

Emotividad y polarización política en discursos parlamentarios: 1965 a 2022

Portada de la tesis

Índice

1	Introducción	5
2	Literatura sobre emotividad y polarización	7
2.1	Literatura general	7
2.2	Caso chileno	7
3	Fuentes de información y preprocesamiento	9
3.1	Textos parlamentarios biblioteca del congreso nacional	9
3.2	Biografías parlamentarias biblioteca del congreso nacional	12
3.3	Votaciones de diputados	12
4	Metodología	14
4.1	Word embeddings	14
4.2	Diccionario LIWC	17
4.3	Identificación de la polaridad	20
4.4	Identificación de tópicos	25
4.5	Polarización política	25
5	Resultados	29
6	Resultados	29
6.1	Estadística descriptiva de polos y tópicos	29
6.2	Regresiones (PENSAR EN UN MEJOR TÍTULO)	34
7	Conclusiones	39
	Referencias	41

Índice de cuadros

1	Total de intervenciones parlamentarias	9
2	Ejemplo de preprocesamiento	11
3	Estadísticos de resumen	11
4	Estadísticos de resumen de los párrafos	12
5	Palabras más cercanas a rojo	15
6	Ejemplo de analogía con Word Embeddings. 5 palabras más cercanas a la analogía	16
7	Total de intervenciones parlamentarias	18
8	15 frases más cognitivas	22
9	15 frases más afectivas	23
10	Especificaciones sin WNOMINATE	36
11	Especificaciones con WNOMINATE	38

Índice de figuras

1	Número de intervenciones parlamentarias por año	10
2	Histograma del número de palabras por párrafo	12
3	Cantidad de votaciones por año en la Cámara de Diputados	13
4	Agrupación de palabras en un espacio bidimensional	16
5	Palabras del diccionario más representativas de cada polaridad	19
6	Nubes de palabras	20
7	Indicadores de afectividad y cognición	21
8	5.000 frases más cognitivas y afectivas	24
9	Proyección en dos dimensiones de los mil primeros textos cognitivos y afectivos	24
10	W-NOMINATE: Puntaje promedio por partido de la primera dimensión . . .	27
11	W-NOMINATE: Puntaje promedio por parlamentario de la primera dimensión	28
12	W-NOMINATE: Puntaje promedio por parlamentario de las dos primeras dimensiones	28
13	Emocionalidad y cognición a lo largo del tiempo	30
14	Emocionalidad y cognición, según tópicos	31
15	Emocionalidad y cognición a lo largo del tiempo	32
16	Emocionalidad y cognición en algunos tópicos a lo largo del tiempo	33
17	Relación entre WNOMINATE y emocionalidad	34

1 Introducción

El uso de recursos afectivos en el arte de la persuasión ha sido objeto de estudio (CITAR a Aristóteles) desde el mundo antiguo hasta la actualidad (BUSCAR CITAR MODERNA). Es un hecho que en casi cualquier controversia o intento de convencer emergen aspectos emotivos como un recurso retórico. Todos los días estamos expuestos a argumentos de tipo emotivo por parte de la publicidad, ya sea a través de música, imágenes o texto. Por supuesto, la política no es la excepción. Dado que lo que está en disputa es el poder y aspectos fundamentales de la vida en sociedad, no es extraño que la actividad política, en gran medida, esté teñida de componentes emotivos.

Pese a la constatación anterior, sabemos muy poco sobre cuál es el rol que cumple la emocionalidad en la decisiones de los políticos y cuál es el balance entre las dimensiones afectivas y cognitivas en el devenir del proceso político y, por ende, en el desarrollo de la vida en sociedad. La escasa información se debe en parte a la inexistencia de fuentes de información estructurada y al desconocimiento por parte de los científicos sociales de técnicas de análisis provenientes del campo de la computación, posibles de ser usadas para analizar grandes volúmenes de texto. Esta situación ha venido cambiando aceleradamente durante los últimos años y las técnicas de Procesamiento de Lenguaje Natural (NLP) se encuentran cada vez más establecidas como estrategia de investigación dentro de las ciencias sociales.

Justamente, este trabajo contribuye a superar las 2 barreras antes mencionadas. En primer lugar, se pone a disposición un set de datos semi estructurado de discursos parlamentarios con información desde 1965 a 2022, construido a partir de información web. En segundo lugar, se incorporan técnicas de NLP para analizar más de 200.000 textos, lo cual sería imposible mediante métodos tradicionales del enfoque cualitativo en ciencias sociales.

Siguiendo la metodología de XXX, se utiliza una estrategia de *deep learning* que permite construir representaciones vectoriales de las palabras, denominadas *word embeddings* (CITAR MIKOLOV Y GENTE DE GOOGLE). Los vectores resultantes han demostrado ser capaces de reflejar la semántica de las palabras y dar cuenta de conceptos complejos. Aprovechando esta potencialidad, este trabajo busca construir una dimensión que capture un polo cognitivo y otro emotivo, utilizando para ello conjuntos de palabras clasificadas en un diccionario desarrollado por XXX (CITAR).

Calculando la cercanía de cada discurso respecto de los polos cognitivo y emotivo, se computa un valor sintético de emocionalidad para cerca de 2.6 millones de párrafos, correspondientes a más de 3 décadas de discursos parlamentarios, lo cual permite construir una serie histórica desde 1965 a la actualidad. Utilizando un clasificador automático de textos, se elaboran tópicos de discusión parlamentaria, con el objetivo de identificar cuál es el nivel de emocionalidad presente en cada uno de ellos.

Finalmente, mediante modelos lineales, se aborda la relación entre polarización y emocionalidad en los discursos parlamentarios, utilizando un indicador de la ciencia política denominado WNOMINATE. Una de las principales conclusiones del estudio es que existe una relación positiva entre extremismo ideológico y emocionalidad. En ese sentido, a diferencia de otros estudios que abordan polarización política en el contexto chileno (CITAR

ALGUNAS), el presente trabajo incluye la variable de emocionalidad, de modo de estudiar cuál es la relación entre ambas.

2 Literatura sobre emotividad y polarización

2.1 Literatura general

XXXXX NOTAS PARA TERMINAR LA REDACCIÓN

2.2 Caso chileno

A partir de la década del 2000, varios autores se han avocado a estudiar el fenómeno de la polarización política al interior del congreso, lo cual ha dado forma a un conjunto de trabajos que, sin ser demasiado extenso, permite entender relativamente bien la dinámica parlamentaria desde el retorno a la democracia hasta la actualidad. Es posible distinguir, al menos, dos líneas de trabajo o énfasis al momento de abordar el tema. La primera línea utiliza los votos individuales (*roll-call voting*) para ordenar a los políticos, partidos y conglomerados en un continuo ideológico, a fin de identificar las dinámicas de cohesión y polarización, utilizando mayoritariamente el método WNOMINATE XXXX POOLE WNOMINATE o métodos bayesianos, basados en remuestreos. La segunda línea de trabajos intenta ir un poco más allá de los votos individuales, poniendo atención a la información disponible sobre coautoría de proyectos de ley.

