Detección de sesgos debido al género en decisiones judiciales utilizando Procesamiento de Lenguaje Natural

Christian Javier Ratovicius

Antonela Tommasel

Trabajo presentado para obtener el título de
Magister en Tecnología Informática

Marzo, 2025

Resumen

Desde hace algunos años existe la preocupación por la adopción de perspectivas de género en los procesos y actores del Poder Judicial. Desde un enfoque sociocultural, un estereotipo puede definirse como una visión generalizada o un pre-concepto (o incluso prejuicio) sobre las características de un grupo o sobre los roles que los individuos deben cumplir en la sociedad. En particular, los estereotipos de género se refieren a la construcción social y cultural de hombres y mujeres, en relación con sus diferentes funciones físicas, biológicas y sociales. De acuerdo con diferentes estudios, se ha mostrado que los estereotipos pueden incidir en el razonamiento judicial. La hipótesis de trabajo final es que los sesgos (en particular los relacionados al género) pueden ser identificados a partir de análisis de texto, utilizando de técnicas de Procesamiento de Lenguaje Natural, que permitan analizar sus contextos gramaticales, sintácticos y lingüísticos.

En este contexto, en este trabajo final de maestría, se propone la evaluación de diferentes técnicas de procesamiento de lenguaje natural y *machine learning* para la detección de estereotipos de género en decisiones judiciales de manera automática.

El objetivo final es asistir a la revisión y análisis de resoluciones judiciales con el fin de detectar situaciones y razonamientos que evidencien dichos sesgos de género. La evaluación experimental realizada permitió mostrar que, si bien las aplicaciones como la metodología utilizada se adecuan al objetivo del trabajo, es necesario profundizar en representaciones especificas del lenguaje legal español.

Palabras claves

machine learning, procesamiento de lenguaje natural, sesgos de género

Dedicatoria

A mis padres, que me han dado la vida, la educación y el amor necesarios para llegar hasta aquí. Gracias por su apoyo incondicional, por sus consejos y enseñanzas.

A mis hijos Ámbar, Abril y Abel, que ha sido mis compañeros de viaje en este largo camino. Gracias por su amor, su comprensión y apoyo.

A mis abuelos Salvador, Amelia, Casimira y mi madrina Victoria, que me ayudaron con su amor y aliento para superarme día a día.

Reconocimientos

Quiero agradecer inicialmente a Carlos Neil, quien me acompaño desde el principio, brindándome todo su apoyo y plena confianza, guiarme y aconsejarme siempre.

A mi directora Antonela Tommasel, por ser una excelente profesional, ya que sin ella hubiese sido imposible realizar este trabajo.

Extiendo mi agradecimiento a quienes participaron en el proyecto de investigación que culminó exitosamente con este trabajo.

Índice General

Índice de Gráficos

Índice de Tablas

Capítulo 1

Introducción

En las últimas décadas se ha avanzado en la detección y eliminación de diversos tipos de sesgo, incluyendo el sesgo de género. En este marco, en los últimos años ha cobrado relevancia la incorporación de una perspectiva de género en los diversos ámbitos de la administración pública, lo que abarca también al Poder Judicial, sus procesos y sus actores. Un ejemplo de las acciones del Estado para el abordaje de esta problemática es la introducción de forma obligatoria de la capacitación conocida como “Ley Micaela”[1].

Un estereotipo se define como una creencia generalizada sobre una categoría específica de personas. Representa una expectativa que se puede tener acerca de cada individuo perteneciente a un grupo particular. Este tipo de expectativa puede variar, abarcando aspectos como la personalidad, las preferencias, la apariencia o las capacidades del grupo. Los estereotipos, a menudo, son generalizaciones excesivas, inexactas y resistentes a nueva información. Estos cubren una amplia gama de creencias sociales y desempeñan un papel crucial en la construcción de la identidad social (González Gavaldón, 1999). Según la autora, el concepto de estereotipo es uno de los más controvertidos, aceptando la propuesta de (Mackie, 1973), quien los define como creencias populares sobre los atributos que caracterizan a un grupo social y sobre las cuales existe un acuerdo general.

Los estereotipos de género se definen como creencias, actitudes o expectativas preconcebidas sobre cómo deben comportarse, verse o actuar las personas en función de su género. Estos estereotipos hacen referencia a la construcción de las identidades de hombres y mujeres, basada en las diferencias de sus funciones físicas, biológicas, sexuales y sociales. Aunque pueden tener un

[1] https://www.argentina.gob.ar/generos/ley-micaela

impacto significativo en la vida de las personas, también pueden limitar sus oportunidades y opciones, además de contribuir a la discriminación y la violencia. Por ejemplo, frases como “las mujeres se preocupan” pueden tener consecuencias negativas, ya que refuerzan la mirada condescendiente para con dichos grupos (Timmer, 2016). Si una sociedad estereotipa a las mujeres como emocionales e irrazonables, es más probable que las mujeres sean juzgadas con mayor dureza en situaciones emocionales o se les den menos oportunidades para roles de liderazgo, lo que refleja un sesgo de género. Se trata de un término genérico que abarca tanto a mujeres y sus subgrupos como a hombres y sus subgrupos. Por lo tanto, su significado es fluido y cambia con el tiempo, las culturas y las sociedades. En este sentido, la relación entre los conceptos de “estereotipo”, “prejuicio” y “discriminación” es muy estrecha. Este enfoque parte de una concepción clave: la consideración de que estos términos están íntimamente ligados al concepto de actitud, entendido como un fenómeno compuesto por tres componentes: cognitivo, afectivo y conductual.

La detección de sesgo de género es una de las áreas más prometedoras del Procesamiento del Lenguaje Natural (NLP, de sus siglas en inglés, *Natural Language Processing*), lo que ha generado un gran interés en los últimos años. Sus posibles aplicaciones resultan útiles y diversas, lo que convierte a este campo en un área de interés tanto para diversas instituciones como para medios sociales variados. El crecimiento de la tecnología y el avance en distintas técnicas ha facilitado la proliferación de herramientas para la identificación de este fenómeno, aplicadas en instituciones como el legislativo, el ámbito educativo, entre otros, así como en espacios diversos como artículos de prensa, blogs, redes sociales y foros. El gran volumen de dicho material disponible en la actualidad hace inviable su procesamiento manual, pero, a su vez, fomenta la construcción y perfeccionamiento de métodos automáticos que faciliten esta labor. En este contexto, este proyecto aborda la definición, análisis y evaluación de diversas técnicas para la identificación de sesgo de género de manera automática o semi automática en las decisiones judiciales.

1.1 Hipótesis y objetivos

El Procesamiento de Lenguaje Natural (NLP) es un área de investigación que explora las capacidades de una computadora para comprender y manipular el lenguaje natural, ya sea escrito u oral, con el fin de facilitar su uso a través de la tecnología. En el ámbito jurídico, el NLP se aplica en una variedad de tareas, incluyendo el análisis de documentos legales, la búsqueda de información legal y la revisión de documentos que contienen grandes volúmenes de datos (Bonina, 2020). También se puede observar su utilidad en la predicción de posibles resultados a partir del análisis y revisión de precedentes judiciales previos (por ejemplo, *Lexmachina.com*[2]), entre otras aplicaciones.

En este contexto, la hipótesis de este trabajo sostiene que los *sesgos, en particular los relacionados con el género, pueden ser identificados mediante el análisis de texto utilizando técnicas de NLP que permitan analizar los contextos gramaticales, sintácticos y lingüísticos*. El objetivo de la investigación es *explorar el uso de NLP, aprendizaje automático supervisado y los avances en modelos de lenguaje de gran escala (LLMs) para detectar sesgos en sentencias judiciales*. En particular, se buscará identificar patrones lingüísticos y otras características que puedan revelar sesgos inconscientes contra mujeres u otras minorías de género. Con ello, se espera contribuir a los esfuerzos en curso para promover la justicia y la equidad en el sistema judicial, garantizando que todas las personas reciban un trato justo y equitativo conforme a la ley.

El presente trabajo se basa en los principios establecidos por la publicación (Ratovicius, Diaz-Pace, & Tommasel, 2024), que proporciona una base sólida para el desarrollo de la investigación al ofrecer un enfoque detallado sobre temas relevantes. La aplicación de estos conceptos ha permitido estructurar la metodología y el análisis propuestos en este estudio, garantizando que el marco teórico esté respaldado por fuentes confiables y actualizadas.

A largo plazo, se espera que este desarrollo permitirá asistir a los distintos actores involucrados en el ámbito jurídico, facilitando la revisión y análisis de resoluciones con el objetivo de detectar situaciones y razonamientos que evidencien discriminación por motivos de género. Se prevé que este método, en conjunto con

[2] https://lexmachina.com/legal-analytics/

una herramienta[3] prototipo semi-automática, permitirá reducir los tiempos de análisis de los pronunciamientos judiciales para verificar si presentan los mencionados sesgos, lo cual es obligatorio en virtud de diversas normas de derecho internacional e interno, contribuyendo tanto a la tarea de detección y corrección de sesgos como a la implementación de medidas preventivas frente a dicha problemática. Basándose en los resultados de la herramienta, un experto podría contar con asistencia para realizar estudios sistemáticos sobre determinados organismos (como un juzgado o una fiscalía en particular), sobre una materia específica (como procesos vinculados a violencia doméstica), o sobre un tipo de resolución determinado (como sentencias definitivas).

1.2 Principales contribuciones a la temática

Las principales contribuciones del trabajo son:

- El desarrollo de una metodología que presenta las etapas necesarias para identificar patrones de lenguaje que describan el sesgo de género.

- La creación de un modelo para la detección automática de sesgo de género en sentencias judiciales.

1.3 Estructura General

La estructura general del trabajo está compuesta por seis capítulos, cada uno de los cuales desarrolla los aspectos principales del estudio, describiendo las técnicas y tecnologías utilizadas para desarrollar el prototipo.

[3] bonzokimba/deteccionsesgo: Detección de sesgos debido al género en decisiones judiciales utilizando PNL

En el **Capítulo 2**, Marco Teórico, se contextualiza el trabajo de investigación en su contexto teórico y disciplinario. Se aborda la detección de sesgos de género en el sistema judicial y se exploran las oportunidades de utilizar NLP para hacer frente a estos sesgos. Se inicia con una breve descripción del sesgo de género en la toma de decisiones judiciales, seguido de un análisis de las ventajas de emplear el NLP para detectar y abordar dichos sesgos. Luego, se presenta una explicación detallada de los métodos de NLP utilizados en el estudio y los resultados obtenidos. Finalmente, se discuten las implicaciones de estos resultados y el potencial del NLP para mejorar la toma de decisiones en el sistema judicial.

En el **Capítulo 3**, Trabajos relacionados, se analizan investigaciones previas vinculadas a distintos tipos de detección utilizando técnicas de NLP. A partir de esta revisión, se realiza una comparación entre la propuesta presentada y los trabajos relacionados.

En el **Capítulo 4**, Enfoque de la solución, se establece como objetivo principal el desarrollo de un sistema capaz de detectar el sesgo de género en decisiones judiciales, resumiendo la idea general de la implementación de la solución.

En el **Capítulo 5**, se describe la evaluación experimental llevada a cabo sobre los distintos clasificadores implementados, y se discuten los diversos resultados alcanzados.

En el **Capítulo 6**, finalmente, se presentan las conclusiones del proyecto y las principales líneas de trabajo futuro.

Capítulo 2

Marco teórico

En este capítulo se presentan los conceptos fundamentales relacionados con el desarrollo de esta investigación. Se analiza lo existente sobre los temas relevantes y se establece un punto de partida para la investigación en curso. Se definen los conceptos clave, tales como el sesgo y sus diferentes tipos, así como las etapas del Procesamiento de Lenguaje Natural (NLP), los modelos de representación de texto y las herramientas de aprendizaje automático.

Comúnmente, un *estereotipo* se refiere a una visión generalizada o una preconcepción acerca de los atributos o características de los miembros de un grupo particular, o sobre los roles que esos miembros deben desempeñar. El término "estereotipo" fue utilizado por primera vez en el año (Lippmann, 1922), utilizado para describir un proceso de impresión. Se adaptó metafóricamente en las ciencias sociales para explicar cómo las personas generan preconceptos sobre otras, como si estas fueran una reproducción exacta de un molde. Desde esta perspectiva, los seres humanos no perciben el "mundo exterior" tal y como es; más bien, preconfiguran "imágenes mentales" o estereotipos, que luego utilizan para darle sentido a la realidad que experimentan. Según esta definición, los estereotipos suponen que todos los miembros de un grupo social comparten ciertos atributos o características específicas. Para que una generalización se califique como un estereotipo, no importa si dichos atributos son o no comunes entre los miembros del grupo, ni si estos realmente poseen tales características. El elemento clave es que, al presuponer que el grupo específico tiene esas características, se asume que una persona, por el simple hecho de pertenecer al grupo, actuará conforme a la visión generalizada o preconcepción existente. Así, todas las dimensiones de la personalidad que hacen única a una persona son filtradas a través de este lente, en función de la visión generalizada sobre el grupo con el cual se le identifica.

El sistema judicial desempeña un papel fundamental en la sociedad, siendo clave para defender la justicia y garantizar que las personas reciban un trato justo y equitativo. Sin embargo, diversos estudios (Gillis, 2021) han señalado preocupaciones acerca de la influencia de los prejuicios en las decisiones judiciales, especialmente en cuestiones de género. A pesar de los esfuerzos por promover la neutralidad de género y la objetividad dentro del sistema judicial, existe creciente evidencia de que podría haber un sesgo de género presente en las decisiones judiciales. Estos sesgos provienen de diversas fuentes, como estudios de decisiones judiciales, encuestas a jueces y testimonios de abogados y litigantes.

Uno de los estudios más completos sobre el sesgo de género en las decisiones judiciales (Molina, 2022) demuestra cómo esta problemática se manifiesta a través de las distintas etapas judiciales. Los resultados mostraron que las mujeres tienen más probabilidades de perder casos de custodia, recibir sentencias más cortas por delitos penales y obtener menos compensación en casos de lesiones personales. Asimismo, se observó que los jueces eran más propensos a creer el testimonio de testigos masculinos que el de testigos femeninos.

Otro estudio reciente[4] determinó que los jueces pueden ser incluso más parciales que el público general al tomar decisiones sobre casos en los que se cuestionan los roles de género tradicionales. Los resultados indicaron que los jueces que apoyaban estos roles tradicionales manifestaban una mayor propensión a otorgar la custodia a las madres, incluso cuando los padres tenían las mismas calificaciones. Por otra parte, encuestas realizadas a jueces han revelado que muchos creen que las mujeres son menos competentes que los hombres como abogadas y juezas (Miller, 2018).

Los trabajos literarios sugieren la posible existencia de un sesgo de género en las decisiones judiciales, lo cual tiene un impacto significativo en las vidas de las mujeres, quienes podrían verse más afectadas por las decisiones tomadas en los tribunales. Por tanto, resulta necesario investigar esta situación y trabajar en la reducción o eliminación de los prejuicios sexistas en el ámbito judicial.

En los últimos años, el NLP ha emergido como una herramienta poderosa para descubrir distintos fenómenos, como por ejemplo sesgos, en datos textuales,

[4] https://www.sciencedaily.com/releases/2018/04/180403085049.htm

incluidas las decisiones judiciales. Las técnicas de NLP pueden facilitar la identificación de patrones y tendencias en el uso del lenguaje que podrían revelar sesgos sistémicos o factores problemáticos que contribuyan a resultados injustos.

El uso del NLP se ha implementado en diversos proyectos a nivel internacional. Un ejemplo destacado es MIREL[5], una iniciativa financiada por la Unión Europea que tiene como objetivo desarrollar herramientas y técnicas para la minería y comprensión de textos legales. Otro proyecto relevante es el que se lleva a cabo en la Facultad de Derecho de Harvard[6] en colaboración con Ravel Law, cuyo objetivo es digitalizar la colección completa de jurisprudencia estadounidense, considerada la base de datos más completa y autorizada sobre casos judiciales en Estados Unidos, fuera de la Biblioteca del Congreso.

En Argentina, también se observan desarrollos similares. *Sherlock Legal*[7], por ejemplo, incluye un buscador con inteligencia artificial para interactuar con la base de fallos Dial.com. Además, ofrece funcionalidades como el "Anonimizador automático" y "Fallos similares". Por otro lado, *Prometea*[8], impulsado por el Ministerio Público Fiscal de la Ciudad de Buenos Aires y la Facultad de Derecho de la UBA, para el usuario la interfaz permite, a través de la interacción en un chat con una voz artificial, se puede realizar la consulta de expedientes ingresados. Al indicar el número de expediente, le lee al usuario la carátula del caso e indica una posible solución de este. Es importante destacar que el uso de inteligencia artificial en esta herramienta no reemplaza el juicio humano, sino que actúa como una herramienta de apoyo para los jueces y el personal judicial, optimizando el tiempo y reduciendo el margen de error en procesos rutinarios.

2.1 Tipos de Sesgo

Existen diferentes tipos de sesgos que pueden afectar la toma de decisiones, la percepción de la realidad y la interpretación de la información. En este trabajo, se

[5] https://www.mirelproject.eu/index.html
[6] https://www.lawnext.com/2016/06/harvard-ravel-digitization-project-adds-mass-delaware-cases.html
[7] https://www.sherlocklegal.com.ar/
[8] https://es.wikipedia.org/wiki/Prometea

aborda específicamente el *sesgo de género*, el cual se refiere a creencias preconcebidas sobre las características y comportamientos típicos de hombres y mujeres. Entre los ejemplos más comunes de sesgos de género se incluyen:

- **Discriminación de género**. El trato desigual hacia las personas basado en su género. Esto puede incluir el acceso desigual a oportunidades laborales, educativas y económicas.
- **Exclusión y marginación**. La falta de inclusión de personas de ciertos géneros en ciertos ámbitos y oportunidades. Este tipo de sesgo puede ser especialmente evidente en áreas de liderazgo y toma de decisiones.
- **Acoso sexual**. El comportamiento no deseado de naturaleza sexual que ocurre en el lugar de trabajo, en la educación, en el hogar o en espacios públicos.
- **Violencia de género**. La violencia dirigida hacia personas de un género específico, como la violencia doméstica o la violación.

Estos sesgos pueden conducir a un trato diferencial hacia las personas en función de su género, lo que puede generar consecuencias significativas en términos de igualdad de oportunidades, discriminación y violencia. Combatir este tipo de sesgo es esencial para promover la igualdad de género y erradicar la discriminación. La

Tipos de sesgo	Ejemplos de creencias y características	Reflejo en algunas prácticas
De sexo	Centrados en las diferencias físicas y psicológicas entre hombres y mujeres. Ej.: Los hombres son más fuertes, racionales y firmes, mientras que las mujeres son irracionales, emocionales.	Prohibición de acceso a ciertos trabajos (fuerza física, trabajos nocturnos). Los hombres dominan los espacios públicos y de toma de decisiones, mientras que las mujeres ocupan espacios privados. Las mujeres no son consideradas negociadoras eficaces.
Sexual	Relacionado con las características sexuales que se atribuyen a los géneros, y la interacción sexual entre ellos. Ej.: La sexualidad vinculada a la procreación.	Prohibición del matrimonio entre personas del mismo sexo.
Roles de genero	Comportamientos atribuidos y esperados de hombres y mujeres, basados en construcciones sociales y culturales. Ej.: En las familias, los hombres son proveedores y las mujeres cuidan de los hijos/as.	Normas sociales que dictan la distribución de derechos y obligaciones. Superar estos roles genera una resistencia a las prácticas tradicionales.
Compuesto	Sesgos que interactúan con otros estereotipos, asignando roles y	Negar la posibilidad de adopción o la tenencia de hijos/as a mujeres

	características a grupos diversos de mujeres. Ej.: Mujeres solteras, mujeres lesbianas no son buenas madres.	que no cumplen con los roles tradicionales

Tabla 1 presenta un ejemplo[9] de cómo los sesgos de género se manifiestan en la sociedad.

Un caso importante que ilustra la profundización de los estereotipos de género es el libro "Estereotipos de Género" (Cusack, 2010), donde se analiza cómo los estereotipos asignados a las personas, especialmente a las mujeres, influyen en sus roles y oportunidades dentro de la sociedad. El texto examina cómo estos estereotipos degradan a las mujeres al asignarles roles serviles. Además, se destaca la importancia de erradicar estos estereotipos como una medida crucial para eliminar todas las formas de discriminación hacia las mujeres. En este marco, la *Convención sobre la Eliminación de todas las Formas de Discriminación contra la Mujer (CEDAW)* resalta la obligación de los Estados de eliminar estos estereotipos. Mediante un proceso legal transnacional, se busca desarrollar el concepto de *igualdad transformadora* que permita superar las barreras culturales y sociales que perpetúan la desigualdad de género.

Tipos de sesgo	**Ejemplos de creencias y características**	**Reflejo en algunas prácticas**
De sexo	Centrados en las diferencias físicas y psicológicas entre hombres y mujeres. Ej.: Los hombres son más fuertes, racionales y firmes, mientras que las mujeres son irracionales, emocionales.	Prohibición de acceso a ciertos trabajos (fuerza física, trabajos nocturnos). Los hombres dominan los espacios públicos y de toma de decisiones, mientras que las mujeres ocupan espacios privados. Las mujeres no son consideradas negociadoras eficaces.
Sexual	Relacionado con las características sexuales que se atribuyen a los géneros, y la interacción sexual entre ellos. Ej.: La sexualidad vinculada a la procreación.	Prohibición del matrimonio entre personas del mismo sexo.
Roles de genero	Comportamientos atribuidos y esperados de hombres y mujeres, basados en construcciones sociales y culturales. Ej.: En las familias, los hombres son proveedores y las mujeres cuidan de los hijos/as.	Normas sociales que dictan la distribución de derechos y obligaciones. Superar estos roles genera una resistencia a las prácticas tradicionales.

[9] https://acnudh.org/load/2020/03/Poder-Judicial_PDF-2.pdf

Compuesto	Sesgos que interactúan con otros estereotipos, asignando roles y características a grupos diversos de mujeres. Ej.: Mujeres solteras, mujeres lesbianas no son buenas madres.	Negar la posibilidad de adopción o la tenencia de hijos/as a mujeres que no cumplen con los roles tradicionales

Tabla 1: Ejemplos de estereotipos nocivos

Estos sesgos pueden conducir a un trato diferencial hacia las personas en función de su género, lo que puede generar consecuencias significativas en términos de igualdad de oportunidades, discriminación y violencia. Combatir este tipo de sesgo es esencial para promover la igualdad de género y erradicar la discriminación.

Estos sesgos no solo se manifiestan en el lenguaje de forma explícita, sino también de forma implícita mediante percepciones subconscientes que no se expresan de forma directa en el lenguaje compartido (Greenwald & Banaji, 1995), debiendo ser interpretados en función del conocimiento del contexto y cultura donde se expresa (Schmeisser-Nieto, Montserrat, & Mariona, 2022). Los estereotipos implícitos se manifiestan a través de elecciones lingüísticas sutiles, como el uso de ciertos adjetivos, pronombres o metáforas, que refuerzan roles de género tradicionales. Por ejemplo, la frase "*las mujeres son débiles*" representa un sesgo explícito, mientras que "*la mujer lloró*" representa un sesgo implícito respecto a la debilidad emocional de las mujeres (Tenghao, Faeze, Vered, & Snigdha, 2021).

2.2 Perspectiva de género

La perspectiva de género se entiende como un enfoque que considera el género como una categoría social que influye en las experiencias, oportunidades y resultados de las personas. Reconoce que, debido a las normas y expectativas sociales asociadas con el género, mujeres y hombres experimentan el mundo de manera distinta. Este enfoque no es ajeno a los países de Latinoamérica; aunque algunos han logrado avances significativos, otros recién comienzan a implementar políticas de género.

En **Chile**[10], por ejemplo, ha existido históricamente una invisibilización de la violencia de género, situación que se refleja en el trato que algunas mujeres reciben al intentar denunciar violencia y acoso sexual. A pesar de que el proceso de denuncia es un derecho y una obligación, no siempre se brinda el apoyo necesario a las mujeres. En respuesta a esta problemática, el Ministerio de Justicia de Chile ha comenzado a incorporar de manera gradual la perspectiva de género en los tribunales penales, civiles, laborales y de familia. Este cambio se está implementando a través de estudios que vinculan la perspectiva jurídica con la dimensión social de la violencia de género.

En **Colombia**, la Corte Suprema ha abordado la perspectiva de género en sus decisiones judiciales[11], destacando la importancia de proteger los derechos de las mujeres frente a cualquier tipo de discriminación. La Corte subraya la necesidad de adoptar un enfoque diferencial para proteger a las mujeres contra la discriminación y resalta la obligación de los jueces de equilibrar el poder entre las partes en litigios. Este enfoque tiene como objetivo eliminar los sesgos de género para garantizar la justicia, especialmente en casos de violencia o discriminación. Asimismo, se menciona la importancia de flexibilizar la carga probatoria en estos casos, sin que ello implique una violación al debido proceso judicial.

