MÉTODOS AUTOMATIZADOS DE ANÁLISIS DE CONTENIDO APLICADO A DISCURSOS PARLAMENTARIOS

Ricardo Modesto Vieira[1]

Resumen: al reducir sustancialmente el tiempo y el costo de la búsqueda utilizando textos, los métodos automatizados de análisis de contenido han permitido investigar cuantitativamente grandes colecciones, comodiscursos parlamentarios. Este artículo identifica el marco de conceptos de estanuevafrontera del conocimiento, describe un proyecto de texto típico como se presenta, presenta las principales metodologías y sus aplicaciones a los discursos parlamentarios, muestra los principales desafíos y las mejores prácticas en estaárea de investigación, y finalmente sugiere como aplicar estos métodos a las colecciones de la Casa Legislativa. La intención no es agotar la gama de métodos, técnicas y modelos; pero brindar una guía para que estos métodos, cada vez más populares entre los científicos políticos y sociales, se apliquen adecuadamente a los discursos parlamentarios y las cámaras legislativas.

Palabra clave: análisis de contenido automatizado; texto como se da; Análisis del discurso parlamentario.

INTRODUCCIÓN

La revolución de los grandes datos y la inteligencia artificial proporcionaron grandes oportunidades para que las Cámaras Legislativas ahorren trabajo, como a través de la indexación automática, generen información estratégica, como el posicionamiento ideológico de las partes, y proporcionen ideas importantes, como las de los investigadores Moreira (2016), Izumi (2017) y Schwartz (2018), a través del análisis de contenido automatizado aplicado al discurso parlamentario[2]. Este artículo tiene como objetivo demostrar como los discursos pueden analizarse cuantitativamente con la ayuda de herramientas computacionales, a fin de facilitar la organización y comprensión de esta vasta colección, así como generar indicadores e información estratégica para las Cámaras Legislativas, sus Mesas Directivas y sus Liderazgos.

Los parlamentarios expresan sus opiniones, defienden sus posiciones y hacen proposiciones en palabras. Según el enfoque empírico adoptado, lo que importa es la probabilidad de que aparezcan palabras y expresiones "comunes" e "inusuales" como eventos lingüísticos (MANNING; SCHUTZE, 1999, p. 7). Por lo tanto, a través de la frecuencia con la que estas palabras y expresiones se manifiestan, es posible describir el uso

[1]Titular de Maestría en Ciencias Políticas por la San Diego State University, Estados Unidos. Correo electrónico: ricardomodestovieira@gmail.com

[2]No se trata de las líneas francesas del Análisis del Discurso (cuyas referencias son Michel Pêcheux y Dominique Maingueneau) o el Análisis de Contenido (cuya referencia es Laurence Bardin), sino de la línea estadounidense de procesamiento del lenguaje natural, cuyas referencias son George Kingsley Zipf y Alan Turing. Así, la visión empirista de estos dos autores difería de la lingüística racionalista de Chomsky. En este artículo, el enfoque no está en el significado del discurso o su contenido, sino en el uso de técnicas cuantitativas automatizadas aplicadas a un contenido específico: discursos parlamentarios y debates. El término "discursos y debates" fue definido oficialmente por la Cámara de Diputados como adecuado para referirse tanto al discurso del orador solo en la tribuna, así como en los pequeños y grandes despachos, así como a la discusión de asuntos en la pauta.

del lenguaje en las Cámaras Legislativas. Por supuesto, este enfoque estadístico siempre estará condenado a varios errores y nunca reemplazará al componente humano; sin embargo, con el avance de las herramientas computacionales, los métodos automatizados de análisis de contenido se han convertido en herramientas útiles para explorar una vasta colección, como los discursos parlamentarios.

En este sentido, se analizarán diez métodos diferentes: cuatro métodos de clasificación, que incluyen dos modelos de tema, tres métodos de escala y tres métodos de similitud de texto, y sus aplicaciones para las Cámaras Legislativas. Todos estos métodos utilizan la "bolsa de palabras" como una suposición: lo importante es la frecuencia con la que se encuentran ciertas palabras en el corpus o la colección. Como no existe un método global para el análisis de contenido automatizado (GRIMMER; STEWART, 2013), se eligió su uso en la literatura de Ciencias Políticas y su oportunidad de aplicación a las Cámaras Legislativas en vista de la variabilidad de los posibles propósitos.

Para comprender mejor estas diez técnicas diferentes y sus posibles aplicaciones, primero es necesario explicar los fundamentos del procesamiento del lenguaje natural, así como las técnicas de preprocesamiento de datos que permiten el análisis de contenido automatizado. Después de explicar el preprocesamiento y describir las técnicas y sus aplicaciones, se analizarán los desafíos más comunes relacionados con estos métodos de análisis automatizado aplicados al discurso parlamentario, así como las mejores prácticas recomendadas en la literatura para abordar estos desafíos. Finalmente, se sugerirán algunas posibles aplicaciones de estos métodos a las Cámaras Legislativas.

FUNDAMENTOS Y PREPROCESAMIENTO DE DATOS

En un enfoque empírico, el cerebro humano parte de asociaciones, reconocimiento de patrones y generalizaciones para aprender la estructura detallada del lenguaje natural. Por lo tanto, no existe una facultad innata del lenguaje como herencia genética. No hay diferencia entre el dominio del idioma (conocimiento de la estructura del lenguaje) y el desempeño lingüístico de uno en el mundo. Dado que el lenguaje es inseparable de su contexto social, lo que importa son los patrones comunes que ocurren en el uso del lenguaje (MANNING; SCHUTZE, 1999, p. 5-6).

En este sentido, las palabras y frases se consideran "usuales" e "inusuales", y la frecuencia de los tipos "usuales" demuestra las preferencias que se producen en el uso del lenguaje (MANNING; SCHUTZE, 1999, p. 9). A través de un enfoque estadístico, es posible aprender automáticamente estas preferencias léxicas y las palabras y expresiones que tienden a agruparse y formar su propio campo léxico. Debido a que el significado de las palabras/expresiones está relacionado con el contexto en el que se usan, este conocimiento de las preferencias léxicas se puede explorar para comprender relaciones semánticas más profundas (MANNING; SCHUTZE, 1999, p. 18-19).

Siguiendo este enfoque estadístico, Zipf (1949) descubrió que existe una relación entre la frecuencia de una palabra y su rango en una lista de palabras en el mismo corpus[3]. Esta constante es útil como una descripción aproximada de la distribución de frecuencia de las palabras del lenguaje humano: hay algunas palabras muy comunes, un número intermedio de palabras de frecuencia intermedia y muchas palabras de baja frecuencia[4]. Zipf también obtuvo evidencia empírica de la tendencia de las palabras que pertenecen al mismo contenido a agruparse (MANNING; SCHUTZE, 1999, p. 24-25).

Por lo tanto, para reducir la complejidad y el tamaño del vocabulario, así como para centrarse en lo que es habitual y significativo en el texto, solo se analizan las palabras de frecuencia intermedia. Por lo tanto, es necesario eliminar palabras innecesarias e infrecuentes: aquellas que aparecen en el 99% o menos del 1% de los documentos (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018, citando HOPKINS; KING, 2010; QUINN y col., 2010). Las palabras más comunes generalmente no generan contenido significativo y corresponden a conjunciones, preposiciones, artículos, pronombres y verbos como vinculación.

Cuando no se tiene en cuenta el orden en que aparecen las palabras en el texto, se asume el principio de la bolsa de palabras (BLEI, 2012; GRIMMER; STEWART, 2013). El principio de la bolsa de palabras presenta cada documento como un solo vector con la longitud igual al número de palabras únicas en el texto (ROBERTS y col., 2015; IZUMI; MOREIRA, 2018). Para restaurar cierta importancia al orden de las palabras, uno puede usar bigramas, trigramas (GRIMMER; STEWART, 2013) o colocaciones[5].

También es importante enfatizar la necesidad de eliminar lo que en la literatura se llama palabras vacías (GRIMMER; STEWART, 2013; ROBERTS y col., 2015; CASAS; WILKERSON, 2017). Cada contexto genera una lista diferente de palabras vacías, es decir, palabras, frases y expresiones que se usan con mucha frecuencia en un contexto particular, pero que no generan contenido significativo. Lauderdale y Herzog (2016), por ejemplo, retiraron todos los discursos procesales del proceso legislativo, como los discursos del Presidente, la lectura de las actas y la agenda, las elecciones de la Mesa, las oraciones y los homenajes.