Dentro del primer grupo de trabajos, Saiegh y Alemán (2007), utilizando datos de la cámara de diputados, muestran que en el periodo 1997-2000 las dos grandes coaliciones votan de manera altamente cohesionada, lo que da forma a dos bloques con poca superposición ideológica entre sí. A partir ello, los autores descartan la existencia de un centro político, cuya existencia estructuró el régimen político hasta 1973. Respecto a datos del Senado, Alemán (2008) muestra conclusiones muy similares para el mismo periodo de estudio, es decir, la dinámica de bloques separados ideológicamente se reproduce en el Senado. En el mismo estudio, los autores analizan datos sobre copatrocinio de proyectos de ley, a partir de lo cual concluyen que en la etapa inicial del trámite legislativo, los Senadores enfrentan menos restricciones que en el momento de la votación, lo cual hace posible cruzar fronteras ideológicas y establecer relaciones de cooperación entre coaliciones. En una línea similar, ALEMÁN (2009) analiza el periodo 1961-2006, llegando a la conclusión de que las redes de cooperación entre los partidos se debilitaron en las décadas de los sesenta como resultado del aumento de la polarización política y que estas se recomponen progresivamente desde 1990. También aplicando herramientas del análisis de redes, Morán (2020) muestra que los proyectos de ley exitosos presentan una mayor colaboración entre colaciones, en comparación con el total de los trámites legislativos, para el periodo 2007-2017.

Más recientemente, Fábrega (2022) aplica las ideas sobre ideología y cooperación al contexto de la Convención Constituyente. El autor aporta evidencia de que las metodologías basadas en WNOMINATE o en métodos bayesianos permiten reconstruir adecuadamente el posicionamiento ideológico de los convencionales y entrega datos que respaldan la idea de que quienes cooperan en la realización de propuestas, presentan mayor cercanía ideológica. Uno de los estudios más recientes sobre polarización política describe la situación en el congreso antes y después del plebiscito de salida. Los autores dan cuenta de que se ha

producido un aumento de la polarización en ambos lados del espectro y un debilitamiento del centro político.

La literatura sobre polarización en Chile aborda las dinámicas de cooperación-polarización al interior del congreso, sin embargo, hasta el momento no existen estudios que intenten explicar cuáles son los factores asociados a la polarización. El presente estudio es un intento por avanzar en esta línea de trabajo, aportando evidencia sobre la relación que existe entre polarización y afectividad en la arena política.

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 persona². 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 trabajo³

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 dictadura⁴.

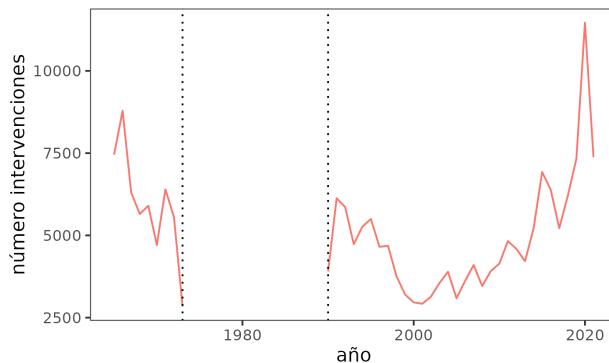
¹En concreto, se desarrolló un código en R y por medio de una tecnología llamada Selenium se simuló un usuario que navegó a través de todos los discursos parlamentarios durante varias horas.

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

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

⁴No se muestran los datos de 2022 en el gráfico debido al bajo número de intervenciones existentes 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



Nota: Debido al momento en el que se hizo la recolección, solo se obtuvieron datos de las primeras semanas de 2022. Se excluyen dichos datos para no distorsionar la comparaciones entre años.

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 similares⁵.

En tercer lugar, se dividieron los discursos en párrafos⁶, 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

⁵Una 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*

⁶El separador utilizado fue el interlineado

⁷Spacy 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
<p><i>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.</i></p>	<p>[‘destaco’, ‘queremos’, ‘trabajar’, ‘proyectos’, ‘parlamentarios’, ‘visto’, ‘busca’, ‘projeto’, ‘ministro’, ‘anticipado’, ‘queremos’, ‘aliviar’, ‘familias’, ‘materia’, ‘crediticia’, ‘compartimos’, ‘espíritu’, ‘quiere’, ‘explica’, ‘queramos’, ‘trabajar’, ‘proyectos’, ‘ley’, ‘empujado’, ‘sacado’]</p>

Nota: Se aplican los siguientes pasos: 1) convertir textos en minúscula; 2) remoción de extractos poco informativos; 3) Separación de textos en párrafos; 4) tokenización; 5) remoción de signos de puntuación y stopwords; 6) selección de palabras relevantes mediante POS

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, lo que quiere decir que, en promedio, una intervención contiene 12,63 párrafos y 233,83 palabras.

Cuadro 3: Estadísticos de resumen

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

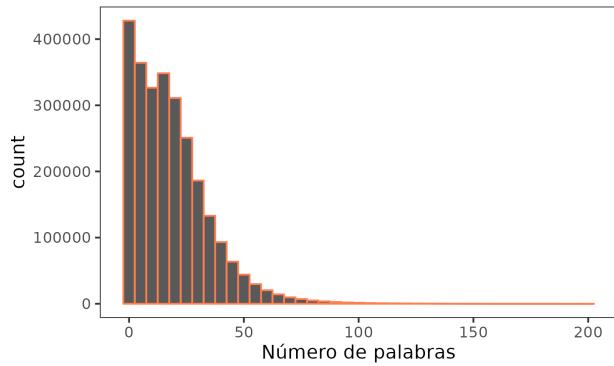
Respecto a los párrafos (unidad de análisis), el cuadro 4 y la figura 2 muestran que el número promedio de palabras es aproximadamente 19 y que, en general, los textos no son demasiado extensos, ya que el 50% tiene 15 palabras o menos y el 90% tiene 39 palabras o menos. El hecho de utilizar una unidad de análisis más desagregada que la intervención, hace más sencilla la identificación en el texto de características distintivas. Estas tienden a desaparecer

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



Nota: Largo de los párrafos luego del procesamiento descrito más arriba.

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

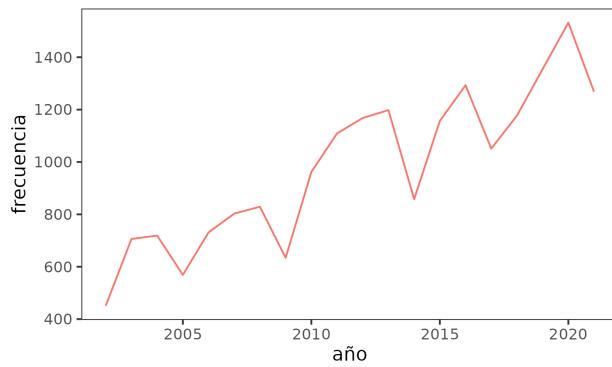
La última fuente de información corresponde a las votaciones en sala de los parlamentarios. Para obtener estos datos se utilizó una API (*Application Programming Interface*) dispuesta por la Cámara de Diputados, mediante la cual fue posible extraer todas las votaciones emitidas en la cámara baja desde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



Nota: Debido al momento en el que se hizo la recolección, solo se obtuvieron datos de las primeras semanas de 2022. Se excluyen dichos datos para no distorsionar la comparaciones entre años.

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

Nota:

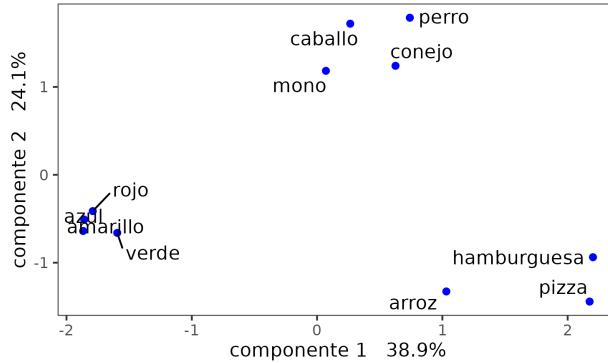
Para encontrar las palabras más cercanas al vector rojo se utiliza similitud coseno

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

⁸Corpus 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>

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.

Figura 4: Agrupación de palabras en un espacio bidimensional



Nota: Dos primeras dimensiones del PCA sobre los vectores correspondientes a las palabras mostradas en el gráfico.