En **México**, la Corte Suprema de Justicia[12], en 2013, publicó el "Protocolo para Juzgar con Perspectiva de Género", consolidando el derecho a la igualdad. Este protocolo promueve que los jueces incorporen la perspectiva de género al tomar decisiones judiciales, buscando identificar y eliminar los sesgos que perpetúan la discriminación, especialmente en casos de violencia. El modelo establece directrices claras sobre cómo evaluar las situaciones considerando las desigualdades de género, con el fin de garantizar justicia para todas las personas, en particular para las mujeres en contextos de vulnerabilidad.

[10] https://www.pauta.cl/nacional/juzgar-con-perspectiva-de-genero-los-fallos-y-jueces-que-han-sido-clave

[11] https://cortesuprema.gov.co/corte/index.php/2023/01/12/dl_sl2936-2022/

[12] https://secretariagenero.poder-judicial.go.cr/images/Sentencias/SentenciasPerspectivaGenero/Modelo_de_incorporacin_de_la_Perspectiva_de_Gnero_en_las_sentencias.pdf

2.3 Procesamiento de Lenguaje Natural

El Procesamiento del Lenguaje Natural (NLP) es un campo interdisciplinario que integra la inteligencia artificial, la computación y la lingüística. Su principal objetivo es facilitar la interacción entre humanos y computadoras a través de los lenguajes naturales, que son los medios de comunicación empleados por las personas, tanto en forma oral como escrita. Este campo resulta crucial para desarrollar sistemas que permitan a las máquinas comprender y generar lenguaje humano, superando las barreras de comunicación entre ambas entidades.

Etapas del Procesamiento del Lenguaje Natural

El procesamiento de textos generalmente sigue una serie de etapas que permiten transformar el lenguaje natural en una representación estructurada que pueda ser analizada por las computadoras, como se puede observar en Figure 1. Es preciso considerar que tanto la secuencia de pasos a seguir como la implementación de cada paso pueden variar dependiendo de la tarea específica a desarrollar.

Figure 1: Etapas PNL

2.3.1 Preprocesamiento de datos

El **preprocesamiento** es un paso fundamental en NLP, ya que consiste en transformar el texto original en un formato que sea más adecuado para su análisis. Este proceso incluye diversas operaciones, tales como la eliminación de caracteres especiales, la conversión a minúsculas y la eliminación de palabras vacías.

- **Eliminación de Caracteres Especiales**. Se refiere a la eliminación de signos de puntuación, símbolos no alfabéticos o cualquier otro carácter que no aporte valor al análisis del texto. Por ejemplo, en la frase "*¿Cuáles son los canales de información que avalan el relato de la víctima?*", la eliminación de caracteres especiales resultaría en:

Original	***Transformación***
¿Cuáles son los canales de información que avalan el relato de la víctima?	Cuáles son los canales de información que avalan el relato de la víctima

- **Conversión de Caso**. Este proceso implica la estandarización del texto mediante la conversión de todas las palabras a minúsculas. Este paso facilita la comparación de palabras sin tener en cuenta su capitalización, como en el caso de la frase anterior, que se convertiría en:

Original	***Transformación***
¿Cuáles son los canales de información que avalan el relato de la víctima?	cuáles son los canales de información que avalan el relato de la víctima

- **Eliminación de Palabras Vacías (*StopWords*)**: Las palabras vacías son aquellas que, aunque comunes en el idioma, no aportan significado relevante para el análisis. Se suelen eliminar para reducir el ruido en los datos. Por ejemplo, la frase siguiente podría reducirse a las palabras más significativas, eliminando términos como "son", "los", "de", entre otros.

Original	***Transformación***
¿Cuáles son los canales de información que avalan el relato de la víctima?	canales información avalan relato víctima

2.3.2 Análisis Léxico

Una vez preprocesado el texto, se procede al **análisis léxico**, que incluye tareas como la tokenización, el stemming y la lematización.

- **Tokenización**. Se divide el texto en unidades mínimas llamadas "tokens", que pueden ser palabras, números u otras unidades significativas. En español, la tokenización es relativamente sencilla debido a la separación natural de las palabras por espacios. No obstante, en idiomas como el japonés, el proceso resulta más complejo debido a la ausencia de espacios entre las palabras.

Original	*Transformación*
¿Cuáles son los canales de información que avalan el relato de la víctima?	‘¿’,’Cuáles', 'son', 'los', 'canales', 'de', 'información', 'que', 'avalan', 'el', 'relato', 'de', 'la', 'víctima’,’?'

- **Stemming y Lematización**. Tanto el *stemming* como la *lematización* tienen el objetivo de reducir las palabras a su forma base, aunque difieren en su enfoque. El *stemming* se encarga de eliminar sufijos y prefijos con el fin de obtener la raíz de la palabra, mientras que la *lematización* se basa en reglas gramaticales para asegurar que la palabra resultante sea una forma válida dentro del idioma.

Original	*Transformación*
¿Cuáles son los canales de información que avalan el relato de la víctima?	'¿', 'cuál', 'ser', 'el', 'canal', 'de', 'información', 'que', 'avalar', 'el', 'relato', 'de', 'el', 'víctima', '?'

2.3.3 Análisis Sintáctico

El análisis sintáctico se centra en la estructura gramatical de las oraciones. Una técnica clave en este proceso es el etiquetado gramatical, también conocido como **POS tagging**, que asigna categorías gramaticales a cada palabra del texto, como

sustantivo, verbo, adjetivo, etc, como puede observarse en la Tabla 2. Este etiquetado facilita el análisis de las relaciones entre las palabras dentro de la estructura de la oración.

Palabra	POS	Dependencia
¿	PUNCT	punct
Cuáles	PRON	nsubj
serían	VERB	ROOT
los	DET	det
canales	NOUN	obj
de	ADP	case
información	NOUN	nmod
que	PRON	nsubj
avalan	VERB	acl
el	DET	det
relato	NOUN	obj
de	ADP	case
la	DET	det
víctima	NOUN	nmod
?	PUNCT	punct

Tabla 2: Ejemplo POS

2.3.4 Representación de texto

El último paso en el procesamiento de texto es la representación numérica. Los modelos tradicionales de representación de texto se centran en el léxico, en las palabras que componen el texto y su frecuencia, considerando, en algunos casos, las palabras cercanas (Melucci & Baeza-Yates, 2011). Sin embargo, al enfocarse exclusivamente en aspectos léxicos, estos modelos pierden información crucial relacionada con el orden y la secuencia de las palabras en el texto. Además, estos enfoques están sujetos al fenómeno conocido como el "*curse of dimensionality*" (maldición de la dimensionalidad), el cual se refiere a los problemas que surgen al analizar y procesar datos en espacios de alta dimensión. A medida que la dimensionalidad de los datos aumenta, la cantidad de muestras necesarias para representar adecuadamente la distribución crece exponencialmente, lo que puede provocar escasez de datos, aumento del ruido y dificultades en el aprendizaje automático.

Aunque los modelos léxicos tradicionales son útiles en diversas tareas, carecen de la capacidad necesaria para abordar cuestiones semánticas, un aspecto fundamental cuando se analizan sentencias con sesgo implícito, como se ha mencionado previamente. Para mitigar los problemas de las representaciones basadas en léxico, han surgido modelos de representación más avanzados basados en *deep learning* emplean técnicas como *word embeddings*, que representan las palabras en un espacio vectorial de alta dimensionalidad, capturando sus significados y relaciones contextuales. Dado que existe una fina línea entre el uso de estereotipos positivos, el refuerzo de estereotipos negativos y la consideración de las realidades para lograr la igualdad (Timmer, 2016), los *embeddings* tienen la ventaja de capturar el contexto necesario para determinar si los patrones presentes sugieren sesgos. Algunas representaciones de este tipo son:

Word2Vec. Es un modelo de aprendizaje profundo (Mikolov, G., K., & J., 2013), que utiliza una red neuronal para aprender representaciones vectoriales de palabras a partir de grandes corpus de texto. Considera las palabras como una unidad ignorando su morfología. Una de sus limitaciones es cuando los idiomas tienen vocabularios muy extensos con muchas palabras poco frecuentes.

FastText. Es una extensión del modelo Word2Vec (Bojanowski, Grave, Joulin, & T., 2017) que también permite aprende representaciones vectoriales de sub-palabras, lo que facilita el manejo de palabras desconocidas y mejora la precisión en tareas de clasificación de texto. Este modelo es compatible con aproximadamente 150 idiomas entrenados con Wikipedia y ofrece modelos para la identificación de idiomas. En lugar de considerar las palabras de manera aislada, se toman en cuenta en conjunto con la representación n-gram de sus caracteres.

BERT. Es un modelo de lenguaje de última generación (Devlin, Chang, Lee, & Toutanova, 2019) que ha tenido un impacto significativo en el campo del procesamiento de lenguaje natural. Utiliza la arquitectura *Transformer*, y se distingue por su capacidad para comprender el contexto de las palabras y frases de manera bidireccional. La innovación principal de este modelo radica en que se entrena mediante dos tareas de procesamiento de lenguaje natural: la predicción de palabras faltantes (llenado de espacios en blanco) y la predicción de la siguiente oración en un texto. Esto permite que el modelo aprenda no solo las relaciones entre las palabras

dentro de una oración, sino también el contexto más amplio en el que estas son utilizadas

BETO. (BERT Español) (Cañete J, 2020) es un modelo de lenguaje BERT diseñado para el idioma español. Fue desarrollado por el equipo ciencias de la computación universidad de chile[13]. Existen dos versiones “BETO-cased” conserva las mayúsculas y minúsculas originales y “BETO-uncased” convierte todo el texto a minúsculas.

paraphrase-multilingual-mpnet-base-v2[14]. Es un modelo preentrenado para la generación de *embeddings* de oraciones, el mismo es multilingüe y soporta más de 50 idiomas. Basado en la arquitectura MPNet[15] (Masked and Permuted Network), este modelo es una mejora de los modelos BERT y RoBERTa, lo que permite capturar releaciones semánticas de manera más eficiente y precisa.

RoBERTalex[16]. Es un modelo de transformer basado en la arquitectura RoBERTa, especializado en el procesamiento de textos jurídicos en el lenguaje español. Fue desarrollado por PlanTL-GOB-ES[17] y está disponible en HugginFace. Este modelo ha sido entrenado con grandes corpus de textos legales en español, optimizando su capacidad para comprender y generar lenguaje natural en el contexto de derecho.

2.3.5 Algoritmos de clasificación

En esta sección se describen los algoritmos de aprendizaje supervisado seleccionados para este estudio.

Gradient Boosting (GB). Es una técnica de aprendizaje automático que combina múltiples modelos (generalmente árboles de decisión) para crear un modelo predictivo más robusto. Comienza con un modelo simple, generalmente un árbol de decisión, que suele presentar alto sesgo y baja varianza. En cada iteración, se añade un nuevo árbol entrenado para corregir los errores del conjunto actual, ajustándose a

[13] https://www.dcc.uchile.cl/
[14] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[15] https://huggingface.co/docs/transformers/model_doc/mpnet
[16] https://huggingface.co/PlanTL-GOB-ES/RoBERTalex
[17] https://huggingface.co/PlanTL-GOB-ES

los residuales. Este enfoque es robusto y permite estrategias de ajuste de pesos para datos desbalanceados. (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, Catboost: unbiased boosting with categorical features, 2018).

Regresión Logística (LR). Este método estadístico es comúnmente utilizado para problemas de clasificación binaria. La regresión logística extiende el concepto de regresión lineal al incorporar una función logística, también conocida como sigmoide, que transforma la salida de la regresión lineal en un valor comprendido en el intervalo [0, 1]. Este valor se interpreta como la probabilidad de pertenencia a una categoría específica. El proceso de ajuste del modelo implica la estimación de los coeficientes óptimos de las variables predictoras, de manera que se maximice la verosimilitud de los datos bajo la distribución binomial o de Bernoulli. Además de su uso en problemas binarios, la regresión logística puede generalizarse a problemas de clasificación multiclase mediante técnicas como la regresión logística multinomial u ordinal.

Support Vector Machines (SVM). Introducidos por (Vapnik & Cortes, 1995), los SVM son algoritmos de aprendizaje supervisado que han demostrado ser altamente efectivos en tareas de clasificación y regresión. La principal característica de SVM es la identificación de un hiperplano que maximiza el margen de separación entre las clases. Este hiperplano es determinado por las muestras más cercanas a la frontera de decisión, denominadas vectores de soporte. SVM no solo se utiliza para problemas lineales, sino que, a través de la aplicación del truco del kernel (*kernel trick*), puede abordar problemas no lineales al mapear los datos a un espacio de mayor dimensión donde puedan ser separables linealmente. A pesar de su potencia, los SVM pueden requerir un ajuste fino de sus parámetros y presentar altos costos computacionales en conjuntos de datos grandes.

Random Forest (RF). Son una técnica de aprendizaje automático basada en el ensamblaje de múltiples árboles de decisión, propuesta por (Breiman, 2001). Esta metodología genera un conjunto de árboles de decisión a partir de muestras aleatorias del conjunto de datos, aplicando un subconjunto aleatorio de características en cada nodo de división. Esta aleatorización contribuye a la reducción del sobreajuste y mejora la capacidad de generalización del modelo. Además, los RF proporcionan

estimaciones de la importancia relativa de las características, lo que facilita la selección y análisis de estas. A diferencia de un único árbol de decisión, que puede ser sensible a variaciones menores en los datos, los RF son más robustos frente a estos problemas y pueden manejar eficazmente conjuntos de datos con alta dimensionalidad y ruido.

Naïve Bayes. Este algoritmo se basa en el teorema de Bayes, que utiliza probabilidades condicionadas para hacer predicciones. Aunque se denomina "naïve" (ingenuo) debido a su suposición de independencia entre las características, ha demostrado ser un enfoque eficiente en tareas de clasificación, especialmente en la clasificación de texto. El modelo calcula la probabilidad de que un documento pertenezca a una clase dada, basándose en la probabilidad de que las palabras presentes en el documento correspondan a dicha clase. A pesar de sus suposiciones simplificadoras, este algoritmo es altamente efectivo y ampliamente utilizado en el campo del procesamiento de lenguaje natural (NLP).

CatBoost. Es un algoritmo de aprendizaje automático basado en el método de boosting de árboles de decisión, diseñado especialmente para manejar eficientemente variables categóricas y evitar el sobreajuste. Presentado en (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, Catboost: unbiased boosting with categorical features, 2018), el mismo puede manejar de manera eficiente valores faltantes. Una de las innovaciones más destacadas es su manejo nativo de variables categóricas. En lugar de depender de codificaciones manuales, el algoritmo genera combinaciones de características categóricas y numéricas de manera automática, utilizando una técnica de ordenamiento que preserva la dependencia causal y reduce el riesgo de overfitting.

2.3.6 Modelos de Lenguaje Grandes (LLMs)

Los Modelos de Lenguaje de Gran Escala (LLMs, del inglés *Large Language Models*) han marcado un avance disruptivo en el campo de la inteligencia artificial, particularmente en las áreas de procesamiento del lenguaje natural (NLP) y la generación de texto. Estos modelos, como GPT-3 (Brown, Mann, & et, 2020) y sus sucesores, son entrenados utilizando grandes volúmenes de datos textuales no

estructurados, aprovechando arquitecturas basadas en *Transformers* (Vaswani, Shazeer, Parmar, & Uszkoreit, 2017), lo que les otorga la capacidad de capturar y aprender representaciones profundas y complejas de los patrones lingüísticos. A través de este entrenamiento masivo, los LLMs pueden modelar con precisión las relaciones semánticas entre palabras, frases y contextos en un espectro diverso de tareas lingüísticas.

Uno de los avances clave que han permitido los LLMs es su capacidad para realizar tareas de NLP mediante una interacción contextual más flexible y natural. A diferencia de los enfoques tradicionales, donde los modelos requerían entrenamiento especializado para tareas específicas, los LLMs permiten el aprendizaje en contexto con solo formular instrucciones en lenguaje natural, lo que constituye un cambio de paradigma. Este enfoque, conocido como "*in-context learning*", se refiere a la capacidad del modelo para adaptarse a nuevas tareas durante la inferencia, basándose únicamente en ejemplos o instrucciones provistas dentro de la entrada. Esto facilita una amplia gama de aplicaciones sin necesidad de realizar un entrenamiento supervisado adicional (T. Brown et al., 2020).

La técnica de "prompting", que se refiere al diseño de entradas textuales que orientan el comportamiento del modelo hacia tareas específicas, ha facilitado la adopción de estos modelos en diversos dominios. A través de "*prompts*" o indicaciones bien estructuradas, los LLMs pueden ser guiados para generar texto coherente y relevante en una variedad de contextos, (Chang, Xu, & Luo, 2024). Este enfoque también ha abierto las puertas a la democratización del uso de modelos de NLP avanzados, permitiendo que usuarios no expertos puedan interactuar con sistemas sofisticados de manera intuitiva, sin necesidad de conocimientos técnicos profundos (Fagbohun, Harrison, & Dereventsov, 2024).

En este contexto, los LLMs pueden utilizarse tanto para la generación de representaciones de texto como para tareas de clasificación. Para la representación, los LLMs pueden transformar el texto en vectores densos mediante el uso de *embeddings* contextuales, obtenidos a través de las capas internas del modelo, como en el caso de las representaciones de *embeddings* mencionadas anteriormente. Por otro lado, para la clasificación, los LLMs pueden ser utilizados para clasificar textos

mediante *prompting*, considerando diferentes modalidades, como *zero-shot* (se le proporciona al LLM una instrucción clara para clasificar un texto sin ejemplos previos, confiando en su conocimiento entrenado) o *few-shot* (se incluyen en el *prompt* algunos ejemplos de texto con sus respectivas clasificaciones para guiar al LLM antes de pedirle que clasifique un nuevo texto).

2.4 Métricas de Evaluación de Modelos

En esta sección se explican las métricas, las cuales se utilizan para evaluar la calidad y el rendimiento de los modelos. Estas métricas proporcionan una forma cuantitativa de medir qué tan bien un modelo está realizando una tarea específica. Las métricas son esenciales por varias razones:

- **Comparación de Modelos**. Las métricas permiten comparar diferentes modelos para la misma tarea, lo que ayuda a seleccionar el modelo más eficiente.
- **Evaluación del Rendimiento**. Ofrecen una medida objetiva del rendimiento de un modelo, permitiendo observar mejoras a lo largo del tiempo o en función de distintos ajustes.
- **Identificación de Áreas de Mejora**. Las métricas permiten identificar áreas (por ejemplo, conjuntos de instancias con determinadas características) en las que el modelo requiere ajustes o mejoras para optimizar su desempeño.

Para asegurar una evaluación confiable del modelo, es fundamental contar con una estrategia adecuada para la división de los datos. Separar el conjunto de datos en entrenamiento y prueba permite estimar el rendimiento del modelo en datos no vistos, evitando mediciones sesgadas.

La validación cruzada es una técnica crucial en el aprendizaje automático utilizada para evaluar y ajustar modelos de manera robusta. Una de las formas más comunes de sobreajustar un modelo es entrenarlo y evaluarlo en el mismo conjunto de datos. Para abordar este problema, la validación cruzada ofrece soluciones

eficaces. Las estrategias más utilizadas incluyen KFold, StratifiedKFold, GroupKFold y StratifiedGroupKFold.

En su forma más simple, esta técnica divide el conjunto de datos en varias particiones o "folds" para entrenar y evaluar el modelo en diferentes subconjuntos de los datos. El proceso de **k-fold** cross-validation divide el conjunto de datos en **k** partes, entrenando el modelo con **k-1** de ellas y evaluándolo con la parte restante. Este proceso se repite K veces, alternando el fold de evaluación. Sin embargo, cuando el conjunto de datos está desbalanceado, existe el riesgo de que algunos folds contengan una representación no equilibrada de las clases, lo que puede conducir a una evaluación poco representativa del modelo. En este contexto, **StratifiedKFold**[18] en lugar de dividir los datos en *folds* (particiones) de manera aleatoria (como hace *KFold*), asegura que cada *fold* contenga una proporción similar de muestras para cada clase que en el conjunto total de datos. Así, si una clase tiene menos ejemplos, se asegura de que estén representados en todos los *folds*. Esta técnica resulta especialmente útil cuando se trabaja con conjuntos de datos con una distribución desbalanceada de las clases. Esto es crucial para evaluar los modelos de manera justa, particularmente cuando se trata de problemas de clasificación multiclase o binaria.

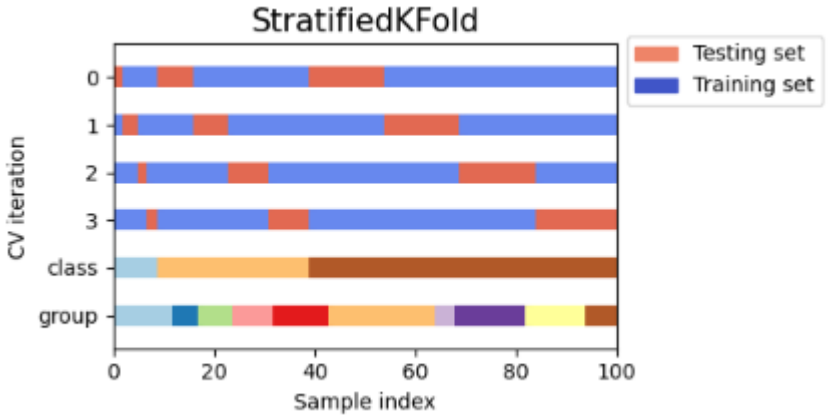

Figure 2: StrafiedkFold

Precisión

La precisión es una métrica que evalúa la proporción de predicciones correctas sobre el total de predicciones realizadas. Se calcula como la razón entre los

[18] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

verdaderos positivos (TP) y la suma de los verdaderos positivos y falsos positivos (FP), es decir, la proporción de instancias correctamente clasificadas como positivas respecto al total de instancias predichas como positivas. La fórmula para calcularla es la siguiente.

$$Precision = \frac{TP}{TP + FP}$$

Ecuación 1:Precisión

Por ejemplo, si un modelo de clasificación binaria predice 100 instancias como positivas, y de esas 100, 80 son realmente positivas (verdaderos positivos) y 20 son negativas (falsos positivos), entonces la Precisión del modelo sería de 0.8 o 80%.

La precisión es especialmente útil cuando el costo de los falsos positivos es alto, ya que se enfoca en minimizar este tipo de errores. Sin embargo, una de sus limitaciones es que no considera los falsos negativos, lo que puede ser problemático en contextos donde es crucial detectar todas las instancias positivas[19]

Sensibilidad (Recall)

La sensibilidad, también conocida como *recall* o tasa de verdaderos positivos, mide la capacidad de un modelo para identificar correctamente las instancias positivas. Se define como la proporción de verdaderos positivos (TP) sobre la suma de los verdaderos positivos y los falsos negativos (FN).

$$Recall = \frac{TP}{TP + FN}$$

Ecuación 2: Formula Recall

Por ejemplo, si un modelo de clasificación binaria identifica 100 instancias como positivas, y de esas 100, 80 son realmente positivas (verdaderos positivos) y 20 son negativas (falsos negativos), entonces el *recall* del modelo sería de 0.8 o 80%. El *recall* es crucial cuando el costo de los falsos negativos es alto, ya que maximiza la cantidad de instancias positivas correctamente identificadas. No obstante, no toma en cuenta los falsos positivos, lo que podría ser una desventaja en algunas

[19] https://www.themachinelearners.com/metricas-de-clasificacion/.

aplicaciones, especialmente cuando es importante reducir los errores de predicción en ambas direcciones.

MCC

El coeficiente de correlación de Matthews (MCC) [20] , inventado por Brian Matthews en 1975, se utiliza para entender qué tan bien un modelo de aprendizaje automático puede predecir resultados. Dicho coeficiente (Chicco & Jurman, 2020), suele ser más confiable que AUC-ROC en conjuntos de datos desbalanceados. Este coeficiente tiene en cuenta verdaderos negativos (las predicciones correctas de que algo no sucedería), verdaderos positivos (las predicciones correctas de que algo sucedería), falsos negativos (las predicciones incorrectas de que algo no sucedería) y falsos positivos (las predicciones incorrectas de que algo sucedería). Se calcula con la siguiente fórmula:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Ecuación 3: Formula MCC

El **MCC** tiene valores que oscilan entre -1 y 1, donde:

- 1 indica una predicción perfecta.
- 0 indica que el modelo no tiene mejor rendimiento que una predicción aleatoria.
- -1 sugiere un rendimiento completamente erróneo, es decir, una clasificación opuesta a la realidad.