Después de eliminar las palabras vacías, es necesario reducir la variabilidad de las palabras mediante la derivación o la lematización. La derivación es la reducción de la palabra a su radical eliminando su final, como en plurales o conjugaciones verbales. De hecho, la derivación es una aproximación de un concepto lingüístico llamado lematización, que busca reducir las palabras a

[3]Un conjunto de textos se llama *corpus*. Varias de estas colecciones de texto se denominan *corpora* (MANNING; SCHUTZE, 1999, p. 6).

[4]Debido a que la línea es demasiado baja para la mayoría de las calificaciones bajas y demasiado alta para las calificaciones superiores a 10.000, Mandelbrot obtuvo una relación más precisa entre la clasificación y la frecuencia, utilizando otros tres parámetros de texto como variables (MANNING; SCHUTZE, 1999, p. 25).

[5]Una colocación es una expresión de uso común donde el todo es mayor que la suma de las partes. Cualquier expresión que las personas repitan es candidata a colocación. Los dos patrones de colocación de palabras más comunes son "sustantivo-adjetivo" y "sustantivo-sustantivo" (MANNING; SCHUTZE, 1999, p. 29-31).

su forma básica y agruparlas usando un algoritmo más complejo que identifica el origen de la palabra y devuelve solo su lema o raíz (GRIMMER; STEWART, 2013; ROBERTS y col., 2015; CASAS; WILKERSON, 2017). En portugués, se utiliza una adaptación del algoritmo de Porter (1980) para la derivación (IZUMI; MOREIRA, 2018).

Después de la derivación, las ocurrencias individuales de cada palabra se denominan tokens (MANNING; SCHUTZE, 1999, p. 22) y el contenido del documento finalmente está listo para convertirse en datos cuantitativos. Vale la pena recordar que antes del preprocesamiento es necesario obtener[6], codificar[7] y tratar[8] los documentos. La bolsa de palabras es el método más simple y más utilizado para convertir cada documento en un solo vector cuyo valor está determinado por la ausencia o presencia de un token en el documento, la frecuencia de estos tokens o la frecuencia normalizada por el tamaño del documento (DIERMEIER al al. ., 2012). El objetivo es crear una matriz de documentos y términos (Document Term Matrix o DTM) en el que cada fila representa un documento y cada columna representa un token único. Dado que cada celda de la matriz indica el número de veces que el token indicado en la columna aparece en el documento indicado en la fila, cada documento está representado por un vector único (ROBERTS y col., 2015; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

MÉTODOS Y APLICACIONES

Después de obtener, codificar, procesar y preprocesar los discursos parlamentarios, es posible utilizar varios métodos de análisis de contenido automatizado para diferentes propósitos, como ahorrar trabajo, generar información estratégica y hacer descubrimientos científicos. Estos métodos se pueden dividir en tres conjuntos, de acuerdo con los objetivos y tareas que pretenden lograr: (1) clasificación; (2) dimensionamiento y (3) similitud entre textos. El primero propone clasificar textos o documentos en categorías conocidas o desconocidas; el segundo propone estimar el lugar de los actores usando una escala; y el tercero propone medir la similitud y/u homogeneidad de textos o partes de estos textos.

También hay dos enfoques diferentes: métodos supervisados y no supervisados. La principal diferencia entre estos enfoques es que, en supervisado, es necesario especificar de antemano la estructura conceptual de los textos, mientras que en el no supervisado se usa un modelo para

[6]Los documentos se pueden obtener de varias maneras: (1) datos abiertos en notación de objetos de script Java (JSON), lenguaje de marcado extensible (XML), valores separados por comas (CVS), etc.; (2) documentos de diversos tipos convertidos en datos editables mediante reconocimiento óptico de caracteres (OCR); (3) métodos de raspado de datos web, en los que la computadora accede a páginas web, copiando y organizando su contenido; (4) plataformas en línea como el Mechanical Turk de Amazon, y aplicaciones (API) que permiten solicitar contenido seleccionado de una base de datos (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

[7]La codificación es la forma en que la computadora traduce caracteres individuales y únicos en bytes para que la máquina pueda leer el texto (ROBERTS y col., 2015; IZUMI; MOREIRA, 2018). Para los caracteres latinos, se utiliza la codificación UTF-8 (GRIMMER; STEWART, 2013).

[8]Se requieren documentos que estén formateados de manera similar. El formato idéntico permite escribir un solo script para extraer contenido más específico de múltiples documentos a la vez (CASAS; WILKERSON, 2017).

encontrar un resumen de baja dimensionalidad que mejor explique los documentos observados , y sólo es necesario informar de antemano el número de categorías (ROBERTS y col., 2015). La importancia de esta dimensión que el método recupera en el caso de los discursos parlamentarios depende del contexto político, ya que las preferencias y motivaciones para el discurso varían según este contexto.

Sin embargo, los resultados de la baja dimensionalidad de los indicadores elaborados a través del análisis de contenido automatizado han sido reconocidos como una característica importante de la toma de decisiones del Congreso, especialmente cuando se considera que el uso de palabras en el debate político también varía de acuerdo con el tema debatido[9] y el tipo de escala política o ideológica estimada (LAUDERDALE; HERZOG, 2016). Por lo tanto, la baja dimensionalidad en el Congreso implica, por ejemplo, que votar por un representante en un tema será un buen predictor de su elección de voto en otro tema no relacionado (DIERMEIER at al., 2012).

1.1 Métodos de clasificación

Los métodos de clasificación pueden ser supervisados y no supervisados. Los métodos de aprendizaje supervisados utilizan la frecuencia con la que aparecen las palabras en el texto para clasificar documentos en categorías predeterminadas o para medir el grado en que los documentos pertenecen a categorías específicas. El algoritmo luego "aprende" como clasificar los documentos en estas categorías utilizando un conjunto de entrenamiento. Es decir, el algoritmo usa las características del documento para clasificarlas en categorías. Debido a que existen estadísticas claras que resumen el desempeño del modelo, los métodos supervisados son más fáciles de validar (GRIMMER; STEWART, 2013).

Dentro de los métodos supervisados, aquellos que previamente requieren la identificación de las palabras que separan las clases se denominan ‘‘métodos de diccionario’’. Los métodos de diccionario utilizan la frecuencia relativa de las palabras clave para medir la presencia de cada categoría en el texto. Utilizando una lista de palabras asociadas con el tono de voz (diccionario anotado), así como la frecuencia relativa en la que aparecen estas palabras clave, es posible medir el tono de un documento (GRIMMER; STEWART, 2013). Sin embargo, el análisis de sentimientos, otra área importante de la investigación de clasificación, donde el objetivo es ordenar el texto ordinariamente (de negativo a positivo, por ejemplo) en lugar de categóricamente, puede emplearse utilizando métodos supervisados y no supervisados (CASAS; WILKERSON, 2017).

Los modelos de asociación mixta no supervisados o los modelos de temas son un conjunto de modelos generativos bayesianos que codifican la estructura específica del problema en una

[9]La asociación política para ciertas palabras depende del debate en el que se usaron esas palabras (dado que una palabra que implica una posición de izquierda en un debate puede implicar una posición de derecha en otro debate), pero también algunas palabras se usan de manera similar para denotar el posicionamiento en muchos debates.

estimación de categoría (GRIMMER; STEWART, 2013). En otras palabras, son un conjunto de algoritmos que tienen como objetivo descubrir e identificar documentos con información temática (BLEI, 2012). Los algoritmos analizan palabras de texto para encontrar temas, como estos temas están conectados entre sí y como cambian con el tiempo. Los algoritmos no tienen información sobre estos temas y los documentos no están etiquetados con temas o palabras clave (BLEI, 2012). Dado que estadísticamente un tema es una función probabilística de las palabras, para estimar un tema, los modelos usan la coincidencia de palabras entre los documentos (GRIMMER; STEWART, 2013).

Por lo tanto, estos métodos no supervisados utilizan características de texto sin imponer categorías predeterminadas, utilizando solo suposiciones de modelado y propiedades de texto para estimar un conjunto de categorías y asignar simultáneamente documentos (o partes de documentos) a estas categorías (GRIMMER; STEWART, 2013).) Las distribuciones de temas interpretables surgen al calcular la estructura oculta que probablemente generó la colección de documentos observados. Por lo tanto, el proceso que genera los temas define una probabilidad de distribución en relación con variables aleatorias observables y no observables (BLEI, 2012).