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

$$\text{reina} \approx \text{rey} - \text{hombre} + \text{mujer} \quad (1)$$

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
conde	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 en *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

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:

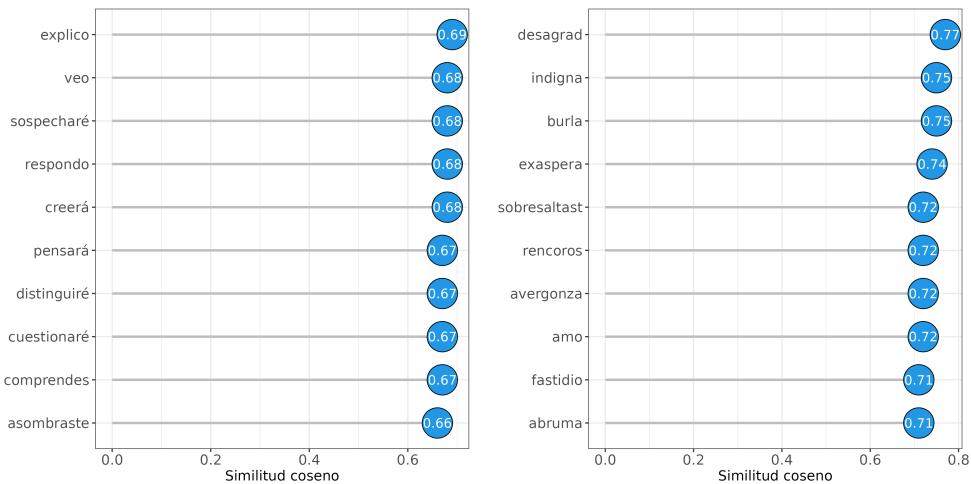
1. Se calcula el centroide de cada una de las matrices
2. Se calcula la similitud coseno de cada uno de los vectores con su respectivo centroide
3. Se ordenan las palabras de menor a mayor similitud
4. Se conserva el 20% de palabras en cada polaridad⁹

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

⁹Para 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.

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.

Figura 5: Palabras del diccionario más representativas de cada polaridad



Nota: Se compara la similitud de cada palabra respecto al centroide, utilizando similitud coseno.

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 193.205¹⁰ 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*, *demonstrar* 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 positivas¹¹.

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

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

Figura 6: Nubes de palabras



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:

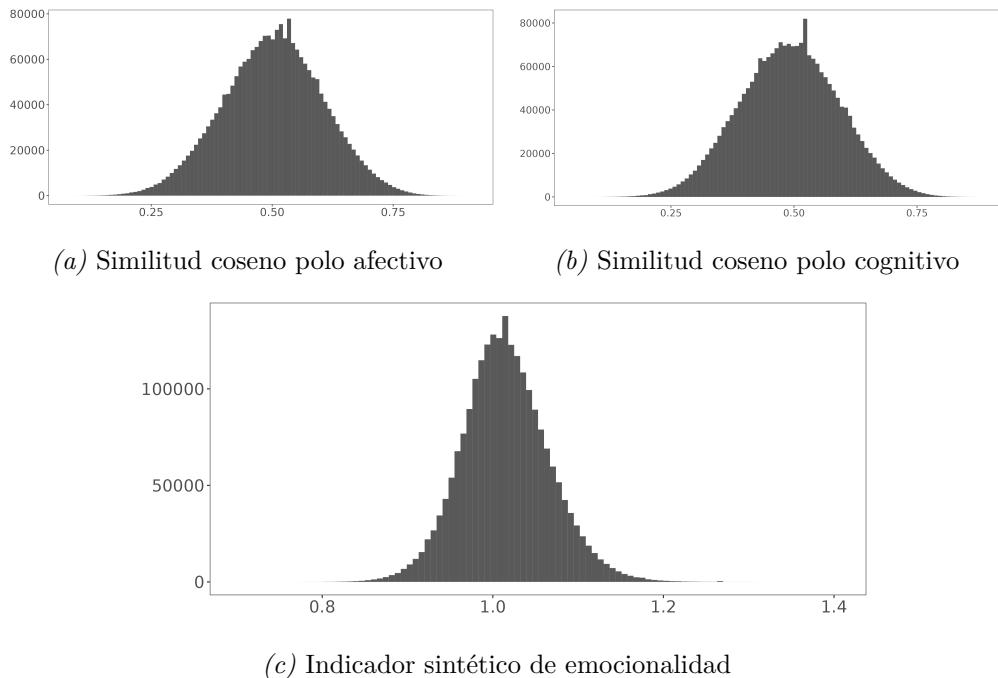
$$Y_i = \frac{\text{sim}(d_i, A) + b}{\text{sim}(d_i, C) + b} \quad (2)$$

Donde A representa al vector del polo afectivo y C , al vector cognitivo. La expresión $\text{sim}(v, w) = (v \cdot w) / (\|v\| \|w\|)$ 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 , $\text{sim}(d_i, A)$ y $\text{sim}(d_i, C)$. 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



Nota: Construido sobre la base de todos los párrafos del corpus

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
<ul style="list-style-type: none">- <i>después fijaremos la fecha exacta.</i>- <i>más adelante analizaré los otros artículos del proyecto.</i>- <i>los numerales de que consta son los siguientes.</i>- <i>eso es lo que tenemos y lo que buscamos modificar.</i>- <i>aquí efectuamos una modificación que señalaré más adelante.</i>- <i>aquí dijimos revisemos esto veamos qué ocurre.</i>- <i>porque cambiamos la ubicación de ese artículo.</i>- <i>¡si no los tenemos ahora ni los tendremos mañana si no llegamos a acuerdo.</i>- <i>entonces votamos y establecemos un plazo para las.</i>- <i>además tenemos proyectos aprobados.</i>- <i>que los pocos instrumentos que tenemos los utilizaremos para llegar a un acuerdo con el gobierno eso haremos.</i>- <i>corresponden a las indicaciones números 233 y 234.</i>- <i>eso lo discutiremos cuando llegue el texto respectivo.</i>- <i>tenemos que analizar cómo lo hacemos para adelante.</i>- <i>vamos a decidir si sacamos o no de la tabla el proyecto.</i>

Cuadro 9: 15 frases más afectivas

text
<ul style="list-style-type: none">- <i>desocupación de los jóvenes un 30 por ciento de la población juvenil.</i>- <i>la sensación de inseguridad en la población.</i>- <i>sea una indolencia brutal total.</i>- <i>sembrando un clima de inquietud de inseguridad de violencia.</i>- <i>entonces más fragmentación más inestabilidad más desgobierno.</i>- <i>para alentarlos al desorden y a la anarquía.</i>- <i>a con esfuerzo físico excesivo.</i>- <i>en consecuencia con el mismo ánimo solidario reflejo mi preocupación por aquello.</i>- <i>acá hay lluvias en demasía y sequías excesivas hay falta de agua en el verano y exceso en el invierno.</i>- <i>f la campaña del terror desatada por los latifundistas.</i>- <i>sin el ánimo de disminuir la importancia de la iniciativa.</i>- <i>además ésa es una inquietud de numerosos sectores de nuestra ciudadanía.</i>- <i>esta situación ha contribuido a la exacerbación y recrudecimiento de otros males socialmente nefastos creando un clima de inseguridad desconfianza y desesperanza.</i>- <i>¡eso y no el boicot es lo que está provocando la escasez de alimentos.</i>- <i>cierta hilaridad es manifestación de nerviosismo.</i>

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

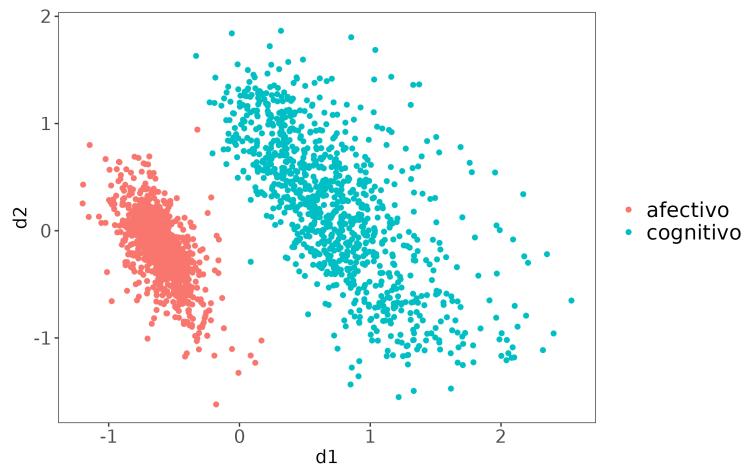
(b) Palabras polo afectivo

Nota: Fueron removidas las *stopwords* 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 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.

