a un dato sintético (una predicción de la red generativa) o a un dato original. Tal como señalan Clark et al. (2020), este modelo presenta reminiscencias de las redes generativas adversarias (GAN), sin embargo, existen algunas diferencias que la distancian de dicho diseño.

El modelo recibe como entrada un texto y una serie de tópicos considerados relevantes. La respuesta consiste en un vector que contiene la probabilidad que la red le asigna a cada tópico. Para cada texto se seleccionó el tópico con probabilidad más alta, el cual se usó como etiqueta. Los tópicos considerados fueron: 1) salud, 2) educación, 3) deporte, 4) medioambiente, 5) impuestos, 6) cultura, 7) pensiones, 8) sindicalismo, 9) transporte, 10) familia y 11) aborto.

4.5 Polarización política

Para incluir una medida de polarización política se utilizó un modelo proveniente de la ciencia política llamado W-NOMINATE, cuya formulación inicial fue realizada por Poole y Rosenthal (1983)[13] que permite posicionar a cada político en un continuo ideológico a partir de sus votaciones en el congreso.

La idea central del modelo es que los legisladores tienen un punto ideológico ideal, de modo que mediante sus decisiones de voto intentarán minimizar la distancia respecto a dicho punto ideal. Poole y Rosenthal proponen que la función de utilidad de los políticos depende de un componente determinístico y de un componente de shocks aleatorios. Se asume que las personas intentarán maximizar su utilidad, mediante votaciones que minimicen la distancia respecto a su punto ideal, sujeto a un componente aleatorio.

Considerando estas ideas, la utilidad U del legislador i en la votación j, por haber votado afirmativamente (representado por el subíndice y) es:

$$U_{ijy} = u_{ijy} + \epsilon_{ijy} \tag{3}$$

$$u_{ijy} = \beta exp[\frac{\sum_{k=1}^{s} w_k^2 d_{ijyk}^2}{2}] \tag{4}$$

u_{ijy} representa la parte determínistica de la utilidad del legislador, mientras que ϵ_{ijy} representa el componente estocástico. El término d_{ijyk}^2 es la distancia euclidiana entre el punto ideal x_i del político en la dimensión k y la posición z_{jyk} resultante de haber votado afirmativamente el proyecto de ley:

$$d_{ijyk}^2 = \sum_{k=1}^{s} (x_{ik} - z_{jyk})^2) \tag{5}$$

[13]El modelo inicial de Poole y Rosenthal fue bautizado como NOMINATE. Con el tiempo comenzaron a surgir variaciones de la idea original, lo que dio lugar a los modelos D-NOMINATE, W-NOMINATE t DW-NOMINATE. En la actualidad, W-NOMINATE es el más utilizado y, por ende, con implementaciones en lenguajes de programación

Tanto el peso w como β deben ser estimados, partiendo de valores de 0.5 y 15, respectivamente. w representa la poderación de cada dimensión política, mientras que el término β corresponde a la importancia que tiene la parte determinística de la utilidad. Así, valores altos de β implican una pérdida de relevancia del componente aleatorio.

Si bien el modelo puede utilizarse para obtener, s cantidad de dimensiones, por lo general se utilizan las dos primeras, ya que se ha demostrado empíricamente que no se requiere más que ello para generar agrupaciones coherentes. De hecho, en muchos casos es suficiente la primera dimensión para resumir el comportamiento político de las coaliciones. En ese sentido, el modelo puede ser entendido como estrategia de reducción de dimensionalidad, ya que típicamente se parte con cientos o miles de votaciones, las cuales son reducidas a una o 2 dimensiones.

A continuación, se muestran los puntajes promedio obtenidos para algunos de los partidos más relevantes a nivel nacional a lo largo de todo el periodo disponible. Se observa que la primera dimensión del modelo genera una agrupación satisfactoria respecto a la tradición política de los últimos 20. El posicionamiento de los partidos de derecha (RN, UDI y EVOPOLI) y de izquierda (PS, PPD, RD, PCS, PC) refleja relativamente bien la configuración ideológica de la Cámara de Diputados. Es interesante notar que el modelo es capaz de dar cuenta de movimientos en el comportamiento de los partidos, como por ejemplo, el caso de la Democracia Cristiana (DC), cuyo posicionamiento pasa desde un alineamiento con los partidos de izquierda y centro izquierda, hacia un viraje hacia la derecha en los últimos años.