Los métodos de clasificación y las plantillas de temas se pueden usar para una variedad de propósitos. Estos métodos se pueden usar para estudiar cuestiones legislativas, temas y agendas, como comprender cómo cambian las agendas políticas con el tiempo (DIERMEIER en al., 2012). Los métodos supervisados pueden servir para la categorización e indexación automática de documentos, ahorrando trabajo, así como el análisis de sentimientos y el posicionamiento de los actores políticos. Los métodos de clasificación no supervisados pueden responder preguntas sobre el funcionamiento interno del gobierno, la influencia de diferentes grupos políticos y actores en la formulación de políticas públicas, el posicionamiento político de los actores, etc. (ROBERTS y col., 2015).

Ejemplos prácticos recientes incluyen Moreira (2016), quien investigó la relación gobierno-oposición basada en discursos parlamentarios a través del modelo de tema del Modelo de Agenda Expresada y descubrió que el énfasis temático de los diputados no sigue la relación verificada gobierno-oposición en el contexto de los votos nominales; Izumi (2017), quien estimó las posiciones políticas en el Senado a través del clasificador NaiveBayes (análisis de sentimientos) aplicado a los discursos del Senado; y Diermeierat al. (2012), que utilizaron máquinas de vectores de soporte (SVM) junto con la base de datos de votación nominal para predecir la posición ideológica de los parlamentarios, así como para medir el grado de cohesión dentro del partido en el Senado. En comparación con NaiveBayes, los resultados de Diermeier demuestran que el SVM es superior para clasificar el posicionamiento ideológico.

1.1.1 NaiveBayes

El NaiveBayes es uno de los métodos de clasificación supervisada más utilizados. Aunque parte de una suposición ingenua, el modelo supone que las palabras se generan independientemente para una categoría dada (*thenaiveassumption*), cuando en realidad el uso de las palabras está altamente correlacionado en cualquier conjunto de datos, el modelo proporciona un método alternativo útil para asignar documentos a categorías predeterminadas (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). Para colecciones grandes, el clasificador DecisionTree se puede usar junto con NaiveBayes para aumentar la precisión (KOHAVI, 2011).

Para que el método supervisado funcione, primero debe (1) construir un conjunto de capacitación, en el que haya (1.1) la creación de un esquema codificado manualmente y (1.2) la selección aleatoria de documentos de muestreo (como regla general entre 100 y 500 documentos); luego (2) aplicar el método de aprendizaje supervisado para que el algoritmo "aprenda" – aprendiendo la relación entre características y categorías en el conjunto de entrenamiento y luego usándolo para inferir etiquetas en el conjunto de prueba – como clasificar el documentos en las categorías que usan el conjunto de entrenamiento; y (3) validar la salida del modelo comparando la salida de codificación automatizada con la salida de codificación manual. Los métodos de aprendizaje supervisados son mucho más fáciles de validar, con estadísticas claras que resumen el uso del modelo. Después de estos tres procedimientos, es posible clasificar con éxito los otros documentos (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018).

1.1.2 Máquinas de vectores de soporte (SVM)

El método de máquinas de vectores de soporte (SVM) es un método supervisado que utiliza un algoritmo de clasificación de texto para extraer los términos más indicativos de las posiciones conservadoras y liberales en el discurso legislativo y para predecir las posiciones ideológicas de los parlamentarios. El SVM se basa en el principio de minimización del riesgo estructural de la teoría del aprendizaje estadístico. Primero, se analizaron los discursos de los 25 senadores estadounidenses más liberales y conservadores, utilizando los puntajes del DW-Nominate (voteview.com), que miden el posicionamiento ideológico basado en votos nominales, para entrenar el algoritmo de clasificación (DIERMEIER, y col., 2012).

En la fase de entrenamiento, una categoría fue etiquetada arbitrariamente como "negativa" y la otra como "positiva". Debido a que los puntos de datos en cada categoría son separables por un hiperplano, hay dos hiperplanos paralelos donde la distancia entre los puntos en cada categoría es la más larga posible. Estos puntos de datos se denominan vectores de soporte, y la distancia entre los dos hiperplanos paralelos se denomina "margen". La tarea del SVM en la fase de entrenamiento es encontrar los dos hiperplanos de separación para que el margen sea máximo. Luego, se utilizó la misma metodología para investigar senadores moderados, capacitando al clasificador en senadores moderados (DIERMEIER at al., 2012).

1.1.3 Asignación de Dirichlet latente (LDA)

La asignación de Dirichlet latente (LDA) es el tema más simple o modelo de datos agrupados de asociación mixta. La idea detrás de la LDA es que cada documento es una mezcla de varios temas en diferentes proporciones. Además del supuesto de bolsa de palabras, existe el supuesto de que el número de temas es conocido y fijo. Técnicamente, el modelo supone que los temas se generan primero, antes que los documentos (BLEI, 2012; GRIMMER; STEWART, 2013).

Para cada documento del corpus, las palabras se generan en dos pasos: (1) elegir aleatoriamente una distribución sobre los temas; (2) para cada palabra en el documento (2.1) elija aleatoriamente un tema de la distribución del primer paso y (2.2) elija aleatoriamente una palabra de la distribución correspondiente sobre el vocabulario. La distribución que se usa para dibujar la distribución de temas previa al documento se denomina distribución de Dirichlet, y el resultado de Dirichlet se usa para asignar palabras del documento para diferentes temas. Todos los documentos comparten el mismo conjunto de temas, pero cada documento muestra estos temas en una proporción diferente (BLEI, 2012).

1.1.4 Modelo de tema estructural (STM)

El STM también es una plantilla de tema de asociación mixta, pero proporciona una forma flexible de incorporar metadatos asociados con el texto, como cuándo se escribió el texto, dónde se escribió, quién lo escribió, las características del autor, etc., como covariables en el análisis de documentos (ROBERTS y col., 2015). La inclusión de metadatos de documentos sigue y amplía el Modelo de Tema Dinámico, un modelo en el que la probabilidad de observar un tema cambia con el tiempo, y el Modelo de Agenda Expresada, modelo que incluye información sobre los autores de documentos, suponiendo que cada autor se divide su atención a un conjunto de temas. Las covariables también permiten compartir metadatos dependiendo de la frecuencia de los temas, como la probabilidad de que las mujeres hablen un tema específico o usen ciertas palabras en relación con los hombres. El STM también se distingue de la LDA al reemplazar la distribución de Dirichlet con una distribución logística normal, como en el Modelo de Temas Correlacionados, para estimar la correlación entre los temas. Con esto, es posible dibujar una red de temas correlacionados para un modelo de tema estructurado, utilizando un tema como predictor de temas en un corpus dado (GRIMMER; STEWART, 2013; ROBERTS y col., 2015).

1.2 Métodos de dimensionamiento

Los métodos de dimensionamiento se utilizan para ubicar a políticos y partidos en espacios ideológicos continuos (KLUVER, 2009; DIERMEIER y col., 2012; GRIMMER; STEWART, 2013; LAUDERDALE; HERZOG, 2016; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018). Estos métodos se basan en la suposición de que las tendencias ideológicas de los actores políticos determinan lo que se discute en los textos: la suposición de dominio ideológico en el

habla (GRIMMER; STEWART, 2013). Pero esto puede no ser cierto porque, como lo ha demostrado Mayhew (1974), los políticos participan regularmente en reclamos de crédito no ideológicos. Por lo tanto, los métodos de dimensionamiento funcionarán mejor si van acompañados de métodos que separen las declaraciones ideológicas y no ideológicas (CASAS; WILKERSON, 2017).

El uso más conocido de estos métodos es el proyecto Vote View (voteview.com). Debido a que la ideología no es directamente observable, Poole y Rosenthal (1991) desarrollaron un modelo espacial bidimensional (puntajes D-Nominate) para clasificar a los parlamentarios por sus votos nominales. La primera dimensión representa la visión tradicional izquierda-derecha asociada con el papel del gobierno en la economía y la redistribución del ingreso. La segunda dimensión representa cuestiones de interferencia estatal en la vida privada, la esclavitud y, posteriormente, cuestiones de derechos raciales y civiles. El modelo puede clasificar correctamente el 85% de las decisiones de votación individuales de cada miembro del Congreso (DIERMEIER y col., 2012).

Otra aplicación popular es el Manifesto Project, que utiliza la decodificación manual para clasificar a más de 1,000 partidos en 50 países según sus manifiestos políticos (manifesto-project.wzb.eu). En este caso, la estimación puntual ideal se utiliza para clasificar las partes como puntos de referencia en una escala ideológica izquierda-derecha (DIERMEIER en al., 2012).