Este coeficiente es especialmente útil en conjuntos de datos desbalanceados, ya que tiene en cuenta tanto los aciertos como los errores en todas las clases, proporcionando una visión más global del rendimiento del modelo.

[20] https://statologos.com/coeficiente-de-correlacion-de-matthews/

Capítulo 3

Trabajos relacionados

Este apartado ofrece una visión general de estudios previos que han abordado la problemática del *sesgo de género*, especialmente en la detección de estereotipos a través de diversas metodologías y enfoques.

Diversos estudios han reportado avances significativos en el desarrollo de técnicas para identificar y mitigar el sesgo de género, con un enfoque predominante en el idioma *inglés* que, a diferencia del español, no tiene género gramatical explícito. Estos estudios han explorado la representación del género en textos, utilizando desde enfoques estadísticos tradicionales hasta el uso de modelos de lenguaje preentrenados, para detectar sesgos implícitos y explícitos en corpus textuales. Sin embargo, a pesar de estos avances, pocos estudios han investigado este fenómeno en el idioma español y mucho menos en el contexto de decisiones judiciales, donde el impacto del lenguaje sesgado puede tener implicaciones legales y sociales de gran alcance.

En el ámbito judicial, estudios recientes han comenzado a explorar cómo el sesgo puede manifestarse en sentencias y fallos legales. De todas formas, los estudios que abordan específicamente el sesgo de género en fallos judiciales en español siguen siendo escasos.

Este trabajo se propone llenar ese vacío, proporcionando un análisis más profundo del sesgo de género en textos legales en español, y contribuyendo a la creación de herramientas que permitan una detección más efectiva y contextualizada de dichos sesgos, teniendo en cuenta las particularidades del idioma.

3.1 Sesgos en textos legales judiciales

En esta sección nos centramos en los sesgos específicos que se manifiestan en el lenguaje utilizado en los textos legales y judiciales.

En el estudio realizado por (Elliott, Chen, & Ornaghi, 2020) se plantea una pregunta de investigación central: *¿cómo influye el sesgo de género en el comportamiento de los jueces?* Este trabajo se centra en dos aspectos clave: el impacto del sesgo en las decisiones judiciales y en el tratamiento de las juezas.

En primer lugar, se explora cómo este sesgo afecta las decisiones judiciales. El estudio revela que los jueces con un mayor sesgo de género tienden a emitir fallos más conservadores, particularmente en cuestiones relacionadas con la ampliación de los derechos de las mujeres. De hecho, se estima que una desviación estándar en la inclinación de género del juez incrementa la probabilidad de votar en contra de los derechos de las mujeres en un 4,5%. En segundo lugar, el estudio examina cómo el sesgo de género afecta las interacciones de los jueces con sus colegas mujeres. Los resultados indican que este sesgo también influye en el trato hacia las juezas, evidenciado en un mayor número de revocaciones de decisiones cuando el juez es una mujer, así como en una menor citación de sus opiniones. Estas observaciones destacan la importancia del género no solo en las decisiones judiciales, sino también en la dinámica profesional entre jueces y juezas.

Metodológicamente, este trabajo emplea *word embeddings* (GloVe) para analizar el texto, utilizando un enfoque novedoso en la identificación de sesgos de género a partir de decisiones judiciales. Los datos provienen de diversas fuentes, entre ellas: recolección aleatoria de juicios, *Appeals Court Attribute Data*, y el *Federal Judicial Center*. En conclusión, los resultados reflejan que los jueces con mayores sesgos de género no solo votan de forma más conservadora en casos de derechos de la mujer, sino que también son más propensos a revocar decisiones cuando las juezas son las responsables de emitir dichas decisiones.

En el estudio de (Sevim, Sahinuç, & Koç, 2022) se analiza la presencia de prejuicios sexistas en los modelos de *word embeddings*, con especial énfasis en su uso dentro del ámbito jurídico. Para ello, los autores emplean el modelo *Law2Vec*, entrenado específicamente con corpus jurídicos, y comparan diversas técnicas de aprendizaje automático como los RF, SVM, y redes de memoria a largo plazo (LSTM) para la predicción de decisiones de casos judiciales. El estudio también aborda el desarrollo de un sistema de obtención de información para leyes mediante técnicas

de NLP. A través de un análisis exhaustivo de corpus jurídicos de distintos países, detectaron un sesgo de género generalizado en los textos legales, independientemente de los métodos de medición o del país de origen.

Los resultados de los métodos basados en *word embeddings* mostraron que los algoritmos de eliminación de sesgo (*Hard Debiasing* and *Soft Debiasing)* tuvieron éxito al reducir los niveles de sesgo a cifras insignificantes en comparación con los niveles de sesgo medidos antes de la eliminación del sesgo. Sin embargo, los métodos indirectos basados en las palabras vecinas de un término objetivo arrojaron resultados contradictorios, sugiriendo que los algoritmos de reducción de sesgo no tuvieron un impacto significativo. Estos hallazgos destacan la necesidad de continuar investigando en este ámbito, con el fin de desarrollar algoritmos más robustos y efectivos para la eliminación de sesgos en el NLP aplicado al derecho.

3.2 Modelos de lenguaje para el dominio legal

A continuación, nos enfocamos en los avances y aplicaciones de modelos de lenguaje en el ámbito legal.

aymurai/flair-ner-spanish-judicial[21]. Este modelo fue desarrollado por collective.ai[22] como parte del proyecto AymurAI[23] de DataGenero[24]. Tiene como objetivo abordar la falta de datos disponibles en el sistema judicial sobre fallos relacionados con la violencia de género en América Latina. Su objetivo es generar confianza en el sistema de justicia y mejorar el acceso a la justicia para mujeres. Aún es un prototipo y solo se está implementando en el Tribunal Penal N°10 en la Ciudad de Buenos Aires, Argentina, sus capacidades están limitadas a la recopilación y

[21] https://huggingface.co/aymurai/flair-ner-spanish-judicial
[22] https://www.collectiveai.io/
[23] https://www.aymurai.info/
[24] https://datagenero.org/

análisis de datos semiautomatizados, y los resultados pueden estar sujetos a limitaciones como la calidad, consistencia, así como la disponibilidad de estos. El mismo se entrenó con un conjunto de datos cerrado de un tribunal penal argentino, lo que garantiza que esté adaptado al contexto legal y cultural específico.

3.3 Detección de sesgos en otros dominios

Como se mencionó previamente, el análisis de sesgos no se limita exclusivamente al ámbito judicial. Con el fin de ampliar el contexto y ofrecer una visión más integral, a continuación, se presentan estudios que han abordado este desafío en otros dominios

En (Chen, Xu, Lui, & Guo, 2020) se propone una metodología general para cuantificar y mitigar sesgos en textos, abordando tanto estereotipos explícitos como implícitos sin depender de atributos sensibles predefinidos. Utilizan un enfoque de tres módulos: extracción de características, medición de discrepancias y mitigación de sesgos. Para la extracción, aplican *bag-of-ngrams* para detectar sesgos explícitos y representaciones basadas en *autoencoders* para capturar estereotipos implícitos en los datos. Luego, cuantifican la diferencia entre el conjunto de datos sesgado y uno de referencia utilizando la Discrepancia Media Máxima (MMD), compara la distribución del conjunto de datos (sesgado) con la de un conjunto de datos de referencia (neutral), si MMD es cercano a 0 significa distribuciones similares. El estudio se basa en dos conjuntos de datos: comentarios de *Twitter*, que reflejan sesgos ideológicos explícitos, y titulares de noticias, donde los sesgos son más implícitos. Los resultados muestran que los sesgos explícitos se detectan mejor con características léxicas, mientras que los sesgos implícitos requieren técnicas más avanzadas.

En (Dacon & Liu, 2021) se presenta un estudio sobre el sesgo de género en los artículos de noticias. El estudio se basa un conjunto de datos de más de 296.000

artículos de noticias en inglés (de las colecciones MIND[25] y NCD[26]), en las cuales se analizan las representaciones de hombres y mujeres. El análisis mostró que las mujeres se encontraban sub-representadas en las noticias, y que cuando se las menciona, a menudo se las representa en roles estereotipados. Los autores encontraron que las mujeres representan solo el 29% de los autores y el 33% de las fuentes citadas en los artículos de noticias. Asimismo, las mujeres tienen más probabilidades de ser mencionadas en relación con su apariencia física o sus relaciones personales, mientras que los hombres tienen más probabilidades de ser mencionados en relación con sus logros profesionales o su experiencia. Asimismo, se determinó que el sesgo de género en las noticias es más pronunciado en ciertos temas, como los negocios y la política. Los autores encontraron que las mujeres representan solo el 29% de los autores y el 33% de las fuentes citadas en los artículos de noticias.

En (Plaza del Arco, Molina Gonzalez, Ureña Lopez, & Valdivia, 2020) se abordó la identificación de sexismo en redes sociales mediante un sistema de aprendizaje multitarea (MTL) basado en BERT. El enfoque MTL permite entrenar varias tareas relacionadas en paralelo, como la clasificación de sentimientos, emociones y detección de lenguaje ofensivo, lo que mejora la capacidad del sistema para detectar sexismo. Esta investigación participó en la competencia EXIST 2021, donde se desarrollaron dos subtareas: en la primera, se buscaba clasificar si un texto era sexista, mientras que la segunda se enfocaba en categorizar el tipo de sexismo presente en el mensaje. Los autores utilizaron diversos corpus, incluyendo el conjunto de datos EXIST en inglés y español, además de otros como InterTASS, EmoEvent, HatEval y MEX-A3T, los cuales contienen datos de *Twitter* relevantes para tareas relacionadas con la polaridad, emociones y lenguaje ofensivo. Los resultados obtenidos en la primera subtarea lograron una precisión del 78%, ocupando el segundo lugar, y en la segunda subtarea alcanzaron un F1 de 56.67%, ubicándose en tercer lugar. Los autores concluyeron que el uso de tareas relacionadas como la clasificación de polaridad y la detección de lenguaje ofensivo contribuyó a mejorar la identificación de sexismo. Sin embargo, señalaron que la combinación de la clasificación de emociones con la detección de sexismo no fue tan efectiva en la

[25] https://microsoftnews.msn.com/
[26] https://www.huffpost.com/

segunda subtarea, sugiriendo la inclusión de tareas adicionales como la detección de ironía o sarcasmo para futuras mejoras.

En (Nishtha, Sameep, Taneea, & Yatin,, 2018) se enfocaron en la identificación y eliminación de estereotipos de género en las películas de Bollywood. Los autores analizaron tanto los guiones como los pósteres de películas para estudiar cómo se representan los géneros masculino y femenino, utilizando técnicas de NLP y análisis de imágenes. Para llevar a cabo este análisis, procesaron datos de más de 4,000 películas desde 1970 hasta 2017, extraídas de Wikipedia, así como 880 tráileres de películas y guiones de 13 películas. El análisis detectó estereotipos de género en descripciones y acciones asignadas a los personajes, en general, los hombres tendían a ocupar roles más centrales y ser descritos con verbos activos, mientras que las mujeres se representaban en función de su apariencia o en relación con personajes masculinos. Además, los autores desarrollaron un algoritmo llamado DeCogTeller que es capaz de modificar los guiones para eliminar sesgos de género, mediante el ajuste semántico de las historias, creando versiones más neutrales. Una limitación importante del estudio es que algunos cambios propuestos por el sistema, como intercambiar roles de género en ciertas profesiones o contextos, no siempre son plausibles. Por ello, los autores sugieren mejorar el sistema para manejar casos más complejos y realistas, como la jerarquía ocupacional.

En (Ferrer, van Nuenen, Such, & Criado, 2021) se desarrolló un enfoque basado en *embeddings* para identificar y categorizar sesgos lingüísticos en la plataforma Reddit[27]. Los autores analizaron si ciertos términos se asociaban de manera con sesgos lingüísticos propios de cada comunidad dentro de diferentes subreddits[28]. Para ello, trabajaron con varios corpus extraídos de distintos subreddits, y entrenaron un modelo de word embeddings utilizando Word2Vec para cada uno de ellos. Este enfoque permitió capturar sesgos relacionados con atributos como género, etnicidad y religión. Una vez entrenados los modelos, se emplearon palabras objetivo, relacionadas con los sesgos específicos que se deseaba analizar. El análisis permitió descubrir palabras sesgadas y agruparlas en categorías utilizando algoritmos de clustering como k-means. Además, se llevó a cabo un análisis de polaridad o

[27] https://www.reddit.com/
[28] https://support.reddithelp.com/hc/es-es/articles/204533569--Qu%C3%A9-son-las-comunidades-o-subreddits

sentimiento para determinar si los sesgos identificados presentaban connotaciones negativas o positivas.

En (Cryan, y otros, 2020) se reexaminaron la detección de estereotipos de género en el contexto de herramientas modernas, evaluando la eficacia de enfoques basados en léxicos y en aprendizaje automático. Su trabajo aportó tres contribuciones principales: (i) el desarrollo de un léxico actualizado de estereotipos de género que reflejó el lenguaje y las interpretaciones contemporáneas, (ii) la recopilación de un corpus de 4.333 artículos etiquetados por humanos y (iii) el entrenamiento de un modelo de clasificación basado en BERT. Los resultados mostraron que, a pesar de los esfuerzos por mejorar los modelos léxicos, los enfoques basados en BERT ofrecieron un desempeño significativamente superior, incluso con un conjunto de datos de tamaño moderado.

Esta reseña de estudios destaca tanto los avances como las limitaciones en la detección de sesgos de género en diferentes dominios y metodologías. Como se mencionó previamente se observa que el idioma *inglés* ha sido el más estudiado, dado que es comúnmente utilizado en la investigación sobre procesamiento de lenguaje natural (NLP) y no presenta la complejidad gramatical que tiene el español, especialmente en lo que respecta a la marcación explícita de género. Sin embargo, los avances en otros idiomas, como el *español*, aún son limitados, lo que representa una brecha importante en la investigación y aplicación de modelos para la detección de sesgos en lenguajes con estructuras gramaticales más complejas. En particular, el entrenamiento de modelos en corpus específicos, como los textos judiciales en español, puede ofrecer resultados más precisos y ajustados a la realidad local, permitiendo la detección de sesgos de género que son específicos de esa área del lenguaje y el contexto legal. A continuación, la Tabla 3 muestra un resumen de dichos trabajos.

Cita	Tarea	Idioma	Representación	Herramienta	Corpus	Resultados

(Elliott, Chen, & Ornaghi, 2020)	Detección de estereotipos de género en apelaciones judiciales	Inglés	Word embeddings (GloVe)	Metodología propia de cálculo de sesgo	Apelaciones de juicios	Los jueces tienden a votar de manera más conservadora en casos de derechos de las mujeres. Además, son más propensos a revocar decisiones de jueces de distrito cuando estos son mujeres y citan menos opiniones escritas por juezas.
(Cryan, y otros, 2020)	Detección de estereotipos de género en artículos online	Inglés	BERT, Word embeddings (Word2Vec, Glove, FastText)	Análisis de texto	Artículos online	Deep Learning muestra mejores resultados en la detección de estereotipos de género en comparación con enfoques basados en léxicos.
(Ferrer, van Nuenen, Such, & Criado, 2021)	Detección de prejuicios lingüísticos	Inglés	Word embeddings (Word2Vec)	Análisis crítico del discurso, IAT (Pruebas de Asociación Implícita)	Reddit	El estudio concluye que los sesgos lingüísticos existen y pueden rastrearse mediante técnicas de IA.
(Nishtha, Sameep, Taneea, & Yatin,, 2018)	Estudio de estereotipos basados en el género en roles de hombres y mujeres	Inglés	Word embeddings (Word2Vec)	Dense CAP, Word Graph Technique	Páginas Wiki, películas	Se observa una alta precisión en la predicción de género con conjuntos de datos de entrenamiento pequeños.
(Plaza del Arco, Molina Gonzalez, Ureña Lopez, & Valdivia, 2020)	Detección de discurso de odio hacia inmigrantes y mujeres	Español e ingles	Léxico, BERT Word embeddings, (WordPiece embeddings)	Machine Learning (LR, SVM, BN, DT, EV), Deep Learning (RNN)	InterTASS, EXIST, EmoEvent, HatEval	El método de "voting" obtuvo los mejores resultados en la detección de discurso de odio.

(Stanczak & Augestein, 2021)	Análisis de publicaciones sobre prejuicios de género	Inglés	Word embeddings (Word2Vec, FastText,) BERT	WEAT, SEAT, Bias Amplification	EEC Dataset, Multilingual Template-Based Dataset	El estudio abarca hasta junio de 2021 y observa una disminución en el sesgo de género en los últimos años.
(Dacon & Liu, 2021)	Análisis del sesgo de género implícito y explícito en distintos dominios	Inglés	Lexicones	Análisis de Resonancia Centralizada (CRA)	MIND, NCD	El sesgo de género aumenta en las secciones de política y negocios, mientras que en temas relacionados con la familia se observa un predominio femenino.
(Chen, Xu, Lui, & Guo, 2020)	Cuantificación y mitigación de sesgos en datos de lenguaje natural	Inglés	Bag-of-n-grams, Representaciones profundas (ARAE)	Discrepancia Media Máxima (MMD), Deep Generative Models	Comentarios de Twitter, Titulares de noticias	Los sesgos explícitos se detectan mejor con bag-of-ngrams, mientras que las representaciones profundas capturan sesgos implícitos. La mitigación de sesgos es más efectiva en los sesgos léxicos.
(Sevim, Sahinuç, & Koç, 2022)	Evaluación de sesgos de género en word embeddings utilizados en NLP	Inglés	Law2Vec, Law2VecNew	KNN, ECC	HUDOC, EEC	Se concluye que los word embeddings en NLP perpetúan estereotipos de género y discriminación involuntaria.

Tabla 3: Resumen Trabajos Relacionados

Capítulo 4

Enfoque metodológico

La identificación de sesgos de género en textos legales presenta un reto considerable, debido a que dichos sesgos no siempre se manifiestan de manera explícita. Frecuentemente, dichos sesgos emergen de forma sutil a través de construcciones lingüísticas que perpetúan estereotipos dentro de contextos específicos, lo cual puede inducir interpretaciones o conclusiones sesgadas. Por ejemplo, una sentencia que afirme: "*La víctima, al ser mujer, probablemente exageró la gravedad del incidente debido a su naturaleza emocional y su tendencia a dramatizar situaciones*" incluye un sesgo evidente, identificable por un sistema de análisis automático mediante el reconocimiento de palabras clave. No obstante, existen casos más insidiosos, como en la frase: "*Se presume que la madre tendrá una mayor capacidad para cuidar y proporcionar un entorno estable a los hijos; por lo tanto, se le concede la custodia principal en el caso de divorcio*". En este último ejemplo, el análisis resulta más complejo, ya que se parte de la suposición implícita de que la mujer, en virtud de su rol de madre, es inherentemente más adecuada para el cuidado de los hijos, sin una evaluación objetiva de las competencias parentales individuales.

Para abordar estos desafíos, el desarrollo de un modelo basado en técnicas de Procesamiento del Lenguaje Natural (NLP) requiere un análisis contextual sofisticado que permita detectar patrones de sesgo de forma precisa. Es importante destacar que los fragmentos que contienen sesgos tienden a representar una proporción relativamente pequeña del texto completo de una sentencia. En este trabajo, se ha implementado un enfoque metodológico que integra técnicas de NLP, algoritmos de aprendizaje automático supervisado (clasificación) y modelos de lenguaje de gran escala (LLMs), con el propósito de desarrollar un modelo de clasificación binaria capaz de identificar sesgos de género en resoluciones judiciales. La Figure 3 ilustra los componentes clave de este pipeline metodológico.

El diseño metodológico responde a las siguientes preguntas de investigación:

- **PI-1**. ¿Cuál es la representación de texto más eficaz para una clasificación tradicional?

- **PI-2**. ¿Es la clasificación basada en LLMs comparable a los enfoques de clasificación tradicionales? ¿Existen diferencias significativas en los resultados producidos por distintos LLMs?

- **PI-3**. ¿Pueden los LLMs generar explicaciones coherentes y comprensibles para los humanos en relación con la detección de sesgo?

Para implementar esta metodología, el proyecto se desarrolla en Python (versión 3) utilizando Google Colab como entorno de trabajo. Este servicio basado en Jupyter Notebook, proporciona una infraestructura computacional robusta, eliminando la necesidad de configuraciones locales y facilitando el acceso a recursos computacionales avanzados

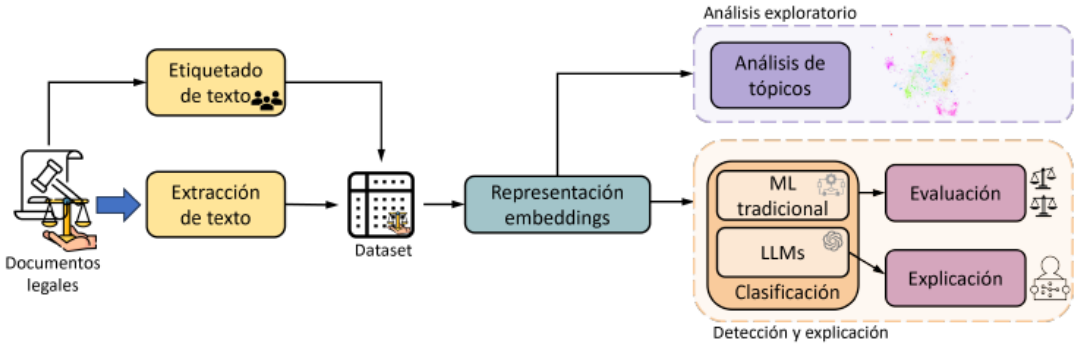

Figure 3: Esquema del pipeline de procesamiento propuesto

4.1 Obtención de texto

El punto de partida de este estudio es un conjunto de documentos legales, específicamente sentencias judiciales, en los cuales se han etiquetado fragmentos de texto que contienen indicios de sesgo de género. Dichas sentencias suelen encontrarse en formatos como PDF o documentos de texto (por ejemplo, Office), esto implica la necesidad de automatizar tanto la extracción como la limpieza del texto. Esta última incluye la eliminación de caracteres no deseados y la supresión de elementos recurrentes como encabezados y pies de página. En lugar de procesar cada sentencia como una unidad indivisible, se ha optado por tomar el párrafo como

unidad de análisis, lo que permite individualizar de manera más precisa las expresiones que podrían evidenciar la presencia de sesgos. Posteriormente, a cada párrafo se le asigna una etiqueta binaria que indica la presencia o ausencia de sesgo.

El análisis se llevó a cabo sobre sentencias judiciales de Argentina, recopiladas por el "*Observatorio de Sentencias Judiciales*" de la "*Articulación Regional Feminista*". Este observatorio recoge decisiones judiciales de varios países de la región (Argentina, Bolivia, Chile, Colombia, Ecuador, México y Perú), con el propósito de evaluar el grado de cumplimiento con la *Convención para la Eliminación de todas las Formas de Discriminación contra la Mujer*. Se seleccionaron las sentencias en las que se detectó la presencia de sesgos, las cuales estaban acompañada por un resumen que describía los párrafos afectados por dichos sesgos. A partir de estas descripciones, se revisaron y anotaron las sentencias seleccionadas, con el objetivo de construir la solución de referencia (*ground truth*). En total, se analizaron 12 sentencias judiciales, que contenían un total de 1.094 párrafos, de los cuales 41 (3,74%) presentaban indicadores de sesgo de género. Esta distribución revela un desbalance significativo en las clases, con una proporción de 1:26 entre los párrafos positivos (con sesgo) y negativos. Aunque podría considerarse el uso de técnicas de aumento de datos (*data augmentation*) para mitigar este desbalance, se decidió dejar su implementación para trabajos futuros.

Dado que las sentencias se encuentran documentadas en formato PDF, fue necesario desarrollar un script para la extracción automática del texto. Para generar el corpus en formato JSONL, se utilizó la librería de Python "PyMuPDF"[29], que facilita la extracción, análisis y manipulación de documentos PDF. Esta herramienta permitió separar el contenido textual de otros elementos como imágenes, líneas o "ruido" no textual.