Cabe destacar que los textos asociados al polo cognitivo se agrupan de manera más dispersa, lo cual indica que dicha categoría no funciona tan bien como la del polo emotivo, donde se puede observar un mayor nivel de cercanía entre los puntos. Un procedimiento diferente en la etapa de selección de palabras del diccionario, podría mejorar esta clasificación.

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



Nota: Proyección en 2 dimensiones mediante PCA de los 1000 párrafos más cognitivos y afectivos

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 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.

Es importante mencionar que una debilidad de este método, a diferencia de un enfoque basado en *topic modeling* (CITAR), es que sin importar el contenido del texto, el clasificador siempre asignará una clase y hará lo que mejor pueda respecto a lo que haya aprendido en el entrenamiento y a los 11 tópicos seleccionados. Ello quiere decir que en muchos casos la etiqueta puede no coincidir con el texto, pero dado el gran volumen de documentos, se espera que en el agregado funcione razonablemente bien. Esto no sucedería en una estrategia de *topic modeling*, como LDA o LSA, mediante la cual un algoritmo genera tantos grupos sea necesario para optimizar alguna función objetivo. La desventaja radica en que la cantidad de tópicos resultante puede ser muy alta, lo cual implica un arduo trabajo manual para “bautizar” a cada uno de los tópicos, según el criterio del investigador.

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

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

¹³El 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 y DW-NOMINATE. En la actualidad, W-NOMINATE es el más utilizado y, por ende, con implementaciones en lenguajes de programación

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:

$$U_{ijy} = u_{ijy} + \epsilon_{ijy} \quad (3)$$

$$u_{ijy} = \beta \exp\left[\frac{\sum_{k=1}^s w_k^2 d_{ijyk}^2}{2}\right] \quad (4)$$

u_{ijy} representa la parte determinística de la utilidad del legislador, mientras que ϵ_{ijy} representa el componente estocástico. El término d_{ijyk}^2 es la distancia euclíadiana 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:

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

Tanto el peso w como β deben ser estimados, partiendo de valores de 0.5 y 15, respectivamente. w representa la ponderación 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.

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

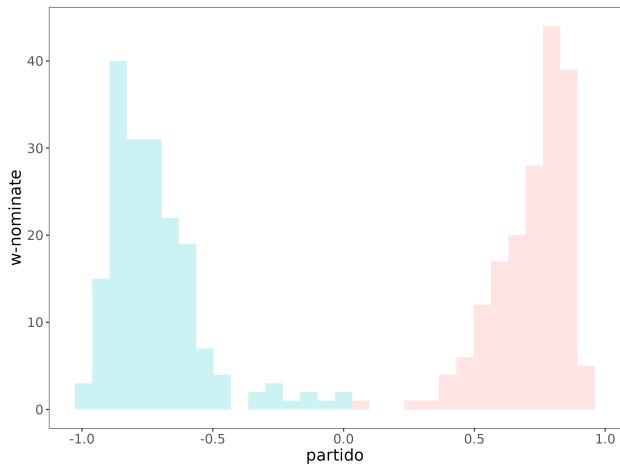


Nota: Construido utilizando las votaciones de la Cámara de Diputados disponibles en el *web service* del sitio del Congreso Nacional.

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

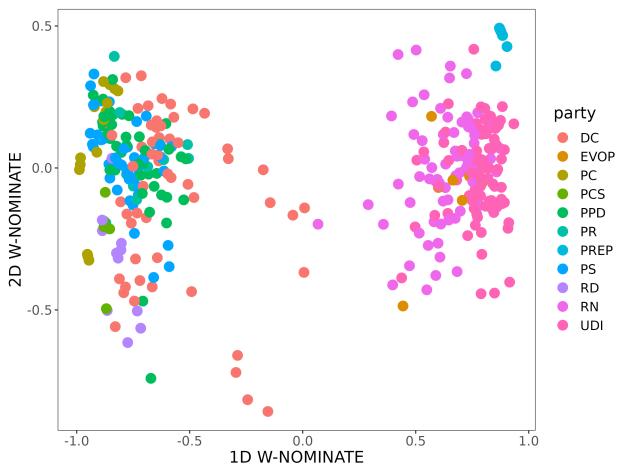
convergencia entre ambas polos, lo cual se explica principalmente por el posicionamiento de los parlamentarios del Partido Demócrata Cristiano. Ello se observa con bastante claridad en la figura 12, donde los puntos rojos corresponde a la posición de los parlamentarios de dicha colectividad en un espacio de dos dimensiones.

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



Nota: Construido utilizando las votaciones de la Cámara de Diputados disponibles en el *web service* del sitio del Congreso Nacional.

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



Nota: Construido utilizando las votaciones de la Cámara de Diputados disponibles en el *web service* del sitio del Congreso Nacional.

5 Resultados

6 Resultados

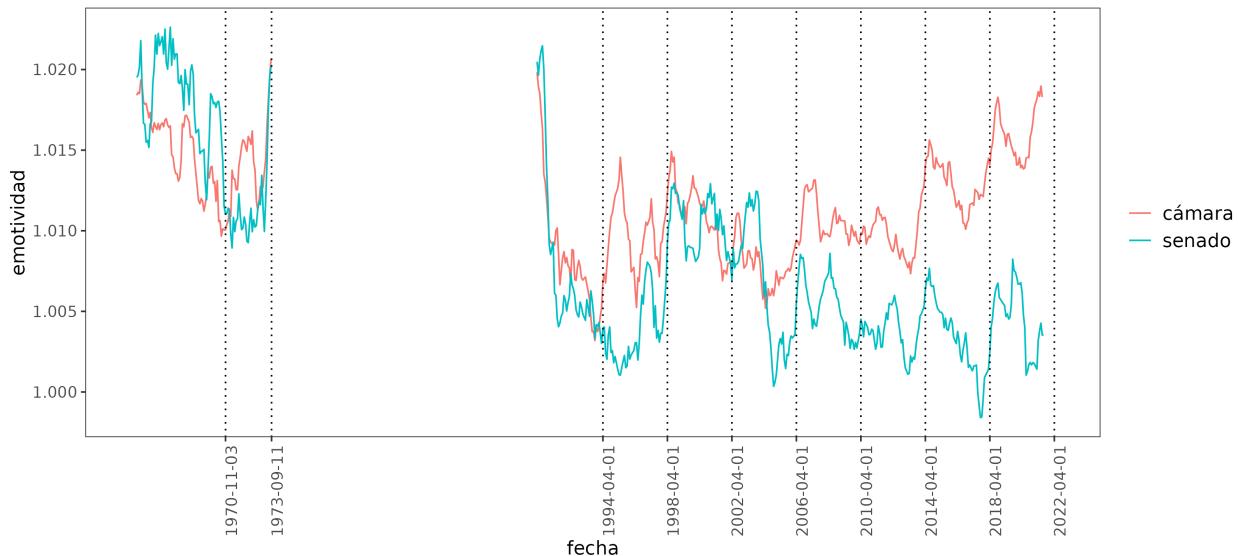
En el presente apartado se presentan los principales resultados obtenidos a partir del análisis de los textos parlamentarios.