Además, Lauderdale y Herzog (2016) estimaron las posiciones políticas individuales de cada parlamentario del EE. UU. utilizando el método de dimensionamiento Wordshoal, que combina el Wordfish y el análisis de factor Bayesiano. Los resultados de esta investigación también sugieren que las posiciones políticas individuales de los parlamentarios son predictores precisos de la cohesión dentro del partido y el comportamiento disidente en países donde los sistemas electorales ofrecen fuertes incentivos para los votos personales (LAUDERDALE; HERZOG, 2016). Finalmente, Pritoni (2014) analizó las dificultades para medir la influencia de los grupos de interés extrayendo su posición política y comparándola con productos legislativos, a través de métodos de dimensionamiento.

1.2.1 Wordscore

El Wordscore es un algoritmo de escala supervisado y un caso especial de método de diccionario. El primer paso es la selección de textos de referencia que definen las posiciones políticas en el espacio como liberales y conservadoras. Los textos de referencia (entrenamiento) se utilizan para generar una puntuación para cada palabra. La puntuación mide la velocidad relativa a la que se usa cada palabra en los textos de referencia. Esto crea una medida de qué tan bien la palabra separa a los parlamentarios liberales y conservadores. Luego, la puntuación de palabras se usa para dimensionar los textos restantes (GRIMMER; STEWART, 2013). Es decir, el segundo paso es generar puntajes para las palabras de los textos de referencia en función de la posición política asignada a priori y ponderada por la probabilidad de observarla en un

documento. Se puede usar el mismo procedimiento para la escala ideológica izquierda-derecha (IZUMI; MOREIRA, 2018).

El Wordscore se basa en una serie de suposiciones: (1) las posiciones políticas se reflejan en la frecuencia relativa de las palabras utilizadas dentro y entre textos; (2) el significado de las palabras permanece estable en el tiempo; (3) todas las palabras tienen el mismo peso en el proceso de estimación; y (4) todas las palabras de interés están contenidas en los textos de referencia (KLUVER, 2013). Todavía es necesario abordar la definición de la dimensión política a investigar y elegir un conjunto de textos de referencia con estimaciones de posición política conocidas, preferiblemente de una fuente independiente (KLUVER, 2013). También es importante que los textos de referencia utilicen el mismo léxico que los textos a ser probados, que cubren todo el espectro ideológico y que tienen un conjunto diverso de palabras (IZUMI; MOREIRA, 2018).

1.2.2 Wordfish

El Wordfish es un algoritmo no supervisado que estima la importancia de las palabras para discriminar posiciones políticas según la teoría de respuesta a artículos (IRT). El modelo supone una distribución de Poisson para el recuento de palabras y supone que la probabilidad de mirar una palabra en un documento es independiente de la posición de las otras palabras en el mismo documento. Por lo tanto, se puede usar para descubrir palabras que distinguen posiciones en un espectro político (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). Wordfish no requiere texto de referencia en contraste con Wordscore. En este caso, el investigador debe definir la dimensión política a analizar, seleccionar los documentos que abordan esta dimensión política y eliminar todos los pasajes de texto que no se refieren a la dimensión investigada (PRITONI, 2014). Por lo tanto, se requiere una validación cuidadosa para confirmar que se ha identificado el espacio ideológico deseado (GRIMMER; STEWART, 2013).

1.2.3 Wordshoal

El Wordshoal es un modelo de factores jerárquicos para hablar en debates legislativos que combina dos enfoques en una sola estrategia de estimación. El primero es limitar el análisis a los discursos sobre un solo tema legislativo, manteniendo constante la variación tópica. El segundo enfoque es combinar muchos discursos sobre muchos temas legislativos en un documento para cada legislador o partido. Es decir, en la primera etapa, el modelo utiliza la escala de texto de Wordfish existente para medir la variación del uso de palabras en cada tema por separado. En la segunda etapa, el análisis factorial bayesiano se utiliza para construir una escala común a partir de las posiciones de debate específicas estimadas en la primera etapa. Por lo tanto, el modelo presenta los resultados de las estimaciones basadas en el Wordfish para cada debate, y luego utiliza estas estimaciones como datos para el modelo de agregación de la segunda etapa para evaluar si un parlamentario está generalmente a la derecha o izquierda de otro parlamentario en

un conjunto de debates sobre temas heterogéneos (LAUDERDALE; HERZOG, 2016).

Más específicamente, el modelo utiliza una escala unidimensional del Wordfish aplicada a un conjunto de textos dentro de un solo debate político, manteniendo así una variación constante del uso de palabras basada en temas. Este modelo de escala de Poisson aplicado a cada debate da como resultado una estimación específica del debate sobre la posición relativa de cada parlamentario. Habiendo estimado la posición expresada para todos los parlamentarios sobre un tema dado, el modelo agrega dimensiones específicas de debate que involucran diferentes subconjuntos de legisladores en un número menor de dimensiones que incluyen a todos los legisladores. Dado que este enfoque no depende de variar el uso del discurso en ningún debate para estimar posiciones en una dimensión latente de desacuerdo, es posible generar una (o más) posición general latente para cada legislador altamente predictivo (LAUDERDALE; HERZOG, 2016).

La suposición empírica del Wordshoal es que el desacuerdo político se refleja más clara y consistentemente en la variación dentro del debate sobre el uso de palabras que en la variación en el uso de palabras en varios debates. Una de sus principales innovaciones es que el Wordshoal permite que el significado y el poder discriminatorio de una palabra dada varíen de un debate a otro. La variación en el uso de palabras entre discursos es tanto una función del tema de debate como una función de la posición que adopta un legislador. Además, el método proporciona estimaciones de incertidumbre significativas de las posiciones agregadas de los legisladores, teniendo en cuenta la frecuencia con la que los parlamentarios hablaron y la consistencia con la que expresaron sus posiciones en los debates (LAUDERDALE; HERZOG, 2016).

1.3 Métodos de similitud de texto

La reutilización de texto consiste en descubrir instancias de similitud en el uso del lenguaje. La característica distintiva de los algoritmos de reutilización de texto es que valoran explícitamente la secuencia de palabras al juzgar la similitud del documento (CASAS; WILKERSON, 2017). Sin embargo, dado que la semejanza del corpus es inherentemente multidimensional (serán similares de alguna manera y diferentes en otras), una medida de similitud solo tiene sentido al comparar dos corpus homogéneos. Por lo tanto, la similitud sólo puede interpretarse a la luz de la homogeneidad del cuerpo. En este sentido, se puede usar la misma medida para similitud y homogeneidad comparando la distancia entre dos cuerpos (distancia dentro del cuerpo) (KILGARRIFF; ROSE, 1998).

Se pueden usar métodos de similitud para comparar documentos en su conjunto o encontrar pequeños fragmentos de texto que coincidan entre dos documentos, como los que se encuentran en la legislación extraída de múltiples fuentes. Para estudiar si los proyectos de ley del Congreso reutilizan textos de otros proyectos de ley del Congreso, Wilkerson y col. (2015) utilizaron el algoritmo Smith-Waterman para comparar cadenas de texto de proyectos de ley del Congreso de

los Estados Unidos presentados desde 1990. Además de revelar estándares sobre parlamentarios que presentan proyectos similares en una legislatura o entre legislaturas, el algoritmo puede usarse para determinar en qué medida el lenguaje que introduce un legislador coincide con el de otros legisladores (BURGESS y col., 2016).

Ya Hertel-Fernández y Kashin (2015) utilizaron métodos de similitud para seguir los orígenes de las proposiciones y desentrañar la influencia de los grupos de interés en el proceso legislativo (CASAS; WILKERSON, 2017). En la misma línea, Burgess y col. (2016) desentrañó la reutilización del texto en las propuestas al usar el ElasticSearch para limitar el número de comparaciones para detectar la influencia de los grupos de interés en la proliferación de legislaciones estatales en los Estados Unidos (Proyecto de Detector de Influencia Legislativa: dssg.uchicago.edu/lid/).

Finalmente, los métodos de similitud también pueden usarse para posicionar a parlamentarios y partidos en un espectro ideológico. En este sentido, Schwartz (2018) utilizó la técnica X^2 para probar la similitud de los discursos de Expedientes Grandes y Pequeños, a través de la frecuencia de palabras y colocaciones en conjuntos de discursos de partidos políticos (PT, PSDB, PMDB, PSOL, PCdoB y PTB), comparados dos por dos.

