Universidad de Oviedo

Programa de Doctorado en Informática

TECNOLOGÍAS SEMÁNTICAS EN EL ÁMBITO POLÍTICO-LEGISLATIVO

Doctorando:

Francisco Adolfo Cifuentes Silva

Oviedo, 2025

Universidad de Oviedo

Escuela de ingeniería informática

Programa de Doctorado en Informática

TECNOLOGÍAS SEMÁNTICAS EN EL ÁMBITO POLÍTICO-LEGISLATIVO

Doctorando:

Francisco Adolfo Cifuentes Silva

Director:

Jose Emilio Labra Gayo

Oviedo, 2025

Esta tesis está dedicada a dos personas:

A mi abueli, Emilia de Silva, que me ha acompañado desde el cielo durante esta larga aventura.

A ese niño flaco morenito, ese bien porfiado, que nació en el sur de la capital un domingo de lluvia, que cuando pequeño copiaba letras sin entender y cuando grande nunca se enteró que había un límite para hacer los sueños realidad.

Agradecimientos

A quienes están día a día conmigo, mi *core*, mi compañera de vida y esposa Karina y mis hijas Fernanda y Magdalena, que me acompañan en mis días buenos y malos.

A mi padre, Raúl Cifuentes, por darme fuerza y ánimo cada vez que podría, y a mi madre, María Eugenia Silva, por apoyarme y siempre confiar en mi.

A mi director de tesis José Emilio Labra, por su ayuda y apoyo cada vez que lo he necesitado, en esos momentos clave, y también por la confianza que siempre ha depositado en mi.

Al profesor Hernán Astudillo, quien me ayudó a ver muchas veces lo que no podía, y con su experiencia me afianzó en el desarrollo de la tesis.

A la Bilblioteca del Congreso Nacional de Chile (BCN) y sus autoridades, por apoyar mis estudios de doctorado mediante su política de capacitación institucional, como también por permitirme abrir los espacios para desarrollar mis ideas.

A don Manuel Alfonso Pérez, por haber confiado en mi desde un inicio y ver valor en este trabajo.

A mis amigas y amigos de las distintas áreas de la BCN que pusieron un granito de arena. En especial a Felipe Rivera, Mauricio Amar, Ángelo Palli, Carlos Medel, Daniela Benavente, Ignacio Riquelme, Denisse Espinace, Felipe Escobar, Francesca Alfaro, Karem Orrego, Mariano Ferrero, Renzo Dinali, Marek Hoehn, Eduardo Goldstein, David Manríquez y Carolina Reyes, quienes me apoyaron de distintas maneras, tanto en las etapas de pruebas experimentales como en discusiones conceptuales y técnicas.

A los doctores Patricia Reyes y Edgar Castillo por haber leído y revisado con muy buen criterio este documento, y permitirme presentar esta tesis.

A mis hermanos Jaime, Pablo y Omar por estar siempre conmigo, y apoyarme de distintas maneras durante este camino, y a mi cuñada Lorena por apoyarme y recibirme en mis inicios profesionales.

Y a Mario Alonso Puig, por enseñarme de forma anónima y desinteresada que por muy difícil que se vea, siempre hay algo mínimo que podemos hacer para aportar a cumplir nuestros objetivos.

Resumen

Actualmente el ámbito político-legislativo, se enfrenta a un creciente volumen de datos proveniente de diversas fuentes, al tiempo que una amplia variedad de actores requieren analizar de forma eficiente esta información para la toma de decisiones y acciones. Es en este contexto que las Tecnologías Semánticas, que en este trabajo son entendidas como la unión de los campos tecnológicos de Web Semántica, Minería de Textos y Análisis de Redes Sociales, surgen como una solución que permite habilitar nuevas formas de análisis, permitiendo dentro de otros, estructurar, vincular y dotar de significado a datos heterogéneos, integrarlos y convertirlos en información útil a través de automatización y visualización.

De esta manera, esta tesis trata sobre análisis político-legislativo utilizando tecnologías semánticas, tomando como base para la investigación el trabajo realizado en la Biblioteca del Congreso Nacional de Chile (BCN) sobre el Congreso Nacional chileno, donde se utilizan tecnologías semánticas para procesar datos de distintos tipos (no estructurados, semiestructurados y estructurados) permitiendo generar en la actualidad productos y servicios para la ciudadanía y la comunidad de usuarios del Congreso Nacional.

En la necesidad de introducir nuevas herramientas para la asesoría parlamentaria utilizando tecnologías semánticas, se han diseñado e implementado herramientas de análisis orientadas a responder tres preguntas reales del ámbito político-legislativo: 1) ¿quién cumple un rol clave en el contexto de un tema específico?, 2) ¿Cuál es el nivel de cohesión política de un grupo frente a un tema particular?, y 3) ¿cuáles son los temas de mayor relevancia para un representante?. La hipótesis del trabajo sostiene que es posible automatizar el procesamiento de datos no estructurados mediante este tipo de tecnologías para responder preguntas de análisis político-legislativo.

Para proveer datos para el análisis, primeramente se describe el contexto tecnológico en torno a las tecnologías semánticas en la BCN y se evalúan distintas fuentes de datos disponibles, seleccionando aquella más idónea para el trabajo de investigación a fin de equilibrar el alcance de la tesis con la riqueza del conjunto de datos en cuanto a sus características.

Posteriormente, se explora una serie de experiencias en el campo del análisis político-legislativo mediante tecnologías informáticas, se describe un marco de trabajo de las tecnologías semánticas para su aplicación en el ámbito político-legislativo, y se diseñan y describen tres instrumentos de análisis que se basan en este marco de trabajo: un visualizador de temas de interés parlamentario, un visualizador de rol clave asociado a un tema específico y un visualizador de cohesión política, donde cada uno ha sido desarrollado utilizando diversos componentes, fuentes de datos y técnicas de analítica. Cada uno de estos instrumentos permite responder una de las preguntas previamente descritas, y en conjunto inducen la validación de la hipótesis general.

Habiendo desarrollado estos instrumentos, se realiza un experimento con un grupo de usuarios expertos del área de la asesoría parlamentaria, donde se evalúa la percepción de acierto de los instrumentos de análisis desarrollados, el cual arroja que de los tres instrumentos, dos responden de forma totalmente satisfactoria las preguntas de investigación y un tercer instrumento posee una valoración satisfactoria pero con un margen de mejora mayor. El análisis de los resultados de la evaluación permite concluir que las preguntas de investigación definidas pueden ser respondidas con un nivel de acierto satisfactorio y en consecuencia, que la hipótesis general puede ser validada con base en estos resultados.

Finalmente, a lo largo del documento se hace referencia a los artículos desarrollados durante la investigación, permitiendo añadir en el capítulo de conclusiones una serie de hallazgos asociados al uso de tecnologías semánticas aplicadas al ámbito político-legislativo.

Índice general

Lista de Acrónimos

AKN Akoma-Ntoso. 6, 47, 112

AUC Area Under Curve. 153

BCN Biblioteca del Congreso Nacional de Chile. , 5–7, 24, 111, 112

BERT Bidirectional Encoder Representations from Transformers. 35

GAT Graph Attention Networks. 35

GE Grupo de Expertos. 28, 60, 62, 64, 67, 69, 75, 76, 81, 87, 90, 95, 100, 107–110, 115–117

HL Historia de la Ley. 5

IA Inteligencia Artificial. 39, 56, 111

LDA Latent Dirichlet Allocation. 23, 38, 50, 155

LP Labor Parlamentaria. 5

MLP Multi Layer Perceptron. 35, 131

NER Named Entity Recognition. 48, 49, 70

NoSQL Not Only SQL. 83

OWL Ontology Web Language. 46

PCA Análisis de Componentes Principales. 35

PLN Procesamiento de Lenguaje Natural. 47, 48, 50, 131

RDF Resource Description Framework. 34, 44, 46

RDFS RDF Schema. 46

RNN Redes Neuronales Recursivas. 35

ROC Receiver Operating Characteristic. 153

SHACL Shapes Constraint Language. 46

ShEx Shape Expressions. 46

SNA Social Network Analysis. 36, 37, 50, 83, 84

SPARQL SPARQL Protocol and RDF Query Language. 46, 49, 82

STM Structural Topic Model. 38

SVM Support Vector Machine. 35, 38, 131

TF Term Frequency. 38, 48

TF-IDF Term Frequency-Inverse Document Frequency. 48, 132, 155

TI Tecnologías de la Información. 29, 39

URI Uniform Resource Identifier. 44, 49

WSD Word-sense disambiguation. 48

XML Extensible Markup Language. 47

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1 La asamblea del sábado

La asamblea del sábado convocó a un diverso grupo de vecinos en la plaza de armas. Algunos eran campesinos que cultivaban sus tierras, otros eran maestros que enseñaban a los jóvenes el catecismo y la lengua española, y varios eran comerciantes que vendían telas y especias. Todos compartían la misma preocupación: un nuevo proyecto para dividir las tierras comunes y construir un nuevo camino real podría despojarles de sus tierras.

Para entender mejor las implicaciones de esta propuesta, decidieron realizar un exhaustivo análisis político. Consultaron los documentos de don Ambrosio, el gobernador, discutieron las intenciones de los corregidores, hablaron con varios funcionarios locales y hasta con el cura. Aprendieron a reconocer promesas vacías y a detectar las intrincadas manipulaciones que beneficiaban a los aristócratas en detrimento de los más humildes.

Armados con este conocimiento, asistieron a la siguiente reunión en el cabildo. Gracias a sus razonamientos claros y fundamentados, lograron que la autoridad colonial reconsiderara el proyecto. El resultado fue que las tierras comunes quedaran protegidas y se mejoraran las rutas comerciales sin perjudicar a los campesinos.

A raíz de este análisis, comprendieron que su participación activa era fundamental en la defensa de sus bienes y derechos. Unidos por esta experiencia, se comprometieron a mantenerse informados y a recordar siempre que, cuando se comprende la política, ésta puede ser una herramienta poderosa para el bienestar de todos.

Anécdota del Chile colonial

La política, vista como una de las más antiguas actividades sociales del ser humano, goza de ubicuidad y gran impacto en la vida de las personas, dado que es a través de los mecanismos que ésta provee, cómo es que se organiza la sociedad, pasando desde familias o tribus, hasta naciones y continentes.

Las primeras referencias a los conceptos en torno a la política se remontan al siglo V antes de Cristo en la antigua Grecia, cuando pensadores como *Herótodo de Halicarnaso* comenzaban a utilizar conceptos como el de *democracia* para describir los sistemas políticos. Posteriormente, en el siglo IV antes de Cristo, el filósofo griego *Platón* en su obra *La República*, planteaba los primeros fundamentos e ideas de lo que comenzaba a sistematizar bajo el concepto de política, estableciendo en ese entonces bases filosóficas para el análisis político. Pocos años más tarde, *Aristóteles* en su obra *La Política*, describe las formas de gobierno y la dinámica del poder,

sentando las bases para lo que hoy conocemos como *Análisis político*.

El concepto de *democracia*[1], describe una idea universal de gobernanza basada en la participación ciudadana; por lo que *nación democrática* integra el concepto de democracia al contexto de nación, con estructuras, instituciones y normas que reflejan los valores democráticos. En una nación democrática, la política se encuentra estrechamente vinculada al concepto de poder público, ya que es a través de esta que el poder se delega en representantes encargados de tomar decisiones en nombre de la ciudadanía. Este concepto de poder público constituye la base fundacional de los Poderes del Estado, idea desarrollada por el filósofo político francés *Charles de Montesquieu* en el siglo XVIII. En su obra *El espíritu de las leyes*, Montesquieu describe la separación funcional del poder público en tres ramas: el *Poder Ejecutivo*, encargado de administrar el Estado, implementar y hacer cumplir las leyes; el *Poder Judicial*, responsable de interpretar y aplicar la ley; y el *Poder Legislativo*, dedicado a crear las normas que rigen la vida de la nación.

En el seno del Poder Legislativo, los representantes de la ciudadanía debaten ideas, respaldan puntos de vista, posturas ideológicas e intereses colectivos. Estos debates abarcan temas de diversa naturaleza (cíclicos, coyunturales, programáticos, entre otros) y quedan registrados en múltiples medios: la prensa escrita o digital, las redes sociales en las que participan y, documentación oficial, durante las sesiones de trabajo legislativo a través de las instituciones a las que pertenecen (Congreso, Asamblea de Representantes o Parlamento, por mencionar algunas), variando la denominación de estas corporaciones según la nación, la cultura y el sistema político de cada país.

En este marco, el registro de los debates e interacciones legislativas, donde convergen posturas ideológicas, intereses ciudadanos y decisiones programáticas, se convierte en una fuente invaluable para comprender las dinámicas del poder político. Estos registros, no solo evidencian las tensiones y negociaciones propias de la toma de decisiones, sino que también ofrecen una base concreta para analizar las relaciones de poder, las alianzas estratégicas y otros elementos que configuran el orden de la "red social". Entonces es en esta información donde reside gran parte del registro físico del *poder político*, entendiendo el poder político como la capacidad de un actor, para influir, dirigir o controlar las acciones, decisiones y recursos dentro de una estructura social, con el objetivo de organizar y gobernar una comunidad.

El poder político, constituye el eje central del análisis político, y por su naturaleza social y dinámica se encuentra en constante tensión y sujeto a transformaciones que reflejan las complejidades de un orden social. En este contexto, el análisis político emerge como una herramienta indispensable para comprender las causas, equilibrios, negociaciones y disputas que lo estructuran, y al mismo tiempo, la actividad legislativa se consolida como uno de los escenarios privilegiados donde la dinámica del poder político se manifiesta y evoluciona, reflejando los intereses, tensiones y acuerdos que surgen de la interacción entre diversos actores estatales y sociales. Por ello, el análisis político encuentra en el ámbito legislativo un campo de estudio esencial, dado que las negociaciones, las normas y las decisiones adoptadas en este espacio revelan cómo se ejerce, equilibra y transforma el poder.

De este modo, el análisis político-legislativo tiene como objetivo brindar diagnósticos precisos sobre los actores, los temas y las relaciones de poder asociadas a determinados grupos que intervienen en el proceso de generación de la ley, por lo que resulta valioso para un amplio

[1]En la antigua Grecia, la democracia no contaba con la aprobación general de los pensadores, quienes la consideraban propensa a la inestabilidad y al desorden debido a la posible falta de cualificaciones de los gobernantes elegidos por el pueblo. Aunque algunos la defendían, como Heródoto o Pericles, otros, como Sócrates, Platón y Aristóteles, se mostraban críticos y la consideraban una forma de gobierno inferior o inadecuada.

espectro de interesados, dentro de los cuales están:

- *Actores políticos:* Los parlamentarios y partidos políticos necesitan de un análisis detallado para preparar propuestas de ley coherentes con las demandas ciudadanas y las realidades sociales, como también para impulsar el posicionamiento de actores o temas en diversos contextos. Esta información les ayuda a anticipar las repercusiones de sus iniciativas, a negociar con otras bancadas y a diseñar estrategias de comunicación efectivas ante el electorado.

- *Asesores parlamentarios:* Contar con información de análisis político-legislativo permite realizar tareas de asesoría de forma proactiva por oferta, al detectar preliminarmente las necesidades y temas de interés de los parlamentarios, e identificando oportunamente las tendencias de la coyuntura política para anticipar la evolución del debate público. También esta información facilita la elaboración de argumentos sólidos y la verificación de datos en tareas de asesoría como minutas o discursos, lo que fortalece la credibilidad de los legisladores frente a la opinión pública, como también analizar cuotas de votaciones y posiciones de los legisladores, para prever alianzas y conflictos.

- *Ciudadanía y bases votantes:* Resulta fundamental contar con información de análisis político-legislativo al momento de ejercer un voto informado, como también para participar en debates públicos. Información acerca de las posturas o la labor de sus representantes, o entender el alcance de las decisiones legislativas en temas como derechos sociales, inversión pública o políticas de salud, favorece la formación de opiniones críticas e impulsa acciones ciudadanas.

- *Investigadores y académicos:* Profesores, investigadores y estudiantes de ciencias sociales, derecho, política y otras disciplinas se benefician del acceso a información sistemática sobre la producción legislativa y su impacto social. Estos insumos son esenciales para la investigación empírica, la discusión de teorías y la generación de nuevo conocimiento.

- *Organizaciones ciudadanas:* Organizaciones que representan intereses específicos - medioambientales, culturales o de derechos humanos, entre otros - necesitan conocer el panorama de la actividad legislativa para actuar e incidir en los procesos de toma de decisiones. El análisis político-legislativo les ofrece datos y argumentos para sustentar sus demandas ante las autoridades, como también para coordinar campañas de incidencia con actores específicos y basadas en evidencia.

- *Sector privado:* Las empresas, tanto grandes como pequeñas, se ven directamente afectadas por cambios en la legislación que pueden influir en áreas como la fiscal, el comercio exterior o la regulación laboral. Disponer de información de análisis político-legislativo les permite planificar inversiones y operaciones de manera más certera, participar proactivamente en la elaboración de políticas públicas que favorezcan la competitividad y el crecimiento, posicionar actores relevantes a sus intereses y promover acciones dirigidas a influir en la toma de decisiones *(lobby)*.

Con base en esto, resulta especialmente interesante analizar la información oficial que emana del Poder Legislativo, como también puede resultar interesante explorar aquella difundida por medios como la prensa y las redes sociales, y de esta manera responder preguntas que abarcan diversas dimensiones de análisis, tales como las siguientes:

- ¿Quién cumple un rol clave en el contexto de un tema específico?
- ¿Cuál es el nivel de cohesión política de un grupo frente a un tema particular?
- ¿Cuáles son los temas de mayor relevancia para un representante?

Para despejar este supuesto y establecer fuentes de datos que permitan responder estas preguntas, un análisis exploratorio (desarrollado en el Capítulo 2) muestra que los datos oficiales sobre debate parlamentario cuentan con las características idoneas para responder de forma precisa preguntas como las planteadas por esta investigación. Solo para adelantar, algunas de estas características son un relativo equilibrio en el número de documentos por parlamentario, ser una fuente de libre acceso o que las temáticas abordadas se relacionen al trabajo político-legislativo.

Es entonces cuando las *Tecnologías Semánticas* emergen como un eslabón articulador para procesar y estandarizar los datos, facilitando el análisis. Para efectos de esta tesis, el concepto de Tecnologías Semánticas combina tres grandes áreas técnicas de la informática: la Web Semántica, que aporta modelos para representar el conocimiento y mecanismos estandarizados de interoperabilidad; la Minería de Textos, que facilita la extracción automatizada de información relevante desde grandes volúmenes de documentos no estructurados; y el Análisis de Redes Sociales (SNA), que permite modelar, analizar y visualizar relaciones complejas entre actores políticos y entidades relevantes. Esta integración tecnológica posibilita transformar conjuntos aislados de datos en conocimiento estructurado, interoperable y listo para análisis cuantitativos rigurosos y reproducibles. Considerando esto, se propone el diseño de un marco tecnológico integrado que reúne los componentes y mecanismos de interoperabilidad necesarios para aplicar Tecnologías Semánticas al análisis político-legislativo. Este marco permitirá disponer, de manera integral y estandarizada, de datos para análisis cuantitativos, facilitando así el desarrollo de herramientas capaces de responder tanto preguntas definidas previamente como nuevas interrogantes que surjan en este ámbito.

De esta manera, el enfoque planteado de esta tesis es validar que a través del uso de tecnologías semánticas, es posible procesar datos no estructurados y generar análisis político-legislativo de utilidad, a la vez que se mejora su reproducibilidad y eficiencia. Para validar esto, se implementarán tres instrumentos específicos que utilizan tecnologías semánticas desde la fase de procesamiento de datos hasta su implementación final, cada uno orientado a responder una de las preguntas de análisis previamente planteadas, los cuales en conjunto serán evaluados por un grupo de usuarios expertos del ámbito político-legislativo. Ya que el ámbito político-legislativo responde a realidades locales nacionales, todo el desarrollo del trabajo experimental y la descripción de los pormenores técnicos, se realizará en el contexto del Congreso Nacional chileno, tomándolo como caso de estudio.

1.2 Chile como caso de estudio

1.2.1 El contexto político-legislativo chileno

Desde el punto de vista político-legislativo, Chile es una república donde se ejerce el modelo de división de los tres poderes del estado definida por Montesquieu. El Poder Legislativo está conformado por un Congreso Nacional de carácter bicameral, compuesto por dos organismos independientes: la Cámara de Diputadas y Diputados (cámara baja) integrado por 155 parlamentarios, y el Senado (cámara alta) integrado por 50 parlamentarios. Adicionalmente, dentro

del Congreso Nacional existe un tercer organismo no legislador, la Biblioteca del Congreso Nacional, el cual realiza labores de asesoría, documentación y apoyo a la labor legislativa. En este espacio, se debaten y aprueban las leyes que rigen la vida del país, así como se ejerce la fiscalización de las autoridades. Pese a su importancia institucional, el presupuesto que el Congreso recibe para el año 2025 representa menos del 0,2% del total nacional, lo que refleja un gasto relativamente limitado en comparación con el conjunto de la administración pública. Esta información es posible de obtener desde la visualización del presupuesto[2], una herramienta basada en datos abiertos y visualizaciones [Cifuentes-Silva et al., 2020] desarrollada durante la investigación.

Desde el punto de vista del funcionamiento, cada una de las cámaras del Congreso chileno sesiona tanto en sala de sesión, momento en que todos los parlamentarios de la cámara participan, como también sesiona en comisiones temáticas, instancia donde participan grupos reducidos de parlamentarios que realizan evaluación de los proyectos de ley en torno al tema de la comisión. A partir tanto de las sesiones de sala, como en las sesiones de comisión de ambas corporaciones, se generan documentos que dan cuenta de la actividad legislativa, la cual es desarrollada dentro de un espacio de tiempo que se denomina *legislatura*. Una legislatura corresponde al periodo de sesiones, que por lo general comienza el 11 de marzo de un año, hasta el 10 de marzo del año siguiente.

En el caso de las sesiones de sala, se emanan una variedad de documentos dentro de las que destaca el *diario de sesión* como documento central de la sesión. El diario de sesión es un documento que recopila tanto la transcripción de todo el debate legislativo realizado en la sala de sesión, como también todos los documentos anexos que se presentan en la sesión, sean estos de una variedad de tipos como oficios, proyectos de ley, solicitudes de distinta índole y muchos otros. A la fecha, el diario de sesión es transcrito por taquígrafos a partir tanto de la sesión en sala como de correcciones posteriores apoyadas por grabaciones en video de la sesión.

En el caso de las sesiones de comisión, la realidad es distinta a las sesiones de sala, ya que a la fecha no existe un diario específico con la transcripción de la sesión, pero sí existen otros documentos clave tales como los vídeos de la sesión, y en el caso de la tramitación de proyectos de ley, los *informes de comisión* asociados al proyecto.

El primer Congreso Nacional de Chile se instaló el 4 de julio de 1811, y su primera acta de sesión se redactó al día siguiente, el 5 de julio. En ese entonces, el Congreso chileno estuvo compuesto por una sola cámara e integrado por 41 diputados. Esta primera versión del Congreso implementó importantes iniciativas, entre ellas, la declaración de libertad de vientres, que liberó de la esclavitud a los hijos de esclavos; la creación del Instituto Nacional y la Biblioteca Nacional; la proclamación de la libertad de comercio; la abolición de los derechos parroquiales; el Reglamento de Instrucción Primaria y la Ley de Libertad de Prensa. Además, promulgó el Reglamento para la organización del poder ejecutivo provisorio de Chile, aprobado el 14 de agosto de 1811, que buscaba definir la separación de poderes y establecer la supremacía del Congreso en las decisiones públicas [BCN, 2025].

Luego de más de 200 años de historia y de una serie de eventos que han moldeado la realidad político-legislativa, la conformación actual del Congreso difiere considerablemente tanto en su estructura orgánica, en el número de representantes y en la diversidad de temas que en él se discuten. En la actualidad, el Congreso se compone de dos cámaras: la Cámara de Diputados y el Senado. Cada una de estas cámaras cuenta con un conjunto de comisiones legislativas permanentes que facilitan el análisis y la discusión de temas específicos; en total, existen 56

[2]Disponible en `https://www.bcn.cl/presupuesto`

comisiones, de las cuales 28 pertenecen al Senado y 28 a la Cámara de Diputados, además de un número variable de comisiones investigadoras que se constituyen según la necesidad de abordar cuestiones puntuales.

La actividad legislativa actual se evidencia en el número de sesiones que se llevan a cabo en cada legislatura: durante los últimos 10 años, la Cámara de Diputados ha celebrado en promedio 143,1 sesiones, mientras que el Senado ha alcanzado aproximadamente 118,9 sesiones. Esta dinámica se traduce en una labor legislativa constante, reflejada también en las 92,1 leyes que se aprueban en promedio anualmente .

El flujo de iniciativas es igualmente notable, pues cada año ingresan un promedio de 630 proyectos de ley en forma de mociones y 83 en forma de mensajes presidenciales desde el poder ejecutivo. Esta intensa actividad genera una gran producción documental, la que equivale a un promedio de 2.566 documentos oficiales generados anualmente. Este volumen de información resalta la importancia de contar con sistemas que faciliten la organización y procesamiento para un posterior análisis de datos.

Un elemento crucial para el estudio y preservación de la memoria política de Chile es el archivo histórico de los diarios de sesión del Congreso, resguardado por la BCN. Este archivo, que se remonta a 1811, ha sido digitalizado a partir de 1965, encontrándose gran parte disponible en formatos como Word y XML, lo que ha permitido su procesamiento y transformación en nuevos productos.

1.2.2 Historia de la Ley y Labor Parlamentaria

Es en este contexto que en 2011 la BCN inició un proyecto para automatizar la elaboración de la *Historia de la Ley* a partir de los diarios de sesión disponibles en texto. Una Historia de la Ley (HL) es la recopilación de todos los documentos generados durante la tramitación legislativa de una ley; desde la iniciativa que da origen al proyecto de ley, pasando por su discusión en el Congreso, los informes de las comisiones parlamentarias que lo estudiaron y las transcripciones de los debates en las salas de sesiones, reuniendo su trazabilidad dentro del proceso legislativo. El objetivo de la HL es recopilar el llamado *espíritu de la ley*, posibilitando su interpretación de manera precisa en relación con el alcance y sentido que se otorgó a la norma cuando fue discutida. Este instrumento jurídico es particularmente útil tanto para los jueces al preparar sentencias como para los abogados cuando utilizan determinadas normas para fundamentar sus argumentos.

De manera similar, la Labor Parlamentaria (LP) es una recopilación de toda la actividad legislativa realizada por un parlamentario durante el ejercicio de su cargo, siempre que haya sido registrada en medios impresos pertenecientes al poder legislativo, como una moción parlamentaria, un diario de sesiones o un informe de comisión.

En Chile hasta ese entonces, ambos productos eran elaborados por analistas jurídicos solo para solicitudes específicas, procesando manualmente cada documento relacionado con una ley.

Para la elaboración electrónica y automatizada de ambas recopilaciones documentales, se requiere disponer de una base de datos granular que registre todos los documentos del proceso legislativo donde se haga cualquier referencia a proyectos de ley o parlamentarios, permitiendo posteriormente extraer y recuperar selectivamente qué se discutió en torno a un proyecto de ley que se convertirá en ley, así como lo que un determinado legislador ha dicho en cualquier contexto.

Por esta razón, se propuso como solución técnica el marcaje de los diarios de sesión y otros documentos del proceso legislativo, transformando el texto a formato XML. Este procedimiento

permitiría identificar y etiquetar las secciones y subsecciones en las que se abordara un proyecto de ley específico, una persona determinada o temas particulares.

Para llevar a cabo esta tarea, se evaluaron distintas soluciones técnicas basadas en XML disponibles en ese momento, siendo el actual estándar Akoma-Ntoso (AKN) el que presentaba mayor proyección. Este esquema permitía la incorporación de marcas mediante tecnologías de Web Semántica, como URIs y ontologías, maximizando el uso de estándares de interoperabilidad. Además, proporcionaba una semántica precisa para el marcado del debate parlamentario, abarcando todo tipo de metadatos y documentos generados durante el proceso legislativo, incluidas las intervenciones parlamentarias, ofreciendo mecanismos de extensión en caso de ser necesario.

A partir de esta necesidad técnica es que en la BCN se decide implementar el primer portal de datos abiertos legislativos de Latinoamérica [Cifuentes-Silva et al., 2011], que incluía entre otros tipos de datos, la versión en datos abiertos del sistema Ley Chile[3] (todas las normas de la base de datos legal chilena en RDF), un conjunto de reseñas biográficas parlamentarias en RDF (con URIs para personas, organismos, cargos y otros), así como un conjunto de ontologías que modelaban dichos datos. Esta infraestructura constituía un insumo crítico para el desarrollo de un sistema basado en estándares de Web Semántica.

Procesamiento de documentos legislativos

Habiendo definido el estándar de marcaje, para llevar a cabo el procesamiento de documentos se diseñaron procesos de negocio que fueron implementados en un entorno de *workflow*. Estos procesos dividen y sistematizan las tareas necesarias de planificación, ejecución, aseguramiento de la calidad (QA) y publicación, entre otras. Un esquema simplificado del flujo de proesamiento de documentos y su transformación en los productos HL y LP se presenta en la figura 1.1.

Este proceso, cuyos aspectos técnicos se detallan en un trabajo desarrollado durante la investigación [Cifuentes-Silva and Labra Gayo, 2019], culmina con la generación de dos elementos fundamentales para este estudio, adicionales a los productos HL y LP, que obviamente constituyen la motivación principal de toda la infraestructura desarrollada. El primer elemento es el conjunto de documentos de debate parlamentario marcados y enriquecidos en formato AKN, a partir de los cuales es posible extraer una amplia variedad de información estructurada. El segundo es la representación en RDF de los datos extraídos desde el formato AKN, que constituye un insumo primario para el desarrollo de una ilimitada gama de aplicaciones.

Este conjunto de datos abiertos enlazados generados a partir de las intervenciones parlamentarias, se suma a una amplia variedad de datos del ámbito político-legislativo que publica la BCN a través del portal de datos abiertos `datos.bcn.cl`.

Datos disponibles del ámbito político-legislativo

La BCN pone a disposición de la comunidad una amplia variedad de datos abiertos a través de su portal de datos. Esta oferta incluye tanto conjuntos de datos como ontologías, disponibles mediante URIs desreferenciables y consultas SPARQL a través del endpoint accesible en dicho portal. Un ejemplo del uso de estos datos se presenta en la figura 1.2, que muestra una consulta SPARQL diseñada para obtener las militancias y personas registradas en la base de datos RDF de la BCN.

[3] `https://www.bcn.cl/leychile`

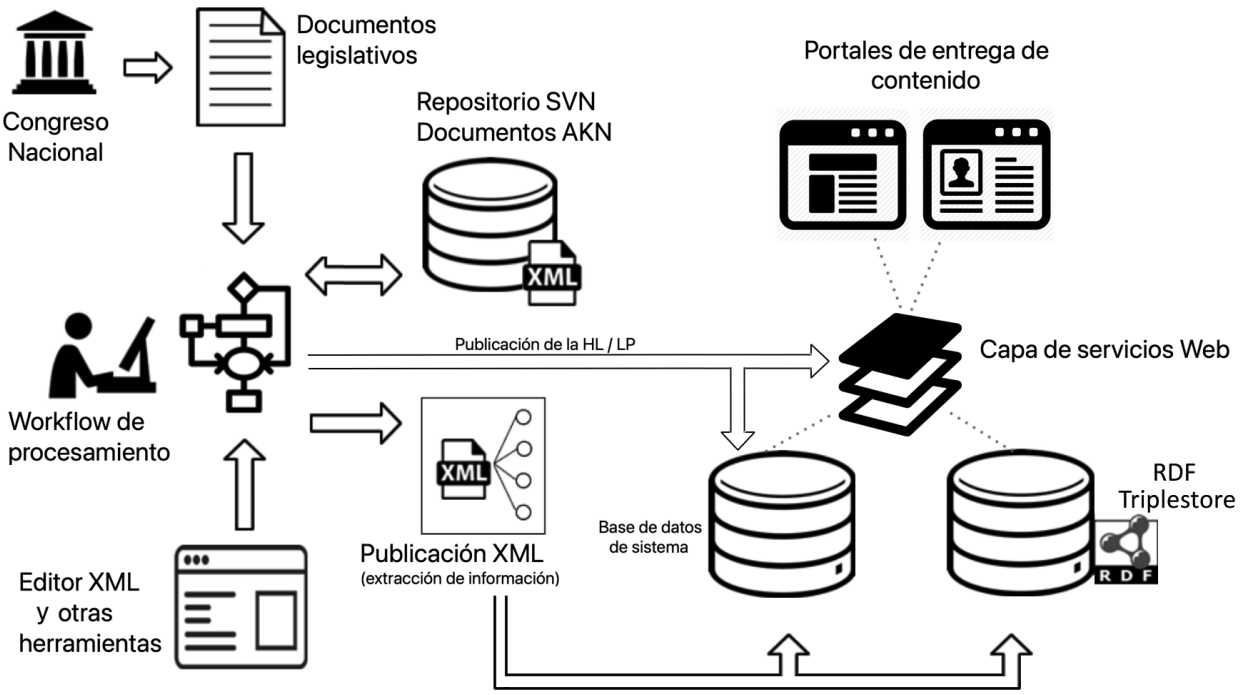

Figura 1.1: Proceso de generación de la Historia de la Ley y la Labor Parlamentaria

En este ejemplo, los datos de personas, disponibles como Linked Open Data, están definidos con base en la ontología de Biografías Parlamentarias[4] descrita en RDF. Al ser los datos abiertos, es posible obtener una representación de ellos en URIs desreferenciables. La figura 1.3 muestra una representación del recurso RDF en HTML para una persona, en particular, del actual presidente de Chile, Gabriel Boric Font.

Lo que se muestra es que gracias a la utilización de tecnologías de datos abiertos y del modelo RDF, es posible disponer de un nivel de granularidad detallado en los datos sobre personas y partidos políticos, lo que habilita un análisis exhaustivo y flexible. Esta granularidad comprende información sobre personas, partidos políticos, militancias, periodos parlamentarios, cargos, lugares geográficos, normas jurídicas, proyectos de ley, documentos legislativos (como diarios de sesión e informes de comisión), intervenciones parlamentarias y otras entidades relevantes. El uso de RDF no solo permite estructurar esta información de manera interoperable y estandarizada, sino que también facilita su integración con otros conjuntos de datos, tanto internos como externos, lo que amplía significativamente las posibilidades de análisis en el ámbito legislativo. Esta capacidad de representar relaciones complejas y de aplicar consultas semánticas mediante SPARQL ha sido fundamental para el desarrollo de los análisis presentados en esta investigación.

Adicionalmente, para los experimentos realizados en esta tesis, la BCN dispone de dos bases de datos complementarias de carácter no abierto: una referida a noticias de prensa, que será descrita en detalle en la sección 2.4, y otra que contiene datos de redes sociales (Twitter) de parlamentarios, cuyo detalle se presentará en la sección 2.5. Todas estas fuentes de datos, que serán evaluadas en su utilización, enriquecen el análisis al proporcionar contextos políticos, mediáticos y sociales que permiten estudiar la interacción entre el debate legislativo formal y su reflejo en los medios de comunicación y redes sociales.

[4]Accesible en `http://datos.bcn.cl/ontologies/bcn-biographies/doc/`

Figura 1.2: Consulta SPARQL para obtener datos de parlamentarios

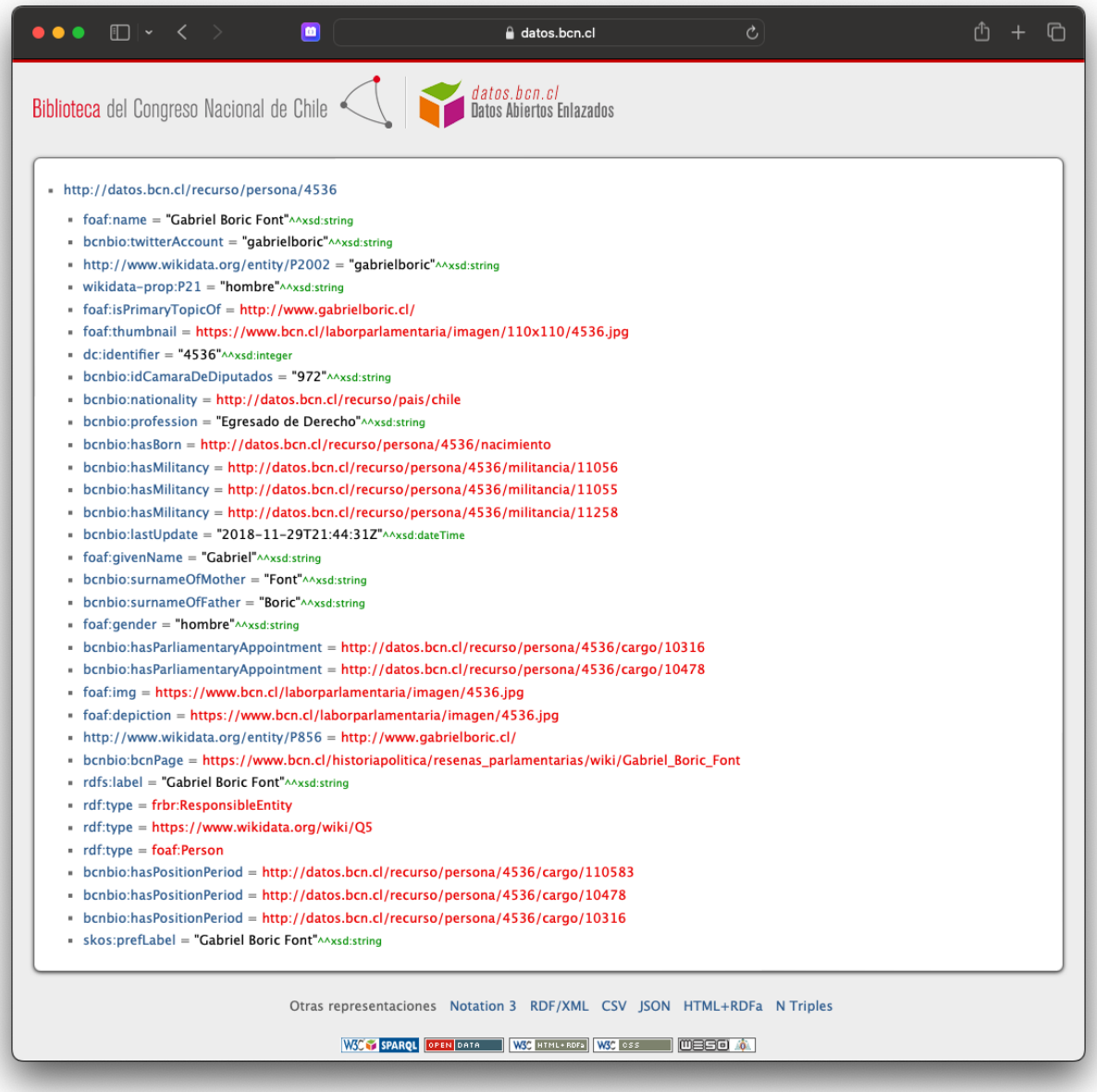

Figura 1.3: Representacion del recurso RDF de los datos de una persona

Capítulo 2

Análisis exploratorio preliminar

2.1 Introducción

El presente capítulo tiene como objetivo realizar un análisis exploratorio preliminar de los datos que servirán como base para el desarrollo del estudio. Para realizar esta sección, primeramente se caracterizarán las distintas fuentes de información consideradas, evaluando su pertinencia, calidad e idoneidad para los análisis previstos, aunque dado el volumen de datos y su variedad (en cuanto a distintos parlamentarios y contexto político), se seleccionará una ventana de un año para el análisis. Dado que se analizarán datos asociados al ámbito legislativo, se utilizará un periodo legislativo también denominado legislatura. La legislatura seleccionada corresponde a la número 367 que va desde el 11 de marzo de 2019 hasta el 10 de marzo de 2020.

De esta manera, se examinarán las bases de datos inicialmente propuestas, identificando su estructura, volumen y cobertura temática, con el propósito de justificar su inclusión o exclusión en el estudio. Este proceso es fundamental para asegurar que las decisiones metodológicas se fundamenten en un conocimiento detallado de los datos disponibles, permitiendo anticipar posibles limitaciones. Para este fin, se utilizarán estadísticas descriptivas e indicadores que faciliten la comparación de los conjuntos de datos, además de su forma y características, intentando develar posibles debilidades dentro de los conjuntos de datos, lo que permitirá establecer una base sólida para los análisis posteriores.

2.2 Caracterización de la conformación del Congreso Nacional Chileno

El Congreso Nacional Chileno está compuesto por la Cámara de Diputadas y Diputados (cámara baja) y el Senado (cámara alta). A la fecha de análisis[1], la cantidad de parlamentarios corresponde a 198 integrantes, de los cuales 155 corresponden a la cámara baja y 43 a la cámara alta. El gráfico 2.1 muestra la distribución de las edades de los parlamentarios por sexo por cada una de las cámaras. Los gráficos permiten observar que para todos los rangos etarios, la frecuencia de parlamentarias es más baja. En particular, a la fecha de estudio, un 22,7% de las parlamentarias son de sexo femenino (45 personas), mientras que un 77,3% son de sexo masculino (153 personas).

[1] 11 de marzo de 2019

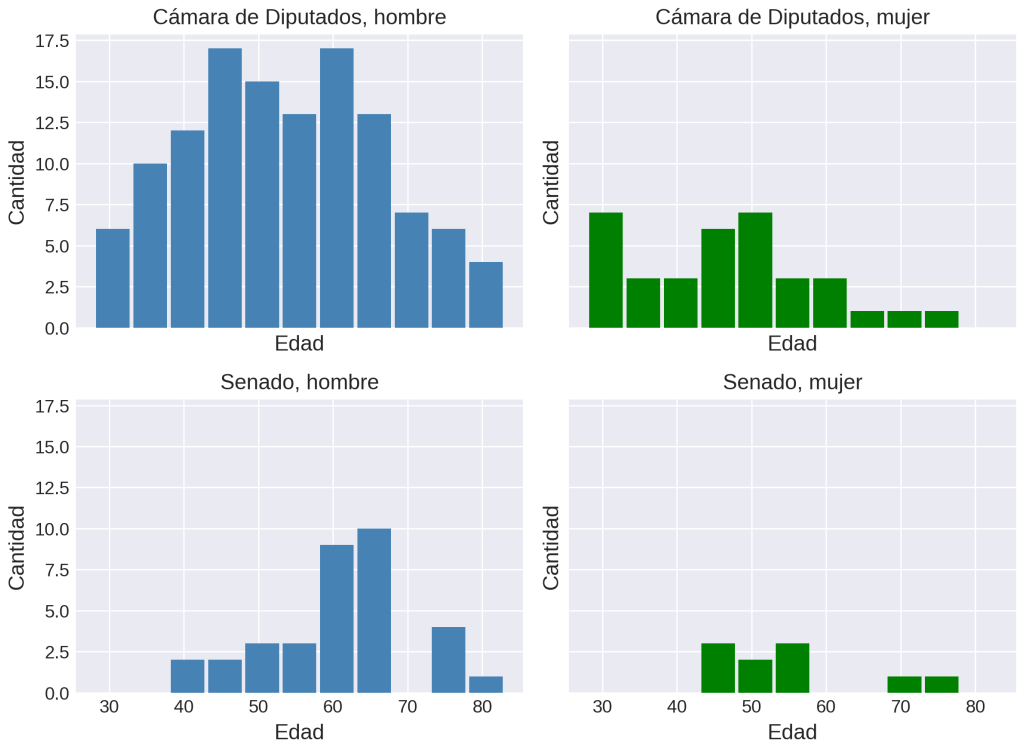

Figura 2.1: Distribución de edades de parlamentarios por cámara y sexo

En cuanto a la distribución por partido político, a la fecha de estudio el Congreso chileno está compuesto por 17 partidos políticos más un grupo de parlamentarios y parlamentarias independientes (aunque normalmente cuentan con una afinidad política determinada y conocida). La figura 2.2 muestra la distribución por partido político y sexo de los parlamentarios analizados, destacando con tonos más claros a los independientes.

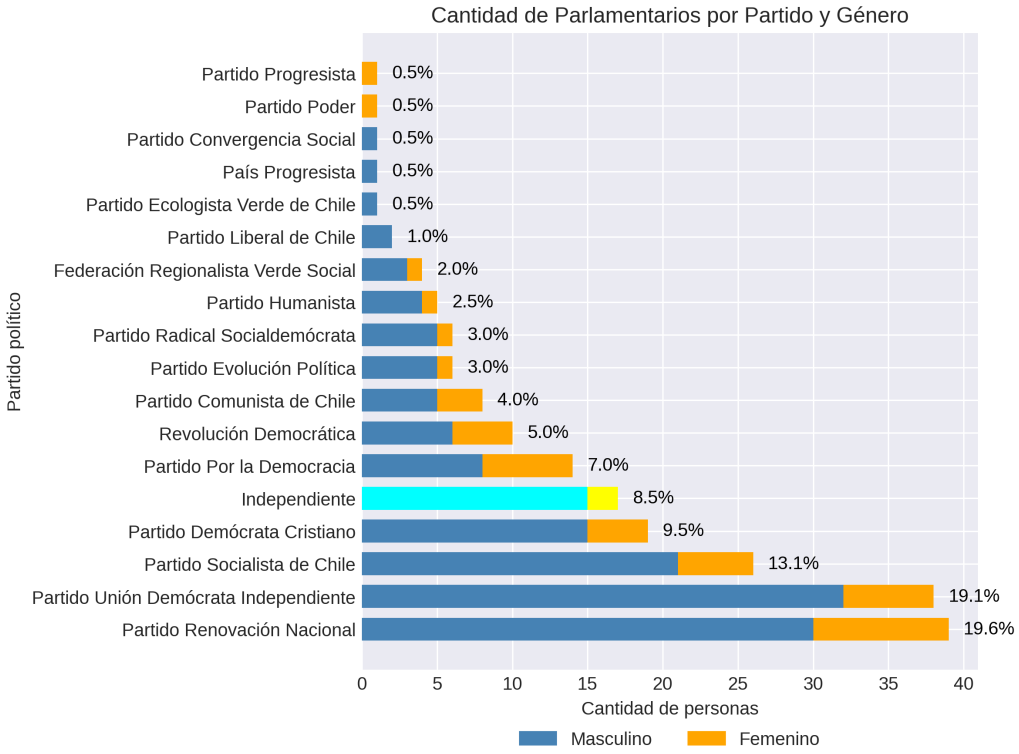

Figura 2.2: Distribución de parlamentarios por militancia a partido político

2.3 Documentos del Congreso Nacional

2.3.1 Conjunto de documentos fuente

Acorde a los datos expuestos en el portal de datos abiertos, en el periodo seleccionado (legislatura 367) existen 118 sesiones del senado y 161 sesiones de la cámara de diputados, de los cuales se cuenta con 263 diarios de sesiones, siendo 108 diarios generados por el Senado y 155 por la Cámara de Diputadas y Diputados. Esta diferencia entre el número sesiones y diarios de sesión puede deberse a sesiones fracasadas (por falta de quorum), secretas o suspendidas, sesiones que no generaron un diario de sesión.

Los diarios de sesiones, que en su fuente están disponibles en formato de texto plano, han sido procesados para obtener una representación en XML bajo el estándar Akoma-Ntoso y posteriormente se ha generado una representación en RDF del documento, lo cual se presenta en la sección 1.2.

En cada uno de estos diarios, que son una transcripción de la sesión llevada en la sala del Congreso, se describe todo el debate parlamentario llevado a cabo en la sesión, además de la asistencia, las votaciones sobre diversos temas (proyectos de ley, proyectos de acuerdo, acusaciones y otros) y los distintos documentos que se dan cuenta en la sesión, tales como mociones parlamentarias, oficios a autoridades u organismos sobre distintas materias, mensajes presidenciales y comunicaciones de distinta índole.

El proceso de marcaje llevado a cabo desde el texto plano hasta su representación en XML, permite incorporar una serie de marcas sobre el texto que permiten identificar secciones estructurales del documento, sub documentos, entidades y metadatos de cada sección y subsección. De esta manera, el documento fuente, es enriquecido con marcas que posteriormente permiten

identificar y extraer la información que se requiere, tal como las intervenciones de parlamentarios específicos, la asistencia a la sesión o las votaciones sobre la tramitación de un proyecto de ley. De hecho, para el desarrollo de esta tesis se ha utilizado un conjunto de intervenciones obtenidas desde el endpoint SPARQL `http://datos.bcn.cl/sparql`, las cuales han sido extraídas desde los documentos XML.

2.3.2 Intervenciones parlamentarias

Una *intervención* en el contexto de un documento legislativo, corresponde al registro en texto de cuando un legislador toma la palabra durante una sesión o comisión para presentar una idea, defender una postura o realizar una consulta. Este texto corresponde a la transcripción fidedigna de lo que dijo el o la parlamentaria y es extraído directamente desde un diario de sesión, informe de comisión u otro. De forma similar, una *participación* corresponde a cualquier forma de pronunciamiento de un parlamentario en el documento legislativo, ya sea mediante una intervención, una autoría, una adherencia a una iniciativa u otro, por lo cual representa un caso más general de intervenir. De esta manera, una intervención puede considerarse un caso particular de participación, aunque para este trabajo de investigación se utilizarán ambos conceptos de forma indistinta.

El conjunto de documentos analizado contiene 19.990 textos de participaciones de parlamentarios en sesiones de sala, las cuales dan cuenta del debate realizado tanto en el Senado (4.733 intervenciones) y la Cámara de Diputadas y diputados (15.257 intervenciones). Dentro de este número, durante el análisis de los datos se identificó un total de 258 participaciones asociadas a personas que no corresponden al periodo, lo cual representa un 1,29% de los datos. Al ser los textos de participaciones de distintos tipos tales como intervenciones en sala, oficios o mociones, muchos de ellos están asociados a más de una persona, razón por la cual el número de participaciones de distintos parlamentarios corresponde a 29.045. La tabla 2.1 muestra las estadísticas descriptivas del número de participaciones por diarios de sesiones asociados a cada una de las cámaras.

Conjunto	**N**	**Media**	**Mín.**	**Q1**	**Med.**	**Q3**	**Máx.**
Senado	4.733	43	2	28	40	48	254
Cámara de Diputadas y Diputados	15.257	99	8	39	71,5	113	447
Congreso Nacional	19.900	76	2	33	50	88,75	447

Tabla 2.1: Estadísticas descriptivas de participaciones en diarios de sesión

En la figura 2.3 los gráficos de cajas muestran la distribución de documentos por persona tanto a nivel completo del Congreso Nacional, como por cada una de las cámaras. Se visualiza que aunque el número de sesiones y en consecuencia de documentos es menor en el Senado, los totales de intervenciones son muy similares entre miembros de ambas cámaras. A nivel general dentro de este conjunto de datos, se calcula un coeficiente de variación (CV) de *0,53* asociado al número de intervenciones por persona. Una tabla con las estadísticas descriptivas de número de intervenciones por persona y cámara se presenta en la tabla 2.2.

Como se visualiza, para el periodo en los documentos del Senado existen participaciones para 53 personas, y en la cámara baja para 163, teniendo en cuenta que en ambos casos estos números exceden al número de integrantes. Esto significa que los diarios de sesiones muchas

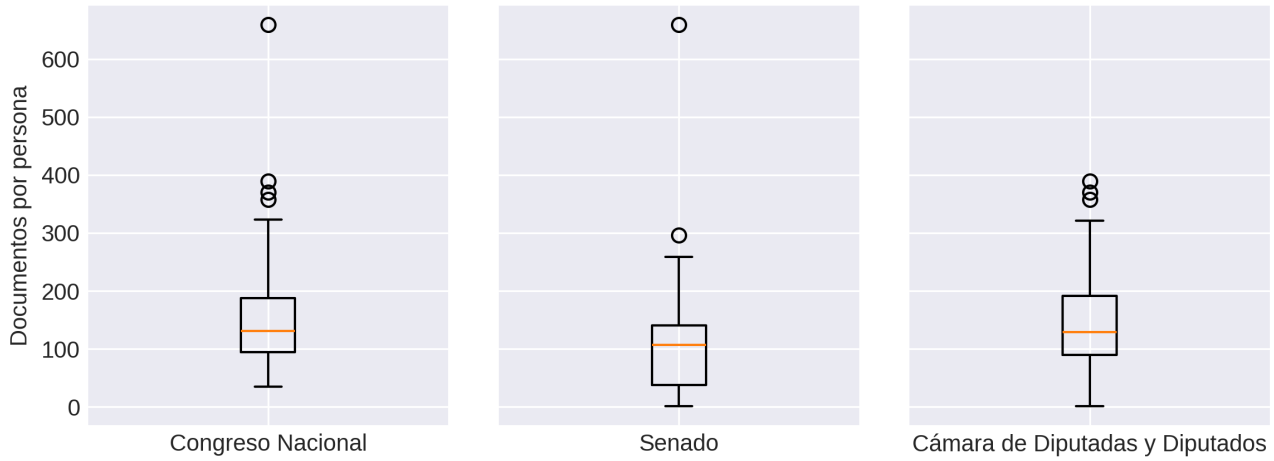

Figura 2.3: Distribución de participaciones por persona en el Congreso Nacional de Chile

Conjunto	N	Media	Mín.	Q1	Med.	Q3	Máx.	Std.	CV
Senado	53	111,7	1	38	107	140	659	106,2	0,95
Cámara de Diputadas y Diputados	163	140,4	1	90	129	191	389	73,4	0,52
Congreso Nacional	198	145,4	35	94	131	187,3	659	76,8	0,53

Tabla 2.2: Estadísticas descriptivas de número de intervenciones de parlamentarios por cámara

veces incorporan documentos, por ejemplo mociones, donde figuran parlamentarios de ambas cámaras.

Otro dato relevante es que un 89,3% de los documentos pertenecen a una única persona (17.782), y solo un 10.7% de los documentos (2.118 intervenciones) pertenecen a más de una persona. El gráfico de la figura 2.4 muestra la distribución de este fragmento de los datos.

Desde el punto de vista del análisis del texto, la tabla 2.3 muestra las estadísticas descriptivas asociadas al número de palabras por documento representativas de la muestra. Complementariamente, los gráficos de la figura 2.5 muestran la distribución de palabras por documento, como un histograma en escala logarítmica, que muestra los totales de palabras por documento donde se permiten visualizar claramente los valores atípicos.

Total intervenciones	Media	Mínimo	Q1	Mediana	Q3	Máximo
27.192	607	5	71	286	870	31.714

Tabla 2.3: Estadísticas descriptivas de total de palabras por participación

La figura 2.6 muestra el gráfico de deciles de contribución, que permite observar cómo se distribuye la cantidad total de documentos asociados a parlamentarios, agrupados en segmentos de igual tamaño (deciles). Cada barra representa el porcentaje del total de documentos que aporta un 10% específico de los parlamentarios, ordenados desde aquellos con más documentos hasta aquellos con menos. En este contexto, el gráfico muestra que el 10% de parlamentarios

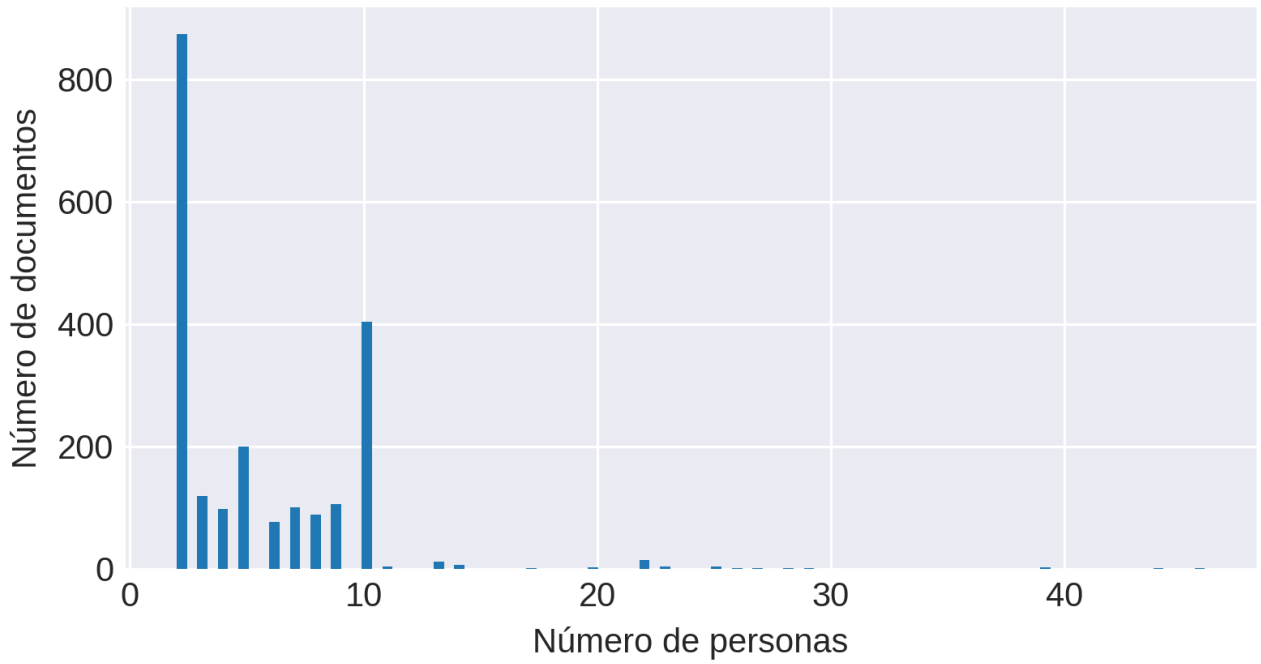

Figura 2.4: Documentos asociados a más de una persona

que más documentos tiene asociados en este conjunto aporta con el 20% del total de documentos del conjunto, mientras que el 10% que menos tiene aporta con un 5.5% de los documentos.

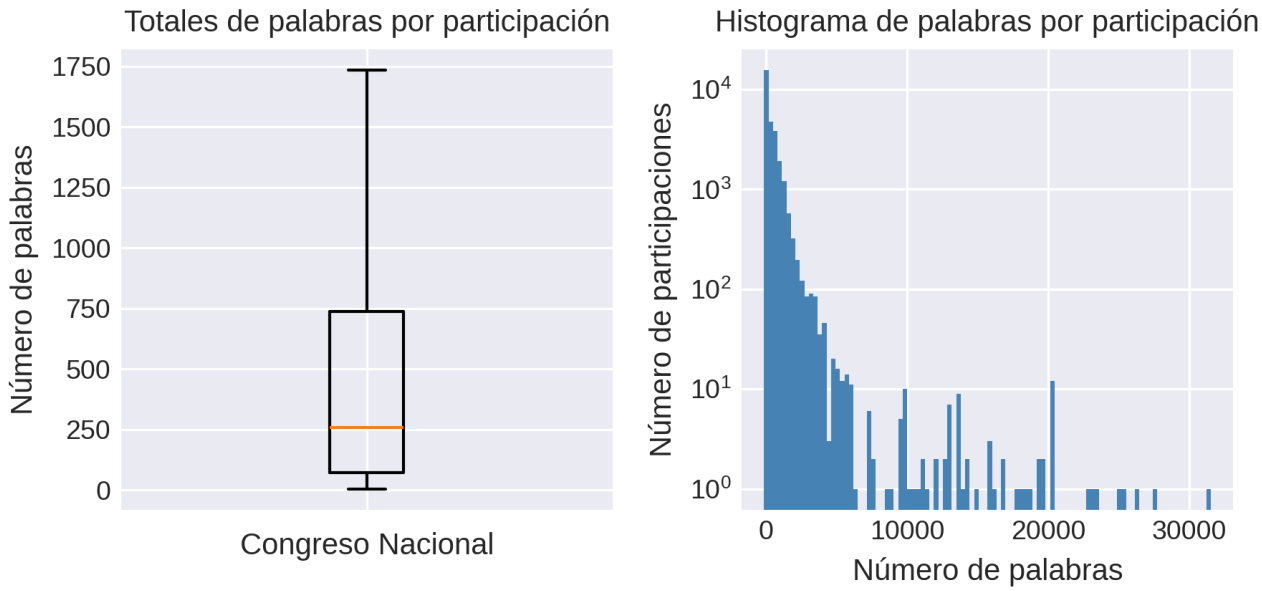

Figura 2.5: Distribución de palabras por participación

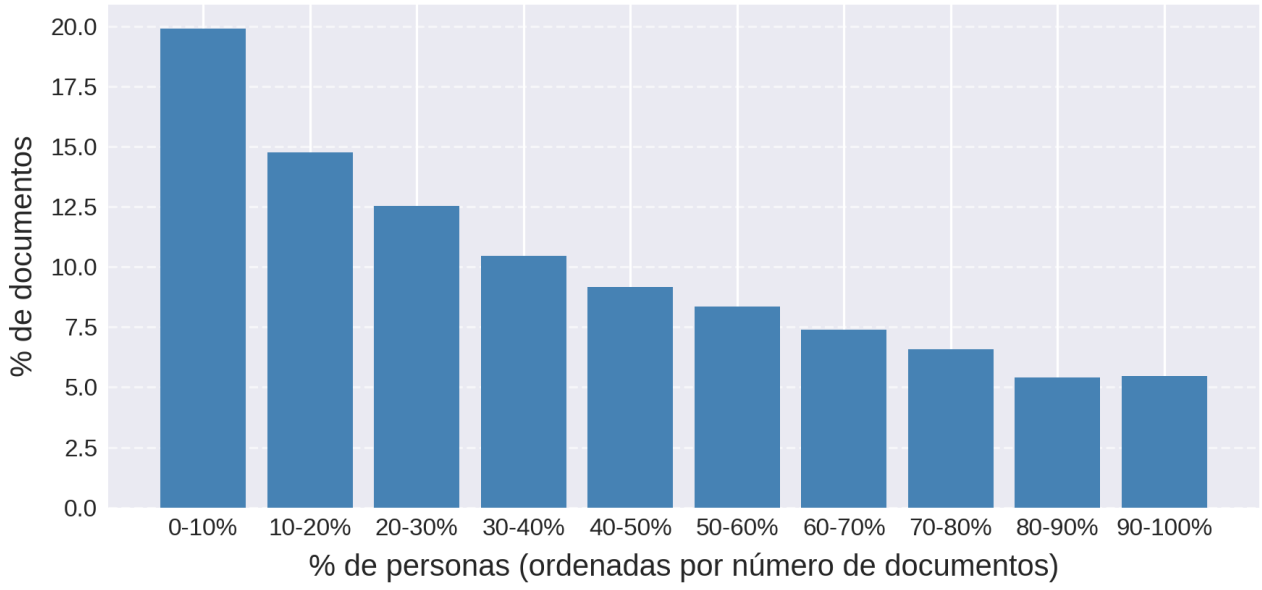

Figura 2.6: Contribución por deciles de personas por número de participaciones

2.4 Documentos de prensa

2.4.1 La base de datos de noticias

Para el desarrollo de la investigación se analizó una base de datos con noticias de medios nacionales en línea acotando la búsqueda al periodo de estudio, esto es entre el 11 de marzo de 2019 al 10 de marzo de 2020. Este conjunto de datos estuvo compuesto por 273.421 registros de noticias de 90 medios en línea (diarios, radios, portales de noticias, portale de gobierno y otros), distribuidos en 4 conglomerados de medios más un grupo de medios independientes. La figura 2.7 muestra un gráfico que da cuenta de los datos analizados por conglomerado de medios de prensa.