Una vez completada esta etapa, se obtuvo como producto final un corpus estructurado en cuatro columnas: la columna "página" indica el número de página de la sentencia; "texto" contiene el contenido del párrafo; "highlight" señala los fragmentos donde los expertos han identificado el sesgo de género; y finalmente,

29 https://pypi.org/project/PyMuPDF/

"doc" hace referencia al nombre del documento de la sentencia. La Tabla 4 presenta un ejemplo de la estructura del corpus:

página	texto	highlight	doc
38	Señaló también la defensa contradicciones de la menor en la declaración en Cámara Gesell, que no especificó, y cuya transcendencia o incidencia tampoco advierte el suscripto. Por lo demás, no se entiende por qué razón resulta llamativo al esmerado letrado que hayan concurrido al mismo hotel, ni qué sospecha sobre la veracidad de la niña puede surgir de esta circunstancia. Que no haya contado a nadie el suceso o que ocultase su embarazo, tampoco demuestra que la relación haya sido consentida. Puede afirmarse que hay más razones para ocultar un abuso violento, que una relación que no lo es. En los abusos suele primar la vergüenza, las amenazas, o el temor a la opinión de terceros, en cambio en las relaciones consentidas, aunque sean ilícitas o moralmente prohibidas, se da más la jactancia, o la confesión amistosa o cómplice con amigas o hermanas.	En los abusos suele primar la vergüenza, las amenazas, o el temor a la opinión de terceros, en cambio en las relaciones consentidas, aunque sean ilícitas o moralmente prohibidas, se da más la jactancia, o la confesión amistosa o cómplice con amigas o hermanas.	_Causa N° 4566 contraViĕctor Alejandro Soliĕs Chambi.pdf

Tabla 4: Resumen corpus de sentencias

4.2 Preprocesamiento

Una vez obtenidos los datos, es fundamental someterlos a un proceso de preprocesamiento. Este paso es crucial para normalizar las instancias, con el fin de convertirlas en un formato adecuado para ser utilizado por los diferentes algoritmos de aprendizaje automático. En general, las sentencias judiciales sin procesar representan un conjunto de datos ruidosos debido a la variabilidad en los formatos utilizados en cada caso. Estos documentos contienen elementos característicos como

números de sentencia, referencias a juzgados y direcciones, que deben ser extraídos de manera adecuada para su procesamiento. La normalización de estas características resulta imprescindible para su correcta utilización en los diversos modelos de clasificación. Para estandarizar el conjunto de datos y reducir su tamaño, se sigue una serie de pasos específicos de preprocesamiento.

El preprocesamiento de las sentencias judiciales incluyó las siguientes tareas:

- Conversión de todas las palabras a minúsculas.
- Eliminación de signos de puntuación.
- Remoción de caracteres especiales.
- Eliminación de palabras con menos de tres letras.
- Eliminación de palabras vacías (*stopwords*) en español.
- Tokenización del texto.
- Lematización, es decir, reducción de las palabras a su forma base o raíz gramatical.
- Aplicación de *stemming*, técnica utilizada para remover y reemplazar los sufijos de las palabras.

Todas estas tareas se explicaron con detalle en el Capítulo 2. En la Tabla 5, se puede observar el resultado de este procesamiento.

Original	A fs. 558/560 vta. los Dres. H.Y. y A.V. por la representación unificada de las codemandadas, expresan sus agravios ratificando y convalidando en primer lugar las cuestiones planteadas en el recurso antes relatado. Entienden, además, que el presente amparo refiere a la protección de un derecho individual, en este caso sobre la actora Mirta Graciela del Valle Sisnero, cuyo progreso podría crear, en su criterio, un importante precedente de alcance colectivo.
Preprocesado	representación unificada codemandadas expresan agravios ratificando, convalidando primer cuestiones planteadas recurso relatado entienden además presente amparo refiere protección individual actora mirta graciela valle sisnero cuyo progreso podría crear criterio portante precedente alcance colectivo

Original	A fs. 565/569 vta. la Sra. Defensora Oficial Civil Nº 4 con testa los agravios y solicita el rechazo de los recursos conforme los argumentos que allí explicita.
Preprocesado	defensora oficial civil testa agravios solicita rechazo recursos conforme argumentos allí explicita

Original	También el art. 13 de la Constitución de la Provincia recepta la garantía en el orden local cuando declara que "Todas las personas son iguales ante la ley, sin distinción por razón de nacimiento, raza, sexo, religión, opinión o cualquier otra condición o circunstancia personal o social... Garantizase la igualdad del hombre y la mujer y el ejercicio pleno de sus derechos económicos, sociales, culturales y políticos".
Preprocesado	constitución provincia recepta garantía orden local declara todas personas iguales distinción razón nacimiento raza sexo religión opinión cualquier condición circunstancia personal social garantízase igualdad hombre mujer ejercicio pleno derechos económicos sociales culturales políticos

Original	A fs. 552/556 vta. Ahynarca S.A. expresa agravios y en primer término destaca la diferencia entre ser condenado por discriminación y resultar alcanzado por los efectos jurídicos de una acción colectiva. Respecto al primer punto, niega haber incurrido en una conducta discriminatoria en relación a la Sra. Sisnero, quien sólo se habría limitado a entregar a la empresa una solicitud de trabajo. Además, señala que la certificación contable agregada a la causa prueba que la empresa no incorporó a choferes desde mucho antes del pedido de la actora y hasta la actualidad.
Preprocesado	ahynarca expresa agravios primer término destaca diferencia condenado discriminación resultar alcanzado efectos jurídicos acción colectiva primer punto niega incurrida conducta discriminatoria sisnero sólo limitado entregar empresa solicitud trabajo además señala certificación contable agregada causa empresa incorporó choferes pedido actora actualidad

Tabla 5: Ejemplos de sentencias y su versión preprocesada

4.3 Extracción de Características

La extracción de características es un paso clave para convertir los datos textuales en representaciones numéricas que los algoritmos de aprendizaje automático puedan procesar. Esta etapa resulta fundamental para el éxito de cualquier sistema de NLP, ya que traduce el texto en vectores numéricos, facilitando su comprensión y análisis por parte de los algoritmos.

Para este propósito, se utilizan, *CountVectorizer* y *TF-IDF Vectorizer*, como primera medida para convertir el texto en vectores numéricos.

4.3.1 CountVectorizer

Se utiliza la clase, `sklearn.feature_extraction.text.CountVectorizer`[30] de la biblioteca sklearn. A continuación, en la Tabla 6, se detallan los principales parámetros utilizados en el pipeline de procesamiento:

Nombre Parámetro	Descripción Parámetro	Valor del Parámetro
max_df	Este parámetro controla la frecuencia máxima permitida para los términos. Si un término aparece en más del max_df porcentaje de los documentos, será ignorado. Esto es útil para eliminar palabras muy comunes (como artículos o preposiciones) que no aportan información relevante al modelo.	1.0
min_df	Controla la frecuencia mínima para los términos. Si un término aparece en menos de min_df documentos, será ignorado. Esto es útil para eliminar palabras demasiado raras que no generalizan bien.	1
ngram_range	Especifica el rango de n-gramas que se extraerán. Un n-grama es una secuencia de n palabras consecutivas. Este parámetro acepta una tupla que indica el valor mínimo y máximo de n en los n-gramas.	(1,1)
stop_words	Especifica las palabras que deben ser eliminadas del texto antes de la tokenización. Puede ser una lista personalizada de palabras o un valor predefinido como 'english', que elimina las palabras comunes del inglés.	None
max_features	Establece un límite en el número máximo de características (palabras únicas) que se deben tener en cuenta. El vectorizador seleccionará las palabras más frecuentes hasta alcanzar este número.	None
binary	Si se establece en True, en lugar de contar el número de apariciones de una palabra en un documento, simplemente se marcará con un 1 si la palabra aparece al menos una vez, o 0 si no aparece. Este parámetro es útil para generar vectores binarios	False
lowercase	Determina si se deben convertir todas las palabras a minúsculas antes de la tokenización. Por defecto está en True	True
analyzer	Controla si el análisis se realiza a nivel de palabras o caracteres. Los valores posibles son 'word' para tokenización por palabra o 'char' para tokenización a nivel de caracteres.	word
tokenizer	Permite definir una función personalizada para realizar la tokenización. Esto es útil si se quiere un procesamiento especializado más allá de la tokenización por defecto de CountVectorizer	None
vocabulary	Especifica un diccionario de vocabulario fijo en lugar de que el vectorizador lo construya automáticamente a partir del corpus de documentos	None

Tabla 6: Parámetros CountVectorizer

[30] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

4.3.2 TFIDFVectorizer

Para la técnica de TF-IDF (*Term Frequency-Inverse Document Frequency*), se empleó la clase `sklearn.feature_extraction.text.TfidfVectorizer`[31]. Los parámetros más relevantes utilizados se presentan a continuación en la Tabla 7.

Nombre Parámetro	Descripción Parámetro	Valor del Parámetro

[31] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

max_df	Controla la frecuencia máxima permitida para los términos. Si un término aparece en más del max_df porcentaje de los documentos, será ignorado. Esto ayuda a eliminar términos que son demasiado comunes.	1.0
min_df	Controla la frecuencia mínima para los términos. Si un término aparece en menos de min_df documentos, será ignorado. Esto es útil para eliminar palabras demasiado raras que pueden no aportar valor.	1
ngram_range	Especifica el rango de n-gramas a extraer. Un n-grama es una secuencia de n palabras consecutivas. El parámetro recibe una tupla que define el rango mínimo y máximo.	(1, 1)
stop_words	Este parámetro especifica una lista de palabras que deben ser eliminadas del texto antes de la tokenización. Puede usar una lista personalizada o un conjunto predefinido para eliminar palabras comunes	None
max_features	Este parámetro limita el número de características (palabras únicas) a considerar. Selecciona las max_features palabras más relevantes basadas en su puntuación TF-IDF.	None
norm	Define la norma que se utilizará para normalizar los vectores de salida. Las opciones más comunes son 'l2' (la norma euclidiana) y 'l1' (la norma de Manhattan)	l2
use_idf	Determina si se utiliza el término inverso de frecuencia de documentos (IDF) al calcular los pesos de los términos. Si se establece en False, solo se utilizará la frecuencia de términos (TF), ignorando la IDF.	True
smooth_idf	Si está establecido en True, se agregará un suavizado (añadiendo 1 al denominador) al calcular el IDF para evitar divisiones por cero cuando un término aparece en todos los documentos	True
sublinear_tf	Si se establece en True, aplica una escala sublineal a la frecuencia de términos, reemplazando los conteos de frecuencia por 1 + log(tf). Esto reduce el impacto de palabras que ocurren con mucha frecuencia en un documento	False
vocabulary	Este parámetro permite especificar un vocabulario fijo, en lugar de dejar que el TfidfVectorizer lo construya automáticamente a partir de los documentos. Se debe pasar un diccionario donde las palabras son las claves y los valores son sus índices	None

Tabla 7: Parámetros TfidfVectorizer

4.3.3 Word2vec

Para la representación vectorial de las palabras, se entrena uno de los modelos utilizando Word2Vec a través del módulo `models.word2vec` de la biblioteca Gensim[32]. Esta técnica de *embeddings* de palabras permite capturar relaciones semánticas

[32] https://en.wikipedia.org/wiki/Gensim

entre las palabras utilizando los algoritmos Skip-Gram y CBOW. La Tabla 8 muestra los valores configurados.

Nombre Parámetro	Descripción Parámetro	Valor del Parámetro

vector_size	Este parámetro especifica la dimensionalidad de los vectores de palabras generados por el modelo. Un tamaño más grande puede capturar relaciones más complejas, pero también aumenta el costo computacional.	100
window	Define el tamaño del contexto en el que se consideran las palabras vecinas para aprender las representaciones de las palabras objetivo. Por ejemplo, si window=5, el modelo considerará 5 palabras a la izquierda y 5 a la derecha de la palabra objetivo	5
min_count	Controla el umbral mínimo de frecuencia. Las palabras que aparecen menos veces que el valor de min_count en el corpus serán ignoradas durante el entrenamiento. Este parámetro ayuda a reducir el ruido y las palabras raras	1
sg	Determina el algoritmo de entrenamiento. Si sg=0, se utiliza el modelo de Bag of Words (CBOW). Si sg=1, se emplea el Skip-Gram. El CBOW es más rápido, mientras que el Skip-Gram es más adecuado para representar palabras raras.	0
workers	Especifica el número de hilos de procesamiento (threads) que se utilizarán para entrenar el modelo. Esto permite aprovechar el paralelismo en máquinas con múltiples núcleos para acelerar el entrenamiento.	4
epochs	Número de veces que el modelo pasa por el conjunto completo de datos durante el entrenamiento. Más épocas pueden permitir al modelo aprender mejor, pero también aumentan el tiempo de entrenamiento.	5
negative	Este parámetro controla el número de muestras negativas que se usarán para el entrenamiento con negative sampling. Si negative=0, se desactiva el negative sampling. Un número mayor de muestras negativas puede mejorar la calidad del modelo, pero a costa de aumentar el tiempo de entrenamiento.	5
hs	Define si se debe utilizar la estrategia de hierarchical softmax para entrenar el modelo. Si hs=1, se utiliza hierarchical softmax; si hs=0, se emplea el negative sampling (controlado por el parámetro negative)	0
alpha	Es la tasa de aprendizaje inicial. La tasa de aprendizaje disminuye linealmente durante el entrenamiento desde alpha hasta min_alpha	0.025
max_vocab_size	Si se proporciona, este parámetro establece un límite máximo al tamaño del vocabulario. Se utiliza para restringir el tamaño del vocabulario cuando se trabaja con conjuntos de datos grandes, eliminando las palabras más raras hasta alcanzar el límite establecido.	None

Tabla 8: Parámetros Word2Vec

4.3.4 FastText

Para la técnica FastText, se utiliza el módulo `models.fasttext` de Gensim. Este modelo extiende Word2Vec añadiendo la capacidad de representar palabras fuera del vocabulario (OOV) y se entrena de manera similar a Word2Vec. En la Tabla 9 se pueden observar los parámetros que se utilizaron para las pruebas.

Nombre Parámetro	Descripción Parámetro	Valor del Parámetro
vector_size	La dimensionalidad de los vectores de palabra, es decir, el número de características que representa cada palabra	100
alpha	La tasa de aprendizaje inicial. Se reduce linealmente hasta min_alpha	0.025
min_alpha	La tasa de aprendizaje mínima	0.0001
window	Tamaño de la ventana de contexto, es decir, el número de palabras vecinas a tener en cuenta para entrenar	5
min_count	Ignora palabras que aparezcan menos veces que este valor.	1
sample	Umbral para downsampling de palabras más frecuentes. Valores más pequeños suprimen las palabras más frecuentes.	0.001
sg	Si usar el algoritmo Skip-gram (sg=1) o el algoritmo CBOW (sg=0).	0
hs	Si usar Softmax jerárquico (hs=1) o muestreo negativo (hs=0).	0
negative	Número de ejemplos negativos a tomar para el entrenamiento de Skip-gram.	5
epochs	Número de iteraciones sobre el corpus de entrenamiento.	5
word_ngrams	Si agregar o no n-gramas de palabras al vocabulario, en lugar de solo palabras individuales.	1
workers	Número de hilos o procesos paralelos usados para el entrenamiento	4
compute_loss	Si se debe calcular y almacenar la pérdida del modelo durante el entrenamiento.	False

Tabla 9: Parámetros FastText

Estos modelos de aprendizaje de representaciones distribuidas de palabras que, a pesar de capturar similitudes semánticas entre palabras, se consideran enfoques simples a nivel palabra, ya que no toman en cuenta el contexto específico en el que las palabras aparecen. En estos modelos, cada palabra es asignada a un único vector, independientemente de su sentido o uso en diferentes contextos. *Word2Vec* genera estos vectores a partir de patrones de co-ocurrencia en grandes corpus de texto, pero sigue siendo limitado en su capacidad para desambiguar

significados polisémicos. *FastText*, aunque mejora sobre *Word2Vec* al utilizar sub-palabras y capturar información morfológica, sigue tratándose de una representación estática de las palabras.

La Tabla 10 resume los diferentes tipos de representaciones de texto evaluadas para los párrafos de las sentencias analizadas. Tamaño representa la dimensión del vector con el que se representa cada párrafo de una sentencia.

	Tipo	Tamaño
TF-IDF	Bag-of-words	1000
multilingual-mpnet-base-v2	Embedding multilingual a nivel sentencia	768
Beto	Embedding contextual a nivel palabra.	768
RobertaLex	Embedding contextual a nivel palabra.	768
Word2vec	Vector denso a nivel palabra	100
Fasttext	Vector denso a nivel palabra	100
CountVectorizer	Bag-of-words	1000

Tabla 10: Representación de texto utilizadas

En la Tabla 11, se presenta un resumen detallado del corpus de datos utilizado en el presente trabajo.

Total, de sentencias	12
Total, de párrafos	1.094
Párrafos promedio por sentencia	91,17
Párrafos con sesgo de género	41 (3,74%)
Párrafos sin sesgo de género	1.053 (96,26%)
Promedio párrafos por sentencia con sesgo	4
Promedio párrafos por sentencia sin sesgo	88
Proporción entre párrafos con y sin sesgo	1:26

Tabla 11: Resumen Corpus

Como se mencionó anteriormente el desbalanceo es evidente para la clase sin sesgo.

4.4 Análisis de datos

En esta etapa, se lleva a cabo un análisis detallado de los datos con el objetivo de comprender las características clave del conjunto de datos, lo que facilita decisiones informadas sobre cómo preprocesar los datos y qué tipo de modelo será más adecuado para el análisis posterior. A continuación, se presentan los análisis realizados en diversos aspectos de los datos.

.

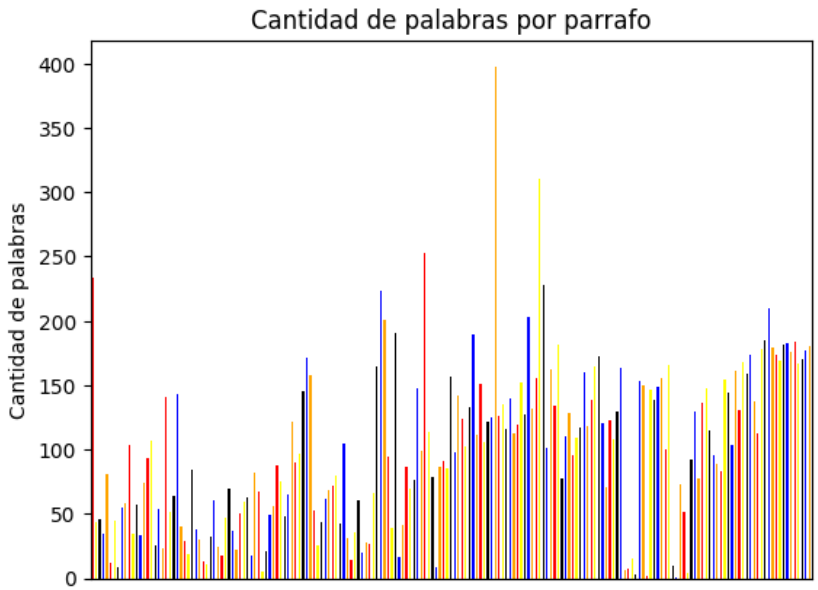

Figure 4: Cantidad de palabras por párrafo

Distribución de Palabras por Párrafo

Una vez procesados los datos, la Figure 4 presenta la distribución de palabras por párrafo. Con una cantidad de palabras que varía entre 1 y 186, un promedio de 64 palabras por párrafo y un desvío de 49.

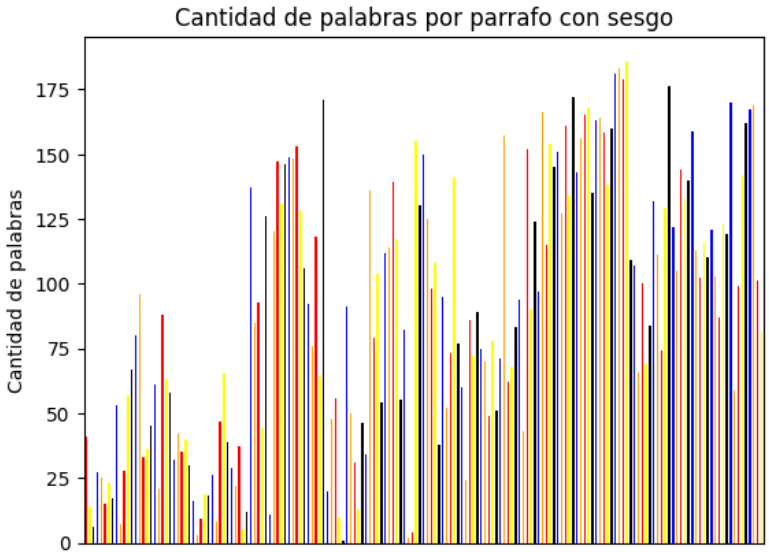

Figure 5: Cantidad de palabras por párrafo con sesgo

Distribución de Palabras por Párrafo con sesgo

Se observa en la Figure 5. Figure 5el valor promedio es 43.85, lo que indica el centro aritmético de los datos. El valor más bajo en el conjunto de datos es 6, mientras que el valor más alto observado en los datos es 136 y el desvío estándar es de 26.80

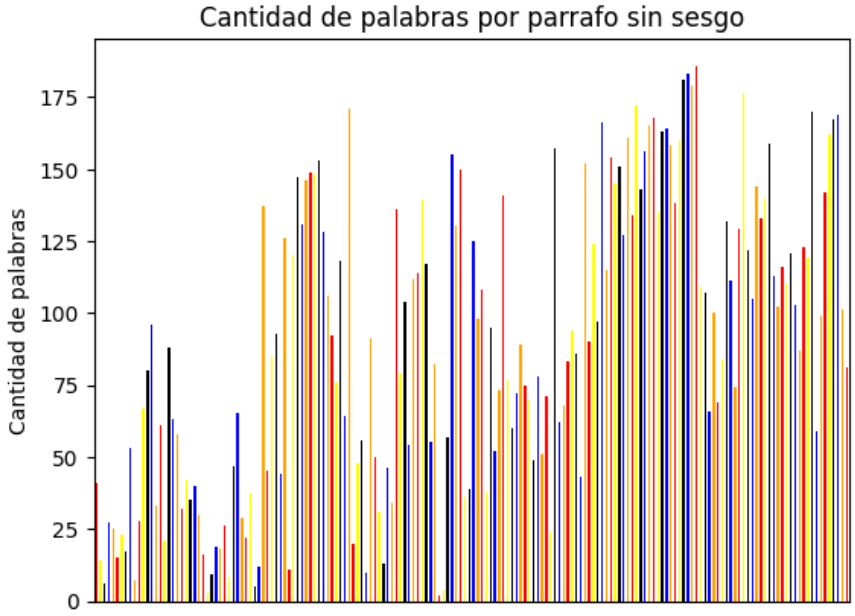

Figure 6: Cantidad de palabras por párrafo procesado sin sesgo

Distribución de Palabras por Párrafo sin sesgo

Los datos que se muestran en Figure 6 con valores que oscilan entre 1 y 186. La media de 63.88 y el desvío estándar 49.36.

Análisis de Distribución Basada en "POS" (Partes de la Oración)

A continuación, se analiza la distribución de palabras basadas en su categoría gramatical, comenzando con los sustantivos, verbos, adjetivos, y sustantivos de género. Para detectar si una palabra se refiere a un género tanto masculino o femenino se utilizó la librería de spaCy[33] utilizando el modelo preentrenado *es_core_news_lg*[34], esto permitió realizar un análisis de los sustantivos, pronombres y adjetivos que suelen llevar marcas de género en español, la librería etiqueta las palabras con morfología, lo que te permite identificar la marca de género en cada palabra.

Palabras Femeninas Más Frecuentes

La Figure 7 presenta las 20 palabras femeninas más frecuentes en el corpus de sentencias. Se observa que, como era de esperar, el término “situación” ocupa la primera posición con 129 la primera posición, algunos términos que se desprenden de este análisis son pareja con 106 y herida con 103 apariciones.