1.3.1 Smith-Waterman

El algoritmo Smith-Waterman está diseñado para encontrar subsecuencias similares dentro de largas cadenas de ADN, pero también se puede usar para encontrar fragmentos de un documento que son similares a fragmentos de otros documentos mediante una puntuación de alineación basada en tres parámetros: coincidencia coincidencia, falta de coincidencia y brecha. Este algoritmo es una buena opción para comparar un número relativamente pequeño de documentos, pero puede llevar mucho tiempo ejecutarlo en un corpus grande (BURGESS et al, 2016).

Para resolver este problema, Burgess et al (2016) utilizaron el motor de búsqueda ElasticSearch, configurado con la función de puntuación de Lucene estándar, para ordenar documentos para una consulta determinada y así filtrar el conjunto de documentos ejecutados mediante la identificación de un subconjunto de los documentos del corpus tienen más probabilidades de contener texto similar al del documento de consulta. Se ha demostrado que el filtrado aumenta la eficiencia porque el algoritmo de alineación local compara solo los documentos devueltos por el módulo de búsqueda, sin sacrificar la precisión en las tareas de similitud de documentos (BURGESS et al, 2016).

1.3.2 Coseno

Como ya se mostró, cada documento puede representarse mediante un vector, cuya longitud es igual al número de palabras únicas en el texto. Por lo tanto, se supone que cuanto mayor sea la similitud en la frecuencia relativa de las palabras utilizadas, mayor será la similitud de contenido entre textos. Con dos vectores "u" y "v", es posible calcular la similitud a través del producto interno entre ellos, porque cuanto mayor es el producto interno entre ellos, mayor es la frecuencia de las mismas palabras. Como esta medida sigue siendo problemática, la solución es dividir el producto interno por el producto de las longitudes del vector, que está matemáticamente representado por el coseno del ángulo formado entre los vectores "u" y "v" (IZUMI; MOREIRA, 2018).

1.3.3 X^2

Kilgarriff y Rose (1998) presentaron un método para evaluar la similitud de los corpora, llamado Similitud Conocida de Corpora, y probaron enfoques comúnmente discutidos en la literatura: medidas de entropía cruzada, Spearmen y X^2. Para el tamaño del corpus utilizado, un subconjunto del Corpus Nacional Británico que contiene 300,000 palabras periodísticas y periódicas divididas en 10,000 pares de palabras, los enfoques X^2 y Spearmen funcionaron mejor[10] que cualquiera de las medidas de entropía cruzada; entre los dos, X^2 superó a Spearmen. Para cada una de las palabras más comunes, los autores calcularon el número de ocurrencias esperadas en cada corpus si ambos corpora eran muestras aleatorias del mismo corpus. Como un corpus nunca es una muestra aleatoria de palabras, la diferencia en la frecuencia de cada palabra entre dos corpora tiende a aumentar, pero no aumenta en el orden de magnitud, como ocurre con las frecuencias brutas (KILGARRIFF; ROSE, 1998).

[10]La prueba de confiabilidad (llamada "estándar de oro") de los métodos se realizó teniendo como parámetro la comparación de dos conjuntos de pares por los codificadores (KILGARRIFF; ROSE, 1998).

Cuadro 1- Resumen de los métodos automatizados de análisis de contenido

Método	Tecnica	Descripción y aplicación
Clasificación	NaiveBayes	Método supervisado que clasifica los documentos en categorías conocidas de un conjunto de entrenamiento. El clasificador Decision Tree se utiliza para aumentar la precisión en caso de grandes colecciones. Se puede utilizar para la indexación automática, la clasificación de los temas más debatidos, el análisis de sentimientos y el posicionamiento ideológico de los parlamentarios.
	SVM	Un método supervisado, Support Vector Machines (SVM) se utiliza junto con la base de datos de votación nominal para predecir la posición ideológica de los parlamentarios, para correlacionar lo que dijeron y cómo votaron, así como para medir la consistencia ideológica de los partidos y grado de cohesión intrapartidaria en el Congreso.
	LDA	Método no supervisado, Lattent Dirichlet Allocation (LDA) clasifica los documentos sin tener que especificar categorías por adelantado, pero debe ingresar el número de categorías.
	STM	Método no supervisado, el Modelo Structural Topic (STM) le permite ingresar metadatos como covariables, combinando el Modelo de Agenda Expresada y el Modelo Dynamic Topic. Se puede utilizar para clasificar los temas más debatidos, así como para comprender las posiciones políticas, los patrones de liderazgo y las influencias en el proceso legislativo.
Dimensionamiento	Wordscore	Método supervisado que posiciona políticamente documentos en dimensiones conocidas de un conjunto de capacitación. Se puede utilizar para posicionar a los partidos políticos y parlamentarios en una escala ideológica, como izquierda-derecha o liberal-conservadora, a través de sus discursos.
	Wordfish	Método no supervisado que posiciona políticamente los documentos sin tener que especificar dimensiones por adelantado, pero debe ingresar el número de dimensiones. Puede usarse para predecir posiciones políticas y posicionar partidos políticos y parlamentarios en una escala ideológica.
	Wordshoal	Wordshoal combina Wordfish y el análisis factorial bayesiano para estimar la posición de cada parlamentario entre sí dentro del partido y entre los partidos en función de sus discursos. También se puede usar para estimar la coherencia del posicionamiento parlamentario a lo largo del tiempo, así como la cohesión dentro del partido y el comportamiento disidente.
Similitud de texto	Smith-Waterman	Smith-Waterman es un algoritmo que mide la reutilización de partes de un texto y se utiliza para medir la similitud de secciones entre dos textos. Se puede utilizar para identificar la fuente de partes de propuestas, informes, enmiendas y legislación aprobada. También se puede usar para medir la influencia de los grupos de interés en los productos legislativos.
	Coseno	La técnica del coseno supone que cuanto mayor sea la similitud en la frecuencia relativa de las palabras utilizadas, mayor será la similitud entre dos textos en su conjunto. Se puede usar para identificar cuán similares son dos proposiciones o enmiendas.
	X^2	El estadístico X^2 se utiliza para medir la similitud y/o la homogeneidad entre los cuerpos. X^2 calcula el número de ocurrencias de las palabras más comunes en cada corpus y mide la diferencia entre la frecuencia de estas palabras en los dos corpus. Se puede usar para medir la identidad ideológica entre partidos y bloques de partidos en función del discurso de los parlamentarios.

Fuente: El autor (2018)

DESAFÍOS METODOLÓGICOS Y MEJORES PRÁCTICAS

Dado que los métodos automatizados de análisis de contenido no son un sustituto de la lectura humana, se requiere una validación cuidadosa de los resultados (GRIMMER; STEWART, 2013), basada en la replicabilidad de los resultados y la codificación manual rigurosa, preferiblemente por más de un codificador. Por lo tanto, la validación es un componente crítico de cada proyecto de texto como dado (CASAS; WILKERSON, 2017). Para los métodos de clasificación y dimensionamiento supervisados, es importante demostrar que la clasificación computarizada replica la codificación manual. Un clasificador se puede refinar mediante un libro de códigos y las iteraciones de codificación manual (GRIMMER et al, 2018). Sin embargo, en métodos no supervisados, no existe dicho estándar de oro, la validación se produce a medida que los parámetros se ajustan para examinar nuevos resultados (GRIMMER et al, 2018). Una forma de hacerlo es examinar el grado de cohesión y la distinción de palabras de cada tema (ROBERTS et al, 2014; CASAS; WILKERSON, 2017). Los métodos no supervisados también requieren la validación de que las mediciones producidas corresponden a los conceptos reivindicados (GRIMMER; STEWART, 2013).

Vale la pena recordar que, además de ser multidimensionales, los textos son más flexibles que otros tipos de variables, creando una gama más amplia de propiedades de texto potenciales para ser analizadas y validadas (GRIMMER et al, 2018). El tema de la multidimensionalidad del texto también es mencionado por Lauderdale y Herzog (2016), Roberts y col. (2015), Diermeier y col. (2012) y Kilgarriff y Rose (1998). Como todo el texto es multidimensional, es necesario elegir una o unas pocas dimensiones (representación de baja dimensión) para comprender el corpus y hacer inferencias. Esta limitación es inherente a los tres conjuntos de métodos[11]. Por supuesto, el problema de la multidimensionalidad en el caso de la posición ideológica también enfrenta la dificultad de operacionalizar conceptos complejos como ideología, izquierda-derecha, liberal-conservador, etc.