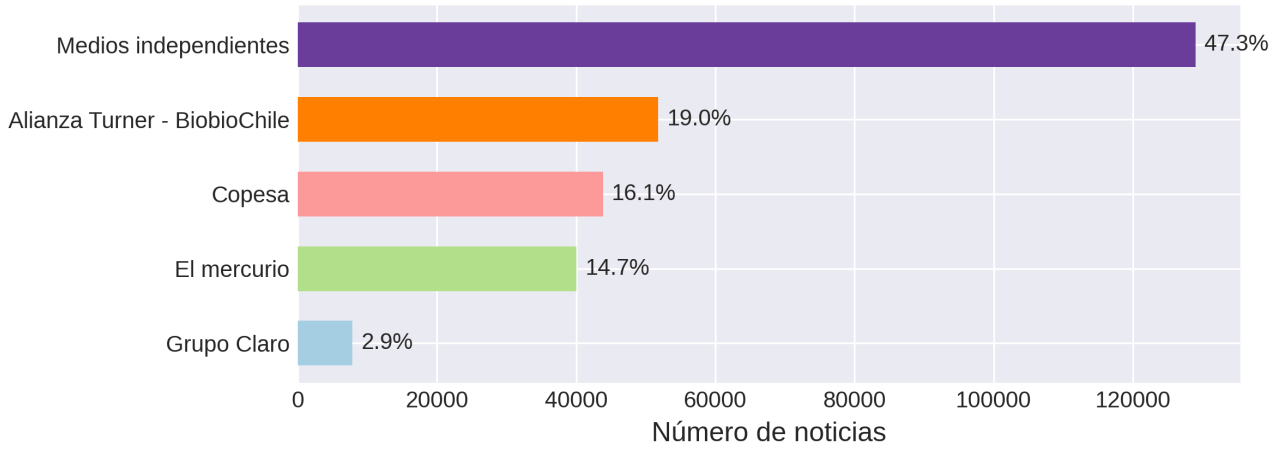

Figura 2.7: Porcentaje de noticias analizadas por conglomerado

A nivel de documentos por persona, en el conjunto de datos existen 20.845 noticias que mencionan a parlamentarios (solo un 7,6% del total de noticias), los que equivale a 41.387 intervenciones en prensa. De esta manera, la tabla 2.4 muestra las estadísticas descriptivas del número de noticias por persona y los graficos de la figura 2.8 muestran la distribución del número de noticias por persona. En este contexto, se observa una dispersión más amplia que en los datos del Congreso Nacional, lo que se verifica también mediante el coeficiente de variación (*0,96*), presentando valores atípicos muy por encima del límite del cuarto cuartil. Esto pone de manifiesto diferencias en la presencia mediática de los parlamentarios en los medios analizados y revela la incidencia de factores políticos y comunicacionales, no evidentes a simple vista, que pueden sesgar el conjunto de datos y, en consecuencia, los criterios que distinguen a unos parlamentarios de otros.

Total noticias	Media	Mínimo	Q1	Mediana	Q3	Máximo	CV
20.485	209,02	1	73,5	152	271,25	1.606	0,96

Tabla 2.4: Estadísticas descriptivas sobre número de noticias por persona

Desde el punto de vista del análisis del texto, los gráficos de la figura 2.9 muestran la distribución de palabras por noticia, como un histograma en escala logarítmica, que muestra los totales de palabras por documento donde se permiten visualizar claramente los valores atípicos.

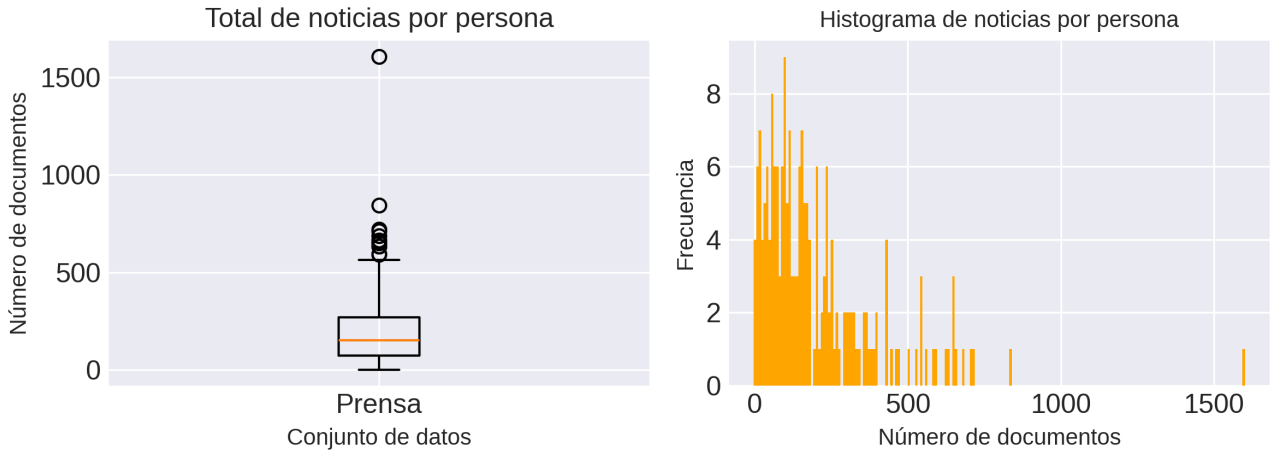

Figura 2.8: Distribución de noticias por persona

Desde la perspectiva de la contribución a los datos por persona, la figura 2.10 muestra el gráfico de deciles de contribución, lo que permite observar cómo se distribuye la cantidad total de noticias asociados a los distintos parlamentarios, agrupados en segmentos de igual tamaño (deciles). Cada barra representa el porcentaje del total de documentos que aporta un 10% específico de los parlamentarios, ordenados desde aquellos con más documentos hasta aquellos con menos. En este contexto, el gráfico muestra que el 10% de parlamentarios que más documentos tiene asociados en este conjunto aporta con más del 30% del total de documentos del conjunto, mientras que el último decil aporta con menos de un 2% de los documentos, lo cual muestra un desequilibrio importante considerando además que este último decil agrupa una cantidad mayor al 10% efectivo de los parlamentarios por la forma en que se calculan los deciles.

2.4.2 Análisis del contenido de la prensa

Por otro lado, habiendo realizado una exploración no eshaustiva, es válido indicar que cuando en una noticia se habla de más de un parlamentario, las temáticas más frecuentes de las noticias corresponden a conflictos o colaboraciones entre los intervinientes, equilibrio de opiniones y exposición por su caracter de rostros de los sectores políticos. Esta percepción es coherente con estudios previos identificados acerca de la cobertura mediática de la prensa en política [Semetko and Valkenburg, 2000, Yildirim et al., 2022, Curry et al., 2024].

También el muestreo realizado y las estadísticas sobre noticias por parlamentario, permiten concluir que algunos parlamentarios concentran significativamente más atención mediática en desmedro de otros, ya sea a raíz de estrategias de posicionamiento [Željko Poljak, 2024] o debido a su rol como *"figuras de poder"* de los sectores políticos a los cuales pertenecen [Aelst et al., 2010], siendo frecuentemente consultados como portavoces oficiales o referentes para ciertos temas específicos tales como políticas sociales, políticas económicas o temas valóricos. Además, es un hecho que algunos parlamentarios han alcanzado notoriedad pública previa como líderes de opinión provenientes de movimientos sociales, tales como la "Bancada estudiantil"[2] o del ámbito municipal, lo que les permite acumular un capital político y comunicacional relevante

[2]Grupo de parlamentarios que emergieron como dirigentes estudiantiles secundarios o universitarios a partir de movilizaciones sociales en chile en la década del 2010.

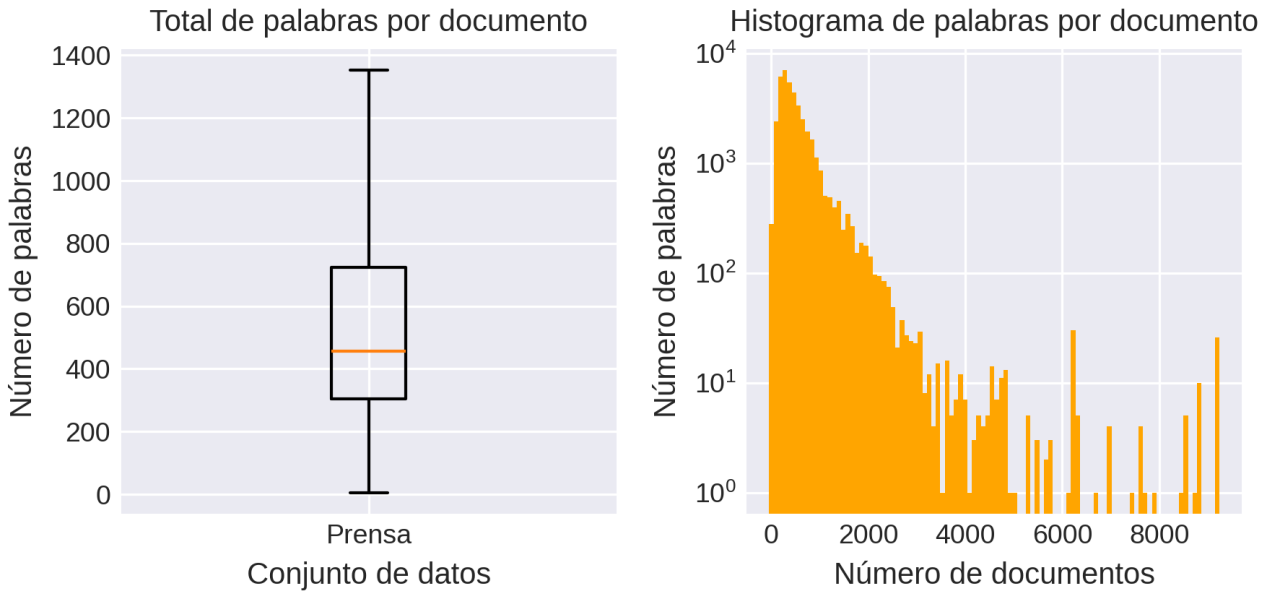

Figura 2.9: Distribución de palabras por noticia

que la prensa tiende naturalmente a destacar. Estos parlamentarios, al contar con trayectorias políticas reconocidas y redes de influencia consolidadas, resultan más atractivos para los medios de comunicación que aquellos menos conocidos, cuya presencia mediática queda restringida a contextos más locales o específicos. Por lo mismo, es frecuente encontrar que parlamentarios con menos notoriedad nacional, pero con fuerte arraigo territorial, reciban escasa cobertura en comparación con figuras más visibles o controvertidas que capturan el interés de audiencias amplias y variadas.

Además de lo anterior, es importante considerar que los medios de comunicación están influenciados por grupos de poder asociados a distintos sectores políticos, lo cual también incide en la notoriedad y tendencia respecto a las figuras públicas mencionadas en las noticias, así como en la perspectiva desde la cual se presentan los hechos. Esta influencia puede determinar la frecuencia, el tratamiento y el tono con que ciertas personalidades políticas aparecen en la prensa, generando sesgos comunicacionales que refuerzan o debilitan la relevancia pública de los parlamentarios.

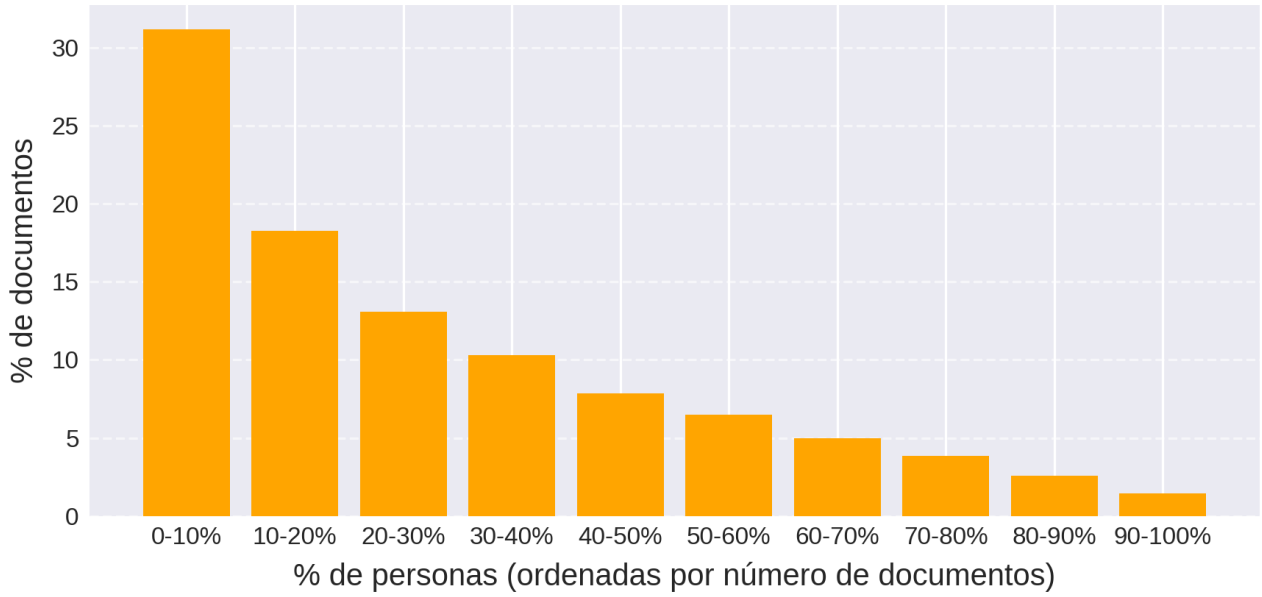

Figura 2.10: Contribución por deciles de personas por número de noticias

2.5 Datos de redes sociales

2.5.1 La base de datos de tweets

El tercer conjunto de datos explorado corresponde a una muestra relevante de documentos recopilados mediante la red social Twitter (actualmente X), durante el mismo periodo en estudio, esto es entre el 11 de marzo de 2019 hasta el 10 de marzo de 2020, mediante el uso de la biblioteca java Twitter4J[3]. Estos documentos, incluyen tweets asociados a las cuentas de los parlamentarios que se tenía constancia de una cuenta oficial, sean tweets emitidos por un parlamentario, retuiteados por ellos (acción de reenviar un mensaje para aumentar su difusión), tweets que el parlamentario haya puesto *me gusta*, o tweets que hayan mencionado al parlamentario mediante su nombre de usuario utilizando @. Para el periodo, se cuenta con un total de 149.472 registros, de los cuales 133.685 pertenecen a personas del periodo.

El total de parlamentarios que no tienen registros en este dataset corresponde a 8 personas, lo que equivale a un 4% de parlamentarios del periodo. La tabla 2.5 muestra las estadísticas descriptivas sobre tweets por persona[4], donde se visualiza el coeficiente de variación de los datos (CV) en un valor de *1.03*. De la misma forma, los gráficos de la figura 2.11 muestran un alto grado de dispersión en el número de tweets por persona, pensando en poder utilizar este conjunto de datos para compararlas.

Dada la limitación técnica que imponía Twitter respecto al número de caracteres máximos que se podían utilizar (en un principio 140 caracteres, luego 280 y actualmente 4000), la distribución del total de palabras de los mensajes asociados a parlamentarios se muestran en los gráficos de la figura 2.12, donde el histograma permite visualizar una distribución bimodal.

Respecto la contribución a los datos por persona, la figura 2.13 muestra el gráfico de deciles de contribución, lo que permite observar cómo se distribuye la cantidad total de tweets asociados a

[3] `https://twitter4j.org/`

[4] Los cálculos de la tabla consideran solo aquellos que tienen cuenta de Twitter a la fecha de consulta

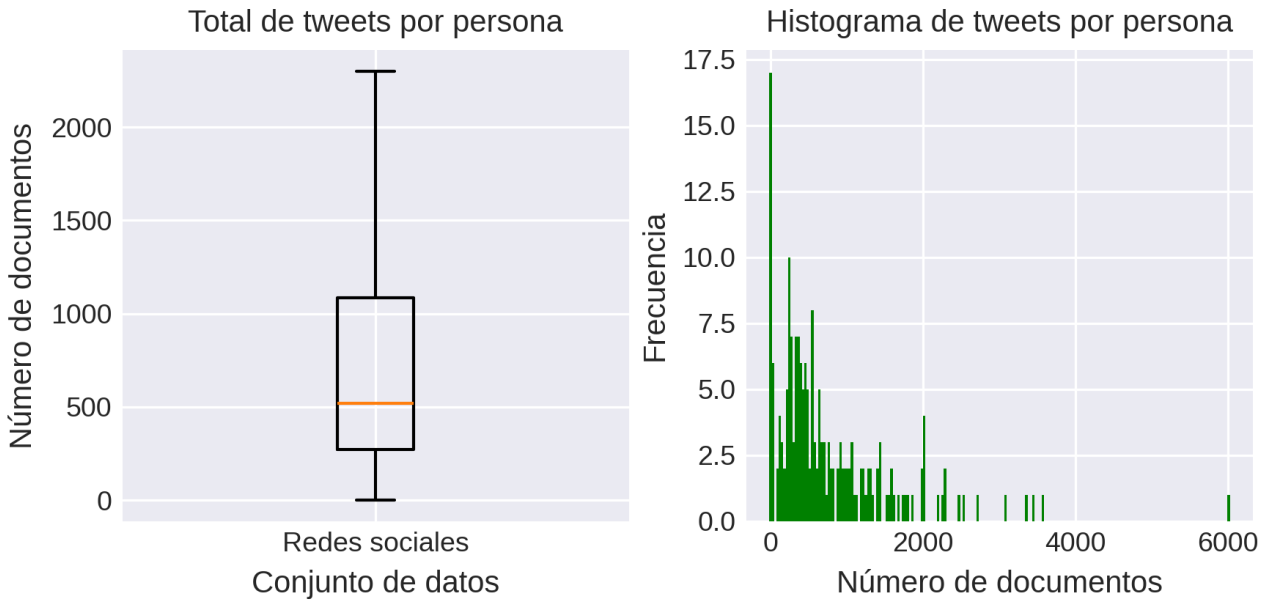

Figura 2.11: Distribución de tweets por persona

Total tweets	Media	Mínimo	Q1	Mediana	Q3	Máximo	CV
133.685	791,17	2	271,25	520	1083,25	6043	1,03

Tabla 2.5: Estadísticas descriptivas sobre número de tweets por persona

los distintos parlamentarios, agrupados en segmentos de igual tamaño (deciles). Como se indica para los casos anteriores, cada barra representa el porcentaje del total de documentos que aporta un 10% específico de los parlamentarios, ordenados desde aquellos con más documentos hasta aquellos con menos. De esta manera, el gráfico muestra que el 10% de parlamentarios que más documentos tiene asociados en este conjunto, al igual que el conjunto de prensa, aporta con más del 30% del total de documentos del conjunto, mientras que el último decil aporta con menos del 1% de los documentos, lo cual también muestra un desequilibrio importante considerando además que este último decil agrupa una cantidad mayor al 10% efectivo de los parlamentarios por la forma en que se calculan los deciles, y que además un 4% de los parlamentarios no tienen una cuenta deTwitter. Un dato relevante es que de los 149.472 documentos identificados del periodo un 31% (37.096) fueron realizados por mujeres, lo cual indica que las mujeres utilizan en mayor proporción las redes sociales que los hombres, ya que solo un 22,7% de los parlamentarios son mujeres.

2.5.2 Análisis del contenido de las interacciones en Twitter

También a partir de una exploración no exhaustiva, se observa que los temas abordados por los parlamentarios en sus mensajes de Twitter se vinculan a distintos tipos de información. Entre ellos destacan contenidos relacionados con la agenda legislativa, interacción con la ciudadanía mediante respuestas a preguntas, hilos explicativos, apoyos a campañas electorales [Santander et al., 2017, Rodríguez et al., 2018], apoyo a iniciativas ciudadanas (uso de hashtags), participación en conversaciones dentro de su propio sector político, debates en tono conflictivo con adversarios de

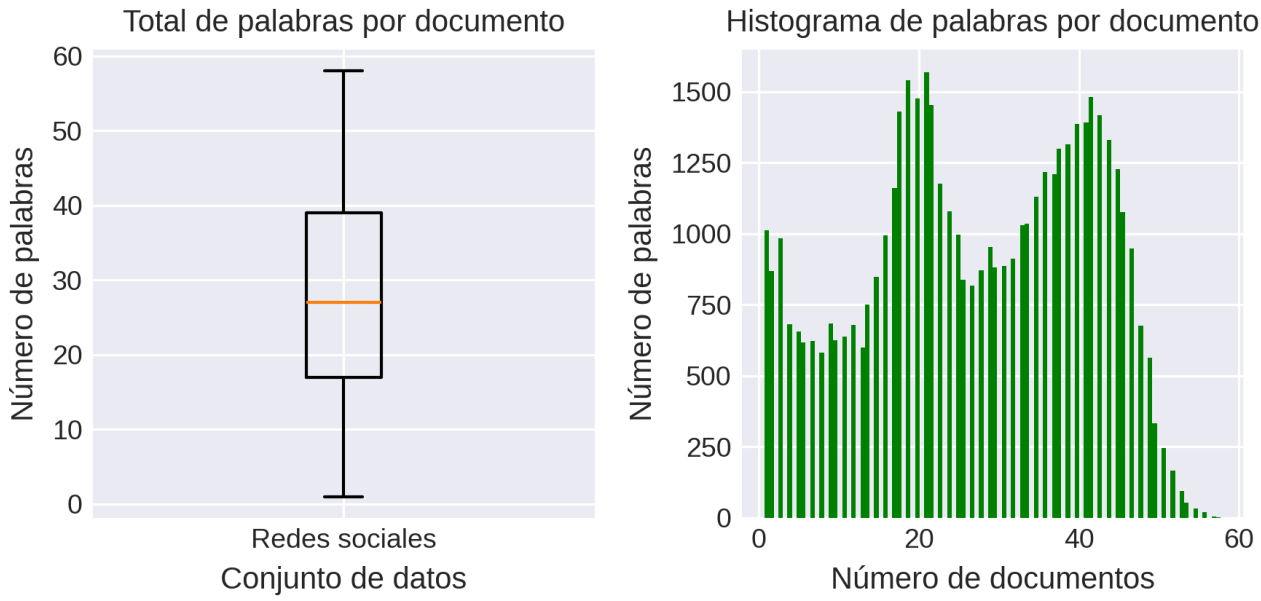

Figura 2.12: Distribución de palabras por tweet

otros sectores y comentarios sobre asuntos de índole personal, entre otros. Muchas de estas ideas, igualmente han sido observadas por estudios recientes donde se analizan las interacciones de los parlamentarios en la red social [Agarwal et al., 2019, Valle et al., 2022]

También aunque en menor medida, dado que en la base de datos se registran menciones a parlamentarios que no han sido realizadas por ellos mismos pero que están marcadas como asociados a ellos, existe el riesgo de interpretar erróneamente la información, atribuyendo categorías temáticas a personas y documentos que no reflejan en la realidad tales formas de pensar. Esta situación dificulta una correcta categorización de los mensajes y complejiza el proceso, ya que requiere incorporar pasos adicionales para discriminar entre el contenido efectivamente generado por los parlamentarios y aquel que no lo es.

En este contexto, la combinación de diversos tipos de contenido y la brevedad de los mensajes podrían introducir ruido en la clasificación temática de los datos para su posterior análisis agregado. Además, dado que la discusión en la red social carece de pautas temáticas, límites o marcos generales, y no existe un esquema jerárquico predefinido para organizar los temas, resulta difícil establecer categorías consistentes. Esta complejidad se ve acentuada por el hecho de que, en numerosos casos, la coyuntura y los temas de alto impacto social capturan la atención de la mayoría de los parlamentarios. Todo ello podría afectar negativamente la capacidad de comparar o agrupar a los parlamentarios conforme a los criterios analíticos propuestos.

Otros temas relevantes asociado a los datos de redes sociales como Twitter y que afectan la pureza de la información que se pretende extraer, es la importante utilización de *bots* en política [Castillo et al., 2019], y el sesgo algorítmico para fines específicos que puede ser introducido por quienes controlan la red social. Ejemplos de ello son algunos trabajos recientes (en progreso) [Graham and Andrejevic, 2024] que investigan eventuales sesgos durante las elecciones de los Estados Unidos en 2024 y antes de las elecciones federales alemanas [Prama et al., 2025], los que concluyen que es posible que exista sesgo de sobrerepresentación de algunos grupos promovido por la red social.

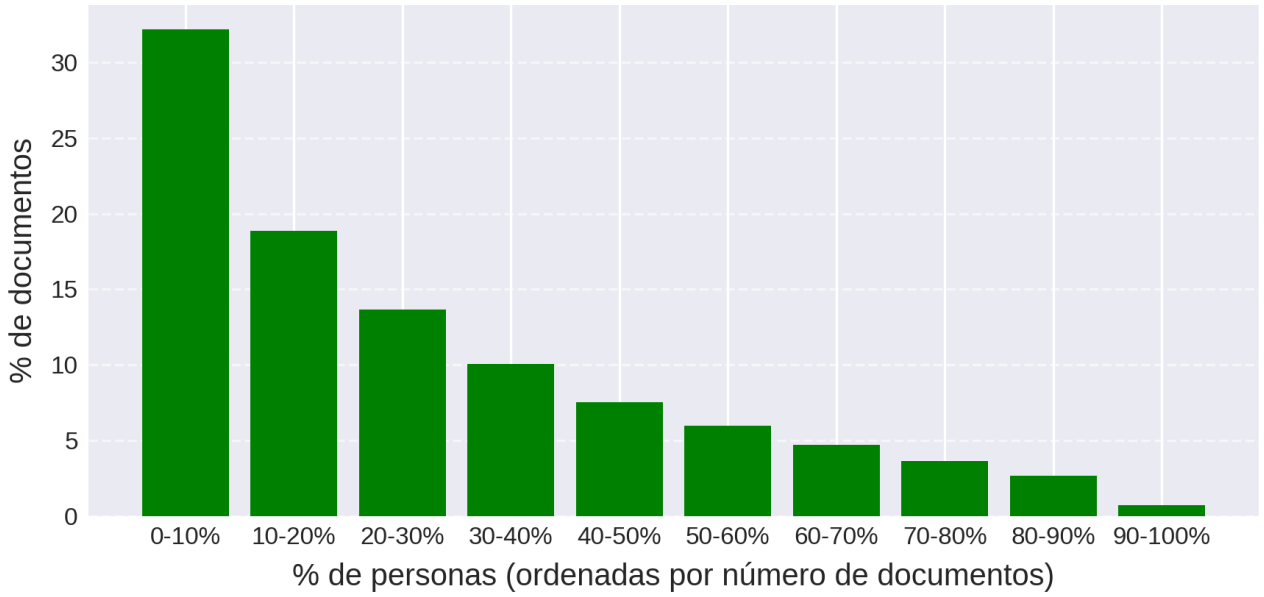

Figura 2.13: Contribución por deciles de personas por número de tweets

2.6 Conclusiones del muestreo preliminar

El presente ejercicio de análisis exploratorio tuvo como objetivo central verificar la idoneidad de los distintos conjuntos de datos disponibles para la elaboración de instrumentos de análisis basados en tecnologías semánticas. Este paso fue crucial para asegurar que los datos seleccionados permitieran responder adecuadamente las preguntas de investigación, validar la hipótesis general y, en definitiva, garantizar el éxito de la investigación.

2.6.1 Equilibrio de los conjuntos de datos

Como primer paso, se analizó la distribución de documentos por parlamentario en cada conjunto de datos, utilizando la distribución por deciles (figura 2.14). A partir de estos datos, se construyó la curva de Lorenz (figura 2.15), que permite visualizar la desigualdad acumulada en comparación con un equilibrio perfecto (la diagonal principal del gráfico).

El análisis mostró que los datos del Congreso Nacional son los más equilibrados, mientras que los conjuntos de Prensa y Redes Sociales presentan una concentración desigual: en ambos casos, el 20% de los parlamentarios (36 personas) acumula aproximadamente la mitad de los documentos, dejando al 80% restante con la otra mitad. Además, debido al método de cálculo de deciles, la composición de los grupos no es exacta, lo que introduce una leve distorsión adicional.

De forma complementaria, se calculó el coeficiente de Gini para cada conjunto, evidenciando que el debate parlamentario es el más equilibrado (Gini = 0,2698), mientras que Prensa y Redes Sociales mostraron valores significativamente más altos (0,4702 y 0,4983, respectivamente). Estos resultados se resumen en la tabla 2.6.

Este desequilibrio podría afectar negativamente los análisis posteriores, ya que tendería a sobrerrepresentar a los parlamentarios con más documentos. Tal sesgo distorsionaría métricas e instrumentos de evaluación, como los gráficos y modelos de análisis. Por el contrario, la mayor homogeneidad del conjunto del Congreso Nacional sugiere una representación más balanceada

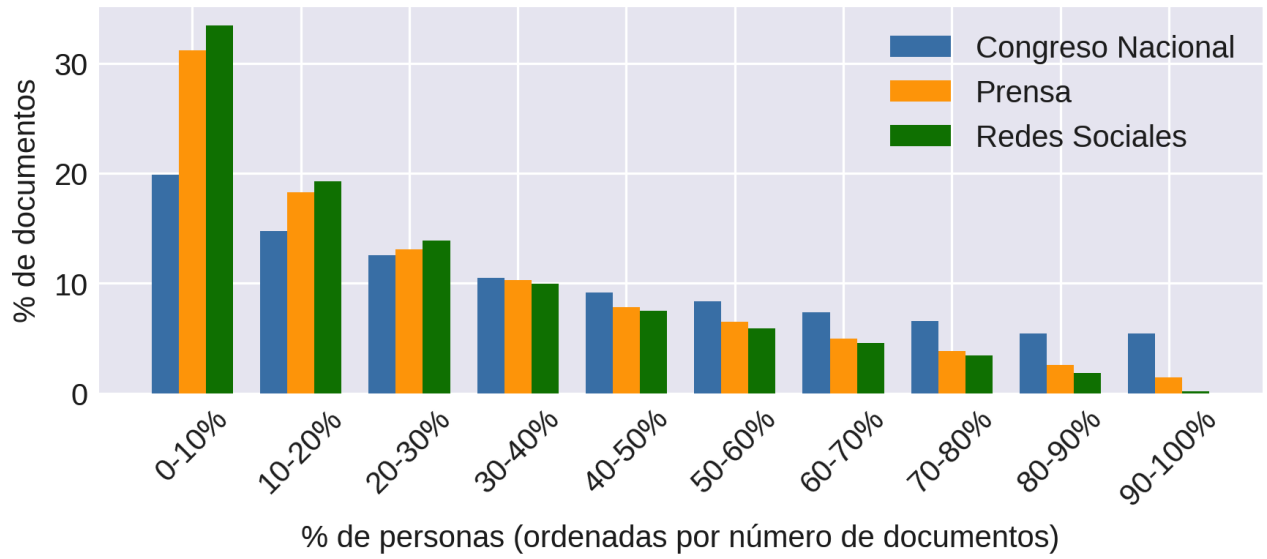

Figura 2.14: Comparación de contribución por deciles de personas por conjuntos de datos

y metodológicamente adecuada.

2.6.2 Análisis del contenido y relevancia temática

Para comparar el contenido temático de los tres conjuntos, se implementó un modelo de tópicos Latent Dirichlet Allocation (LDA) que fue aplicado a todo el conjunto. Las consideraciones de diseño, preprocesamiento y los detalles de la implementación del modelo de tópicos se describen en el anexo I. Una vez analizados los datos, los resultados revelaron que:

- De los tópicos identificados en el óptimo de coherencia, solo 10 de los 25 tópicos correspondían a temas legislativos.
- Los documentos pertenecientes a los conjuntos de Congreso Nacional y Prensa cubrían todos los temas legislativos, mientras que los de Redes Sociales solo abordaban la mitad.
- Los tipos de información asociados a los tópicos variaban significativamente entre los conjuntos, reflejando diferencias en la naturaleza de los datos.

En cuanto al contenido de los datos, se identificaron subtipos de documentos en Prensa y Redes Sociales que aportan ruido o que no se relacionan directamente con el quehacer parlamentario ni con las preguntas de investigación. Esta diversidad no se presenta en los datos del debate parlamentario, que, al ser registros oficiales de comunicación legislativa, reflejan de forma directa y sin intermediaciones la posición de los parlamentarios. Además, es posible establecer que el uso exclusivo del conjunto del Congreso Nacional evita la incorporación de agendas editoriales (en el caso de la prensa) o conversaciones informales (en redes sociales), reduciendo así el riesgo de análisis sobre información descontextualizada o irrelevante. El gráfico de radar en la figura 2.16 muestra la cobertura de los distintos tipos de tópicos asociados a cada conjunto de datos, los cuales fueron claisficados de forma manual durante el análisis.

2.6.3 Permisos de uso y consideraciones legales

Respecto al nivel de permisos de acceso y uso de los datos, el debate parlamentario es de acceso público a través de los portales de datos abiertos de BCN y el Congreso Nacional chileno, a diferencia de los datos de prensa y redes sociales, en los cuales existen restriciones de derechos de autor que impiden el uso y difusión de los datos. También desde el punto de vista del derecho a rectificación, la base de datos de tweets no implementa mecanismos que implementen el derecho a eliminar información incorrecta generada por los parlamentarios, lo cual es una debilidad que puede afectar el análisis.

Característica	**Congreso Nacional**	**Prensa**	**Redes sociales**
Coeficiente de Gini	**0,2698**	0,4702	0,4983
Documentos por persona	**Más equilibrado**	Menos equilibrado	Menos equilibrado
Permisos de uso	**Libre uso**	Propiedad intelectual	Restricciones de uso y derecho al olvido
Nº de temas legislativos en tópicos	**10 de 10**	**10 de 10**	5 de 10
Tipos de información	Debate político-legislativo	Por posicionamiento, coyuntural, patrocinado	Interacción con ciudadanía, debate, personal
Requiere Procesamiento Adicional	**No**	Sí	Sí
Sesgo	**No**	Grupos económicos	Potencial
Largo medio de los textos (palabras)	250	450	30
Utilidad potencial	**Alta**	**Alta**	**Alta**

Tabla 2.6: Tabla resumen de valoraciones en muestreo preliminar de conjuntos de datos

2.6.4 Conclusión general

En conclusión, desde un punto de vista del análisis de datos, el reducir la heterogeneidad de fuentes a una sola fuente oficial facilita el análisis, mitiga la *varianza exógena*[5] reduciendo el número de procesos necesarios para preparación de datos, a la vez que permite atribuir con

[5]Variación que se explica por factores externos al sistema en estudio

mayor confianza los atributos que se desea asignar a los objetos de análisis, mejorando el nivel de fiabilidad de los experimentos y garantizando su coherencia metodológica. Acorde a la tabla 2.6, el conjunto de datos del Congreso Nacional provee un corpus completo, equilibrado y directamente alineado con los objetivos de la tesis, mientras que las otras fuentes, aunque presentan un alto potencial de uso, añaden complejidad y amplian el alcance, incorporando más ruido, sesgos y riesgos que superan sus eventuales beneficios en el marco de este estudio.

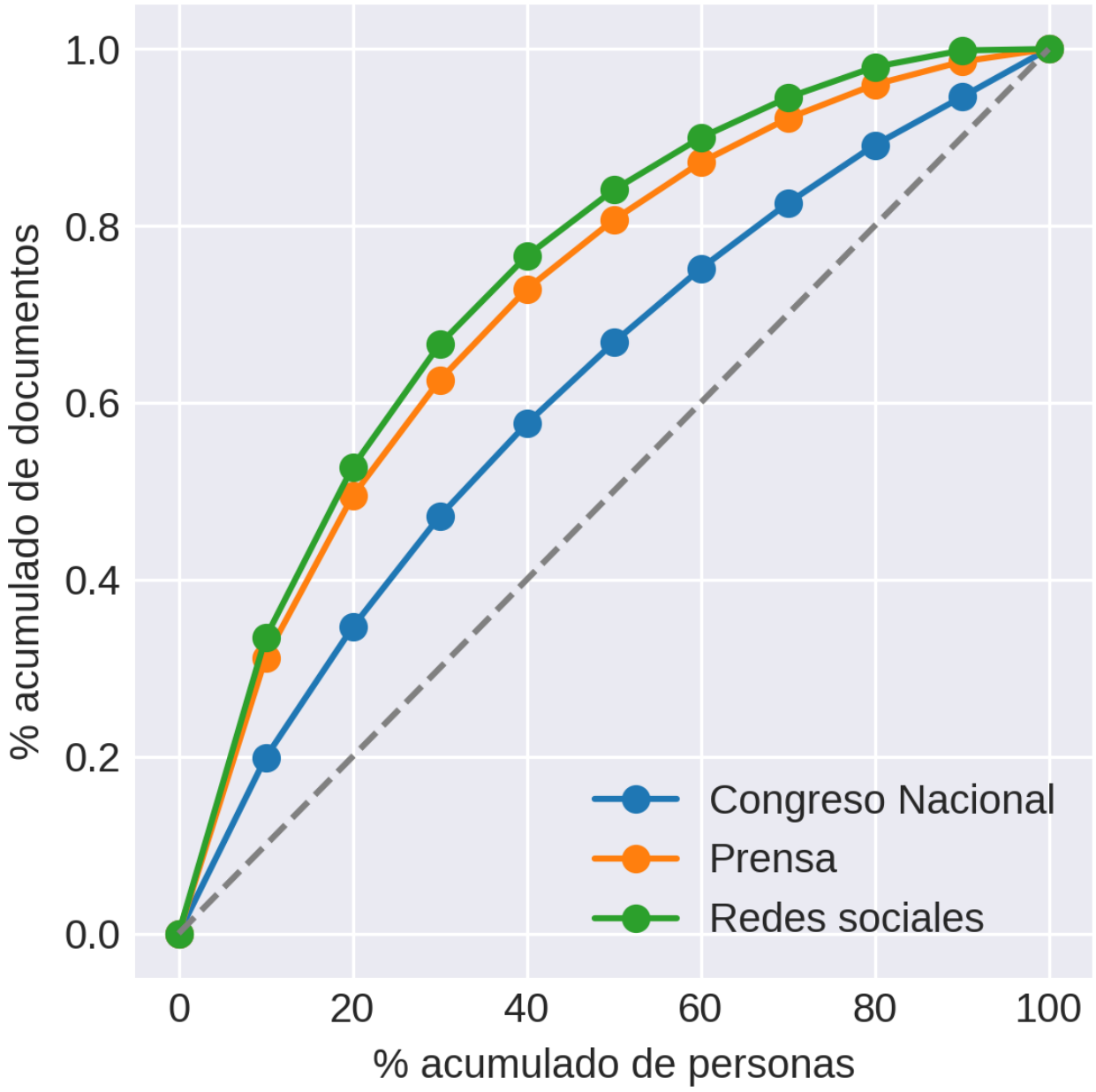

Figura 2.15: Curva de Lorenz de desigualdad de conjuntos de datos

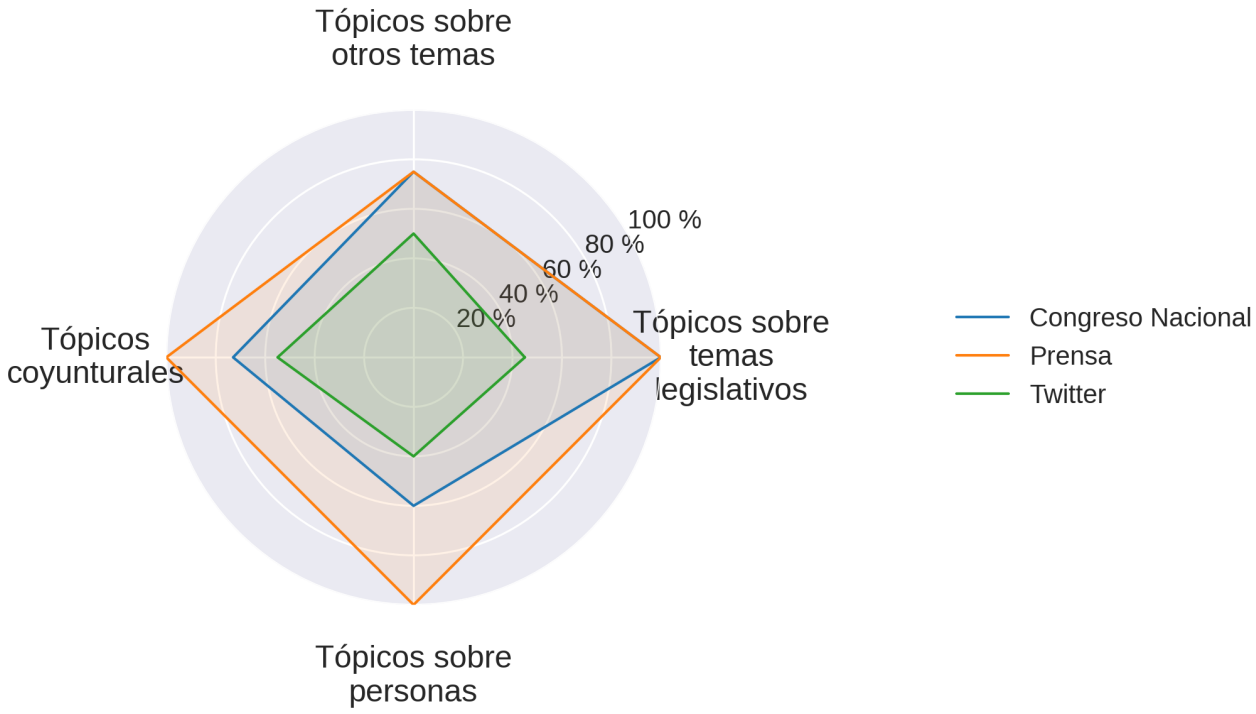

Figura 2.16: Cobertura de tipos de tópicos asociados a los distintos conjuntos de datos

Capítulo 3

Hipótesis y objetivos del trabajo

Como se ha visto hasta el momento, ya se dispone de un conjunto de datos para analizar y de un conjunto de tecnologías que deberán ser puestas a prueba para validar su pertinencia y uso. En consecuencia, a continuación se presenta la hipótesis del trabajo de investigación y su trasfondo investigativo con base en las preguntas de investigación definidas.

3.1 Hipótesis de investigación

La hipótesis del presente trabajo de investigación es la siguiente:

Las Tecnologías Semánticas permiten automatizar el procesamiento de datos no estructurados para generar respuestas a consultas de análisis político-legislativo.

O dicho de forma extensa, *que disponiendo de datos no estructurados provenientes de diversas fuentes, es posible aplicar Tecnologías Semánticas bajo procesos automatizados, para la extracción y anotación de metadatos, como también para la generación de relaciones entre entidades y metadatos, tales que permitan consultar y describir información agregada y desagregada, proveyendo un instrumento de análisis cuantitativo útil para el análisis político-legislativo.*

Al tratarse de una hipótesis abierta, se reconoce desde un inicio que el objeto de estudio no puede reducirse fácilmente a relaciones lineales o causales evidentes. En este contexto, la investigación se plantea desde una lógica inductiva y flexible, privilegiando el análisis y la generación de conclusiones con base en casos de prueba relacionados que permitan establecer puntos de soporte, así como inferir patrones emergentes en el contexto experimental.

En consecuencia, para abordar la hipótesis de manera integral y metodológicamente rigurosa, resulta necesario formular preguntas específicas que guíen la exploración en diferentes dimensiones del fenómeno estudiado, que para nuestro caso en resumen es que *"es posible generar análisis cuantitativo en el ámbito político-legislativo de forma automatizado desde texto plano"*. Estas preguntas cumplen una función clave, ya que permiten operacionalizar la hipótesis general, delimitando áreas concretas de indagación y ofreciendo marcos de referencia que faciliten el proceso analítico destinado a descubrir nuevas comprensiones sobre el tema.

3.2 Preguntas de investigación

Para validar la hipótesis, se establecerán tres preguntas de investigación, de las muchas posibles, que en conjunto abarcan aspectos representativos del análisis político-legislativo, permitiendo ponderar una conclusión general. Las preguntas de investigación son las siguientes:

RQ1: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuáles son los temas de mayor relevancia para un representante?*

RQ2: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuál es el nivel de cohesión política de un grupo frente a un tema particular?*

RQ3: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas quién cumple un rol clave en el contexto de un tema específico?*

Cada una de estas tres preguntas está asociada a distintas dimensiones del análisis político-legislativo que pueden ser respondidas mediante una revisión y análisis manual exhaustivo de documentos relacionados. La idea entonces es validar que es posible hacer el análisis de forma automatizada para generar el mismo efecto, mediante tecnologías semánticas.

3.3 Objetivos

Con base en la hipótesis y a las preguntas de investigación planteadas, a continuación se exponen los objetivos del presente estudio, los cuales se dividen en general y específicos.

3.3.1 Objetivo general

Demostrar la viabilidad y fiabilidad del uso de Tecnologías Semánticas para la construcción de instrumentos que permitan representar y generar analisis político-legislativo de forma automatizada.

3.3.2 Objetivos específicos

Para conducir hacia el logro del objetivo general, se definen los siguientes los objetivos específicos:

1. Identificar y seleccionar fuentes de datos idoneas del ámbito político-legislativo para procesar mediante tecnologías semánticas.

2. Diseñar un marco de trabajo basado en tecnologías semánticas para la extracción automatizada de información desde documentos de texto.

3. Diseñar e implementar instrumentos de análisis político-legislativo utilizando los datos procesados mediante el marco de trabajo y evaluar su nivel de acierto por un Grupo de Expertos (GE).

4. Analizar las evaluaciones del GE, y responder las preguntas de investigación asociadas a cada instrumento.

5. Validar o rechazar la hipótesis de investigación con base en las respuestas obtenidas por las preguntas de investigación.

La consecusión de los objetivos específicos, permitirá responder cada pregunta de investigación para, en conjunto, analizar y responder si la hipótesis planteada es válida o no puede ser aceptada como cierta.

Capítulo 4

Metodología

4.1 Introducción

En este capítulo se describe el enfoque metodológico empleado en la investigación, así como los pasos y procedimientos que se llevaron a cabo durante el diseño experimental, tomando como referencia las pautas definidas por [Hernández et al., 2014]. En primer lugar, se abordará la justificación del enfoque metodológico destacando sus fundamentos teóricos y su relevancia en el estudio. Posteriormente, se explicará el diseño experimental, la selección de participantes y el tipo de muestra, detallando las técnicas de muestreo y los datos recogidos. También se expondrá el entorno donde se desarrolló el experimento, las tecnologías utilizadas y las medidas adoptadas para el control de sesgos y la validación de los datos. Por último, se especificarán las limitaciones identificadas y se reflexionará sobre los aspectos éticos involucrados a lo largo del proceso.

4.2 Enfoque metodológico

Para el desarrollo de la investigación, se ha adoptado un enfoque experimental, en donde a través de la aplicación del marco de trabajo, se procesen datos del contexto político-legislativo para permitir la confección de tres instrumentos de análisis de distintos tipos, los que plantean declaraciones factuales detectadas a partir de los datos, que posteriormente son validadas por un grupo de usuarios expertos mediante preguntas definidas en una escala de Likert.

En detalle, el proceso metodológico desarrollado en la investigación es el siguiente:

1. *Definición del marco de trabajo de las tecnologías semánticas*

   Se realizó un diseño conceptual del marco de trabajo de las tecnologías semánticas (Capítulo 6), el cual está basado en distintos tipos de componentes pertenecientes a los conjuntos de tecnologías seleccionadas, describiendo fuentes de datos pertinentes, y considerando aspectos de explicabilidad algorítmica, flujos de información e interoperabilidad.

2. *Obtención de datos para procesamiento* Es necesario contar con datos que puedan ser recopilados y que permitan ser posteriormente procesados para obtener información de análisis. Este procedimiento se llevó a cabo con tres fuentes de datos, tal como se describe en el Capítulo 2, de las cuales posterior a su exploración inicial, se decidió utilizar solo una de ellas.

3. *Aplicación del marco de trabajo sobre el conjunto de datos* Se aplicó el marco de trabajo definido en el Capítulo 6 sobre el conjunto de datos seleccionado, generando los datos para análisis que permiten el diseño de los instrumentos de evaluación y visualización agregada de la información, dejando la información disponible en una base de datos para consulta.

4. *Diseño y desarrollo de los instrumentos de evaluación*

   Tal como se describe en el Capítulo 7, se desarrollaron tres instrumentos de evaluación donde se utiliza la información procesada por el marco de trabajo, los cuales se construyeron llevando a cabo un proceso estructurado que incluyó las siguientes etapas:

   (a) *Definición de preguntas de investigación*: En la etapa inicial de la investigación, se formularon preguntas de investigación específicas vinculadas a la hipótesis general, diseñadas para que en conjunto permitan evaluar su validez.

   (b) *Identificación de datos requeridos*: Se determinaron los conjuntos de datos necesarios para responder a cada pregunta de investigación.

   (c) *Diseño de instrumentos de evaluación*: Se diseñaron herramientas específicas que permitan responder a las preguntas de investigación, utilizando datos procesados a través del marco de trabajo previamente definido.

   (d) *Desarrollo de la aplicación*: Se diseñó y programó una aplicación para la recolección de datos, el cálculo de métricas relevantes y se realizaron ajustes finales.

5. *Validación con usuarios expertos*

   El proceso de validación con un GE se llevó a cabo considerando los siguientes pasos:

   (a) *Selección de expertos*: Identificación y selección de especialistas del ámbito político-legislativo, incluyendo asesores y funcionarios con experiencia comprobada en el análisis de información y procesos legislativos.

   (b) *Aplicación de instrumentos de evaluación*: Para obtener respuestas estandarizadas, se utilizaron herramientas de medición basadas en una escala de Likert para evaluar la precisión de las predicciones factuales generadas por el sistema.

   (c) *Obtención de datos cuantitativos*: Se recopilaron las puntuaciones proporcionadas por los expertos, reflejando su valoración sobre la calidad y precisión de las predicciones, lo que permite posteriormente realizar un análisis descriptivo objetivo de los resultados.

6. *Análisis de Resultados*

   Se llevó a cabo un procesamiento estadístico de los datos recolectados mediante análisis exploratorio y descriptivo, lo cual se muestra en el Capítulo 8. En primer lugar se realizó un análisis descriptivo utilizando medidas de tendencia central (media, mediana) y dispersión (desviación estándar) para analizar tanto las respuestas como los tiempos de respuesta. Conjuntamente se llevó a cabo el uso de análisis exploratorio mediante gráficos que permita visualizar y caracterizar los fenómenos observados en los datos. Luego, se realizó un análisis de correlación entre distintas variables, así como comparaciones entre los subgrupos de expertos. Este análisis permitió identificar relaciones significativas entre los datos y evaluar la posible existencia de sesgos derivados de las características

de los usuarios en relación con la valoración de los instrumentos. Dado que por cada caso se plantearán dos preguntas mediante el instrumento, se calculó el coeficiente alfa de Cronbach sobre los datos de cada instrumento para medir si las dos preguntas miden un concepto común. Finalmente, los hallazgos fueron interpretados en función de los objetivos definidos, generando conclusiones fundamentadas en la evidencia cuantitativa obtenida.

4.3 Diseño experimental

Para el diseño experimental, se ha considerado que la pregunta fundamental de investigación *¿Es posible aplicar Tecnologías Semánticas mediante procesos automatizados a datos no estructurados para realizar análisis político-legislativo?*, puede ser respondida con base en evaluar los resultados de procesar datos mediante instrumentos que responden preguntas más concretas y específicas, en el fondo, que evalúan algún aspecto específico de interés en el ámbito político-legislativo. De esta manera, cada instrumento planteará un escenario asociado a un aspecto de interés a analizar, mostrando los resultados de datos procesados de forma automatizada mediante elementos pertenecientes al marco de trabajo. Estos resultados se plantearán como declaraciones factuales que serán presentadas a usuarios expertos, quienes validarán el resultado propuesto por la herramienta basado en una escala de conformidad de cinco (5) puntos (escala de Likert [Likert, 1932]), donde las categorías de conformidad con la respuesta varían entre *Totalmente de acuerdo*, *Parcialmente de acuerdo*, *No tengo claro*, *Parcialmente en desacuerdo* y *Totalmente en desacuerdo*. Las escalas de Likert son generalmente consideradas ordinales, lo que implica que sus valores reflejan un orden, pero las distancias entre los puntos no son necesariamente equivalentes. Por ejemplo, la diferencia entre "2" y "3" podría no ser igual a la diferencia entre "4" y "5". No obstante, en la práctica las escalas de cinco o más puntos suelen tratarse como aproximadamente intervalares, particularmente cuando se calculan promedios de respuestas individuales o se analizan datos de grandes muestras [Matas, 2018]. La figura 4.1 muestra la escala de Likert de forma gráfica además de los valores asociados a cada elemento.

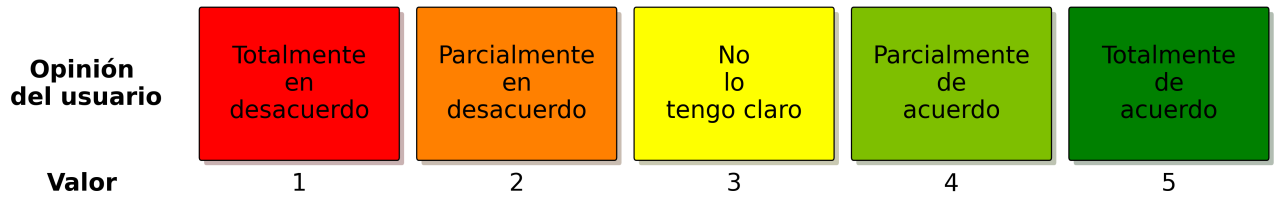

Figura 4.1: Escala de Likert utilizada

Para el caso de estudio, se han definido tres aspectos a evaluar:

1. Identificación de intereses de parlamentarios con base en el texto de transcripción de sus intervenciones en la sala de sesiones del Congreso Nacional.
2. Detección de los niveles de cohesión en los distintos temas y grupos de parlamentarios, identificado en los proyectos de ley en votación.
3. Identificación de parlamentarios con roles clave dentro de una red de parlamentarios interconectados.

De este modo, para cada aspecto a evaluar se han formulado dos preguntas específicas, consideradas como los objetos de estudio y definidas como las variables dependientes. Estas preguntas permitirán validar las declaraciones generadas de forma automática. Por su parte, las variables independientes de cada tipo de pregunta están determinadas por los diferentes instrumentos utilizados, los cuales se describirán en detalle en el Capítulo 7.

4.4 Entorno experimental

4.4.1 Características funcionales

Para el desarrollo del experimento con los usuarios expertos, se ha implementado una aplicación Web que permite el acceso y visualización del banco de preguntas, la responder desde el navegador y una página con información general que contiene la descripción detallada de la información que se recopila, el uso que se le va a dar a las respuestas y el objetivo de la aplicación.

La aplicación, previa autenticación, permite acceder a la lista de preguntas, las cuales se presentan mediante tres cuadrículas navegables de forma de poder visualizar de forma cómoda en una sola pantalla las más de 300 preguntas dispuestas en los 3 instrumentos.

Cada instrumento, presenta una pregunta y un texto introductorio sobre la teoría detrás del aspecto en evaluación, permitiendo responder bajo la mecánica de la escala de Likert, registrando además el tiempo que le toma a cada usuario responder.

Una descripción detallada de las interfaces de usuario se presenta en el Capítulo 7, tanto a nivel de la interfaz de acceso a las preguntas como a nivel de cada instrumento desarrollado .

4.4.2 Tecnologías del entorno

La aplicación se ha construido bajo una pila tecnológica de código abierto, utilizando una base de datos MySQL, código PHP para el desarrollo de programación en el servidor, y para el desarrollo de la interfaz de usuario se ha utilizado el framework de desarrollo Angular 11 en complemento a la biblioteca de gráficos D3.

Si bien para recopilar la evaluación de los usuarios se ha desarrollado una aplicación Web, es importante destacar que la implementación de la aplicación no se ajusta de forma perfecta a dispositivos móviles como teléfonos, por lo cual, su utilización se ha recomendado en computadores personales o tabletas.

4.5 Muestra y datos recogidos

Para el desarrollo del muestreo, se solicitó la colaboración de un grupo de profesionales de las áreas de ciencias sociales, ciencias jurídicas y asesoría parlamentaria. Estos expertos, debido a su labor en asesoría parlamentaria, poseen un profundo conocimiento del contexto político nacional en Chile. Todos ellos desempeñan funciones en la Biblioteca del Congreso Nacional de Chile y constituyen la muestra del experimento. A partir de ahora, se les denominará Grupo de Expertos (GE).

En este contexto, el tipo de muestreo utilizado es no probabilístico, y la técnica específica aplicada corresponde al muestreo por juicio experto.

Para la selección de los expertos, se consideró invitar únicamente aquellos profesionales con al menos cinco años de experiencia en asesoría parlamentaria. Este criterio se basa en el estudio de [Ericsson et al., 1993], donde se establece que una persona alcanza el nivel de experto tras 10.000 horas de práctica deliberada. Dado que una jornada laboral típica comprende 40 horas semanales y un año laboral tiene aproximadamente 50 semanas, cada año de trabajo equivale a 2.000 horas de experiencia. Así, en un período de cinco años se alcanza el umbral de 10.000 horas, consolidando el nivel de expertise mínimo requerido para este estudio.

En este escenario, el GE al que se solicitó su participación estuvo conformado por 17 personas, de las cuales solo 13 efectivamente colaboraron con el desarrollo del estudio.

En cuanto al perfil profesional de los integrantes del GE, se distinguen dos grupos principales. El primero está compuesto por 7 abogados expertos en la elaboración de la Historia de la Ley y en el análisis de la Labor Parlamentaria. Estos profesionales poseen un conocimiento profundo sobre las intervenciones de los parlamentarios, sus perfiles biográficos, sus intereses legislativos, así como sobre el proceso de formación de la ley, incluyendo las etapas iniciales asociadas a los proyectos de ley. El segundo grupo está integrado por 5 sociólogos y 1 periodista. Este último, dado que desempeña funciones equivalentes a las de los sociólogos y solo respondió seis instrumentos del total de casos, ha sido considerado dentro del mismo grupo para efectos analíticos. Todos ellos se desempeñan en labores de asesoría parlamentaria y mantienen un contacto directo con los parlamentarios, tanto en el contexto de las comisiones legislativas como a través del sistema de solicitudes de informes que opera en la ventanilla de asesoría parlamentaria.

4.5.1 Justificación del tipo de muestreo

El enfoque de muestreo por juicio experto es el más apropiado en este caso porque se requiere conocimiento especializado en el ámbito político-legislativo local. Por esta razón, los participantes deben ser expertos y conocer el ámbito político chileno, lo que hace inviable o poco relevante un muestreo aleatorio con usuarios no expertos. También es importante recalcar que, dado que el objetivo es validar criterios cualitativos y/o especializados, se prioriza la calidad y relevancia de las respuestas sobre la generalización estadística a toda la población.

4.5.2 Datos recogidos

Se dejará a disposición de cada experto la aplicación con 322 de casos divididos en tres tipos:

- *Temas de interés parlamentario:* 212 casos
- *Cohesión política:* 70 casos
- *Roles clave:* 40 casos

Cada caso permitirá responder dos preguntas, definidas por cada tipo de caso, las cuales son respondidas mediante escala de Likert definida anteriormente. Dada la cantidad de casos, se solicitó a cada experto responder una cantidad de preguntas acorde a su disponibilidad de tiempo, no estableciendo ningún mínimo, y dejando como máximo el número total de casos. En cada caso, se recopilará el tiempo total que toma cada usuario en evaluar y responder las dos preguntas planteadas. El detalle sobre los datos recopilados y analizados de desarrolla en la sección 8.2 del Capítulo 8.

4.6 Control de sesgos

Durante el desarrollo del experimento se ha precavido la generación de sesgos en las siguientes dimensiones:

- *Sesgos en la generación de casos:* Durante la generación de los distintos tipos de casos asociados a cada instrumento, se ha realizado una verificación aleatoria que ha permitido validar la objetividad de las declarlaciones factuales planteadas.

- *Sesgos en los instrumentos de evaluación:* Se han diseñado los instrumentos y preguntas sin sugerir preferencias en las respuestas y generando una escala de respuestas que permite moverse entre total acuerdo a total desacuerdo.

- *Sesgos en la selección del grupo de expertos:* Se han seleccionado expertos de tres áreas funcionales distintas (Asesoría parlamenaria, Estudios y Servicios Legislativos) que desarrollan labores de asesoría parlamentaria, lo cual asegura heterogeneidad en las perspectivas para evaluar cada caso. También, todos los expertos seleccionados tienen un mínimo de 5 años de experiencia en sus labores, y como servidores del Congreso Nacional ejercen bajo el principio de prescindencia política, lo que implica que no presentan conflictos de interés ni representan sectores políticos que puedan afectar a su evaluación.

- *Sesgos en la administración del experimento:* Se ha ocultado a los expertos la lógica subyacente que explica el cálculo de cada caso en caso de existir, con la idea de que las respuestas eventualmente no se vean influenciadas por los datos numéricos. De la misma manera, no existe un orden predeterminado ni en la disposición de las preguntas ni en la obligación de las respuestas, es decir, cada usuario puede ingresar a responder la pregunta que estime pertinente.

- *Sesgos en el análisis de datos:* Se realizará un análisis de covariables que permita determinar si el tiempo utilizado, el sexo o el área de los distintos expertos influye en la respuesta.

4.7 Limitaciones del estudio

Las limitaciones identificadas en este estudio se describen a continuación:

- A raíz de utilizar muestreo por juicio experto, los resultados podrían no ser completamente generalizables a otros contextos o GE fuera de la Biblioteca del Congreso Nacional de Chile. Esto debido principalmenta a que los resultados pueden estar influenciados por sesgos individuales o limitaciones en la diversidad de perspectivas. Además, aunque la selección de expertos se basa en la regla de las 10.000 horas antes mencionada, el concepto de *"experticia"* en política y asesoría parlamentaria puede ser más complejo y depender de otros factores cualitativos.

- En la misma línea, la implementación del marco de análisis se ha realizado con base en el sistema legislativo Chileno, por lo cual puede requerir ajustes si se pretende aplicar a otros países o sistemas legislativos con estructuras diferentes.

- Debido a que gran parte de los datos procesados se ha realizado de forma automatizada, es posible que del total de casos donde se utiliza clasificación automática, identificación de entidades, extracción de conceptos u otros, existan errores no detectados que puedan influir levemente en las tendencias generales de los casos planteados en los instrumentos de análisis. Esto puede ser en gran parte a causa de ambigüedades del texto en el contexto legislativo y político.
- El estudio se ha realizado utilizando datos de un periodo de tiempo específico, por lo cual no representan la realidad actual, ni reflejan los cambios de opinión ni tendencias que se puedan haber desarrollado a la fecha.
- Si bien en un inicio se consideró la utilización de tres conjuntos de datos: prensa, redes sociales y documentos oficiales del trabajo legislativo, finalmente se optó por analizar solo datos de los documentos oficiales del Congreso, principalmente para evitar errores de interpretación en los datos y simplificar el análisis. En detalle, de esto se da cuenta en la sección de exploración preliminar de datos, Capítulo 2.
- Existe una limitación asociada a la subjetividad en la interpretación de resultados. A pesar del enfoque cuantitativo, el análisis de cohesión política y alineamiento ideológico sigue dependiendo en parte de decisiones metodológicas sobre cómo se definen y miden estos conceptos. Si bien el estudio presenta enfoques definidos con datos y métricas claramente descritas, desde las ciencias sociales pueden existir enfoques que divergen del planteado en este estudio.