[33] https://spacy.io/
[34] https://spacy.io/models

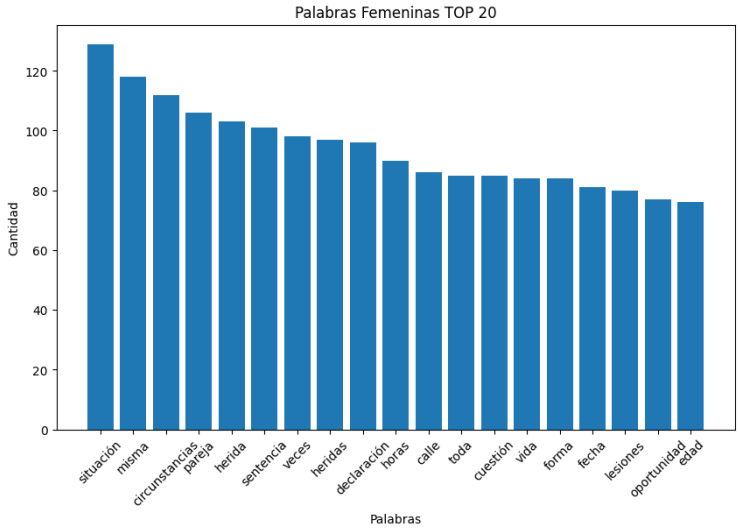

Figure 7: Palabras femeninas TOP 20

Palabras Femeninas Más Frecuentes con sesgo

El análisis de la Figure 8 estos datos revelan que "existencia" es el término más frecuente, con 10 menciones. Le siguen "edad", "opinión", "relaciones" y "declaración", cada uno con 7 apariciones. "Situación" aparece 6 veces, mientras que "conclusión", "razón" y "dudas" tienen 5 menciones. Otros términos importantes incluyen "contradicciones", "circunstancias", "amenazas" y "propia", con 4 menciones cada uno. Este conjunto refleja temas vinculados a la valoración de pruebas, las circunstancias y las conclusiones legales en los casos analizados

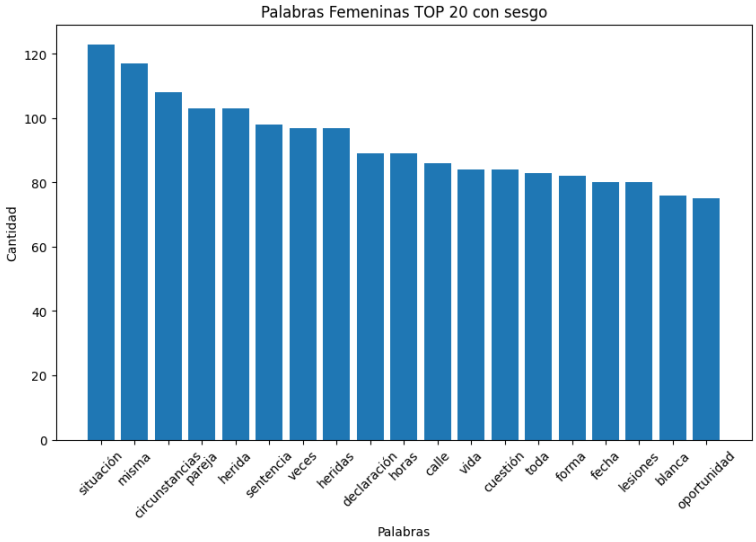

Figure 8: Palabras femeninas TOP 20 con sesgo

Palabras Femeninas Más Frecuentes sin sesgo

En Figure 9, el término "situación" es el más frecuente, con 123 menciones, seguido por "misma" con 117 y " circunstancias" con 102. Palabras como "pareja”,” herida“ con 103, "sentencia" 98 también destacan. Otras palabras como "fecha" 80 y "blanca" 76, mostrando la importancia del tiempo y las decisiones en los casos subrayando las consecuencias judiciales y personales en los hechos analizados.

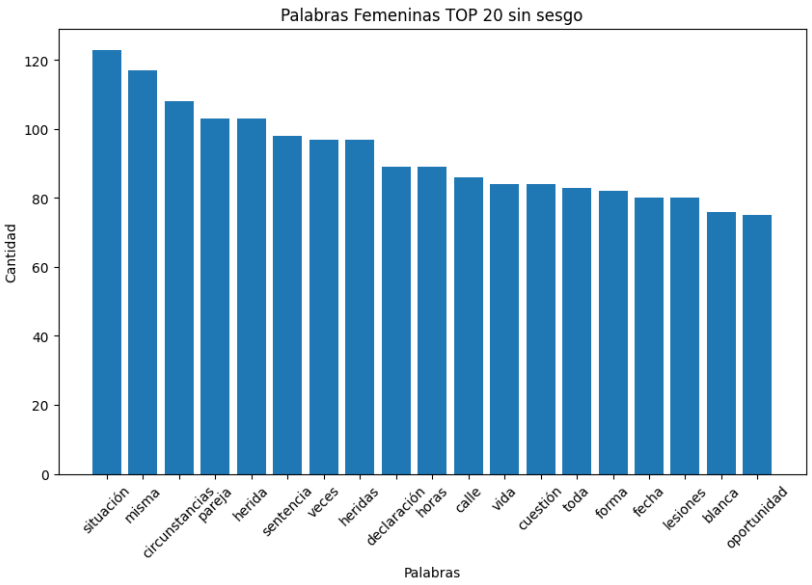

Figure 9:Palabras femeninas TOP 20 sin sesgo

Se

Palabras Masculinas Más Frecuentes

La Figure 10 muestra las 20 palabras masculinas más utilizadas en las sentencias. Entre las palabras más destacadas encontramos términos como “código”, “juicio” y “hijos”. La distribución de las palabras más frecuentes es bastante uniforme, sin palabras que sobresalgan significativamente de las demás.

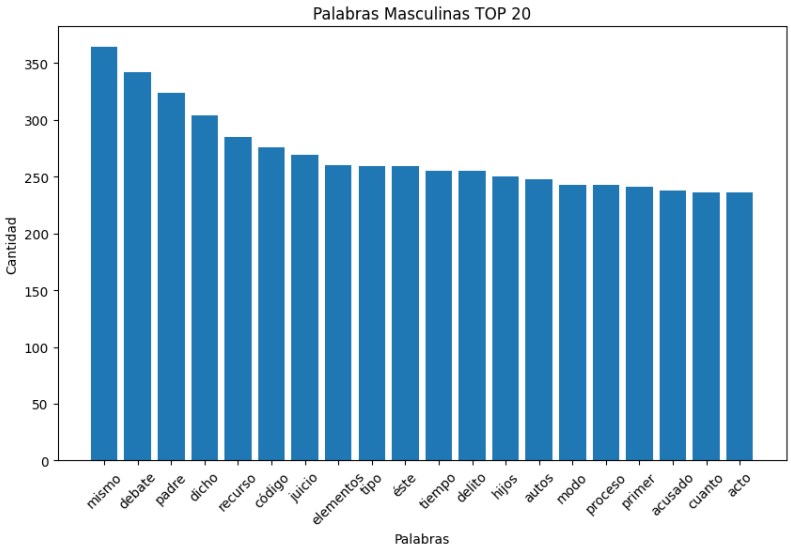

Figure 10: Palabras Masculinas TOP 20

Palabras Masculinas Más Frecuentes sin sesgo

El análisis de la Figure 11 los datos revelan que el término "hechos" es el más frecuente, con 121 menciones, seguido de "derecho" con 110 y "tipo" con 101. Otros términos importantes incluyen "dicho" 95, "visto" 94 y "tiempo" 92. Además, "éste" y "días" aparecen 90 veces cada uno, mientras que "hijo" y "domicilio" tienen 88 menciones. Palabras como "homicidio" 81, "delito" 8) y "elementos" 79 reflejan un enfoque en la descripción de los crímenes y los sujetos involucrados en los casos evaluados.

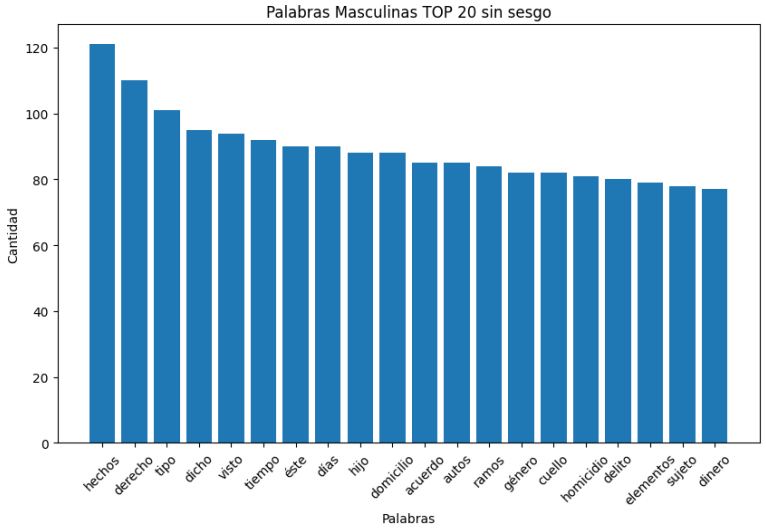

Figure 11: Palabras Masculinas TOP 20 sin sesgo

Palabras Masculinas Más Frecuentes con sesgo

En este conjunto de datos la Figure 12, el término "relato" es el más frecuente, con 10 menciones, seguido de "hechos" con 7 apariciones. Palabras como "primer" y "embarazo" figuran 6 veces cada una, mientras que "casos", "ocurrido", "abuso", "suceso" y "autor" tienen 5 menciones. Otros términos importantes incluyen "testimonios", "dichos", "acceso" y "acto", cada uno con 5 apariciones. Además, palabras como "párrafo", "consentimiento", "elementos" y "aborto" aparecen 4 veces. Este conjunto refleja un enfoque en la narración de los hechos, el abuso y el consentimiento en los casos analizados.

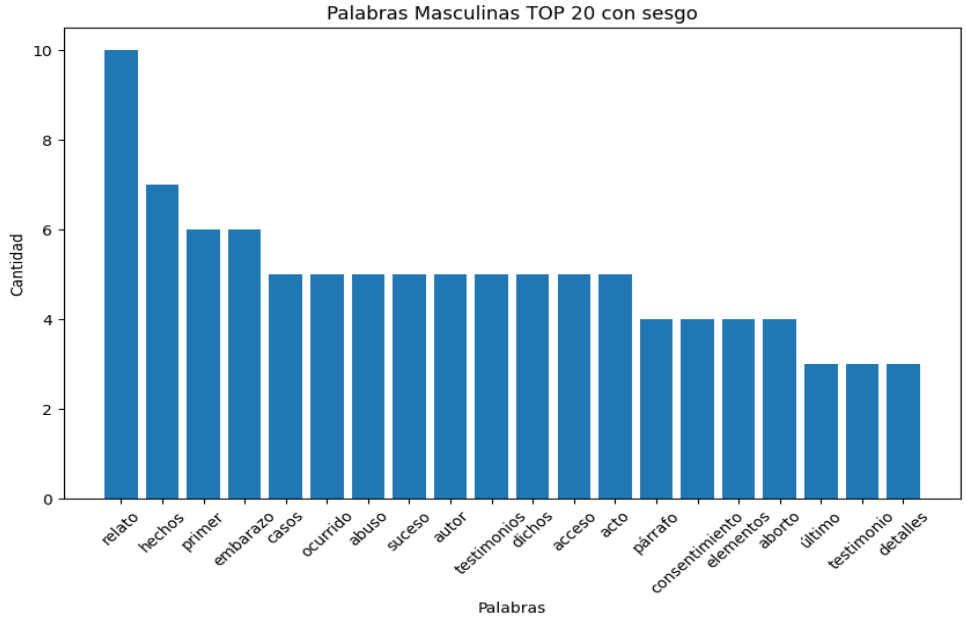

Figure 12: Palabras Masculinas TOP 20 con sesgo

Verbos Más Frecuentes

En la Figure 13Figure 13, se analizan los verbos más utilizados en las sentencias. Estos verbos aparecen en su forma tal como están en el corpus, sin transformación alguna. En el análisis, no se detecta una regla común que explique la predominancia de ciertos verbos, lo que sugiere que su uso es más aleatorio o relacionado con contextos específicos.

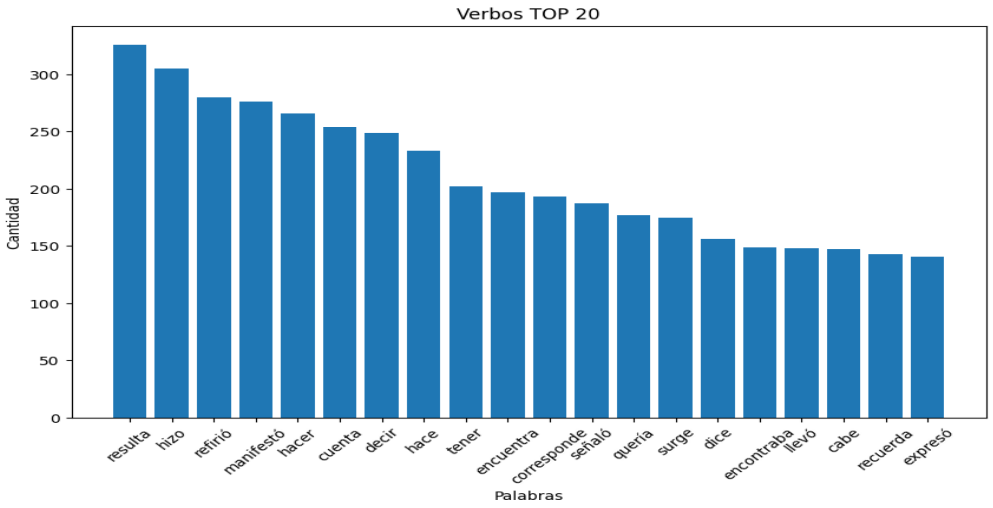

Figure 13:Top 20 Verbos

Verbos Más Frecuentes con sesgo

En este análisis la Figure 14 se muestra, el término "soler" es el más frecuente, con 12 menciones, seguido por "resultar", "existir" y "referir", cada uno con 8 apariciones. Otros términos significativos incluyen "contar" 7, "tratar" 7 y "sostener" 6. Además, palabras como "considerar" 5, "advertir" 5, "hacer" 5, "aludir" 5, "reconocer" 5, "describir" 5 y "surgir" 5 destacan por su relevancia en el contexto de acciones, percepciones y descripciones. Finalmente, verbos como "encontrar" 3 y "expresar" 3 sugieren la importancia de la búsqueda y la comunicación en los casos analizados.

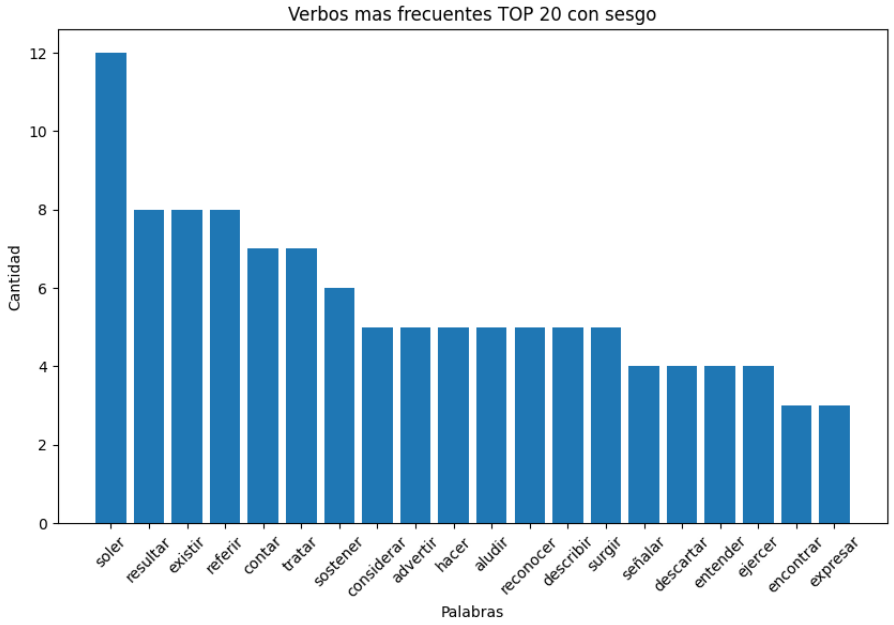

Figure 14: Top 20 Verbos con sesgo

Verbos Más Frecuentes sin sesgo

En este análisis la Figure 15, el término "hacer" es el más frecuente, con 421 menciones, seguido por "decir" con 326 y "encontrar" con 245 apariciones. Otros términos significativos incluyen "llevar" 197, "saber" 191 y "quedar" 185. Además, "recordar" 172 y "llegar" 154 destacan por su relevancia en el contexto de acciones y procesos. Otros términos como "salir" 135, "llamar" 134 y "dejar" 129 sugieren la importancia de interacciones y movimientos en las situaciones analizadas. Finalmente, verbos como "pasar" 128, "hablar" 127 y "manifestar" 122 indican la relevancia de la comunicación y la expresión en los casos estudiados.

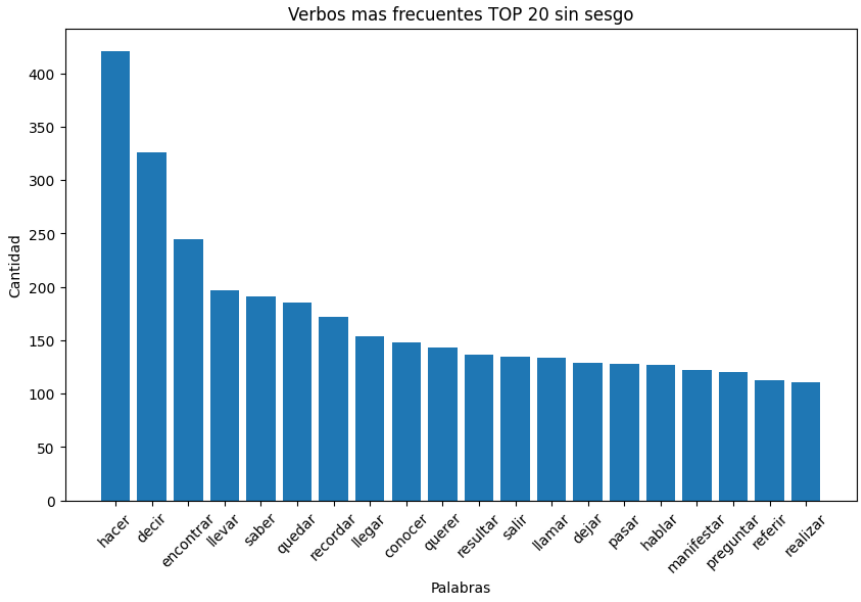

Figure 15: Top 20 Verbos sin sesgo

Adjetivos Más Frecuentes

La Figure 16 muestra los 20 adjetivos más frecuentes en las sentencias. Se destaca la aparición del adjetivo “sexual”, lo que sugiere que muchas de las sentencias tienen connotaciones relacionadas con cuestiones de género y sexualidad. Otros adjetivos siguen una distribución más uniforme.

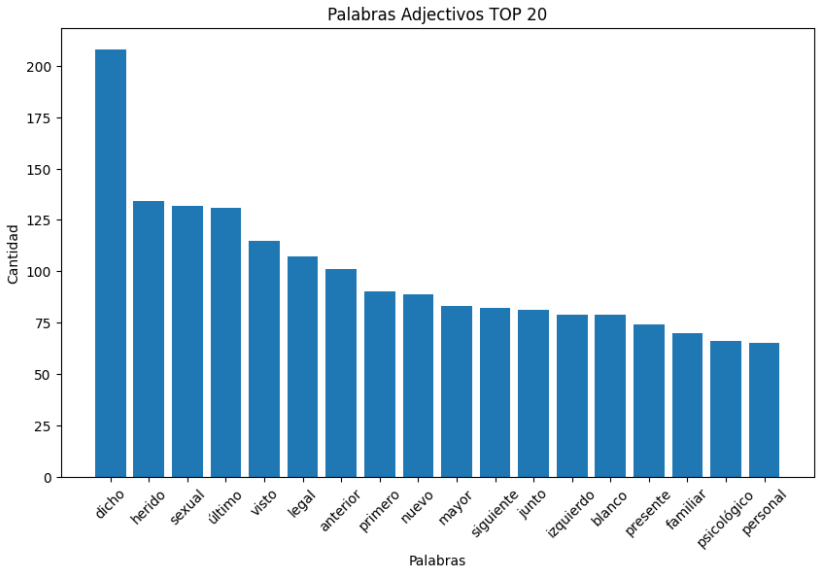

Figure 16 Adjetivos TOP 20

Adjetivos Más Frecuentes sin sesgo

En este análisis la Figure 17, el término "dicho" es el más frecuente, con 205 menciones, seguido por "primero" con 175 y "herido" con 134 apariciones. Otros términos significativos incluyen "último" 126, "sexual" 123 y "visto" 114. Además, "legal" 107 y "fiscal" 99 destacan por su relevancia dentro del contexto legal. Palabras como "blanco", "anterior", "derecho" y "judicial" aparecen entre 96 y 98 veces, reflejando aspectos tanto físicos como judiciales. Otros términos como "nuevo" 89, "presente" 86 y "declarante" 84 sugieren la importancia de la secuencia temporal y la participación de testigos en los casos analizados.

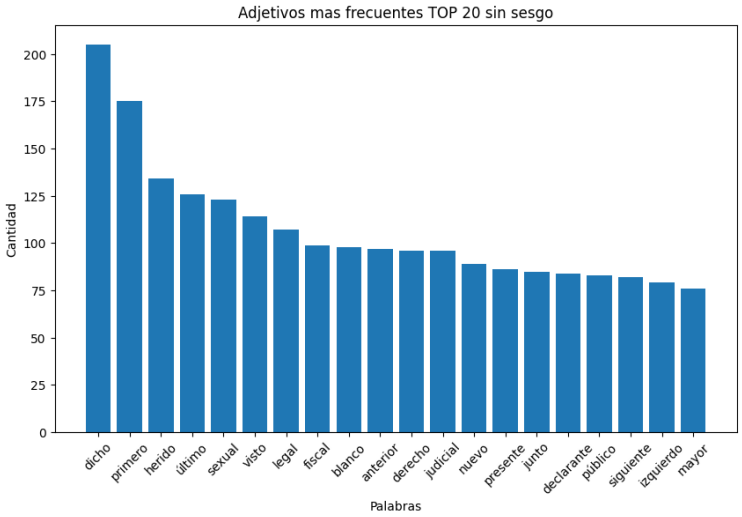

Figure 17: Adjetivos TOP 20 sin sesgo

Adjetivos Más Frecuentes con sesgo

El análisis de la Figure 18 muestra que el término "sexual" es el más frecuente, con 10 menciones. A continuación, "dicho" aparece 8 veces, seguido de "mayor" y "ocurrido" con 7 menciones cada uno. Otros términos destacados incluyen "primero" y "carnal" con 6 apariciones, mientras que "último", "creíble" y "violento" registran 5 menciones. Asimismo, palabras como "posible", "físico", "chambi" y "anterior" tienen 4 apariciones. Este conjunto de términos refleja temas relacionados con la violencia, la credibilidad de los hechos y el proceso judicial, con un enfoque en lo físico y emocional.

4o

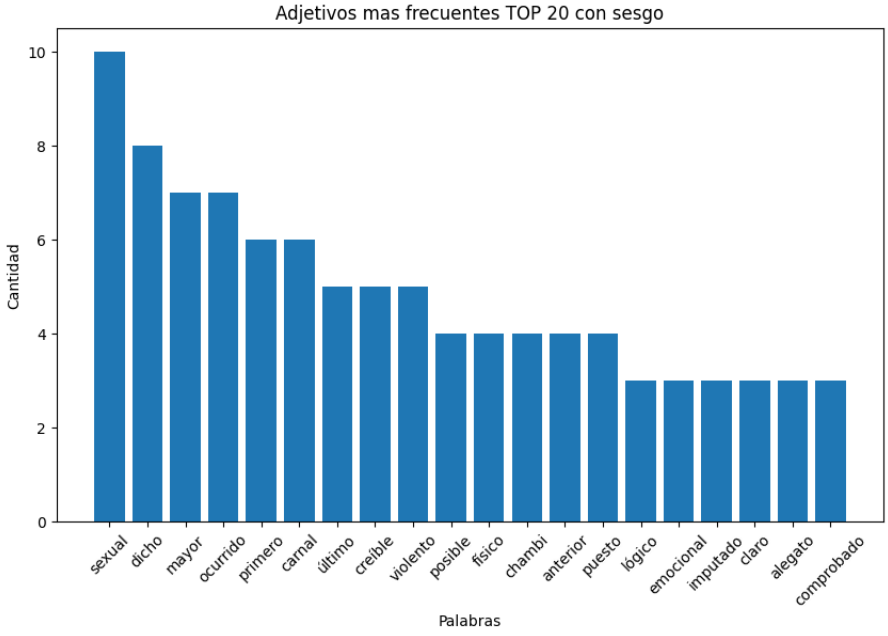

Figure 18: Adjetivos TOP 20 con sesgo

Sustantivos Femeninos Más Frecuentes

En la Figure 19, se presentan los 20 sustantivos femeninos más frecuentes. Los términos más comunes incluyen “situación”, “circunstancias” y “mujer”, lo que refleja la temática abordada en este trabajo. La distribución es relativamente uniforme entre los sustantivos femeninos.