Por lo tanto, es importante validar los resultados con eventos del mundo real y los eventos esperados (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017) vinculándolos, por ejemplo, con hechos, datos cuantitativos y resultados de la actividad legislativa (SCHWARTZ, 2018).) Otra buena práctica de validación es usar diferentes algoritmos para el mismo propósito y comparar si los resultados son similares (CASAS; WILKERSON, 2017 citando a QUINN et al, 2010, GRIMMER; KING, 2011, ROBERTS et al, 2014). En este sentido, los métodos de aprendizaje supervisado se pueden utilizar para validar o generalizar los resultados

[11]En la clasificación es necesario definir los temas o al menos el número de temas. En el dimensionamiento es necesario definir las dimensiones a medir. En la similitud del corpus, dos textos serán similares de alguna manera y diferentes en otros, ya que la similitud solo puede interpretarse a la luz de la homogeneidad del corpus, es decir, no es apropiado, por ejemplo, comparar discursos parlamentarios con manuales técnicos (KILGARRIFF; ROSE, 1998).

proporcionados por métodos no supervisados (GRIMMER; STEWART, 2013). También existe una compensación entre el grado de generalización del concepto y la validez de los indicadores empíricos (PRITONI, 2014). Es tentador generalizar incluso cuando la propiedad del texto a analizar es más específica. Esto aumenta la relevancia teórica pero disminuye la fidelidad del indicador (GRIMMER et al, 2018).

Un problema común relacionado con la validación se llama sobreajuste. Cuando se usan los mismos documentos para descubrir propiedades de texto, es común malinterpretar los resultados. Para resolver este problema, una solución es separar los conjuntos de entrenamiento y prueba (diseño de muestra dividida). Al dividir las muestras y separar un conjunto de entrenamiento para usar en el descubrimiento (establecer resultados potenciales) y un conjunto de prueba en la estimación (análisis), la dependencia entre encontrar las propiedades del texto a analizar y su efecto causal (GRIMMER y col., 2018). Para Grimmer y Stewart (2013), el procedimiento de validación ideal sería dividir los datos en tres subconjuntos. El primero sería el subconjunto de prueba, donde se ajustaría el modelo. El segundo sería el subconjunto de validación, codificado a mano y utilizado para evaluar el rendimiento del modelo. Finalmente, el modelo se aplicaría para clasificar el tercer subconjunto.

Casas y Wilkerson (2017) también consideran útil entrenar el algoritmo en un conjunto de textos ya etiquetados antes de probar su precisión en un conjunto desconocido. Sin embargo, creen que repetir este proceso una y otra vez, usar diferentes conjuntos para el entrenamiento y las pruebas, y luego agregar la validación de resultados (validación cruzada múltiple) es un enfoque aún mejor. Además del diseño de muestra dividida, otra forma de evitar el sobreajuste es usar diferentes algoritmos para demostrar si hay grupos similares (CASAS; WILKERSON, 2017, citando a QUINN et al, 2010; GRIMMER ; KING, 2011 y ROBERTS et al, 2014).

Además del sobreajuste, también existe el problema de la inestabilidad del tema en los métodos no supervisados. Para evitar esta inestabilidad, es saludable utilizar resultados de diferentes modelos del mismo algoritmo. Casas y Wilkerson (2017), por ejemplo, utilizaron 17 modelos de asignación latente de Dirichlet (LDA), variando el número de temas entre 10 y 90, en incrementos de cinco, para producir 850 temas (10 + 15 + 20... + 90) Para determinar qué temas eran consistentes, primero calcularon la similitud del coseno para todos los pares de temas (lo que resultó en 722.500 puntajes de similitud) y luego usaron el algoritmo SpectralClustering para agrupar los 850 temas basados en la similitud del coseno y verificar qué temas permanecieron constantes (CASAS; WILKERSON, 2017).

Otras dos características típicas de los discursos parlamentarios son, como demuestran Lauderdale y Herzog (2016), la escasez y la selección de oradores. Teniendo en cuenta que en la vida real solo unos pocos parlamentarios hablan en ciertos debates, como resultado, la Matriz de Documentos y Términos (DTM) está dispersa. Otro punto es el control del partido y el control de la agenda. Para evitar este sesgo, Moreira (2016) usó solo discursos de Expedientes Pequeños, y

Schwartz (2018) usó discursos de Expedientes Pequeños y Grandes, porque en estos dos momentos, los parlamentarios son libres de expresarse sobre cualquier tema. Lauderdale y Herzog (2016) también encontraron que en los sistemas con fuertes incentivos para el voto personal y no partidista, los parlamentarios hablan más libremente porque los partidos reconocen la necesidad de reconocer los nombres de los parlamentarios.

POSIBLES APLICACIONES Y RECOMENDACIONES

Hay varias aplicaciones posibles de métodos automatizadod de contenido para las Cámaras Legislativas. Los métodos supervisados, por ejemplo, pueden ser más adecuados para ahorrar trabajo, como a través de la indexación o clasificación automáticas, y los métodos no supervisados para hacer descubrimientos y proporcionar información sobre estas clasificaciones (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017). Sin embargo, también son métodos complementarios, ya que el aprendizaje supervisado se puede utilizar para validar o generalizar los hallazgos proporcionados por métodos no supervisados (GRIMMER; STEWART, 2013).

En el caso de los métodos supervisados, es obvio que puede ahorrar trabajo mediante la indexación automática o al menos semiautomática, si el servidor confirma el término sugerido por el algoritmo. Entre los diversos procesos de indexación realizados por la Cámara de Diputados, por ejemplo, se encuentran la indexación de discursos parlamentarios (realizados por el Departamento de Taquigrafía, Revisión y Redacción o DETAQ), proposiciones (llevadas a cabo por el Centro de Documentación e Información o CEDI) y artículos periodísticos (realizados por la Secretaría de Comunicación Social o SECOM). Además de la indexación, otros procesos, como unir y distribuir propuestas a los comités (llevados a cabo por la Secretaría General de la Mesa Directiva o SGM), también podrían ganar mucho con el uso de métodos supervisados. Además de ahorrar trabajo, la automatización también aportaría una mayor uniformidad a la indexación de la cámara y, en consecuencia, facilitaría la búsqueda y recuperación de información.

Por lo tanto, se recomienda utilizar un método supervisado, como NaiveBayes, para implementar la automatización de los procesos de indexación; así como refinar el clasificador a través de un libro de códigos, como la base del tesauro de la cámara, y validar los resultados de las pruebas mediante iteraciones de codificación manual. Aún así, es saludable mantener el componente humano en el proceso de trabajo para revisar los resultados presentados por la máquina. Para evitar el sobreajuste, se recomienda utilizar un diseño de muestra dividida [12].

[12]También puede usar el método sin supervisión de STM para descubrir nuevas palabras y frases que los miembros usan con frecuencia, pero no en la indexación. Este puede ser el caso con neologismos y expresiones como "ideología de género", "escuela sin partido", "pixuleco", etc.

Otra aplicación de análisis automatizado para las Cámaras Legislativas es la provisión de información estratégica diversa. ¿Qué tan importante es saber qué temas fueron más debatidos por la Cámara durante la legislatura? ¿Cuál es la posición de los diputados y partidos en cada uno de estos temas? ¿Es esta una legislatura mayoritariamente de derecha o izquierda? ¿Liberal o conservador? ¿Existe una identificación ideológica en las partes? ¿Los parlamentarios votan de acuerdo con esta identificación? ¿Hay coherencia en lo que hablan y cómo votan los parlamentarios? ¿Existe una cohesión partidaria en los discursos parlamentarios? ¿Hay una alineación entre los discursos de base y de oposición? ¿Quién influye más en el proceso legislativo y en los discursos de los parlamentarios? El poder ejecutivo? Los partidos? La base electoral? ¿Los grupos de interés?

La cantidad de información que se puede extraer de los discursos parlamentarios, especialmente dada la combinación de estos discursos con los productos y resultados de la actividad legislativa, parece interminable. Para fines didácticos, es posible dividir esta información en tres conjuntos: (1) temas de debate, (2) posicionamiento en una escala política o ideológica, y (3) influencia en el proceso legislativo. Dado que el uso de palabras en el debate político puede variar según el tema discutido, la clasificación de los temas también sirve como una dimensión para analizar el posicionamiento y la influencia política.