4.8 Aspectos éticos

La presente investigación utiliza información pública sobre las sesiones del Congreso Nacional de Chile, así como datos públicos de parlamentarios (biografías, intervenciones, partidos políticos) y proyectos de ley (votaciones), junto a aportes de un GE, cuyas respuestas y datos se manejan de forma agregada y bajo su consentimiento informado. A continuación, se detallan los principales aspectos éticos considerados:

- *Grupo de Expertos (GE)*: Se ha recabado consentimiento informado para usar las respuestas que han proporcionado en los instrumentos de análisis. Dichas respuestas solo se tratarán en conjunto (forma agregada) y no se divulgará información que permita identificar personalmente a ningún miembro del GE.
- *Fuentes de datos*: Toda la información relativa a sesiones del Congreso, intervenciones, votaciones, proyectos de ley, etc., procede de fuentes oficiales y son de acceso público. Se respetan las licencias y términos de uso establecidos por las fuentes oficiales (Senado, Cámara de Diputadas y Diputados, Biblioteca del Congreso Nacional), garantizando un uso legítimo y no manipulado de la información disponible.
- *Anonimización de datos*: Cuando se presenten resultados o conclusiones derivadas del trabajo con el GE, no se publicará información que permita identificar a una persona específica. Los datos se tratarán en conjunto, para salvaguardar la privacidad de cada participante. De la misma forma, al presentar información sobre parlamentarios y sus relaciones, también se aplicará anonimización para evitar exponer información real de los parlamentarios a la comunidad científica.

- *Difusión de la investigación*: Los resultados globales y las contribuciones metodológicas se publicarán sin exponer información sensible de los expertos. Una vez finalizado el estudio, compartirá con los miembros del GE los hallazgos principales, en concordancia con los principios de colaboración y reciprocidad científica. Al mismo tiempo, se presentarán conclusiones basadas en la evidencia, cuidando de no omitir datos relevantes y de evitar sesgos en la interpretación de los hallazgos.

- *Aspectos normativos*: La investigación se adhiere a los lineamientos éticos y normativos de la Universidad de Oviedo, así como a las normas de uso de información pública definidas por la Biblioteca del Congreso Nacional de Chile.

4.9 Cronograma de la investigación

A continuación se describe el cronograma de las actividades realizadas durante la investigación:

Etapa	**Actividades principales**	**Período**
Inicio de la investigación	• Ajustar objetivos y preguntas de investigación y definir alcances e hipótesis • Revisión de la literatura en el contexto de herramientas de análisis político-legislativo utilizando Tecnologías Semánticas. • Implementación de primeras pruebas y experimentación con Tecnologías Semánticas aplicadas a texto no estructurado para la generación de metadatos automáticos.	Primer año y medio de investigación
Definición del marco experimental	• Diseño y documentación del marco de trabajo • Implementación de primeras pruebas, recopilación de datos y análisis sobre conjuntos de datos seleccionados • Procesamiento de datos de prueba mediante técnicas seleccionadas • Implementación de primeras herramientas de análisis como pruebas de concepto • Selección y descarte de pruebas de concepto	Segundo año y medio de investigación
Diseño de instrumentos de evaluación	• Diseño de instrumentos de evaluación con datos procesados • Desarrollar las herramientas experimentales para evaluación • Selección de usuarios expertos • Aplicación de instrumentos de evaluación a usuarios expertos	Tercer y cuarto año de investigación
Análisis de datos y redacción	• Procesar los datos generados con los instrumentos de evaluación mediante herramientas estadísticas • Contrastar resultados con las preguntas de investigación e hipótesis planteada • Desarrollo estado del arte actualizado • Redacción del documento final de tesis, elaborar conclusiones y discusión	Quinto año

Tabla 4.1: Cronograma de la investigación

Capítulo 5

Estado del Arte

Nos paramos sobre hombros de gigantes

5.1 Introducción

A lo largo de la historia, el análisis político-legislativo ha enfrentado diversos desafíos, dando lugar a soluciones parciales desde distintos enfoques, cada uno de los cuales ha abordado aspectos específicos del problema. A continuación, se presentan los principales trabajos y enfoques desarrollados a lo largo del tiempo, desde los primeros estudios que marcaron el inicio del análisis político-legislativo hasta la actualidad, donde las tecnologías de la información desempeñan un papel fundamental en su evolución y aplicación.

Para la elaboración de este estado del arte se establecieron criterios de inclusión basados en la relevancia directa de cada estudio con respecto a las prreguntas de investigación y al contexto específico de este trabajo. Aunque inicialmente se identificaron numerosas publicaciones que aparentaban tener algún grado de relación con el tema, tras una revisión detallada se determinó que muchas de ellas no cumplían con los requisitos mínimos de pertinencia, bien porque su enfoque no se ajustaba al objeto de estudio o porque los resultados que presentan no ofrecen aportes significativos al caso analizado. En consecuencia, se han descartado dichos estudios para garantizar que las referencias incluidas reflejen aportes sustanciales y verdaderamente asociados a la presente investigación.

Sin perjuicio de lo anterior, y a pesar de los criterios de exclusión basados en la relevancia directa al caso de estudio, se ha decidido incorporar un apartado específico sobre usos controvertidos de la Tecnologías de la Información (TI) en el ámbito político. Esta inclusión obedece a la necesidad de exponer desafíos éticos vistos en la práctica, y al mismo tiempo develar la lógica utilizada en su base tecnologica. Desde una perspectiva más general, analizar estos escenarios, aunque en apariencia distantes del caso de estudio principal, permite poner alertas en los vacíos regulatorios, la responsabilidad legal y los riesgos asociados a la adopción de algoritmos aplicados al procesamiento de datos de personas en el contexto político, mostrando con claridad la complejidad de armonizar intereses individuales, el interés público, la protección de derechos fundamentales y la innovación tecnológica.

5.2 Aparición del análisis político-legislativo

Con la expansión en el mundo de la separación de los poderes del Estado y, en consecuencia, la aparición de parlamentos en diversos países, se generaron nuevos procesos para la elaboración de leyes y la estructuración de los cuerpos legislativos. Aunque existen registros de reflexiones y análisis sobre los parlamentos desde la antigüedad, el estudio sistemático de su funcionamiento y dinámica tomó mayor impulso en la era moderna.

Las primeras transcripciones de debates parlamentarios que se conocen se remontan al año 1771 en Londres, donde se comenzaron a elaborar los denominados *Hansard*, una transcripción corregida del debate realizado en el parlamento británico, la cual fue adoptada posteriormente por el resto de países de la *Commonwealth of Nations* - países con lazos históricos a la corona británica. En españa, las actas y diarios de sesiones comenzaron a ser registradas desde el año 1808[1] (Actas de Bayona) y en Chile, se cuenta con diarios de sesiones desde 1810 a la fecha[2].

Con estos insumo, lo que hoy denominamos análisis político-legislativo, entendido como un estudio basado en métodos científicos y sistemáticos, comenzó a consolidarse en el siglo XIX, en paralelo con la institucionalización de la ciencia política como disciplina académica.

De la misma manera, a fines del siglo XIX y principios del XX, se crearon las primeras revistas académicas y obras de referencia sobre ciencia política y derecho constitucional, que incluyeron análisis del proceso legislativo, la representación y el comportamiento de los legisladores. Por mencionar algunas de las más relevantes, se encuentran las revistas académicas sobre ciencia política y derecho *Revue Politique et Parlementaire* (Francia, 1894), *The American Political Science Review* (EE.UU., 1906) , *Zeitschrift für Politik* (Alemania, 1907) o *The Journal of Politics* (EE.UU., 1939), y también obras de referencia como *Lehre vom modernen Staat* [Bluntschli, 1886] o *Congressional Government* [Wilson, 1885].

Dado que en la primera mitad del siglo XX la ciencia política empezó a profesionalizarse, comenzaron a emerger subcampos como la teoría política, la política comparada y, de manera progresiva, los estudios sobre parlamentos. Durante este periodo, por medio de autores como Robert Dahl, Giovanni Sartori o Nelson Polsby, se consolidó el estudio empírico de los congresos y parlamentos, y aparecieron metodologías cuantitativas y cualitativas para su análisis.

5.2.1 Antes de la era de la información

La *Era de la información*[3] es una etapa histórica marcada por el uso intensivo de tecnologías digitales para la creación, procesamiento y distribución del conocimiento. Si bien no existe una delimitación temporal exacta, es posible afirmar que esta Era comienza a consolidarse entre las décadas de 1980 y 1990, con la masificación de los computadores personales y la amplia adopción de Internet.

Antes de la irrupción de los procesadores de texto, y posteriormente de la Web, las tareas de obtención y análisis de información legislativa asociada al Congreso se realizaban de forma completamente manual. Los investigadores debían realizar observación directa (también denominado *soaking and poking*) o acudir a bibliotecas, archivos y hemerotecas para recopilar documentos como diarios de sesiones, informes de comisión y proyectos de ley. Este proceso, intensivo en tiempo y recursos humanos, implicaba una lectura detallada y la transcripción manual de los datos relevantes. Funciones clave como la elaboración de resúmenes, la identificación

[1] `https://app.congreso.es/est_sesiones/`

[2] Diarios de sesiones antiguos en la Biblioteca del Congreso Nacional de Chile `https://bcn.cl/cd2rEN`

[3] `https://es.wikipedia.org/wiki/Era_de_la_informacin`

de materias o la extracción de palabras clave y entidades nombradas (personas, instituciones, fechas) se llevaban a cabo artesanalmente, lo que limitaba considerablemente la capacidad de análisis y el acceso oportuno a la información. Si bien estos métodos manuales sentaron las bases del análisis político-legislativo contemporáneo, carecían de la automatización y estandarización que hoy posibilitan las tecnologías semánticas, el uso de metadatos estructurados y el uso general de herramientas de IA. Dentro de las obras de referencia más relevantes de este periodo es posible encontrar las siguientes:

Congress and the Presidency [Polsby, 1965], quien analiza documentos oficiales tanto del Congreso como del Poder Ejecutivo estadounidense, como también datos históricos sobre comportamiento legislativo, mezclando fuentes cuantitativas y cualitativas, para realizar análisis político-legislativo en aspectos como análisis de liderazgos , colaboraciones y conflictos en aprobación de leyes, dentro de otros;

Congress: The Electoral Connection [Mayhew, 1974], quien utiliza registros del Congreso estadounidense (incluyendo discursos, proyectos de ley, votaciones y actas) para analizar cualitativa y cuantitativamente el comportamiento legislativo, con la idea de fundamentar que los congresistas actúan principalmente para ser reelectos;

Home Style: House Members in Their Districts [Fenno et al., 1978], donde el autor documenta su observación directa acompañando a varios congresistas estadounidenses en sus distritos para buscar patrones en sus comportamientos; y *Handbook of Legislative Research* [Loewenberg et al., 1985] , que ofrece un panorama de los métodos de estudio de las instituciones legislativas, incluidas técnicas de archivo, análisis de actas y métodos comparados anteriores al uso de informática.

5.3 Avances en la Era de la Información

Desde la segunda mitad del siglo XX, los avances en las tecnologías digitales y de la información han permitido que progresivamente se reemplacen muchas de estas tareas humanas asociadas al análisis político-legislativo, por tareas automatizadas o al menos basadas en tecnologías informáticas.

5.3.1 Experiencias en Web Semántica Legislativa

Desde la aparición del campo de la Web Semántica como disciplina tecnológica, las iniciativas relacionadas con el ámbito legislativo han experimentado un crecimiento sostenido, impulsadas tanto por esfuerzos gubernamentales como por contribuciones académicas [Reyes Olmedo, 2017]. Diversos gobiernos y parlamentos - incluido Chile [Cifuentes-Silva et al., 2011] - han implementado marcos y estándares para promover el uso de metadatos semánticos y formatos estructurados, facilitando así la interoperabilidad legislativa, mejorando la transparencia y fomentando la participación ciudadana mediante el uso de datos abiertos. Estas iniciativas enmarcan en una familia de conceptos que es necesario mencionar, dentro de las cuales está el Gobierno Electrónico o *e-Government* (uso de las tecnologías de información por parte del Estado para mejorar la eficiencia administrativa, facilitar el acceso a los servicios públicos y promover la transparencia institucional), Parlamento Abierto u *Open-Parliament* (una forma de interacción entre el poder legislativo y la ciudadanía que incorpora énfasis en los procesos de transparencia, participación ciudadana, rendición de cuentas y colaboracion), Parlamento Electrónico o *e-Parliament* (asociado a transformar digitalmente el quehacer legislativo mediante la adopción de tecnologías semánticas, estándares abiertos y plataformas interoperables) y también Partic-

ipación Electrónica o *e-Participation* (que se refiere al conjunto de mecanismos digitales que permiten a la ciudadanía involucrarse de manera activa en los procesos de formulación, discusión y evaluación de políticas públicas y proyectos legislativos).

Estas iniciativas reflejan la importancia creciente de adoptar tecnologías como Linked Open Data (LOD), ontologías y esquemas de documentos compartidos para hacer interoperables contenidos legislativos, fortaleciendo la democracia y al mismo tiempo acercar a la ciudadanía con los diversos actores legislativos.

En este contexto, a continuación se describen las iniciativas más relevantes en el mundo sobre Web Semántica en el ámbito legislativo.

Identificador Europeo de Legislación

El *Identificador Europeo de Legislación*[4] (en inglés European Legislation Identifier - ELI) [Francart et al., 2019] es un estándar adoptado voluntariamente por los países e instituciones de la Unión Europea, basado en plantillas de URIs que permiten identificar de forma unívoca y accesible en línea la legislación de la UE y de sus Estados miembros. Además, proporciona un marco para describir recursos legislativos mediante metadatos semánticos, lo que facilita la recuperación, comparación y vinculación de documentos legales en distintos contextos transnacionales, al tiempo que se integra con iniciativas de datos abiertos y fomenta la interoperabilidad. En la práctica, el ELI se concreta en tres aspectos clave:

- Identificadores web (URI) para la información jurídica.
- Metadatos que describen el contenido y la naturaleza de la legislación.
- Un lenguaje específico que posibilita el intercambio de datos legislativos en formatos procesables automáticamente.

Akoma-Ntoso

Akoma-Ntoso (acrónimo de *Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies*) [Palmirani and Vitali, 2011] es un estándar OASIS[5] LegalDocML, diseñado para la creación y el intercambio de documentos parlamentarios y jurídicos, cuyo objetivo principal es mejorar la accesibilidad, interoperabilidad y reutilización de documentos normativos mediante un modelo formal y estandarizado que respeta las prácticas jurídicas locales y permite su integración con tecnologías de la Web Semántica. Su enfoque permite describir con precisión la estructura lógica y semántica de los textos legales, e incorpora ejemplos de ontologías y taxonomías que enriquecen el contenido y permiten marcaje de metadatos y análisis semánticos avanzados.

La adopción de Akoma-Ntoso se ha materializado de diversas formas, tanto a nivel de estándar de interoperabilidad para compartir información, como también dentro del proceso de producción de la norma y documentación legislativa (como debates del Congreso). En este contexto, destacan sus aplicaciones prácticas en congresos y parlamentos en América (Brasil, Uruguay, Argentina, Chile, Nicaragua, Estados Unidos), Europa (Italia, Reino Unido, Alemania, Francia, Luxemburgo) y África (Sudáfrica, Kenia), donde ha contribuido a modernizar la gestión documental y promover la transparencia legislativa.

[4] `https://eur-lex.europa.eu/eli-register/about.html?locale=es`

[5] OASIS, acrónimo de Organization for the Advancement of Structured Information Standards `https://www.oasis-open.org`

LEOS Project

El *Proyecto de Software Libre para Edición de Legislación* (en inglés Legislation Editing Open Software Project - LEOS Project) es una iniciativa de la Comisión Europea que ofrece una plataforma abierta para la redacción de textos legislativos y sus anexos, incorporando estructuras semánticas y promoviendo el trabajo colaborativo en línea. Integra funcionalidades como comentarios, sugerencias, control de versiones y coedición, todo en un entorno unificado que busca facilitar la elaboración normativa. Su enfoque estructural es deliberadamente restrictivo, no por rigidez sino para orientar a los redactores en el cumplimiento de normas formales, reducir errores y asegurar la coherencia documental.

LEOS se basa en el formato XML de Akoma-Ntoso, e implementa específicamente el subesquema AKN4EU[6], que garantiza la interoperabilidad entre las instituciones de la Unión Europea y los Estados miembros. Además, promueve el uso de metadatos comunes y estándares como RDF y XML para facilitar la trazabilidad, el versionado y la reutilización de contenido. El proyecto incluye documentación técnica, guías de desarrollo y acceso abierto a su código fuente, fomentando su adopción y evolución colaborativa.

Marcaje de documentos legislativos en otros países

Si bien Akoma-Ntoso, junto con su ecosistema de componentes, iniciativas y plataformas asociadas, puede considerarse el estándar de facto para el marcaje de documentos legislativos en Europa, América Latina y África, la situación varía significativamente en otras regiones del mundo. A continuación, la tabla 5.1 presenta una muestra representativa de países a nivel global que incluye potencias internacionales, economías regionales clave, naciones con fuerte influencia cultural y países en procesos de transición institucional o en vías de desarrollo, donde se describe a nivel país (no divisiones territoriales de estos) el uso de distintos enfoques para marcaje de documentos legislativos.

[6] `https://op.europa.eu/en/web/eu-vocabularies/akn4eu`

País	Estándares	Plataformas donde se utiliza
Potencias internacionales		
China	PDF/HTML oficiales sin esquema XML conocido	Base de datos Legal China `https://flk.npc.gov.cn`
Corea del Sur	RDF LOD, Esquema XML propio, sin Akoma-Ntoso	Centro Nacional de Información Jurídica de Corea `https://www.law.go.kr/`, Portal de Linked Open Data, `https://lod.law.go.kr/` Portal de datos abiertos legislativos `https://open.law.go.kr/`
Estados Unidos	XML (United States Legislative Markup - USLM) compatible con Akoma-Ntoso	Government Publishing Office `https://www.govinfo.gov`
Japón	XML propio e-LAWS, exploración con Akoma-Ntoso	Portal e-Gov `https://laws.e-gov.go.jp`
Rusia	PDF/Texto, no fue posible verificar más	Portal Oficial de Información Legal `https://www.pravo.gov.ru`
Economías regionales clave		
Arabia Saudita	PDF/HTML	Utilizan la Ley Islámica, sin un único portal que concentre toda la información legislativa
Emiratos árabes	PDF/HTML	Cuentan con un diario oficial (UAE Official Gazette) `https://dlp.dubai.gov.ae/en/Pages/OfficialGazette.aspx`
India	PDF/HTML	Debate legislativo cámara baja (Lok Sabha) `https://eparlib.nic.in`, Debate legislativo cámara alta (Rajya Sabha) `https://rsdebate.nic.in`, Legislación sistema India Code `https://www.indiacode.nic.in`
México	PDF/HTML	Senado de México `https://www.senado.gob.mx/66/diario_de_los_debates` Cámara de Diputados`https://cronica.diputados.gob.mx`
Países con fuerte influencia cultural		
España	XML, RDF, ELI	Portal de datos abiertos de BOE `https://www.boe.es/datosabiertos/`, Debate parlamentario `https://www.congreso.es/es/datos-abiertos`

5.4 Analítica en el ámbito político-legislativo

La analítica en el ámbito político-legislativo se posiciona actualmente como uno de los tópicos más atractivos en el estado del arte, debido principalmente a las enormes posibilidades que ofrece para descubrir patrones, tendencias y hallazgos relevantes mediante el análisis sistemático de grandes volúmenes de datos. Su atractivo radica no solamente en la riqueza del tipo de información analizada, sino especialmente en la capacidad creativa del ser humano para analizar, interpretar y transformarla en conocimiento significativo, lo que permite comprender con mayor profundidad los procesos político-legislativos y revelar fenómenos que hasta ahora permanecían subyacentes.

Es de gran relevancia mencionar que no se encontraron trabajos que combinen las tecnologías semánticas, definidas bajo el concepto de esta tesis, aplicadas al ámbito político-legislativo ni para Chile, ni en general. Las propuestas más cercanas a la desarrollada en esta tesis, pero aplicadas solo a un contexto general y sin indagar en el ámbito político-legislativo se asocian al campo del *Social Semantic Web Mining*[Prabhu et al., 2019], donde se utilizan tecnologías similares pero aplicadas a medios sociales (Social Media), orientado a mejorar el análisis en el ámbito de los medios sociales o también llamadas redes sociales.

Dicho lo anterior, a continuación se realiza una revisión al estado del arte en el contexto de soluciones a las preguntas de investigación definidas en la sección 3.2.

5.4.1 Detección de Posturas ideológicas

La detección de posturas ideológicas en el ámbito político-legislativo corresponde a un conjunto de técnicas analíticas orientadas a identificar y caracterizar las posiciones que adoptan los parlamentarios frente a diversas propuestas, proyectos o ideas discutidas en un contexto legislativo. Este tipo de análisis permite identificar posiciones políticas, inferir tendencias ideológicas y evaluar niveles de cohesión dentro de agrupaciones parlamentarias, a partir de distintos enfoques metodológicos. A continuación, se describen los principales métodos presentes en el estado del arte:

- Un primer estudio presenta el modelo de Wordscores[7], una técnica estadística supervisada que permite asignar un puntaje (*score*) para estimar la ubicación ideológica de documentos de texto a partir del patrón de uso de palabras, sin requerir etiquetado manual ni uso de métodos convencionales de análisis cualitativo como *codificación* (normalmente usado en ciencias sociales). Este enfoque resulta altamente replicable, eficiente y aplicable a múltiples casos de uso en política, facilitando el análisis sistemático de grandes volúmenes de texto. Mediante el uso de técnicas clásicas de procesamiento de lenguaje natural, tales como la eliminación de palabras vacías (*stopwords*), análisis de frecuencias de palabras y uso de bolsas de palabras (*bag-of-words*), integradas al modelo Wordscores, se procesaron manifiestos de partidos políticos del Reino Unido e Irlanda, logrando replicar estimaciones previas de posicionamiento ideológico etiquetados manualmente. Posteriormente, el modelo fue utilizado exitosamente en contextos lingüísticos distintos, como en textos políticos alemanes o discursos legislativos, evidenciando su versatilidad. Una de las principales fortalezas del método es su capacidad para generar medidas de incertidumbre asociadas a las estimaciones, lo que permite evaluar rigurosamente la significancia de las diferencias entre posturas políticas inferidas [Laver et al., 2003].

[7] `https://www.tcd.ie/Political_Science/wordscores`

- Un segundo estudio muestra la generación de una red estructurada de opiniones políticas, *OpinioNetIt* [Awadallah et al., 2012], a través de tripletas RDF del tipo actor, postura, tópico, extraídas automáticamente desde fragmentos y citas de noticias en línea sobre temas controvertidos (por ejemplo, "el conflicto en Siria"). El sistema utiliza el *dependency parsing* [Chen and Manning, 2014] de Stanford NLP, desambiguación de entidades con YAGO[8], y expansión semántica con WordNet[9] para enriquecer citas con sinónimos y antónimos. Las opiniones se modelan en Resource Description Framework (RDF), y los tópicos se organizan jerárquicamente mediante recursos como Debatepedia y Wikipedia. Para asignar citas y facetas a tópicos finos, el sistema emplea un clasificador kNN basado en similitud estadística entre textos, utilizando vectores de unigramas y bigramas y una función de *query likelihood*, lo cual permite realizar análisis avanzados sobre la dinámica de opiniones políticas, como la detección de cambios de opinión política, disidentes ideológicos y sesgos mediáticos, a través de consultas SPARQL y visualizaciones de datos.

- Otro estudio relevante presenta el uso de Redes Neuronales Recursivas (RNN) para la detección de ideología política en textos. En él se parte de la premisa de que, debido a la naturaleza jerárquica del lenguaje, las RNN son capaces de modelar la composición semántica del texto, aprovechando la estructura gramatical para identificar con mayor precisión las sutilezas sintácticas y semánticas que reflejan la posición ideológica en cada oración. Para fortalecer la capacidad del modelo, se emplean transcripciones de debates parlamentarios anotadas manualmente a nivel de frases y oraciones, lo que permite un entrenamiento más detallado y favorece la diferenciación fina de matices ideológicos y polaridades discursivas. Los resultados experimentales demuestran que la incorporación de esta información composicional mejora significativamente el rendimiento en la clasificación de orientaciones políticas (por ejemplo, liberal o conservadora), superando a los enfoques tradicionales tanto en conjuntos de datos previamente establecidos como en corpus específicamente construidos para este estudio, lo que evidencia la versatilidad y efectividad del modelo [Iyyer et al., 2014] .

- Otro caso asociado al uso de redes neuronales, es mediante la combinación del modelo Bidirectional Encoder Representations from Transformers (BERT) con Redes Neuronales de Atención sobre Grafos (Graph Attention Networks (GAT)). Por una parte, BERT se utiliza para codificar y generar representaciones semánticas enriquecidas de los textos parlamentarios, aprovechando su capacidad para comprender el contexto lingüístico y las relaciones semánticas profundas entre palabras y frases. Por otra, las redes neuronales basadas en grafos con atención permiten integrar información contextual, propagando de manera explícita las relaciones existentes en el texto entre parlamentarios. La idea tras esta combinación es facilitar la capturar de forma más precisa de estructuras complejas e interacciones semánticas presentes en los discursos políticos. Para evaluar el desempeño de este modelo, se emplearon textos extraídos del conjunto de datos ParlVote, que contiene transcripciones de debates parlamentarios etiquetadas según sentimiento positivo o negativo [Abercrombie and Batista-Navarro, 2020]. Los resultados experimentales confirman que esta propuesta supera significativamente a métodos tradicionales como Support Vector Machine (SVM), Multi Layer Perceptron (MLP), BERT-MLP o Deepwalk [Perozzi et al., 2014], posibilitando un análisis detallado y agregado de la cohesión

[8]Una base de conocimiento general abierta `https://yago-knowledge.org`
[9]Base de datos léxica abierta para inglés `https://wordnet.princeton.edu`

ideológica en diferentes grupos parlamentarios [Glavaš et al., 2017].

- Un enfoque estadístico alternativo para inferir posiciones ideológicas propuesto por [GovTrack.us, 2013] se basa en el análisis de patrones de coautoría entre parlamentarios estadounidenses, utilizando técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA). Este método construye una matriz donde cada celda representa la cantidad de veces que un parlamentario fue coautor o copatrocinó proyectos de otro, calculando luego la descomposición en valores singulares de dicha matrizs, la cual permite identificar patrones ocultos en esa información y organizar a los parlamentarios según sus similitudes en el comportamiento de coautoría. A partir de este análisis, se obtiene una puntuación en un espacio latente que refleja diferencias de comportamiento ideológico, sin considerar explícitamente el contenido de los proyectos ni la militancia partidaria. Si bien, acorde a los autores no se interpreta directamente como una escala izquierda-derecha, se observan con alta precisión bloques ideológicos y niveles de moderación entre parlamentarios.

- De la mano de los mismos autores anteriores [GovTrack.us, 2013] se definen otros mecanismos para identificar liderazgo parlamentario. Este enfoque utiliza el algoritmo *PageRank*, originalmente diseñado por Google para establecer un ranking de páginas Web en buscadores, para calcular una puntuación de liderazgo a partir de las relaciones de coautoría de proyectos de ley. La idea central es que un parlamentario será considerado líder si muchos otros copatrocinan sus proyectos, especialmente si esos copatrocinadores también son influyentes. Los resultados permiten distinguir parlamentarios que tienden a liderar (aquellos cuyos proyectos son ampliamente apoyados) de aquellos que siguen (quienes apoyan más de lo que reciben apoyo). Gráficamente, estas puntuaciones generan distribuciones donde los líderes suelen ubicarse en posiciones extremas del espectro ideológico, lo que sugiere una posible asociación entre liderazgo y posturas ideológicas marcadas.

- Otro enfoque basado en una técnica denominada regularización bayesiana (*bayesian shrinkage*), se basa en un procedimiento de selección de palabras discriminantes para comparar discursos de grupos políticos [Monroe et al., 2017]. Para ello estima, para cada palabra, cuán distintiva es entre dos grupos (p. ej. partidos) ajustando por varianza y rareza. Posteriormente, la regularización bayesiana reduce la influencia de palabras extremadamente frecuentes o muy raras, minimizando sesgos del clásico tf-idf, lo cual permite una visualización de palabras clave partidistas, lo que ayuda a comprender mejor las diferencias políticas de fondo. Como caso de uso se presenta un experimento aplicado a debates del Senado de EE. UU., donde el método identifica términos que capturan diferencias de género, partido y orientación distributiva con mayor estabilidad que enfoques basados solo en frecuencias.

- Un último caso es el denominado escalamiento dentro de debates (*within-debate scaling*) [Lauderdale and Herzog, 2016], que estima la posición política de cada legislador a partir del texto de sus intervenciones en sala de sesión. El método modela, para cada palabra, dos componentes simultáneos: a) la frecuencia propia en el tema tratado en ese debate y b) la inclinación ideológica del orador. Al separar tema y postura, el modelo compensa las diferencias léxicas que surgen porque cada cuestión (economía, salud, defensa, etc.) tiene su vocabulario característico. De este modo se evita que la estimación de la posición política se distorsione por el simple cambio de asunto. Como aplicación empírica, los autores analizan todos los discursos a esa fecha del *Dáil Éireann* (Parlamento irlandés).

El procedimiento revela con claridad la polarización entre gobierno y oposición, distingue matices dentro de un mismo partido y muestra cómo las brechas ideológicas se amplían o reducen ante crisis económicas y ciclos electorales.

5.4.2 Análisis de redes sociales parlamentarias

El análisis de redes sociales parlamentarias corresponde a un conjunto de técnicas metodológicas destinadas a estudiar las relaciones e interacciones entre parlamentarios dentro de un contexto legislativo, identificando patrones de colaboración, oposición y alineamiento político en torno a proyectos de ley, votaciones o debates específicos. Este tipo de análisis permite revelar la estructura subyacente de las interacciones parlamentarias, identificar actores clave, evaluar la cohesión o fragmentación interna en grupos políticos, e inferir dinámicas relacionales que influyen en la toma de decisiones legislativas. A continuación, se presentan las experiencias y métodos más relevantes que se encuentran en el estado del arte:

- Mediante el uso de Social Network Analysis (SNA) aplicado a datos provenientes de Twitter, un estudio comparativo sobre 11 países europeos más Estados Unidos, propone caracterizar la estructura y los patrones de interacción digital entre parlamentarios, explorando cómo estos reflejan o difieren de las formas institucionales de la democracia. A través de la construcción de grafos dirigidos basados en relaciones de *following* (seguir el perfil) entre parlamentarios, se examinan métricas estructurales como densidad de los grafos, modularidad, centralización y homofilia partidaria, lo que permite inferir niveles de polarización, cohesión intra e interpartidaria y centralidad de actores en el medio digital. El enfoque permite observar si los patrones de comunicación en redes sociales replican las divisiones ideológicas formales del parlamento o si surgen nuevas configuraciones distintas a las estructuras tradicionales. Los resultados muestran diferencias significativas entre países, sugiriendo que el uso parlamentario de Twitter está condicionado tanto por factores institucionales como por dinámicas culturales y tecnológicas, y que el análisis de estas redes ofrece un complemento empírico relevante para entender las manifestaciones contemporáneas de la representación política en el entorno digital [Praet et al., 2021].

- Mediante el uso de SNA aplicadas a la autoría y patrocinio[10] de proyectos de ley en la Cámara de Representantes de los Estados Unidos, se propone un modelo cuantitativo para medir la influencia de los legisladores en el éxito legislativo a partir de su posición estructural en la red de coautorías. A través de la construcción de grafos donde los nodos representan congresistas y las aristas indican vínculos de copatrocinio, se calculan métricas como centralidad, intermediación y cercanía para estimar el impacto individual en la probabilidad de que una propuesta se convierta en ley. Este enfoque permite superar las limitaciones de los análisis centrados exclusivamente en el contenido de las mociones, integrando dimensiones relacionales que revelan dinámicas de colaboración estratégica dentro del proceso legislativo. Los resultados evidencian que ciertas posiciones en la red, más allá de la afiliación partidaria o la antigüedad, tienen una correlación significativa con la aprobación de proyectos, lo que convierte a este modelo en una herramienta relevante para el estudio del poder informal y la eficacia política en entornos legislativos complejos [Sotoudeh et al., 2024].

[10] A diferencia del caso chileno, en este caso existe la relación de patrocinio, en la cual un parlamentario no se asocia al proyecto como autor o coautor, sino como patrocinante.

- También a través del análisis de redes de co-autoría legislativa, se examina la influencia de los *caucus* (bancadas parlamentarios) en la dinámica colaborativa del Congreso brasileño, proponiendo un enfoque exploratorio que combina métricas de redes y alineamientos institucionales. La red se construye a partir de proyectos de ley con más de un autor, donde los nodos representan parlamentarios y las aristas indican coautorías compartidas. El análisis incorpora medidas como densidad, modularidad y centralidad, permitiendo evaluar en qué medida los caucus estructuran la cooperación legislativa y contribuyen a la formación de bloques cohesivos. Esta perspectiva relacional complementa las formas tradicionales de agrupación partidaria, revelando cómo las afinidades temáticas o ideológicas expresadas en estos frentes parlamentarios pueden tener un impacto significativo en la organización interna del Congreso. Los hallazgos sugieren que los caucus no solo actúan como vehículos de articulación política, sino también como ejes de influencia dentro de la red legislativa, modelando patrones de colaboración que trascienden las fronteras partidarias convencionales [Nery and Mueller, 2022].

- Para Chile, el estudio de [Morán, 2020] examina la evolución de la cooperación y la polarización en la Cámara de Diputados de Chile para el periodo 2006-2017, mediante técnicas de SNA aplicadas a redes de coautoría de proyectos de ley. Se construyeron dos tipos de redes: una considerando todos los proyectos y otra solo aquellos que fueron discutidos y votados en sala (proyectos exitosos). La información fue extraída por *web scraping* desde el portal del Congreso Nacional, incluyendo 4312 proyectos de ley. Para medir la cohesión, se emplearon métricas como densidad, coeficiente de agrupamiento (clustering) y longitud de caminos más cortos, evaluando también la presencia de estructuras de tipo *small-world*[11]. Para capturar la polarización, se segmentó a los parlamentarios en coaliciones sobre la cual se calculó la métrica de modularidad y se aplicó el algoritmo *walktrap* [Pons and Latapy, 2006] para detección de comunidades, midiendo la proporción de comunidades con composición transversal entre coaliciones. Los resultados muestran que aunque las redes de todos los proyectos son más densas, presentan una fuerte polarización interna. En contraste, las redes derivadas de proyectos exitosos exhiben menor modularidad y mayor presencia de comunidades mixtas entre coaliciones, indicando mayor cooperación transversal. Esto sugiere que, mientras la agenda legislativa tiende a reforzar la disciplina partidaria en la votación, actúa en sentido contrario al seleccionar colaboraciones más transversales en las iniciativas que logran avanzar. El estudio destaca cómo la configuración partidaria y el poder de *gatekeeping*[12] influyen en la estructura relacional del Congreso, y cómo el análisis de redes permite capturar dimensiones de cooperación política invisibles a través de otras métricas legislativas más tradicionales.

5.4.3 Detección de intereses parlamentarios

La detección de intereses parlamentarios en el contexto legislativo corresponde a un conjunto de tareas orientadas a identificar los temas sustantivos que captan la atención de los legisladores y sobre los cuales estos participan activamente en el debate parlamentario. Este enfoque se

[11] Un tipo de red que presenta alta agrupación entre los nodos tendiendo a formar grupos cerrados, y al mismo tiempo es posible llegar de un nodo a otro por pocas conexiones.

[12] En el contexto legislativo, el gatekeeping se refiere a la capacidad de ciertos actores u organismos para permitir o bloquear que una propuesta legislativa avance en el proceso. Es decir, es el poder de decidir qué se discute y qué queda fuera del debate.

basa en el análisis automatizado de documentos legislativos, intervenciones en el hemiciclo y registros asociados a la actividad parlamentaria. Para ello, se emplean técnicas de minería de texto, clasificación, procesamiento de lenguaje natural y, más recientemente, modelos de lenguaje, con el objetivo de extraer y categorizar los tópicos presentes en discursos, proyectos de ley, documentos de fiscalización y otras fuentes textuales. El propósito central de este tipo de análisis es inferir los focos temáticos predominantes en la acción legislativa y mapear las materias que configuran la agenda del debate político. A continuación, se presentan las principales experiencias y metodologías reportadas en la literatura especializada.

- Mediante el uso de técnicas de minería de texto combinadas con algoritmos de clasificación supervisada y clustering, se propone un sistema automatizado para analizar documentos legislativos del Congreso de Taiwán con el fin de representar el desempeño temático de cada legislador. A partir de un corpus compuesto por interpelaciones, discursos en sesiones, propuestas legislativas y mociones transitorias, se implementa un preprocesamiento basado en segmentación morfosintáctica utilizando CKIP[13] (generar tokens y asignar su categoría gramatical), extracción de frases nominales y cálculo de frecuencias Term Frequency (TF). Posteriormente, se aplica una estrategia de clustering en dos etapas (jerárquico + k-means) para identificar patrones temáticos, seguida por un modelo SVM entrenado con etiquetas definidas por expertos en ciencia política. Este enfoque permite clasificar los documentos en tres dimensiones definidas como dirección geográfica (5 categorías), grupos objetivo (30 grupos) y áreas temáticas legislativas (29 temáticas) [Lin et al., 2015].

- Mediante la aplicación de técnicas de minería de texto orientadas a la extracción automática de palabras clave, un estudio sobre las elecciones legislativas portuguesas de 2022 propone una metodología para analizar y visualizar los programas electorales de 16 partidos políticos. El enfoque parte del preprocesamiento lingüístico de los documentos, incluyendo tokenización, lematización y filtrado de términos irrelevantes y stopwords, seguido por la construcción de representaciones vectoriales de los textos. A continuación, se aplica un algoritmo extendido de extracción de palabras clave que combina criterios de frecuencia, coocurrencia y especificidad temática para identificar los términos más representativos tanto a nivel de documento individual como del conjunto global. Esta información se traduce en visualizaciones como nubes de palabras y gráficos comparativos, que permiten contrastar los énfasis que da cada partidos en sus discursos y declaraciones. El estudio demuestra que la incorporación de técnicas de NLP y análisis cuantitativo de texto en el ámbito electoral posibilita una evaluación sistemática y escalable de las agendas políticas, facilitando su comparación temática y fortaleciendo las capacidades de monitoreo automatizado del discurso programático [Campos et al., 2023].

- Aunque no es exactamente relacionado, otro estudio muestra la experiencia al combinar modelado de tópicos mediante LDA y Structural Topic Model (STM) con análisis cualitativo asociado al análisis de políticas públicas [Isoaho et al., 2021]. Partiendo de grandes corpus de documentos gubernamentales y debates legislativos de Finlandia, el método aplica LDA/STM para descubrir temas latentes y, a la vez, guía a los investigadores con una heurística práctica (formulación de preguntas, depuración del corpus, elección del número de temas y verificación manual) para que los resultados sean teóricamente significativos. Tras la extracción temática, los autores recomiendan volver al texto original

[13] `https://ckip.iis.sinica.edu.tw`

mediante lecturas dirigidas y codificación cualitativa, de modo que los tópicos identificados se interpreten en función de modelos conceptuales asociados a las políticas públicas. Esta estrategia se ilustra con un estudio de planes nacionales de energía, donde el modelado de temas permite rastrear la evolución de narrativas sobre sostenibilidad y seguridad energética a lo largo del tiempo y entre actores.

- Igualmente al caso anterior, otro estudio basado en STM aplicado a textos políticos provenientes de varios países e idiomas [Lucas et al., 2015] integra metadatos (como fecha, medio o contexto institucional) directamente en el proceso de estimar los temas, permitiendo comparar cómo varía la agenda temática entre partidos y en el tiempo. Los autores discuten retos de traducción automática y normalización lingüística, proponen flujos de trabajo reproducibles (descarga, limpieza, traducción, modelado y validación) y ofrecen recomendaciones de software. Demuestran la utilidad del enfoque con un análisis de artículos de prensa sobre China publicados por distintas agencias, evidenciando diferencias sistemáticas de cobertura asociadas a la línea editorial y al momento histórico. El STM facilita así el examen conjunto de volumen temático y de efectos de covariables en contextos multilingües, ampliando las posibilidades de la investigación en política comparada.

5.5 Usos controvertidos de tecnologías de la información en el ámbito político

La utilización de TI, y en particular algoritmos de Inteligencia Artificial (IA) en el ámbito político, ha generado un intenso debate sobre la forma en que estas tecnologías pueden utilizarse para influir en la percepción pública y en los distintos fenómenos políticos, tales como manifestaciones sociales o procesos electorales. A continuación se describen las principales estrategias de análisis utilizadas en el ámbito político y casos conocidos de su utilización, que han generado controversia en la opinión pública.

5.5.1 Estrategias de análisis en el ámbito político

En la actualidad, las campañas políticas recurren a un abanico de estrategias analíticas que buscan profundizar en las motivaciones, actitudes y comportamientos de la ciudadanía. Tanto en el diseño de mensajes como en la ejecución de tácticas de persuasión, el análisis de información detallada —ya sea sobre preferencias individuales o sobre segmentos de la población— cobra una relevancia creciente. Este apartado abordará algunas de esas estrategias en el ámbito político, centrándose en el papel que adquieren enfoques más específicos como la psicografía y la microsegmentación, considerados cada vez más indispensables para comprender y enfrentar los desafíos contemporáneos en materia de comunicación política.

Análisis Psicográfico

Desde el ámbito del marketing, se conoce como Análisis Psicográfico al estudio y clasificación de las personas según sus actitudes, aspiraciones y otros criterios de tipo psicológico. Se trata de un enfoque que combina métodos cualitativos y cuantitativos para detectar rasgos comunes en grupos de consumidores o usuarios, con el objetivo de segmentar el mercado a partir de factores psicológicos.

El propósito práctico del Análisis Psicográfico es reconocer patrones de personalidad, motivaciones y valores, a fin de diseñar contenidos persuasivos ajustados a las características cognitivas y emocionales de cada individuo [Jeria Cánovas and Wall Opazo, 2005]. La idea principal es comprender qué impulsa a una persona a tomar ciertas decisiones o a mostrar afinidad con determinados mensajes (es decir, el "por qué" de su comportamiento), con miras a desarrollar estrategias de comunicación o marketing altamente personalizadas.

Para ello, se analizan tanto aspectos cognitivos (creencias, actitudes) como factores emocionales (preferencias, temores), recurriendo a datos de perfiles en redes sociales y a técnicas de minería de texto (por ejemplo, análisis de sentimientos y clasificación) aplicadas sobre contenido generado por los usuarios (comentarios, respuestas o tweets[14]). Además, se emplean algoritmos de Big Data para descubrir patrones subyacentes en grandes volúmenes de información.

Con el fin de segmentar personalidades de manera más precisa, el Análisis Psicográfico suele valerse de modelos de personalidad como Big Five (o OCEAN[15]) y MBTI (Myers-Briggs Type Indicator). Dichos modelos permiten clasificar y agrupar a los individuos según sus rasgos psicológicos, lo que resulta útil para predecir conductas, preferencias y patrones de toma de decisiones en diversos contextos, incluido el político.

Microsegmentación

Por otro lado, la *microsegmentación* [Kotler and Keller, 2006] es la práctica de dividir una audiencia amplia en subconjuntos muy específicos, atendiendo a variables demográficas, de comportamiento y/o características psicográficas, o también responde el *"quién es"* o *"cómo se agrupan"* los usuarios. El objetivo de esta técnica es diseñar mensajes y estrategias de comunicación o marketing altamente personalizados, para que cada grupo reciba contenidos ajustados a sus motivaciones, preferencias y rasgos particulares. En el ámbito político, por ejemplo, se emplea la microsegmentación para adaptar discursos o anuncios electorales a segmentos muy concretos de votantes, maximizando el impacto del mensaje y favoreciendo la persuasión. A nivel tecnológico, esta técnica es posible de implementar mediante la recopilación de datos de usuarios provenientes de redes sociales, encuestas, historiales de navegación o bases de datos propias de empresas (de ahí la importancia de qué a quién le permitimos el uso de nuestros datos) utilizando algoritmos de clustering, análisis factorial o mediante algoritmos de recomendación tales como filtrado colaborativo (*Collaborative Filtering*) o filtrado basado en contenidos (*Based-Content Filtering*) [Adomavicius and Tuzhilin, 2005, Ansari et al., 2000], entre otros.

Con estas dos técnicas, es posible generar un mapeo preciso de la audiencia que, aplicado en el ámbito político, posibilita dirigir mensajes persuasivos capaces de ajustarse a las características psicológicas de cada usuario.

5.5.2 Casos de uso controvertidos

Campañas de Barack Obama en Estados Unidos

La estrategia tecnológica de la campaña de elección de Barack Obama en 2008 estuvo fuertemente apoyada por medios sociales [Bimber, 2014], pero no fue hasta su reelección en 2012, que se implementó el sistema Narwhal, una plataforma que integraba datos de votantes, donantes y

[14] "Tweet" es el término que describe las publicaciones en la red social X, antes conocida como Twitter.

[15] Este acrónimo corresponde a los cinco rasgos principales: Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism

voluntarios en un solo sistema [MIT, 2012]. Con un marcado énfasis en Big Data [Sudhahar et al., 2015], se recopilaron grandes volúmenes de información para alimentar algoritmos de análisis predictivo, dentro de ellas:

- *Historiales de voto*: Se utilizaron registros públicos de participación electoral que, sin revelar por quién se votó, permiten conocer la frecuencia con que un ciudadano ejerce su derecho a sufragio y, en algunos casos, su afiliación partidaria. Esta información, accesible bajo determinadas condiciones legales en diversos estados de EE. UU., resulta estratégica para estimar la propensión al voto y modelar la probabilidad de persuasión individual dentro de la campaña.

- *Redes sociales*: Se analizaron interacciones públicas o semipúblicas en plataformas como Twitter y Facebook, respetando los límites legales y las políticas de privacidad vigentes. A partir de los intereses declarados, los comentarios, los grupos seguidos o los temas compartidos, fue posible inferir afinidades temáticas y niveles de involucramiento político, aportando insumos clave para la personalización de mensajes y contenidos [Gerodimos and Justinussen, 20

- *Patrones de donación*: El análisis de donaciones a campañas pasadas permitió identificar tanto la disposición a donar como la frecuencia, montos y canales preferidos. Estos datos facilitaron la construcción de modelos predictivos capaces de estimar la probabilidad de una nueva donación bajo determinados estímulos, optimizando así las estrategias de recaudación y segmentación de mensajes financieros.

- *Encuestas y estudios cualitativos*: A través de encuestas telefónicas y presenciales, tanto estructuradas como abiertas, se recolectó información valiosa para validar perfiles psicográficos y evaluar las reacciones emocionales ante distintos tipos de discurso. Esta retroalimentación directa permitió ajustar el enfoque comunicacional según los valores y preocupaciones de distintos segmentos del electorado.

- *Datos sociodemográficos públicos*: Se integraron variables provenientes de fuentes oficiales como el Censo o el Servicio de Impuestos Internos estadounidense, incluyendo nivel educativo, ingresos, situación laboral, tipo de vivienda y composición del hogar. Esta información permitió enriquecer los modelos de segmentación y aumentar la precisión de las predicciones en distintos territorios y grupos sociales.

- *Bases de datos comerciales*: Se incorporaron datos adquiridos legalmente a través de data brokers especializados, como *Catalist*[16] y *TargetSmart*[17], que ofrecían perfiles agregados sobre hábitos de consumo, tenencia de vehículos, historial de mudanzas y otras variables comportamentales útiles para afinar la microsegmentación y definir mensajes contextualmente relevantes que ayudaron a inferir intereses, nivel económico y valores culturales.

- *Actividades en terreno*: La información recolectada mediante campañas puerta a puerta y llamadas telefónicas fue registrada sistemáticamente, incluyendo percepciones sobre el candidato, nivel de interés y disposición a votar. Estos datos, ingresados en tiempo real al sistema centralizado, permitieron retroalimentar los modelos analíticos y optimizar el despliegue territorial de la campaña.

[16] https://catalist.us/
[17] https://targetsmart.com/

Considerando todas estas fuentes de datos, el siguiente pilar fue el uso de microsegmentación, lo cual permitió dividir la base de votantes en segmentos muy específicos, recibiendo cada uno mensajes políticos afines a sus intereses e inquietudes. En paralelo, se realizaron pruebas del tipo A/B test de forma constante para refinar asuntos de correo, ofertas de donación y guiones de llamadas telefónicas. Con ello, se maximizó la eficacia de la recaudación de fondos y se optimizó la logística para movilizar a los votantes el día de la elección. Otro elemento clave fue la integración de un equipo diverso de desarrolladores, científicos de datos y expertos en marketing online, quienes diseñaron herramientas y paneles de control que se actualizaban en tiempo real con cada interacción. Todo ello se tradujo en un modelo centralizado de gestión de campaña, donde la toma de decisiones se basó en datos y la segmentación profunda de la ciudadanía, logrando un grado de personalización sin precedentes en la política estadounidense.

Asimismo, es importante destacar que, pese al uso intensivo de datos y las técnicas de Big Data que sustentaron gran parte de la estrategia, la campaña no recurrió a información obtenida de manera ilegal ni manipuló bases de datos sin consentimiento. Toda la recolección y el tratamiento de los perfiles de votantes, donantes y voluntarios se llevaron a cabo siguiendo criterios de transparencia y respeto a la privacidad, lo que garantizó la legitimidad de la información empleada en la toma de decisiones.

El caso de Cambridge Analytica

El caso de Cambridge Analytica constituye uno de los episodios más emblemáticos en el uso controvertido de técnicas de análisis psicográfico y microsegmentación con fines electorales. Durante la campaña presidencial de Estados Unidos en 2016, los equipos asociados al entonces candidato Donald Trump aplicaron modelos de segmentación que se apoyaban en perfiles psicológicos construidos a partir de datos masivos extraídos de redes sociales. La consultora obtuvo acceso no autorizado a los datos personales de decenas de millones de usuarios de Facebook, sin su consentimiento explícito, y los correlacionó con métricas psicográficas basadas en la taxonomía de los Big Five, mediante algoritmos de aprendizaje automático.

Esta infraestructura analítica permitió emitir alrededor de 50.000 mensajes políticos personalizados por día, dirigidos a segmentos altamente específicos del electorado, con el propósito de reforzar percepciones o inducir cambios sutiles en la conducta electoral. El sistema infería rasgos de personalidad a partir de "me gusta", comentarios e interacciones en línea, y ajustaba el contenido de los mensajes en función de la vulnerabilidad cognitiva o emocional de cada perfil.

El escándalo se hizo público en 2018, generando una fuerte reacción institucional y mediática en torno a la privacidad de los datos y la manipulación mediante algoritmos en plataformas digitales. A raíz de esto, tanto Cambridge Analytica como Facebook fueron objeto de investigaciones y sanciones, y el caso impulsó reformas regulatorias y nuevas exigencias de transparencia en la gestión de datos personales. Este episodio reveló el potencial de la IA para amplificar prácticas de manipulación política, poniendo en evidencia la necesidad urgente de marcos normativos que garanticen el uso ético de tecnologías de segmentación en procesos democráticos.

Otros casos controvertidos conocidos

Además de Cambridge Analytica, se han documentado otros episodios donde el uso combinado de *big data*, IA y publicidad política hiperdirigida se utilizó para traspasar los límites éticos de la comunicación electoral:

El referéndum del Brexit en el Reino Unido Durante la campaña *Vote Leave* en 2016, consultoras vinculadas a AggregateIQ emplearon bases de datos provenientes de redes sociales, registros de consumo y censos locales para construir perfiles psicográficos de los votantes británicos. A través de la plataforma de anuncios de Facebook se difundieron miles de variaciones de mensajes sobre inmigración, gasto público y soberanía [Hall, 2022]. Investigaciones parlamentarias posteriores señalaron que se desconocía la procedencia de los datos utilizados para la campaña, posibles violaciones de los topes de gasto electoral y falta de consentimiento informado de los usuarios, lo que motivó multas de la *Information Commissioner's Office* e impulsó exigencias de trazabilidad de los anuncios políticos en línea [UK Parliament, 2019].

La campaña presidencial de Jair Bolsonaro en Brasil Durante la campaña de Jair Bolsonaro en 2018 en Brasil existen al menos tres aspectos que han sido documentados con claridad. Primero, distintos estudios coinciden en señalar un uso intensivo de WhatsApp durante la campaña presidencial. Segundo, tanto su relevancia como su centralidad en el debate público fueron ampliamente destacadas por analistas y académicos. Y tercero, diversos trabajos identifican a la plataforma como un canal clave para la circulación de contenidos falsos y estrategias de desinformación. En esa línea, un informe elaborado por IDEIA Big Data para Avaaz reveló que el 98% de los votantes de Bolsonaro estuvo expuesto a una o más noticias falsas difundidas por WhatsApp, y que el 90% consideró que esas historias eran verdaderas [Brito Cruz et al., 2019].

Capítulo 6

Marco de trabajo de las Tecnologías Semánticas

6.1 Introducción

El procesamiento automatizado de información político-legislativa plantea desafíos complejos asociados a la naturaleza no estructurada de los documentos, la diversidad de fuentes y formatos, y la necesidad de proveer mecanismos de análisis que permitan una comprensión significativa de la actividad parlamentaria. En este contexto, la presente investigación propone un marco de trabajo sustentado en el uso de Tecnologías Semánticas, con el objetivo de abordar dichas problemáticas mediante una solución integrada y escalable.

Las *Tecnologías Semánticas* se definen aquí como la confluencia de tres disciplinas clave:

- *Web Semántica*: que entrega mecanismos para dotar de significado e interoperabilidad a los datos mediante tecnologías como URIs, RDF, OWL y SPARQL.
- *Minería de Textos*: que habilita la identificación y extracción automática de elementos presentes en el texto tales como estructura, entidades, conceptos, relaciones y metadatos desde fuentes textuales.
- *Análisis de Redes Sociales*: un conjunto de técnicas de análisis orientadas a la exploración estructural y relacional de vínculos entre entidades, para el caso de este trabajo en contextos político-legislativos.

Este capítulo se organiza en dos partes complementarias. La primera sección aborda el marco conceptual y teórico, presentando los conceptos tecnológicos que sustentan el enfoque adoptado, incluyendo conceptos de las tres disciplinas de estudio. En la segunda sección se presenta el marco de trabajo técnico, describiendo la arquitectura diseñada, los componentes definidos y los criterios metodológicos que guían su implementación e interconexión.

6.2 Marco conceptual y teórico

6.2.1 Web Semántica

La Web Semántica surgió alrededor del año 2000, basada en la idea de extender el enlace entre dos páginas existentes en la Web (no semántica), que carece de significado más allá de una

simple conexión (en HTML, la etiqueta <a href...>), a una en la que los enlaces tengan un significado específico y legible por máquinas. En ese entonces se pensaba en prospectiva con la creación de agentes inteligentes capaces de navegar, entender y resolver problemas humanos explorando la Web. Para lograr esto, la Web Semántica propone una pila tecnológica basada en dos elementos fundamentales derivados de la Web tradicional:

1. *Uniform Resource Identifier (URI)*: Es una cadena de caracteres que se utiliza como identificador de recursos desde un punto de vista abstracto y se diferencia ligeramente de las URL, que localizan documentos. Esto significa que un único recurso puede tener un URI, pero varias URLs, dependiendo de las representaciones del recurso. Por ejemplo, si un artículo se publica en un URI, al acceder a él se puede obtener una representación legible por máquinas para una aplicación, mientras que un usuario humano podría obtener una copia legible en su idioma.

2. *RDF*: Es un formato declarativo que permite describir un recurso definido por un URI a través de atributos y relaciones con otros recursos, y se puede implementar en múltiples sintaxis, como JSON, XML o CSV. La idea detrás del uso de URIs y RDF es que tanto los datos como los modelos de datos (vocabularios, taxonomías y ontologías) pueden describirse en el mismo formato y bajo un modelo de interoperabilidad basado en HTTP.

Linked Data

Con base en los elementos técnicos base URI y RDF, en 2006 el creador de la Web, Tim Berners-Lee, propuso la Web de Datos y los Datos Abiertos Enlazados (*Linked Open Data*) [Berners-Lee, 2006], promoviendo la adopción internacional de estándares para la publicación de datos utilizando tecnologías de la Web Semántica a nivel gubernamental. El sitio *The Linked Open Data Cloud* (`lod-cloud.net`) recoge desde 2007 la composición de los distintos conjuntos de datos vinculados en el mundo a través de tecnologías de Web Semántica, en específico RDF. A lo largo del tiempo, la nube de datos abiertos registrada en el sitio ha crecido desde 12 conjuntos de datos enlazados abiertos en 2007 (figura 6.1) hasta a 1.357 conjuntos de datos a marzo de 2025 (figura 6.2).

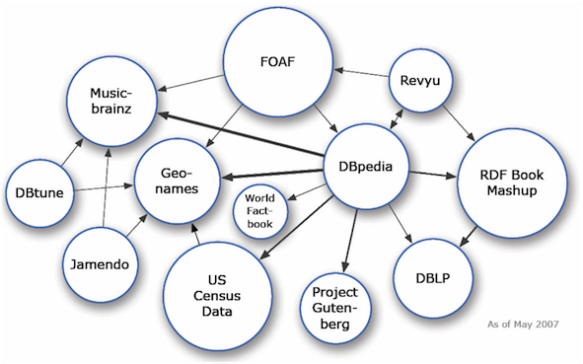

Figura 6.1: Diagrama lod-cloud.net a mayo de 2007

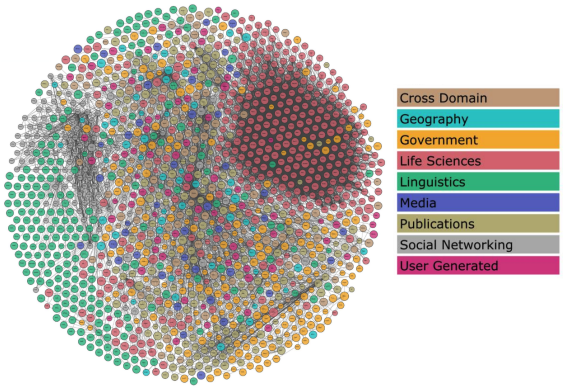

Figura 6.2: Diagrama lod-cloud.net a marzo de 2025

Knowledge Graphs

Siguiendo la definición propuesta en la literatura especializada, un grafo de conocimiento (*Knowledge Graph*) corresponde a una estructura de datos orientada a representar y transmitir conocimiento sobre el mundo real, donde los nodos modelan entidades y los bordes capturan relaciones entre ellas [Hogan et al., 2021]. Esta estructura se construye sobre un modelo de grafos, como grafos dirigidos con aristas etiquetadas o grafos de propiedades. El conocimiento representado puede provenir de fuentes externas o inferirse del propio grafo, y se expresa mediante enunciados simples (ejemplo: "Santiago es la capital de Chile") o cuantificados (ejemplo: "todas las capitales son ciudades"). Mientras que los enunciados simples pueden integrarse directamente como aristas, los cuantificados requieren representaciones más expresivas, como ontologías o reglas. En este contexto, es posible aplicar métodos deductivos para inferir nuevo conocimiento, o bien técnicas inductivas que permitan enriquecer progresivamente el grafo a partir de patrones observados. Un grafo de conocimiento, en el contexto de la Web Semántica, está descrito en RDF, y puede ser integrado de forma natural con otros mediante recursos compartidos como nodos o propiedades.

Negociación de contenido

El concepto de *negociación de contenido* en el contexto de la Web Semántica, se refiere al mecanismo que permite acceder a diferentes representaciones de un mismo recurso identificado por una URI. Si bien una URI define un recurso, este puede ser descrito utilizando diversas sintaxis RDF (como RDF/XML, Turtle o JSON-LD), por lo que la URI no debería estar acoplada a una forma específica de representación. Asociar explícitamente la URI a un formato particular constituiría un error conceptual, al confundir el recurso con su representación.

La negociación de contenido entra en juego cuando un cliente accede a una URI sin especificar el formato deseado. En ese caso, el servidor evalúa las cabeceras HTTP [Berners-Lee et al., 1996] que acompañan la solicitud, las cuales incluyen información como los tipos de contenido aceptados, la codificación de caracteres y el idioma preferido. Con base en estas cabeceras, el servidor puede responder utilizando el código HTTP 303 "See Other", redirigiendo al cliente hacia una URI alternativa que contiene la representación más adecuada del recurso. Posteriormente, el cliente accede a esta nueva URI y recibe la representación solicitada con una respuesta exitosa (código 200 "OK"). La figura 6.3 explica a nivel gráfico el mecanismo de negociación de contenido.

SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) [W3C, 2013] es el lenguaje estándar para la consulta y manipulación de datos representados en RDF. Como uno de los pilares fundamentales de la Web Semántica, SPARQL permite extraer, filtrar, transformar y combinar datos estructurados mediante patrones de tripletas, facilitando la interoperabilidad entre fuentes de datos heterogéneas, ya sea que estén almacenadas nativamente como RDF o expuestas como tales mediante capas intermedias. Su especificación define tanto la sintaxis como la semántica del lenguaje, soportando consultas de diversos tipos (SELECT, CONSTRUCT, ASK, DESCRIBE), así como patrones requeridos u opcionales, conjunciones y disyunciones, subconsultas, agregaciones, negaciones, generación de valores mediante expresiones, y restricciones por grafo de origen. Los resultados pueden entregarse como conjuntos de resultados tabulares o

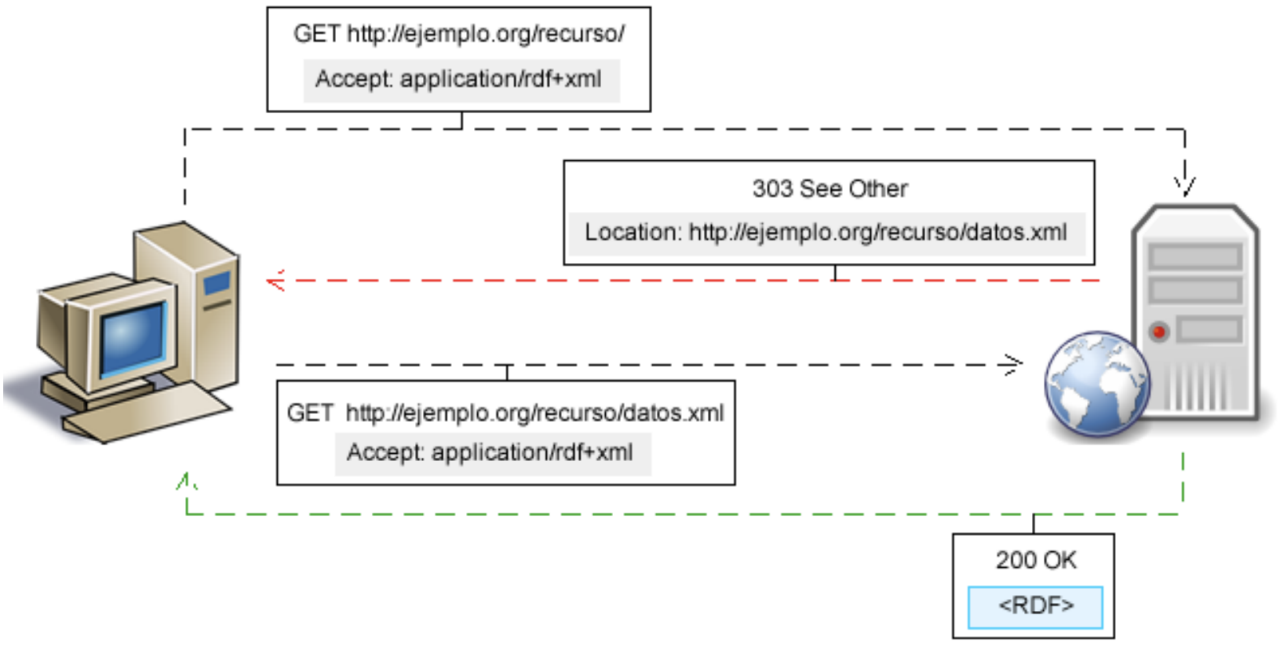

Figura 6.3: Mecanismo de negociación de contenido

como grafos RDF, lo que permite una amplia flexibilidad para integrar y explotar conocimiento distribuido en la Web.