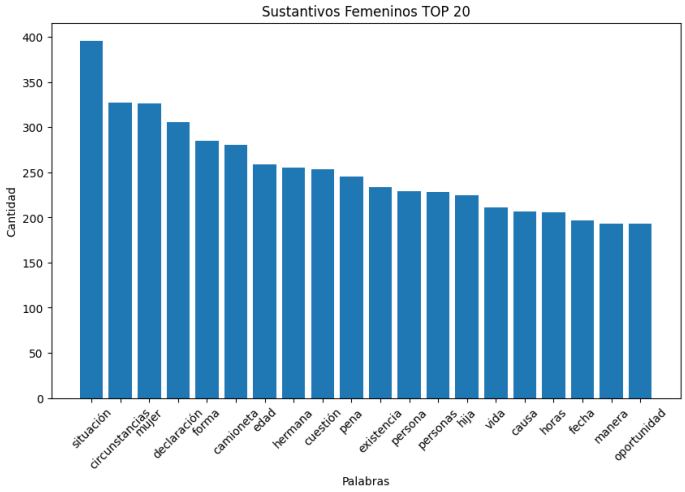

Figure 19: Top20 Sustantivos Femeninos

Sustantivos Femeninos Más Frecuentes con sesgo

En este conjunto de la Figure 20, el término "existencia" es el más frecuente con 10 menciones. A continuación, se destacan "edad", "opinión", "relaciones" y "declaración" con 7 apariciones cada uno. Otros términos importantes incluyen "situación" 6, "conclusión" y "razón" con 5 menciones cada uno, seguidos por "dudas" y "contradicciones" con 4 apariciones.

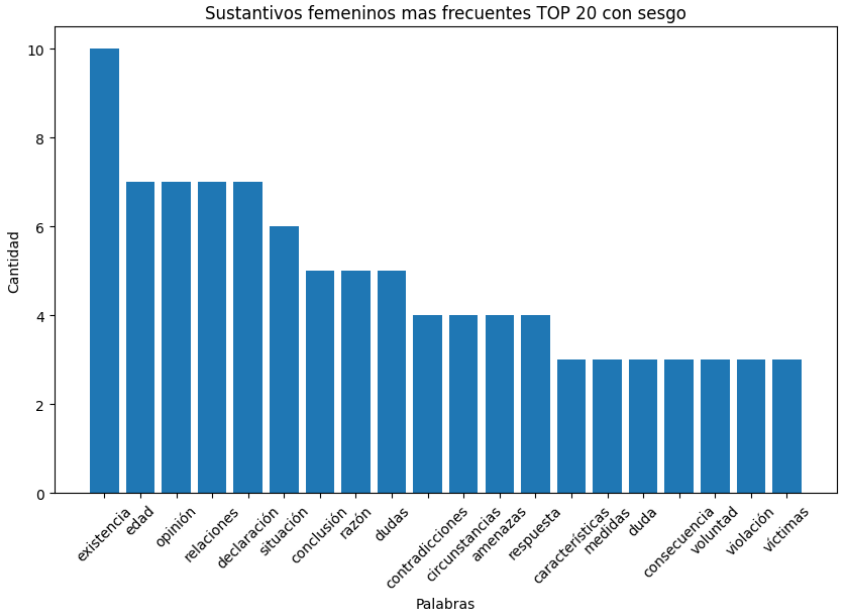

Figure 20: Sustantivos Femeninos Top 20 con sesgo

Sustantivos Femeninos Más Frecuentes sin sesgo

El análisis de la Figure 21 los datos muestran que el término "situación" es el más frecuente, con 123 menciones, seguido de "circunstancias" con 107 y "veces" con 93. Otros términos destacados incluyen "horas" (92), "declaración" (88), "sentencia" (86) y "panadería" (86). También son relevantes palabras como "forma" (85), "vida" (81), "lesiones" (80) y "cuota" (79).

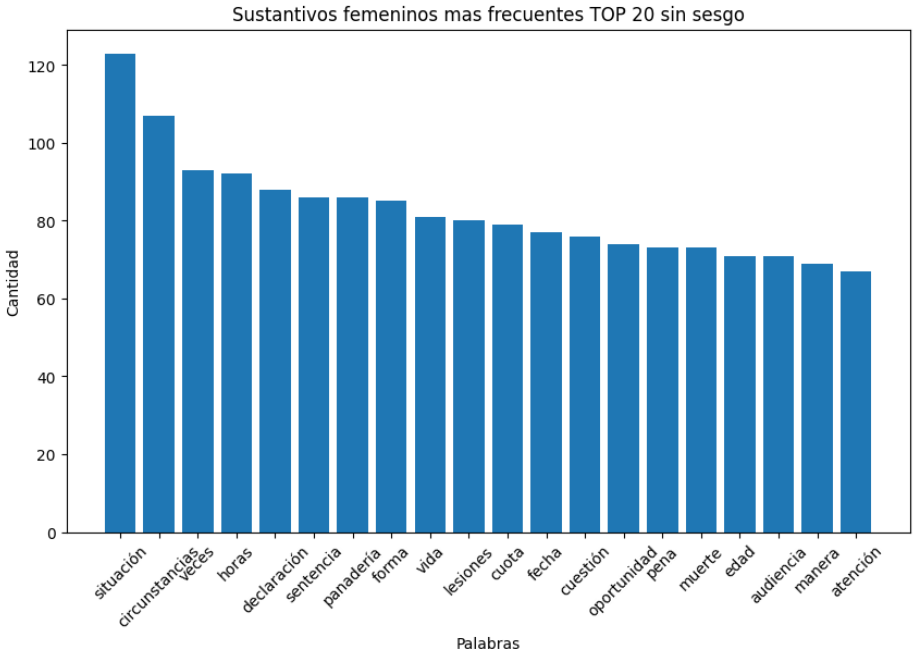

Figure 21: Sustantivos Femeninos Top 20 sin sesgo

Sustantivos Masculinos Más Frecuentes

Finalmente, la Figure 22 ilustra los 20 sustantivos masculinos más frecuentes. En este caso, se observa que muchos de los sustantivos están relacionados con tecnicismos jurídicos, como “debate”, “padre”, “recurso” y “juicio”. Esto indica que la mayoría de los términos masculinos son propios del ámbito judicial.

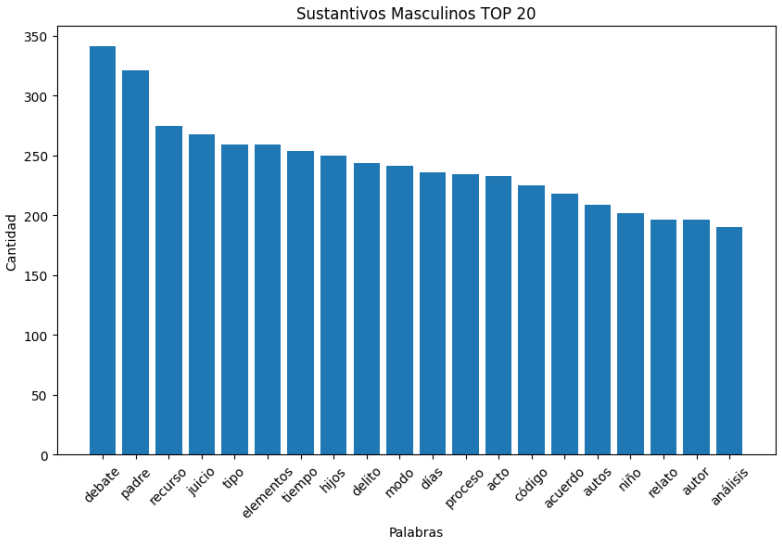

Figure 22: Top 20 Sustantivos Masculinos

Sustantivos Masculinos Más Frecuentes sin sesgo

El análisis de la Figure 23 los datos revelan que el término "relato" es el más frecuente, apareciendo en 9 ocasiones. A continuación, los términos "casos", "suceso", "embarazo", "testimonios", "acceso" y "acto" presentan una frecuencia de 5 menciones cada uno. Términos como "párrafo", "hechos", "consentimiento" y "autor" aparecen 4 veces, mientras que otros como "testimonio", "detalles" y "resultado" figuran 3 veces.

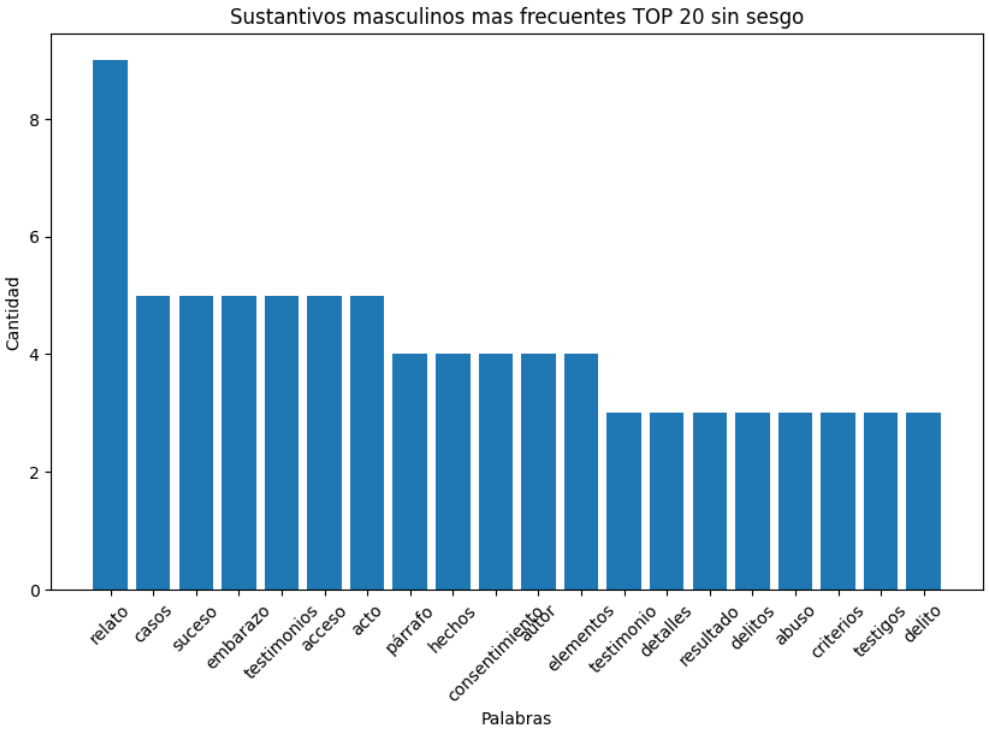

Figure 23: Top 20 Sustantivos Masculinos sin sesgo

Sustantivos Masculinos Más Frecuentes con sesgo

En la Figure 24 análisis de los datos revela las palabras más frecuentes en los párrafos evaluados. El término "tipo" aparece con una frecuencia de 101, seguido de "tiempo" con 92 menciones y "días" con 90. Términos como "acuerdo" y "elementos" aparecen 83 y 79 veces, respectivamente.

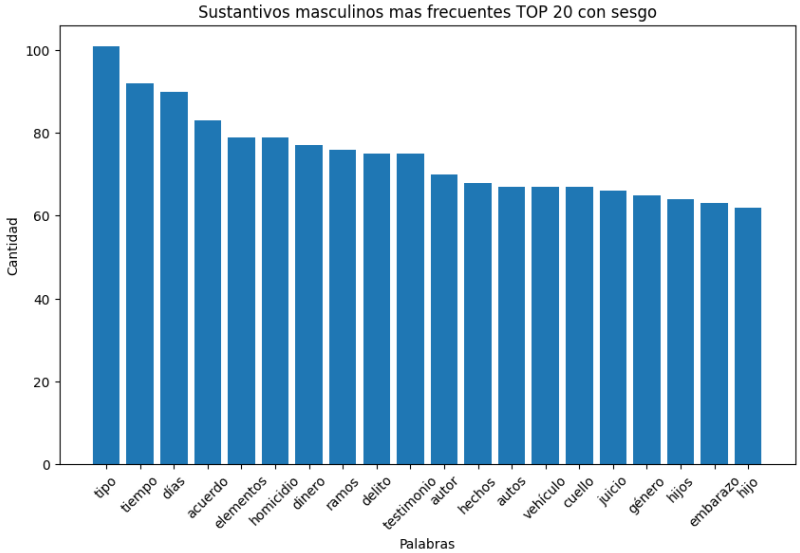

Figure 24: Top 20 Sustantivos Masculinos con sesgo

4.5 Clasificación basada en modelos tradicionales

Dado el dominio de estudio, donde predominan los párrafos sin sesgo (clase negativa), el conjunto de datos presenta un desbalance entre las clases, lo que puede llevar a un sesgo en la clasificación hacia la clase mayoritaria. Este desbalance puede hacer que el modelo no logre identificar correctamente las instancias de la clase positiva (sesgo de género). Para abordar este problema, se requieren enfoques específicos (por ejemplo, ajuste de pesos de clase) que permitan a un clasificador aprender de manera equilibrada las instancias de la clase minoritaria y, por ende, identificar con mayor precisión el sesgo de género en el texto. Además, la elección cuidadosa de características y la validación cruzada son fundamentales para garantizar la robustez y la capacidad de generalización de los modelos resultantes.

4.6 Optimización de Hiperparámetros

La optimización de hiperparámetros busca optimizar el desempeño de cada modelo, garantizando un balance adecuado entre sesgo y varianza, así como una mejor capacidad de generalización en distintos escenarios de clasificación, y una reducción del riesgo de sobreajuste. Esta etapa implica la selección de parámetros clave que no se aprenden directamente de los datos, sino que deben ser ajustados antes de entrenar el modelo.

4.6.1 Métodos de Optimización:

- **Grid Search**: Implica una búsqueda exhaustiva a través de un espacio predefinido de hiperparámetros, evaluando todas las combinaciones posibles. Aunque es preciso, este enfoque es computacionalmente costoso.
- **Random Search**: Selecciona muestras aleatorias del espacio de hiperparámetros, lo que lo hace más eficiente que Grid Search, especialmente cuando algunos parámetros tienen un impacto marginal.
- **Métodos Bayesianos**: Modelan la función objetivo para encontrar combinaciones prometedoras de manera iterativa y eficiente.

4.6.2 Parámetros a optimizar

En el modelo de regresión logística, se consideraron varios parámetros de ajuste. El parámetro `logreg__C` corresponde al coeficiente de regularización C, el cual controla la magnitud de los coeficientes del modelo con el objetivo de mitigar el sobreajuste. Por otro lado, `logreg__solver` define el algoritmo de optimización empleado para ajustar el modelo, siendo `'liblinear'` y `'lbfgs'` algunos de los valores posibles. Además, el parámetro `logreg__class_weight` permite ajustar el peso de las clases en la función de pérdida de la regresión logística, lo que resulta especialmente relevante en escenarios con conjuntos de datos desbalanceados; sus valores posibles incluyen `None` y `balanced`.

En el caso Naïve Bayes Gaussiano, se ajustó el parámetro `var_smoothing`, cuyo propósito es mejorar la estabilidad del modelo frente a problemas numéricos que puedan surgir en el cálculo de probabilidades. Para el modelo de RF, se exploraron diversas configuraciones para optimizar su desempeño. El parámetro `rf__n_estimators` controla la cantidad de árboles que componen el bosque aleatorio, mientras que `rf__max_depth` define la profundidad máxima permitida para cada árbol. Asimismo, `rf__min_samples_split` establece el número mínimo de muestras requerido para dividir un nodo interno en el árbol de decisión, y `rf__min_samples_leaf` especifica el número mínimo de muestras que debe contener una hoja o nodo terminal. En el caso de Naïve Bayes con suavizado de Laplace, el parámetro `nb__alpha` regula el suavizado de Laplace, una técnica

utilizada para evitar la asignación de probabilidades nulas a eventos no observados en el conjunto de entrenamiento.

Para el SVM, los principales hiperparámetros ajustados incluyen `C`, que controla el grado de regularización del modelo e influye en el margen de separación de las clases, y `kernel`, que define el tipo de núcleo utilizado para transformar los datos a un espacio de mayor dimensionalidad, con valores comunes como `'linear'` y `'rbf'`. Adicionalmente, gamma regula el coeficiente del núcleo en el caso de kernels no lineales, afectando la influencia de las instancias de entrenamiento en la función de decisión. Por último, el parámetro `class_weight` ajusta el peso de las clases en la función de pérdida, lo que resulta útil en escenarios donde las clases están desbalanceadas. Por el contrario, para CatBoost no se realizaron ajustes de hiperparámetros, salvo el ajuste de los pesos de clase, ya que este parámetro tiene un impacto significativo en el rendimiento del modelo.

4.7 Evaluación y Validación

En el contexto de la evaluación y validación de modelos de clasificación, especialmente cuando se trabaja con conjuntos de datos desbalanceados, es esencial garantizar que el modelo no solo se ajuste bien a los datos de entrenamiento, sino que también sea capaz de generalizar correctamente a datos no vistos.

Para abordar este problema, se utilizó **StratifiedKFold** con 5 *folds*, que aplica la estratificación para garantizar que la proporción de ejemplos de cada clase en cada *fold* sea similar a la proporción general del conjunto de datos. De este modo, tanto los conjuntos de entrenamiento como los de validación mantienen una distribución más fiel de las clases, lo que asegura una evaluación más robusta y representativa del desempeño del modelo

4.8 Clasificación basada en LLMs

En este enfoque, los Modelos de Lenguaje Grande (LLMs) juegan un papel central al realizar la clasificación del texto. Su desempeño depende en gran medida de cómo se define su *rol* en la tarea y cómo se estructura el *prompt* proporcionado al modelo.

Rol del LLM: Se asignó el rol de "*abogado penalista*" para determinar si un fragmento de sentencia judicial presenta sesgo de género,de la siguiente manera:

```
Sos un abogado penalista y tenés la tarea de determinar si el fragmento de sentencia judicial dado manifiesta sesgo de género.
```

Luego, a partir de la tarea a realizar (por ejemplo, de clasificación), se espera que el LLM generalice en base a su comprensión de las indicaciones proporcionadas y pueda abordar una gama amplia de problemas, adaptándose a diferentes situaciones. En este caso, para comunicar al LLM qué se espera que haga se definieron las siguientes instrucciones e indicaciones respecto a la salida:

```
[DESCRIPCION-TAREA] Realizá las siguientes tareas:
```

1. Identificá la categoría más probable del texto proporcionado.
2. Asigná el texto a dicha categoría.
3. Retorná la respuesta en formato JSON con la clave “label” y su valor correspondiente.

Los LLMs han mostrado buenas capacidades de generalización en escenarios *zero-shot* y *few-shot* learning. A diferencia de tareas de NLP tradicional, como la clasificación de texto, en las que se entrena un modelo para clasificar un conjunto predefinido de clases o etiquetas, en estos escenarios el modelo no es entrenado sobre los datos disponibles (si es que existen), sino que recibe un *prompt* que da información sobre los (pocos) datos disponibles para realizar la tarea.

Estrategias de *Prompting*

Zero-shot prompting. En esta estrategia, el modelo realiza la tarea sin haber visto ejemplos previos. De esta forma, el *prompt* definido incluye únicamente el texto a clasificar:

```
Se te dará la siguiente información:
   • Un texto delimitado por comillas triples.
   • Una lista de categorías a las cuales puede asignarse el
     texto. La lista está delimitada por corchetes. Las categorías
     en la lista se encuentran entre comillas simples y separadas
     por comas.

[DESCRIPCION-TAREA]

Lista de categorías: {labels}
Texto: “{text}”
Tu respuesta en formato JSON:
```

La Figure 25 presenta un ejemplo de clasificación utilizando el prompt anterior. En este caso, el LLM recibe el texto a analizar, las categorías potenciales (como "sesgo" y "no sesgo") e instrucciones sobre el formato esperado para la respuesta.

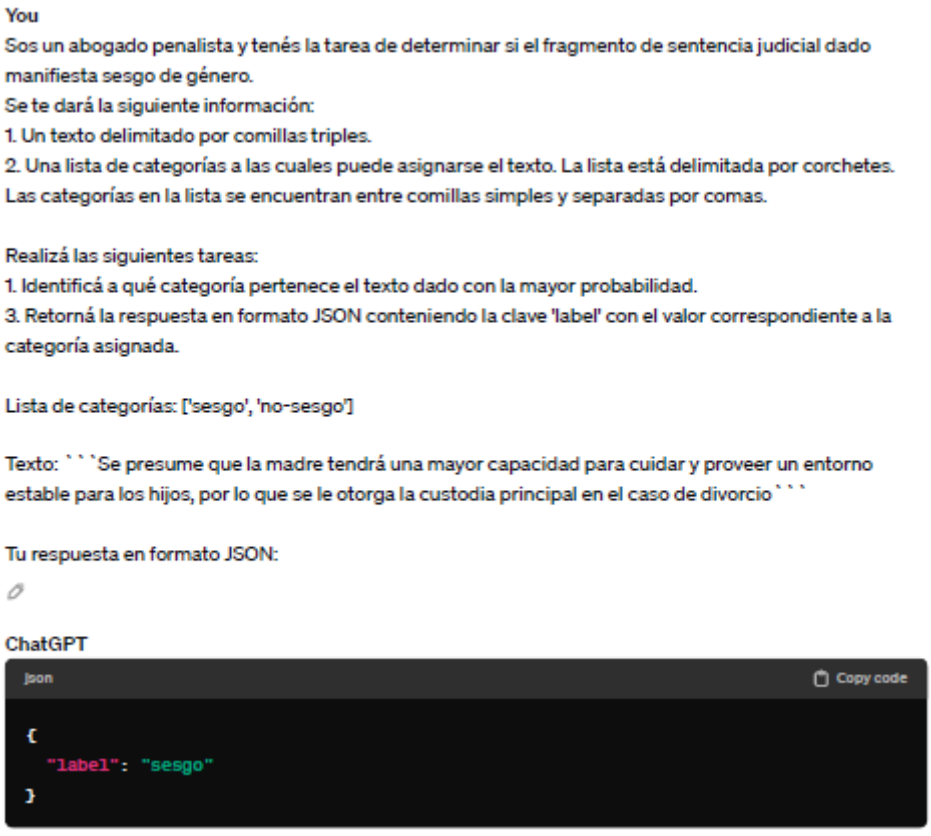

Figure 25 : Ejemplo de clasificación de un párrafo siguiendo la estrategia zero-shot con GPT-3.5

Few-shot prompting. En esta estrategia, se incorporan un número reducido de ejemplos de entrenamiento dentro del prompt. Estos ejemplos ayudan al modelo a generar respuestas más precisas y contextualizadas. Un aspecto relevante es que el

uso de estrategias **few-shot** puede ser más confiable que el **fine-tuning** para la incorporación de nuevo conocimiento en el LLM (O. Ovadia, 2023). Para el caso de few-shot, el prompt se modifica para incorporar los ejemplos a utilizar como entrenamiento:

```
Se te dará la siguiente información:

   • Un texto delimitado por comillas triples.
   • Una lista de categorías a las cuales puede asignarse el
     texto. La lista está delimitada por corchetes. Las categorías
     en la lista se encuentran entre comillas simples y separadas
     por comas.
   • Ejemplos de texto de entrenamiento y sus categorías
     correspondientes. Los ejemplos están delimitados por comillas
     triples. Para cada texto, se indica la categoría asignada.
     Estos ejemplos se deben usar como datos de entrenamiento.

[DESCRIPCION-TAREA]

Lista de categorías: {labels}
Texto: “{text}”
Tu respuesta en formato JSON:
```

Considerando la variabilidad de respuestas que pueden obtenerse de acuerdo con la selección de ejemplos de entrenamiento (J. Liu, May 2022) , se definieron dos escenarios:

- **Escenario Estático**. En este caso, los ejemplos de entrenamiento son fijos para todos los textos a analizar. La selección de los ejemplos se eligió porque habían sido clasificados erróneamente por el *zero-shot*, buscando una distribución balanceada de ejemplos de ambas clases.

A continuación, se muestra un ejemplo completo del *prompt* utilizando esta estrategia.

```
System: Sos un abogado penalista y tenés la tarea de determinar si
las sentencias judiciales tienen sesgo de género.

Human:
Se te dará la siguiente información:
1. Un texto delimitado por comillas triples.
```

```
2. Una lista de categorías a las cuales puede asignarse el texto. La lista está delimitada por corchetes. Las categorías en la lista se encuentran entre comillas simples y separadas por comas.
3. Ejemplos de texto de entrenamiento y sus categorías correspondientes. Los ejemplos están delimitados por comillas triples. Para cada texto, se indica la categoría asignada. Estos ejemplos se deben usar como datos de entrenamiento.

Realizá las siguientes tareas:
1. Identificá a qué categoría pertenece el texto dado con la mayor probabilidad.
2. Asigná el texto a dicha categoría.
3. Retorná la respuesta en formato JSON conteniendo solo la clave 'label' y el valor correspondiente a la categoría asignada.

Lista de categorías: ['sesgo', 'no-sesgo']

Textos de entrenamiento:

Texto ejemplo: ```Por todo lo expuesto, opino que corresponde hacer lugar a los recursos extraordinarios interpuestos y revocar el fallo apelado a fin de que, por intermedio de quien corresponda, se dicte uno nuevo de acuerdo a derecho. ```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```Recurso extraordinario interpuesto por el Dr. Pedro Eugenio Despouy Santoro, abogado defensor de María Cecilia Leiva.```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```Los Jueces Javier Anzoátegui y Marcela Rodríguez dijeron: Por sus razones y fundamentos, adhieren en su totalidad al voto del magistrado que lidera el acuerdo.```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```Contra dicho pronunciamiento, la Defensora Oficial, Dra. Fabiana Vannini, interpuso recurso de casación (fs. 127/150 del legajo 103.123). ```
Categoría de texto ejemplo: no-sesgo

Texto: ```Por lo expresado, se observa que, en el caso a estudio, no se da ninguna circunstancia extraordinaria que neutralice el trascendente significado que se concede al vínculo como agravante del delito. En consecuencia, y por los fundamentos expresados precedentemente corresponde casar parcialmente el fallo dejando sin efecto la sentencia impugnada en cuanto la existencia de circunstancias extraordinarias de atenuación y se declara a M. E. P. responsable penalmente del delito de homicidio en los términos del art. 80 inc. 1°. ```
```

```
Tu respuesta en formato JSON:
```

- **Escenario Dinámico**. Los ejemplos se eligen de acuerdo con el texto a analizar (Pezeshkpour P., 2020). Para cada texto se eligen sus N vecinos más cercanos[35], de acuerdo con su representación de *embeddings* previamente descripta. La cercanía o semejanza se calculada mediante la distancia Euclídea. Este enfoque se fundamenta en la premisa de que cuanto más similares son los ejemplos, más útil será la información que aportan al modelo para generar respuestas satisfactorias (B. Pecher, 2024).