Por lo tanto, la clasificación de temas surge como el primer desafío para la Cámara para proporcionar información estratégica de calidad a través del análisis de contenido automatizado. Entre los diversos enfoques posibles, se sugiere utilizar métodos supervisados y no supervisados como complementarios, utilizando un algoritmo para validar los resultados del otro. En este sentido, hay dos formas posibles: (1) utilizar el método STM no supervisado para la clasificación y validar los resultados con el método supervisado NaiveBayes/DecisionTree; o (2) usar NaiveBayes/DecisionTree para la clasificación, y después validar los resultados con el STM[13]. En ambos sentidos, es posible utilizar el resultado binomial/trinomial para rechazar la hipótesis nula con respecto al uso de los dos algoritmos, así como para validar los resultados con hechos, eventos y resultados de la actividad legislativa.

Además, se recomienda que utilice un diseño de muestra dividida para NaiveBayes, así como refinar el clasificador y validar el conjunto de entrenamiento mediante iteraciones de codificación manual. En el caso del STM, se recomienda su uso, siguiendo las enseñanzas de Casas y Wilkerson (2017) y en vista de los 32 temas utilizados por la Cámara en su tabla de clasificación[14] – al menos seis modelos diferentes con incrementos de ocho temas cada uno (8, 16, 24, 32, 40,

[13]STM es el método no supervisado más recomendado para permitir el uso de metadatos como covariables, como el tiempo de conversación. Otra característica importante de este método es la posibilidad de crear un campo léxico, como una red de palabras, alrededor del tema. El Naive Bayes/Decision Tree, a su vez, es un algoritmo bien probado cuyo uso ya está dominado por algunos servidores de la Cámara..

[14]Como los 32 temas son muy amplios, la definición de subtemas es esencial para comprender el debate parlamentario y facilitar la clasificación de conjuntos de proposiciones.

48). Por lo tanto, puede abordar el problema de inestabilidad del tema al agrupar todos los temas resultantes de las seis plantillas y verificar que temas permanecen consistentes.

Debemos recordar que los métodos automatizados cuentan la frecuencia de las palabras en el habla, sin tener en cuenta el tiempo del habla. Por lo tanto, un congresista puede hablar la misma palabra docenas de veces en un discurso de cinco minutos, y otro puede hablar solo unas pocas veces en un discurso importante. Aunque el uso de campos léxicos para definir temas mitiga este problema, es aconsejable medir el tiempo de debate por tema, considerado el inicio y el final del debate como metadatos del discurso, y también usarlo como indicador de tema más debatido en la Cámara en un período determinado. Se cree que es posible lograr estos resultados utilizando un enfoque en el que el tiempo (período inicial y final) de los discursos se usa como covariable en el método STM. En este mismo sentido, es posible medir el tiempo de debate invertido en un subtema, en un conjunto de proposiciones o en una política pública. Por lo tanto, el tiempo para hablar puede verse como un indicador de poder, al medir el dominio de los partidos y parlamentarios en el trabajo de la Cámara, así como un indicador de desempeño al medir el tiempo total invertido en debatir las políticas públicas.

Sin embargo, los temas discutidos también reflejan las ideologías presentes en el Parlamento. Como diría Bakhtin (1988), el discurso parlamentario es también un contenido ideológico, se refiere a algo fuera de sí mismo, al establecer intereses y niveles de dominación, transformando el Parlamento en un escenario donde se libran las batallas de diversos intereses nacionales. En este sentido, las ideologías se expresan no necesariamente hablando de manera diferente sobre los mismos problemas, sino también sobre temas diferentes (DIERMEIER y col., 2012). Por lo tanto, buscamos comprender a través de modelos estadísticos como estos contenidos ideológicos se manifiestan en formas lingüísticas y como reflejan y/o refractan dichos contenidos.

Empíricamente, la ideología nos permite predecir la posición política de un parlamentario sobre un tema a través de su posición sobre otro tema no relacionado (DIERMEIER y col., 2012). Tradicionalmente, los politólogos usan los votos nominales para estimar las posiciones individuales de los parlamentarios. Sin embargo, la votación a menudo es solo simbólica y no está registrada nominalmente. También hay un fuerte control del partido en las encuestas para evitar votos discrepantes. En este sentido, es más preciso estimar la diversidad de posiciones de los parlamentarios a través del conjunto de sus discursos (LAUDERDALE; HERZOG, 2016). Los discursos parlamentarios serían, por lo tanto, un "producto" que demuestra la conexión electoral.

Dado que las ideologías pueden manifestarse a través de la elección de temas, frases o palabras en contextos particulares, primero es necesario definir que se entiende por "ideología". Pragmáticamente, según Converse (1964), la ideología es un conjunto de creencias que guía el posicionamiento del individuo en varios temas (DIERMEIER y col., 2012). Esta es una visión similar a lo que Cancian (2007) defiende para la investigación empírica: la ideología se describe

como el conjunto de ideas, valores o creencias que guían la percepción y el comportamiento de las personas sobre diversos temas (SCHWARTZ, 2018).

Al definir la ideología como un conjunto de creencias que guían el posicionamiento parlamentario, es posible examinar si las posiciones ideológicas de los parlamentarios expresadas en sus discursos determinan sus votos, en vista de las limitaciones institucionales como el control de la agenda y los partidos (DIERMEIER y col., 2012). Si el voto se explica más por una ideología política preexistente (como se expresó en el discurso) que por factores institucionales[15], conocer la posición de los parlamentarios sobre un conjunto de temas es predictivo para conocer su visión para otras preguntas no correlacionadas[16](DIERMEIER y col., 2012).

Sin embargo, dada la necesidad de reelección parlamentaria - especialmente en sistemas con fuertes incentivos para el voto personal y no partidista - los parlamentarios necesitan, además de tomar posiciones sobre diversos temas, anunciar base electoral y obtener crédito. En este punto, Casas y Wilkerson (2017) están de acuerdo con Mayhew (1974) y concluyen: no todo discurso parlamentario es ideológico. No es coincidencia que los resultados de la encuesta de Diermeier et al (2012) mostraron que tanto los republicanos como los demócratas pasan la mayor parte de sus discursos felicitando a los votantes[17]. Por lo tanto, la necesidad de separar tanto como sea posible el contenido no ideológico del corpus es evidente, eliminando, por ejemplo, discursos de procedimiento, homenajes y agradecimientos.

Una buena práctica es observar las variaciones en el uso de palabras por tema antes de analizar las variaciones por posicionamiento. También es aconsejable utilizar dimensiones específicas en cada debate, luego resumir la variación en estas dimensiones específicas de cada debate utilizando dos dimensiones generales – como izquierda/derecha, conservador/liberal o gobierno/oposición. Esto nos permite recuperar la multidimensionalidad en las estimaciones de preferencias con etiquetas de temas. Otra buena práctica es medir la variación por tema y por parlamentario, luego fusionar todos los temas de los mismos parlamentarios para recuperar la multidimensionalidad (LAUDERDALE; HERZOG, 2016).

Como estas prácticas dependen de una definición detallada de las dos (o más) dimensiones generales utilizadas, se recomienda crear un grupo de estudio, con la participación de consultores de la Cámara, para crear un diccionario anotado por tema y por política pública, con las definiciones y el campo léxico de lo que sería izquierda y derecha, así como conservador y liberal,

[15]Diermeier y col. (2012) compara lo que dijeron los senadores estadounidenses y como votaron y concluye que votar y debatir son expresiones diferentes pero correlacionadas, de un mismo sistema de creencias ideológicas.

[16]Dado que se ha logrado una precisión del 94% en la clasificación de senadores extremistas y 52% de senadores moderados (además de reconocimiento), a través de experimentos adicionales de que los senadores moderados son versiones atenuadas de senadores extremistas en lugar de una categoría distinta, se cree que existe un campo léxico distinto para los legisladores conservadores y liberales. A pesar de los muchos temas discutidos en el Congreso, los conservadores y liberales siempre hablan sobre cualquiera de estos temas de manera distinta y estable (DIERMEIER y col., 2012).

[17]Los resultados indicaron que los republicanos tienden más a dar discursos de felicitación (17%), seguidos de discursos de promoción (8%), educación (7%) y familiares (6%). Los demócratas también dan más discursos de felicitación (9%), seguidos de discursos sobre educación (9%), salud (7%) y familia (5%).

en cada uno de estos temas, en vista de la indexación de la Cámara y los trabajos anteriores sobre el asunto[18]. En términos de métodos, los tres conjuntos analizados pueden medir el posicionamiento con éxito. Hay experimentos con el SVM - que se pueden usar junto con los votos nominales para este propósito – con el Wordshoal – que combinan el Wordfish con el análisis factorial bayesiano, para estimar posiciones – y con la X^2 – el enfoque de similitud de texto combinado con el análisis factorial bayesiano para el mismo propósito. Se recomienda utilizar diferentes métodos para validarse entre sí, como se recomienda en la literatura.