6.1 Estadística descriptiva de polos y tópicos

Para tener una primera mirada del fenómeno estudiado es útil observar los datos en perspectiva histórica. El siguiente gráfico muestra los niveles de emotividad desde 1965 hasta 2022 (con una ventana de 17 años). Es posible identificar una serie de fenómenos interesantes:

- 1) Respecto al periodo 1965-1973 se registra en promedio, un nivel más alto de emotividad que en el periodo postdictadura. Entre 1965 y 1970, durante el gobierno de Frei Montalva, se registra una caída del nivel de emotividad en ambas cámaras. La interpretación de dicha caída no es sencilla, ya que al no contar con datos anteriores, no es posible saber con certeza si la disminución se explica por fenómenos sociales del momento o por movimientos estacionales relacionados con los cambios de gobierno. Respecto a esto último, es perfectamente plausible que los niveles de emotividad hayan alcanzado niveles altos al comienzo del gobierno de Frei y hayan comenzado a caer conforme se acercaba el fin del periodo presidencial. De hecho, al comenzar el gobierno de la Unidad Popular también pareciera producirse un aumento de la emotividad, lo cual empuja a pensar en la hipótesis de un comportamiento cíclico. Ahora bien, el aumento notorio de la emotividad registrado en los meses previos a septiembre de 1973, hace pensar en una explicación un poco más coyuntural vinculada a la situación política que vivía el país en aquella época.
- 2) Luego del retorno a la democracia, se registran altos niveles de emotividad, muy cercanos a los que existían en septiembre de 1973. Este nivel desciende rápidamente durante los primeros meses del gobierno de Aylwin.
- 3) Desde el retorno a la democracia, el comportamiento de la emotividad en la Cámara de Diputados presenta un comportamiento cíclico asociado al recambio de parlamentarios. Las líneas punteadas marcan los meses en los que asumen sus funciones los diputados recién electos y, en general, dichos momentos tienden a coincidir con peaks de emotividad, después de lo cual se evidencia una caída.
- 4) La cámara de diputados, a mediados de la primera década de los dosmil, comienza a distanciarse de la trayectoria del Senado hacia niveles más altos de emotividad, acercándose a los niveles existentes en el periodo 1965-1973.

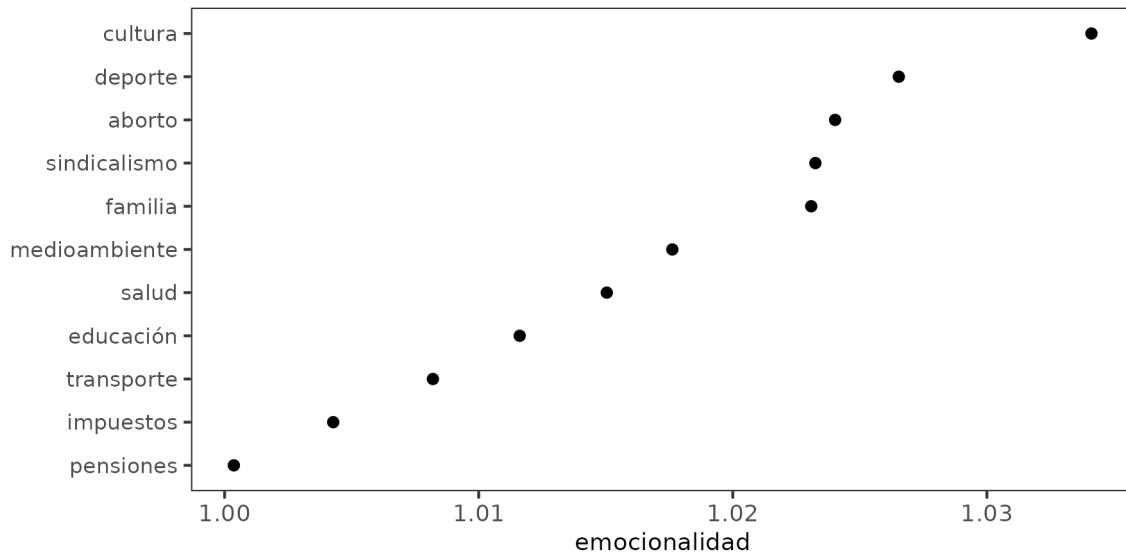
Figura 13: Emocionalidad y cognición a lo largo del tiempo



Nota: El gráfico fue construido con los textos que tienen al menos 4 *tokens*. El formato de las fechas es año-mes-día

En la figura 14 se muestra el promedio de emocionalidad en cada uno de los tópicos detectados. Se puede observar que el ordenamiento de los tópicos guarda relación con las materias que usualmente presentan un carácter más técnico. Así, pensiones, impuestos y transporte toman valores más cercanos al polo cognitivo, mientras que aborto, cultura y deporte, dan cuenta de mayores nivel de emocionalidad.

Figura 14: Emocionalidad y cognición, según tópicos

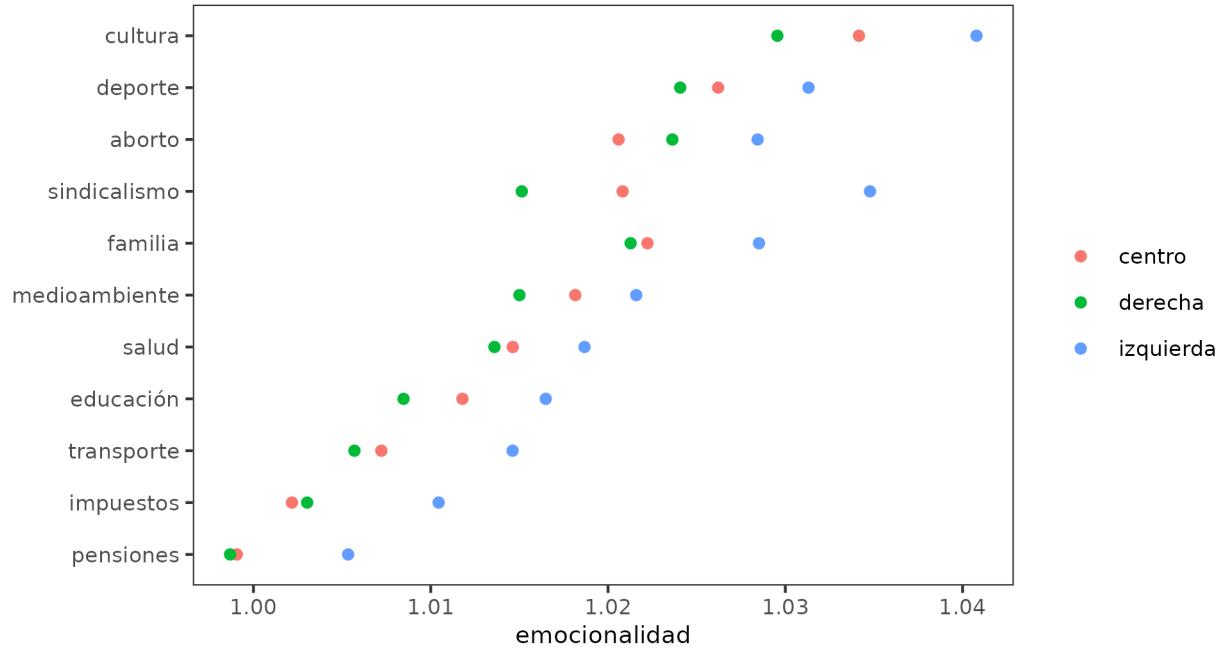


Nota: El gráfico fue construido con los textos que tienen al menos 4 *tokens*. El formato de las fechas es año-mes-día

La figura 15 muestra el promedio de emocionalidad en cada uno de los tópicos detectados, según el posicionamiento político (derecha, izquierda o centro) de los parlamentarios¹⁴. Se puede observar que los discursos emitidos por políticos de centro y derecha no tienen una gran diferencia, mientras que los de izquierda sí se distancian de los demás. De hecho, en todos los tópicos, los discursos ligados a la izquierda son los que muestran un mayor valor de emocionalidad. Es interesante el caso del tópico sindicalismo, cuyo valor en el polo de izquierda se distancia bastante del de los demás, lo cual arroja luces de que este tema sigue siendo sensible para dicho grupo.