A continuación, se muestra un ejemplo completo del *prompt* utilizando esta estrategia.

```
Prompting: System: Sos un abogado penalista y tenés la tarea de
determinar si las sentencias judiciales tienen sesgo de género.

Human:
Se te dará la siguiente información:
1. Un texto delimitado por comillas triples.
2. Una lista de categorías a las cuales puede asignarse el texto.
La lista está delimitada por corchetes. Las categorías en la lista
se encuentran entre comillas simples y separadas por comas.
3. Ejemplos de texto de entrenamiento y sus categorías
correspondientes. Los ejemplos están delimitados por comillas
triples. Para cada texto, se indica la categoría asignada. Estos
ejemplos se deben usar como datos de entrenamiento.

Realizá las siguientes tareas:
1. Identificá a qué categoría pertenece el texto dado con la mayor
probabilidad.
2. Asigná el texto a dicha categoría.
3. Explicá por qué asignaste el texto a dicha categoría.
4. Retorná la respuesta en formato JSON conteniendo la clave
'label' con el valor correspondiente a la categoría asignada, la
clave 'explicación' con la explicación de la asignación a la
categoría y la clave 'keywords' con las palabras que resuman los
tópicos clave de la explicación.

Lista de categorías: ['sesgo', 'no-sesgo']
```

[35] La selección se realizó utilizando el algoritmo de Nearest Neighbors https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html

Textos de entrenamiento:

Texto ejemplo: ```"S , J M s/ abuso sexual -art. 119 3° párrafo-" CSJ 873/2016/CS1 Suprema Corte: 1 El Superior Tribunal de Justicia de Rio Negro, por mayoría, rechazó los recursos de casación interpuestos por la Defensora de Menores e Incapaces y la parte querellante, contra la sentencia por la que la Sala A de la Cámara en lo Criminal de Viedma absolvió a J M S en orden al delito de abuso sexual agravado por el acceso carnal y el aprovechamiento de la situación de convivencia preexistente -artículo 119, párrafos primero, tercero y cuarto, del Código Penal(fs. 578/589 del principal). ```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```IV Si bien la apreciación de la prueba constituye, por vía de principio, facultad propia de los jueces de la causa y no es susceptible de revisión en la instancia extraordinaria (Fallos: 332:2659), la Corte puede conocer en los casos cuyas particularidades hacen excepción a esa regla con base en la doctrina de la arbitrariedad (Fallos: 327:5456 y sus citas) ya "S , J M si abuso sexual -art. 1193° párrafo-" CSJ 873/2016/CS1 que con ésta se procura asegurar las garantías constitucionales de la defensa en juicio y el debido proceso exigiendo que las sentencias sean fundadas y constituyan una derivación razonada del derecho vigente con aplicación a las constancias efectivamente comprobadas en la causa (Fallos: 315:2969; 321:1909; 326:8; 327:5456; 334:725, considerando 4° y sus citas). ```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```En ese aspecto, cabe poner de relieve la doble condición de la niña, tanto de menor de edad como de mujer, que la vuelve particularmente vulnerable a la violencia (conf. Corte Interamericana de Derechos Humanos, "Caso González y otras -' Campo Algodonero' vs. México", sentencia del 16 de noviembre de 2009, parágrafo 408; en el mismo sentido, "Caso Veliz Franco y otros vs. Guatemala", sentencia del 19 de mayo de 2014, parágrafo 134). ```
Categoría de texto ejemplo: no-sesgo

Texto ejemplo: ```Pienso que por haber hecho hincapié en esos aspectos -el supuesto desinterés, hipotéticas contradicciones y la omisión de detalles que ni siquiera se ocupó de particularizarla mayoría se apartó de los estándares internacionales mencionados para el juzgamiento de esta clase de hechos, y relativizó el relato de la niña a pesar de que, conforme lo "S , J M s/ abuso sexual -art. 1193° párrafo-" CSJ 873/2016/CSl valoró el voto en minoría, los informes psicológicos descartaron la presencia de elementos fabulosos y de tendencia a la fabulación, sus maestras destacaron su honestidad, y aquélla expuso -en los términos que le

```
permitió su edad y desarrollo información precisa, relevante y sustancial acerca del lugar en que ocurrieron los abusos denunciados, cómo se desarrollaron, los concretos actos en que consistieron y las palabras que intercambió con el imputado (fs. 581 vta., 583 vta. y 585 vta.). ```
Categoría de texto ejemplo: no-sesgo

Texto: ```"S , J M s/ abuso sexual -art. 119 3º párrafo-" CSJ 873/2016/CS1 Suprema Corte: 1 El Superior Tribunal de Justicia de Rio Negro, por mayoría, rechazó los recursos de casación interpuestos por la Defensora de Menores e Incapaces y la parte querellante, contra la sentencia por la que la Sala A de la Cámara en lo Criminal de Viedma absolvió a J M S en orden al delito de abuso sexual agravado por el acceso carnal y el aprovechamiento de la situación de convivencia preexistente -artículo 119, párrafos primero, tercero y cuarto, del Código Penal(fs. 578/589 del principal). ```

Tu respuesta en formato JSON:
```

Como los LLMs se pueden ajustar a instrucciones en lenguaje natural para proveer respuestas “útiles”, también pueden generar explicaciones sobre dichas respuestas (S. Huang, 2023). Por ejemplo, al analizar el sentimiento de un texto, el modelo puede retornar no solo la valencia del sentimiento sino también una explicación, listando palabras o frases cargadas de sentimiento que se encuentren en el texto analizado. Para el dominio de sesgo de género, se modificó la descripción de la tarea de forma tal que, luego de realizar la predicción de la categoría, el LLM genere una explicación que justifique dicha categoría.

El nuevo *prompt* (la parte agregada está subrayada) se formula como:

```
[DESCRIPCIÓN-TAREA] Realizá las siguientes tareas:
1. Identificá a qué categoría pertenece el texto dado con la mayor probabilidad.
2. Asigná el texto a dicha categoría.
3. Explicá por qué asignaste el texto a dicha categoría.
4. Retorná la respuesta en formato JSON conteniendo la clave ’label’ con el valor correspondiente a la categoría asignada y la clave ’explicación’ con la explicación de la asignaci´on a la categoría.
```

Los modelos de LLM utilizados, tanto para clasificación como para explicaciones, se resumen en la Tabla 12. Los LLMs fueron accedidos mediante el framework **LangChain**[36], configurando su temperatura en 0 para reducir el azar en las respuestas[37]. Cada texto fue procesado en una interacción separada con Langchain, y no se utilizó memoria conversacional.

Familia del Mod.	Modelo	Creador	Año	Tokens
GPT	GPT-3.5-turbo	OpenAI	2022	4.096 max
Llama2	Llama2-7b	Meta	2023	4.096 max
Mistral	Mistral-7B	Mistral AI	2023	8.000 max
Gemma	Gemma-2b	Google	2024	2048 max

Tabla 12 LLMs utilizados

4.9 Otras herramientas

Los modelos entrenados fueron complementados con dos herramientas de terceros: Themis[38] y GenBit[39].

Themis se promociona como un corrector ortográfico gramatical para eliminar el sesgo de género y fomentar la escritura inclusiva. Se basa en detectar términos que considera sesgados de un diccionario y en realizar sugerencias para adaptar el texto. Los términos con sesgo son aquellos que son morfológicamente masculinos, ya que la herramienta aclara que los términos en femenino no se consideran excluyentes. Sin embargo, la herramienta no parece tomar en cuenta el contexto de uso, lo que no siempre resulta adecuado. Se considera que un párrafo contiene sesgo si contiene alguno de los términos del diccionario utilizado por Themis.

GenBit es una herramienta de Microsoft dedicada a medir el sesgo en texto, con el fin de determinar si el género está distribuido de manera uniforme en los datos. Para esto, analiza la asociación entre una lista predefinida de palabras relacionadas

[36] https://python.langchain.com/v0.2/docs/introduction/
[37] https://platform.openai.com/docs/api-reference/chat
[38] https://themis.es/
[39] https://github.com/microsoft/responsible-ai-toolbox-genbit

con el género y otras palabras en el corpus mediante estadísticas de coocurrencia. Para determinar la presencia de sesgo, se utilizó el umbral (*threshold*) predefinido por la herramienta.

4.10 Métricas de evaluación

La evaluación se basó en **precisión** y **recall** tanto para la clase positiva (sesgo) y su versión ponderada de acuerdo con el peso de las clases[40]. Dada la naturaleza de la tarea, se considera que el *recall* es más relevante que la precisión. También se incluyó el **Coeficiente de Matthews (MCC).** Sobre el conjunto de métricas de clasificación se realizó un análisis estadístico por pares ($\alpha = 0.01$)

[40] Todas las metricas fueron implementadas utilizando scikit-learn: https://scikit-learn.org/stable/modules/model_evaluation.html

Capítulo 5

Evaluación experimental

En este capítulo se enfoca en la evaluación experimental, la cual se refiere al proceso de evaluar y medir el rendimiento, la efectividad o la validez del sistema propuesto. Esta es una etapa crítica en la investigación, ya que permite validar y respaldar los resultados presentados en el trabajo.

5.1 Modelado de tópicos

El modelado de tópicos es una técnica de análisis de texto cuyo objetivo es identificar y comprender los temas latentes y transversales dentro de un conjunto de documentos. Este enfoque permite descubrir patrones ocultos y extraer información significativa acerca de los temas tratados. Un tópico se describe generalmente mediante un conjunto de palabras o frases asociadas con una temática particular. Los tópicos no siempre están distribuidos de manera uniforme entre los documentos. Un único documento puede abarcar varios tópicos, y un tópico puede estar presente en diversos documentos (por ejemplo, una sentencia o un párrafo de una sentencia).

Por ejemplo, la Figure 26 ilustra los principales tópicos identificados por BERTopic[41] en los párrafos de un cuerpo de sentencias. Cada tópico está representado por un conjunto de cinco palabras clave que resumen su temática. En esta representación, cada punto en la figura corresponde a un párrafo, y estos puntos están representados mediante *embeddings* generados por BETO. La disposición espacial en 2D de los *embeddings* refleja la distancia relativa entre los mismos, lo que facilita la interpretación de la proximidad temática de los párrafos.

[41] https://maartengr.github.io/BERTopic/

La proyección en 2D es especialmente útil para entender la distribución y densidad de los diferentes tópicos dentro del cuerpo de las sentencias. Además, permite una visualización clara de aquellos párrafos que podrían aludir de manera más o menos explícita a estereotipos de sesgo de género. De esta forma, el modelado de tópicos no solo facilita el diagnóstico visual del corpus de sentencias, sino que también actúa como una fase preliminar para la posterior clasificación de los textos según los sesgos detectados

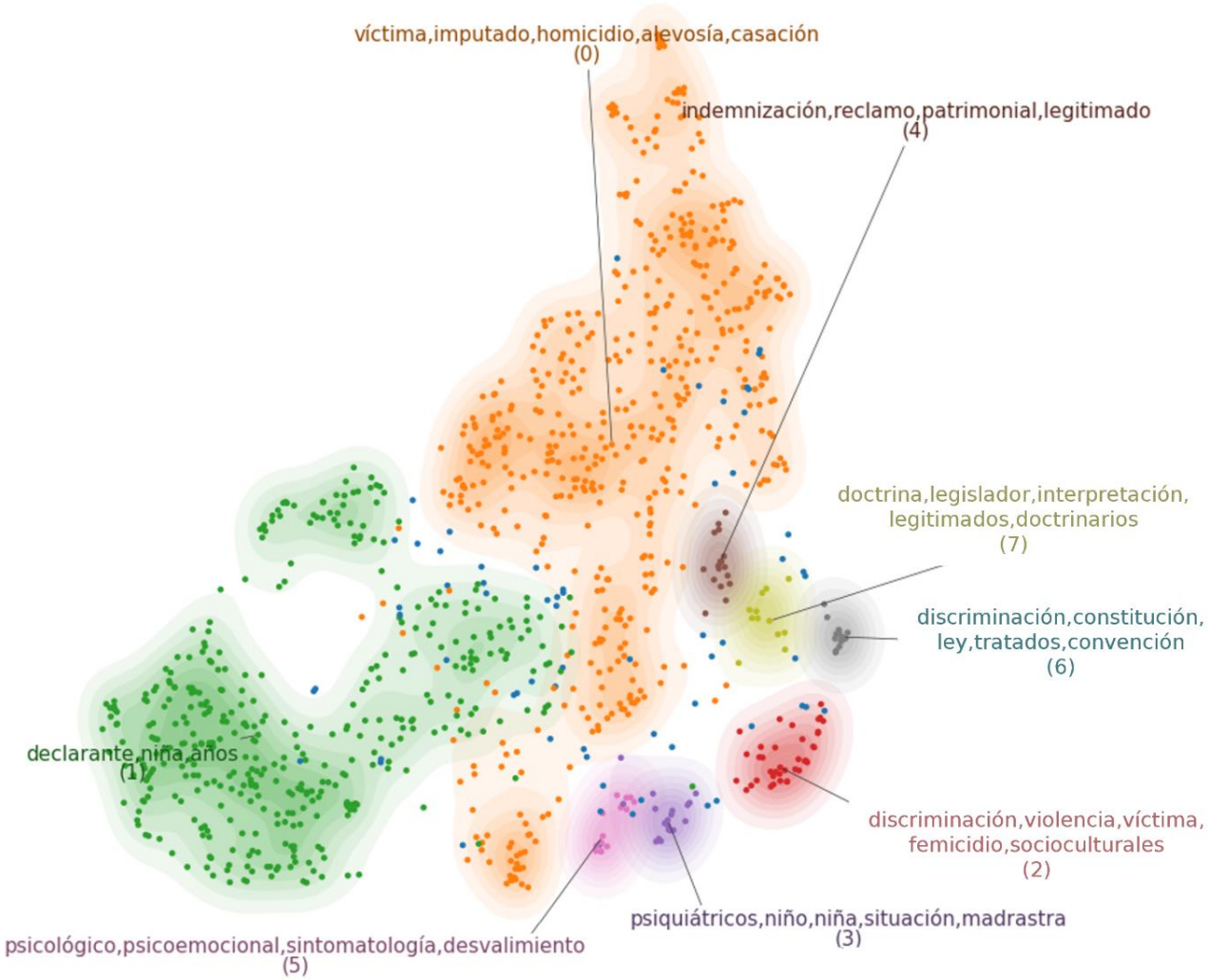

Figure 26: Análisis de tópicos generado por BERTopic sobre un cuerpo de sentencias. Proyección 2D con UMAP sobre embeddings de BETO.

Figure 27 : Cuerpo de párrafos con sesgo (en rojo) y sin sesgo (en gris) de género. Proyección UMAP sobre embeddings BETO

En particular, se identificaron los siguientes 7 tópicos: **0)** Referente a testimonios de víctimas en delitos contra la integridad sexual, y la relevancia de las declaraciones de los testigos, relacionados con las medidas judiciales tomadas respecto a situaciones de violencia. **1)** Relativo a la presentación de pruebas durante un juicio, tanto de la fiscalía como de la defensa. **2)** Asociado a hechos de violencia física explícita a razón de disputas por roles y reclamos de cuidado. **3)** Referente a violencia de género, y destacando la importancia de no ignorar las manifestaciones de violencia de este tipo, y de reconocer las condiciones que la constituyen. **4)** Enfocado en situaciones de violencia psicológica, sus efectos, y a los roles ocupados por víctimas y victimarios. **5)** Relativo a reclamos de igualdad de trato ante daños materiales y morales para ambos géneros. **6)** Asociado a efectos psicológicos y sus consecuencias sobre la vulnerabilidad personal y social.

Para contextualizar la tarea de clasificación, la Figure 27 presenta la distribución de las instancias de la clase positiva (en rojo), para los *embeddings* de BETO, donde

puede apreciarse el desbalanceo entre las clases. Es interesante destacar que la distribución de párrafos en los tópicos no resulta uniforme, siendo los tópicos 0 y 1 los que concentraron la mayor cantidad de párrafos. En ambos tópicos se observaron manifestaciones implícitas de sesgos, como parte de los alegatos de cada caso. De manera complementaria, los tópicos **2** y **4** refieren explícitamente a indicadores de sesgo de género, como estereotipos de roles de género, por ejemplo, el rol de cuidadora de una madre respecto a un menor a su cargo. Estos escenarios motivan la realización de un análisis semántico, más que basado en palabras clave, del texto para su clasificación. En los restantes cuatro tópicos no se evidencia la existencia de párrafos con sesgo.

5.2 Clasificación tradicional (P1-1)

5.2.1 Desempeño de Modelos de Clasificación en Representaciones Textuales

Se probaron varios métodos de optimización (GridSearchCV, Randomized SearchCV, Optuna[42]), siendo el más satisfactorio **RandomizedSearchCV** [43]**,** debido a su balance entre resultados y costo computacional.

Dado el gran desbalanceo de clases mencionado en el capítulo anterior, se empleó la técnica **SMOTE**[44]**,** para mejorar el balance de las clases y evitar el sobreajuste. Esto permitió mejorar la capacidad de los algoritmos para aprender patrones de la clase minoritaria.

La Tabla 13 representa un resumen de los resultados obtenidos, en el que se incluye, además, una clasificación basada en la clase mayoritaria como referencia. Esta clasificación sirve para establecer un punto de comparación, proporcionando los resultados mínimos alcanzables bajo dicho enfoque.

[42] https://optuna.org/
[43] https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[44] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Estrategia	Binaria (cla. sesgo)		Ponderada		Coef.
	Precisión	Recall	Precisión	Recall	Matthews
Mayoría	0.00	0.00	0.926	0.963	0.00
Random	0.4	0.5	0.96	0.954	0.424

Tabla 13: Resultados Baseline

A continuación, se presentan los resultados en función de las representaciones numéricas del texto.

Modelo	Sin SMOTE Media ± Dev.	Con SMOTE Media ± Dev.
BOW Log	**0.2580 ± 0.0899**	0.0439 ± 0.1216
BOW SVM	0.1210 ± 0.1366	0.0856 ± 0.0252
BOW Rand.Forest	0.0696 ± 0.1391	0.0420 ± 0.0877
BOW N. Bayes	0.1220 ± 0.1146	0.1320 ± 0.1425
BOW CatBoost	0.0875 ± 0.1237	0.0608 ± 0.0727
TF-IDF Log	**0.2443 ± 0.1170**	0.1874 ± 0.0871
TF-IDF SVM	0.2181 ± 0.1328	0.0850 ± 0.1439
TF-IDF Rand Forest	0.0669 ± 0.1405	0.0860 ± 0.1322
TF-IDF N. Bayes	0.0000 ± 0.0000	0.0989 ± 0.0168
TF-IDF CatBoost	0.0620 ± 0.0877	0.0168 ± 0.0989
Word2Vec Log	0.1094 ± 0.0617	0.1198 ± 0.0746
Word2Vec SVM	0.1220 ± 0.0555	0.1378 ± 0.0407
Word2Vec N. Bayes	-0.0063 ± 0.0812	-0.0007 ± 0.0738
Word2Vec Rand Forest	0.0000 ± 0.0000	**0.9981 ± 0.0023**
Word2Vec CatBoost	0.0000 ± 0.0000	0.0686 ± 0.0591
FastText Log	0.0503 ± 0.0622	0.0488 ± 0.1601
FastText SVM	0.0276 ± 0.2132	0.0163 ± 0.0929
FastText Rand.Forest	-0.0038 ± 0.0075	0.8938 ± 0.0340
FastText N. Bayes	0.0617 ± 0.1835	0.0502 ± 0.1678
FastText CatBoost	-0.0066 ± 0.0082	**0.9168** ± 0.0082

Tabla 14: Resumen MCC C. Tradicional

En la Tabla 14 se puede observar el análisis comparativo de distintos algoritmos y técnicas de representación de textos, se evaluaron BOW, TF-IDF, Word2Vec y FastText. A continuación, se destacan los mejores resultados obtenidos para cada representación, tomando en cuenta el MCC (Coeficiente de Correlación de Matthews): BOW: El mejor valor de MCC fue de 0.2580 utilizando el algoritmo Logístico sin aplicar la técnica de balanceo SMOTE. TF-IDF: El mayor MCC obtenido fue de 0.2443, también con el algoritmo Logístico y sin SMOTE. Word2Vec: Destacó con un MCC de 0.9981, utilizando el algoritmo Random Forest con SMOTE, siendo el

valor más alto entre todos los embeddings. FastText: El valor más alto de MCC fue de 0.9168, obtenido con el algoritmo CatBoost y aplicando SMOTE.

De estos resultados, Word2Vec con Random Forest y SMOTE fue la combinación que proporcionó el mejor rendimiento, alcanzando un MCC de 0.9981, superando significativamente al resto de los métodos.

Model	Sin SMOTE Media ± Dev.	Con SMOTE Media ± Dev.
BOW Log	**0.2711 ± 0.0735**	0.0916 ± 0.0916
BOW SVM	0.1085 ± 0.0916	0.0884 ± 0.0047
BOW Rand.Forest	0.0444 ± 0.0889	0.0872 ± 0.0556
BOW N. Bayes	0.1571 ± 0.1066	0.1374 ± 0.1268
BOW CatBoost	0.0444 ± 0.0629	0.1058 ± 0.0540
TF-IDF Log	**0.2718 ± 0.1083**	0.2103 ± 0.0762
TF-IDF SVM	0.2282 ± 0.1298	0.0752± 0.0946
TF-IDF Rand Forest	0.0444 ± 0.0889	0.0476 ± 0.0673
TF-IDF N. Bayes	0.0000 ± 0.0000	0.2578 ± 0.1361
TF-IDF CatBoost	0.0444 ± 0.0629	0.1058 ± 0.0075
Word2Vec Log	0.1240 ± 0.0366	0.1369 ± 0.0491
Word2Vec SVM	0.1418 ± 0.0374	0.1545 ± 0.0293
Word2Vec N. Bayes	0.0635 ± 0.0299	0.0651 ± 0.0282
Word2Vec Rand Forest	0.0000 ± 0.0000	**0.9991 ± 0.0012**
Word2Vec CatBoost	0.0000 ± 0.0000	0.0741 ± 0.0524
FastText Log	0.0890 ± 0.0283	0.0836 ± 0.0722
FastText SVM	0.0821 ± 0.1006	0.0672 ± 0.0378
FastText Rand.Forest	0.0000 ± 0.0000	0.9467 ± 0.0170
FastText N. Bayes	0.0964 ± 0.0841	0.0866 ± 0.0751
FastText CatBoost	0.0000 ± 0.0000	**0.9588 ± 0.0085**

Tabla 15: Resumen F1-score C. Tradicional

En la Tabla 15 el análisis de rendimiento de los distintos modelos y técnicas de representación de texto, se evaluaron los mejores F1-scores obtenidos. Los resultados mostraron diferencias significativas en la efectividad de cada embedding, tanto con cómo sin la aplicación de la técnica de balanceo de datos SMOTE.

BOW: El mejor F1-score fue de 0.2711, utilizando el algoritmo de regresión logística sin aplicar SMOTE. Esto indica que BOW logra un rendimiento moderado sin técnicas de balanceo. TF-IDF: El mayor valor de F1-score fue 0.2718, también con el algoritmo de regresión logística y sin SMOTE, mostrando un rendimiento similar a

BOW, pero ligeramente superior. Word2Vec: El mejor F1-score fue significativamente mayor, alcanzando 0.9991 con el algoritmo Random Forest y aplicando SMOTE. Este resultado destaca la capacidad de Word2Vec combinado con algoritmos de ensamble y balanceo de datos para capturar patrones complejos en los textos. FastText: El mejor F1-score obtenido fue de 0.9588, utilizando el algoritmo CatBoost con SMOTE, siendo también muy alto pero inferior al de Word2Vec.