Ontologías

En el contexto de la Web Semántica, una ontología es una especificación formal y explícita de un conjunto de conceptos, propiedades y relaciones relevantes dentro de un dominio determinado (*o una especificación formal de una conceptualización*[Gruber, 1993]). Su propósito es estructurar y dar significado a los datos, utilizándose como esquemas y modelos de datos, facilitando la interoperabilidad, la inferencia automática y el razonamiento sobre la información distribuida en la web. Para su construcción y representación, se utilizan RDF y lenguajes como RDF Schema (RDFS) [Guha and Brickley, 2014], que permite definir jerarquías básicas de clases y propiedades, y Ontology Web Language (OWL) [owl, 2012], que ofrece una mayor expresividad para modelar restricciones, cardinalidades, equivalencias y axiomas lógicos más complejos. Juntas, estas tecnologías permiten que los agentes automatizados comprendan, consulten y razonen sobre los datos de forma semánticamente enriquecida.

Validación estructural de grafos RDF

Dado que los grafos de conocimiento integran datos provenientes de fuentes heterogéneas y con distintos niveles de calidad, se hace necesario garantizar su consistencia estructural. A medida que estos grafos crecen en complejidad, se vuelve fundamental contar con mecanismos que permitan validar que los datos cumplen con reglas específicas de forma, cardinalidad y estructura esperada. En este marco, han surgido lenguajes de validación estructural específicos para RDF, entre los que destacan Shape Expressions (ShEx) y Shapes Constraint Language (SHACL). Ambos permiten definir y comprobar restricciones sobre nodos y propiedades en

grafos RDF, pero lo hacen con enfoques y sintaxis distintas, ofreciendo distintas ventajas según el caso de uso.

ShEx [Prud'hommeaux et al., 2014] es un lenguaje compacto y declarativo diseñado específicamente para describir y validar la forma de nodos dentro de un grafo RDF. Su diseño se centra en la simplicidad de expresión y en la facilidad de uso para validar estructuras esperadas, mediante definiciones que indican qué propiedades deben o pueden estar presentes en una entidad, junto con sus tipos, cardinalidades y posibles valores. ShEx es particularmente útil en etapas tempranas de desarrollo o para generar validaciones automáticas a partir de grafos existentes, permitiendo detectar rápidamente errores estructurales o inconsistencias en los datos.

Por su parte, SHACL [Knublauch et al., 2017] es una recomendación oficial del W3C que proporciona un marco más amplio y flexible para la validación de datos RDF, incluyendo capacidades para definir restricciones tanto estructurales como semánticas. SHACL permite describir "shapes" o formas que actúan como plantillas contra las cuales se evalúan los nodos del grafo, y soporta expresiones complejas, validaciones condicionales, integración con SPARQL y reutilización modular. Gracias a su robustez y respaldo institucional, SHACL se ha convertido en una opción ampliamente adoptada en entornos de producción que requieren validaciones más sofisticadas o integradas con reglas de negocio.

Estos lenguajes conforman el núcleo de los mecanismos de validación estructural en entornos RDF, ofreciendo garantías formales sobre la calidad y adecuación de los datos, aspecto clave para aplicaciones semánticas robustas y confiables.

XML

Extensible Markup Language (XML) es un lenguaje de marcado flexible para texto plano, diseñado para representar información estructurada de forma jerárquica y legible tanto por humanos como por máquinas. Su función principal es la codificación de datos en un formato estandarizado y autodescriptivo, el cual es ampliamente utilizado para el intercambio de información entre sistemas. En el contexto de la Web Semántica y el marcaje de documentos, XML actúa tanto como base sintáctica sobre la cual se construyen representaciones de recursos como RDF/XML, como también en formatos especializados como AKN, permitiendo anotar semánticamente contenido textual, estructurar documentos legislativos, y facilitar su posterior procesamiento, integración y reutilización. Adicionalmente, XML tiene múltiples formas de uso, tales como configuración de sistemas, representación de datos bibliográficos, interoperabilidad en servicios Web y almacenamiento de datos estructurados entre otros.

6.2.2 Minería de Textos

La Minería de Textos emergió como un área interdisciplinaria que combina técnicas de Procesamiento de Lenguaje Natural (PLN), lingüística, aprendizaje automático y recuperación de información, con el propósito de extraer conocimiento útil desde datos textuales no estructurados. Su desarrollo responde a la necesidad de automatizar la comprensión y el análisis de textos en dominios donde el volumen de información excede la capacidad humana de revisión manual. En el contexto de sistemas relacionados al ámbito político-legislativo, la minería de textos permite identificar entidades, relaciones, emociones, patrones lingüísticos y tópicos recurrentes en distintos tipos de documentos, tales como legislación o intervenciones parlamentarias. Un aspecto transversal de los elementos de este conjunto de tecnologías, son los mecanismo de evaluación de la calidad, donde se utilizan métricas previamente establecidas, las que se definen en

el anexo H.

A continuación se repasan las principales herramientas y técnicas de la minería de textos, aplicadas al ámbito del trabajo.

Técnicas de preprocesamiento

Dentro de las primeras etapas del procesamiento en Minería de textos, se encuentra un amplio conjunto de técnicas de PLN que actúan como fase preliminar para segmentar, normalizar y estructurar el contenido textual. Estas técnicas son fundamentales para facilitar tareas posteriores, como la extracción de información o la clasificación automática. A continuación se describen algunas de las técnicas más utilizadas en el análisis de textos político-legislativos:

- *Tokenización:* consiste en dividir un texto en unidades mínimas denominadas *tokens*, que pueden ser partes de palabras (*multi-word token expansion*), palabras, o frases con un número de palabras determinado (*n-gramas*). Es el paso inicial en la mayoría de los flujos de procesamiento lingüístico.

- *Lematización:* Consiste en reducir una palabra a su forma canónica o lema (ejemplo: "viviendo" → "vivir"), lo cual permite agrupar diferentes formas flexionadas de una misma palabra bajo una única representación, facilitando el posterior análisis.

- *Stemming:* En parte similar a la lematización, el stemming aplica reglas heurísticas para eliminar afijos (ejemplo: "corriendo" → "corr"), permitiendo agrupar variantes bajo una misma forma para tratarlas como equivalentes, lo cual ayudará a reducir la dimensionalidad del texto.

- *POS tagging:* También conocido como *etiquetado morfosintáctico* o *etiquetado de parte de la voz*, es la asignación de categorías gramaticales (sustantivo, verbo, adjetivo, etc.) a cada palabra de un texto, lo que permite comprender su función dentro de una oración.

- *Eliminación de palabras vacías:* También conocidas como *stopwords*, son palabras que no agregan un significado relevante y por lo general son frecuentemente utilizadas dentro de cualquier texto.

- *Desambiguación léxica:* También conocido como Word-sense disambiguation (WSD), es un proceso de identificar el significado correcto de una palabra con múltiples acepciones con base en su contexto.

- *Resolución de correferencias:* Es la tarea de encontrar todas las expresiones que hacen referencia a la misma entidad en un texto, lo cual es útil para tareas de PLN tales como la generación automática de resúmenes, la respuesta a preguntas y la extracción de información.

- *Análisis de frecuencia:* Esta tarea consiste en calcular la frecuencia con que aparecen elementos como palabras, entidades o n-gramas en una misma ventana de contexto. Este análisis permite ponderar los elementos con "mayor peso" en el texto, bajo distintas perspectivas, tales como: TF, Term Frequency-Inverse Document Frequency (TF-IDF) o Bolsa de palabras (*bag of words*).

- *Análisis de co-ocurrencia:* Se enfoca en calcular la frecuencia con que aparecen conjuntamente (co-ocurren) palabras, entidades o n-gramas en una misma ventana de contexto. El análisis de estas co-ocurrencias resulta útil para identificar términos relevantes y posibles relaciones semánticas entre los elementos del texto. Algunos de las técnicas para análisis de co-ocurrencia más utilizadas son: matrices de co-ocurrencia, *Pointwise Mutual Information* o el algoritmo de ventana deslizante (*sliding window*).

Reconocedor de Entidades Nombradas

Un Named Entity Recognition (NER) es una herramienta que identifica y clasifica entidades mencionadas en un texto, como nombres propios, fechas, lugares, eventos o leyes, asociándolas con uno o varios tipos de entidad, en función de una probabilidad de acierto o puntaje. Para la elaboración de un NER es fundamental contar con texto previamente etiquetado (normalmente bajo el esquema BIO - BILOU [Ratinov and Roth, 2009]) que permita entender los patrones del texto. Esto es fundamental, ya que el acierto en la detección de las entidades va a depender, además de la calidad intrínsica de los algoritmos y modelos de aprendizaje automático utilizados (como *Conditional Random Fields* o modelos de aprendizaje tipo *Transformers*), de la calidad de los datos de entrenamiento, de su idioma o de su forma de redacción (distintos tipos son por ejemplo: un discurso parlamentario, una noticia, un comentario de redes sociales o un artículo técnico) dentro de otros.

Marcador estructural de texto

Se define como detección estructural del texto a la tarea de identificar grupos de cadenas consecutivas que, desde la perspectiva de un lector humano, corresponden a elementos como títulos, subtítulos, párrafos, secciones (conjuntos de párrafos bajo un mismo título o subtítulo), enumeraciones, listas y otras estructuras que se puedan definir dependiendo del contexto de aplicación. Para el caso de los textos legislativos, se define un tipo especial de elemento estructural denominado *intervención*, en el cual se describe lo hablado por una persona, pudiendo estar compuesto por uno o más párrafos consecutivos. De esta manera, se define el marcador estructural como la herramienta encargada de ejecutar la detección estructural, agregando marcas al texto que indican el inicio y fin de cada elemento estructural. En el caso de estudio, dichas marcas se representan en formato XML.

Dentro de las estrategias principales utilizada para la detección de secciones estructurales y jerárquicas en documentos de texto, está el uso combinado de expresiones regulares y la aplicación de reglas que encapsulan lógica programática, las cuales se ejecutan al detectar patrones específicos en el texto. Esta solución resulta especialmente práctica en el contexto de los documentos generados en el Congreso Nacional, ya que suelen seguir ciertas normas de redacción estandarizadas. De este modo, la herramienta permite identificar secciones estructurales de primer, segundo y tercer nivel, secuencias de elementos mediante listas numeradas y no numeradas (incluso anidadas), así como las participaciones de parlamentarios. También alternativamente, es posible en la actualidad hacer uso de aplicaciones basadas en Transformers, de la misma forma en que se realiza el reconocimiento de entidades.

Desambiguación de entidades

La desambiguación de entidades, o *Entity Linking*, consiste en identificar menciones a entidades específicas en un texto (como personas, organizaciones, lugares, fechas o leyes) y vincularlas

automáticamente con su identificador (URI) en un grafo de conocimiento (para el caso de estudio, a través de un endpoint SPARQL). A diferencia del NER, que detecta y clasifica el tipo de entidad mencionada en el texto, la desambiguación busca resolver la ambigüedad asociada a nombres que pueden referirse a distintos individuos u objetos según el contexto e identificar a qué individuo u objeto específico se hace referencia. Para aumentar la precisión de la herramienta, se pueden definir datos complementarios de contexto que permitan reducir el conjunto de alternativas posibles al seleccionar la URI correspondiente para cada etiqueta a identificar. De esta manera, suponiendo un uso en documentos del Congreso Nacional (y que por ejemplo se quisieran desambiguar nombres de personas), algunos datos de contexto útiles pueden ser la fecha de la sesión, la cámara del documento, el número de la sesión o el periodo legislativo.

Clasificación de textos

La clasificación de textos es el proceso mediante el cual se asignan etiquetas o categorías predefinidas a documentos en función de su contenido. Su utilización, al igual que sus mecanismos de implementación, abarcan un amplio abanico de escenarios y opciones. En el contexto de este trabajo, se explorarán los mecanismos comúnmente utilizados en el contexto de los instrumentos desarrollados, los cuales están enfocados en la clasificación de texto de intervenciones parlamentarias, y que se asocian al uso de algoritmos de clasificación basados en *aprendizaje supervisado*.

El aprendizaje supervisado se caracteriza por "aprender" a realizar una tarea específica a partir de un conjunto de datos de ejemplo. De esta manera, un clasificador puede aprender a clasificar un texto en dos o más categorías a partir de datos de ejemplo clasificados en esas dos o más categorías.

Para materializar el aprendizaje, un clasificador puede ser implementado usando distintas estrategias que permitan encontrar los patrones que servirán para la identificación y asignación correcta de las categorías al texto. Estas estrategias se dividen en dos grandes aspectos a explorar durante el desarrollo del clasificador, que son las siguientes:

1. *Selección de modelos y algoritmos*: Son herramientas matemáticas y computacionales que procesan los datos de entrenamiento y ajustan parámetros internos para optimizar su capacidad predictiva. Los algoritmos de aprendizaje automático (*Machine Learning*) abarcan desde árboles de decisión, modelos lineales (como regresión) hasta técnicas más avanzadas, como redes neuronales profundas o algoritmos de refuerzo. El proceso de selección del algoritmo a utilizar es fundamental para obtener un buen resultado en la clasificación, por lo cual, se utilizan métricas estándar para validar y comparar su calidad.

2. *Ingeniería de características:* es el proceso de seleccionar, transformar y crear variables relevantes a partir de datos brutos (texto plano) con el objetivo de mejorar el rendimiento y la eficacia de los modelos de aprendizaje automático. Implica identificar las características más significativas, eliminando aquellas redundantes o irrelevantes, y generando nuevas representaciones que capturen patrones y relaciones complejas dentro de los datos. Este proceso incluye técnicas como la normalización, codificación de variables categóricas, generación de interacciones entre variables, creación de agregados estadísticos y extracción de características a partir de datos no estructurados como texto, imágenes o series temporales. Una buena ingeniería de características es clave para maximizar la precisión

del modelo y reducir su complejidad, ya que permite representar de manera eficiente la información relevante del problema que se desea resolver.

Análisis de sentimientos: Una aplicación práctica de la clasificación de textos es el Análisis de sentimientos (también denominado minería de opinión), el cual utiliza técnicas de PLN para inferir la orientación emocional de un texto (positiva, negativa o neutral) u otros tipos de información subjetiva. Aunque su aplicación en documentos legislativos es menos directa, puede ser útil para detectar tonos críticos/consensuales o de apoyo/rechazo en intervenciones parlamentarias, como también para generar características que apoyen otros procesos de clasificación.

Modelado de Tópicos

El modelado de tópicos es una técnica estadística utilizada para identificar automáticamente los temas latentes presentes en una colección de documentos, sin necesidad de anotaciones previas (no supervisada). Su objetivo es descubrir patrones de coocurrencia de palabras que tienden a agruparse en torno a conceptos subyacentes en el texto, permitiendo así representar cada documento como una distribución probabilística sobre un conjunto de tópicos. Uno de los algoritmos más representativos en este ámbito es LDA [Blei et al., 2003], que asume que cada documento está compuesto por una mezcla de temas, y que cada tema está caracterizado por una distribución de palabras. Esta técnica resulta especialmente útil en contextos de análisis exploratorio de grandes volúmenes de texto, como corpus legislativos, ya que permite identificar automáticamente áreas temáticas recurrentes, trazar la evolución de ciertos temas a lo largo del tiempo o agrupar documentos según afinidades semánticas.

6.2.3 Análisis de Redes Sociales

El SNA surgió como una metodología interdisciplinaria que permite representar, modelar y estudiar las interacciones y estructuras relacionales dentro de sistemas sociales complejos. Su aparición se remonta aproximadamente a la década de 1930, con trabajos sobre *sociometría* [Moreno, 1934] donde se introduce la idea de mapear gráficamente las relaciones interpersonales dentro de grupos sociales, mediante herramientas como el *sociograma*. Sin embargo, fue recién en la década de 1960 cuando se incorporaron formalmente herramientas de la teoría de grafos [Harary et al., 1965], lo que permitió dotar al análisis de redes de una base matemática y computacional sólida. Posteriormente, en la década de 1970, el SNA se consolidó como un campo interdisciplinario, impulsado por la creación de la *International Network for Social Network Analysis (INSNA)*[1], lo que contribuyó a su expansión y formalización como disciplina científica.

El SNA busca modelar las relaciones entre las entidades de un sistema con el objetivo de describir la estructura de la red social resultante, tanto de forma gráfica como mediante métricas capaces de explicar determinados patrones estructurales. De esta manera, es posible estudiar fenómenos como el impacto de la estructura en el funcionamiento general del sistema, la influencia que ejercen actores individuales sobre un grupo o la evolución que experimenta la red a lo largo del tiempo. Para ello, los constructos básicos de una red a modelar son los *nodos* representan entidades (tales como personas, organizaciones o documentos) y los *enlaces* que representan relaciones de diversa índole. En el ámbito político-legislativo, el análisis de redes sociales constituye una herramienta clave para analizar las formas en que los parlamentarios

[1] `https://www.insna.org/about-us`

se organizan y conectan, permitiendo develar las estructuras por donde se canaliza el poder político. En este contexto, a continuación se presentan las principales métricas y mecanismos utilizados durante el desarrollo de la investigación.

Fundamentos básicos

Para describir las redes, lo primero es describir sus elementos básicos, esto es nodos y enlaces. Como se menciona anteriormente, las redes sociales se modelan matemáticamente como grafos, donde un *nodo* puede ser cualquier entidad que sea modelable (real o abstracta) y *enlaces* entre los nodos, que describen una relación entre dos nodos. Según el tipo de grafo, los enlaces pueden ser dirigidos (desde A hasta B) o no dirigidas. También los enlaces pueden tener pesos que representen la intensidad o frecuencia de las interacciones.

Existen múltiples tipos de grafos, los cuales se utilizan dependiendo de las características del problema a representar, como también acorde a los datos disponibles, cada uno con distintas características y particularidades. Para el caso de esta investigación, se utilizaron dos tipos de grafos no dirigidos: uno simple y otro denominado grafo bipartito.

- Un *grafo simple no dirigido* (figura 6.4) es un tipo de grafo compuesto por un conjunto de nodos y un conjunto de enlaces que conectan pares de nodos, donde cada enlace es bidireccional, no existe más de un enlace entre el mismo par de nodos (o puede existir pero no es considerado), y no se permiten enlaces que conecten un nodo consigo mismo (es decir, no existen lazos). Formalmente, un grafo simple no dirigido $G = (V, E)$ está definido por un conjunto de nodos V y un conjunto de enlaces E, donde cada enlace es un subconjunto no ordenado de dos nodos distintos de V. Esta estructura resulta adecuada para modelar relaciones simétricas entre entidades, tales como amistad mutua, colaboración conjunta o coautoría en proyectos legislativos.

- Un *grafo bipartito* (figura 6.5) es un tipo de grafo cuyos nodos pueden dividirse en dos conjuntos disjuntos, de manera tal que cada enlace conecta exclusivamente un nodo de un conjunto con un nodo del conjunto opuesto. Formalmente, un grafo $G = (V, E)$ es bipartito si el conjunto de nodos V puede particionarse en dos subconjuntos V_1 y V_2, cumpliéndose que $\forall (u, v) \in E$, se tiene $u \in V_1$ y $v \in V_2$ o viceversa, y no existen enlaces entre nodos del mismo conjunto. Esta estructura resulta particularmente útil para modelar sistemas en los cuales las relaciones solo se producen entre dos tipos diferentes de entidades, como en redes de colaboración autor-proyecto de ley o parlamentario-partido político.

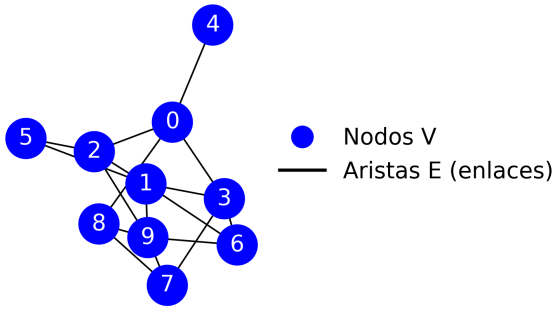

Figura 6.4: Grafo simple no dirigido

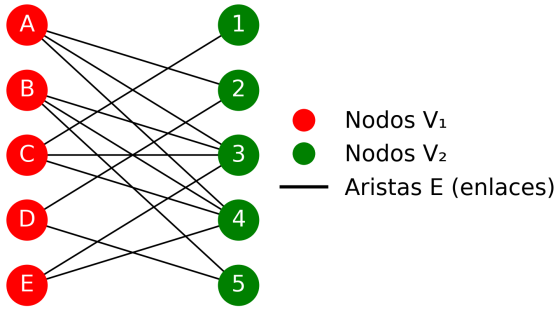

Figura 6.5: Grafo bipartito

A partir de un grafo bipartito es posible la construcción de lo que se denomina un *grafo proyectado*. Esta operación consiste en generar un nuevo grafo que preserva únicamente los nodos de uno de los dos conjuntos originales, estableciendo una conexión directa entre dos nodos si comparten al menos un vecino en el conjunto opuesto. El resultado es un grafo simple o ponderado que refleja indirectamente las relaciones mediadas a través de los nodos eliminados, permitiendo analizar patrones de interacción, similitud o colaboración dentro de un solo conjunto. La figura 6.6 muestra el proceso de proyección del grafo bipartito de la figura 6.5, el cual da como resultado el grafo proyectado de la figura 6.7, el cual es un grafo simple no dirigido.

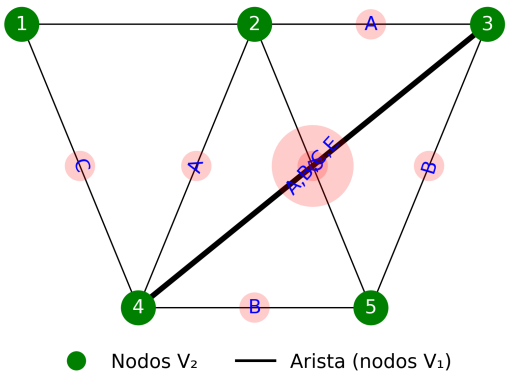

Figura 6.6: Proceso de proyección en función de relaciones con nodos del otro tipo

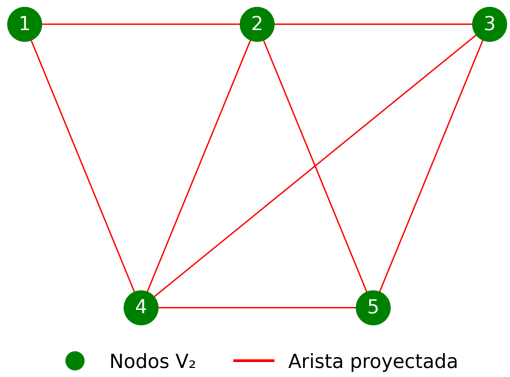

Figura 6.7: Grafo proyectado a partir de un grafo bipartito

Para el esta investigación, se utilizó este esquema de grafos proyectados con base en grafos bipartitos, donde los conjuntos disjuntos utilizados fueron documentos y parlamentarios, y las relaciones entre estos dos conjuntos permitieron generar los grafos entre parlamentarios.

Métricas y algoritmos sobre grafos

El análisis de redes expresadas en grafos se apoya en una amplia variedad de métricas y algoritmos, los cuales permiten abordar las estructuras desde múltiples perspectivas y para diversos fines de análisis. Desde un punto de vista matemático, es importante destacar que existe un isomorfismo entre grafos y matrices, es decir, toda red puede representarse mediante matrices y, recíprocamente, una matriz puede expresar una red. Esta equivalencia facilita tanto la operatoria como el cálculo de métricas mediante técnicas algebraicas. Con esto en cuenta, a continuación se presentan las métricas más relevantes utilizadas durante la investigación.

Caracterización general de un grafo Un primer aspecto fundamental es la caracterización global de un grafo en función de sus propiedades estructurales básicas. Entre las principales medidas se encuentran:

- *Tipos de nodos*: Se refiere a la naturaleza de las entidades representadas, si es que el grafo conecta entidades de distintos tipos o clases.
- *Tipos de conexiones*: Se refiere a la naturaleza y sentido de las relaciones que se establecen entre los nodos. En este contexto pueden existir relaciones uni o bidireccionales, y al mismo tiempo, en un mismo grafo pueden existir relaciones con variedad de significados.

- *Número de nodos*: Cantidad de nodos que forman parte del grafo.
- *Número de conexiones*: Cantidad de enlaces existentes entre los nodos.
- *Número de componentes*: Cantidad de subgrafos independientes que conforman el grafo, es decir, grupos de nodos conectados entre sí pero aislados respecto a otros grupos.
- *Número de puentes (bridges)*: Cantidad de enlaces o nodos cuya eliminación provoca la fragmentación del grafo, separando componentes que, de otro modo, permanecerían conectados.

Métricas asociadas a nodos A nivel de nodo individual, es posible calcular diversas métricas que permiten describir su rol estructural dentro de la red:

- *Grado (Degree)*: Número de conexiones directas de un nodo. Indica su nivel de actividad o interacción en la red.
- *Centralidad de grado (Degree Centrality)*: Cuantifica la importancia de un nodo en función del número de enlaces que posee respecto al total de nodos posibles.
- *Centralidad de intermediación (Betweenness Centrality)*: Mide la frecuencia con la cual un nodo actúa como intermediario en las rutas más cortas que conectan a otros nodos. Representa su capacidad de control o influencia en los flujos de información.
- *Centralidad de cercanía (Closeness Centrality)*: Evalúa qué tan cerca está un nodo del resto, considerando la suma de las distancias más cortas hacia todos los demás nodos.
- *PageRank*: Mide la relevancia de un nodo en función de la importancia de sus vecinos, ponderando no solo el número de enlaces entrantes, sino también su calidad. Es especialmente útil en redes dirigidas o de flujo de información.

Métricas asociadas a componentes o enlaces Para caracterizar la estructura de la red o las propiedades de los enlaces, existen las siguientes métricas:

- *Densidad de la red (Density)*: Relación entre el número de conexiones existentes y el número máximo posible de conexiones. Indica qué tan completa o conectada está la red.
- *Coeficiente de clustering (Clustering Coefficient)*: cuantifica qué tanto está de agrupado (o interconectado) un nodo con sus vecinos. En la práctica, mide la tendencia de los nodos a formar triángulos, es decir, grupos de tres nodos mutuamente conectados.
- *Coeficiente de transitividad (Transitivity)*: Proporción de triángulos cerrados respecto a todas las tríadas posibles en la red, evaluando el grado global de cohesión local.
- *Coeficiente de asortatividad (Assortativity Coefficient)*: Evalúa la tendencia de los nodos a conectarse con otros nodos que comparten características similares, como el grado.
- *Modularidad (Modularity)*: Mide la fortaleza de la división de la red en comunidades o clústeres, comparando la densidad de enlaces dentro de las comunidades con la densidad de enlaces esperada al azar.

6.3 Marco de trabajo técnico

El Marco de trabajo técnico que se presenta a continuación constituye la base tecnológica que habilita el desarrollo de aplicaciones orientadas al análisis político-legislativo, mediante la integración de datos heterogéneos provenientes principalmente de texto plano con datos abiertos enlazados. Esta arquitectura permite transformar información dispersa y no estructurada en recursos de datos procesables y vinculados semánticamente, posibilitando la extracción de conocimiento sobre fenómenos legislativos y políticos. En este contexto, se describen los componentes tecnológicos involucrados, junto con una referencia al aspecto de la explicabilidad algorítmica asociada a sistemas del ámbito político-legislativo, lo cual constituye un elemento clave en procesos de toma de decisiones informadas y transparentes como la en estudio.

6.3.1 Componentes del marco de trabajo

El marco de trabajo se divide en cinco bloques principales, cada uno con características específicas. La figura 6.8 es una representación esquemática de los bloques, donde es posible ubicar cada componente dentro de un flujo genérico que permite describir las etapas por las que pasan los datos crudos hasta que se transforman en información valiosa de análisis. A continuación, se describe cada bloque desde un punto de vista funcional, en el contexto de su uso en el análisis político-legislativo.

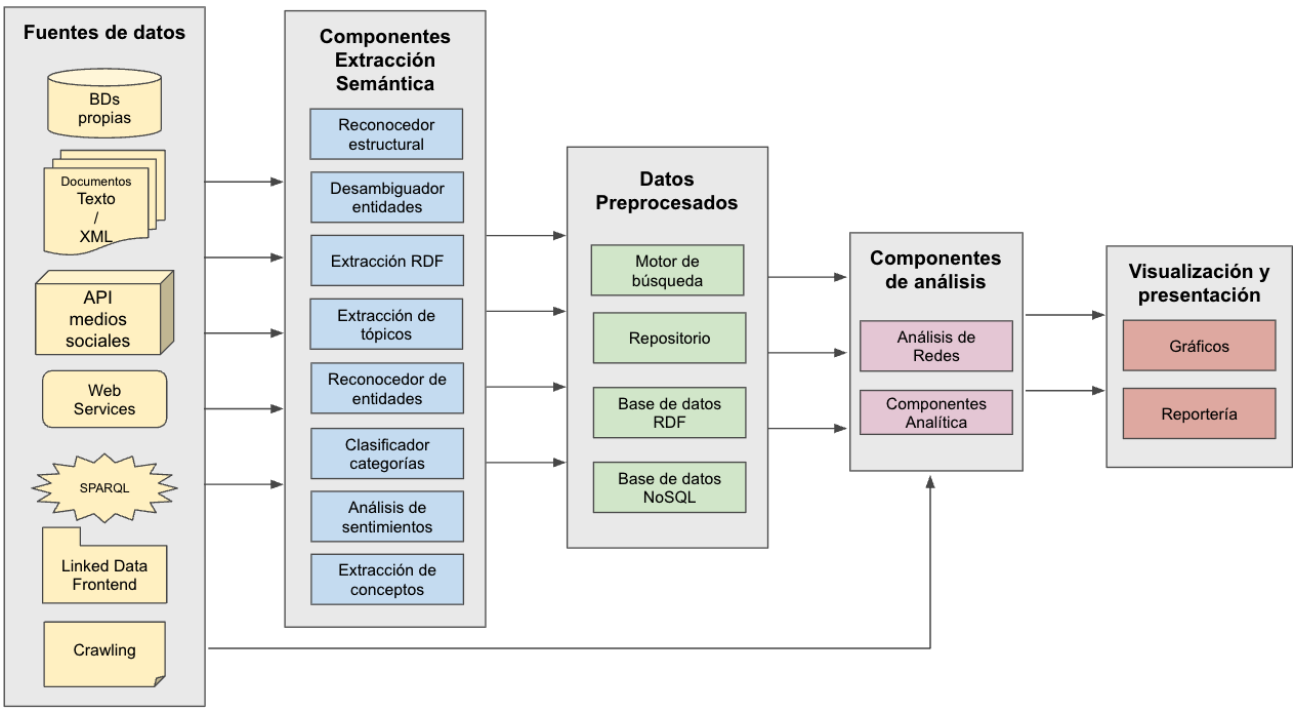

Figura 6.8: Componentes del marco de trabajo de las tecnologías Semánticas

Fuentes de datos

Este bloque actúa como punto de partida del marco de trabajo, y es donde se realiza captura y consolidación de información tanto no estructurada (textual), semiestructurada (xml) como

estructurada, proveniente de diversas fuentes vinculadas al ámbito político y legislativo. En esta etapa, es indispensable asegurar que los datos a recuperar sean representativos, actualizados y suficientemente ricos en contenido para permitir un análisis significativo.

La fuentes de datos, que pueden incluir debates parlamentarios, noticias, redes sociales y otros, permitirán construir una base informacional robusta que sirve como insumo para los procesos posteriores de extracción semántica y análisis, agregando contenido factual (texto, metadatos y otros datos estructurados) que luego será normalizado y enriquecido.

Dentro de los elementos pertenecientes a este bloque, se encuentran bases de datos propias de negocio, documentos de texto y XML, APIs a medios sociales, Servicios Web que proveen acceso a datos, endpoints de datos abiertos enlazados, datos en RDF accedidos mediante negociación de contenido y crawlers que permiten recuperar datos desde páginas, portales y sistemas web.

Componentes de extracción semántica

Una vez recolectados, los datos textuales ingresan a la capa de componentes de extracción semántica, la cual tiene como objetivo transformar los datos crudos en objetos de análisis que se harán persistentes en la capa de datos preprocesados, los cuales posteriormente podrán ser integrados en distintos almacenes de datos y ser utilizados para consulta y/o razonamiento.

Dentro de las tareas comunes que se realizan en los componentes de esta capa se encuentra la preparación del texto, donde se ejecutan tareas como normalización, eliminación de inconsistencias o corrección de errores de codificación. Con el texto preparado, los distintos componentes de esta capa aplican su lógica específica para realizar operaciones de procesamiento de lenguaje natural, que permiten ejecutar tareas humanas que han sido automatizadas, las que se orientan a enriquecer el contenido con metadatos de distinto tipo.

Algunas de las tareas más frecuentes en los componentes de este bloque son el marcaje estructural, que permite la identificación e incorporación de marcas en secciones estructurales dentro del documento; el marcaje de entidades nombradas, que permite la identificación de entidades (como personas, organizaciones y otros tipos) a las que se hace mención en el texto; la desambiguación de entidades, que permite la identificación y enlace entre una entidad en el texto y un identificador de la entidad existente en una base de datos específica; la identificación de análisis de sentimiento, que permite establecer la carga emotiva de un texto o de una parte del texto; la clasificación temática del texto, que permite clasificar el texto en una o más categorías; la extracción de RDF, proceso que por lo general se realiza deste archivos XML o JSON entre muchos otros.

Datos preprocesados

El tercer bloque del marco de trabajo corresponde a la capa encargada de almacenar los datos ya enriquecidos y disponibilizarlos para su uso posterior. A diferencia de la capa de origen, esta cumple una función técnica clave al garantizar la consulta eficiente y flexible de los datos, sirviendo como sustento tanto para los componentes de análisis como para las herramientas de visualización.

Desde el punto de vista funcional, esta capa proporciona mecanismos avanzados de acceso y explotación de la información, como búsquedas textuales optimizadas, consultas facetadas, acceso a grafos semánticos mediante SPARQL, Cypher o GraphQL, razonamiento sobre datos RDF, acceso a objetos vía bases NoSQL, y representación conforme a estándares bibliotecológicos como Dublin Core, METS o Schema.org.

En términos de interoperabilidad, esta capa opera mediante servicios web HTTP, destacando el uso de endpoints SPARQL, GraphQL-LD y otros mecanismos basados en HTTP soportados por motores como Apache SolR, Elasticsearch, OpenSearch, MongoDB, Blazegraph y Neo4J, entre otros.

Componentes de análisis

En esta capa de componentes se encuentran los componentes que permiten procesar los datos semánticamente enriquecidos, y generar cálculos para representaciones analíticas, que posteriormente van a ser visualizadas en gráficos o reportes.

Estos componentes de análisis resuelven tareas específicas con base en los datos, tales como el cálculo de la polarización o alineamiento político, la probabilidad de aprobación de un proyecto de ley, la relevancia de ciertos temas para determinados individuos o grupos u otras variadas, el cálculo de métricas sobre objetos de análisis dispuestos en grafos y múltiples otras aplicaciones. La idea principal de este bloque es enmarcar los componentes que facilitan la identificación de tendencias y herramientas que pueden ser utilizadas como soporte para la toma de decisiones.

Visualización y presentación

Los componentes de este bloque tienen como objetivo la visualización de la información generada en los componentes de análisis. La idea es relevar la importancia que tiene el facilitar el acceso y la comprensión de los datos y resultados analíticos a distintos tipos de usuarios, desde expertos en análisis legislativo hasta ciudadanos interesados, actuando como puente entre el procesamiento computacional de los datos y su interpretación humana, promoviendo la transparencia y el uso efectivo de la información.

Flujos de información e interoperabilidad

El marco de trabajo basado en tecnologías semánticas permite el diseño y la implementación de flujos de información desacoplados e interoperables. Estas características son posibles gracias a una arquitectura de componentes modulares que se comunican mediante estándares abiertos como servicios HTTP y que intercambian datos a través de URIs desreferenciables, SPARQL y negociación de contenido.

El diseño de los flujos se concibe como una secuencia de pasos que comienza en las fuentes de datos: estructuradas, semiestructuradas o no estructuradas, y avanza hacia componentes de almacenamiento, procesamiento, inferencia, análisis y visualización. Cada uno de estos pasos se orquesta mediante consumo directo a la capa anterior, lo que permite su composición, sustitución o evolución sin necesidad de alterar el resto del sistema.

Uno de los aspectos fundamentales que permite esta flexibilidad es el uso de URIs como identificadores de recursos. Este enfoque permite resolver problemas tradicionales de heterogeneidad e integración de datos al posibilitar el enlace de entidades provenientes de múltiples fuentes bajo un mismo esquema referencial. Así, datos de diferentes orígenes pueden ser conectados semánticamente, favoreciendo la construcción de flujos de información distribuidos, pero compatibles.

Al mismo tiempo, la interoperabilidad se ve reforzada mediante el uso de endpoints SPARQL como interfaz de consulta unificada, los cuales permiten tanto el acceso directo a los datos como la federación con otros repositorios distribuidos, permitiendo consultas compuestas sobre datos de distintos organismos o dominios. Complementariamente, la negociación de contenido permite

servir representaciones de recursos en distintos formatos (RDF/XML, Turtle, JSON-LD, RDFa, etc.) dependiendo del consumidor.

Adicionalmente, el uso de componentes de caché juega un rol crucial para asegurar el rendimiento del sistema en escenarios donde los datos son altamente dinámicos o las consultas son costosas computacionalmente. Estos componentes permiten almacenar resultados intermedios de procesamiento, evitando así sobrecargar fuentes remotas o repetir operaciones complejas.

6.3.2 Explicabilidad algorítmica en el ámbito político-legislativo

Este marco de trabajo no solo permite especificar aplicaciones y procesos de automatización de tareas analíticas complejas, sino que también permite garantizar la trazabilidad y reproducibilidad de los resultados obtenidos en los casos de uso. En ese contexto, el concepto de explicabilidad de sistemas y algoritmos asociados al ámbito político es fundamental por razones técnicas, éticas y de transparencia, sobre todo cuando alguna de las ramas de la IA interviene. En política esto es particularmente importante debido a su naturaleza altamente sensible, donde los resultados algorítmicos no solo afectan decisiones individuales, sino que también pueden generar repercusiones institucionales y sociales de gran escala. De hecho, trabajos en este campo proponen la creación de marcos legales y normativos que exijan transparencia y justificación en las decisiones automatizadas (el concepto de *derecho a la explicación*), así como la implementación de mecanismos de supervisión y rendición de cuentas[Maclure, 2021, Kim et al., 2024]. A continuación se mencionan algunos de los principales puntos clave que explican la relevancia de la explicabilidad en el ámbito político-legislativo:

- *Sensibilidad del dominio político*: las decisiones políticas generalmente no son neutras, están cargadas de ideologías, intereses y consecuencias. Por lo tanto, cuando un algoritmo produce un resultado, por ejemplo calcular un valor, emitir una recomendación o clasificar algún texto, debe ser posible explicar cómo y por qué se llegó a ese resultado, especialmente cuando estos no son favorables para ciertos actores, ya que la ocultación de los criterios puede generar sospechas fundadas de manipulación, sesgo o injusticia. Por ejemplo, en un caso hipotético donde un análisis sugiera que un parlamentario tiene baja relevancia, que un grupo muestra desalineación, o que una propuesta carece de viabilidad, puede desatar críticas o tensiones institucionales, por lo cual se vuelve indispensable justificar los fundamentos metodológicos de cada resultado.

- *Confianza y transparencia*: la legitimidad de las decisiones políticas está estrechamente asociada a su nivel de transparencia, y si estas decisiones están basadas en información generada a partir de datos procesados, es indispensable contar con los antecedentes que las avalen. Por ejemplo, un sistema de apoyo a la toma de decisiones legislativas basado en algoritmos debe ser auditable y comprensible para que tanto parlamentarios, asesores legislativos y ciudadanos confíen en su uso, permitiendo la validación científica.

- *Control de sesgos*: Si bien durante el uso de IA los sesgos en los algoritmos son un riesgo siempre latente, el describir el modelo arquitectónico de las soluciones basadas en datos, permite mitigar en parte la aparición de estos sesgos debido a que existen puntos de control donde es posible verificar la trazabilidad de los datos. Esto es un tema relevante, ya que cuando esto no es posible, incluso modelos de lenguaje grandes como ChatGPT han demostrado sesgos ideológicos en algunos contextos[Hartmann et al., 2023].

Al mismo tiempo, es importante a la hora de desarrollar algoritmos aplicados a la política, el garantizar que los sistemas no reproduzcan ni profundicen inequidades existentes, ya sea por el uso de fuentes sesgadas, o por manipulación de los mismos algoritmos.

Sin perjuicio de esto, vale la pena destacar que el aspecto de explicabilidad algorítmica, no es por si solo una solución a lo que se denomina *el problema de la caja negra*. Trabajos recientes indican que la relación entre la confianza con los modelos, su precisión y el nivel de explicabilidad, es más compleja de lo que se piensa. No basta con ofrecer explicaciones para que los usuarios confíen en un sistema, ni una mayor precisión garantiza, por sí sola, una mejor aceptación. Por el contrario, ciertos tipos de explicaciones pueden incluso disminuir la confianza si son percibidas como inconsistentes, poco intuitivas o si revelan un razonamiento que entra en conflicto con el juicio humano[Papenmeier et al., 2022, Lipton, 2018]. En entornos altamente críticos como el político-legislativo, donde los resultados que apoyan la toma de decisiones pueden tener consecuencias normativas, sociales o reputacionales, estos aspectos se vuelven aún más sensibles, por lo cual requieren de un diseño explicativo cuidadoso, transparente y validado bajo criterios de utilidad y legitimidad.

Capítulo 7

Fase experimental

7.1 Introducción

Este capítulo presenta los instrumentos desarrollados para llevar a cabo la fase experimental del trabajo de investigación, los cuales se enmarcan en el uso de Tecnologías Semánticas y datos abiertos legislativos. En particular, se utilizarán tres herramientas analíticas: (i) el Panel de Visualización de Temas de Interés Parlamentario, (ii) el Panel de Visualización e Indicadores de Cohesión Política y (iii) el Visualizador de Rol Clave en el Contexto de un Tema Legislativo. Cada uno de estos instrumentos será descrito de forma individual, detallando su propósito específico, la pregunta de investigación que lo fundamenta, su funcionamiento técnico y metodológico, los datos empleados en la fase experimental, el flujo de procesamiento que integra los componentes del marco semántico, así como el los fundamentos conceptuales que orienta su diseño.

Dado que los datos utilizados hacen referencia a parlamentarios identificados con partidos políticos, cuya filiación puede reflejar cargas ideológicas, existe el riesgo de una interpretación inadecuada o un uso fuera de contexto del contenido presentado. Con el objetivo de evitar este tipo de malentendidos o un uso no deseado de la información, durante la exposición de capturas de pantalla correspondientes a los distintos instrumentos, se emplearán nombres ficticios de parlamentarios. Esta decisión busca preservar la neutralidad del análisis y anonimizar a las personas eventualmente expuestas en el documento.

7.2 Ingreso al entorno experimental

El ingreso al entorno experimental se realiza a través de una pantalla de autenticación que permite controlar el acceso al GE. Una vez autenticado correctamente, el usuario es redirigido a la página principal del entorno, donde se presentan los tres instrumentos de evaluación en forma de cuadrículas interactivas. La figura 7.1 muestra una captura de esta pantalla principal, mientras que el Diagrama de Actividades 7.2 ilustra las posibles interacciones que el usuario puede realizar dentro del sistema.

En esta interfaz principal, las preguntas están organizadas en tres conjuntos, cada uno asociado a uno de los instrumentos de evaluación desarrollados en esta investigación. Cada conjunto se presenta como una cuadrícula de celdas de colores diferenciados: una cuadrícula verde para el instrumento de Visualización de Rol Clave en el Contexto de un Tema Legislativo (Instrumento 3), una cuadrícula amarilla correspondiente al Panel de Visualización e Indicadores

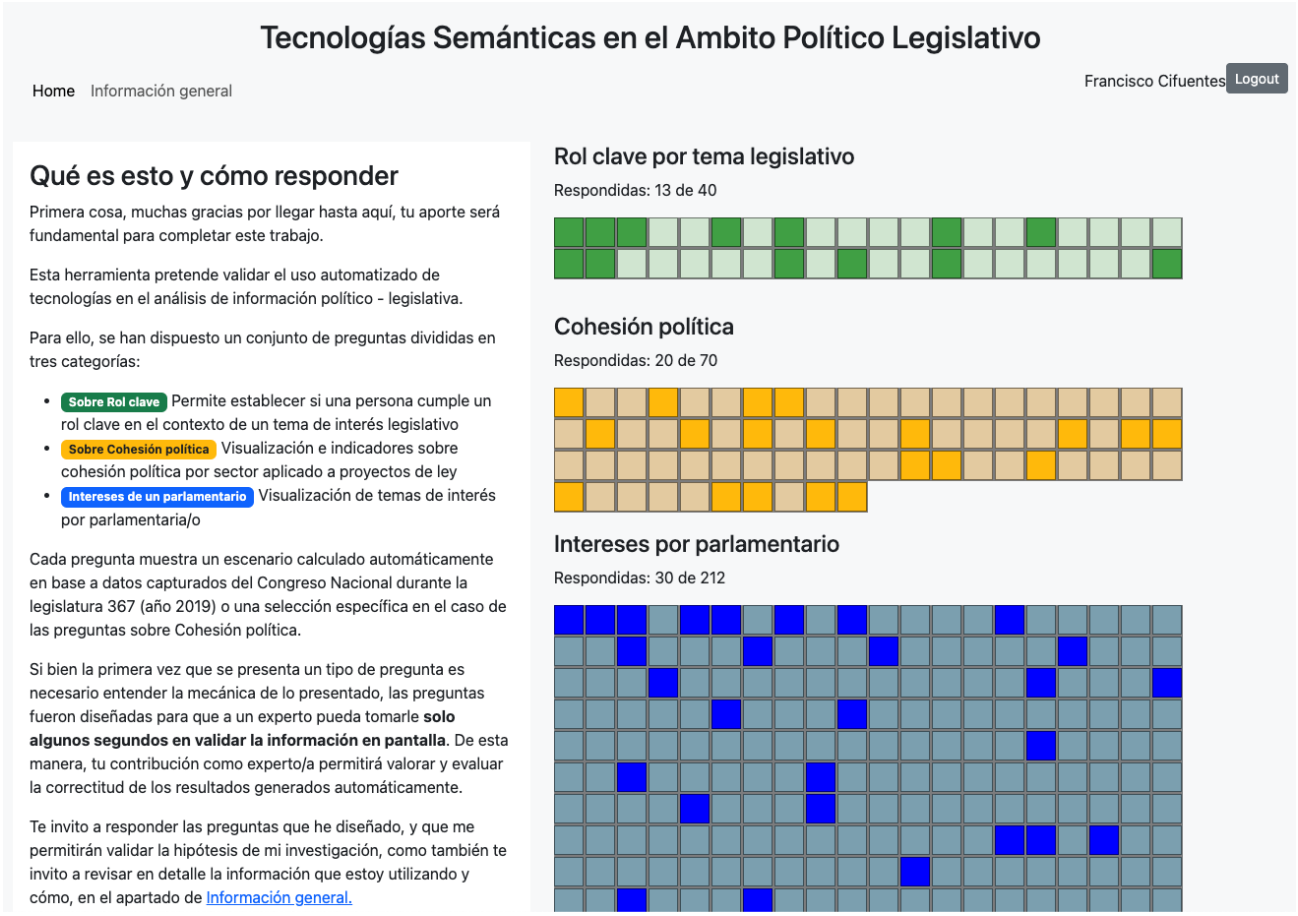

Figura 7.1: Pantalla principal de acceso a las preguntas

de Cohesión Política (Instrumento 2) y una cuadrícula azul asociada al Panel de Visualización de Temas de Interés Parlamentario (Instrumento 1).

Cada celda de estas cuadrículas representa una pregunta. Al posicionar el puntero del ratón sobre una celda, se despliega una descripción breve de la pregunta correspondiente, facilitando así la orientación del usuario antes de acceder a su contenido. Además, el color de fondo de cada celda indica visualmente el estado de la pregunta: las celdas con colores más suaves corresponden a preguntas no respondidas, mientras que aquellas con colores más intensos indican que la pregunta ha sido contestada.

Bajo el título de cada conjunto de preguntas, se muestra dinámicamente el texto *respondidas X de N*, que indica cuántas preguntas han sido respondidas (X) respecto del total disponible para ese instrumento (N), proporcionando así un indicador del avance individual en la fase experimental. A continuación se describe cada uno de los instrumentos desarrollados para la fase de experimentación.

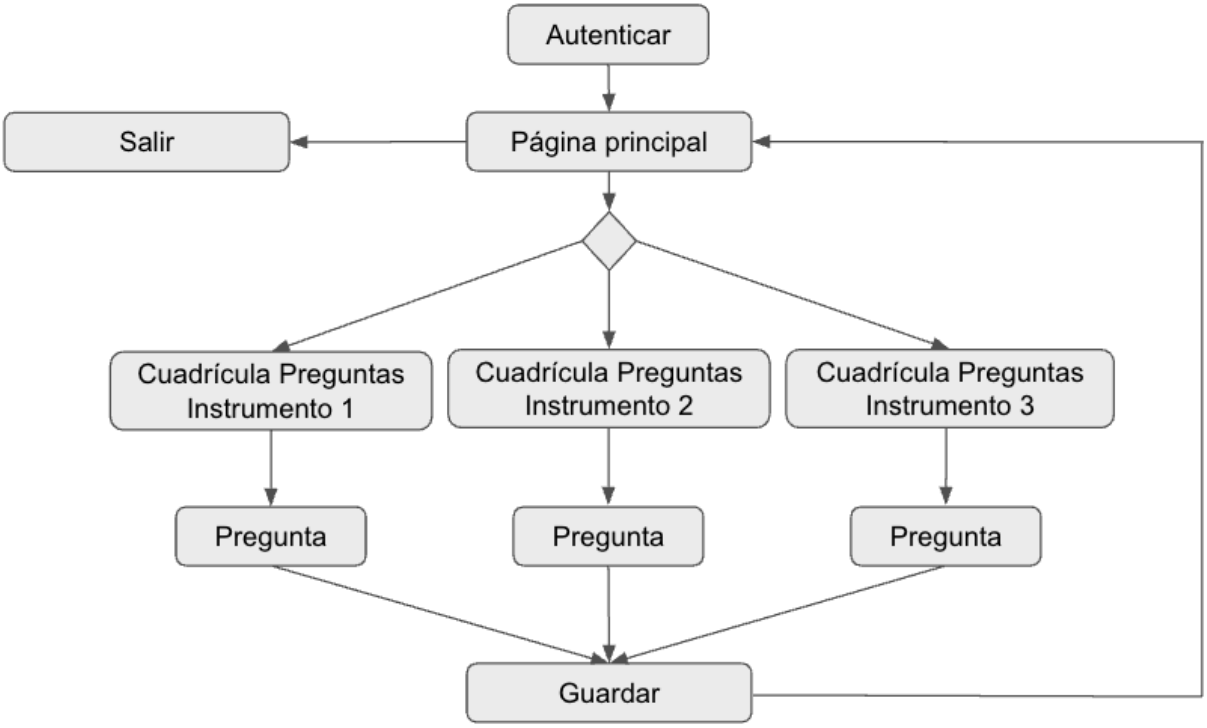

Figura 7.2: Diagrama de posibles flujos dentro de la aplicación

7.3 *Instrumento 1:* Panel de visualización de temas de interés parlamentario

7.3.1 Descripción del instrumento

Esa herramienta presenta al usuario un gráfico con los temas de interés de un parlamentario en un periodo específico, detectados mediante el análisis del texto de sus intervenciones en la sala de sesión, a través de clasificación automática de texto. La figura 7.3 muestra la interfaz de usuario del instrumento desarrollado y presentado al GE. A continuación, la tabla 7.1 describe los principales aspectos del instrumento.

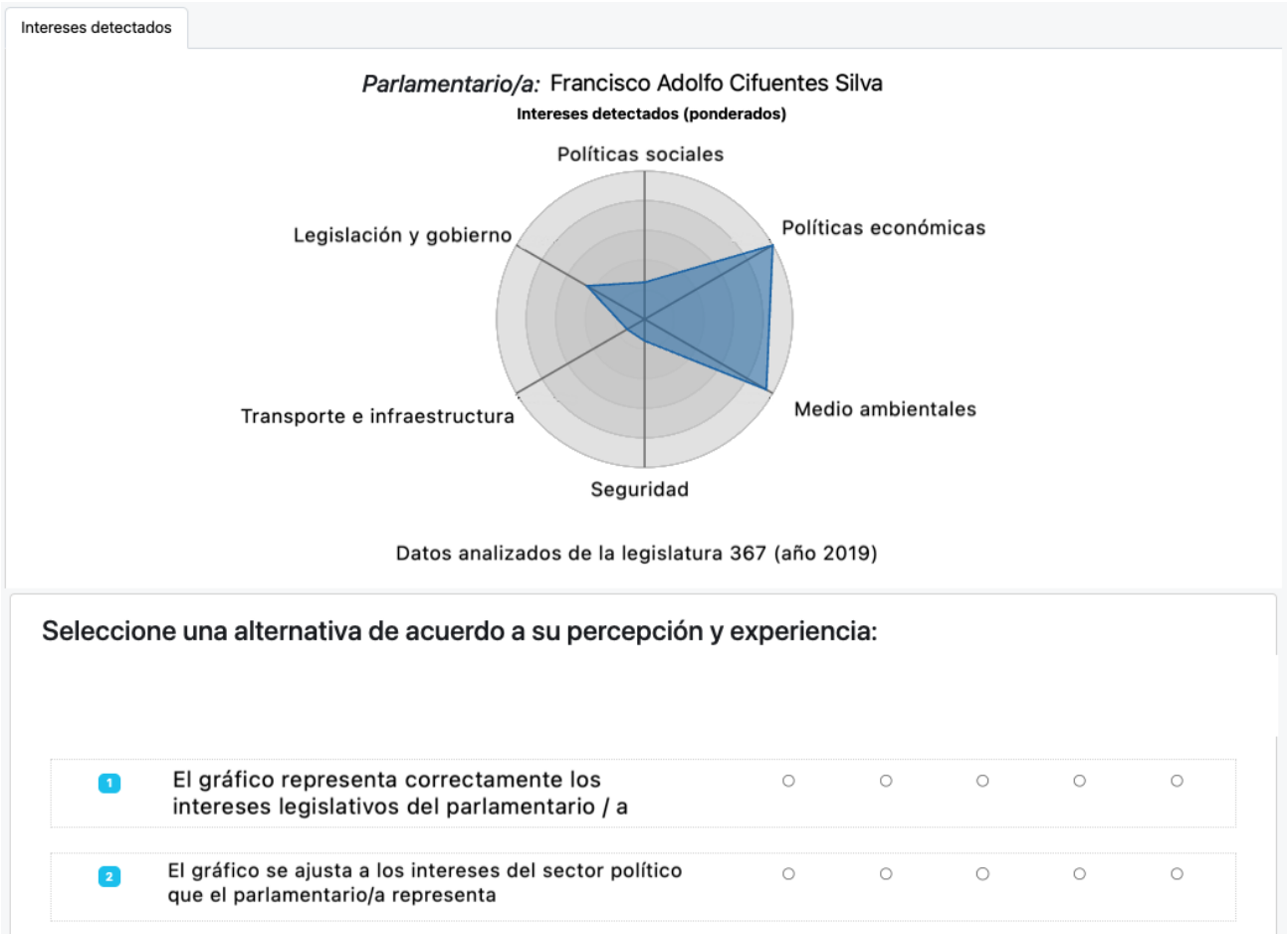

Figura 7.3: Diagrama de radar para identificar los intereses parlamentarios detectados

7.3.2 Flujo de procesamiento

En primer lugar, un servicio de consulta al endpoint SPARQL recupera el conjunto de URIs de los Diarios de Sesión relevantes para el período de estudio. Con esas URIs como referencia, el sistema descarga el texto plano de cada diario y lo somete a un módulo de marcaje estructural, que identifica secciones, encabezados y metadatos internos. Sobre el texto ya estructurado se ejecuta un proceso de NER, seguido inmediatamente por la desambiguación de entidades, con lo cual se garantiza que personas, organizaciones y lugares queden vinculados a identificadores únicos dentro del grafo de conocimiento.

Una vez que el documento ha sido enriquecido semánticamente, el flujo continúa con la extracción de la sección Participación, que contiene las intervenciones de los parlamentarios. A partir de esta sección, se generan triples RDF que representan tanto el contenido textual como los metadatos de cada participación, los cuales se almacenan en una base de datos RDF. Luego, el texto de cada intervención es procesado por un clasificador supervisado, que lo etiqueta con una o más categorías temáticas definidas en la ontología de intereses legislativos. Estas etiquetas se persisten, y el procedimiento se repite para todas las participaciones recuperadas.

Finalmente, un componente de analítica consolida los resultados por parlamentario y genera la visualización: un gráfico de telaraña que muestra la distribución relativa de los temas identificados. De este modo, el usuario puede inspeccionar, en una sola mirada, los focos de interés legislativo de cada parlamentario durante el período analizado. La figura 7.4 representa los

Objetivo del instrumento	Representar los temas de interés más relevantes para un parlamentario, que han sido identificados mediante la aplicación del marco de trabajo, y permitir una evaluación por criterio experto.
Objetivo de investigación	Validar el mecanismo tecnológico de detección de intereses mediante la identificación no supervisada de temas de interés presentes en el texto.
Pregunta de investigación	**RQ1**: ¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuáles son los temas de mayor relevancia para un representante?
Conjunto de datos de prueba	19.990 Intervenciones asociadas a 263 Diarios de Sesión de la Cámara de Diputados (155) y del Senado (108) de Chile correspondientes a la legislatura 367 (2019-03-11 al 2020-03-10) en formato texto.
Tecnologías Semánticas asociadas al instrumento	1. Reconocimiento de entidades nombradas 2. Marcaje estructural del texto 3. Clasificación de texto 4. Desambiguación de entidades 5. Ontologías y taxonomías de términos 6. Visualización de datos 7. Marcaje estructural del texto
Preguntas en el instrumento	1. El gráfico representa correctamente los intereses legislativos del parlamentario/a 2. El gráfico se ajusta a los intereses del sector político que el parlamentario/a representa
Posibles usos	• Identificar información de asesoría a los parlamentarios acorde a sus intereses para ofrecer asesoría parlamentaria de forma proactiva. • Identificar parlamentarios relacionados a temas específicos, para focalizar esfuerzos de influencia en favor de determinados intereses (*lobby*).

Tabla 7.1: Tabla resumen de descripción del instrumento 1

elementos activos del marco de trabajo durante el procesamiento de los datos del instrumento 1.

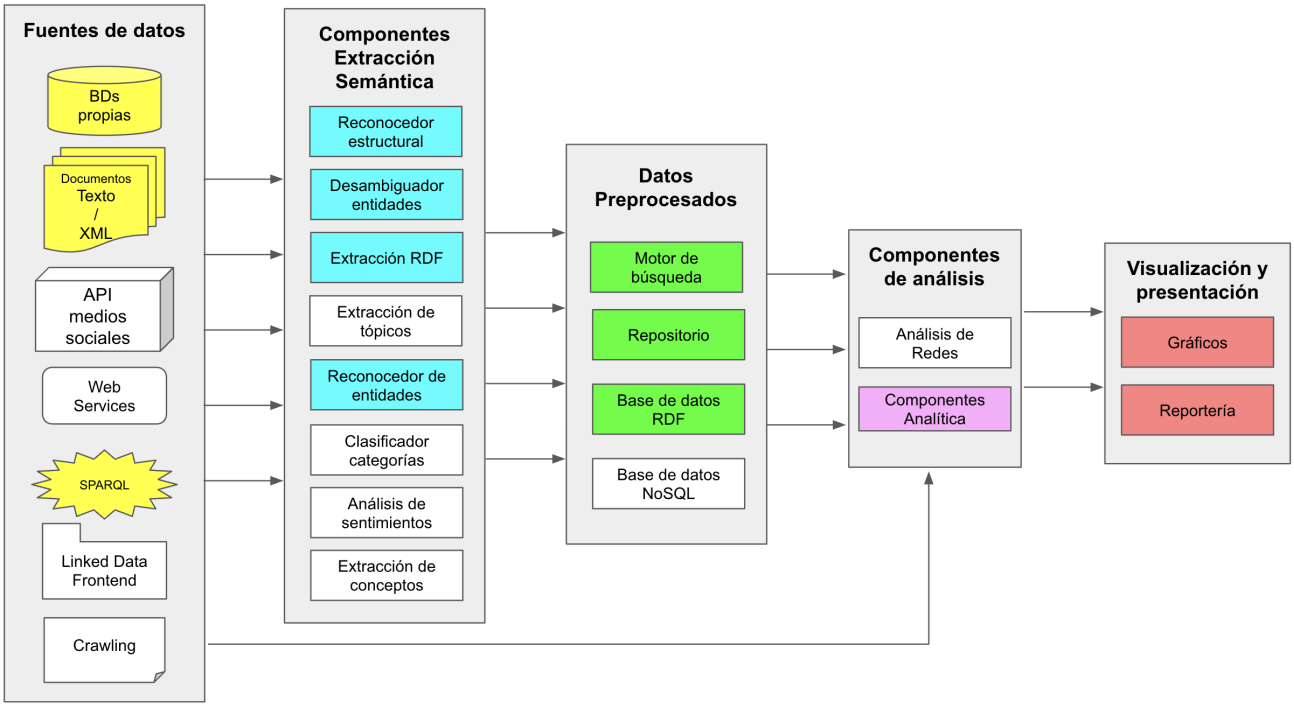

Figura 7.4: Componentes activos del marco de trabajo para desarrollo del instrumento 1

7.3.3 Fundamentos de diseño

Para diseñar este instrumento se partió de la necesidad de caracterizar el perfil temático de cada parlamentario en función del contenido de sus intervenciones legislativas. Con este fin, se propuso identificar los documentos asociados a cada parlamentario y clasificarlos según una taxonomía de temas legislativos previamente definida. Este enfoque permite, al aplicar el mismo proceso a todos los parlamentarios, distinguirlos entre sí según sus áreas de interés predominantes.

Para llevar a cabo esto, se definió la utilización de una taxonomía de temas legislativos, la implementación de un clasificador que permita la clasificación de las intervenciones en las categorías antes definidas y la visualización agregada de los intereses legislativos por parlamentario.