¹⁴Construido con base en la militancia de cada parlamentario

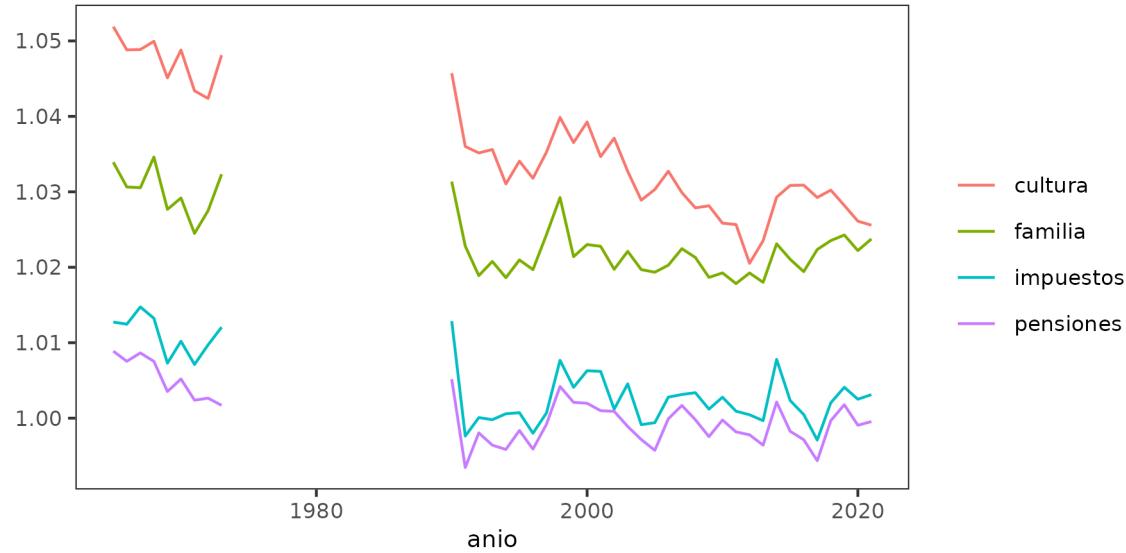
Figura 15: Emocionalidad y cognición a lo largo del tiempo



Nota: fsdfsfsdf

A fin de dar cuenta del nivel de emocionalidad a través del tiempo, la figura 16 muestra una selección de tópicos en cada uno de los extremos. Se puede observar que el nivel de emocionalidad en cada uno de los tópicos no experimenta cambios demasiado notorios. Así, por ejemplo, no se advierte que un tópico con alta emocionalidad, como familia, de pronto, adquiera valores al nivel de pensiones, lo que hace pensar que las materias tratadas en el congreso tienen un comportamiento estructural respecto al nivel de emocionalidad con el aquellas son abordadas.

Figura 16: Emocionalidad y cognición en algunos tópicos a lo largo del tiempo

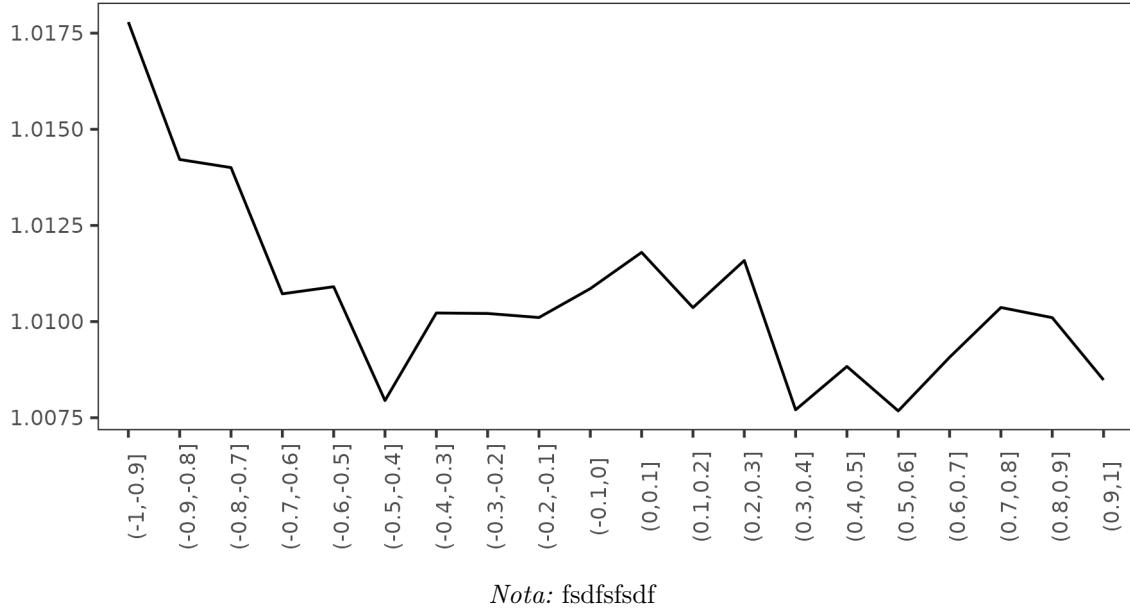


Nota: fsdfsfsdf

Con el objeto de entender un poco mejor la relación entre posicionamiento político y emocionalidad, es posible utilizar los valores de WNOMINATE. Este indicador, al posicionar ideológicamente a todos los políticos desde izquierda a derecha¹⁵, permite estudiar la relación utilizando una variable continua. La figura da cuenta de una cierta relación negativa entre posicionamiento y emocionalidad. Así, parlamentarios posicionados más hacia el extremo izquierdo del espectro presentan un mayor nivel de emocionalidad, la cual va decreciendo conforme el indicador WNOMINATE avanza hacia posiciones más de derecha. Esto contrasta con la evidencia encontrada por Gennaro y Ash (2021) para EEUU, país en el que esta relación presenta una forma de U, de modo que todos los discursos en los extremos están vinculados a altos niveles de emocionalidad y no solo los de izquierda, como se aprecia en la figura 17.

¹⁵El indicador se mueve entre -1 y 1. En el caso de este estudio, valores negativos corresponden a posiciones de izquierda, mientras que valores positivos a posiciones de derecha. Para más detalles, ver el apartado metodológico

Figura 17: Relación entre WNOMINATE y emocionalidad



6.2 Regresiones (PENSAR EN UN MEJOR TÍTULO)

Con el objeto de analizar el efecto de algunas variables en el indicador construido, a continuación se muestran dos grupos de modelos lineales. La diferencia entre ambos grupos dice relación con el número de observaciones consideradas. En el primer grupo (cuadro 10) se incluyen todos los textos para los cuales se cuenta con información de tópicos, es decir, alrededor de 1.2 millones de registros (de un total de 2.4 millones). El segundo grupo de modelos (cuadro 11) contiene una muestra reducida, compuesta únicamente por los registros para los que se cuenta con el valor de WNOMINATE, correspondiente a 344.559 textos.