Además de medir el posicionamiento, estimar las posiciones políticas y predecir la votación, también es posible analizar la coherencia de un parlamentario, observando la variación de su posicionamiento a lo largo del tiempo y la diferencia entre sus discursos y sus votos nominales, así como la cohesión de partidos, bloques y frentes parlamentarios. La cohesión de los partidos puede medirse mediante la aglomeración de diputados en la escala ideológica o la similitud del discurso. En este sentido, Diermeier y col. (2012) encontró evidencia de que la afiliación partidaria y la clasificación ideológica están altamente correlacionadas en los Estados Unidos, calculando la relación de concordancia "*kappa*" como una medida de consistencia entre las etiquetas ideológicas de los senadores y la afiliación partidaria (hubo concordancia casi perfecta: *kappa* = 0.932). Un tipo especial de cohesión es la alineación de la base del gobierno, como lo ha demostrado Schwartz (2018). En este sentido, también es aconsejable centrarse en el uso de diferentes métodos y enfoques para medir el grado de cohesión del partido, ya que este tipo de información es de gran importancia para los líderes de la Cámara.

Además de los temas de debate y posicionamiento en el espectro ideológico, hay un tercer uso para los métodos automatizados: identificar las influencias en los documentos y los resultados del proceso legislativo. En estudios más recientes, la influencia se conceptualiza como control sobre los resultados, más específicamente la diferencia entre lo que los actores involucrados expresamente desean y el resultado[19] (PRITONI, 2014). Puede ser difícil desentrañar lo que los actores involucrados realmente quieren, pero es posible identificar a los actores que han influido en las políticas públicas al conocer el origen de las ideas que se expresan en la legislación aprobada (BURGESS et al, 2016). Por lo tanto, se hace posible y relevante identificar la influencia del Poder Ejecutivo, los partidos, los bloques y los frentes parlamentarios, así como los grupos de interés, en el proceso legislativo, a través de la reutilización del texto.

En este sentido, Burgess et al (2016) utilizaron el algoritmo Smith-Waterman y el mecanismo de búsqueda ElasticSearch para rastrear el origen de las proposiciones. Este mismo enfoque se puede utilizar para verificar la reutilización del texto en las opiniones y enmiendas parlamentarias.

[18]Como el libro de Noberto Bobbio, "Derecha e Izquierda" (1994) y el proyecto voteview.com.

[19]Sin embargo, Pritoni (2014) afirma que los actores involucrados pueden exagerar sus reclamos iniciales para facilitar la negociación y lograr el mejor resultado posible por el momento, independientemente de si la intención es cambiar o mantener la política pública. Por lo tanto, sólo los grupos de interés que desean mantener las políticas públicas mostrarían su verdadera intención, que es no cambiar nada.

Hertel-Fernández y Kashin (2015) utilizaron la combinación de tres métodos, que implican la similitud de bigramas y trigramas, LDA y SVM, para identificar la influencia de los grupos de interés en la ley estatal de los Estados Unidos. Sin embargo, estos enfoques se centran en productos legislativos, como propuestas, opiniones, enmiendas y estudios, y en la legislación misma[20]. Se recomienda el uso del enfoque de Burgess et al (2016) en documentos como propuestas y enmiendas. Además de proporcionar información valiosa sobre la reutilización de texto, los resultados de esta aplicación también pueden contribuir al proceso de anexar propuestas y considerar enmiendas.

REFERENCIAS BIBLIOGRÁFICAS

BAKHTIN, Mikhail. Marxismo e filosofia da linguagem. São Paulo: Hucitec, 1995.

BLEI, D. ProbabilisticTopicModels. **Communications of the ACM**, New York, NY, USA, v. 55, n. 4, p. 77-84, 2012.

BURGESS, M. y col. **The Legislative Influence Detector**: Finding Text Reuse in State Legislation. *KDD*, San Francisco, CA, USA, 2016.

DIERMEIER, D. at al. Language and ideology in Congress. **Br. J. Polit. Sci**. 42(1):31–55, 2012.

GRIMMER, J. We are all social scientists now: how big data, machine learning, and causal inference work together. **PS: Political Science & Politics**, Cambridge, UK, v. 48, n. 1, p. 80-83, 2015.

GRIMMER, J.; STEWART, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. **Political Analysis**, Oxford, v. 21, n. 3, p. 267-297, jan. 2013.

GRIMMER, J. y col. How to Make Causal Inferences using Texts. *Working paper*, 2018.

HERTEL-FERNANDEZ, A.; KASHIN, K. 2015. **Capturing business power across the states with text reuse**. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 16–19, 2015.

IZUMI, M. **Velhas questões, novos métodos**: posições, agenda, ideologia e dinheiro na política brasileira. 2017. 113 f. Tese (Doutorado em Ciência Política) – Universidade de São Paulo, São Paulo, 2017.

KILGARRIFF, A.; ROSE, T. Measures for corpus similarity and homogeneity. **Information Technology Research Institute Technical Report Series**, Brighton, p. 46-52, jul. 1998.

KLUVER, H. Measuring interest group influence using quantitative text analysis. **Eur. Union Polit.**, v.10, n. 4, p.535– 49, 2009.

[20]Con respecto a los discursos parlamentarios, se pueden utilizar métodos similares para rastrear el origen de las palabras, expresiones y argumentos en los debates parlamentarios. Tal enfoque, por ejemplo, puede servir para analizar la similitud entre el discurso de los diputados y los discursos de los representantes de los grupos de interés.

KLUVER, H. **Lobbying in the European Union**: Interest groups, lobbying coalitions and policy change, Oxford, Oxford University Press, 2013.

KOHAVI, R. **Scaling Up the Accuracy of Naive-Bayes Classifiers**: a Decision-Tree Hybrid. **Data Mining and Visualization Silicon Graphics**, Inc. Mountain View, CA, USA, 2011.

LAUDERDALE, B., CLARK, T. Scaling politically meaningful dimensions using texts and votes. **Am. J. Polit. Sci**., v.58, n. 3, p.754–71, 2014.

LAUDERDALE, B. HERZOG, A. Measuring political positions from legislative speech. **Political Analysis**, Cambridge, UK, v. 24, n. 3, p. 374-394, 2016.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge: The MIT Press, 1999.

MAYHEW, D. Congress: **The Electoral Connection**. Yale University Press, 1974.

MOREIRA, D. **Com a palavra os nobres deputados**: frequência e ênfase temática dos discursos dos parlamentares brasileiros. 2016. 204 f. Tese (Doutorado em Ciência Política) – Universidade de São Paulo, SP, 2016.

MOREIRA, D.; IZUMI M. O Texto como Dado: Desafios e Oportunidades para as Ciências Sociais. **Revista Brasileira de Informação Bibliográfica em Ciências Sociais – BIB**. São Paulo, BR, n. 86, 2/2018.

POWER, T.; ZUCCO, C. Estimating ideology of Brazilian legislative parties, 1990-2005: a research communication. **Latin American Research Review**, Pittsburgh, v. 44, n. 1, p. 218-246, 2009.

PRITONI, A. How To Measure Interest Group Influence: Evidence From Italy. **ECPR Joint Sessions of Workshops**, Salamanca, Spain, 2014.

ROBERTS, M. y col. The structural topic model and applied social science. **Advances in neural information processing systems workshop on topic models**: computation, application, and evaluation. Cambridge, MA: Harvard University, 2013.

ROBERTS, M. y col. Computer-Assisted Text Analysis for Comparative Politics. **Political Analysis**, Cambridge, UK, v. 23, p. 254-277, 2015.

SCHWARTZ, F. **Análise do Discurso Parlamentar por Meio da Técnica do Processamento de Linguagem Natural**: Abordagem Estatística e Aprendizagem de Máquina. 2018. 76 f. Pesquisa (Pós-Doutorado em Tecnologia) – Universidade de Brasília, BSB, 2018.

CASAS, A.; WILKERSON, J. Large-scale computerized text analysis in political science: Opportunities and challenges. **Annual Review of Political Science**, Palo Alto, v. 20, p. 529-544, 2017.

WILKERSON, J. y col. Tracing the flow of policy ideas in legislatures: a text reuse approach. **Am. J. Polit. Sci**., v.59, n. 4, p.943–56, 2015.

Artículo recibido el 27/12/2018
Artículo reenviado el 20/02/2019
Artículo aceptado para publicación el 12/03/2019