Taxonomía de temas de interés legislativo

Para establecer el conjunto de categorías temáticas legislativas, se tomó como punto de partida la lista de comisiones parlamentarias permanentes existentes en ambas cámaras del Congreso Nacional de Chile. A partir de la revisión de sus ámbitos de acción, se elaboró una lista de conceptos asociados a políticas legislativas, los cuales sirvieron como base para definir una jerarquía temática de dos niveles, presentada en la figura 7.5.

El proceso de construcción de esta jerarquía comenzó con un análisis comparativo entre las comisiones permanentes del Senado y de la Cámara de Diputadas y Diputados, generándose un pareo temático entre ambas cámaras, el cual se encuentra disponible en la figura A.1 del anexo A. A partir de este pareo, se derivó una lista de temas legislativos específicos, los que posteriormente fueron organizados en seis grandes categorías temáticas de nivel superior.

Estas seis macrocategorías constituyen el núcleo de la taxonomía utilizada e integran la base sobre la cual se implementó un clasificador multiclase, orientado a la clasificación automática

de intervenciones parlamentarias y su posterior visualización agregada. Cabe destacar que la definición de dos niveles de agregación en la categorización de los temas de interés legislativo responde en parte, a la necesidad de presentar una visualización legible y manejable para el usuario. Esto permite mostrar en pantalla un conjunto acotado de temas en lugar del total, con la posibilidad de desplegar el detalle si se requiere. Lo anterior cobra especial relevancia considerando que la jerarquía completa abarca 39 temas, una cantidad que dificultaría su visualización simultánea sin comprometer la claridad y utilidad de la información presentada.

$$
\text{Temas de Interés Legislativo} = \begin{cases}
\text{Temas Sociales} = \begin{cases} \text{Educación, Salud, Vivienda, Protección Social,} \\ \text{Igualdad de Género, Familia, infancia y adolescencia,} \\ \text{Adulto mayor y discapacidad, Cultura, deportes y recreación,} \\ \text{Desarrollo social y pobreza} \end{cases} \\
\text{Temas Económicos} = \begin{cases} \text{Desarrollo Económico, Fiscalidad y Tributación,} \\ \text{Empleo y Mercado Laboral, Comercio Exterior,} \\ \text{Innovación y Tecnología, Agricultura,} \\ \text{Minería, Pesca} \end{cases} \\
\text{Temas Medioambientales} = \begin{cases} \text{Medio Ambiente, Cambio Climático,} \\ \text{Recursos Hídricos, Contaminación y Residuos,} \\ \text{Energías Renovables} \end{cases} \\
\text{Temas de Seguridad} = \begin{cases} \text{Seguridad Ciudadana, Defensa,} \\ \text{Seguridad Nacional, Emergencias y desastres naturales,} \\ \text{Ciberseguridad} \end{cases} \\
\text{Temas de Infraestructura y Transporte} = \begin{cases} \text{Transporte Público y telecomunicaciones,} \\ \text{Infraestructura Urbana,} \\ \text{Desarrollo de Redes de Comunicación,} \\ \text{Energía y Suministro} \end{cases} \\
\text{Temas sobre Políticas, Legislación y Gobierno} = \begin{cases} \text{Regionalización y descentralización, Gobierno,} \\ \text{Relaciones exteriores, Justicia,} \\ \text{Transparencia, Derechos Humanos,} \\ \text{Congreso, Zonas extremas y antártica} \end{cases}
\end{cases}
$$

Figura 7.5: Jerarquía de temas de interés legislativo

Clasificación de intervenciones

Para identificar de forma automática la categoría a la que pertenece cada intervención parlamentaria, se implementó un clasificador multiclase acorde al primer nivel de la Taxonomía de temas de interés legislativo. El modelo fue entrenado sobre un corpus balanceado de 5.400 intervenciones del Congreso Nacional de Chile, manualmente etiquetadas en las seis categorías antes descritas mediante la herramienta OpenRefine.

Se exploraron cuatro algoritmos de clasificación: RidgeClassifier, MLPClassifier, LogisticRegression y LinearSVC, utilizando el framework scikit-learn. El modelo final incorpora car-

acterísticas extraídas mediante TF-IDF y vectores de embeddings generados con el modelo multilingüe MiniLM, tokenizados con el modelo BETO en español. La validación se realizó mediante 10-fold cross-validation, empleando métricas estándar (accuracy, precision, recall y F1-score) para evaluar el rendimiento.

Entre los modelos probados, RidgeClassifier demostró el mejor desempeño, alcanzando un accuracy promedio de 0.824. Este resultado es significativamente superior al de un clasificador aleatorio (aproximadamente 0.16), especialmente considerando la complejidad del lenguaje parlamentario. Las curvas ROC mostraron valores AUC altos para todas las categorías (entre 0.92 y 0.99), con un promedio de 0.96. El análisis Precision-Recall arrojó un micro-average de 0.89, también muy por encima del baseline.

Durante el entrenamiento, se optó por balancear el número de documentos por categoría para evitar sesgos, anticipando que el desequilibrio natural se reflejaría posteriormente al aplicar el modelo sobre la totalidad del corpus, hipótesis que fue confirmada empíricamente.

Dado que el clasificador resultante se utiliza en un entorno de análisis agregado, donde lo relevante es captar las tendencias temáticas predominantes más que la exactitud individual de cada instancia clasificada, los niveles de fiabilidad del clasificador se consideran suficientes.

Un análisis detallado del proceso de experimentación para el desarrollo del clasificador puede ser consultado en el anexo B.

Respecto a la aplicación del clasificador sobre los datos, una vez generado se aplicó sobre el total de documentos. El gráfico en la figura 7.6 muestra la distribución en porcentaje y número de documentos sobre cada una de las seis categorías del primer nivel de la taxonomía. Vale decir que con base en estos datos, se confirma la hipótesis planteada durante la fase de etiquetado manual, asociada a una percepción de un mayor número de documentos asociados a políticas sociales.

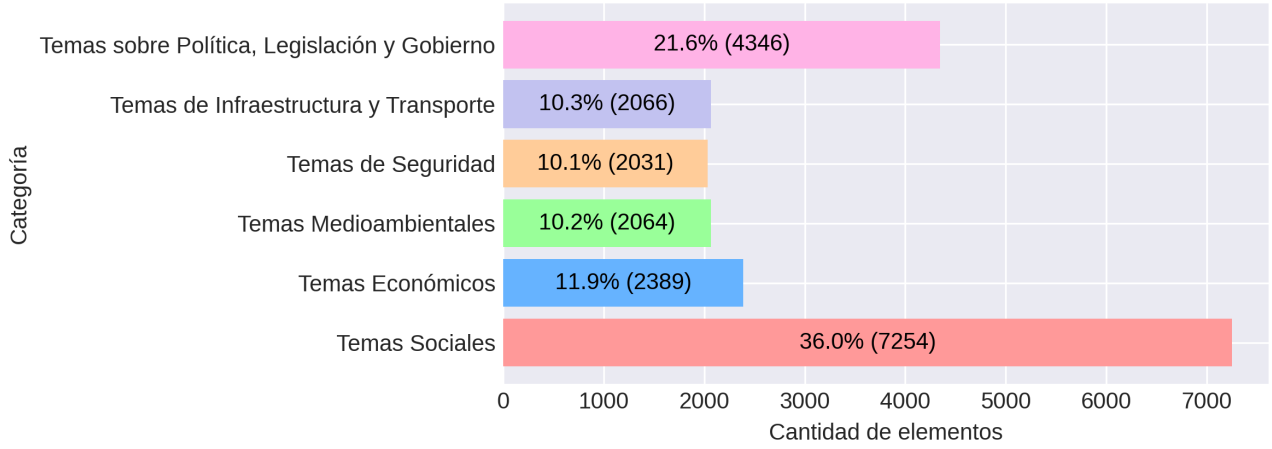

Figura 7.6: Total de intervenciones clasificadas en cada una de las 6 categorías

Visualización de intereses parlamentarios

La visualización de los intereses legislativos de cada parlamentario se realiza mediante un gráfico de radar (o telaraña), el cual permite representar de manera comparativa la distribución temática de sus intervenciones. Para asegurar una interpretación equilibrada de los datos, este gráfico se construye utilizando valores ajustados mediante un proceso de normalización que corrige sesgos inherentes a la agenda legislativa y a la coyuntura política.

En particular, existe un sesgo relevante asociado al alto volumen de documentos vinculados a temáticas de políticas sociales, producto de su centralidad en la agenda del Ejecutivo. Esta sobrerrepresentación puede ocultar intereses legislativos específicos de los parlamentarios y dificultar el análisis comparativo entre distintas áreas temáticas.

El proceso de normalización se basa en el cálculo de un valor de relevancia a partir de un ponderador por categoría temática, determinado a partir del total de intervenciones en cada categoría acumuladas por todos los parlamentarios del conjunto analizado. Este ponderador permite ajustar los valores individuales por categoría, de modo que la visualización refleje de manera más justa y proporcional los intereses temáticos reales de cada representante. El detalle sobre el cálculo del valor normalizado de relevancia se describe en el anexo D.

La figura 7.7 ilustra este enfoque: a la izquierda, se presenta el gráfico de radar con los valores ponderados por categoría; a la derecha, un gráfico de barras que muestra el número bruto de documentos asociados. La diferencia entre ambas representaciones evidencia el impacto del proceso de normalización, que permite corregir distorsiones impuestas por la agenda legislativa dominante y resaltar intereses menos visibles pero igualmente relevantes en la actividad parlamentaria.

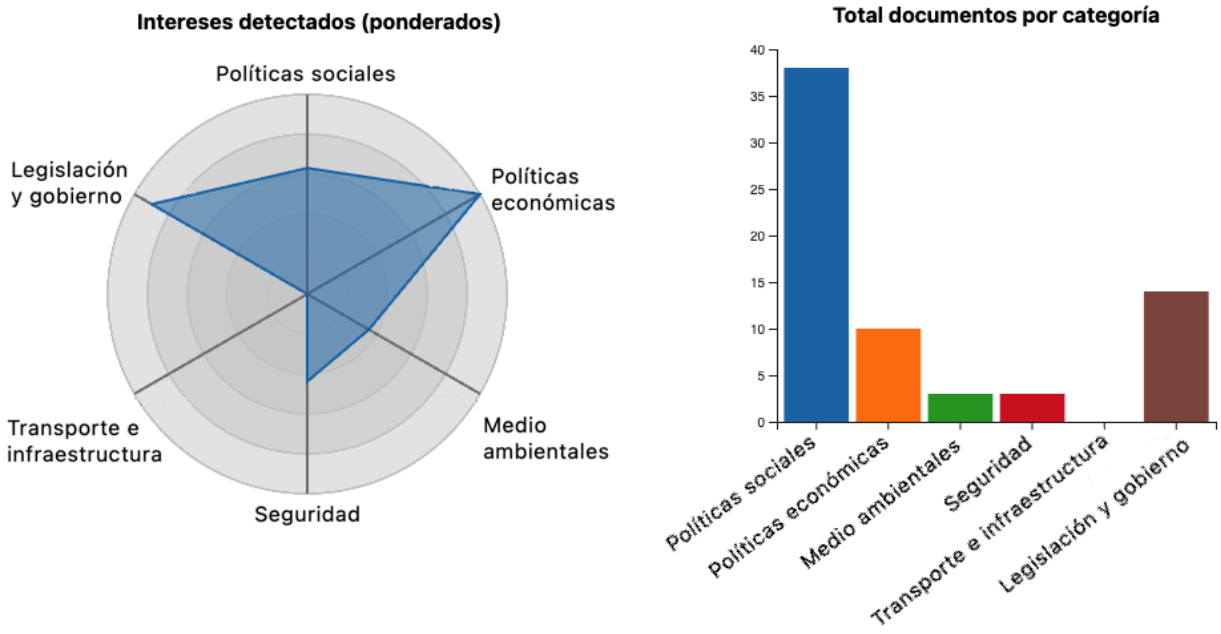

Figura 7.7: Visualización de intereses legislativos

7.4 *Instrumento 2:* Panel de visualización e indicadores sobre cohesión política

7.4.1 Descripción del instrumento

Este instrumento permite visualizar de forma gráfica la cohesión política entre parlamentarios a partir de sus votaciones en proyectos de ley. Mediante una representación basada en grafos, el panel muestra cómo se agrupan o distancian las posturas de los distintos parlamentarios en relación con cada votación, y al mismo tiempo dos gráficos de tipo *media dona o velocímetro*, muestran el porcentaje de polarización y alineamiento calculado en la votación, lo cual refleja la el tipo de cohesión política implícito en el proyecto de ley. La figura 7.8 muestra la interfaz de usuario del instrumento desarrollado y presentado al GE. A continuación, la tabla 7.2 describe los principales aspectos del instrumento.

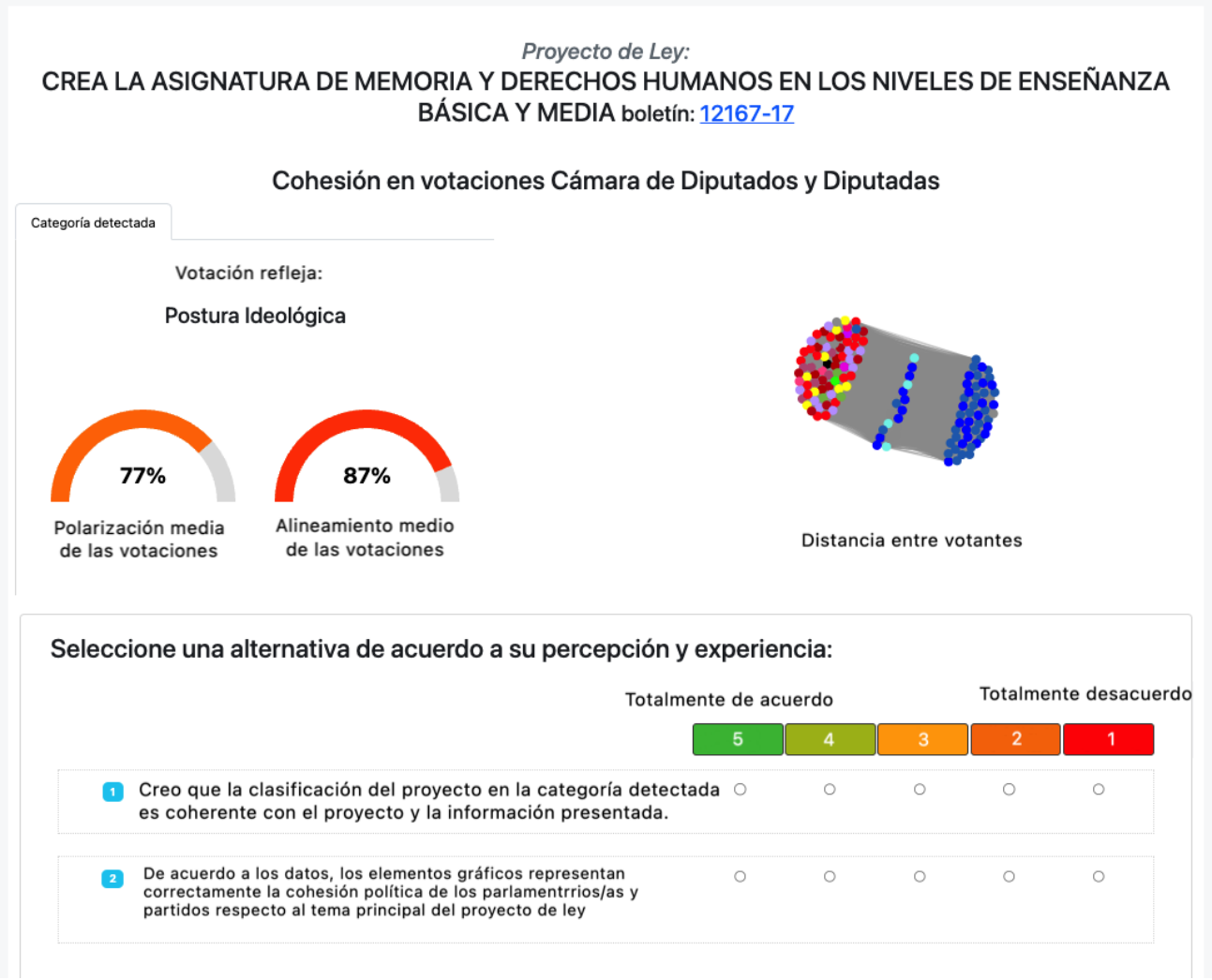

Figura 7.8: Visualización e indicadores sobre cohesión política

Objetivo del instrumento	Representar los niveles de cohesión en los distintos temas y grupos de parlamentarios representados por los proyectos de ley en votación y permitir una evaluación por criterio experto.
Objetivo de investigación	Validar el mecanismo tecnológico que establece el nivel de cohesión de forma cuantitativa entre los distintos temas representados por proyectos de ley y grupos de parlamentarios.
Pregunta de investigación	**RQ2**: ¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuál es el nivel de cohesión política de un grupo frente a un tema particular ?
Conjunto de datos de prueba	70 Proyectos de Ley y sus votaciones provenientes de la base de datos abiertos de BCN, seleccionados de forma equitativa acorde a las cuatro categorías definidas.
Tecnologías Semánticas asociadas al instrumento	1. Servicios Web 2. Componentes Analítica 3. Reconocimiento de entidades nombradas 4. Marcaje estructural del texto 5. Desambiguación de entidades 6. Visualización de datos 7. Marcaje estructural del texto
Preguntas en el instrumento	1. Creo que la clasificación del proyecto en la categoría detectada es coherente con el proyecto y la información presentada. 2. De acuerdo a los datos, los elementos gráficos representan correctamente la cohesión política de los parlamentarios/as y partidos respecto al tema principal del proyecto de ley
Posibles usos	• Monitorear la disciplina partidaria por parte de parlamentarios, permitiendo a las directivas de los partidos ver cuán alineados están sus miembros y detectando eventuales disidencias antes de una votación clave. • Detectar de forma temprana quiebres internos al observar caídas en el índice de alineamiento, lo cual es información clave para adoptar medidas de contención. • Identificar formación de nuevas coaliciones, al identificar qué parlamentarios de otros partidos votan de forma similar, facilitando acuerdos legislativos. • Planificar lobby focalizando esfuerzos en parlamentarios con posturas cercanas para maximizar recursos y tiempo.

Tabla 7.2: Tabla resumen de descripción del instrumento 2

7.4.2 Flujo de procesamiento

Para iniciar este flujo, es posible partir desde distintas etapas:

1. Al igual que en el Instrumento 1, si se dispone de textos de diarios de sesiones, se pueden aplicar sobre ellos los componentes de marcaje estructural, reconocimiento de entidades y desambiguación (de personas y proyectos de ley). A continuación, se realiza la identificación y extracción de votaciones sobre proyectos de ley, transformando los datos desde XML a RDF. Aunque este proceso puede requerir control de calidad humano, es técnicamente posible ejecutarlo de forma completamente automatizada. Como resultado, se obtiene un conjunto de votos individualizados por parlamentario, correspondientes a una votación específica de un proyecto de ley. Se generarán tantos conjuntos como votaciones existan en el diario de sesiones procesado.

2. Si se dispone de datos de votaciones provenientes de Servicios Web o de bases de datos abiertas en RDF, estos pueden ser procesados directamente e integrados a una base de datos local para su posterior análisis.

Una vez que los datos de votaciones se encuentran almacenados en la base de datos RDF, se aplican los componentes de análisis encargados de calcular los índices de Polarización y Alineamiento, a partir de los cuales se determina su categoría analítica. Paralelamente, se calculan las distancias de votación entre los parlamentarios, considerando la suma de las distancias entre todos ellos por cada votación (donde se asigna 1 si votan distinto y 0 si votan igual), esto para el conjunto de votaciones disponibles.

Ambos conjuntos de resultados se visualizan en dos formatos: primero, a través de gráficos que permiten evaluar los niveles de polarización y alineamiento detectados; y segundo, mediante un grafo de fuerzas que representa la cohesión entre parlamentarios en torno a la votación de un proyecto de ley.

La figura 7.9 muestra los componentes activos del marco de trabajo durante el procesamiento de los datos correspondientes al Instrumento 1.

7.4.3 Fundamentos de diseño

Para el diseño de este instrumento se partió del supuesto de que las votaciones legislativas sobre proyectos de ley constituyen una expresión fidedigna y categórica de la cohesión entre los parlamentarios en términos de opinión política. En efecto, el voto legislativo representa la manifestación final de la voluntad del legislador y, al mismo tiempo, la materialización de la opinión de sus representados.

Grafo de fuerzas para identificar cohesión

En este contexto, cada punto del grafo representa a un parlamentario, cuyo color identifica el partido político al que pertenece. La escala cromática se construye a partir del índice de tendencia política de cada partido, codificando los extremos ideológicos desde el rojo (izquierda, valor -1) hasta el azul (derecha, valor 1). Los colores amarillo y verde se asignan a nuevas fuerzas políticas emergentes. La figura 7.10 presenta una representación con datos definidos con base en observación (criterio del autor) de los partidos políticos asociados ordenados por este índice de tendencia política. Las conexiones entre los nodos reflejan la similitud en sus patrones de

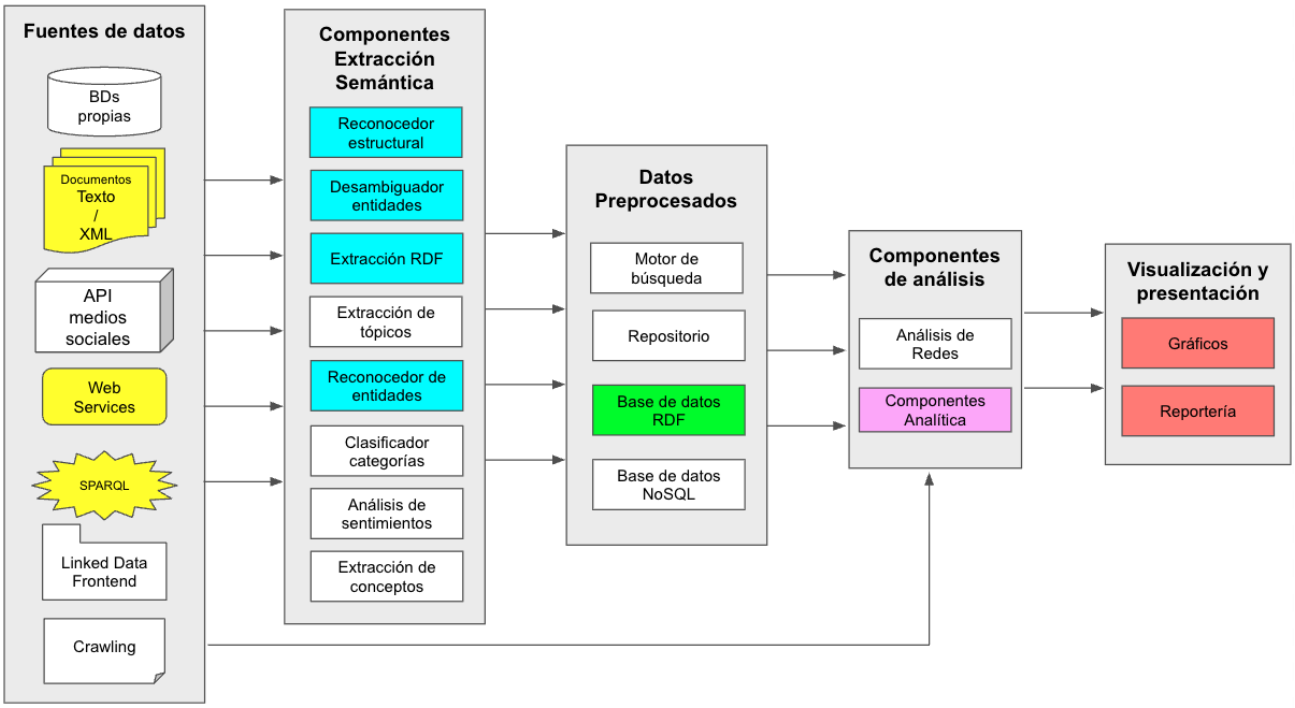

Figura 7.9: Componentes activos del marco de trabajo para desarrollo del instrumento 2

votación: una distancia de 0 implica coincidencia total, mientras que una distancia de 1 indica desacuerdo absoluto.

La cercanía entre puntos en el gráfico señala el grado de alineamiento en las votaciones:

- Puntos próximos denotan alta coincidencia en las decisiones y posturas legislativas similares.
- Puntos distantes reflejan diferencias significativas en las posiciones adoptadas.

A modo de ejemplo, la figura 7.11 muestra los distintos órdenes de conformaciones de parlamentarios para los distintos proyectos de ley presentes en el instrumento.

Polarización y alineamiento político

La idea clave del análisis es caracterizar los proyectos de ley con dos medidas: *alineamiento político* y *polarización*. El alineamiento es una variable de consistencia interna (intra-grupo), y la polarización es una variable de consistencia externa (entre grupos).

A partir de los datos de votaciones se calcularon dos coeficioentes por cada proyecto de ley:

1. *Coeficiente de alineamiento político*: el grado de cohesión en el voto que tienen los miembros del Congreso con respecto a su partido (intra-grupo, solo en el contexto de la votación).
2. *Coeficiente de polarización*: el grado en que la votación divide al grupo de votantes en polos opuestos (entre grupos).

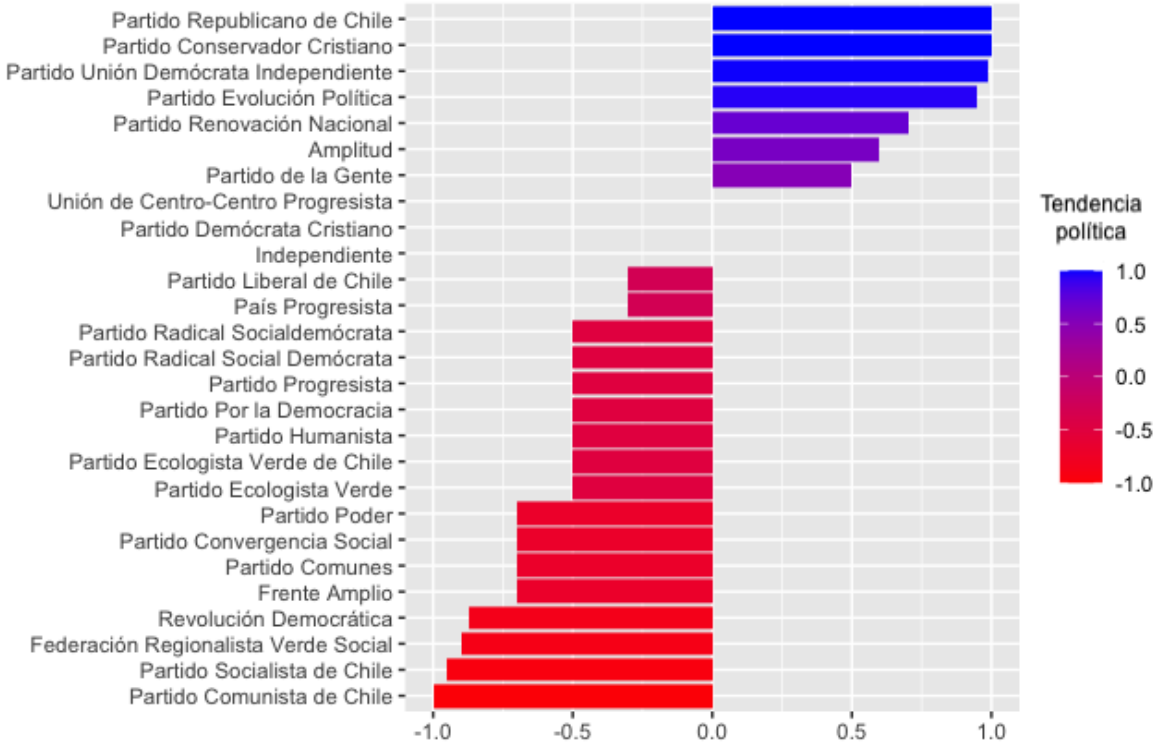

Figura 7.10: Tendencia política percibida asociada a cada partido

Luego, el valor promedio de cada índice se calcula para cada proyecto de ley, con el fin de caracterizarlo con un único valor para cada medida. El anexo E se detalla el procedimiento y las fórmulas de cálculo para el índice de polarización, mientras que en el anexo F se describe el procedimiento y las fórmulas de cálculo para la el índice de alineamiento político utilizado en este trabajo.

De esta manera, junto al grafo, el instrumento incorpora indicadores globales de alineamiento y polarización, los cuales permiten clasificar los proyectos de ley en cuatro categorías analíticas descritas en detalle en los artículos [Cifuentes-Silva et al., 2023] y [Cifuentes-Silva et al., 2024], los cuales son parte de este trabajo:

1. *Consenso técnico*: proyectos con polarización baja y alineamiento alto en la votación; es decir, proyectos donde se estableció un consenso técnico y sin antagonismos políticos en la votación.

2. *Interés temático/local*: proyectos con polarización baja y alineamiento bajo en la votación; es decir, proyectos de interés temático o local, de modo que un parlamentario representa dichos intereses, y el antagonismo surge frente al desinterés de otros miembros del Congreso.

3. *Interés personal*: proyectos con polarización alta y alineamiento bajo en la votación; es decir, muestran una divergencia entre un parlamentario y su partido, lo que sugiere la prevalencia de intereses personales sobre los principios partidarios.

4. *Postura ideológica*: proyectos con polarización alta y alineamiento alto en la votación; es

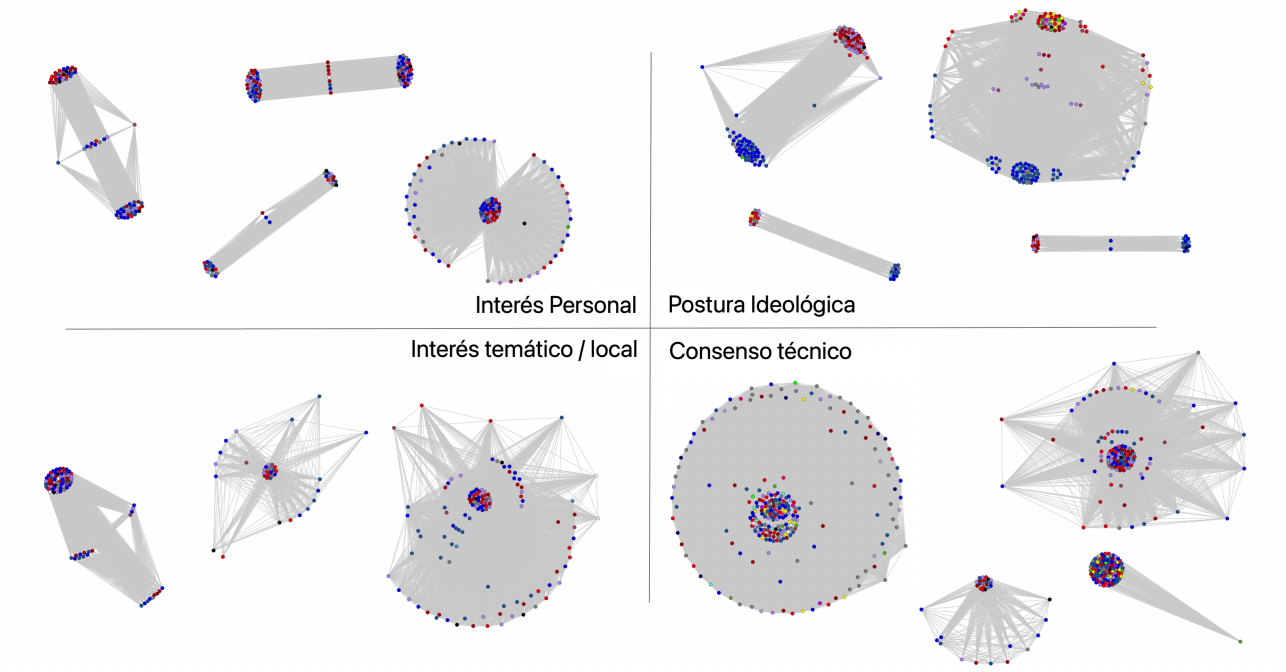

Figura 7.11: Varios grafos de fuerzas representando votaciones de proyectos de ley

decir, reflejan una divergencia en el eje político entre izquierda y derecha, de modo que los votos de los proyectos se ordenan ideológicamente.

De este modo, el instrumento no solo permite visualizar las relaciones entre parlamentarios en función de su comportamiento de voto, sino también interpretar la naturaleza política de los proyectos en discusión, tal como se visualiza en la figura 7.11.

7.5 *Instrumento 3:* Visualizador de rol clave en el contexto de un tema de interés legislativo

7.5.1 Descripción del instrumento

Este instrumento permite representar la posición relativa de los parlamentarios dentro de una red de colaboración temática, construida a partir de la coautoría de participaciones registradas en los diarios de sesiones. Estas participaciones corresponden principalmente a proyectos de ley y proyectos de acuerdo vinculados a una temática específica, definida según la jerarquía establecida en el Instrumento 1 (sección 7.3.3).

La visualización resultante se presenta como un grafo de relaciones entre parlamentarios, donde cada nodo representa a una persona, codificada cromáticamente según su afiliación partidaria, siguiendo la lógica del Instrumento 2 (sección 7.4.3). Los arcos entre nodos representan la coautoría de al menos un documento parlamentario entre los actores conectados. Además del grafo, la visualización incluye dos listas que agrupan automáticamente a los parlamentarios según dos roles clave: (i) los *intermediadores*, cuya posición estructural permite conectar distintos grupos dentro de la red, y (ii) los *líderes intragrupo*, identificados por su centralidad y elevada conectividad dentro de un subgrupo temático. La figura 7.12 presenta la interfaz de usuario del instrumento desarrollado y presentado al GE. Por su parte, la tabla 7.3 describe los principales aspectos técnicos y funcionales del instrumento.

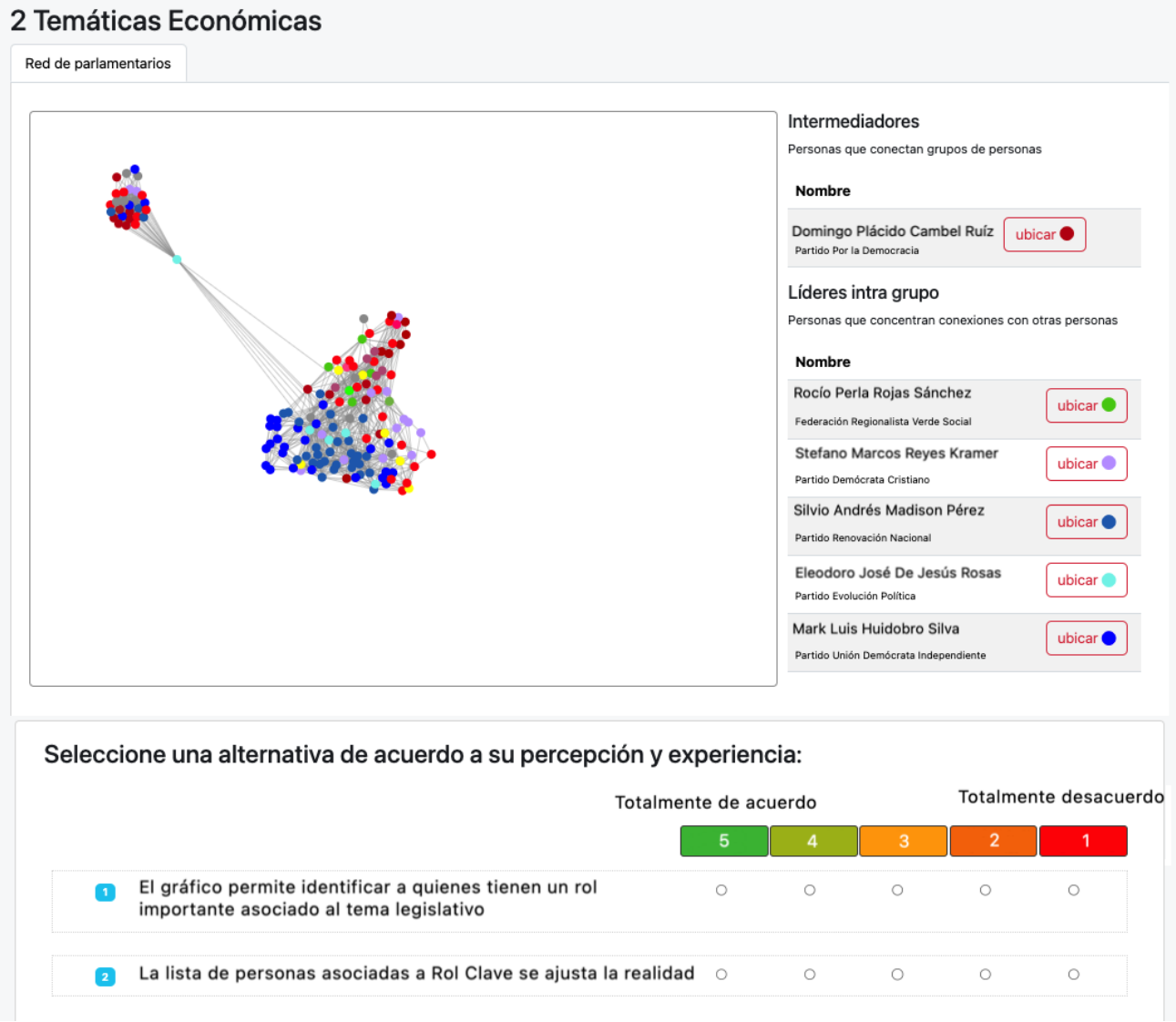

Figura 7.12: Interfaz de usuario del instrumento para detectar roles clave

7.5.2 Flujo de procesamiento

El flujo de procesamiento de este instrumento se basa en el mismo mecanismo de adquisición de datos descrito en el Instrumento 7.3.2, utilizando servicios de consulta al endpoint SPARQL para obtener los diarios de sesiones, los cuales son procesados y clasificados temáticamente de acuerdo con la jerarquía descrita en la sección 7.3.3.

Sin embargo, a diferencia del Instrumento 1, donde el enfoque está centrado en representar los documentos asociados a cada persona, en este caso la lógica se invierte: se parte desde los documentos asociados a un tema específico, para luego identificar las personas que participan en ellos. Este cambio de perspectiva permite construir una red de colaboración temática a partir de una proyección de un grafo bipartito documento-persona hacia un grafo persona-persona, considerando únicamente aquellas intervenciones registradas en documentos clasificados en la categoría temática en análisis.

Los documentos filtrados temáticamente son almacenados en archivos, repositorios o bases de datos Not Only SQL (NoSQL), lo cual facilita su procesamiento posterior por parte de

componentes de analítica.

Posteriormente, se aplican algoritmos de análisis de redes sociales (SNA), específicamente el cálculo de métricas de centralidad como *betweenness centrality* y *degree centrality*, con el fin de caracterizar estructuralmente la posición de cada parlamentario en la red. Estas métricas se combinan con el valor de alineamiento político de cada persona, calculado según la metodología definida en el Instrumento 7.4.3, lo que permite definir dos nuevos indicadores compuestos: el *valor de liderazgo intragrupo* y el *valor de intermediación* entre bloques distintos.

Finalmente, los resultados son presentados al usuario mediante una visualización gráfica tipo grafo, donde los nodos representan a los parlamentarios y los vínculos, su coautoría temática. Un panel complementario entrega una lista de personas que cumplen roles clave dentro de la red, y permite resaltarlas interactivamente en la visualización mediante un botón asociado a cada una de ellas, facilitando así su identificación dentro de la estructura colaborativa. La figura 7.13 presenta los componentes activos del marco de trabajo para el desarrollo de este instrumento.

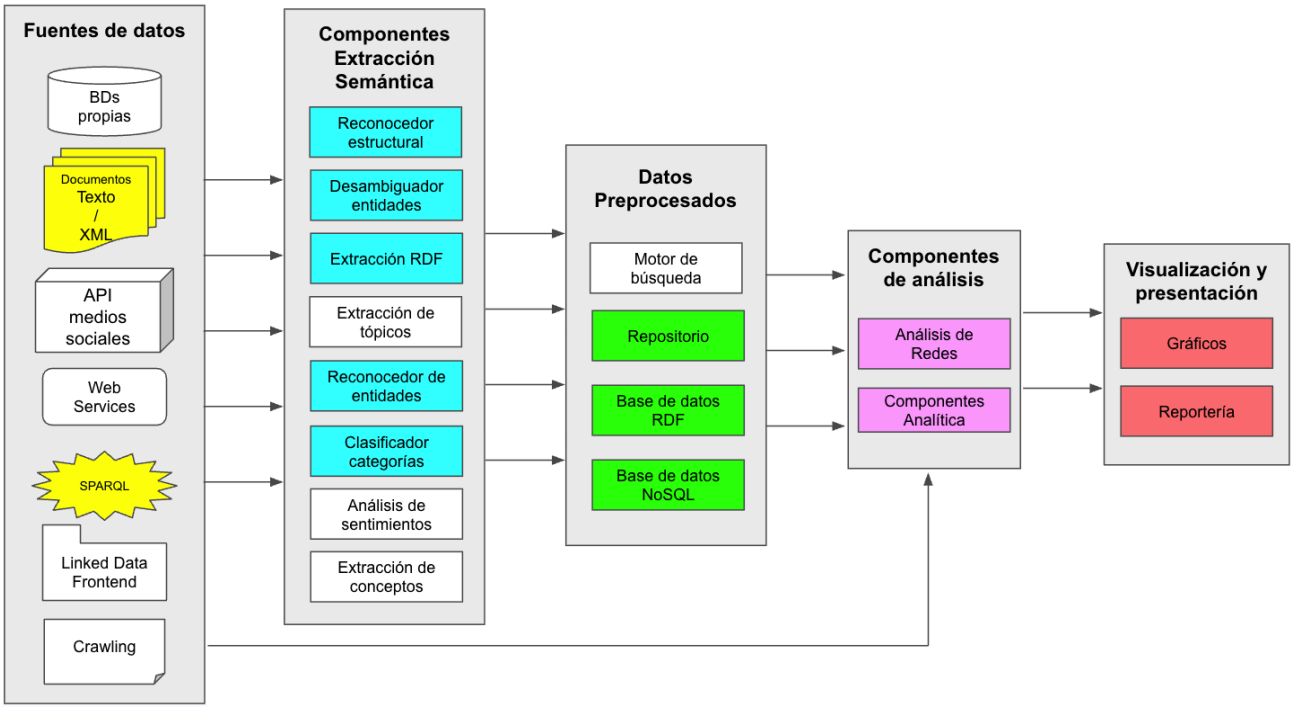

Figura 7.13: Componentes activos del marco de trabajo para desarrollo del instrumento 3

7.5.3 Fundamentos de diseño

Para diseñar este instrumento se consideró la problemática de la identificación de roles clave en una red, teniendo en cuenta que el principal insumo con que se cuenta son documentos asociados a más de una persona, y también teniendo en cuenta que en los instrumentos anteriormente descritos ya se han desarrollado insumos que pueden ser de utilidad.

En ese contexto, la estrategia adoptada fue la de generar una red con dos tipos de nodos conectados: personas y documentos, donde cada documento puede estar conectado a más de una persona mediante una relación de autoría o copatrocinio (ya que aunque son relaciones distintas en la práctica en el Congreso Nacional Chileno no se diferencian). Posteriormente, dado que se requiere analizar roles por tema de interés legislativo, se hace necesario separar el grafo completo

por tema. Para ello, cada documento es clasificado en una de las categorías existentes en la taxonomía de temas legislativos definida en el instrumento 1 sección 7.3.3. Para este fin se hace necesaria la implementación de un clasificador multiclase que implemente todas las categorías de la taxonomía. Luego, se separan todos los grafos por tema, con lo que es posible separar grafos persona-documento por tema de interés legislativo. A partir de estos grafos persona-documento, se calculan proyecciones de grafo persona-persona mediante algoritmos de SNA, en los cuales se puede visualizar la posición de cada persona en el grafo y calcular dentro de otros: métricas a nivel global (de grafo), a nivel de componente (sub grafo) o a nivel de persona. Para complementar el análisis y añadir factores que indiquen liderazgo político fuera del tema bajo análisis, se utiliza la métrica de alineamiento por persona para componer las métricas de intermediadores y líderes intra grupo. A continuación se describen en detalle cada uno de los aspectos previamente mencionados.

Clasificación de texto

Al igual que en el Instrumento 1, este instrumento requiere la implementación de un conjunto de clasificadores basados en la taxonomía de términos definida en la sección 7.3.3. No obstante, en este caso resulta más apropiado utilizar el árbol completo de categorías de temas de interés legislativo, ya que el objetivo es segmentar las relaciones parlamentarias según temáticas específicas. Considerar únicamente el primer nivel de la jerarquía podría generar redes de documentos excesivamente densas (dependiendo del número de documentos), y con alta variabilidad, dificultando así una visualización clara y útil. En este contexto, se considera más apropiado utilizar el segundo nivel de granularidad de la taxonomía, el cual permite construir un mayor número de redes temáticas más específicas. A diferencia del Instrumento 1, este enfoque presenta menos restricciones en cuanto a la visualización de los temas, lo que lo hace más adecuado para el análisis segmentado de relaciones parlamentarias.

Sin embargo, la clasificación de los documentos en el segundo nivel de la taxonomía requirió la adopción de una solución híbrida, principalmente debido a una limitación en los datos disponibles. En particular, durante la fase de etiquetado manual no se logró reunir una cantidad suficiente de documentos representativos para todas las subcategorías del conjunto de datos, lo que dificultó alcanzar el umbral mínimo de calidad (establecido arbitrariamente en un 75% en la métrica F1-Score) necesario para considerar un clasificador como aceptable, y a su vez impidió entrenar un clasificador bajo la misma lógica utilizada para las categorías de primer nivel. En consecuencia, de las 35 subcategorías originalmente definidas, fue posible implementar clasificadores automáticos para un total de 15, mientras que para las 20 restantes solo se logró identificar manualmente un grupo reducido de documentos, suficiente únicamente para construir el grafo correspondiente. Si bien esto puede parecer una debilidad a futuro pensando en la puesta en producción de los clasificadores, el impacto no es tal en el contexto que los datos de entrenamiento utilizados son solo una parte mínima de los existentes (solo la legislatura 367), por lo cual se presume que en el conjunto completo existen suficientes documentos para cada una de las categorías, lo cual habilitaría una implementación automatizada completamente. El anexo C presenta el detalle de la implementación de la clasificación de intervenciones de segundo nivel.

Análisis de Redes Sociales

La red se genera mediante la proyección de un grafo bipartito compuesto por personas y documentos legislativos asociados a un tema particular, lo que permite establecer vínculos entre parlamentarios a partir de su colaboración documentada.

Mediante este grafo, se calculan dos métricas estándar en SNA por cada persona:

- Centralidad de intermediación (*Betweenness Centrality* - BT): la cual mide cuánto un nodo actúa como puente en los caminos más cortos entre otros nodos de la red. En el caso de los parlamentarios, representa la capacidad de esa persona de conectar subgrupos o flujos de información; es decir, qué tanto ocupa posiciones de intermediación en la red. Un valor alto denota que el parlamentario puede influir más allá de sus conexiones inmediatas, facilitando coordinación o acuerdos.

- Centralidad de grado (*Degree Centrality* - DC): la cual mide cuántas conexiones directas tiene un nodo con otros nodos de la red. Para el caso de estudio, refleja cuántas conexiones directas tiene el parlamentario dentro de la red. Un valor alto indica que el parlamentario está altamente conectado, lo que puede reflejar visibilidad, popularidad o actividad dentro de la red.

A partir de los índices calculados para cada persona en el grafo temático, será posible utilizarlos de forma individual o combinada en el diseño de nuevos indicadores compuestos. No obstante, dado que cada métrica se expresa en escalas distintas y responde a lógicas propias (por ejemplo, algunas con máximos relativos y otras con valores absolutos), es necesario someterlas previamente a un proceso de normalización o escalado.

Para este propósito, se aplicará un procedimiento de escalado que transforma los valores originales a un rango común entre 0 y 1 para cada tema analizado. Esto permite no solo facilitar la interpretación de las métricas, sino también asegurar que su comparación o combinación sea válida en términos de magnitud relativa, lo que implica que los valores obtenidos para cada persona serán comparables entre sí, permitiendo identificar quiénes presentan mayor fuerza estructural en la red, ya sea desde la perspectiva de intermediación o centralidad.

Cálculo de índices de intermediación y liderazgo intra grupo

Para el cálculo de índices de intermediación y liderazgo intra grupo, adicionalmente a los índices de BT y DC normalizados, se ha incorporado el coeficiente de alineamiento político definido en el instrumento 2 sección 7.4.3. Si bien es posible calcular este coeficiente por proyecto de ley o incluso por tema, se tomará solo un promedio general por persona en lugar de promedios parciales, para evitar la imputación de datos en caso de que una persona no tenga promedio por tema (por ejemplo un parlamentario nuevo que no haya votado proyectos del tema), ni que se genere sesgo excesivo en casos de que un tema sea muy amplio. De la misma manera, se aplicará el mismo proceso de escalado para que todos los valores de todas las personas queden en el mismo rango entre 0 y 1; con ello, el índice de alineamiento de la persona reflejará qué tan alineada se encuentra respecto a la votación del resto de miembros de su partido en términos generales. En este sentido, el alineamiento normalizado permite medir la coherencia o sintonía de un parlamentario con la línea o postura del grupo, de modo que una alineación alta indica liderazgo en el plano ideológico intra-grupo, pues el parlamentario puede actuar como referente en la postura colectiva.

En este punto, ya con tres indicadores normalizados, se procede a desarrollar los dos índices a implementar:

1. **Índice de liderazgo intra grupo**

$$\text{Liderazgo}_{\text{intra-grupo}} = 0.5\,\text{AP} + 0.2\,\text{BT} + 0.3\,\text{DC}$$

AP : Alineamiento normalizado del parlamentario
BT : Betweenness Centrality normalizada
DC : Degree Centrality normalizada

Donde se prioriza el componente de AP (con peso 0,5) debido a que el liderazgo intra-grupo requiere sobre todo, mantener coherencia con la postura colectiva y encabezar esa línea frente a otros. De la misma manera, los pesos de BC (0,2) y DC (0,3) se ponderan menos aunque siguen siendo relevantes, ya que el liderazgo no solo depende de la coincidencia ideológica con el grupo, sino también de la habilidad de ejercer influencia en la red (puentes o intermediaciones) y del número de contactos directos.

2. **Índice de intermediación**

$$\text{Intermediación}_{\text{intra-grupo}} = 0.2\,\text{AP} + 0.5\,\text{BT} + 0.3\,\text{DC}$$

AP : Alineamiento normalizado del parlamentario
BT : Betweenness Centrality normalizada
DC : Degree Centrality normalizada

En este caso se prioriza el componente de BT (con peso 0,5) debido a que esta métrica es a que mide con precisión la capacidad de conectar grupos distintos actuando como puente. En el caso de DC (0.3), si bien no refleja intermediación en el sentido más estricto, sí indica presencia y visibilidad en la red, lo cual contribuye a la capacidad de ejercer influencia y facilitar conexiones. Finalmente, AP (0.2) se incluye con el peso menor (20%) ya que una cierta sintonía con el grupo puede facilitar el rol articulador, al contar con legitimidad y confianza en la red.

De este modo, una vez calculados los índices para todos los parlamentarios dentro de un tema específico, se identifican aquellas personas que desempeñan un rol clave en la red. Este análisis se realiza seleccionando los parlamentarios que presentan los valores más altos y atípicos en la distribución de cada uno de los índices. Para ello, se ordenan los valores y se aplica un umbral basado en el criterio de valores atípicos superiores, definido como Q3 + 1.5 * IQR (siendo IQR el rango intercuartil. Dado que el interés se centra únicamente en los valores extremos superiores, se consideran solo los outliers por sobre ese umbral. Los puntajes más altos identificados por este método permiten asociar a dichos parlamentarios con los dos roles clave: *Rol de liderazgo intragrupo* y *Rol de Intermediación*.

Objetivo del instrumento	Representar parlamentarios con un rol clave en una red temática de colaboración, que han sido identificados mediante la aplicación del marco de trabajo, y permitir una evaluación por criterio experto.
Objetivo de investigación	Validar el mecanismo tecnológico de identificación de parlamentarios con roles clave en el contexto de un tema específico basado en su posición en la red de colaboraciones.
Pregunta de investigación	**RQ3**: ¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas quién cumple un rol clave en el contexto de un tema específico?
Conjunto de datos de prueba	19.990 Intervenciones asociadas a 263 Diarios de Sesión de la Cámara de Diputados (155) y del Senado (108) de Chile correspondientes a la legislatura 367 (2019-03-11 al 2020-03-10) en formato texto.
Tecnologías Semánticas asociadas al instrumento	1. Reconocimiento de entidades nombradas 2. Marcaje estructural del texto 3. Clasificación de texto 4. Desambiguación de entidades 5. Ontologías y taxonomías de términos 6. Análisis de redes sociales 7. Visualización de datos 8. Marcaje estructural del texto
Preguntas en el instrumento	1. El gráfico permite identificar a quiénes tienen un rol importante asociado al tema legislativo 2. La lista de personas asociadas a Rol Clave se ajusta a la realidad
Posibles usos	• Identificar referentes legislativos que tienen un papel central o articulador en un tema específico • Apoyar procesos de asesoría parlamentaria al mejorar la visión de parlamentarios objetivo en el desarrollo de asesoría por oferta • Visualizar colaboración y segmentación parlamentaria, al representar relaciones entre participantes de distintos partidos políticos, pudiendo identificar alianzas o diferencias por tema • Detectar actores *puente* entre bloques ideológicos, lo que puede ser estratégico para planificar el estrategias para el diálogo político o la construcción de acuerdos

Tabla 7.3: Tabla resumen de descripción del instrumento 3

Capítulo 8

Resultados y análisis de datos

8.1 Introducción

Este capítulo presenta los resultados y análisis de los datos recopilados tras la aplicación de los tres instrumentos de evaluación, que utilizan datos procesados a través del marco de trabajo basado en Tecnologías Semánticas, a un GE del ámbito político-legislativo.

Los datos fueron obtenidos mediante la presentación de escenarios reales, en los cuales los participantes respondieron a preguntas estructuradas en tres tipos de escenarios y a su vez materializadas en los tres instrumentos definidos en el capítulo 7. Las respuestas, formuladas en escala de Likert, permitieron medir la intensidad con la que los expertos consideran correctas las afirmaciones presentadas en cada caso. Cada tipo de instrumento permite validar mediante ejemplos prácticos las preguntas de investigación definidas en la sección 3.2, y en conjunto contribuyen a la validación de la hipótesis general asociada al marco de trabajo propuesto.

El análisis de los datos se ha desarrollado bajo una metodología exploratoria y descriptiva, con el propósito de identificar patrones, tendencias y relaciones en las respuestas de los expertos, así como validar los instrumentos de análisis utilizados. Este enfoque no solo permite caracterizar la percepción de los participantes, sino también establecer vínculos entre las variables que podrían influir en la validación de la hipótesis de la investigación.

De esta manera, los hallazgos obtenidos en este capítulo servirán como base para la discusión de resultados, la evaluación del marco de trabajo de tecnologías semánticas en un contexto real de análisis político-legislativo, y el desarrollo de las conclusiones del presente trabajo.

8.2 Datos, técnicas y herramientas para el análisis

Se utilizó una muestra de 13 usuarios expertos pertenecientes a la Biblioteca del Congreso Nacional (BCN), los cuales desarrollan labores de asesoría parlamentaria y pertenecen a distintas áreas funcionales. Las distintas áreas del conocimiento de los usuarios expertos se enmarcan en las ciencias sociales y juridicas, dentro de los cuales existen mayoritariamenta abogados/as y sociólogos. La sección 4.5 presenta en detalle la descripción del GE.

La recopilación de datos se llevó a cabo mediante una aplicación desarrollada específicamente para este estudio (descrita en el Capítulo 7), la cual permitió capturar tanto las respuestas a las preguntas expuestas en cada instrumento, como también los tiempos que tomó cada usuario en responder.

En total, se administraron 322 casos, distribuidos en 212 correspondientes al instrumento 1

(*Temas de interés Parlamentario*), 70 al instrumento 2 (*Cohesión política*)y 40 al instrumento 3 (*Roles clave*). Durante la fase de evaluación se identificaron errores en el despliegue de 3 preguntas (2 de tipo 2 y 1 de tipo 3), los cuales fueron debidamente considerados y excluídos en el análisis. Para la parte estadística se adoptó un enfoque descriptivo y exploratorio, apoyado en la elaboración de gráficos, el cálculo de medidas de tendencia central y dispersión, y análisis de correlación, complementado con el cálculo del coeficiente alfa de Cronbach para evaluar la consistencia interna y la idoneidad de las preguntas de cada instrumento. Además, se incorporó la variable sexo de los usuarios expertos, como también la variable área del conocimiento asociada a la profesion de los expertos, para explorar posibles diferencias en las respuestas y en los tiempos de respuesta.

Si bien las respuestas en la escala de Likert corresponden a una escala ordinal asociada a la percepción de los usuarios expertos con rango entre *Totalmente de acuerdo* hasta *Totalmente en desacuerdo* (ver sección 4.3), para su análisis se realizó una conversión de los valores categóricos a los siguientes valores numéricos: *Totalmente de acuerdo (valor 5), Parcialmente de acuerdo (valor 4), No lo tengo claro (3), Parcialmente en desacuerdo (valor 2), Totalmente en desacuerdo(1).*

La estructura de la base de datos de respuestas analizada se describe en la tabla 8.1:

Variable	**Tipo de valores**	**Descripción**
Tipo caso	Temas interés parlamentario, Cohesión política , Roles clave	Tipo de caso planteado asociado a las preguntas
Valor 1	Numérica	Respuesta de la pregunta 1 asociada al caso en rango (1,5)
Valor 2	Numérica	Respuesta de la pregunta 2 asociada al caso en rango (1,5)
Tiempo respuesta 1 (t1)	Numérica	Tiempo en segundos para responder la pregunta 1
Tiempo respuesta 2 (t2)	Numérica	Tiempo en segundos para responder la pregunta 2
Tiempo total	Numérica	Tiempo en segundos para responder el caso, corresponde a la suma de t1 y t2
Sexo	Masculino, Femenino	Sexo del usuario que responde
Profesión	Abogado, Sociólogo	Profesión del usuario/a que responde

Tabla 8.1: Descripción de las variables analizadas

Dentro de las medidas estadísticas a utilizar en los análisis de resultados, se encuentran el uso de la media, mediana, desviación estándar, identificación del valor mínimo, identificación del valor máximo, el valor del primer cuartil (Q1), el valor del tercer cuartil (Q3), el coeficiente de variación [Forkman, 2009] y el coeficiente alfa de Cronbach [Everitt, 1998].

8.3 Análisis de datos Instrumento 1

A partir de los datos recopilados mediante el *Panel de visualización de temas de interés parlamentario*, descrito en la sección 7.3, se presenta a continuación un análisis estadístico descriptivo y exploratorio de la información obtenida.

En este experimento, participaron *9* de los 13 usuarios expertos, realizando un total de *770* observaciones.

La tabla 8.2 describe los datos de las respuestas asociados a este instrumento.

8.3.1 Vista general

Estadística	**Valor 1**	**t1** ***(s)***	**Valor 2**	**t2** ***(s)***	**Tiempo total** ***(s)***
Media	4,32	17,94	4,17	9,47	27,41
Mediana	5,00	4,50	5,00	3,00	8,00
Desviación estándar	0,97	65,19	1,00	40,37	83,72
Coeficiente de variación	22,54%	-	24,11%	-	-
Mínimo	1,00	1,00	2,00	1,00	2,00
Máximo	5,00	1170,00	5,00	653,00	1219,00
Q1	4,00	2,00	3,00	2,00	4,00
Q3	5,00	9,00	5,00	5,00	14,00

Tabla 8.2: Estadísticas descriptivas del experimento 1 $N = 770$

Un gráfico que representa los valores de las respuestas se presenta en la figura 8.1.

Para asegurar consistencia entre las preguntas, se calculó el coeficiente *alfa de Cronbach*, el cual entrega un valor de *0.92* para un total de *770* observaciones.

La figura 8.2 muestra un gráfico con los tiempos de respuesta, diferenciados según el valor de la respuesta y el tipo de pregunta. Se optó por omitir los valores atípicos en los diagramas de caja para destacar la distribución central, dado que, como se observa en la tabla 8.2, algunos valores extremos excesivamente altos distorsionan la visualización, afectando la interpretación de los datos.

Los siguientes corresponden a los coeficientes de correlación de Spearman para cada pregunta y su tiempo de respuesta:

- Pregunta 1 *(Valor 1)* vs Tiempo respuesta 1 *(t1)*: **-0.197**
- Pregunta 2 *(Valor 2)* vs Tiempo respuesta 2 *(t2)*: **-0.235**

Acorde a las variables adicionales recopiladas por usuario del GE asociadas a la respuesta, a continuación se presentan análisis agregados a fin de no segmentar el conjunto de datos a tal punto que permita identificar a los participantes.