Para facilitar la interpretación, los valores del indicador de emocionalidad han sido estandarizados. Cabe recordar que valores más altos se encuentran asociados al polo afectivo, mientras que valores bajos se asocian a textos con un carácter más cognitivo. Todas las especificaciones *clusterizan* los errores estándar a nivel de político.

En términos generales, la cuadro 10 muestra relaciones estadísticamente significativas entre los regresores y el indicador de emocionalidad. La única variable sin significancia estadística es edad, por lo cual no se presentan sus coeficientes y solo se incluye como control. La diferencia entre las 4 especificaciones consiste en el tipo de control utilizado para la variable tiempo. En la primera especificación se incorpora un control por efectos fijos de año; en la segunda, dicho control se reemplaza por una variable que indica el año de cada periodo parlamentario; en la tercera especificación se utiliza un control para diferencias entre los periodos pre y post dictadura. Finalmente, la especificación 4 incorpora un de efectos fijos para año-cámara.

Se observa que ser senador (versus diputado) está vinculado a un menor nivel de emotividad

(signo negativo), lo cual coincide con la idea de que la discusión en el Senado tiene un carácter más reflexivo y un poco más distanciado de la política contingente. Otras características que presentan una relación negativa son los tópicos tradicionalmente considerados de carácter más técnico, como impuestos, pensiones y transporte. El signo de los coeficientes para dichos tópicos dan cuenta de una asociación con el polo cognitivo del indicador construido.

Por su parte, las variables que aumentan el nivel de emotividad son sexo (mujer) y una posición de izquierda en el espectro político (comparado con una posición de derecha). Respecto a la categoría “centro” en la variable de posicionamiento político, cabe mencionar que aquella no muestra una diferencia significativa respecto a una posición de derecha.

La segunda especificación incluye un regresor que permite identificar el efecto que tiene el tiempo en cada periodo parlamentario. Dado que en Chile han ocurrido elecciones parlamentarias cada aproximadamente 4 años, es posible clasificar los años desde el 1 hasta el 4, respecto a la elección más cercana. En ese sentido, la categoría 1 representa el primer año del periodo, mientras que la categoría 4 corresponde al último. Tanto la inspección visual, presentada en el apartado anterior, como el signo de los coeficientes muestra que el nivel de emotividad tiende a caer conforme avanzan los periodos parlamentarios. Ello hace suponer la existencia de un comportamiento cíclico a través del cual las intervenciones en el congreso adquieren tonos más afectivos y cognitivos, dependiendo del proceso político.

Otro aspecto a considerar sobre la segunda especificación es la pérdida de significancia estadística de la variable sexo que se produce al reemplazar el control de efectos fijos de año por el del periodo parlamentario. Dicha significancia retorna en la especificación 3, en la cual se agrega un control que separa los datos en periodos pre y postdictadura. Posiblemente, esto pueda deberse a la baja presencia de mujeres en el periodo 1965-1973, la cual comienza a aumentar con el retorno a la democracia, de modo tal que al incluir el control por estos dos periodos, se hace visible el efecto de la variable sexo. Por su parte, el signo positivo de la variable pre-post dictadura indica que el periodo 1965-1973 tiene un efecto positivo sobre el indicador. En ese sentido, los gráficos de la sección anterior que mostraban valores más altos en la etapa predictadura encuentran un correlato en el coeficiente de la regresión.

La especificación número 4 incluye un control por efectos fijos año-cámara, con el objeto de identificar si las trayectorias de cada una de las cámaras está relacionada con el indicador de emocionalidad. Al incluir este control, desaparece el efecto de la variable cámara, lo cual confirma que existen trayectorias diferenciadas (PREGUNTAR ACÁ).

Cuadro 10: Especificaciones sin WNOMINATE

variable	modelo 1	modelo 2	modelo 3	modelo 4
sexo	0.076*** (0.029)	0.049 (0.031)	0.073** (0.029)	0.083*** (0.026)
cámara	-0.077*** (0.017)	-0.063*** (0.02)	-0.078*** (0.018)	0.016 (0.039)
Tópicos				
tópico aborto	0.266*** (0.021)	0.257*** (0.02)	0.263*** (0.02)	0.262*** (0.021)
tópico cultura	0.464*** (0.01)	0.466*** (0.01)	0.468*** (0.01)	0.461*** (0.01)
tópico deporte	0.311*** (0.013)	0.306*** (0.013)	0.308*** (0.013)	0.31*** (0.013)
tópico familia	0.238*** (0.007)	0.239*** (0.008)	0.241*** (0.008)	0.236*** (0.007)
tópico impuestos	-0.157*** (0.009)	-0.148*** (0.01)	-0.155*** (0.009)	-0.158*** (0.009)
tópico medioambiente	0.135*** (0.01)	0.13*** (0.01)	0.136*** (0.01)	0.134*** (0.01)
tópico pensiones	-0.238*** (0.007)	-0.23*** (0.007)	-0.238*** (0.007)	-0.237*** (0.007)
tópico salud	0.076*** (0.01)	0.073*** (0.01)	0.076*** (0.01)	0.071*** (0.01)
tópico sindicalismo	0.206*** (0.012)	0.236*** (0.014)	0.211*** (0.013)	0.206*** (0.012)
tópico transporte	-0.081*** (0.008)	-0.069*** (0.009)	-0.08*** (0.008)	-0.08*** (0.008)
Posicionamiento político				
centro	-0.001 (0.02)	0.029 (0.02)	0.002 (0.02)	0.001 (0.019)
izquierda	0.112*** (0.027)	0.165*** (0.03)	0.114*** (0.028)	0.107*** (0.026)
Periodo parlamentario				
segundo año		-0.035*** (0.008)		
tercer año		-0.057*** (0.008)		
cuarto año		-0.054*** (0.007)		
periodo predictadura			0.162*** (0.024)	
efectos fijos año	Sí	No	No	No
efectos fijos año-cámara	No	No	No	Sí
observaciones	1.217.260	1.217.260	1.217.260	1.217.260

Nota:

La variable dependiente corresponde al indicador de emocionalidad, cuyos valores han sido estandarizados. Valores elevados indican mayor emocionalidad. Todas las especificaciones clusterizan los errores a nivel de político e incluyen un control de edad. Las categorías de referencia para las variables sexo, cámara, tópicos, posicionamiento político y periodo son las siguientes: mujer, diputado, educación, derecha y predictadura, respectivamente. *, ** y *** denotan 90%, 95% y 99% de confianza.

El cuadro 11 contiene especificaciones muy similares a las ya comentadas. La diferencia radica en que estos modelos incluyen polarización política (WNOMINATE). La manera de incorporar esta información ha sido mediante la llave persona-año, es decir, se ha calculado de forma independiente para cada año el valor de WNOMINATE. Ello quiere decir que las intervenciones parlamentarias se vinculan al posicionamiento ideológico que los políticos tenían en el momento en que aquellas ocurrieron.

Es importante recordar que los valores de WNOMINATE se mueven entre -1 y 1, donde valores negativos representan posiciones más a la izquierda y valores positivos más a la derecha. Para poder interpretar los valores de WNOMINATE como una medida de polarización, estos se elevan al cuadrado. De esta manera, números altos indicarán mayor nivel de polarización. Al realizar esta operación se pierde la noción de posicionamiento ideológico, sin embargo, dicha información ya está siendo incluida a través de la variable polo.