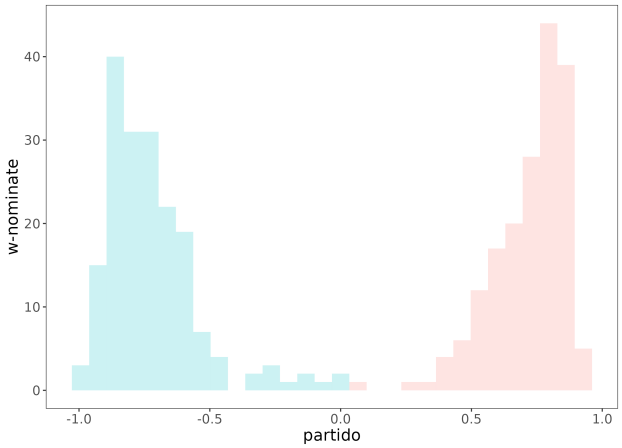
Figura 10: W-NOMINATE: Puntaje promedio por partido de la primera dimensión

Fuente: Elaboración propia con datos de la Cámara de Diputados

Para efectos de este trabajo, se utilizan medidas agregadas de posicionamiento político, clasificando a cada parlamentario en las categorías izquierda y derecha. La figura 11 muestra los puntajes del modelo para el año 2021, considerando los mismos partidos del gráfico anterior, pero ahora a nivel de cada parlamentario. Se observa que los polos de izquierda y derecha se encuentran bien definidos y que existe una pequeña parte en la que se produce convergencia entre ambas polos, lo cual se explica principalmente por el posicionamiento de los parlamentarios del Partido Demócrata Cristiano. Ello se observa con bastante claridad

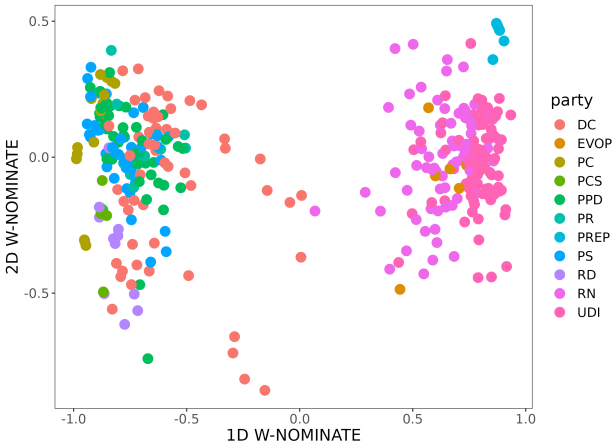
en la figura 12, donde los puntos rojos corresponde a la posición de los parlamentarios de dicha colectividad en un espacio de dos dimensiones.

Figura 11: W-NOMINATE: Puntaje promedio por parlamentario de la primera dimensión



Fuente: Elaboración propia con datos de la Cámara de Diputados

Figura 12: W-NOMINATE: Puntaje promedio por parlamentario de las dos primeras dimensiones



Fuente: Elaboración propia con datos de la Cámara de Diputados

5 Resultados

5.1 Estadística descriptiva polos y tópicos

5.2 Regresión

6 Conclusiones

Mis grandes conclusiones

Referencias

Bojanowski, Piotr, Edouard Grave, Armand Joulin, y Tomás Mikolov. 2016. «Enriching Word Vectors with Subword Information». *CoRR* abs/1607.04606. http://arxiv.org/abs/1607.04606.

Charu C., Aggarwal. 2018. *Neural Networks and Deep Learning. A Textbook*. Springer Cham. https://doi.org/https://doi.org/10.1007/978-3-319-73004-2.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, y Christopher D. Manning. 2020. «ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators». En *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=r1xMH1BtvB.

Gennaro, Gloria, y Elliott Ash. 2021. «Emotion and Reason in Political Language». *The Economic Journal* 132 (643): 1037-59. https://doi.org/10.1093/ej/ueab104.

Hu, Minqing, y Bing Liu. 2004. «Mining and summarizing customer reviews». https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf.

Hutto, C. J., y E. E Gilbert. 2014. «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14)». https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399.

Nielsen, F. Å. 2011. «AFINN». Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby: Informatics; Mathematical Modelling, Technical University of Denmark. http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html.

Pennebaker, J. W., R. L. Boyd, K. Jordan, y K. Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. https://www.liwc.app/static/documents/LIWC2015%20Manual%20-%20Development%20and%20Psychometrics.pdf.

Perez, Jorge, y José Cañete. 2019. «Spanish Word Embeddings». https://github.com/dccuchile/spanish-word-embeddings.

Poole, Keith, y Howard Rosenthal. 1983. «A Spatial Model for Legislative Roll Call Vote Analysis». *American Journal of Political Science* 29 (agosto). https://doi.org/10.2307/2111172.

Skansi, Sandro. 2018. *Introduction to Deep Learning. From Logical Calculus to Artificial Intelligence*. Springer Cham. https://doi.org/https://doi.org/10.1007/978-3-319-73004-2.