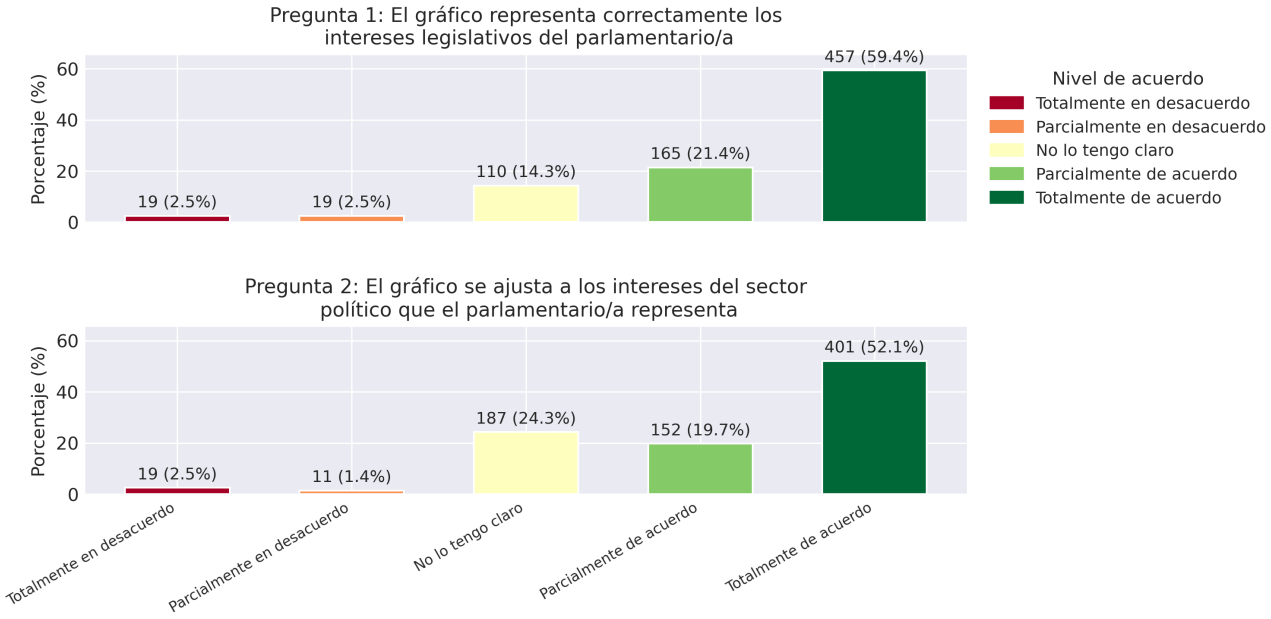

Figura 8.1: Distribución de las respuestas asociadas al instrumento 1

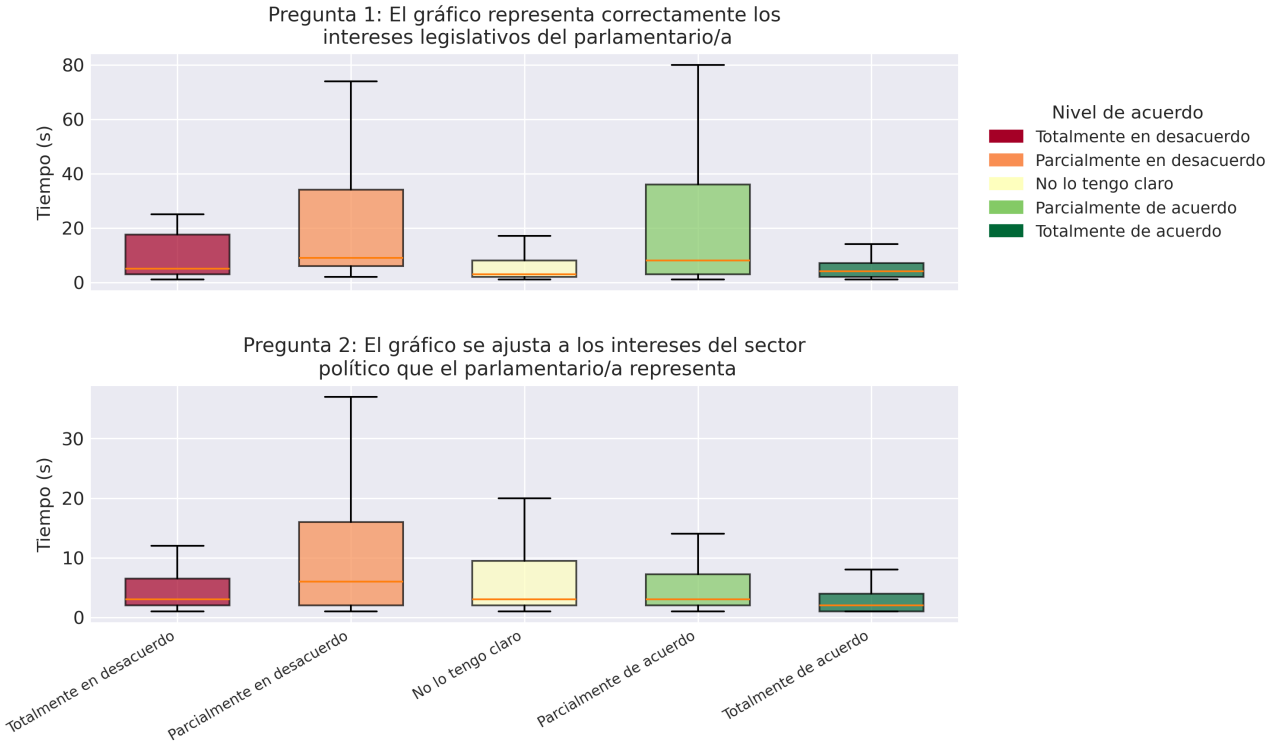

Figura 8.2: Tiempos utilizados en responder preguntas del instrumento 1 (sin valores atípicos)

8.3.2 Análisis agregado por sexo

La figura 8.3 presenta dos histogramas asociados al instrumento 1, cada uno asociado a un tipo de pregunta, que muestran la distribución de las respuestas según el sexo de las personas participantes, al tiempo que permite una comparación visual de la frecuencia de respuestas entre

los dos grupos.

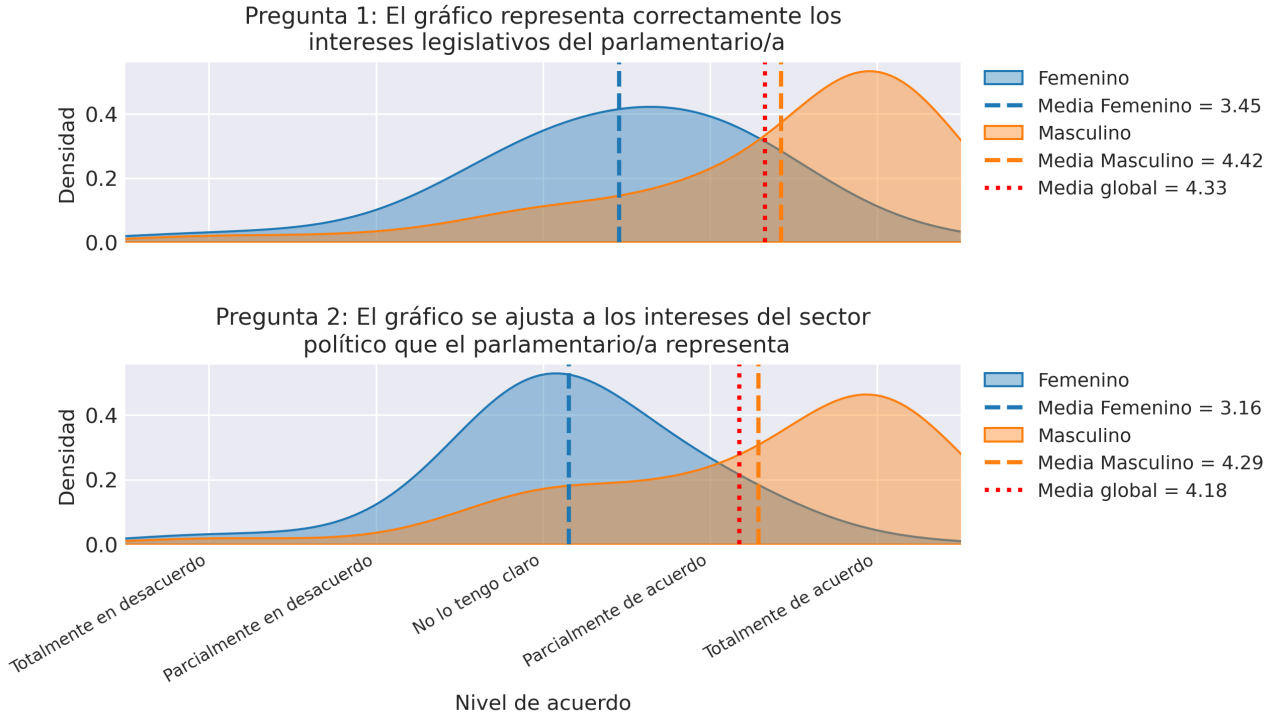

Figura 8.3: Distribución de las respuestas por sexo asociadas al instrumento 1

De la misma manera, los diagramas de calor (*heatmap*) de la figura 8.4, presentan los tiempos de respuestas por sexo, divididos por tipo de pregunta y por valor de respuesta.

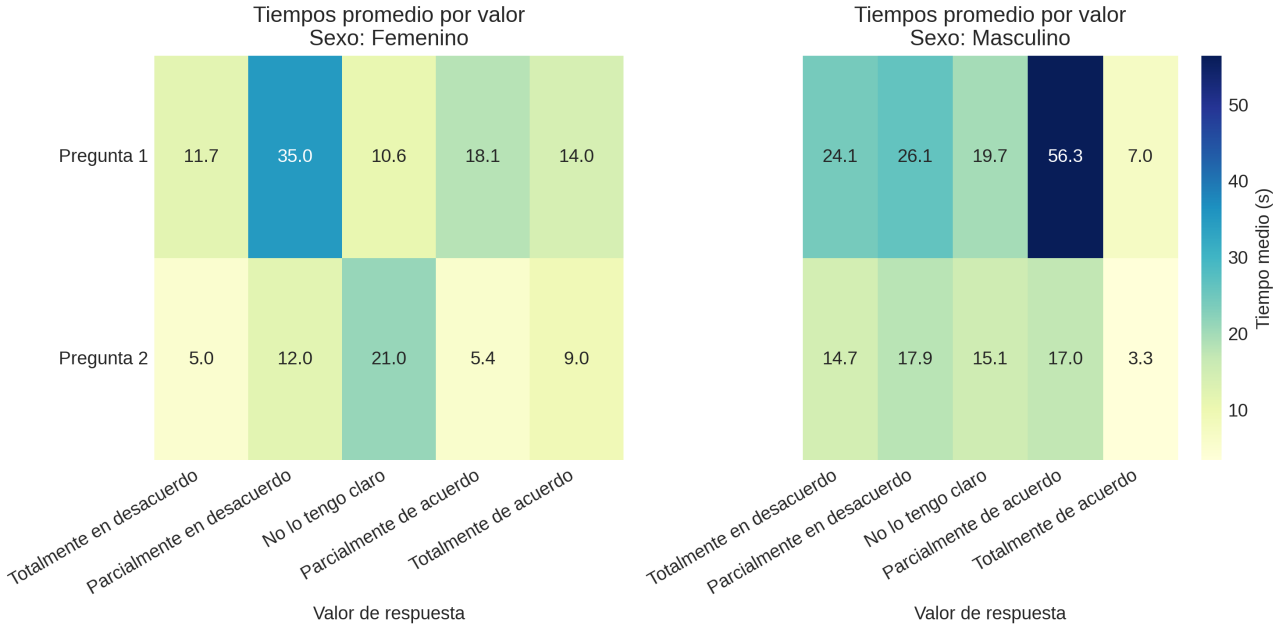

Figura 8.4: Tiempos utilizados por sexo para el instrumento 1 por valor de respuesta

8.3.3 Análisis agregado por profesión

En la misma lógica anterior, la figura 8.5 presenta los histogramas por tipo de pregunta asociado al instrumento 1, donde se muestra la distribución de las respuestas obtenidas, diferenciadas según el profesion de las personas participantes.

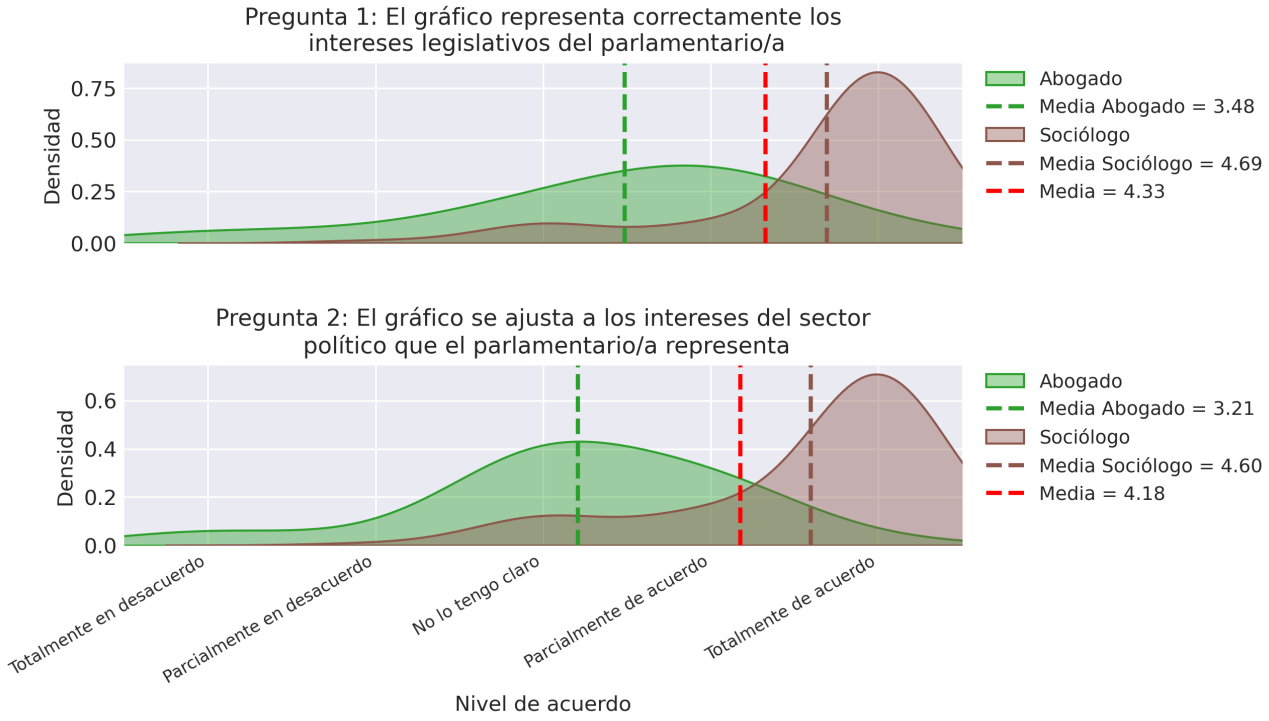

Figura 8.5: Distribución de las respuestas por profesión para el instrumento 1

De forma homóloga, el *heatmap* de la figura 8.6, presenta los de tiempos de respuestas por profesión, divididos por tipo de pregunta y por valor de respuesta.

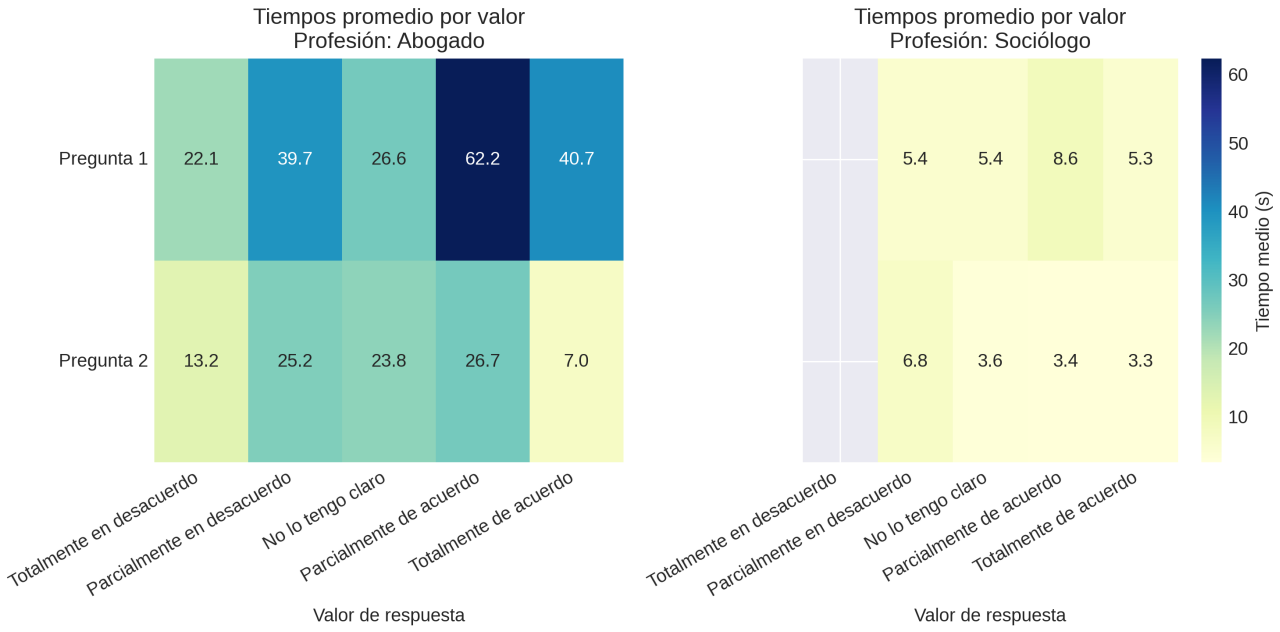

Figura 8.6: Tiempos utilizados por profesión en instrumento 1 por valor de respuesta

8.4 Análisis de datos Instrumento 2

A partir de los datos recopilados mediante el *Panel de visualización e indicadores sobre cohesión política*, descrito en la sección 7.4, se presenta a continuación un análisis estadístico descriptivo y exploratorio de la información obtenida.

En este experimento, participaron *10* de los 13 usuarios expertos, realizando un total de *379* observaciones. La tabla 8.3 describe los datos de las respuestas asociados a este instrumento.

8.4.1 Vista general

Estadística	**Valor 1**	**t1 *(s)***	**Valor 2**	**t2 *(s)***	**Tiempo total *(s)***
Media	4,60	36,50	4,55	14,81	51,32
Mediana	5,00	11,00	5,00	5,00	18,00
Desviación estándar	0,75	111,18	0,78	42,25	127,52
Coeficiente de variación	16,38%	-	17,28%	-	-
Mínimo	1,00	1,00	2,00	1,00	3,00
Máximo	5,00	1610,00	5,00	730,00	1635,00
Q1	4,00	3,00	4,00	2,00	5,00
Q3	5,00	32,50	5,00	15,50	47,50

Tabla 8.3: Estadísticas descriptivas del experimento 2 $N = 379$

Un gráfico que representa los valores de las respuestas se presenta en la figura 8.7.

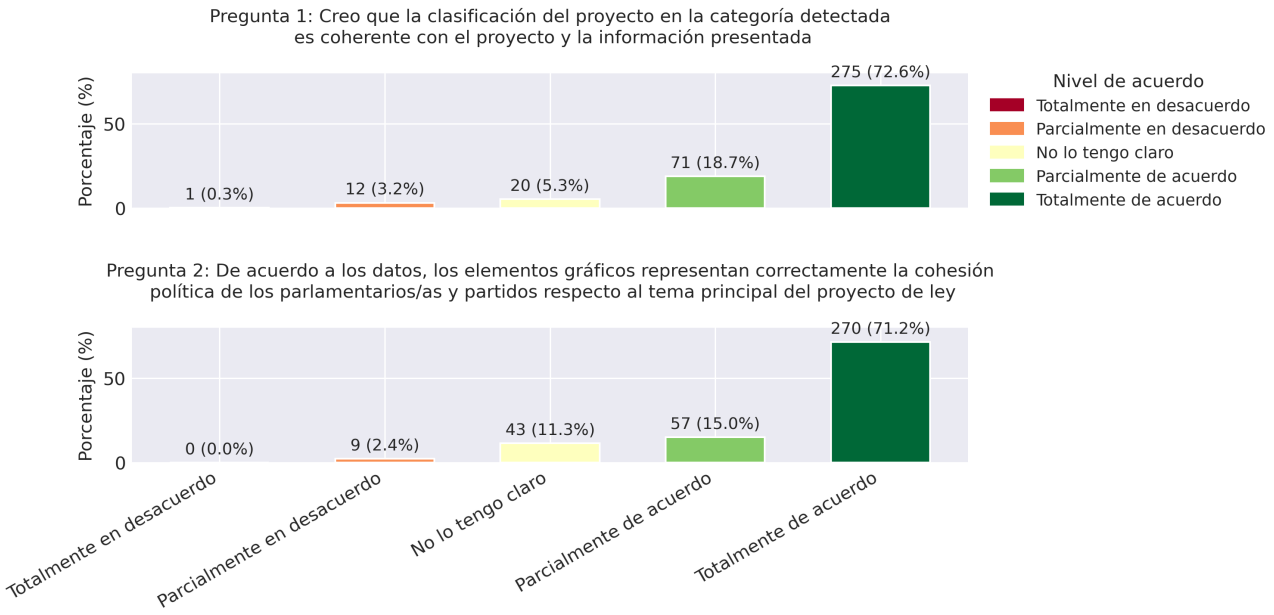

Figura 8.7: Distribución de las respuestas asociadas al instrumento 2

Para asegurar consistencia entre las preguntas, se calculó el coeficiente *alfa de Cronbach*, el cual entrega un valor de *0.92* para un total de *379* observaciones.

Al igual que para el instrumento 1, en el caso del instrumento 2 la figura 8.8 muestra un gráfico con los tiempos de respuesta, diferenciados según el valor de la respuesta y el tipo de pregunta. Igualmente en este caso, se optó por omitir los valores atípicos en los diagramas de caja para destacar la distribución central, dado que, como se observa en la tabla 8.3, algunos valores extremos excesivamente altos distorsionan la visualización, afectando la interpretación de los datos.

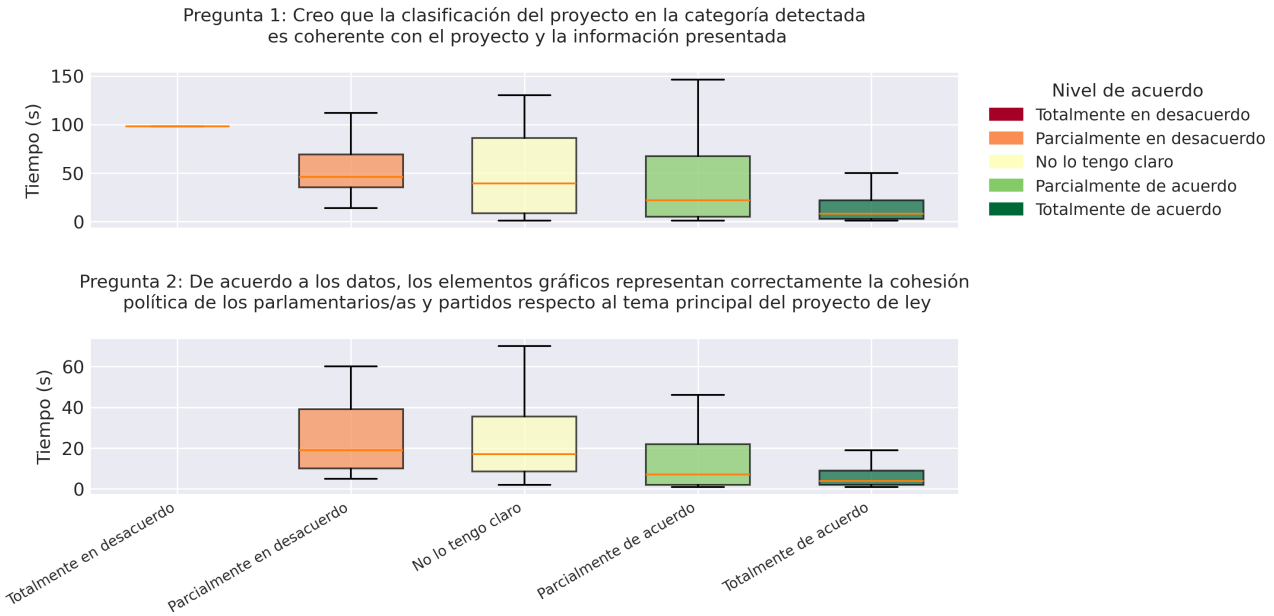

Figura 8.8: Tiempos utilizados en responder asociados al instrumento 2 por valor de respuesta

Los siguientes corresponden a los coeficientes de correlación de Spearman para cada pregunta y su tiempo de respuesta:

- Pregunta 1 *(Valor 1)* vs Tiempo respuesta 1 *(t1)*: **-0.305**
- Pregunta 2 *(Valor 2)* vs Tiempo respuesta 2 *(t2)*: **-0.333**

También para el caso del instrumento 2, dado que son las mismas variables adicionales recopiladas por usuario del GE que en el caso del instrumento 1, a continuación se presentan análisis agregados a fin de no segmentar el conjunto de datos a tal punto que permita identificar a los participantes.

8.4.2 Análisis agregado por sexo

La figura 8.9 presenta un histograma de los valores de respuestas por tipo de pregunta asociado al instrumento 2, diferenciadas según el sexo de los participantes, lo que permite visualizar de manera comparativa las diferencias entre ambos grupos.

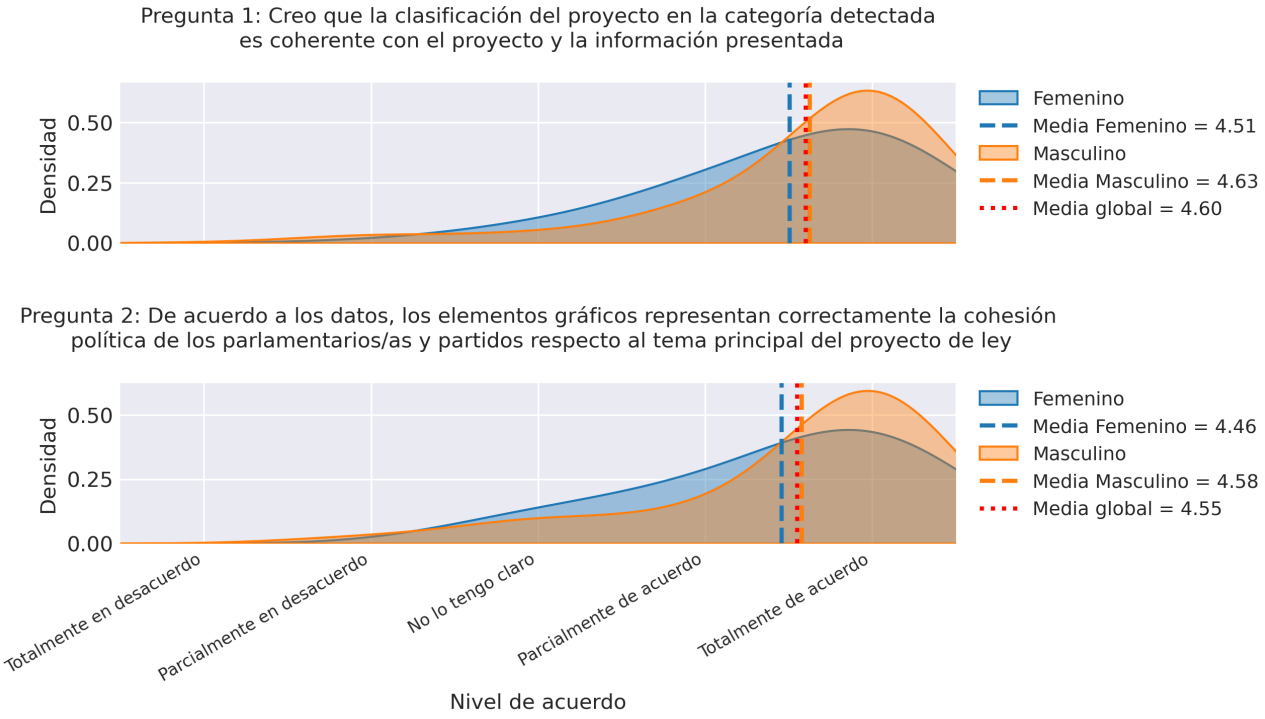

Figura 8.9: Distribución de las respuestas por sexo asociadas al instrumento 2

De la misma manera, los *heatmap* de la figura 8.10, presentan los tiempos de respuestas por sexo, divididos por tipo de pregunta y por valor de respuesta.

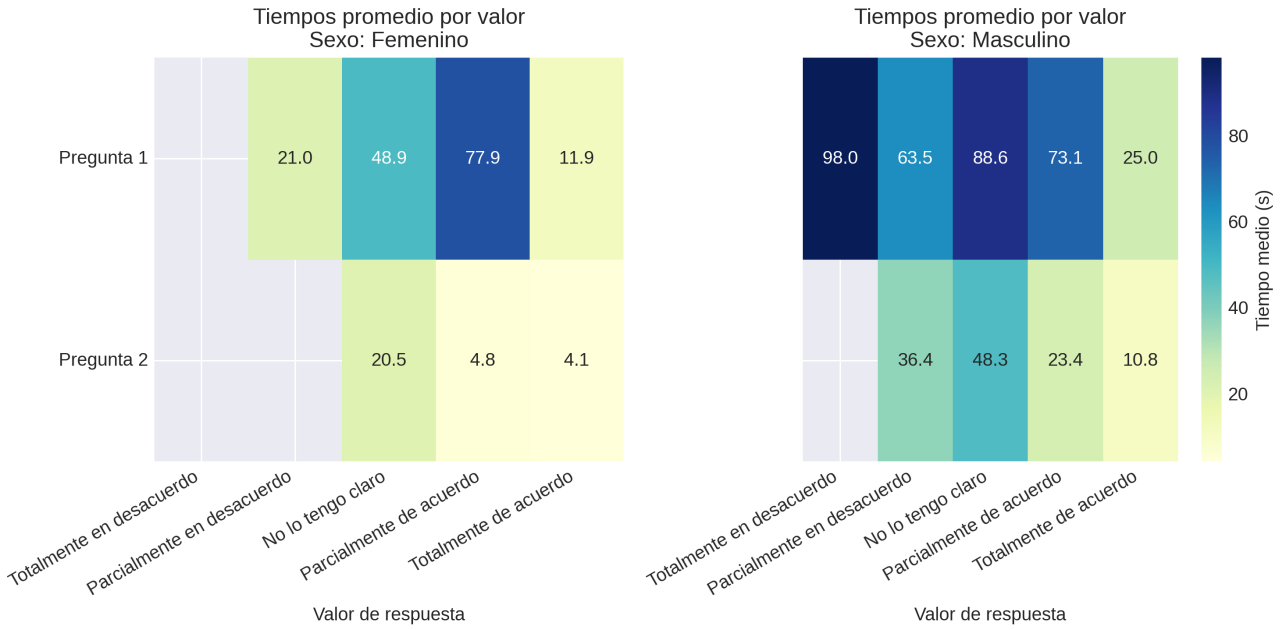

Figura 8.10: Tiempos utilizados por sexo para el instrumento 2 por valor de respuesta

8.4.3 Análisis agregado por profesión

En la misma lógica anterior, la figura 8.11 presenta un histograma por tipo de pregunta asociado al instrumento 2, donde se muestra la distribución de las respuestas obtenidas, diferenciadas según el profesion de las personas participantes.

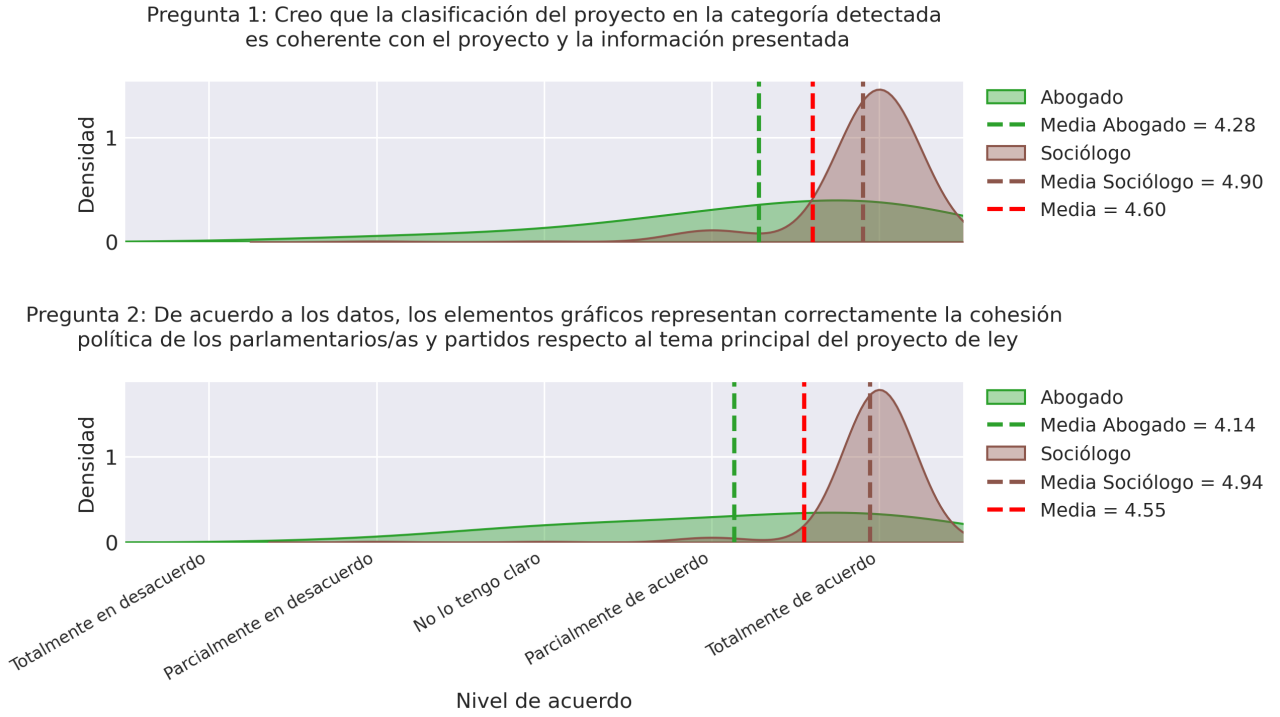

Figura 8.11: Distribución de las respuestas por área del conocimiento para el instrumento 2

De igual forma que para el instrumento anterior, el *heatmap* de la figura 8.12, presenta los de tiempos de respuestas por profesión, divididos por tipo de pregunta y por valor de respuesta.

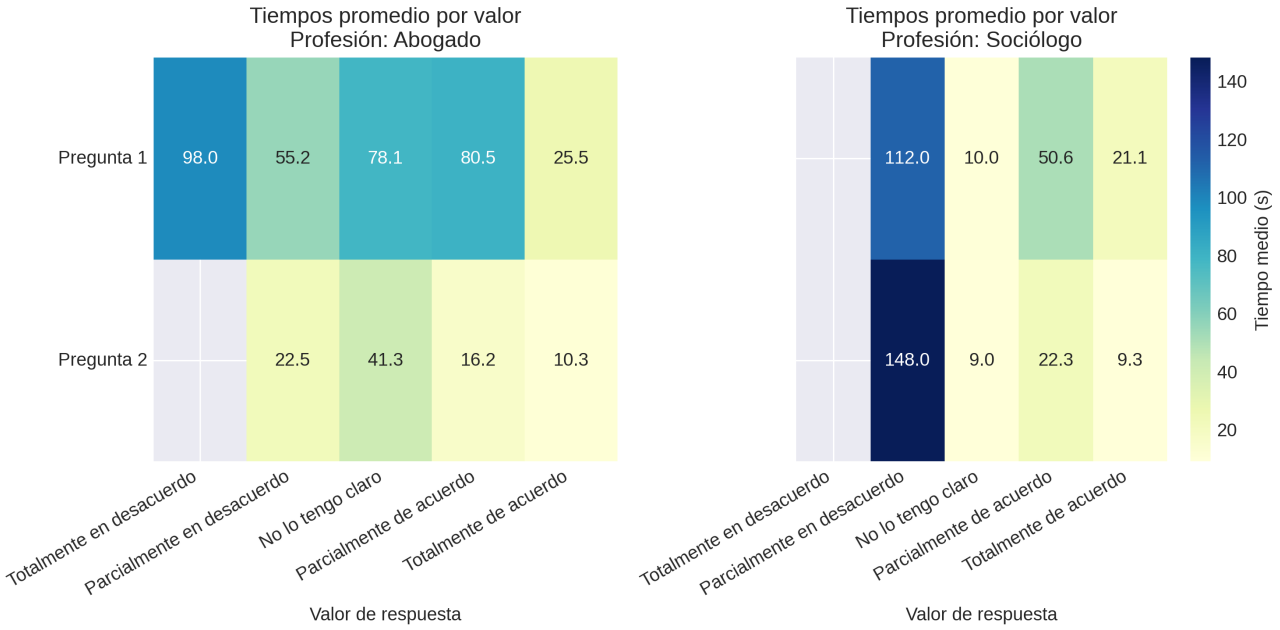

Figura 8.12: Tiempos utilizados por área del conocimiento en instrumento 2 por valor de respuesta

8.5 Análisis de datos Instrumento 3

A partir de los datos recopilados mediante el *Visualización de rol clave en el contexto de un tema de interés legislativo*, descrito en la sección 7.5, se presenta a continuación un análisis estadístico descriptivo y exploratorio de la información obtenida.

En este experimento, participaron los *13* usuarios expertos, realizando un total de *296* observaciones.

La tabla 8.4 describe los datos de las respuestas asociados a este instrumento.

8.5.1 Vista general

Estadística	**Valor 1**	**t1** ***(s)***	**Valor 2**	**t2** ***(s)***	**Tiempo total** ***(s)***
Media	3,89	39,26	3,72	21,58	60,85
Mediana	4,00	14,00	4,00	7,00	22,00
Desviación estándar	1,01	174,52	0,97	70,83	191,69
Coeficiente de variación	26,05%	-	26,04%	-	-
Mínimo	1,00	1,00	1,00	1,00	2,00
Máximo	5,00	2827,00	5,00	853,00	2869,00
Q1	3,00	6,00	3,00	3,00	10,00
Q3	5,00	31,25	5,00	16,00	50,25

Tabla 8.4: Estadísticas descriptivas del experimento 3 $N = 296$

Un gráfico que representa los valores de las respuestas se presenta en la figura 8.13.

Para asegurar consistencia entre las preguntas, se calculó el coeficiente *alfa de Cronbach*, el cual entrega un valor de *0.85* para un total de *296* observaciones.

Al igual que en los instrumentos anteriores, la figura 8.14 presenta un gráfico con los tiempos de respuesta correspondientes al instrumento 3, desagregados según el valor de la respuesta y el tipo de pregunta. Considerando que se observan condiciones similares en la distribución de los datos, se decidió excluir los valores atípicos en los diagramas de caja, con el fin de resaltar la tendencia central. Tal como se evidencia en la tabla 8.4, la presencia de valores extremos excesivamente elevados tiende a distorsionar la visualización y dificulta una interpretación adecuada de los resultados.

Los siguientes corresponden a los coeficientes de correlación de Spearman para cada pregunta y su tiempo de respuesta:

- Pregunta 1 *(Valor 1)* vs Tiempo respuesta 1 *(t1)*: **-0.256**
- Pregunta 2 *(Valor 2)* vs Tiempo respuesta 2 *(t2)*: **-0.345**

Acorde a las variables adicionales recopiladas por usuario del GE asociadas a la respuesta, a continuación se presentan análisis agregados a fin de no segmentar el conjunto de datos a tal punto que permita identificar a los participantes.

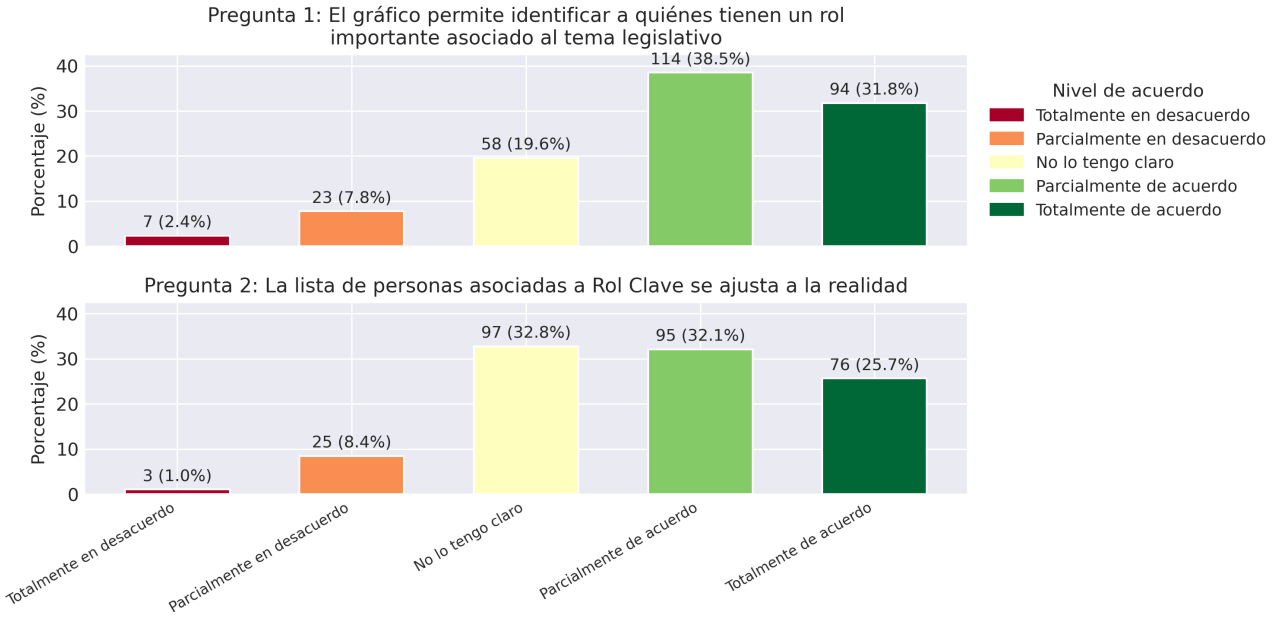

Figura 8.13: Distribución de las respuestas asociadas al instrumento 3

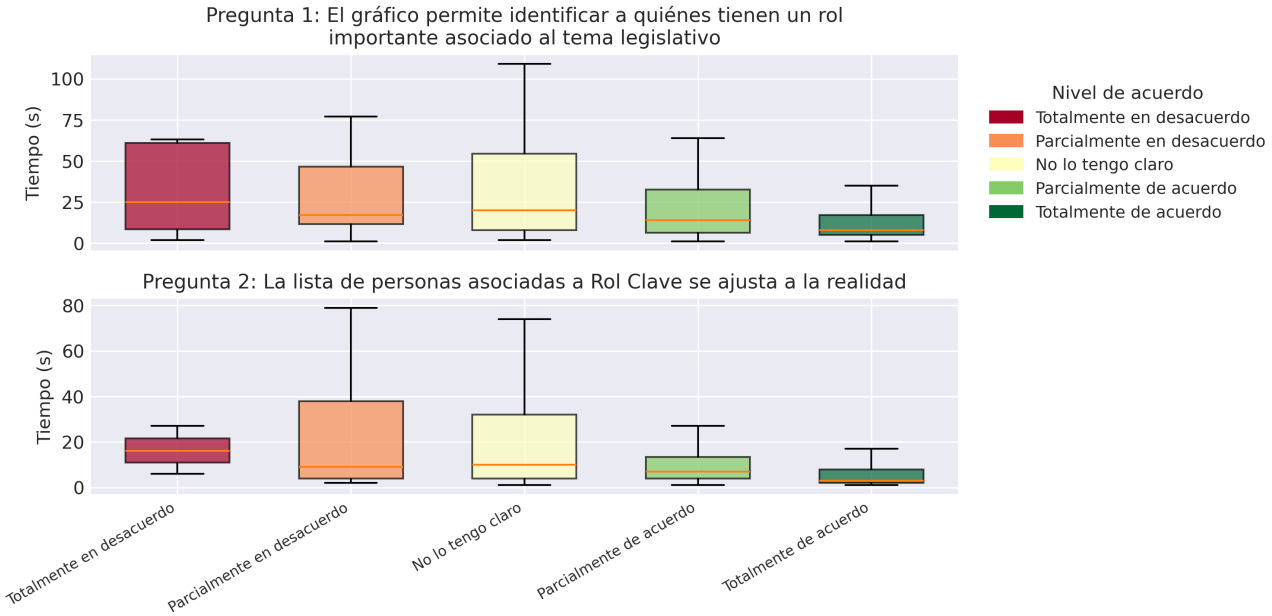

Figura 8.14: Tiempos utilizados en responder asociados al instrumento 3 por valor de respuesta

8.5.2 Análisis agregado por sexo

La figura 8.15 muestra los histogramas por tipo de pregunta correspondiente al instrumento 3, en el que se representa la distribución de las respuestas obtenidas, diferenciadas según el sexo de las personas participantes, permitiendo comparar visualmente la frecuencia de respuestas entre ambos grupos.

De la misma manera, los *heatmap* de la figura 8.16, presentan los tiempos de respuestas por sexo, divididos por tipo de pregunta y por valor de respuesta.

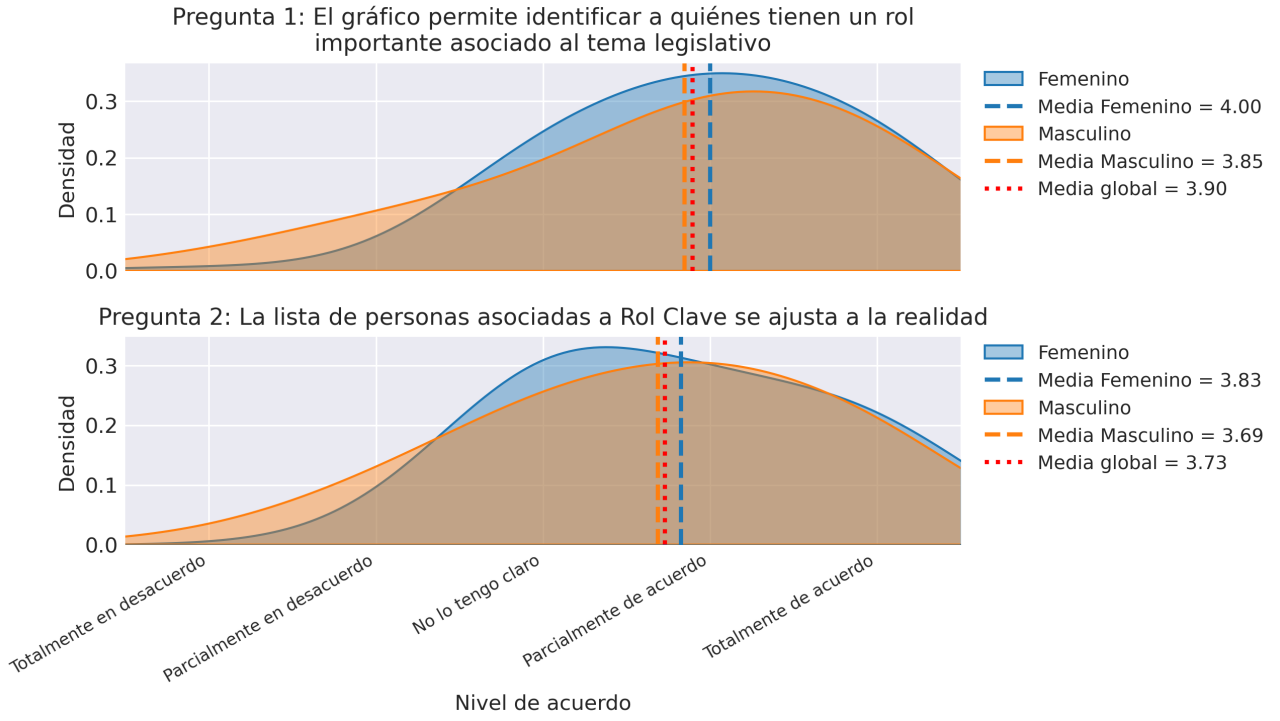

Figura 8.15: Distribución de las respuestas por sexo asociadas al instrumento 3

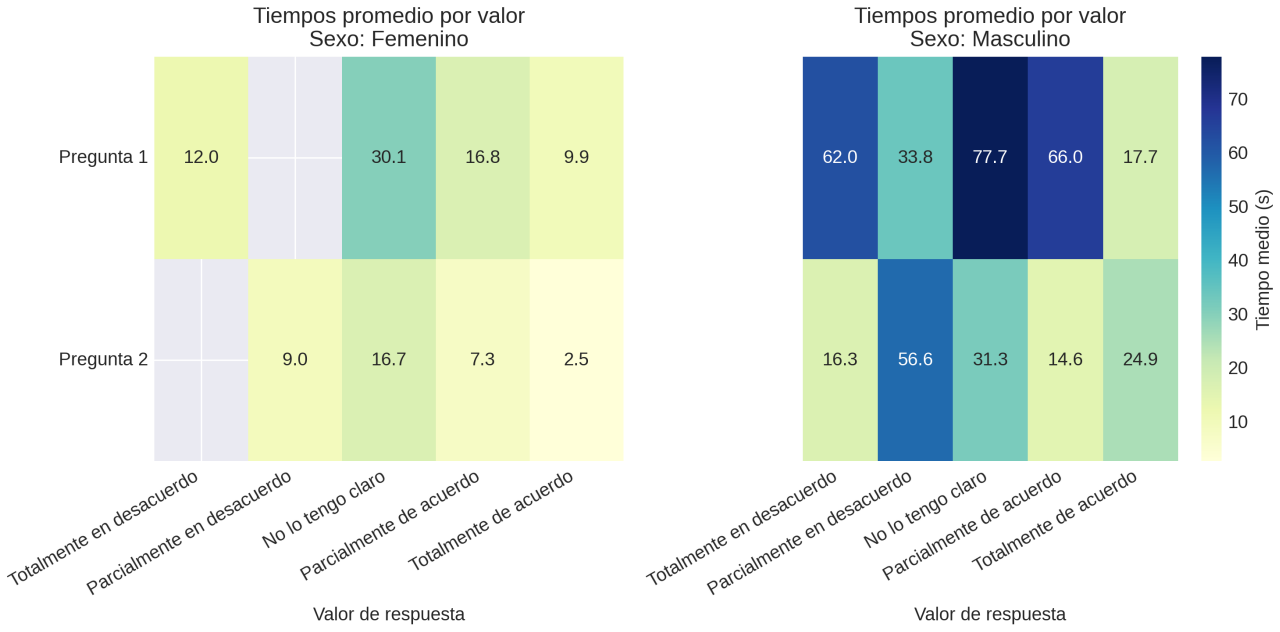

Figura 8.16: Tiempos utilizados por sexo para el instrumento 3 por valor de respuesta

8.5.3 Análisis agregado por profesión

Al igual que para los otros instrumentos, la figura 8.17 presenta un histograma por tipo de pregunta asociado al instrumento 3, donde se muestra la distribución de las respuestas obtenidas, diferenciadas según el profesion de las personas participantes.

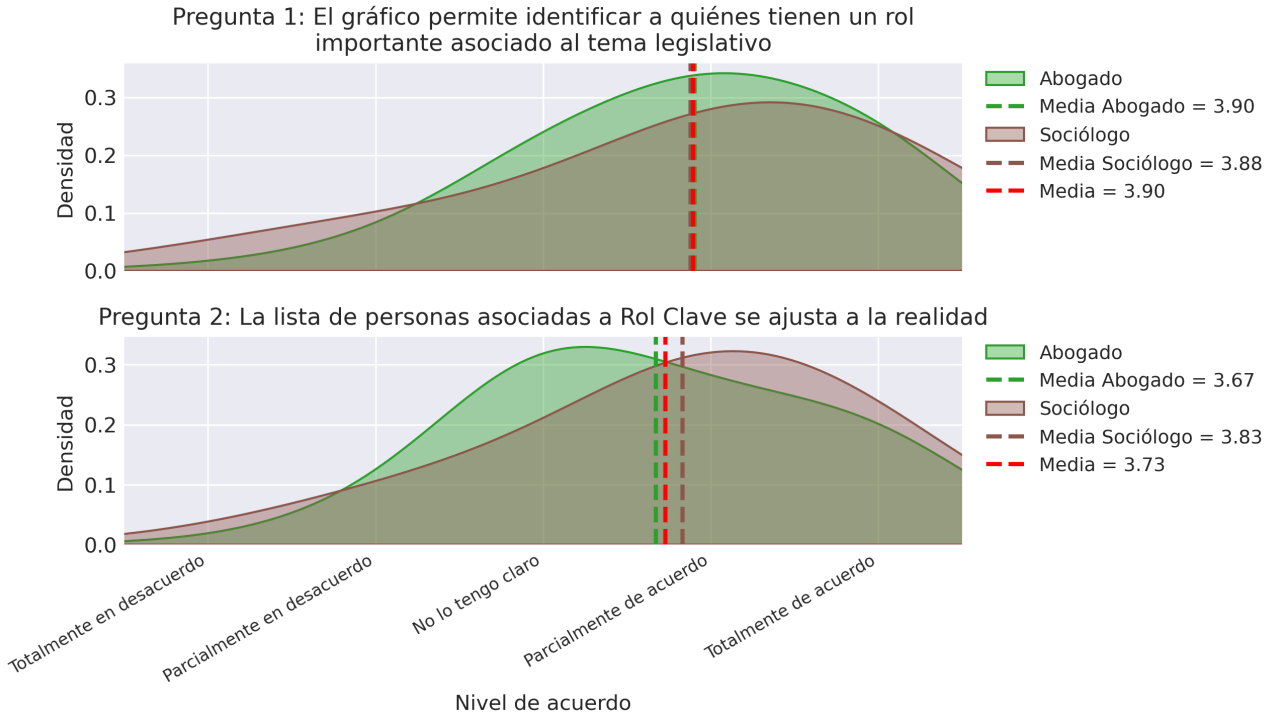

Figura 8.17: Distribución de las respuestas por área del conocimiento para el instrumento 3

De la misma forma, los *heatmap* de la figura 8.18, presentan los de tiempos de respuestas por profesión, divididos por tipo de pregunta y por valor de respuesta.

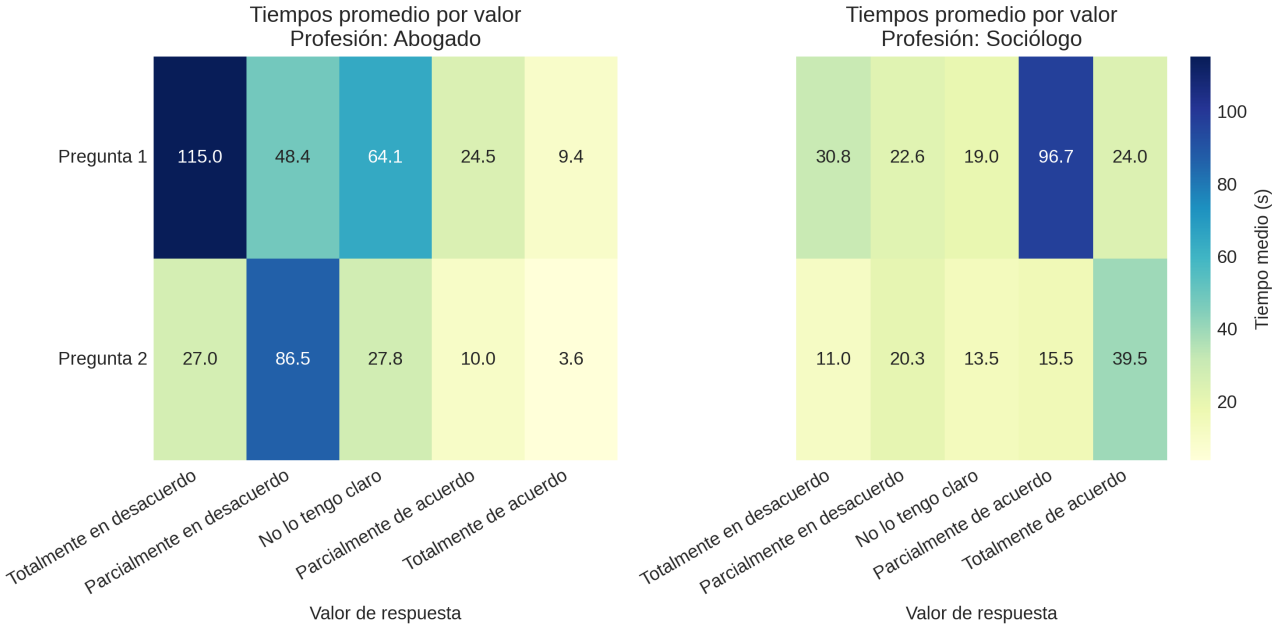

Figura 8.18: Tiempos utilizados por área del conocimiento en instrumento 3 por valor de respuesta

8.6 Integración y análisis comparativo de los instrumentos de evaluación

8.6.1 Variación de las respuestas

Los gráficos de la figura 8.19 muestran por una parte (gráfico superior) los coeficientes de variación calculados para las respuestas en cada instrumento, como también los valores promedio de respuesta (gráfico inferior). Al ver los gráficos, es posible visualizar que existe un índice de variación menor en aquellas preguntas con promedios de respuesta más altos, y de la misma forma, en aquellas preguntas con un promedio de respuesta más bajo (asociadas al Instrumento 3) se muestra un coeficioente de variación más alto.

Es importante recalcar que se presentan los valores promedio asociados a las respuestas, solo como referencia, para hacer un paralelo a los coeficientes de variación. Es importante destacar que si bien los valores promedio de respuesta se presentan en el gráfico en escala numérica con valores que varían entre 1 y 5, en la práctica lo que se mide es una percepción de índole cualitativa y aunque ordinal, categórica, la cual en rigor no puede ser expresada directamente como intervalos de exactamente igual magnitud.

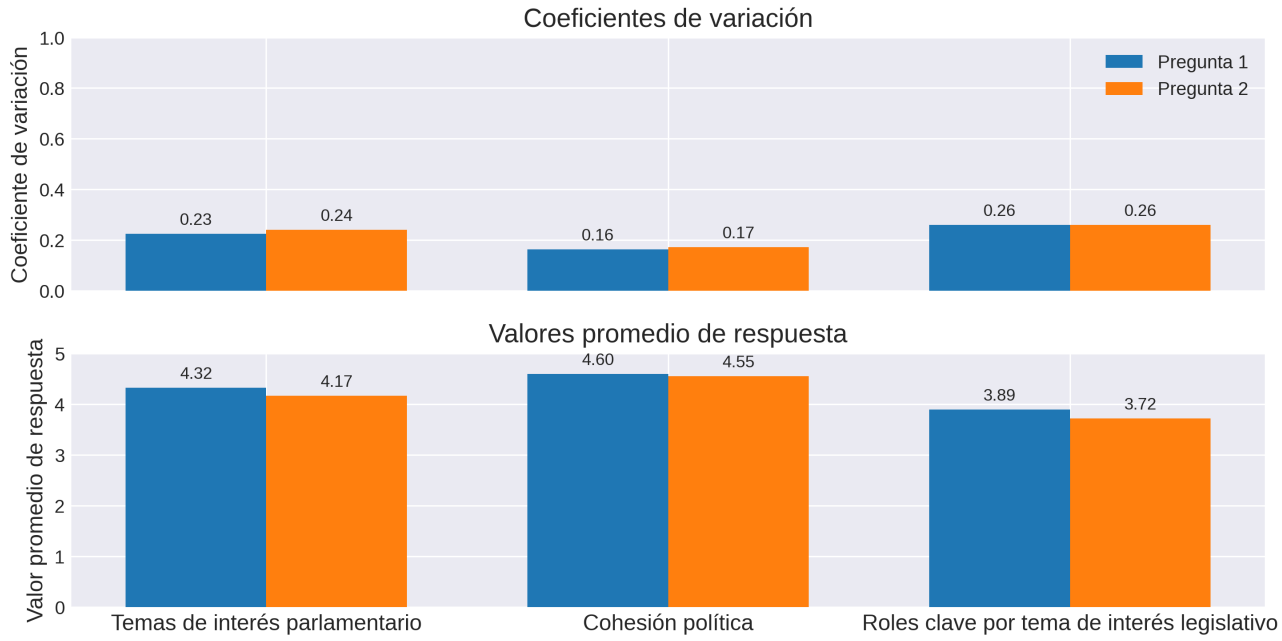

Figura 8.19: Coeficientes de variación y promedios de respuesta para cada instrumento

Si bien es posible establecer rangos para la interpretación del coeficiente de variación, estos pueden varíar de acuerdo al campo de estudio y al problema. Sin perjuidio de eso, el coeficiente de variación permite comparar con criterios homogéneos el grado de dispersión de cada pregunta (e instrumento) relativa a su propia media, lo cual facilita identificar cuál presenta más heterogeneidad en las respuestas y, por tanto, podría requerir un análisis más detallado de su fiabilidad o validez.

8.6.2 Consistencia de los instrumentos

La figura 8.20 presenta los valores del coeficiente alfa de Cronbach obtenidos para cada par de preguntas correspondientes a los instrumentos aplicados. Si bien no existe una única escala universalmente consensuada para su interpretación, el coeficiente se define en el rango [0,1], donde valores más cercanos a uno indican mayor consistencia interna. En este contexto, los tres instrumentos evaluados muestran niveles de fiabilidad que oscilan entre buenos ($\alpha > 0,80$) y excelentes ($\alpha > 0,90$), según los umbrales comúnmente aceptados en la literatura [George and Mallery, 2016]. Estos resultados reflejan una elevada correlación entre las respuestas de los ítems de cada instrumento, lo que respalda su coherencia interna y sugiere que efectivamente abordan aspectos relacionados de un mismo elemento. A su vez, el hecho de que los coeficientes no se aproximen de forma extrema al valor 1 permite asumir que existe un grado razonable de diferenciación entre las preguntas, lo cual resulta deseable cuando se busca capturar dimensiones complementarias dentro de un mismo dominio conceptual sin incurrir en redundancia.

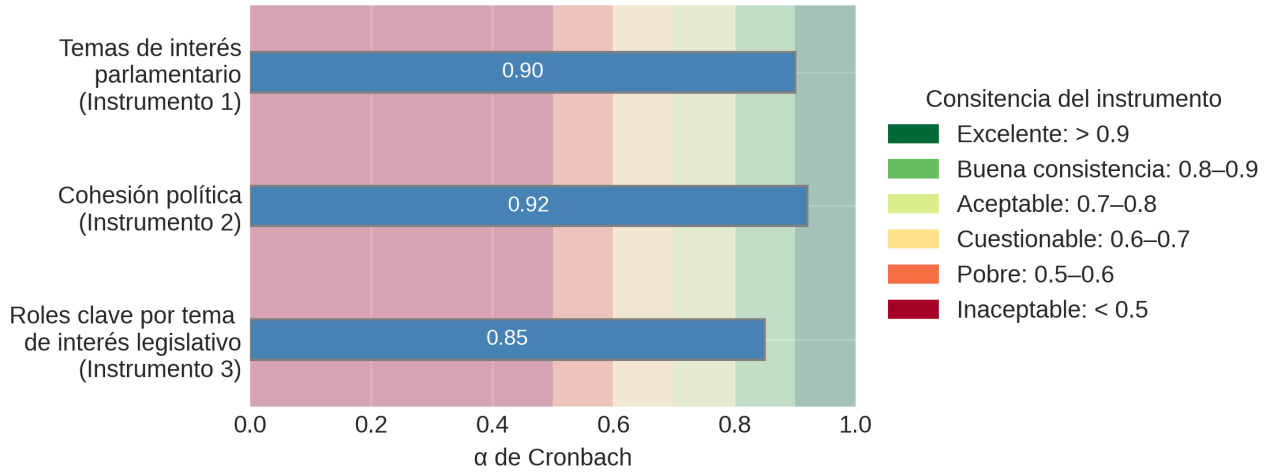

Figura 8.20: Coeficientes de alfa de Cronbach para cada instrumento

8.6.3 Tiempos de las respuestas

La figura 8.21 muestra los gráficos de cajas que presentan las medidas de tendencia central asociadas a los tiempos de respuesta para cada tipo de instrumento. En ellos es posible observar a primera vista que, en general, el instrumento 1 (temas de interés parlamentario) es el que necesita un menor tiempo de respuesta en comparación a los otros dos instrumentos, que poseen gráficas similares. Un gráfico de tiempos totales visualizando los valores atípicos se presenta en el anexo G

8.6.4 Correlación tiempo - valor de las respuestas

La tabla 8.5 muestra los coeficientes de correlación de Spearman entre tiempos de respuesta y valor de la respuesta por cada una de las preguntas y segmentado por tipo de instrumento. En general, se verifica que existe una correlación negativa leve entre el tiempo de respuesta y el valor de la respuesta.