Tanto en la especificación que incluye efectos fijos por año como en la que no, se observa una relación positiva en la variable WNOMINATE. Ello quiere decir que una mayor polarización ideológica estaría asociada a mayores niveles de emotividad en el discurso de los parlamentarios. Esta evidencia encuentra respaldo en una línea de investigación (CITAR) que da cuenta de que los políticos que están en los extremos del espectro ideológico recurren a un recurso basado en argumentación por la vía de la emotividad.

Cuadro 11: Especificaciones con WNOMINATE

variable	modelo 1	modelo 2
sexo	0.087*** (0.029)	0.094*** (0.029)
w2 standard	0.021** (0.009)	0.019** (0.008)
Tópicos		
tópico aborto	0.176*** (0.034)	0.193*** (0.034)
tópico cultura	0.352*** (0.014)	0.354*** (0.013)
tópico deporte	0.317*** (0.023)	0.314*** (0.023)
tópico familia	0.192*** (0.012)	0.192*** (0.012)
tópico impuestos	-0.217*** (0.017)	-0.218*** (0.017)
tópico medioambiente	0.08*** (0.017)	0.083*** (0.017)
tópico pensiones	-0.241*** (0.011)	-0.242*** (0.011)
tópico salud	0.105*** (0.015)	0.112*** (0.015)
tópico sindicalismo	0.078*** (0.016)	0.078*** (0.016)
tópico transporte	-0.094*** (0.015)	-0.093*** (0.015)
Posicionamiento político		
centro	0.029 (0.028)	0.025 (0.029)
izquierda	0.13*** (0.032)	0.156*** (0.033)
efectos fijos año	Sí	No
observaciones	344.559	344.559

Nota:

La variable dependiente corresponde al indicador de emocionalidad, cuyos valores han sido estandarizados. Valores elevados indican mayor emocionalidad. Todas las especificaciones clusterizan los errores a nivel de político e incluyen un control de edad. La variable WNOMINATE ha sido elevada al cuadrado. Las categorías de referencia para las variables sexo, tópicos y posicionamiento político son: mujer, educación, derecha, respectivamente. *, ** y *** denotan 90%, 95% y 99% de confianza.

7 Conclusiones

El presente trabajo explora la relación que existe entre polarización ideológica y emotividad en los discursos parlamentarios, mediante herramientas de procesamiento de lenguaje natural. Explotando un nuevo set de datos de discursos parlamentarios, se aporta evidencia sobre un tema escasamente abordado en el contexto chileno.

Dado que esta es la primera investigación que estudia el fenómeno de la emocionalidad en los discursos parlamentarios chilenos, se ha seguido un enfoque exploratorio, que sin ser totalmente concluyente, entrega evidencia interesante respecto al fenómeno estudiado. La conclusión más relevante es que existe una vinculación entre la polarización ideológica de los parlamentarios y el nivel de emocionalidad de sus intervenciones en el Congreso. Así, políticos ubicados en posiciones más extremas tienden a utilizar discursos con un cariz más cercano al polo emotivo que al cognitivo.

Cabe destacar que a diferencia de lo encontrado en (XXX), de manera gráfica, no existe una relación en forma de U entre polarización ideológica y emotividad. En lugar de ello, se advierte que valores posiciones más de izquierda están asociadas a una mayor nivel de emotividad.

Es un hecho conocido que la discusión política contiene una dimensión emotiva, lo cual no constituye un hecho negativo en si mismo, sin embargo, es relevante conocer cuál es el balance entre emotividad y cognición, ya que el espacio de deliberación política debiese estar guiado principalmente por consideraciones racionales. No es irrelevante el hecho de que posiciones más extremas utilicen como recurso argumentativo la emocionalidad. En ese sentido, una línea de investigación interesante sería una que estudie cuál es el nivel de emocionalidad en los discursos parlamentarios de temas que actualmente generan polarización, como la inmigración y la seguridad pública.

Una segunda conclusión derivada de este estudio es que existe un correlato entre los grandes procesos políticos que ha experimentado el país y el nivel de emocionalidad presente en los discursos parlamentarios. Tanto el quiebre institucional de 1973 como el retorno a la democracia presentan altos niveles de emocionalidad, lo que muestra que existe una conexión entre lo que ocurre en la sociedad y el espacio de poder institucional. Es interesante constatar que los niveles de emocionalidad, en promedio, son más altos en las décadas de los sesentas y setentas, que en el periodo postdictadura. De alguna manera, la efervescencia que existía en la discusión política de aquellos años, caracterizada por el enfrentamiento de proyectos antagónicos de sociedad, se manifestaba en los discursos parlamentarios. Lo contrario sucede con el retorno a la democracia, momento en el cual las tensiones tienden a disminuir, lo que se refleja en menores niveles de emocionalidad. Esta situación podría estar cambiando desde hace algunos años, ya que los datos muestran una modificación en la trayectoria de la emocionalidad, hacia mayores niveles de emocionalidad, al menos en lo que refiere a la Cámara de Diputados.

Una tercera y última conclusión es que los ciclos del proceso político afectan los niveles de emocionalidad del Congreso. En períodos cercanos a elecciones parlamentarias el nivel de emocionalidad muestra aumentos. Pasado el primer año, el nivel comienza a bajar, lo cual

da cuenta de que el fenómeno en estudio tiene una dimensión cíclica y otra de más largo plazo, vinculada esta última a procesos políticos e históricos de carácter más estructural.

Desde el punto de vista metodológico, este trabajo pone a disposición un set de datos semi estructurado con más de 500.000 intervenciones y biografías parlamentarias, que contiene información desde 1965 a 2022. Cabe destacar que el enfoque exploratorio de esta investigación deja un amplio margen para futuros estudios que puedan sacar más provecho del set de datos construido, de modo que los textos reunidos pueden seguir siendo explotados para entender mejor la historia política reciente de nuestro país.

La estrategia escogida se propone incorporar herramientas de NLP, ampliamente utilizadas en la industria y en la ciencia de datos, pero cuya adopción en los campos de la economía y sociología es aún incipiente. Tradicionalmente, las ciencias sociales han estado alejadas de estrategias de investigación basadas en métodos computacionales, pero durante los últimos años se ha producido un cierre de esta brecha tecnológica, lo cual abre un espacio muy prometedor para que la economía y la sociología exploten grandes volúmenes de información no estructurada, ya sea para revisitar viejos temas o para explorar ámbitos nuevos.

Referencias

- Alemán, Eduardo. 2008. «Policy Positions in the Chilean Senate: An Analysis of Coauthorship and Roll Call Data». *Brazilian Political Science Review (online)* 3 (diciembre). <https://doi.org/10.1590/S1981-38212008000100008>.
- ALEMÁN, EDUARDO. 2009. «Institutions, Political Conflict and the Cohesion of Policy Networks in the Chilean Congress, 1961–2006». *Journal of Latin American Studies* 41 (3): 467-91. <https://doi.org/10.1017/S0022216X09990150>.
- 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>.
- Fábrega, Jorge. 2022. «Ordenamiento Ideológico en la Convención Constitucional Chilena». *Revista de Ciencia Política* 42 (1): 127-51. <https://doi.org/10.4067/s0718-090x2022005000106>.
- 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>.
- Morán, Carmen Le Foulon. 2020. «Cooperation and polarization in a Presidential Congress: Policy networks in the Chilean Lower House 2006–2017». *Politics* 40 (2): 227-44. <https://doi.org/10.1177/0263395719862478>.
- 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>.
- Saiegh, Sebastian, y Eduardo Alemán. 2007. «Legislative Preferences, Political Parties, and

- Coalition Unity in Chile». *Comparative Politics* 39 (septiembre). <https://doi.org/10.2307/20434040>.
- 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](https://doi.org/10.1007/978-3-319-73004-2).