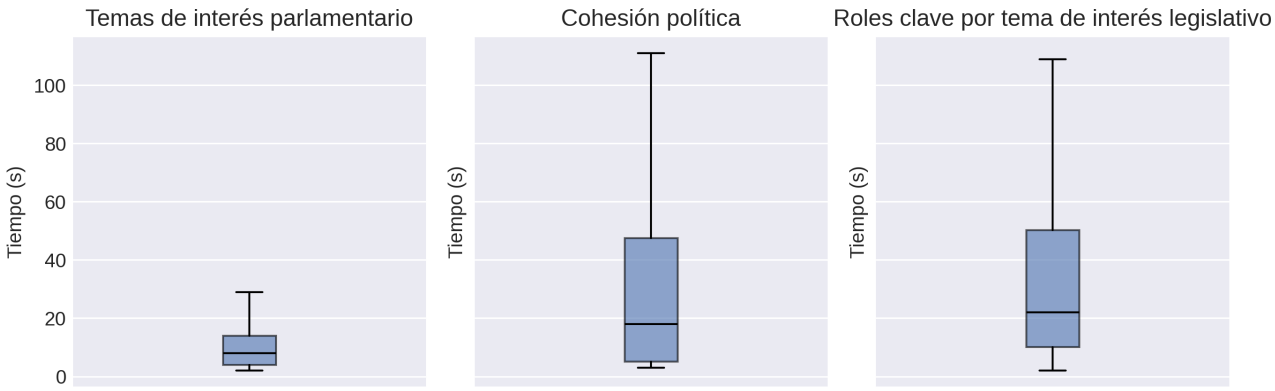

Figura 8.21: Distribución de tiempos de respuesta por tipo de instrumento sin valores atípicos

	Pregunta 1	**Pregunta 2**
Temas de interés parlamentario	-0,197	-0,235
Cohesión política	-0,305	-0,333
Roles clave por tema de interés legislativo	-0,256	-0,345

Tabla 8.5: Correlaciones entre tiempos de respuesta y valoración

Capítulo 9

Discusión de los resultados

9.1 Introducción

El presente capítulo corresponde a la fase de discusión del trabajo, donde se analizan y contrastan los resultados obtenidos a lo largo de la investigación. Particularmente, se pondrá énfasis en los hallazgos empíricos derivados de la aplicación de los tres instrumentos diseñados, abordándolos de manera individual para explorar sus aportes específicos, así como también de una forma conjunta tal que permita identificar patrones comunes y aspectos complementarios de cada uno en torno a la hipótesis. En ese sentido, la discusión se orientará a examinar todos aquellos aspectos que permitan responder las preguntas de investigación formuladas en el Capítulo 3 para validar la hipótesis, evaluando la coherencia de los resultados obtenidos respecto a los esperados.

Junto con ello, este capítulo incorpora elementos clave de las publicaciones realizadas durante el desarrollo de la tesis, así como un contraste del trabajo desarrollado con los trabajos relacionados revisados en la sección de estado del arte. Esta integración busca establecer un diálogo crítico entre los resultados obtenidos y la literatura especializada, resaltando tanto los aportes originales del presente estudio como sus puntos de convergencia con investigaciones previas.

Finalmente, a partir del análisis desarrollado en esta sección, se extraerán las principales conclusiones del trabajo, sentando las bases para la reflexión final y la propuesta de futuras líneas de investigación.

9.2 Discusión de resultados experimentales

9.2.1 Discusión de resultados para el Instrumento 1

Aspectos generales

Este instrumento, en comparación con otros, recoge un mayor volumen de respuestas (770), aunque provienen de un número menor de expertos (9 de los 13 participantes). A nivel general, el *Panel de visualización de temas de interés parlamentario* muestra un buen rendimiento (con medias de respuesta Pregunta 1 = 4,32 y Pregunta 2 = 4,17), lo cual muestra una convergencia entre la opinión del GE con la predicción basada en Tecnologías Semánticas.

En la pregunta 1 (*El gráfico representa correctamente los intereses legislativos del parlamentario/a*), más del 80% de las respuestas fueron positivas (valoración > 4), con un 59,4%

indicando estar *"Totalmente de acuerdo"*, y solo un 5% de respuestas negativas.

Para la pregunta 2 (*El gráfico se ajusta a los intereses del sector político que el parlamentario/a representa*, asociada al sector político del parlamentario), un 71,8% obtuvo una valoración positiva, y un 52,1% alcanzó la máxima valoración. Si bien las respuestas negativas disminuyen a un 3,9%, se observa un aumento del 10% en la incertidumbre del usuario experto, pasando de un 14,3% en la pregunta 1 a un 24,3% en la pregunta 2. Esta diferencia podría deberse a que algunos parlamentarios representan intereses que trascienden su sector político, que el concepto *"sector político"* resulta poco claro, o bien que los usuarios expertos encuentran más difícil juzgar si los intereses declarados reflejan fielmente al sector político del parlamentario.

Análisis de consistencia de las respuestas

Respecto a la consistencia de las valoraciones, el coeficiente de variación se sitúa entre un 22% y 24% en ambas preguntas, mientras que los errores declarados (casos presentados considerados como incorrectos por el GE) no superan el 5%. Esta variabilidad está asociada a las respuestas con valoración intermedia o indicios de duda, y a la dispersión entre quienes se mostraron *"parcialmente"* o *"totalmente de acuerdo"*. En la misma línea, el coeficiente alfa de Cronbach ($\alpha = 0{,}9$) muestra que ambas preguntas recogen aspectos relacionados pero no idénticos, por lo cual se considera que el planteamiento del instrumento es correcto.

Análisis de tiempo de respuestas

En cuanto a los tiempos de respuesta, se observa que las respuestas categóricas (*"Totalmente de acuerdo"* o *"Totalmente en desacuerdo"*) y las que expresan desconocimiento tienden a registrar tiempos más breves. Por el contrario, cuando el evaluador duda, como ocurre en las respuestas *"Parcialmente de acuerdo"* o *"Parcialmente en desacuerdo"*, los tiempos de respuesta muestran mayor dispersión.

Por su parte, el análisis de correlación de Spearman revela una débil correlación inversa entre los valores y tiempos de respuesta, lo cual sugiere que, en casos de mayor reflexión (tiempos más largos), es más probable que la valoración final sea negativa en lugar de positiva.

Análisis agregados por grupo

Al realizar un análisis por sexo de los expertos, se observa que las expertas, en su mayoría, declaran no tener claridad respecto a la preferencia de temas de interés parlamentario, presentando medias de P1 = 3,45 y P2 = 3,16. En contraste, los expertos masculinos entregan medias de valoración más altas (P1 = 4,42 y P2 = 4,29), lo que además revela una brecha que se amplía particularmente en la pregunta 2.

En cuanto al análisis de los tiempos de respuesta según sexo, se advierte que las expertas presentan tiempos relativamente homogéneos para casi todos los valores de la escala de Likert, aunque replican la tendencia general: las respuestas "parcialmente" (valores 2 y 4) requieren más tiempo que las respuestas extremas. En el caso de los expertos masculinos, los tiempos de respuesta son, en general, mayores (manteniendo el mismo patrón observado en el análisis global), salvo en las respuestas *Totalmente de acuerdo*, donde los tiempos de respuesta son notablemente menores.

En una línea paralela, al revisar el análisis agregado por profesión, se observa que los abogados reportan una mayor tendencia a responder *No lo tengo claro* respecto a los intereses legislativos específicos de los parlamentarios, tanto a nivel individual como de sector político

(medias P1 = 3,48 y P2 = 3,21), en comparación con los sociólogos (P1 = 4,69 y P2 = 4,60). Esta diferencia podría explicar en parte la menor participación de expertos en este tipo de pregunta, ya que una mayor tendencia a responder *"No lo tengo claro"* puede ser una causa válida para evitar responder.

Respecto al análisis de tiempos de respuesta por profesión, los abogados registran, en promedio, mayores tiempos de respuesta que los sociólogos. Esta diferencia puede explicarse, al menos en parte, por el mayor grado de duda que manifiestan los abogados en comparación con los sociólogos.

Es posible inferir que, dentro del GE, los sociólogos presentan una mayor experticia en la identificación de los intereses parlamentarios en comparación con los abogados. Esto podría explicarse, principalmente, porque su quehacer profesional suele estar más vinculado al trabajo directo de asesoría parlamentaria, lo que implica una relación más estrecha con los parlamentarios, a diferencia de los abogados, cuyo enfoque se orienta mayormente al análisis legislativo desde una perspectiva general.

A su vez, desde el punto de vista de la implementación, se constata que el grupo de sociólogos otorga una alta valoración a la percepción de los resultados obtenidos mediante este instrumento, lo que respalda su validez y confirma que cumple con el objetivo planteado.

Contraste de resultados con estado del arte

Al comparar los resultados obtenidos mediante el instrumento, con las experiencias descritas en el estado del arte, sección 5.4.3, se verifica que solo la experiencia del Congreso de Taiwan [Lin et al., 2015] se asimila técnicamente al enfoque de clasificación planteado en este trabajo, aunque con dos diferencias fundamentales: 1) el método para definir las categorías, ya que en el caso descrito en establecen las categorías mediante una técnica de clustering más datos de entrenamiento supervisados y para el caso de la tesis categorías son definidas con base en las comisiones parlamentarias permanentes más etiquetamiento manual; y 2) el proceso de normalización aplicado con base en un ponderador por categoría temática desarrollado en nuestro instrumento, el cual permite un ajuste más realista a los intereses parlamentarios, corrigiendo las distorsiones impuestas por la agenda legislativa del ejecutivo, lo cual no existe en ninguno de los trabajos planteados. También destacar que en el caso del Congreso de Taiwan, la evaluación de los expertos fue de 0,84 sobre 1 en cuanto a acierto de temas por legislador, en comparación con la nuestra de 4,32 sobre 5 (lo que se asemeja a 0,86 sobre 1) para la pregunta 1 del instrumento sobre acierto de temas por legislador, lo cual refuerza y valida la decisión del uso del proceso de normalización.

9.2.2 Discusión de resultados para el Instrumento 2

Aspectos generales

Este instrumento obtuvo un total de respuestas (379), realizadas por 10 de los 13 expertos participantes.

A nivel general, el *Panel de visualización e indicadores sobre cohesión política* presenta un rendimiento excelente, con medias de respuesta de 4,60 para la Pregunta 1 y 4,55 para la Pregunta 2, los cuales son muy cercanos al máximo.

En la Pregunta 1 (*Creo que la clasificación del proyecto en la categoría detectada es coherente con el proyecto y la información presentada*), más del 91% de las respuestas fueron

positivas (valoración > 4), con un 72,6% indicando estar *"Totalmente de acuerdo"*, y solo un 3,5% correspondiendo a valoraciones negativas.

En el caso de la Pregunta 2 (*De acuerdo a los datos, los elementos gráficos representan correctamente la cohesión política de los parlamentarios/as y partidos respecto al tema principal del proyecto de ley*), el 86,2% de las respuestas obtuvo una valoración positiva, con un 71,2% alcanzando la máxima puntuación.

Si bien en ambos casos las percepciones negativas son mínimas (menores al 3,3%), se observa, al igual que en el instrumento 1, un leve incremento en la incertidumbre del usuario experto al evaluar ambas preguntas, pasando de un 5,3% en la Pregunta 1 a un 11,3% en la Pregunta 2.

Análisis de consistencia de las respuestas

Respecto a la consistencia de las valoraciones, el coeficiente de variación se sitúa entre un 16,38% y 17,28%, lo cual es bajo para ambas preguntas, mientras que los errores declarados (casos presentados considerados como incorrectos por el GE) no superan el 3,2% y 2,4% respectivamente. En cualquier caso, la variabilidad presentada existe principalmente entre quienes se mostraron *"parcialmente"* o *"totalmente de acuerdo"*, lo cual es positivo para el experimento. En la misma línea, el coeficiente alfa de Cronbach ($\alpha = 0{,}92$) muestra una consistencia excelente, indicando que ambas preguntas recogen aspectos relacionados pero no idénticos, por lo cual se considera de igual manera que el planteamiento del instrumento es correcto.

Análisis de tiempo de respuestas

En cuanto a las distribuciones de tiempos de respuesta, en ambas preguntas se observa un patrón decreciente en la medida en que las respuestas son más positivas (*"Totalmente de acuerdo"*). Esto lo refuerza el análisis de correlación de Spearman, que revela una correlación inversa entre los valores y tiempos de respuesta en ambas preguntas (con valores -0,305 y -0,333; mayores al instrumento 1), lo cual reafirma que, en casos de mayor reflexión (tiempos más largos), es más probable que la valoración final sea negativa en lugar de positiva.

Análisis agregados por grupo

Al realizar un análisis por sexo de los expertos que respondieron, se observa que no existen diferencias fundamentales entre ambos grupos, con medias para el sexo Femenino de P1 = 4,51 y P2 = 4,46 y para el sexo Masculino de P1=4,63 y P2=4,58.

En cuanto al análisis de los tiempos de respuesta según sexo, se advierte que en general las expertas requieren menos tiempos para responder que expertos masculinos. Por otro lado, las expertas presentan un valor alto en el promedio de tiempo cuando responden *"Parcialmente de acuerdo"* (valor 4) para la pregunta 1, lo cual puede estar influenciado por valores atípicos. A diferencia de ellas, los expertos presentan tiempos medios equilibrados para responder todas las categorías, con tiempos mayores en las opciones distintas a la valoración máxima, comportamiento que se replica tanto en la pregunta 1 como en la pregunta 2.

Así mismo, al revisar el análisis agregado por profesión, se observa que ambos grupos se acercan a la tendencia central en ambas preguntas (medias P1 = 4,60 y P2 = 4,55), siendo los abogados (media P1 = 4,28 y P2 = 4,14) quienes reportan una conformidad más baja comparativamente a los sociólogos (media P1 = 4,90 y P2 = 4,94), sin perjuicio de que ambos grupos presenten medias > 4.

Respecto los tiempos de respuesta por profesión, los abogados nuevamente registran valores promedio mayores que los sociólogos, a excepción de la valoración *"Parcialmente de acuerdo"* que presenta valores altos tanto en la pregunta 1 y 2, lo cual acorde a las medidas de tendencia central generales Q3 y desviación estándar, indica que se trata de valores influenciados por valores atípicos.

Contraste de resultados con estado del arte

Al comparar los resultados obtenidos mediante la aplicación del instrumento, con las diversas experiencias descritas en el estado del arte, sección 5.4.1, es posible indicar es que solo un caso de los descritos utiliza un mecanismo distinto al análisis del texto, el cual se basa en análisis de datos de coautoría [GovTrack.us, 2013], pero ninguno utiliza votaciones extraídas desde el texto ni desde otros servicios. Esto permite indicar que el método descrito en este trabajo es innovador y al mismo tiempo, que dada la evaluación, el GE otorga una alta valoración a la percepción de los resultados obtenidos mediante este instrumento, lo que respalda su validez y confirma que cumple con el objetivo planteado. Esta valoración va de la mano con los resultados presentados y validados en los artículos [Cifuentes-Silva et al., 2023, Cifuentes-Silva et al., 2024] donde se presenta este instrumento aplicado al conjunto completo de votaciones de proyectos de ley chilenos.

9.2.3 Discusión de resultados para el Instrumento 3

Aspectos generales

Este instrumento obtuvo un total de respuestas (296), realizadas por los 13 expertos participantes. A nivel general, el *Visualizador de rol clave en el contexto de un tema de interés legislativo* presenta un rendimiento sobre la media, con medias de respuesta de 3,89 para la Pregunta 1 y 3,72 para la Pregunta 2. Estos valores se ubican entre el rango de *No lo tengo claro* y *Parcialmente de acuerdo*.

En la Pregunta 1 (*El gráfico permite identificar a quiénes tienen un rol importante asociado al tema legislativo*), un 70,3% de las respuestas fueron positivas (valoración > 4), un 19,6% indican *No lo tengo claro*, y un 10,8% correspondiendo a valoraciones negativas.

En el caso de la Pregunta 2 (*La lista de personas asociadas a Rol Clave se ajusta a la realidad*), solo un 57,8% de las respuestas obtuvo una valoración positiva, mientras que un 32,8% indican *No lo tengo claro*, con lo que un 9,4% corresponde a valoraciones negativas.

En este contexto, si bien las valoraciones positivas son mayoría, el alto número de valoraciones asociadas a la categoría *No lo tengo claro* ajustan a la baja la percepción de acierto del instrumento.

Análisis de consistencia de las respuestas

Respecto a la consistencia de las valoraciones, los coeficiente de variación se sitúan en un 26,05% y 26,04% respectivamente, lo cual es un valor alto para ambas preguntas, y se condice con la distribución de respuestas. Respecto al coeficiente alfa de Cronbach ($\alpha = 0{,}85$) muestra una consistencia buena, aunque más baja que los otros dos instrumentos. Esto puede deberse a que casi un tercio de las respuestas de la pregunta 2 fueron respondidas en la categoría *No lo tengo claro*.

Análisis de tiempo de respuestas

En cuanto a las distribuciones de tiempos de respuesta, se presenta un escenario similar al del instrumento 2, en que el tiempo de respuesta crece de forma inversa a la valoración, es decir, el análisis de correlación de Spearman revela una correlación inversa entre los valores y tiempos de respuesta en ambas preguntas (con valores -0,256 y -0,345).

Análisis agregados por grupo

A nivel de análisis por sexo, los resultados no muestran diferencias relevantes en ambas preguntas(menos de 0,1 puntos de diferencia entre sexos). En cuanto al tiempo tomado en responder, se observa que el grupo de expertos de sexo masculino presentan tiempos mayores de respuesta, y se verifica que en ambos grupos lo que toma más tiempo es la respuesta *No lo tengo claro*.

Un escenario similar es el análisis agregado por profesion. A partir de los datos se verifica que abogados y sociólogos coinciden en la valoración de ambas preguntas con diferencias mínimas en la percepción de los resultados: abogados (promedios P1=3,90 y P2=3,67) y sociólogos (promedios P1=3,88 y P2=3,83). En cuanto a los tiempos de respuesta, es posible verificar que en general los abogados presentan tiempos de respuesta más altos, mientras que los valores muy por sobre Q3 (para P1=31,25 s y para P2=16 s) se consideran como atípicos.

Contraste de resultados con estado del arte

Al comparar los resultados obtenidos mediante la aplicación del instrumento, con las diversas experiencias descritas en el estado del arte, sección 5.4.2, es posible verificar que el uso de SNA es una técnica ampliamente utilizada para la descripción de la composición estructural de redes en el contexto político, tanto en análisis de datos de coautoría de proyectos de ley, como de datos provenientes de redes sociales (por ejemplo Twitter).

En particular, el instrumento presentado se acerca en parte a la experiencia descrita para la Cámara de Representantes de Estados Unidos [Sotoudeh et al., 2024], donde la posición de un legislador en la red se utiliza para predecir la aprobación de proyectos, y también se aproxima a nuestro enfoque el estudio sobre el Congreso Brasileño [Nery and Mueller, 2022], en que utilizando métricas de alineamiento complementarias a métricas SNA se utilizan para caracterizar la influencia de las bancadas. No obstante, ninguno de los casos presentados presenta un análisis temático (segmentación de los documentos por tema) al mismo tiempo que se identifican roles clave como los planteados en el instrumento (líderes intra-grupo e intermediadores), por lo cual se puede establecer que el diseño del instrumento es original en su objetivo.

9.2.4 Reflexión final de la fase de experimentación

A partir de todo el análisis realizado en la fase experimental, es posible indicar lo siguiente:

- Los tres instrumentos propuestos entregan soluciones viables validadas por el GE, lo que permite confirmar la factibilidad de realizar análisis automatizados útiles en el ámbito político-legislativo mediante el uso de tecnologías semánticas.

- El Instrumento 1, asociado a la detección de temas de interés legislativo, muestra un buen rendimiento, aunque podría ser susceptible de mejora mediante ajustes menores asociados al concepto de *sector político* (P2). En términos generales, su diseño resulta validado,

tanto por la propuesta de una jerarquía conceptual basada en comisiones legislativas como por la normalización de indicadores que permiten visualizar comparativamente temas y actores parlamentarios. Al mismo tiempo, de los tres instrumentos es el que presenta una interfaz más intuitiva y simplificada, lo que se manifiesta en menores tiempos de respuesta comparativamente a los otros dos instrumentos. En este contexto, la pregunta de investigación (*RQ1*) puede responderse afirmativamente, dado que tanto los datos procesados como el instrumento permiten representar de manera satisfactoria los escenarios reales planteados, ajustándose a la realidad.

- El Instrumento 2, asociado a la identificación de cohesión política, alcanzó el nivel más alto de aceptación y fiabilidad. Este resultado se vincula con el mayor grado de desarrollo conceptual y validación preliminar en relación con otros estudios previos. Al mismo tiempo, los resultados positivos del experimento en torno a este experimento validan totalmente el diseño conceptual en torno a su funcionamiento, y permiten responder de forma positiva la pregunta de investigación asociada a él (*RQ2*).

- El Instrumento 3, asociado a la detección de roles clave, si bien presenta un desempeño funcional y es aceptado en términos generales, requiere mejoras en claridad y usabilidad para ser considerado altamente fiable. Esta necesidad queda en evidencia por el alto porcentaje de respuestas asociadas a la opción *No lo tengo claro* durante la fase experimental, y por una percepción homogénea y generalizada entre los distintos sub grupos de evaluadores expertos (considerando sexo, expertiz profesional y tiempos de respuesta). Con todo esto, dado que aun así los resultados son positivos, la pregunta de investigación (*RQ3*) también puede responderse de forma afirmativa, aunque considerando mejoras que permitan esclarecer el alto porcentaje de respuestas bajo la categoría *"No lo tengo claro"*.

- Desde el punto de vista del conjunto de datos utilizado, construido a partir de datos abiertos enlazados y textos legislativos procesados mediante un flujo automatizado que aplica tecnologías semánticas para su enriquecimiento, se puede afirmar que provee de forma óptima los insumos necesarios para construir aplicaciones que permiten responder preguntas de ámbito poítico-legislativo, constituyendose como un el eslabón crítico para el desarrollo de cualquiera de los instrumentos expuestos.

- Sobre el análisis de las valoraciones en los instrumentos de evaluación, de manera transversal se constata empíricamente que las puntuaciones bajas durante la fase experimental requieren más tiempo de respuesta. Este hallazgo es consistente con la literatura, particularmente con la *Teoría prospectiva* [Kai-Ineman et al., 1979], que plantea que los usuarios invierten más tiempo al emitir valoraciones negativas para evitar errores de tipo *"falso negativo"*, y con el concepto de *accountability* en psicología[Lerner and Tetlock, 1999] (hacerse consciente del impacto de las propias decisiones), que postula que criticar implica construir internamente una justificación más elaborada.

- En consecuencia, para reducir la frecuencia de valoraciones bajas, en especial aquellas asociadas a la categoría *No lo tengo claro*, se propone mejorar los instrumentos (en especial el de roles clave) tanto en su representación gráfica (haciéndolos más descriptivos) como en su lógica interna (mediante la revisión, reemplazo o complemento de métricas e indicadores).

- Finalmente, con el objetivo de fortalecer la transparencia y robustez de la fase experimental, se plantea durante la ejecución de experimentos futuros, la incorporación de un conjunto de preguntas ficticias diseñadas como control, con el fin de medir el comportamiento de los evaluadores ante estímulos artificiales. En particular, se podría haber desarrollado el experimento incluyendo una proporción de preguntas con falsos positivos, idealmente en torno a la mitad del conjunto total, para validar con mayor precisión los posibles sesgos cognitivos o interpretativos presentes en las respuestas de los usuarios expertos.

9.3 Experiencias en el uso del marco de trabajo

A partir de los artículos publicado durante la investigación sobre extracción de contenido desde documentos legislativos basado en tecnologías de Web Semántica [Cifuentes-Silva and Labra Gayo, 2019] y sobre experiencias en el uso de IA en la BCN [Cifuentes-Silva et al., 2025], es posible identificar una serie de hallazgos relevantes en cuanto a aspectos operativos, modelado semántico y arquitectura tecnológica:

9.3.1 Reducción de tiempos de procesamiento

- La incorporación de herramientas basadas en Tecnologías Semánticas, especialmente el marcador automático que transforma texto plano en documentos XML estructurados en AKN, permitió reducir en hasta un 65% los tiempos requeridos para elaborar productos como la Historia de la Ley, en comparación con el estudio de línea base [Palmirani and Vitali, 2012].
- Esta mejora se traduce también en un mayor número de marcas por documento, lo que reduce la carga de trabajo manual y aumenta la cantidad de metadatos disponibles. Al mismo tiempo, las estadísticas de uso muestran una disminución sostenida en los tiempos promedio de procesamiento, atribuible tanto al perfeccionamiento de las herramientas como al aprendizaje progresivo de los analistas encargados del marcaje [Cifuentes-Silva and Labra Gayo,

9.3.2 Evaluación de arquitecturas para la entrega de contenidos legislativos

Dado el volumen y granularidad de los productos generados basados en texto procesado bajo Tecnologías Semánticas, dentro de BCN y en el contexto de los proyectos *Historia de la Ley* y *Labor Parlamentaria*, se exploraron distintos enfoques arquitectónicos para el desarrollo de portales de entrega de contenidos (*Content Delivery*) que utilizaban tecnologías de Web Semántica para su funcionamiento. La idea inicial fue basar gran parte de la lógica de programación de los sitios de consulta orientados a la ciudadanía en estas tecnologías a fin de maximizar su uso y adopción. Los distintos enfoques probados y sus resultados fueron los siguientes [Cifuentes-Silva and Labra Gayo, 2019]:

1. *Consultas basadas en razonamiento lógico*: Para acceder a los datos relacionados con intervenciones parlamentarias y tramitación legislativa, se consideró la implementación de consultas SPARQL sobre un dataset modelado por una ontología con varias clases y propiedades que extendían de otras (rdfs:subPropertyOf, rdfs:subClassOf), definiendo a su vez dominios y rangos (rdfs:domain, rdfs:range) en las propiedades, lo cual en la práctica se

traducía en que el dataset contaba con menos triples RDF concretas, pero a la vez se debían inferir muchas otras. Este enfoque buscaba generar un conjunto mínimo de triples en el triplestore RDF, delegando la cobertura de los datos a inferencias dinámicas, ejecutadas en tiempo real por el razonador de la base de datos Virtuoso. Sin embargo, las pruebas realizadas evidenciaron altos tiempos de respuesta e inestabilidad al ejecutar consultas complejas, lo cual es una prueba empírica de los altos costos computacionales asociados a consultas SPARQL complejas [Pérez et al., 2009]. Como resultado, se descartó este enfoque minimalista basado en generación reducida de datos más razonamiento ontológico dinámico.

2. *Consultas sobre materialización previa de inferencias*: Se optó por generar y almacenar anticipadamente las tripletas RDF que anteriormente eran inferidas dinámicamente, con el objetivo de reducir la sobrecarga computacional y temporal asociada al razonamiento en tiempo de consulta. Esta estrategia permitió que las consultas finalizaran exitosamente; sin embargo, los tiempos de respuesta continuaron siendo insuficientes para un entorno de producción, particularmente en el procesamiento de documentos de gran extensión (más de 10.000 páginas). Además, durante esta etapa se identificaron limitaciones importantes en la aplicación de filtros de búsqueda (tanto en atributos como textuales) al emplear exclusivamente tecnologías de Web Semántica. En consecuencia, este enfoque también fue descartado como solución para la entrega de contenido en sistemas orientados al usuario final.

3. *Consulta bajo modelo híbrido*: A partir de las experiencias previas, se optó por una arquitectura híbrida que combina tecnologías de Web Semántica (RDF y SPARQL) para el acceso a datos estructurados, como proyectos de ley, personas, leyes o sus tramitaciones, aprovechando sus capacidades de interoperabilidad, mientras que el acceso a los textos completos de documentos y otros metadatos se realiza a través de una base de datos relacional complementada con índices de texto implementados en Apache Lucene o SolR. Esta solución permite mantener la interoperabilidad mediante el uso de URIs compartidas, al tiempo que garantiza un rendimiento eficiente en consultas sobre grandes volúmenes de información textual, con lo que esta arquitectura es la que actualmente se encuentra operativa en producción.

9.3.3 Análisis de la producción legislativa en años electorales

El análisis de los datos en el *triplestore* RDF publicado en `https://datos.bcn.cl` revela una tendencia sostenida al alza en la generación de triples derivados de documentos legislativos. Sin embargo, esta producción disminuye de forma importante en años electorales, lo que coincide con una menor frecuencia de sesiones del Congreso en dichos periodos [Cifuentes-Silva and Labra Gayo, 2019].

9.4 Otros hallazgos relevantes durante la investigación

9.4.1 Importancia de la validación en la calidad de los datos RDF

En el artículo sobre la visualización del presupuesto nacional publicado como datos abiertos enlazados [Cifuentes-Silva et al., 2020], se presenta un ejercicio de validación de datos RDF utilizando una herramienta para inferir *Shape Expressions* latentes [Fernández-Álvarez et al., 2018]

sobre una base de datos curada y en producción, generada a partir de otro organismo del estado. Este proceso permitió identificar anomalías de diversa índole, las cuales fueron posteriormente corregidas. La experiencia pone de relieve la relevancia de emplear tecnologías como Shape Expressions (ShEx) para asegurar la consistencia estructural y semántica de los conjuntos de datos abiertos, especialmente cuando se trata de información crítica como el presupuesto nacional; y al mismo tiempo advierte sobre lo cuidadoso que se debe ser a la hora de consumir otros datos, que en este caso también son provistos por instituciones de gobierno, para producir nuevos conjuntos de datos abiertos a través de canales propios.

9.4.2 Análisis sobre datos del Congreso Nacional chileno

Dentro de los análisis realizados, durante la investigación, fue posible identificar los siguientes hallazgos asociados al Congreso Nacional Chileno:

- En el trabajo sobre análisis de votaciones [Cifuentes-Silva et al., 2023], y en particular utilizando las métricas de alineamiento y polarización, fue posible identificar que en general los miembros del Senado presentan un comportamiento más disciplinado (mayor alineamiento) que los representantes de La cámara de Diputados y Diputadas, lo cual puede estar explicado por variables tales como un promedio mas alto de edad (lo que se muestra en el gráfico de la figura 2.1 del capítulo 2) o mayor experiencia política (deducida porque en la carrera parlamentaria, lo más frecuente es pasar de la cámara baja a la alta), entre otros. En la misma línea, se identificó que el Senado tiene un comportamiento menos polarizado que en la Cámra de Diputados

- En un trabajo posterior [Cifuentes-Silva et al., 2024], se demuestra cuantitativamente la factibilidad de utilizar los indicadores de alineamiento y polarización aplicados a las votaciones de proyectos de ley, como herramienta para mejorar la tasa de aprobación legislativa. En particular, se identifica un conjunto significativo de iniciativas (70,14% de las analizadas) que pese a contar con amplio consenso (aquellas clasificadas como de *"consenso técnico"* con baja polarización y alto alineamiento), permanecen en tramitación por un periodo excesivo de tiempo (promedio de 667,8 días). La implementación de un mecanismo que priorice este tipo de proyectos podría agilizar su tramitación, contribuyendo así a una mayor eficiencia legislativa y a mejorar la percepción ciudadana respecto al desempeño del Congreso Nacional.

Capítulo 10

Conclusiones y trabajo futuro

10.1 Conclusiones

La presente investigación tuvo por objetivo principal demostrar la viabilidad y fiabilidad del uso de Tecnologías Semánticas para la construcción de instrumentos que permitan representar y generar analisis político-legislativo de forma automatizada.

A partir de un enfoque metodológico que combinó la elaboración de un marco de trabajo, el desarrollo de herramientas tecnológicas y la validación experimental mediante un GE, se logró validar empíricamente que es posible generar soluciones automatizadas que dan respuestas fiables para el análisis, con base en la combinación de datos no estructurados con datos abiertos enlazados.

10.1.1 De los experimentos

Los tres instrumentos diseñados y evaluados ofrecieron evidencia clara para responder afirmativamente las preguntas de investigación planteadas. En particular:

El Instrumento 1, sobre *Temas de interés legislativo* de parlamentarios, logró un nivel de percepción de acierto alto por parte de los expertos evaluadores. Los resultados respaldan que la jerarquía de clasificación de intervenciones basada en comisiones legislativas, junto con visualizaciones de datos, permiten representar de forma adecuada las áreas de interés de cada parlamentario y su sector político, por lo cual el instrumento cumple satisfactoriamente con su propósito. De esta manera la pregunta de investigación **RQ1**: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuáles son los temas de mayor relevancia para un representante?*, se da por respondida de forma afirmativa.

El Instrumento 2, asociado al *Panel de visualización e indicadores sobre cohesión política*, fue el que obtuvo la valoración más alta en la percepción de precisión y coherencia entre el GE, validando tanto la cohesión política de parlamentarios y partidos políticos respecto al tema principal de un proyecto de ley, como a la categoría detectada con base en la clasificación predefinida. También se debe destacar que los resultados empíricos fueron consistentes con los hallazgos expuestos en publicaciones científicas realizadas durante esta tesis, confirmando la validez conceptual del enfoque adoptado. De esta manera, la pregunta de investigación **RQ2**: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas cuál es el nivel de cohesión política de un grupo frente a un tema particular?*, se da por respondida de forma afirmativa.

El Instrumento 3, sobre *Visualización de rol clave en el contexto de un tema de interés legislativo*, presentó resultados positivos, aunque con una percepción de claridad menor en comparación a los otros dos instrumentos. Si bien la mayoría de los expertos valoraron favorablemente su utilidad, se detectó un número elevado de respuestas en la categoría *"No lo tengo claro"*, lo que indica la necesidad de ajustes en la presentación gráfica y en la lógica de determinación de roles. No obstante lo anterior, y teniendo en cuenta consideraciones de mejora sobre este instrumento, la pregunta de investigación **RQ3**: *¿Es posible determinar con base en procesamiento automatizado de datos basado en tecnologías semánticas quién cumple un rol clave en el contexto de un tema específico?*, se da por respondida de forma afirmativa.

En general y asociado a los tres instrumentos, los análisis de las respuestas realizadas por el GE evidenciaron una correlación inversa entre el tiempo de respuesta y las valoraciones asignadas, lo que se traduce a que respuestas de baja valoración o inciertas requieren más tiempo de reflexión por parte de los expertos, lo cual es coherente con lo que señala la literatura especializada. Del mismo modo, en la revisión del estado del arte no se detectaron experiencias equivalentes a las planteadas por esta investigación, por lo cual se establece que todos los experimentos son originales.

10.1.2 De las Tecnologías Semánticas

Desde la perspectiva tecnológica, se demostró que datos en texto plano (no estructurado) en combinación con datos abiertos del ámbito legislativo, procesados mediante Tecnologías Semánticas, particularmente RDF, ontologías, clasificadores de texto y componentes de procesamiento como los asociados al marcaje automático, suficientes para implementar aplicaciones que permitan análisis complejos con un alto grado de acierto.

Desde la perspectiva del procesamiento documental, se demuestra además que los flujos de trabajo que forman parte del marco de trabajo permiten reducir de forma importante los tiempos de procesamiento, lo que a su vez permite reducir costos y aumentar las capacidades de los equipos de trabajo.

Respecto a tecnologías en portales de contenido, se evidenció que el uso de arquitecturas híbridas basadas en Tecnologías Semánticas combinadas a tecnologías tradicionales como bases de datos relacionales e índices textuales, permite una correcta entrega de contenidos, superando las limitaciones identificadas en modelos basados exclusivamente en razonamiento lógico o inferencias materializadas.

Otra conclusión relevante a partir del trabajo realizado durante la investigación fue la validación de mecanismos para el control de calidad de datos RDF, a través de la aplicación de herramientas como ShEx. Se comprobó que incluso en bases curadas y en producción, la validación estructural puede aplicarse para identificar anomalías relevantes, lo cual refuerza la necesidad de incorporar procesos sistemáticos de validación cuando se consumen datos de terceros, incluso si estos provienen de instituciones públicas.

10.1.3 De otros hallazgos

También dentro de otros hallazgos identificados durante el análisis a los datos del Congreso Nacional de Chile, se verifica que el Senado muestra mayores niveles de alineamiento y menor polarización que la Cámara de Diputadas y Diputados, lo cual puede estar asociado a factores como la edad, experiencia o cultura institucional. Además, el análisis arrojó que existe un conjunto importante de proyectos con alto consenso técnico que, sin embargo, permanecen estanca-

dos en tramitación. Esto sugiere que los indicadores de alineamiento y polarización definidos pueden ser utilizados como criterios para agilizar la agenda legislativa, lo que contribuiría a una mayor eficiencia institucional y a mejorar la percepción ciudadana del Congreso.

Una conclusión relevante, aunque no tan evidente, es que al inicio del trabajo se asumía que la información proveniente de distintas fuentes, como prensa, redes sociales y debate parlamentario en el Congreso, era igualmente pertinente para el análisis político-legislativo. Sin embargo, considerando las diferencias observadas en los datos analizados (sección 2), esta suposición resulta al menos cuestionable. Este cuestionamiento surge porque, si bien el debate parlamentario se configura como el espacio donde se expresan los temas y dinámicas más significativas del ámbito político-legislativo, ofreciendo una vitrina equitativa para todos los parlamentarios, tanto la prensa como las redes sociales operan bajo lógicas propias: en el caso de la prensa, siguiendo una dinámica de visibilidad mediática selectiva determinada por criterios editoriales; y en el caso de las redes sociales, en función de la adopción digital que cada parlamentario y su equipo de comunicación o asesores decidan implementar.

10.1.4 Conclusión final

En definitiva, todos estos hallazgos y conclusiones preliminares permiten dar por validada la hipótesis principal, que es: **que disponiendo de datos no estructurados provenientes de diversas fuentes, *sí* es posible aplicar Tecnologías Semánticas bajo procesos automatizados, para la extracción y anotación de metadatos, como también para la generación de relaciones entre entidades y metadatos, tales que permitan consultar y describir información agregada y desagregada, proveyendo un instrumento de análisis cuantitativo útil para el análisis legislativo**.

10.2 Trabajo futuro

A partir de este trabajo se prevén las siguientes líneas de trabajo futuro:

1. A la hora de repetir los experimentos de esta investigación, se plantea la incorporación de mecanismos de control experimental más robustos, como incorporar preguntas ficticias o controles de falsos positivos, que permitan detectar sesgos interpretativos en las respuestas del GE.

2. En la misma línea anterior, una vez que se agregue una nueva ronda experimental, se pretende aumentar el tamaño y la diversidad del grupo experto para reforzar la potencia estadística y detectar posibles sesgos.

3. Asociado al instrumento 1 sobre Temas de interés parlamentario, se prevé que es posible mejorar el instrumento añadiendo algunos elementos de contexto asociados al concepto de "Sector político" (Pregunta 2), para intentar mitigar la variabilidad de las respuestas hacia la opción *"No lo tengo claro"*. La incorporación de estos elementos puede ser directamente añadiendo el índice de tendencia política del parlamentario o indicar su partido, añadiendo al mismo tiempo elementos gráficos que muestren los distintos partidos políticos existentes.

4. Asociado al mismo instrumento 1, se pretende en el corto plazo redactar un artículo científico con el método y los resultados experimentales expuestos en esta investigación

además de incorporar la visualización de la evolución de los temas de interés parlamentario en el tiempo.

5. Asociado al instrumento 2, se pretende redactar una actualización de los artículos científicos ya publicados, incorporando los resultados experimentales obtenidos durante la investigación.

6. En cuanto al índice de tendencia política incorporado en el instrumento 2 y descrito en la sección7.4.3, se prevé la intención de calcular un índice dinámico con base en los datos de votaciones, para todos los partidos políticos y así eliminar el índice basado en la percepción del autor (ya que incorpora un sesgo), y con base en ello generar una nueva publicación científica.

7. Asociado al instrumento 3, primeramente se pretende mejorar su fiabilidad, intentando migrar las valoraciones expertas desde "No lo tengo claro" hacia valores positivos. Para ello, se evaluará la incorporación de nuevos elementos gráficos de apoyo, como también eventualmente probar nuevas métricas que puedan complementar la vista conceptual asociada a la identificación de roles clave.

8. A nivel global, también se prevé la posibilidad de expandir todos estos instrumentos a una ventana temporal, es decir, analizar la evolución de las ventanas estáticas que se presentan actualmente en los instrumentos, por ventanas móviles dependientes de periodos. Esto permitirá visualizar tendencias de comportamiento que pueden ser útiles para generar otros tipos de análisis, tales como análisis prospectivo.

9. También se considera la posibilidad de abrir las herramientas de análisis a otros paradigmas, tales como RAG (Retrieval Augmented Generation), donde es posible combinar modelos de lenguaje con buscadores, modelos de aprendizaje supervizado y grafos de conocimiento, para generar nuevos instrumentos de análisis basados en datos heterogéneos.

Los instrumentos desarrollados están disponibles para ser integrados a la brevedad en flujos de trabajo de análisis parlamentario, aportando evidencia empírica consistente al proceso de toma de decisiones. La disponibilización de estas herramientas de apoyo, aportaría en la disminución de la asimetría de información existente entre el poder legislativo y el poder ejecutivo.

Capítulo 11

Artículos publicados

En este capítulo se presentan los artículos de investigación publicados por el autor durante el periodo de matrícula en el programa de doctorado que están estrechamente relacionados con los objetivos y contribuciones de esta tesis. Cada publicación incluye el título, un resumen que describe los principales hallazgos y aportes, y la referencia a la revista científica o conferencia donde fue publicada. Estos trabajos representan el esfuerzo por difundir los avances logrados durante el desarrollo de la tesis, y han sido validados a través de su aceptación en distintas instancias académicas. La tabla 11.1 presenta un resumen de los artículos como autor principal que son descritos posteriormente.

Año	Título	Medio	Tipo
2019	Legislative Document Content Extraction Based on Semantic Web Technologies A Use Case About Processing the History of the Law	Extended Semantic Web Conference (ESWC2019), Portoroz, Slovenia. LNCS, volume 11503, 2019	Conferencia tipo A
2020	National Budget as Linked Open Data: New Tools for Supporting the Sustainability of Public Finances	Sustainability (MDPI)	Revista JCR IF 2,592
2023	Using polarization and alignment to identify quick-approval law propositions: An open linked data application	International Conference on Applied Informatics (ICAI 2023)	Conferencia tipo B
2024	Toward Efficient Legislative Processes: Analysis of Chilean Congressional Bill Votes Using Semantic Web Technologies	Sringer Nature Computer Science	Revista Scopus
2025	Transforming parliamentary libraries: Enhancing processes and delivering new services with AI	IFLA Journal	Revista JCR IF 1,1

Tabla 11.1: Resumen de publicaciones durante el desarrollo de la tesis

Legislative Document Content Extraction Based on Semantic Web Technologies A Use Case About Processing the History of the Law

Abstract

This paper describes the system architecture for generating the History of the Law developed for the Chilean National Library of Congress (BCN). The production system uses Semantic Web technologies, Akoma-Ntoso, and tools that automate the marking of plain text to XML, enriching and linking documents. These documents semantically annotated allow to develop specialized political and legislative services, and to extract knowledge for a Legal Knowledge Base for public use. We show the strategies used for the implementation of the automatic markup tools, as well as describe the knowledge graph generated from semantic documents. Finally, we show the contrast between the time of document processing using semantic technologies versus manual tasks, and the lessons learnt in this process, installing a base for the replication of a technological model that allows the generation of useful services for diverse contexts.

Autores Francisco Cifuentes-Silva, Jose Emilio Labra-Gayo
DOI: https://doi.org/10.1007/978-3-030-21348-0_36
Conferencia: Extended Semantic Web Conference (ESWC2019), LNCS, volume 11503, 2019 **Nominado como mejor artículo de estudiante**

National Budget as Linked Open Data: New Tools for Supporting the Sustainability of Public Finances

Abstract

This paper presents the visualization of national budget, a tool based on Semantic Web technologies that shows by graphic representations the Chilean budget law published annually, and their execution by each state agency. We describe the processes for consuming open data from the Budget National Agency, and how this data is transformed and published to linked open data, based on a National Budget Ontology. Although similar initiatives have been developed on transparency and public budget around the world, we consider that there is no previous experience showing optimized access mechanisms both for human and machine readable, providing in each case the highest level of aggregation, granularity and interoperability, making it understandable and easy to process complex data and legislation. As part of our analysis, we describe a recent scenario of usage in the context of the socio-political crisis in Chile, where we discuss the possible impact of the linked open dataset and data visualizations for distribution and control of funds, on the premise that this type of tools can support the decision making and sustainability of public finances. Finally, we present the results of our budget knowledge graph and the lessons learned during the development, allowing to replicate the process and enabling potential uses of the published data in other contexts.

Autores Francisco Cifuentes-Silva, Daniel Fernández-Álvarez, Jose Emilio Labra-Gayo
DOI: https://doi.org/10.3390/su12114551
Journal: Sustainability (JCR Impact Factor 2.592), Vol 12(11), 4551 - 2020

Using Polarization and Alignment to Identify Quick-Approval Law Propositions: An Open Linked Data Application

Abstract

Since the return of democracy in 1990 until the end of 2020, Chile's Congress has processed and approved 2404 laws, with an average processing time of 695 days from proposal to official publication. Recent political circumstances have given urgency to identifying those law propositions that might be shepherded to faster approval and those that will likely not be approved. This article proposes to classify law proposals, as well as parliamentarians and political parties, along two axes: polarization (lack of agreement on an issue) and (political) alignment (intra-party coincidence of a group's members regarding specific opinion), yielding four quadrants: (a) "ideological stance" (high polarization, high alignment), (b) "personal interests" (high polarization, low alignment), (c) "thematic interest" (low polarization, low alignment), and (d) "technical consensus" (low polarization, high alignment). We used this scheme to analyze an existing open-linked dataset that records parliamentarians' political parties and their voting on law proposals during 1990–2020. A simple visualization allows identifying a large set of propositions (1,643 = 68%) with technical consensus (i.e., low polarization and high alignment), which

could have been quickly shepherded to approval, but instead took 687 days on average (i.e., essentially the same time as others). Wider adoption of this analysis may speed up legislative work and ultimately allow Congress to serve citizens more promptly.

Autores Francisco Cifuentes-Silva, Jose Emilio Labra-Gayo, Hernán Astudillo and Felipe Rivera-Polo

DOI: https://doi.org/10.1007/978-3-031-46813-1_9

Conferencia: International Conference on Applied Informatics - ICAI 2023

Toward Efficient Legislative Processes: Analysis of Chilean Congressional Bill Votes Using Semantic Web Technologies

Abstract

Between 1990 and 2023, Chile's Congress processed and approved 2738 laws, with an average processing time of 667.8 days from proposal to official publication. Recent political circumstances have underscored the need to identify legislative proposals that can be expedited for approval and which ones are unlikely to be approved at all. This article describes a bottom-up, data-driven classification of voting (and voters) on law proposals, which yield two axis: polarization (lack of agreement on an issue), and (political) alignment (intra-party coincidence of a group's members regarding certain opinion). And four quadrants: "ideological stance" (high polarization, high alignment), "personal interests" (high polarization, low alignment), "thematic interest" (low polarization, low alignment), and "technical consensus" (low polarization, high alignment). We used this scheme to analyze an existing Open Linked Dataset with semantic web technologies (ontologies, RDF Shape expressions, and URI patterns), which records parliamentarians' political parties and their voting on law proposals during 1990–2023. We found that most bills (70.14%) are in the technical consensus quadrant, and could have been quickly shepherded to approval. Wider adoption of this analysis to classify new bills may help to speed up their legislative processing, ultimately allowing Congress to serve citizens in a more timely manner.

Autores Francisco Cifuentes-Silva, Hernán Astudillo , Jose Emilio Labra-Gayo and Felipe Rivera-Polo

DOI: https://doi.org/10.1007/s42979-024-02933-y

Journal: SN Computer Science, 2024

Transforming parliamentary libraries: Enhancing processes and delivering new services with AI

Abstract

The integration of Artificial Intelligence (AI) in libraries offers wide impact on the evolution of information access and management. It allows both streamlining internal processes and transforming the way users interact with information resources, thus enhancing effectiveness and operational efficiency while enriching the user experience. This article presents the experience in incorporating several AI techniques in Chile's Library of Congress (BCN), and describes three initiatives: 1) publishing legislation as linked open data with Semantic Web

technologies, combining machine-readable comprehension to high standards of interoperability; 2) maintaining the history of legislation, via automatic tagging of legislative documentation with natural language processing; and 3) predicting law approval based on current political context, using machine learning. The use of these technologies has allowed BCN to offer a wide variety of knowledge management services, providing useful and timely information for parliamentary work, and automated human-based repetitive tasks for efficient use of public resources.

Autores Francisco Cifuentes-Silva, Hernán Astudillo and Jose Emilio Labra-Gayo
DOI: `https://doi.org/10.1177/0340035225131584`
Journal: IFLA Journal, 2025

Bibliografía

[owl, 2012] (2012). OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[Abercrombie and Batista-Navarro, 2020] Abercrombie, G. and Batista-Navarro, R. (2020). ParlVote: A corpus for sentiment analysis of political debates. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

[Adomavicius and Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

[Aelst et al., 2010] Aelst, P. V., Sehata, A., and Dalen, A. V. (2010). Members of parliament: Equal competitors for media attention? an analysis of personal contacts between mps and political journalists in five european countries. *Political communication*, 27(3):310–325.

[Agarwal et al., 2019] Agarwal, P., Sastry, N., and Wood, E. (2019). Tweeting mps: Digital engagement between citizens and members of parliament in the uk. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):26–37.

[Ansari et al., 2000] Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, 37(3):363–375.

[Awadallah et al., 2012] Awadallah, R., Ramanath, M., and Weikum, G. (2012). Opinions network for politically controversial topics. In *Proceedings of the first edition workshop on Politics, elections and data*, pages 15–22.

[BCN, 2025] BCN, B. (2025). El congreso nacional y sus edificios. historia 1811-1823. `https://www.bcn.cl/historiapolitica/congreso_nacional/historia/index.html?periodo=1811-1823`.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html.

[Berners-Lee et al., 1996] Berners-Lee, T., Fielding, R., and Frystyk, H. (1996). Hypertext transfer protocol–http/1.0. Technical report.

[Bimber, 2014] Bimber, B. (2014). Digital media in the obama campaigns of 2008 and 2012: Adaptation to the personalized political communication environment. *Journal of Information Technology & Politics*, 11(2):130–150.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bluntschli, 1886] Bluntschli, J. C. (1886). *Lehre vom modernen Staat*. JG Cotta.

[Brito Cruz et al., 2019] Brito Cruz, F., Valente, M. H., and Zanatta, R. A. F. (2019). Secrets and Lies: WhatsApp and Social Media in Brazil's 2018 Presidential Election.

[Campos et al., 2023] Campos, R., Jatowt, A., and Jorge, A. (2023). Text mining and visualization of political party programs using keyword extraction methods: The case of portuguese legislative elections. In Sserwanga, I., Goulding, A., Moulaison-Sandy, H., Du, J. T., Soares, A. L., Hessami, V., and Frank, R. D., editors, *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pages 340–349, Cham. Springer Nature Switzerland.

[Castillo et al., 2019] Castillo, S., Allende-Cid, H., Palma, W., Alfaro, R., Ramos, H. S., Gonzalez, C., Elortegui, C., and Santander, P. (2019). Detection of bots and cyborgs in twitter: A study on the chilean presidential election in 2017. In Meiselwitz, G., editor, *Social Computing and Social Media. Design, Human Behavior and Analytics*, pages 311–323, Cham. Springer International Publishing.

[Cañete et al., 2020] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

[Chen and Manning, 2014] Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

[Cifuentes-Silva et al., 2025] Cifuentes-Silva, F., Astudillo, H., and Gayo, J. E. L. (2025). Transforming parliamentary libraries: Enhancing processes delivering new services with artificial intelligence. *IFLA Journal*, 0(0):03400352251315844.

[Cifuentes-Silva et al., 2024] Cifuentes-Silva, F., Astudillo, H., Gayo, J. E. L., and Rivera-Polo, F. (2024). Toward Efficient Legislative Processes: Analysis of Chilean Congressional Bill Votes Using Semantic Web Technologies. *SN Computer Science*, 5(5):604.

[Cifuentes-Silva et al., 2020] Cifuentes-Silva, F., Fernández-Álvarez, D., and Labra-Gayo, J. E. (2020). National budget as linked open data: New tools for supporting the sustainability of public finances. *Sustainability*, 12(11):4551.

[Cifuentes-Silva and Labra Gayo, 2019] Cifuentes-Silva, F. and Labra Gayo, J. E. (2019). Legislative document content extraction based on semantic web technologies. In Hitzler, P., Fernández, M., Janowicz, K., Zaveri, A., Gray, A. J., Lopez, V., Haller, A., and Hammar, K., editors, *The Semantic Web*, pages 558–573, Cham. Springer International Publishing.

[Cifuentes-Silva et al., 2023] Cifuentes-Silva, F., Labra Gayo, J. E., Astudillo, H., and Rivera-Polo, F. (2023). Using Polarization and Alignment to Identify Quick-Approval Law Propositions: An Open Linked Data Application. pages 122–137.

[Cifuentes-Silva et al., 2011] Cifuentes-Silva, F., Sifaqui, C., and Labra-Gayo, J. E. (2011). Towards an architecture and adoption process for linked data technologies in open government

contexts: A case study for the library of congress of chile. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, page 79–86, New York, NY, USA. Association for Computing Machinery.

[Curry et al., 2024] Curry, J. M., Lee, F. E., and Oldham, R. L. (2024). On the Congress Beat: How the Structure of News Shapes Coverage of Congressional Action. *Political Science Quarterly*, 140(1):1–27. _eprint: https://academic.oup.com/psq/article-pdf/140/1/1/56965171/qqae008.pdf.

[Desposato, 2003] Desposato, S. W. (2003). Comparing group and subgroup cohesion scores: A nonparametric method with an application to brazil. *Political Analysis*, 11(3):275–288.

[Ericsson et al., 1993] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363.

[Everitt, 1998] Everitt, B. (1998). The cambridge dictionary of statistics. In *The Cambridge dictionary of statistics*.

[Fenno et al., 1978] Fenno, R. F. et al. (1978). *Home style: House members in their districts*. Little, Brown Boston.

[Fernández-Álvarez et al., 2018] Fernández-Álvarez, D., García-González, H., Frey, J., Hellmann, S., and Gayo, J. E. L. (2018). Inference of latent shape expressions associated to dbpedia ontology. In *ISWC (P&D/Industry/BlueSky)*.

[Forkman, 2009] Forkman, J. (2009). Estimator and tests for common coefficients of variation in normal distributions. *Communications in Statistics—Theory and Methods*, 38(2):233–251.

[Francart et al., 2019] Francart, T., Dann, J., Pappalardo, R., Malagon, C., and Pellegrino, M. (2019). The european legislation identifier. In *Knowledge of the Law in the Big Data Age*, pages 137–148. IOS Press.

[George and Mallery, 2016] George, D. and Mallery, P. (2016). *IBM SPSS statistics 23 step by step: a simple guide and reference*. Routledge, New York, NY, fourteenth edition edition.

[Gerodimos and Justinussen, 2015] Gerodimos, R. and Justinussen, J. (2015). Obama's 2012 facebook campaign: Political communication in the age of the like button. *Journal of information technology & politics*, 12(2):113–132.

[Glavaš et al., 2017] Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017). Cross-lingual classification of topics in political texts. Association for Computational Linguistics (ACL).

[GovTrack.us, 2013] GovTrack.us (2013). Ideology analysis of members of congress.

[Graham and Andrejevic, 2024] Graham, T. and Andrejevic, M. (2024). A computational analysis of potential algorithmic bias on platform x during the 2024 us election.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

[Guha and Brickley, 2014] Guha, R. and Brickley, D. (2014). RDF schema 1.1. W3C recommendation, W3C. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[Hall, 2022] Hall, N.-A. (2022). Understanding brexit on facebook: Developing close-up, qualitative methodologies for social media research. *Sociological Research Online*, 27(3):707–723.

[Harary et al., 1965] Harary, F., Norman, R. Z., Cartwright, D., et al. (1965). *Structural models: An introduction to the theory of directed graphs*, volume 82. Wiley New York.

[Hartmann et al., 2023] Hartmann, J., Schwenzow, J., and Witte, M. (2023). The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

[Hernández et al., 2014] Hernández, R., Fernández, C., and Baptista, P. (2014). Metodología de la investigación mcgraw-hill. *México Df*, pages 217–2.

[Hogan et al., 2021] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J. F., Staab, S., and Zimmermann, A. (2021). *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer.

[Hug, 2010] Hug, S. (2010). Selection effects in roll call votes. *British Journal of Political Science*, 40(1):225–235.

[Isoaho et al., 2021] Isoaho, K., Gritsenko, D., and Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, 49(1):300–324.

[Iyyer et al., 2014] Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

[Jeria Cánovas and Wall Opazo, 2005] Jeria Cánovas, A. and Wall Opazo, C. (2005). Segmentación psicográfica : una aplicación para Chile. Accepted: 2016-12-26T19:03:20Z Publisher: Universidad de Chile.

[Kai-Ineman et al., 1979] Kai-Ineman, D., Tversky, A., et al. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):363–391.

[Kim et al., 2024] Kim, D.-E., Lee, H., Lee, S.-k., and Eom, S.-J. (2024). Conditions for ai systems adoption in public sector: From an accountability perspective. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, dg.o '24, page 42–51, New York, NY, USA. Association for Computing Machinery.

[Knublauch et al., 2017] Knublauch, H., TopQuadrant, Inc., Kontokostas, D., and University of Leipzig (2017). Shapes constraint language (shacl). *W3C Recommendation*, 11:8.

[Kotler and Keller, 2006] Kotler, P. and Keller, K. L. (2006). *Dirección de marketing*. Pearson educación.

[Lauderdale and Herzog, 2016] Lauderdale, B. E. and Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

[Laver et al., 2003] Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

[Lerner and Tetlock, 1999] Lerner, J. S. and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological bulletin*, 125(2):255.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

[Lin et al., 2015] Lin, F.-R., Chou, S.-Y., Liao, D., and Hao, D. (2015). Automatic Content Analysis of Legislative Documents by Text Mining Techniques. In *2015 48th Hawaii International Conference on System Sciences*, pages 2199–2208. ISSN: 1530-1605.

[Lipton, 2018] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

[Loewenberg et al., 1985] Loewenberg, G., Patterson, S. C., and Jewell, M. E. (1985). *Handbook of legislative research*. Harvard University Press.

[Lucas et al., 2015] Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2):254–277.

[Maclure, 2021] Maclure, J. (2021). AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines*, 31(3):421–438.

[Matas, 2018] Matas, A. (2018). Diseño del formato de escalas tipo Likert: un estado de la cuestión. *Revista electrónica de investigación educativa*, 20(1):38–47. Publisher: Universidad Autónoma de Baja California, Instituto de Investigación y Desarrollo Educativo.

[Mayhew, 1974] Mayhew, D. R. (1974). *Congress: The electoral connection*. Yale university press.

[MIT, 2012] MIT (2012). How obama's team used big data to rally voters. *MIT Technology Review*. Accessed: 2025-04-04.

[Monroe et al., 2017] Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2017). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

[Moreno, 1934] Moreno, J. L. (1934). Who shall survive? a new approach to the problem of human interrelations.

[Morán, 2020] Morán, C. L. F. (2020). Cooperation and polarization in a presidential congress: Policy networks in the chilean lower house 2006–2017. *Politics*, 40(2):227–244.

[Nery and Mueller, 2022] Nery, P. F. and Mueller, B. (2022). Co-sponsorship networks in the brazilian congress: An exploratory analysis of caucus influence. *Estudos Econômicos (São Paulo)*, 52:83–111.

[Palmirani and Vitali, 2011] Palmirani, M. and Vitali, F. (2011). Akoma-ntoso for legal documents. *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*, pages 75–100.

[Palmirani and Vitali, 2012] Palmirani, M. and Vitali, F. (2012). Legislative xml: principles and technical tools.

[Papenmeier et al., 2022] Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Trans. Comput.-Hum. Interact.*, 29(4).

[Pérez et al., 2009] Pérez, J., Arenas, M., and Gutierrez, C. (2009). Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, 34(3):1–45.

[Perozzi et al., 2014] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

[Polsby, 1965] Polsby, N. W. (1965). *Congress and the Presidency*. Foundations of Modern Political Science. Prentice-Hall, Englewood Cliffs, NJ. Incluye análisis sobre la relación entre el Congreso y la presidencia en el sistema político estadounidense.

[Pons and Latapy, 2006] Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of graph algorithms and applications*, 10(2):191–218.

[Prabhu et al., 2019] Prabhu, C., Chivukula, A. S., Mogadala, A., Ghosh, R., and Livingston, L. J. (2019). *Social Semantic Web Mining and Big Data Analytics*, pages 217–231. Springer Singapore, Singapore.

[Praet et al., 2021] Praet, S., Martens, D., and Van Aelst, P. (2021). Patterns of democracy? social network analysis of parliamentary twitter networks in 12 countries. *Online Social Networks and Media*, 24:100154.

[Prama et al., 2025] Prama, T. T., Bagchi, C., Kalakonnavar, V., Krauß, P., and Grabowicz, P. A. (2025). Political biases on x before the 2025 german federal election. *arXiv preprint arXiv:2503.02888*.

[Prud'hommeaux et al., 2014] Prud'hommeaux, E., Labra Gayo, J. E., and Solbrig, H. (2014). Shape expressions: an rdf validation and transformation language. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 32–40. ACM.

[Ratinov and Roth, 2009] Ratinov, L. and Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Reyes Olmedo, 2017] Reyes Olmedo, P. (2017). Estándares técnico-jurídicos de gestión para servicios digitales de información legislativa. *Revista chilena de derecho y tecnología*, 6:57 – 95.

[Rice, 1928] Rice, S. A. (1928). *Quantitative methods in politics*. AA Knopf.

[Rodríguez et al., 2018] Rodríguez, S., Allende-Cid, H., Palma, W., Alfaro, R., Gonzalez, C., Elortegui, C., and Santander, P. (2018). Forecasting the chilean electoral year: Using twitter to predict the presidential elections of 2017. In *Social Computing and Social Media. Technologies and Analytics: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part II 10*, pages 298–314. Springer.

[Santander et al., 2017] Santander, P., Elórtegui, C., González, C., Allende-Cid, H., and Palma, W. (2017). Redes sociales, inteligencia computacional y predicción electoral: el caso de las primarias presidenciales de chile 2017. *Cuadernos. info*, (41):41–56.

[Semetko and Valkenburg, 2000] Semetko, H. A. and Valkenburg, P. M. (2000). Framing european politics: A content analysis of press and television news. *Journal of communication*, 50(2):93–109.

[Sotoudeh et al., 2024] Sotoudeh, S., Porter, M. A., and Krishnagopal, S. (2024). A network-based measure of cosponsorship influence on bill passing in the united states house of representatives. *arXiv preprint arXiv:2406.19554*.

[Sudhahar et al., 2015] Sudhahar, S., Veltri, G. A., and Cristianini, N. (2015). Automated analysis of the us presidential elections using big data and network analysis. *Big Data & Society*, 2(1):2053951715572916.

[UK Parliament, 2019] UK Parliament, C. (2019). Disinformation and 'fake news': Final Report - Digital, Culture, Media and Sport Committee - House of Commons.

[Valle et al., 2022] Valle, M. E. D., Broersma, M., and Ponsioen, A. (2022). Political interaction beyond party lines: Communication ties and party polarization in parliamentary twitter networks. *Social Science Computer Review*, 40(3):736–755.

[W3C, 2013] W3C (2013). SPARQL 1.1 Query Language.

[Wilson, 1885] Wilson, W. (1885). *Congressional Government: A Study in American Politics*. Houghton, Mifflin.

[Yildirim et al., 2022] Yildirim, T. M., Thesen, G., Jennings, W., and Vries, E. D. (2022). The determinants of the media coverage of politicians: The role of parliamentary activities. *European Journal of Political Research*.

[Željko Poljak, 2024] Željko Poljak (2024). Give the media what they need: Negativity as a media access tool for politicians. *The International Journal of Press/Politics*, 0(0):19401612241234861.

Anexo A

Comisiones Parlamentarias Permanentes

La figura A.1 muestra dos columnas de elementos pareados, las cuales presentan las comisiones parlamentarias permanentes al 11 de marzo de 2025 en ambas cámaras del Congreso Nacional de Chile, enlazando cada una su equivalente en la otra cámara. Se presentan con flechas verdes aquellas comisiones que dividen su función existente en el Senado en dos funciones en la Cámara de Diputados. Al mismo tiempo, las comisiones que no tienen una equivalencia en las corporaciones, se presentan de color anaranjado, y finalmente, la única comisión permanente bicameral se presenta de color verde.

Figura A.1: Equivalencia entre Comisiones Parlamentarias en el Congreso Nacional de Chile

Anexo B

Clasificación multiclase en 6 categorías

A continuación se describe el experimento y las detalles técnicos realizados para la implementación de un clasificador multiclase para la clasificación de intervenciones parlamentarias basadas en el primer nivel de la taxonomía de temas legislativos definida en la investigación .

B.1 Datos utilizados

Para le elaboración del corpus de entrenamiento se utilizó un subconjunto de 5.400 intervenciones en castellano del Congreso Nacional Chileno, asociadas al periodo en estudio (Legislatura 367) clasificadas manualmente por el autor, lo que se distribuye en 900 intervenciones para cada tipo, las que fueron marcadas utilizando la herramienta *OpenRefine*[1].

Para definir las 6 categorías, se utilizó la taxonomía de temas legislativos desarrollado en la sección 7.3.3.

B.2 Algoritmos utilizados

Para la implementación del clasificador se utilizó el framework abierto *scikit-learn* (1.3.1)[2] y la biblioteca de PLN NLTK[3] (3.9.1), ambas escritas en lenguaje Python, probando los siguientes algoritmos de clasificación:

- *RidgeClassifier*: es un modelo de clasificación basado en un modelo de regresión múltiple, que convierte la salida de la regresión en etiquetas de clase mediante un umbral calculado durante el entrenamiento.
- MLPClassifier: MLP es un algoritmo de clasificación basado en una red neuronal multicapa.
- LogisticRegression: Implementa un modelo de clasificación mediante una función logística con valores entre 0 y 1.
- LinearSVC: Implementa una SVM con núcleo lineal que busca el hiperplano que maximiza el margen entre clases.

[1] `https://openrefine.org/`
[2] `https://scikit-learn.org/`
[3] `https://www.nltk.org/`

B.3 Implementación de características

Con el objetivo de capturar la relevancia semántica de los términos, se implementó un vector de pesos basado en la técnica TF-IDF, la cual permite asignar un valor de relevancia a cada término en un documento en función de su importancia respecto al resto de documentos, excluyendo previamente una lista de stopwords en español proporcionada por la biblioteca NLTK. Adicionalmente, se incorporó una representación vectorial del texto como característica de entrada para los clasificadores, generada mediante un modelo de embeddings multilingüe que incluye lenguaje español. La tokenización se realizó utilizando el modelo entrenado en español *BETO: Spanish Bert* (dccuchile/bert-base-spanish-wwm-uncased) [Cañete et al., 2020], mientras que la obtención de los vectores de embedding se llevó a cabo con el modelo sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [Reimers and Gurevych, 2019].

B.4 Validación de pruebas de clasificación

Para la validación del modelo se empleó una estrategia de validación cruzada de 10 particiones (10-fold cross-validation), en la que el conjunto de datos fue dividido en 10 subconjuntos de aproximadamente 540 intervenciones cada uno. En cada iteración, el modelo se entrena utilizando 9 de las particiones y se evalúa sobre la partición restante. Este procedimiento se repite 10 veces, de modo que cada subconjunto actúa una vez como conjunto de validación. Al finalizar, se calculan las métricas promedio de los 10 ensayos, lo que permite obtener una estimación robusta y generalizable del desempeño del modelo.

Para la evaluación se utilizaron las métricas de Accuracy, Precision, Recall y F1 Score, considerando su capacidad para reflejar distintos aspectos del rendimiento del clasificador en escenarios multiclase.

B.5 Ejecución del experimento

Se probaron los 4 modelos antes mencionados con distintos números de corpus de entrenamiento, variando desde 100 hasta 900 documentos por clase. También, como parte del experimento, y para mejorar la precisión del modelo, se probó la implementación de características adicionales con un modelo de análisis de sentimientos además de estrategias de normalización de texto, las cuales ambas finalmente fueron descartadas.

Luego de haber probado los 4 algoritmos de clasificación con distintas configuraciones, el clasificador que obtuvo mejores resultados fue el clasificador RidgeClassifier. A continuación se describe la configuración final y los valores de su ejecución.

B.6 Resultados

A continuación se muestran los valores de la evaluación obtenidos en cada iteración de la evaluación cruzada, obteniéndose los siguientes valores de *accuracy* por iteración:

Posteriormente, se calcularon las siguientes métricas globales sobre los 10 folds:

- **Accuracy:** 0.824
- **Precision:** 0.824

Fold	Accuracy
Fold 1	0.824
Fold 2	0.813
Fold 3	0.833
Fold 4	0.828
Fold 5	0.811
Fold 6	0.802
Fold 7	0.817
Fold 8	0.830
Fold 9	0.828
Fold 10	0.856

Tabla B.1: Valor de accuracy en cada iteración del entrenamiento de los 6 clasificadores

- **Recall:** 0.824
- **F1 Score:** 0.823

B.6.1 Métricas por categoría

A continuación se muestra el reporte de clasificación detallado para cada una de las 6 categorías temáticas:

Categoría	Precision	Recall	F1-Score	Total documentos
Temas sociales	0.82	0.71	0.76	900
Temas económicos	0.81	0.81	0.81	900
Temas medio ambientales	0.90	0.93	0.91	900
Temas de seguridad	0.81	0.84	0.82	900
Temas de infraestructura y transporte	0.82	0.89	0.85	900
Temas sobre politica, legislación y gobierno	0.77	0.77	0.77	900
Promedio	0.82	0.82	0.82	**N = 5400**

Tabla B.2: Reporte de clasificación para cada categoría temática

La matriz de confusión obtenida en el experimento se presenta en la figura B.1, la cual permite visualizar en la diagonal principal el número de aciertos del clasificador implementado respecto al corpus de entrenamiento.

A continuación, la figura B.2 presenta las curvas ROC para el clasificador desarrollado para cada una de las clases, y luego la figura B.3 presenta la curva ROC micro que permite tener una vista general del clasificador.

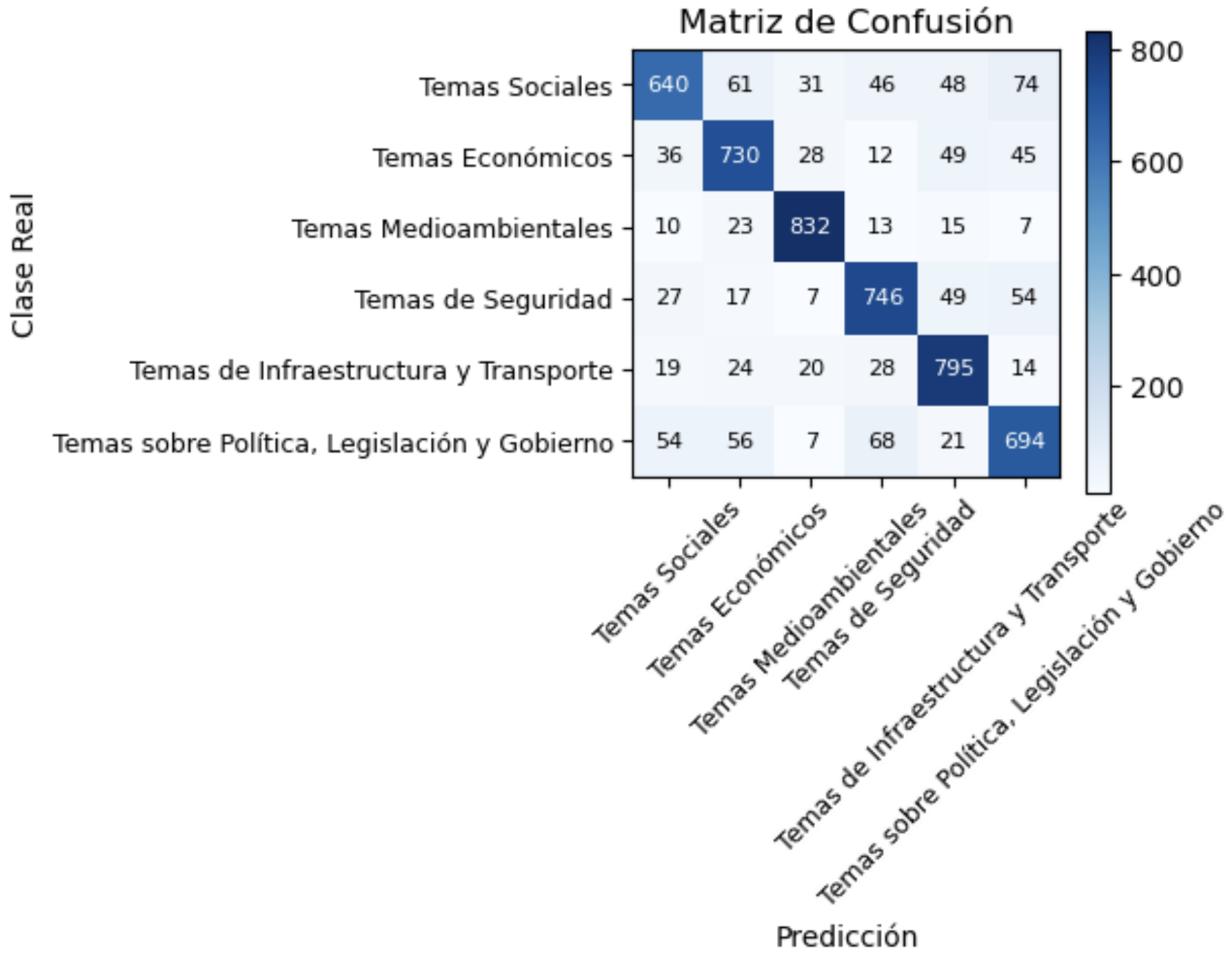

Figura B.1: Matriz de Confusión del experimento.

En este contexto, la figura B.4 presenta las curvas Precision-Recall para el clasificador desarrollado para cada una de las clases, y luego la figura B.5 presenta la curva Precision-Recall micro que permite tener una vista general del clasificador.

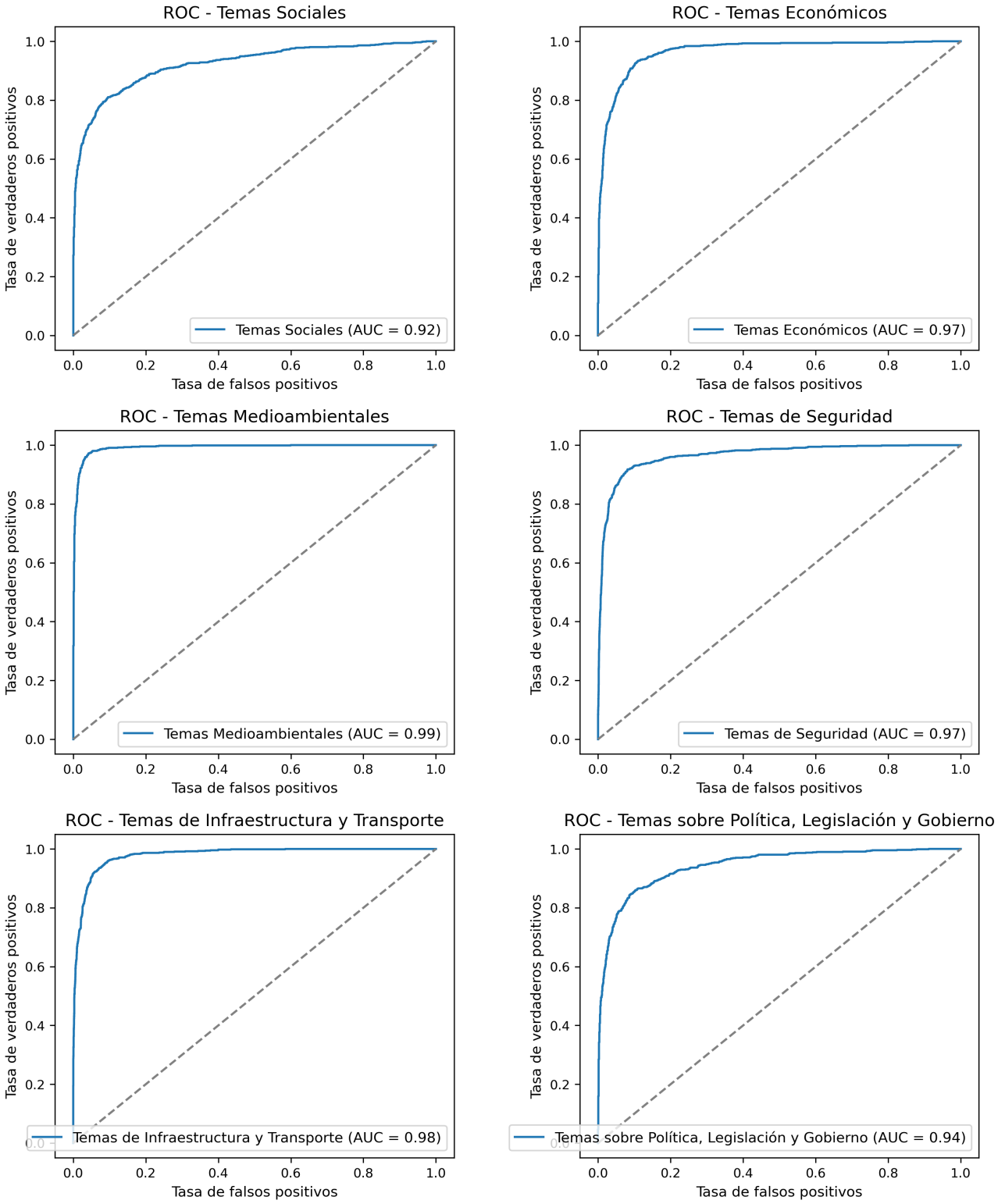

Figura B.2: Curvas ROC para los 6 clasificadores

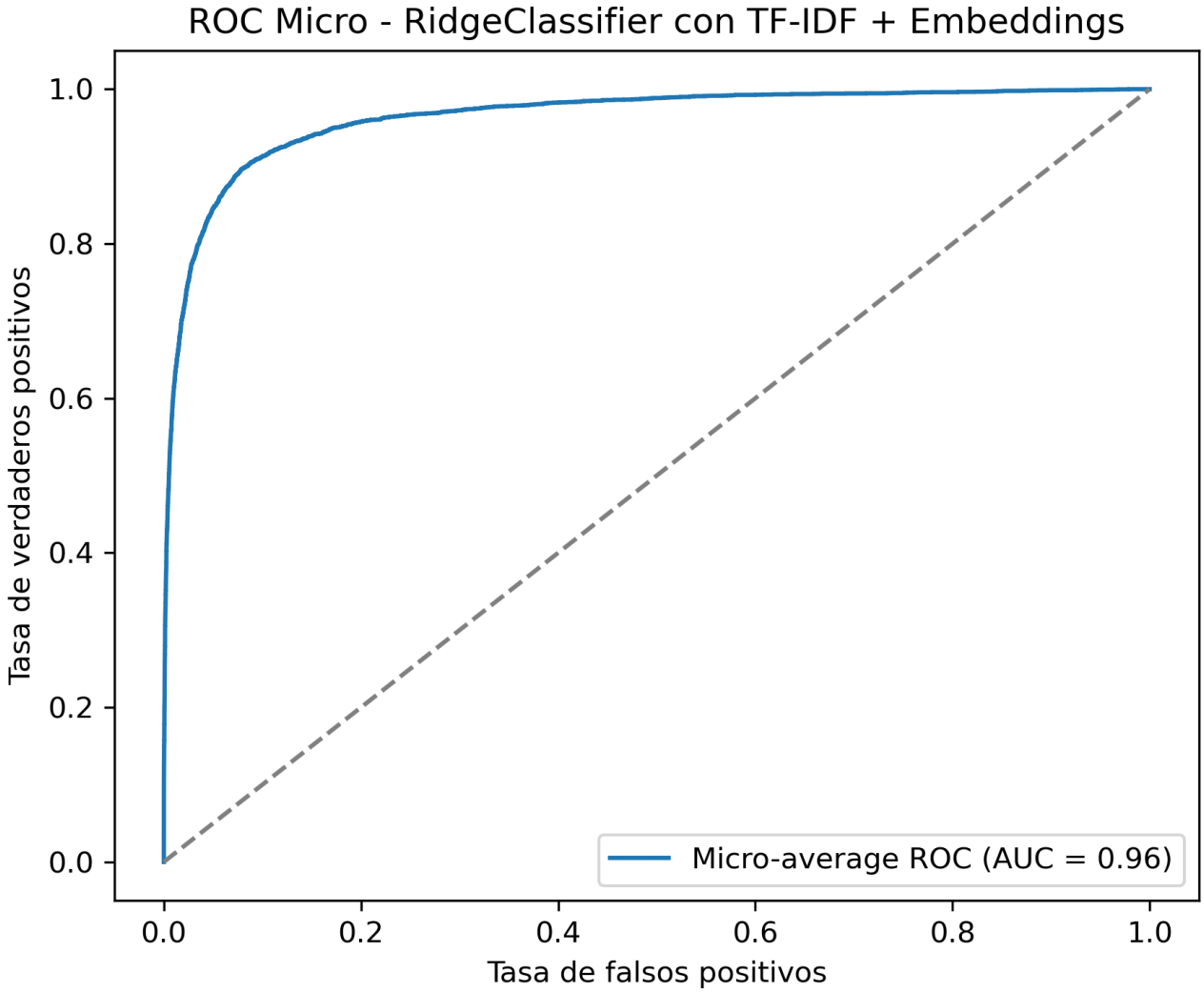

Figura B.3: Curva ROC micro para los 6 clasificadores

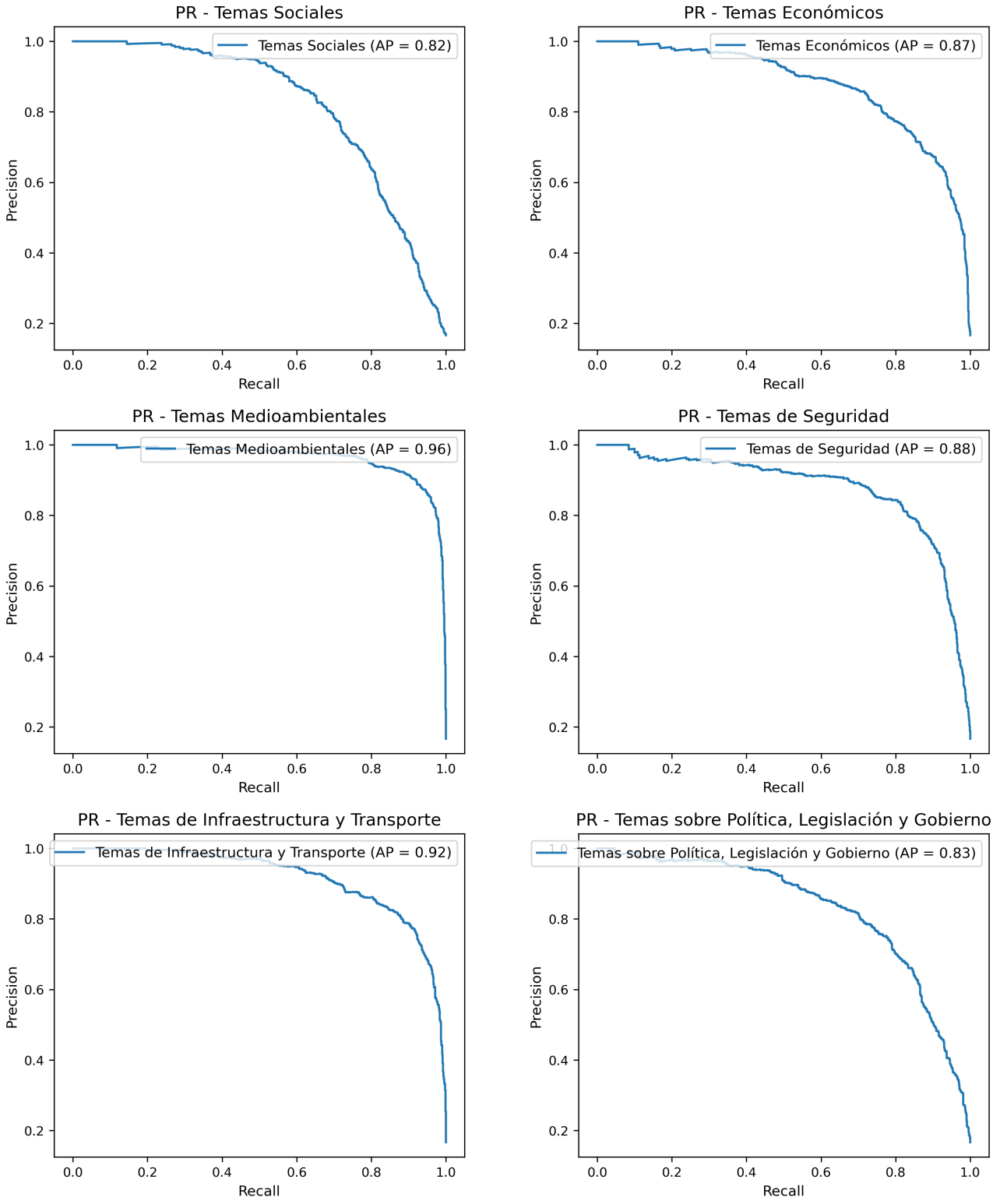

Figura B.4: Curvas PR para los 6 clasificadores

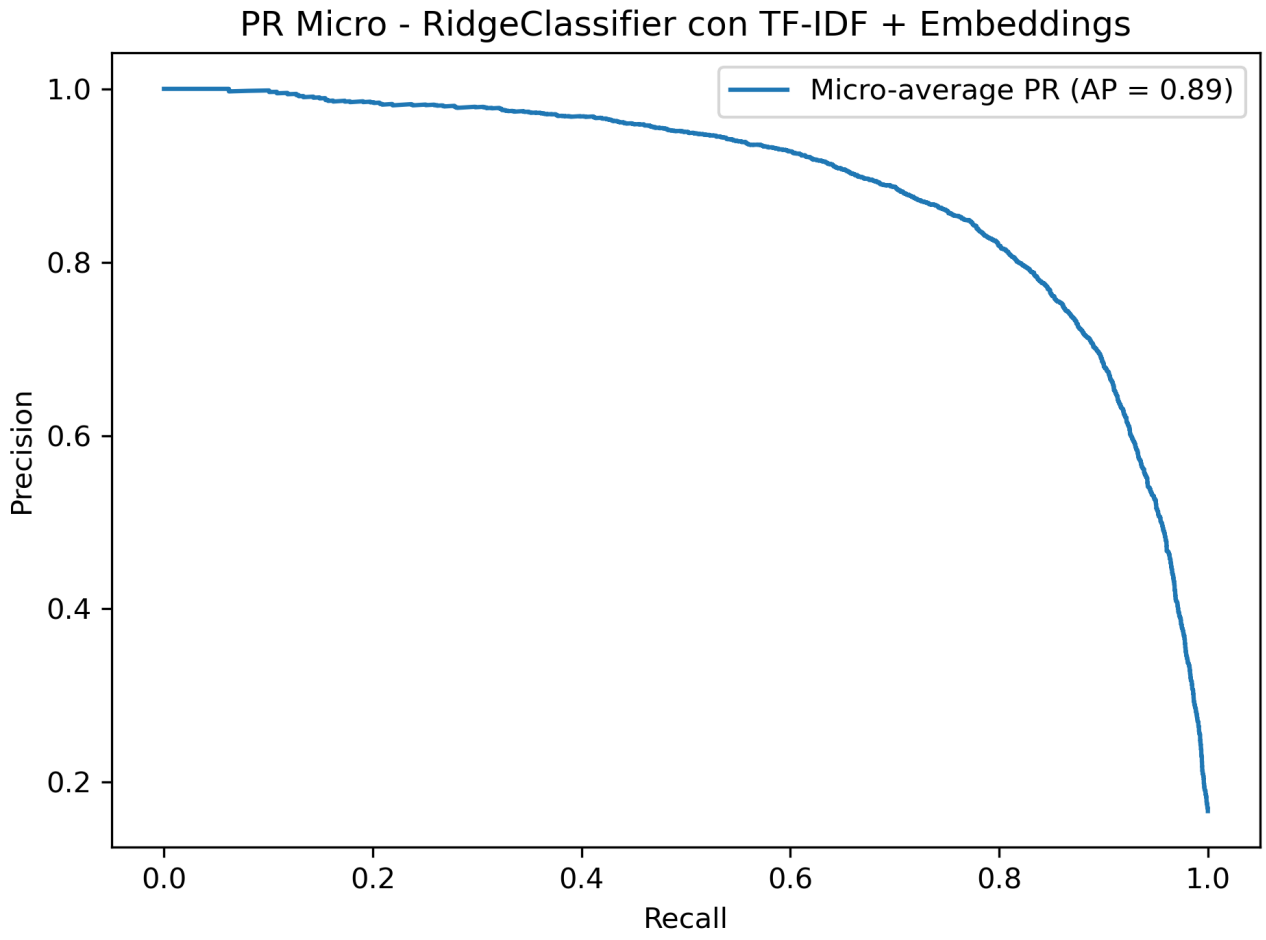

Figura B.5: Curva PR micro para los 6 clasificadores

B.7 Conclusiones de la clasificación en 6 categorías

El modelo RidgeClassifier utilizado en este experimento, en conjunto con la representación de datos mediante TF-IDF y embeddings, muestra un desempeño consistente en la clasificación de las 6 categorías temáticas, alcanzando un *accuracy* global de 0.824. El reporte de clasificación revela que algunas categorías presentan métricas superiores (por ejemplo, las temáticas medioambientales) mientras que otras podrían beneficiarse de futuras mejoras, ya sea a través la incorporación de nuevas características u otras estrategias.

Un *accuracy* global de 0.824 significa que, en promedio, el clasificador acierta en el 82.4% de las instancias, lo cual es una mejora sustancial comparado con un una línea base definida por un clasificador aleatorio (que en un problema con 6 clases balanceadas tendría una precisión alrededor del 16-17%). En el contexto de textos de intervenciones parlamentarias, que a menudo presentan lenguaje formal, tecnicismos y cierta ambigüedad, alcanzar más del 80% de accuracy es un indicador positivo de que el modelo logra capturar gran parte de la información relevante para distinguir entre las diversas temáticas.

Las intervenciones parlamentarias pueden incluir expresiones complejas, ironías, o matices específicos del discurso político, lo cual puede dificultar la clasificación. En este sentido, obtener un 82.4% de *accuracy* es un buen punto de partida, aunque siempre es posible evaluar si se puede lograr una mejora en aquellas categorías donde el desempeño es menor.

Desde el punto de vista del entrenamiento, si bien durante el proceso de etiquetado manual se observó una tendencia hacia una mayor cantidad de documentos en la categoría "Políticas sociales", se optó por equilibrar el número de documentos en cada categoría para evitar sesgos en el proceso de clasificación. Esta decisión se basó en el supuesto de que, una vez implementado el clasificador y aplicado al conjunto completo de documentos, el desequilibrio entre categorías se reflejaría de manera natural en los datos. Esta hipótesis fue posteriormente confirmada de forma empírica.

Respecto al análisis de las curvas ROC, es posible visualizar que en general el clasificador para todas las categorías logra valores altos (entre 0.92 hasta 0.99), con un valor medio de AUC 0.96 (muy alto). La curva asociada a la categoría de Temas sociales muestra el desempeño más bajo con un valor de AUC de 0.92, la cual dentro de todo se explica por la más alta variabilidad de los documentos asociados a esa categoría.

En la misma línea, el análisis de las curvas Precision-Recall muestra que si bien el AUC-PR no es tan alto como el AUC-ROC, el valor Micro-average PR es 0.89, lo cual es alto, considerando un base line 0.16 acorde a la proporción de elementos por categoría en la fase de entrenamiento.

Finalmente, es importante considerar que los textos clasificados con esta herramienta se utilizarán en un contexto de vista agregada, donde lo que se intenta rescatar es la tendencia general de los datos, más que (aunque deseada) la precisión en el detalle de cada clasificación.

Anexo C

Clasificación multiclase en subcategorías

Para la elaboración de los clasificadores en sub categorías, se utilizó el mismo conjunto de datos de intervenciones previamente descrito en el anexo B sobre clasificación multiclase en 6 categorías, pero asociando etiquetas del segundo nivel de la taxonomía de temas legislativos, asignándolas a los documentos de forma manual con base en una búsqueda y marcaje mediante la herramienta OpenRefine.

Para la selección y definición de características, se utilizó el mismo criterio de la clasificación multiclase en 6 categorías (descrito en la sección B.3) con las mismas herramientas y los 4 algoritmos de clasificación distintos, pero añadiendo una característica adicional asociada a la categoría padre, la cual pudo ser obtenida al clasificar el texto previamente con los clasificadores de primer nivel.

Dado que el conjunto de datos está naturalmente desequilibrado por los temas discutidos durante la tramitación legislativa, no fue posible la recolección de un mínimo de documentos que permitiera en todos los casos la implementación de un clasificador automático. En este escenario, se desarrollaron un total de 15 clasificadores de segundo nivel para un total de 35 categorías, y las restante 20 categorías fueron etiquetadas manualmente para la posterior visualización y análisis.

Respecto al criterio para establecer o no el desarrollo de los clasificadores bajo un esquema de aprendizaje supervisado, aunque muchos clasificadores obtenían valores de precisión altos (sobre 0.75), el recall en muchos casos no llegaba al 50%, afectando a la medida F1, y en definitiva haciendo que no se considerara como válido para la clasificación automatizada. De esta manera, la tabla C.1 presenta el reporte del proceso de entrenamiento para los clasificadores finalmente implementados asociados al segundo nivel de la taxonomía.

Clase	**Precisión**	**Recall**	**F1-score**	**Total documentos**
Educación	0.87	0.71	0.79	600
Salud	0.85	0.72	0.78	600
Vivienda	0.89	0.70	0.78	463
Igualdad de género	0.83	0.72	0.77	564
Adulto mayor y discapacidad	0.80	0.57	0.66	76
Cultura, deportes y recreacion	0.81	0.73	0.76	131
Desarrollo economico	0.78	0.92	0.85	595
Minería	0.88	0.51	0.65	43
Pesca	0.85	0.93	0.89	86
Medio ambiente	0.85	0.96	0.90	467
Recursos hídricos	0.77	0.79	0.78	102
Seguridad pública y ciudadana	0.81	0.98	0.89	600
Transporte público y telecomunicaciones	0.83	0.97	0.89	596
Energía y suministro	0.81	0.58	0.68	43
Gobierno	0.74	0.79	0.76	600
Promedio	0.82	0.77	0.79	**N = 5566**

Tabla C.1: Métricas de desempeño por clase en clasificación de 15 categorías

Anexo D

Cálculo del valor normalizado de relevancia por tema de interés

A continuación se describe, el funcionamiento matemático del cálculo del valor normalizado por tema de interés por parlamentario, asociado a las seis categorías temáticas dentro del conjunto de intervenciones.

D.1 Notación

- $C = \{1, \ldots, 6\}$: conjunto de categorías temáticas.
- D: conjunto de todas las participaciones.
- $D_p \subseteq D$: subconjunto de participaciones del parlamentario p.
- participaciones(d, c): número de apariciones de la categoría c en la participación d.

D.2 Totales globales por categoría

Para cada categoría $c \in C$ se suma la cantidad total de documentos pertenecientes a esta en todo el corpus:

$$G_c = \sum_{d \in D} \text{participaciones}(d, c).$$

El total global del corpus es

$$G = \sum_{c \in C} G_c.$$

D.3 Ponderador global

El *ponderador* o peso relativo de la temática c en el corpus se define como

$$w_c = \frac{G_c}{G}, \qquad c \in C.$$

Un valor grande de w_c indica que el tema es frecuente en el conjunto total de documentos.

D.4 Totales por persona

Para un parlamentario p, el número de documentos por categoría es

$$P_{p,c} = \sum_{d \in D_p} \text{participaciones}(d, c),$$

y considerando todas las categorías para una persona

$$P_p = \sum_{c \in C} P_{p,c}.$$

D.5 Valor ponderado por persona y categoría

Se ajusta el conteo personal a la frecuencia global del tema:

$$V_{p,c} = \frac{P_{p,c}}{w_c} = \frac{P_{p,c}\, G}{G_c}.$$

Si $V_{p,c} > P_{p,c}$, el parlamentario habla *más* de lo esperado para esa categoría; si $V_{p,c} < P_{p,c}$, habla menos.

D.6 Relevancia interna preliminar

Para neutralizar el efecto del volumen total del parlamentario, se normaliza dentro de su propio conjunto de participaciones:

$$R'_{p,c} = \frac{V_{p,c}}{P_p}.$$

Cuando $P_p = 0$ (autor sin participaciones) se considera $P_p = 1$ para evitar división por cero.

D.7 Normalización final a porcentaje

Se asegura que las relevancias sumen 100 %:

$$S_p = \sum_{c \in C} R'_{p,c}, \qquad R_{p,c} = \frac{R'_{p,c}}{S_p} \times 100.$$

De esta manera, el vector $\{R_{p,c}\}_{c \in C}$ describe, en porcentaje, la importancia relativa de cada temática para el parlamentario p, considerando tanto su perfil como la distribución global del corpus.

Al mismo tiempo, $R_{p,c}$ (0–100 %) resume qué tan relevante es cada tema para el autor, ajustado por su propio volumen y por la distribución general.

Con estos indicadores se puede identificar rápidamente en qué temáticas cada parlamentario sobresale y comparar perfiles entre distintas personas.

Anexo E

Cálculo de polarización en votaciones de proyectos de ley

En el contexto de las votaciones legislativas, la polarización se definirá como la *falta de acuerdo en un tema*, que provoca que el universo de votantes se agrupe en dos posturas políticamente opuestas (diferencia entre grupos). El nivel de polarización es máximo cuando existen dos grupos con una cantidad equivalente de votantes enfrentados, mientras que es mínimo cuando el universo de votantes emite la misma opción. La Figura E.1 muestra la polarización para varios porcentajes de votos *a favor*/*en contra*.

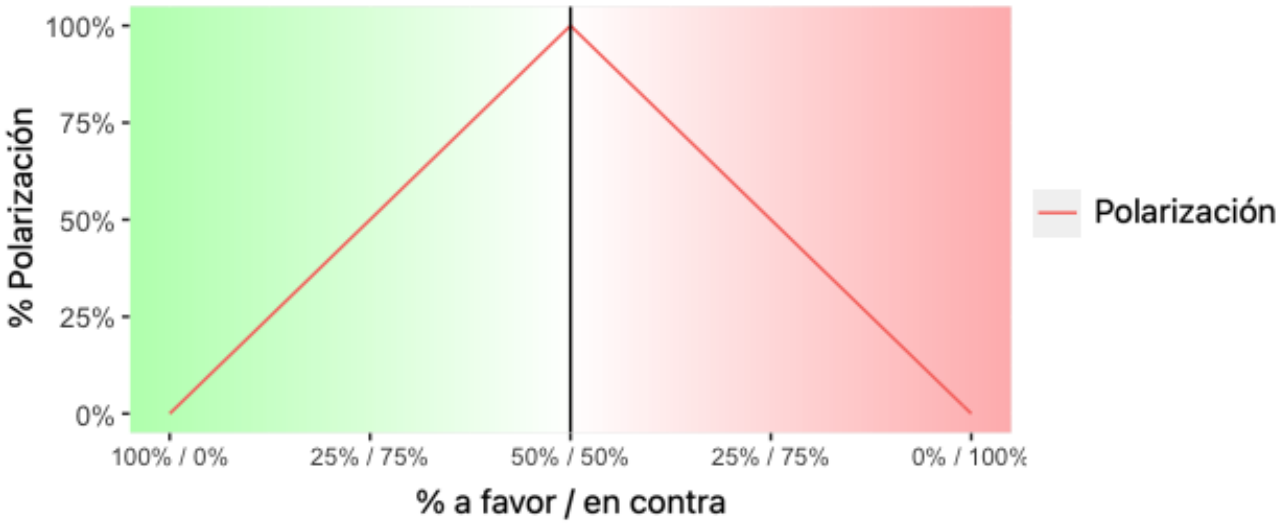

Figura E.1: Comportamiento de la medida de polarización

Para calcular la polarización, solo se consideran los valores extremos (es decir, *a favor* y *en contra*); otros tipos de voto se omiten o se normalizan a una de estas dos opciones. Esto se debe a que el significado de otro tipo de voto siempre depende del contexto político; por ejemplo, la ausencia y la abstención pueden tener fundamentos distintos. En la práctica, la aprobación de la votación se logra al obtener un cierto *quórum*, lo que se traduce en contar con suficientes votos a favor.

Así, la fórmula para calcular el índice de polarización es:

$$C_f = \frac{N_f}{N_f + N_c} \wedge C_c = \frac{N_c}{N_f + N_c} \tag{E.1}$$

donde:

- C_f corresponde al coeficiente de polarización para los votos a favor
- C_c corresponde al coeficiente de polarización para los votos en contra
- N_f corresponde al total de votos a favor
- N_c corresponde al total de votos en contra

$$P_g = 1 - \sigma_p * \sqrt{2} \tag{E.2}$$

donde:

- P_g corresponde al grado de polarización dentro del grupo en la votación
- σ_p corresponde a la desviación estándar del conjunto C_f, C_c

Anexo F

Cálculo de alineamiento político en votaciones de proyectos de ley

Alineamiento político se definirá como una característica que describe el grado de convergencia o coincidencia que ocurre dentro de un grupo de individuos respecto a cierta opinión (consistencia intragrupal). Otros términos que se consideran sinónimos de alineamiento político (o simplemente alineamiento) son cohesión y disciplina partidaria [Hug, 2010].

Esta medida puede utilizarse tanto a nivel grupal (partido político o coalición), personal (miembro del Congreso en función del grupo), por proyecto de ley o por evento de votación. Para el caso de este trabajo, cuando los miembros del Congreso votan sobre proyectos de ley, el alineamiento político describe el grado de similitud en los votos de un grupo de parlamentarios de un mismo partido.

Formalmente, el alineamiento grupal es:

$$\begin{aligned} A_g &= \frac{\sum_{i=1}^{n} \frac{A_i * N_i}{N}}{N} \\ &= \frac{\sum_{i=1}^{n} N_i^2}{N^2} \end{aligned} \tag{F.1}$$

donde:

- A_g: alineamiento grupal;
- A_i: alineamiento del subgrupo de individuos que votaron por la opción i;
- N_i: número total de individuos que votaron por la opción i;
- N: número total de individuos en el grupo.

donde A_i se define como:

$$A_i = \frac{N_i}{N} \tag{F.2}$$

donde:

- A_i: alineamiento dentro del grupo de quienes votaron por la opción i:

- N_i: número total de individuos que votaron por la opción i;
- N: número total de individuos en el grupo.

Para ilustrar, si dentro del mismo grupo, en una votación específica todos los individuos votan en contra, el alineamiento del grupo es 100%, dado que todos votan de la misma manera. En otro escenario hipotético, si la mitad de los individuos del mismo grupo (por ejemplo, el mismo partido) vota a favor y la otra mitad en contra, el alineamiento grupal es 50%, pues globalmente el grupo tuvo una opinión dividida, aunque internamente sí hubo alineamiento.

La literatura de ciencias sociales menciona el Rice Index [Rice, 1928] (y variaciones [Desposato, 2003]) para calcular la cohesión o grado de acuerdo dentro de un evento de votación. Sin embargo, este indicador solo permite obtener una única medida para un grupo completo en análisis (p. ej., un partido político), penalizando al grupo completo por las diferencias internas. El coeficiente de alineamiento político que utilizamos permite asociar un valor independiente a cada persona y voto, así como a todo el proyecto de ley, obteniendo valores más representativos. Esto permite caracterizar a cada miembro del Congreso con medidas asociadas a su alineamiento y al valor de sus votos. Ofrece un rango de aplicación más amplio que el Rice-Index, sin realizar cálculos complejos.

Si analizamos estos casos usando el Rice-Index, el alineamiento máximo sería de 100%, pero si el voto quedara exactamente 50% dividido dentro del grupo, el valor de alineamiento sería 0%. La imagen F.1 describe el comportamiento del Rice-Index, Cos-Rice-Index (variante) y las medidas de Alineamiento vistas como funciones.

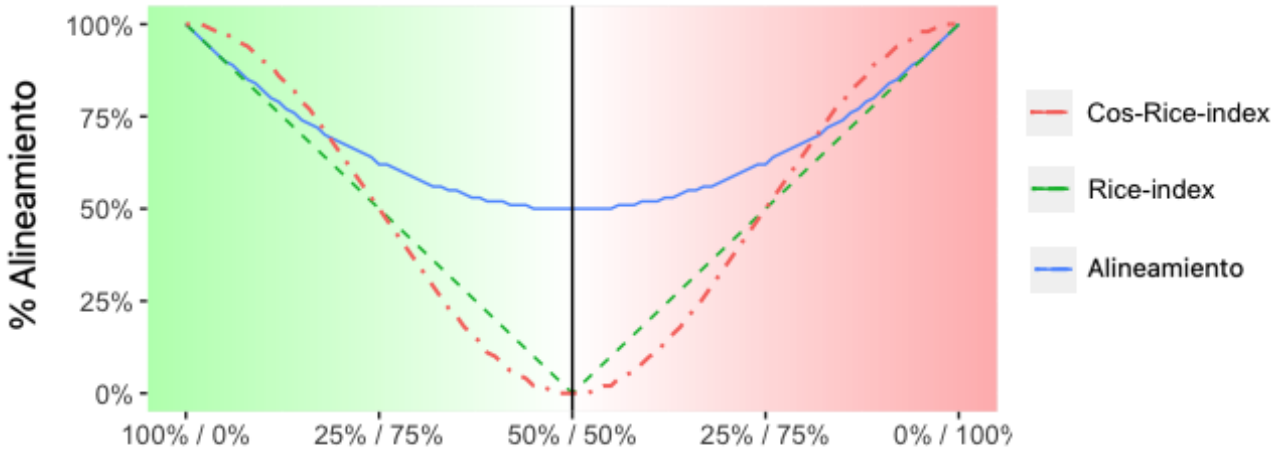

Figura F.1: Comportamiento de las medidas de alineamiento político

Anexo G

Gráfico de tiempos por tipo de pregunta

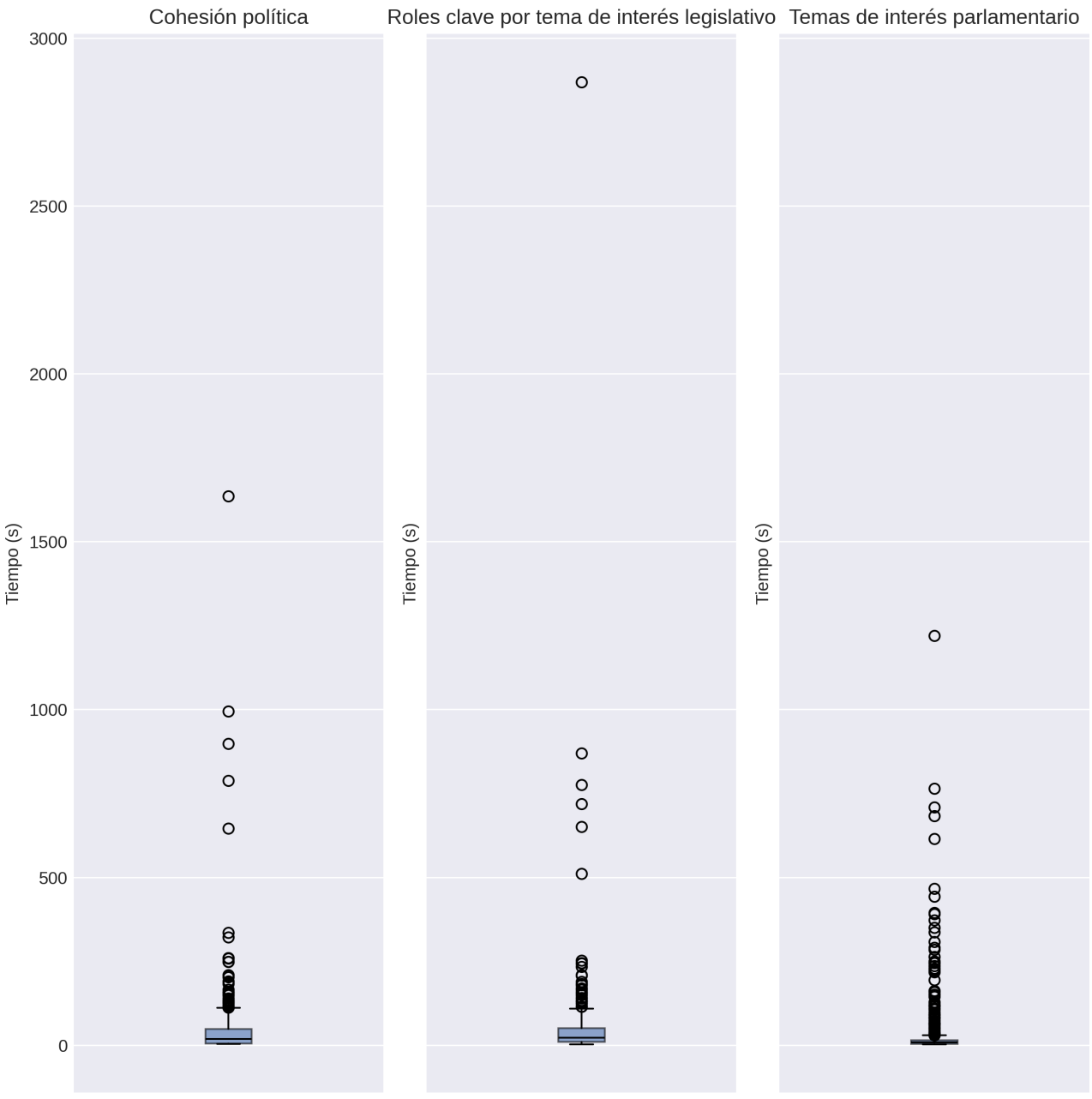

Figura G.1: Distribución de tiempos de respuesta por tipo de instrumento considerando valores atípicos

Anexo H

Métricas

H.1 Métricas de evaluación de resultados en aprendizaje automático

H.1.1 Recall (Exhaustividad)

El *recall* mide la capacidad del modelo para identificar correctamente los casos positivos dentro del conjunto de datos. Se define como

$$\text{Recall} \ = \ \frac{\text{TP}}{\text{TP} + \text{FN}},$$

donde TP (True Positives) son las predicciones positivas correctas y FN (False Negatives) las verdaderas positivas no detectadas.

H.1.2 Precision (Precisión)

La *precision* cuantifica la proporción de predicciones positivas que efectivamente corresponden a casos positivos. Su fórmula es

$$\text{Precision} \ = \ \frac{\text{TP}}{\text{TP} + \text{FP}},$$

donde FP (False Positives) son las predicciones positivas incorrectas.

H.1.3 Accuracy (Exactitud)

La *accuracy* evalúa el porcentaje total de aciertos del modelo sobre todos los casos. Se expresa como

$$\text{Accuracy} \ = \ \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

con TN (True Negatives) siendo las predicciones negativas correctas.

H.1.4 F1 Score (Puntaje F1)

El *F1 score* es la media armónica entre precision y recall, equilibrando ambos aspectos en un único valor. Se calcula mediante

$$\mathrm{F1} \;=\; 2 \times \frac{\mathrm{Precision} \;\times\; \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

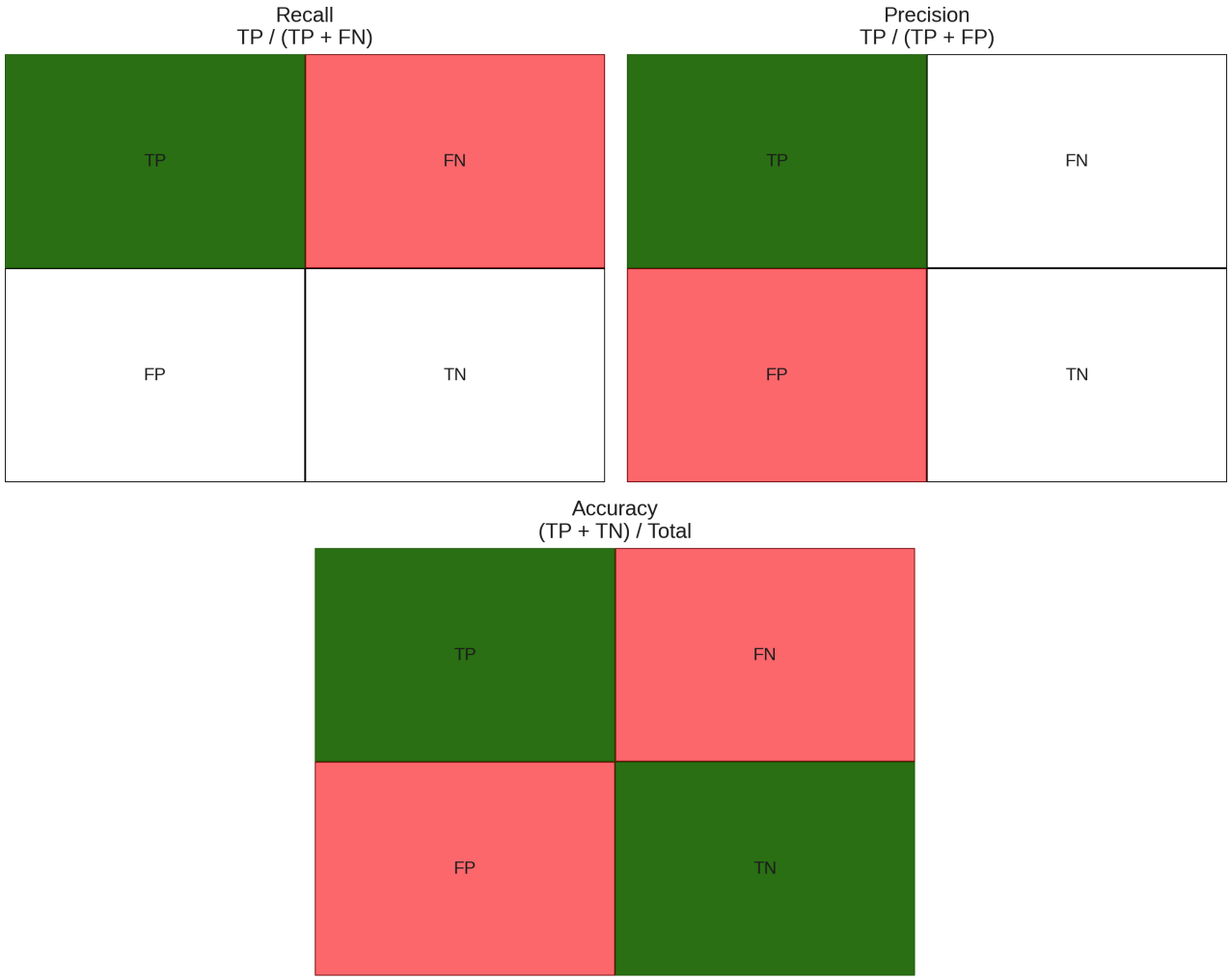

Figura H.1: Métricas de evaluación de resultados en aprendizaje automático

H.1.5 Rango recíproco medio (MRR)

El *Rango Recíproco Medio* (*Mean Reciprocal Rank*, MRR) es una métrica estándar utilizada para evaluar sistemas de recuperación de información o modelos de ranking, asignando un puntaje que varía entre 0 y 1 dependiendo de la posición del primer resultado relevante dentro de una lista ordenada de respuestas, con 0 como valor mínimo de relevancia y 1 valor máximo.

Formalmente, sea Q el conjunto de consultas o instancias de evaluación, y sea rank_i la posición del primer resultado relevante para la consulta i en su lista ordenada de resultados. El MRR se define como:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{H.1}$$

El valor de rank_i es un entero positivo que indica la posición (empezando en 1) del primer resultado relevante para la consulta i. Si el resultado relevante no aparece en la lista, su contribución puede considerarse cero o excluirse del promedio, dependiendo del criterio adoptado.

H.1.6 Curvas ROC y Área bajo la curva (AUC)

Las curvas Receiver Operating Characteristic (ROC) en el contexto de clasificación, muestran cómo se comporta un modelo al separar las clases bajo distintos umbrales de decisión, principalmente mostrando la relación entre tasas de verdaderos y falsos positivos. A partir de su cálculo es posible visualizar el Area Under Curve (AUC), el cual es un medidor de desempeño que entrega valores entre 0 y 1, donde 0 es un resultado sin aciertos, y 1 es un resultado perfecto. En general, resultados cercanos a 1 indican que el modelo posee una buena separabilidad.

La curva ROC representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR), definida como:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{H.2}$$

donde TP, FP, TN y FN representan verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, respectivamente.

El valor de AUC corresponde al área bajo esta curva ROC y puede interpretarse como la probabilidad de que el modelo asigne una puntuación más alta a una instancia positiva que a una negativa seleccionadas al azar. Matemáticamente, un AUC de 1.0 indica una separación perfecta, mientras que un valor de 0.5 corresponde a un clasificador aleatorio.

Esta métrica es especialmente útil en contextos con clases desbalanceadas, ya que no depende de un umbral de decisión específico y evalúa la capacidad de discriminación del modelo a lo largo de todos los posibles umbrales.

H.1.7 Curvas Precision-Recall

En la misma linea anterior, la curva de Precision-Recall permite visualizar el equilibrio entre precisión y recall, es decir qué tanto cae una de las métricas cuando aumenta otra, por ejemplo, qué tanto cae la precisión cuando se aumenta el recall (exhaustividad) o cuánto cae la exhaustividad cuando cae la precisión. A partir de esto es posible calcular el área bajo la curva de Precision-Recall (AUC-PR o AP), en donde un valor vercano a 1 representa un excelente balance entre precisión y recall y un balor bajo indica que el modelo recupera muchos falsos positivos o deja escapar muchos verdaderos positivos.

Anexo I

Análisis de tópicos LDA en muestreo preliminar

Para el análisis de los tres conjuntos de datos (Congreso Nacional, Prensa y Twitter), se implementó un modelo de tópicos mediante LDA utilizando la biblioteca Gensim[1] escrita en lenguaje Python.

Lo primero que se realizó para el modelado de los tópicos, es la carga de los tres conjuntos de datos asociados a parlamentarios. A nivel de preprocesamiento se realizó una normalización y limpieza del texto, eliminación de palabras vacías en español y la utilización de un vectorizador TF-IDF.

En un principio, se implementaron modelos de tópicos diferentes por cada conjunto de datos, pero dado que podrían existir tópicos específicos solo en algunos de los conjuntos, no se prosperó en este enfoque, además porque no sería posible comparar los modelos de tópicos al variar sus palabras. De esta manera, el enfoque adoptado fue el desarrollo de un modelo de tópicos común asociado a la unión de los conjuntos de datos.

Debido a que la cantidad de palabras de los documentos del Congreso Nacional y de Prensa eran, en términos generales similares, pudieron ser incorporados a la fase de entrenamiento de forma directa, mientras que los datos provenientes de Twitter, tuvieron que ser preprocesados para ser unidos en documentos de largo aproximado en 300 palabras. Esto fue necesario porque los largos de los tweets eran mucho más cortos que en el caso de los otros dos conjuntos, y en muchos casos extrenadamente cortos, lo cual derivaba en que el algoritmo no era capaz de calcular sobre ellos y la ejecución terminaba fallando.

Con lo anterior en consideración, el primer paso para el cálculo de los tópicos fue calcular el número óptimo de típicos del conjunto, por lo cual se iteró el algoritmo buscando desde 2 hasta 50 tópicos a fin de maximizar la medida de *coherencia semántica*, la cual pondera con valores más altos al número de tópicos que muestra un grado de relación más fuerte entre palabras y temas. Con esto se pudo identificar que, en función de esta medida, el número óptimo de tópicos era de 25 tópicos en el conjunto de datos unidos. La figura I.1 muestra el cálculo del valor de coherencia semántica para los distintos números de tópicos, su punto máximo y el punto donde la curva deja de subir.

El primer ejercicio luego de generar el modelo de tópicos fue la obtención del tópico principal (con mayor probabilidad) por documento, a fin de verificar la distribución de los tópicos más relevantes por conjunto de datos. El gráfico de la figura I.2 muestra la distribución porcentual de

[1] `https://radimrehurek.com/gensim/index.html`

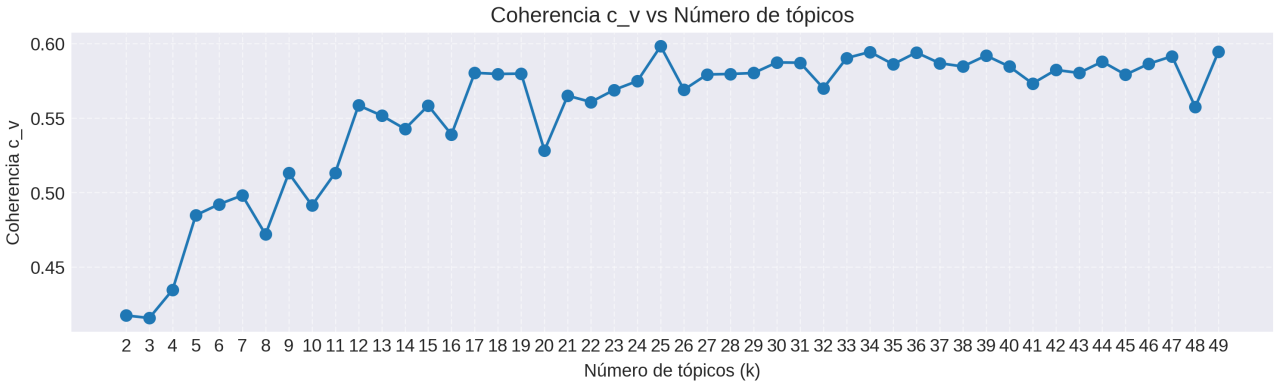

Figura I.1: Selección de número óptimo de tópicos mediante medida de coherencia

documentos por el tópico de mayor probabilidad detectado. En este caso, es posible visualizar que el tópico 4 y el 12 son los tópicos de mayor relevancia para twitter, los cuales agrupan casi el 80% de los datos de twitter.

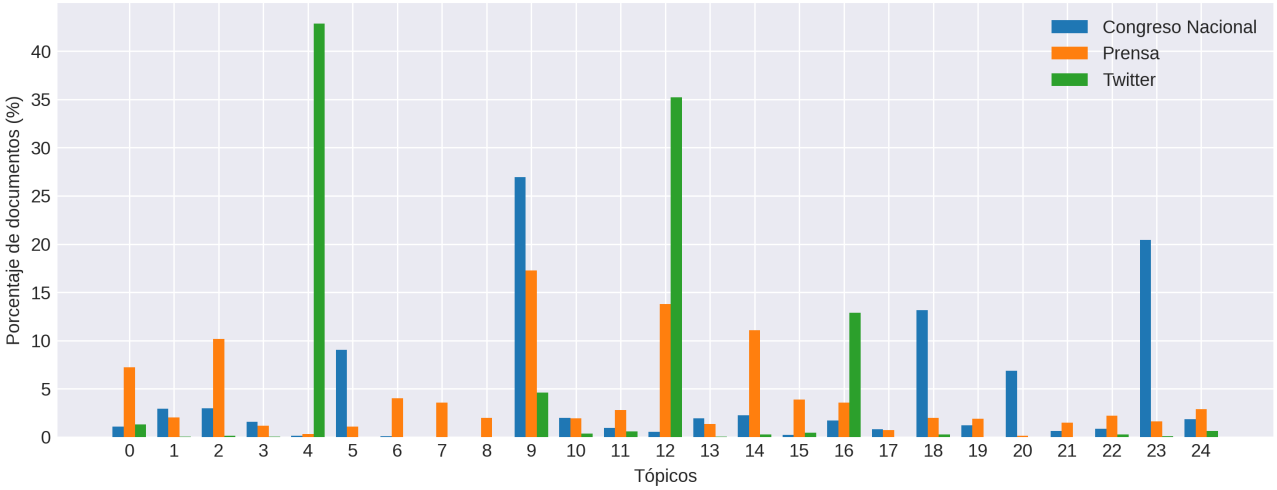

Figura I.2: Distribución porcentual de documentos por tópico de mayor probabilidad

Posteriormente, se ejecutó el modelo identificando los tópicos con una probabilidad > 0.1 en cada documento, lo cual permitió identificar múltiples tópicos por documento, asumiendo que un documento puede tener asociado más de un tópico. En este escenario, la distribución de tópicos se muestra en el gráfico de la figura I.3, donde se visualiza una distribución más homogénea que la anterior, pero de todas maneras, la cobertura que ofrece twitter sobre los dsitintos tópicos es muy inferior a los otros dos conjuntos.

El gráfico de la figura I.4 muestra un gráfico de termita donde se presentan las palabras (eje Y) asociadas a los tópicos identificados (eje X). El radio de cada circunferencia representa un mayor volumen de documentos donde se presenta esa palabra.

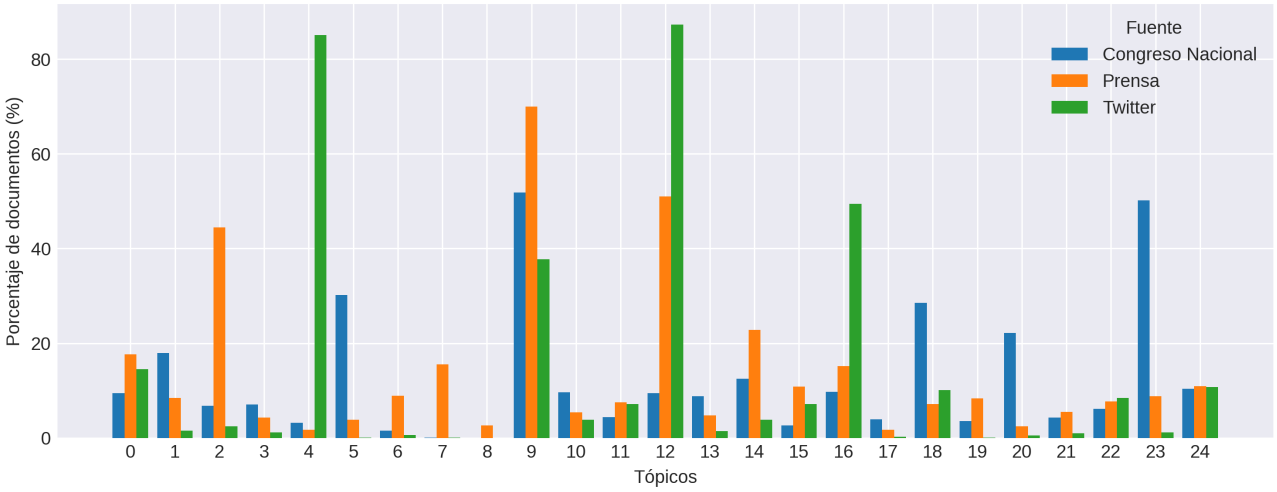

Figura I.3: Distribución porcentual de documentos asociados a múltiples tópicos

Figura I.4: Gráfico de termitas para visualización de tópicos (parte superior)

Figura I.5: Gráfico de termitas para visualización de tópicos (parte inferior)