En resumen, el mejor rendimiento general lo obtuvo Word2Vec con el algoritmo Random Forest y SMOTE, alcanzando un F1-score de 0.9991.

Modelo	**Sin SMOTE (Recall) Media ± Dev.**	**Con SMOTE (Recall) Media ± Dev.**
BOW Log	0.2444 ± **0.0798**	0.2694 ± 0.2432
BOW SVM	0.0722 ± 0.0592	**0.9528 ± 0.0580**
BOW Rand.Forest	0.0250 ± 0.0500	0.3167 ± 0.2017
BOW N. Bayes	0.1750 ± 0.1275	0.1000 ± 0.0935
BOW CatBoost	0.0238 ± 0.0337	0.2894 ± 0.1122
TF-IDF Log	0.2944 ± 0.1033	0.1972 ± 0.0648
TF-IDF SVM	0.1972 ± 0.1292	0.0500 ± 0.0612
TF-IDF Rand Forest	0.0250 ± 0.0500	0.0256 ± 0.0363
TF-IDF N. Bayes	0.0000 ± 0.0000	**0.3444 ± 0.1870**
TF-IDF CatBoost	0.0256 ± 0.0363	0.0733 ± 0.0026
Word2Vec Log	0.6083 ± 0.0565	0.5083 ± 0.1219
Word2Vec SVM	0.4389 ± 0.1250	0.4389 ± 0.1250
Word2Vec N. Bayes	0.4469 ± 0.2560	0.4469 ± 0.2560
Word2Vec Rand Forest	0.0000 ± 0.0000	**0.9991 ± 0.0019**
Word2Vec CatBoost	0.0000 ± 0.0000	0.0495 ± 0.0350
FastText Log	0.5306 ± 0.1736	0.5000 ± 0.4472
FastText SVM	0.4000 ± 0.4899	0.7000 ± 0.4000
FastText Rand.Forest	0.0000 ± 0.0000	0.9440 ± 0.0188
FastText N. Bayes	0.5000 ± 0.4472	0.5000 ± 0.4472
FastText CatBoost	0.0000 ± 0.0000	**0.9839 ± 0.0088**

Tabla 16: Resumen Recall C. Tradicional

En el análisis de Recall la Tabla 16 muestra que, utilizando diferentes técnicas de representación de texto, los mejores resultados obtenidos son los siguientes:

BOW: El mejor valor de Recall fue 0.9528, alcanzado por el modelo SVM con SMOTE. TF-IDF: El valor más alto de Recall fue 0.3444, utilizando Naive Bayes con SMOTE. Word2Vec: El valor más destacado fue 0.9991, obtenido con Random Forest y SMOTE, lo que indica un rendimiento casi perfecto en la clasificación de la clase positiva. FastText: El mejor Recall fue de 0.9839, utilizando CatBoost con SMOTE, también mostrando un rendimiento sobresaliente. El modelo Word2Vec combinado con Random Forest y SMOTE fue el que alcanzó el mejor rendimiento en términos de Recall, con un valor de 0.9991.

Modelo	Sin SMOTE Media ± Dev.	Con SMOTE Media ± Dev.
BOW Log	0.3450 ± 0.1698	0.0553 ± 0.0473
BOW SVM	0.3000 ± 0.3674	0.0463 ± 0.0025
BOW Rand.Forest	0.2000 ± 0.4000	0.0506 ± 0.0323
BOW N. Bayes	0.1444 ± 0.0953	0.2400 ± 0.2245
BOW CatBoost	**0.3333 ± 0.4714**	0.0655 ± 0.0353
TF-IDF Log	0.2654 ± 0.1351	0.2467 ± 0.1323
TF-IDF SVM	0.3067 ± 0.1679	0.2400 ± 0.3878
TF-IDF Rand Forest	0.2000 ± 0.4000	**0.3333 ± 0.4714**
TF-IDF N. Bayes	0.0000 ± 0.0000	0.2079 ± 0.1116
TF-IDF CatBoost	0.1667 ± 0.2357	0.1976 ± 0.0438
Word2Vec Log	0.0695 ± 0.0220	0.0795 ± 0.0303
Word2Vec SVM	0.0850 ± 0.0234	0.0949 ± 0.0190
Word2Vec N. Bayes	0.0342 ± 0.0158	0.0352 ± 0.0150
Word2Vec Rand Forest	0.0000 ± 0.0000	**0.9991 ± 0.0019**
Word2Vec CatBoost	0.0000 ± 0.0000	0.1500 ± 0.1080
FastText Log	0.0486 ± 0.0154	0.0461 ± 0.0399
FastText SVM	0.0458 ± 0.0561	0.0356 ± 0.0201
FastText Rand.Forest	0.8000 ± 0.4000	**0.9496 ± 0.0196**
FastText N. Bayes	0.0535 ± 0.0467	0.0478 ± 0.0416
FastText CatBoost	0.6000 ± 0.4899	0.9351 ± 0.0124

Tabla 17: Resumen Precisión C. Tradicional

La Tabla 17 se muestran diferentes técnicas de representación de texto, los mejores resultados obtenidos son los siguientes: BOW: El mejor valor de Precisión fue 0.4714, alcanzado por el modelo CatBoost sin SMOTE. TF-IDF: El valor más alto de Precisión fue 0.4714, utilizando Random Forest con SMOTE. Word2Vec: El valor más destacado fue 0.9991, obtenido con Random Forest y SMOTE, lo que demuestra

una capacidad excepcional de clasificación precisa. FastText: El mejor valor de Precisión fue 0.9496, utilizando Random Forest con SMOTE. El modelo Word2Vec combinado con Random Forest y SMOTE nuevamente fue el que alcanzó el mejor rendimiento, esta vez en términos de Precisión, con un valor de 0.9991.

Embedding	Modelo	Mejores hiperparámetros
BOW	Logistic Reg.	{'logreg__C': 0.01, 'logreg__class_weight': 'balanced', 'logreg__solver': 'lbfgs'}
BOW	SVM	{'svm__C': 100, 'svm__class_weight': None, 'svm__kernel': 'poly'}
BOW	Random Forest	{'rf__class_weight': 'balanced', 'rf__max_depth': 10, 'rf__min_samples_split': 10, 'rf__n_estimators': 50}
BOW	Naive Bayes	{'nb__alpha': 0.1, 'nb__fit_prior': False}
BOW	CatBoost	{'catboost__depth': 6, 'catboost__iterations': 200, 'catboost__learning_rate': 0.05}
TF-IDF	Logistic Reg.	{'logreg__C': 0.01, 'logreg__class_weight': 'balanced', 'logreg__solver': 'liblinear'}
TF-IDF	SVM	{'svm__C': 0.1, 'svm__class_weight': None, 'svm__kernel': 'linear'}
TF-IDF	Random Forest	{'rf__class_weight': 'balanced', 'rf__max_depth': None, 'rf__min_sample 'rf__min_samples_split': 2, 'rf__n_estimators': 200}
TF-IDF	Naive Bayes	{'nb__alpha': 0.5}
TF-IDF	CatBoost	{'catboost__depth': 6, 'catboost__iterations': 100, 'catboost__learning_rate': 0.05}
Word2Vec	Logistic Reg.	{'logreg__C': 100, 'logreg__class_weight': 'balanced', 'logreg__solver': 'liblinear'}
Word2Vec	SVM	{'svm__C': 10, 'svm__class_weight': 'balanced', 'svm__kernel': 'rbf'}
Word2Vec	Random Forest	{'random_forest__max_depth': 10, 'random_forest__min_samples_split': 2, 'random_forest__n_estimators': 50}
Word2Vec	Naive Bayes	{'naive_bayes__var_smoothing': 1e-09}
Word2Vec	CatBoost	{'catboost__depth': 6, 'catboost__iterations': 100, 'catboost__learning_rate': 0.1}
FastText	Logistic Reg.	{'logreg__C': 100, 'logreg__class_weight': 'balanced', 'logreg__solver': 'liblinear'}
FastText	SVM	{'svm__C': 100, 'svm__class_weight': None, 'svm__kernel': 'linear'}
FastText	Random Forest	{'rf__class_weight': 'balanced', 'rf__max_depth': None, 'rf__min_samples_split': 2, 'rf__n_estimators': 500}
FastText	Naive Bayes	{'nb__alpha': 0.5}

FastText	CatBoost	{'catboost__depth': 4, 'catboost__iterations': 100, 'catboost__l2_leaf_reg': 5, 'catboost__learning_rate': 0.1}

Tabla 18: Resumen mejores hiperparámetros

En la Tabla 18 proporciona una visión integral de las configuraciones más efectivas para cada tipo de modelo y *embedding*, lo que es crucial para optimizar el rendimiento de los sistemas de clasificación en tareas de procesamiento de lenguaje natural.

5.2.2 Clasificación de Herramientas Literatura.

Como se puede observar, en la Tabla 19 los resultados más bajos fueron obtenidos por Themis y Genbit. En el caso de Themis, esto puede deberse a la fuerte relación que establece entre el género masculino de las palabras y la existencia de sesgo, lo que provoca que cualquier párrafo que contenga una palabra masculina sea etiquetado como sesgado. Si bien esta estrategia permite identificar algunos párrafos con sesgo, también genera una gran cantidad de falsos positivos (evidenciado por la baja precisión) y no permite detectar párrafos que contienen manifestaciones sutiles de sesgo. Por otra parte, en el caso de Genbit, aunque mejora el *recall*, sigue presentando una gran cantidad de falsos positivos. Considerando que el umbral por defecto para detectar sesgo es alto, esta situación podría deberse a diferencias entre el dominio de los textos utilizados en su entrenamiento y los párrafos evaluados, lo que genera diferencias en el vocabulario y registro utilizados.

Algoritmo	**Binaria (clase sesgo)**		**Ponderada**		**Coef.**
	Precisión	**Recall**	**Precisión**	**Recall**	**Matthews**
GenBit	0.047	0.146	0.929	0.857	0.019
Themis	0.029	0.098	0.926	0.843	-0.017

Tabla 19: Resultados de clasificación de sesgo de genero

Los resultados obtenidos para la clasificación de texto utilizando diferentes algoritmos sobre un conjunto de datos etiquetado con entidades extraídas por el modelo **aymurai/flair-ner-spanish-judicial** (Tabla 20) revelan una variabilidad significativa en el desempeño de los modelos evaluados.

La Regresión Logística mostró el mejor rendimiento global, con un F1-score de 0.2648 y un MCC de 0.2482. Sin embargo, aún existe margen de mejora en precisión (0.3255) y recall (0.2444). Este modelo utilizó una regularización balanceada con C=0.01. El modelo SVM, con el mismo parámetro de regularización y un kernel lineal, presentó un desempeño inferior, con un F1-score de 0.1133 y un MCC de 0.1162. Su bajo recall (0.0722) indica dificultades para diferenciar entre clases. RF tuvo un rendimiento aún más bajo, con un F1-score de 0.0444 y un MCC de 0.0696, lo que sugiere que el modelo no logró aprender adecuadamente las características de los datos. Aunque su precisión fue moderada (0.2000), el recall extremadamente bajo (0.0250) evidencia una clara incapacidad para identificar los ejemplos positivos. Finalmente, el modelo Naïve Bayes superó ligeramente a RF, pero quedó por debajo de la Regresión Logística. Su F1-score fue de 0.1714 y el MCC de 0.1471, reflejando limitaciones similares en la identificación de casos de sesgo.En resumen, para este conjunto de datos, la Regresión Logística resultó el modelo más prometedor, aunque todos los algoritmos presentan desafíos significativos para lograr una alta precisión y *recall*, lo cual podría estar relacionado con la complejidad y el desbalanceo en las clases del conjunto de datos.

Modelo	**MCC**	**F1-score**	**Recall**	**Precision**	**Mejores Hiperparámetros**
Logistic Regression	0.2482	0.2648	0.2444	0.3255	C: 0.01, solver: lbfgs, class_weight: balanced
SVM	0.1162	0.1133	0.0722	0.2667	C: 0.01, kernel: linear, class_weight: balanced
Random Forest	0.0696	0.0444	0.025	0.2	N/A
Naïve Bayes	0.1471	0.1714	0.15	0.2	alpha: 0.1, fit_prior: True

Tabla 20: Resultados aymurai/flair-ner-spanish-judicial

5.2.3 Representaciones de embeddings

El uso de *embeddings* mejoró los resultados previos, particularmente en términos de precisión, aunque el *recall* fue similar, tal como puede observarse en la Tabla 21. Esto indica que las alternativas difieren en la cantidad de falsos positivos que encuentran. Dado que la mayor precisión se obtuvo con las representaciones de

embeddings, es posible inferir que analizar individualmente el vocabulario no es suficiente para detectar sesgo; en su lugar, es necesario realizar un análisis semántico que reconozca expresiones implícitas de sesgo. Se observaron diferencias estadísticamente significativas entre RobertaLex, y BETO, favoreciendo a los primeros.

Los modelos de *embeddings* entrenados específicamente para el idioma español obtuvieron mejores resultados que el modelo multilingual. Si bien estos han mostrado utilidad para tareas que involucran múltiples idiomas, donde es relevante el análisis de la semántica compartida entre ellos, dado que su entrenamiento incluye textos de dominio general, pueden no ser la mejor opción para tareas monolinguales en dominios específicos, como el legal, o en tareas específicas como la detección de discursos de odio (Joshi, Velankar, & Patil, 2022). Adicionalmente, se observó que las diferencias entre RobertaLex y BETO son estadísticamente significativas con un tamaño de efecto pequeño, lo cual muestra la utilidad de contar con *embeddings* entrenados en textos del dominio de la tarea. El clasificador utilizado para estas representaciones es *CatBoost*, un algoritmo de *gradient boosting* conocido por su robustez y capacidad para manejar datos desbalanceados.

Modelo	Binaria (clase sesgo)		Ponderada		Coef.
	Precisión	Recall	Precisión	Recall	Matthews
multilingual	1.0	0.375	0.977	0.977	0.605
Beto	0.75	0.375	0.968	0.973	0.519
RobertaLex	0.8	0.5	0.975	0.977	0.622

Tabla 21: Representación Embeddings

5.3 Clasificación con LLMs (P2-1)

En la Tabla 22 se presentan los resultados de las alternativas de clasificación *zero-shot* (zero) y *few-shot* (estático y dinámico) para los LLMs evaluados. Para el caso *few-shot-dinámico*, se evaluaron diversas configuraciones modificando los *embeddings* empleados para la representación semántica, incluyendo BETO, robertalex, SentenceTransformer y bert multilingual, de los cuales BETO alcanzó los mejores resultados. En cuanto a la selección de ejemplos para el enfoque *few-shot*,

se consideraron configuraciones tanto balanceadas entre las dos clases como no balanceadas. Finalmente, tras comparar las métricas coseno y euclidiana en el contexto de BETO, se decidió utilizar la métrica euclidiana, dado su mejor rendimiento en los experimentos, lo que se refleja en las métricas reportadas. Teniendo en cuenta este análisis, se reportan los resultados seleccionando los vecinos utilizando BETO y la métrica euclideana, dado que fue la que obtuvo el mejor desempeño.

Modelo/Configuración	Binaria (clase sesgo)		Ponderada		Coef.
	Precisión	Recall	Precisión	Recall	Matthews
GPT-zero	0.077	0.78	0.953	0.644	0.164
Llama2-zero	0.038	0.923	0.92	0.098	-0.009
Mistral-zero	0.064	0.512	0.938	0.692	0.088
gema-zero	0.055	0.8	0.949	0.479	0.102
GPT-estático	0.084	0.903	0.959	0.589	0.191
Llama2-estatico	0.075	0.307	0.933	0.839	0.085
Mistral-estático	0.073	0.773	0.952	0.614	0.149
Gema-estático	0.037	1	0.001	0.037	0
gpt-dinamico	0.093	0.732	0.952	0.722	0.189
Llama2-dinamico	0.078	0.601	0.927	0.577	0.087
Mistral-dinámico	0.107	0.615	0.941	0.765	0.181
Gema-dinámico	0.141	0.725	0.943	0.779	0.251

Tabla 22: Resultados de alternativas de clasificación

En lo que respecta al *zero-shot*, los mejores resultados balanceando precisión y *recall* fueron obtenidos por *GPT*, para el cual se observaron diferencias estadísticamente significativas respecto de *Gemma* y *Llama2*. Si bien *Llama2* alcanzó el *recall* más alto, lo obtuvo a expensas de determinar que casi todos los párrafos contenían sesgo (mostrando así una alta sensibilidad al lenguaje (Zhang, He, Ji, & Lu, 2024) lo que genero muchos falsos positivos, y en consecuencia el peor valor del MCC.

Comparando *zero-shot* con *few-shot-estático*, excepto para *Gemma*, las alternativas *few-shot* obtuvieron un mejor valor para el MCC. En el caso de *Llama2*, si bien el *few-shot* disminuyó el *recall* para la clase positiva en 66%, también introdujo una mejora en precisión de un 100%. Para *Gemma*, el incremento en *recall* fue acompañado por una disminución en la precisión, dado que todos los textos fueron clasificados como sesgo. Estos resultados permiten inferir que la definición de un

contexto para la tarea, es decir, agregar ejemplos relacionados, puede ayudar al LLM a realizar una mejor clasificación.

En lo que respecta a las alternativas de *few-shot*, la opción *dinámica* mejoró el balance entre precisión y *recall* para todos los LLMs. Excepto para *Llama2*, se observan valores de recall más bajos, con una disminución promedio del 22%, pero se logró mejorar la precisión en todos los casos, con un aumento promedio del 84%. Este mejor balance en la clasificación se ve acompañado por mejoras en el MCC. Para *GPT* y *Gemma*, las diferencias respecto de la variante estática fueron significativas. Asimismo, todos los modelos mostraron diferencias significativas respecto a las alternativas *zero-shot* correspondientes. En cuanto a las métricas ponderadas, en promedio, el *recall* se incrementó en un 143%. Aunque la disminución en el *recall* de la clase positiva implica que se deja sin identificar una mayor cantidad de textos, el incremento en precisión reduce los falsos positivos, lo que potencialmente permite disminuir los esfuerzos en una tarea manual posterior.

En general, para el método few-shot, se observaron mejoras promedio en precisión del 87% (mínimo de 19% para *GPT* y máximo de 155% para *Gemma*) y en el MCC del 178% (mínimo de 15% para *GPT* y máximo de 845% para *Gemma*). Asimismo, para *F-measure*, se observaron mejoras promedio del 70% (mínimo de 16% para *GPT* y máximo de 128% para *Gemma*). Estos resultados muestran el efecto diferenciado que el contexto de la tarea tiene sobre cada LLM, siendo importante no solo agregar ejemplos, sino también cómo se definen, ya que elegir los textos más cercanos a clasificar permitió mejorar la precisión, el balance entre precisión y *recall*, y el MCC respecto a elegirlos de forma arbitraria.

Mientras que la clasificación tradicional tendió a mejorar la precisión a expensas de valores de *recall* más bajos, los LLMs tendieron a producir valores de *recall* más altos a expensas de numerosos falsos positivos. Las diferencias observadas fueron estadísticamente significativas a favor de los modelos tradicionales. Estos resultados implican que, aunque los LLMs poseen un buen potencial para las tareas de análisis y clasificación de texto, todavía no logran superar a modelos de *embeddings* (como BERT y derivados) en dominios específicos. Este

hallazgo coincide con reportes recientes de la literatura (Qiu & Jin, 2024) (Chuang & Tang, 2023).

Finalmente, en relación con la valoración cualitativa de las respuestas en términos de cuan fiel fue su formato respecto a lo que se solicitaba en el *prompt*, los modelos más responsivos fueron *GPT* y *Gemma*. Por el contrario, *Mistral* y *Llama2* a menudo devolvieron respuestas cambiando los nombres de las claves solicitadas en el formato JSON, y *Llama2* alternaba entre español e inglés, entre otras incidencias.

5.4 Explicaciones provistas por los LLMs (P1-3)

Mas allá de la capacidad de clasificación de textos, los LLMs pueden generar explicaciones textuales que acompañan el resultado de su clasificación. De los cuatro LLMs analizados, únicamente GPT generó respuestas coherentes y relativas al contexto del párrafo bajo clasificación para todos los casos. En Gema y Mistral solo algunas de las explicaciones resultaron coherentes, mientras que para Llama2 la mayoría de las respuestas contenían repeticiones y alucinaciones.

La Figure 29 presenta la separación en tópicos para las explicaciones dadas por *GPT*, considerando aquellos párrafos que fueron correctamente clasificados. Para este análisis se utilizó BETO como representación principal. Se observa que la mayoría de las explicaciones del LLM respecto a la clase positiva aparecen agrupadas (tópico 2), y se separan del resto de los tópicos con explicaciones relacionadas a la clase negativa. Asimismo, en la figura Figure 28 se puede ver la representación de *Llama2*, donde se identifican varias agrupaciones, cada una representando un conjunto de párrafos relacionados temáticamente. Los tópicos más destacados incluyen palabras clave como "*casación, sentencia, adhiero, recurso*" (tópico 2), que sugiere un foco en procedimientos judiciales, y "*imputado, víctima, hechos, homicidio*" (tópico 1), relacionado con casos criminales. Otros tópicos se centran en términos anatómicos (como "*cubital, subclaviculares, maxilar*" en el tópico 6), mostrando la variedad de temas abordados en las sentencias analizadas. La dispersión y densidad

de puntos refleja la diversidad y frecuencia de aparición de estos tópicos en los documentos procesados.

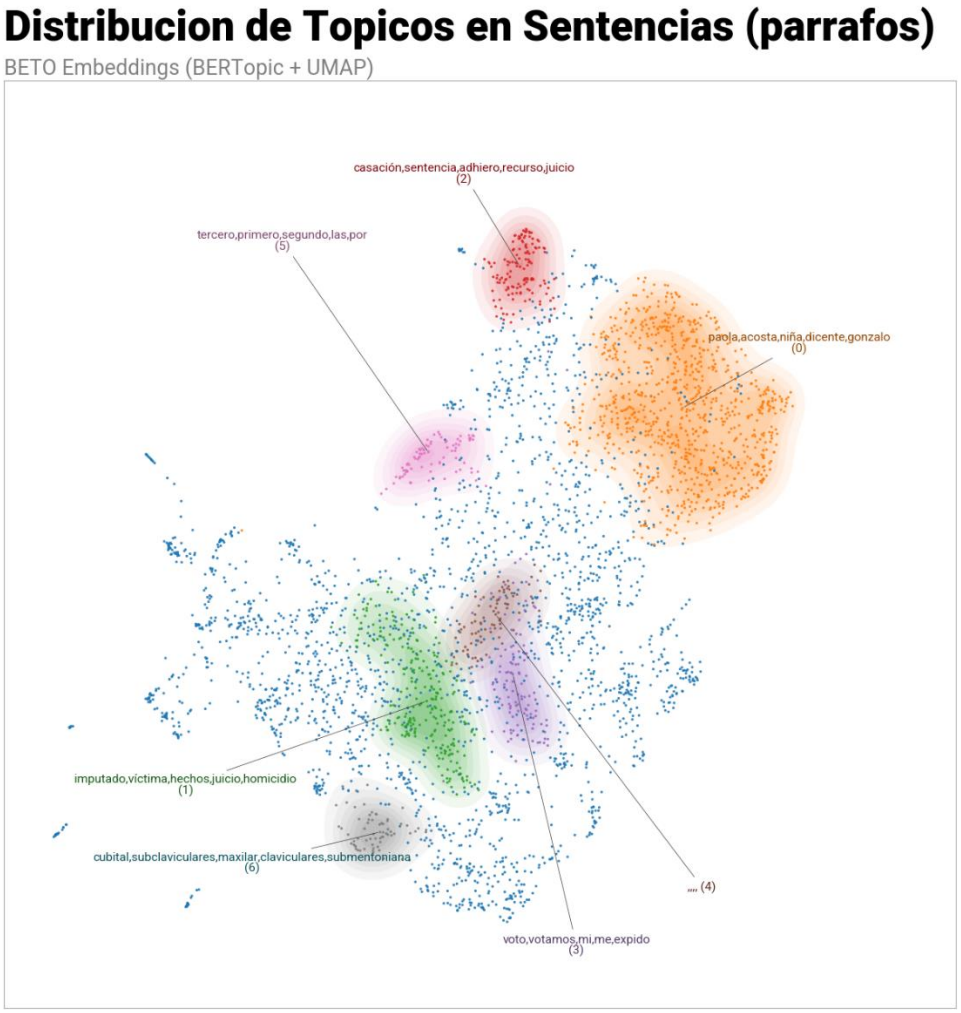

Figure 28: Resultados Llama2 Bertopic Proyección UMAP sobre embeddings BETO

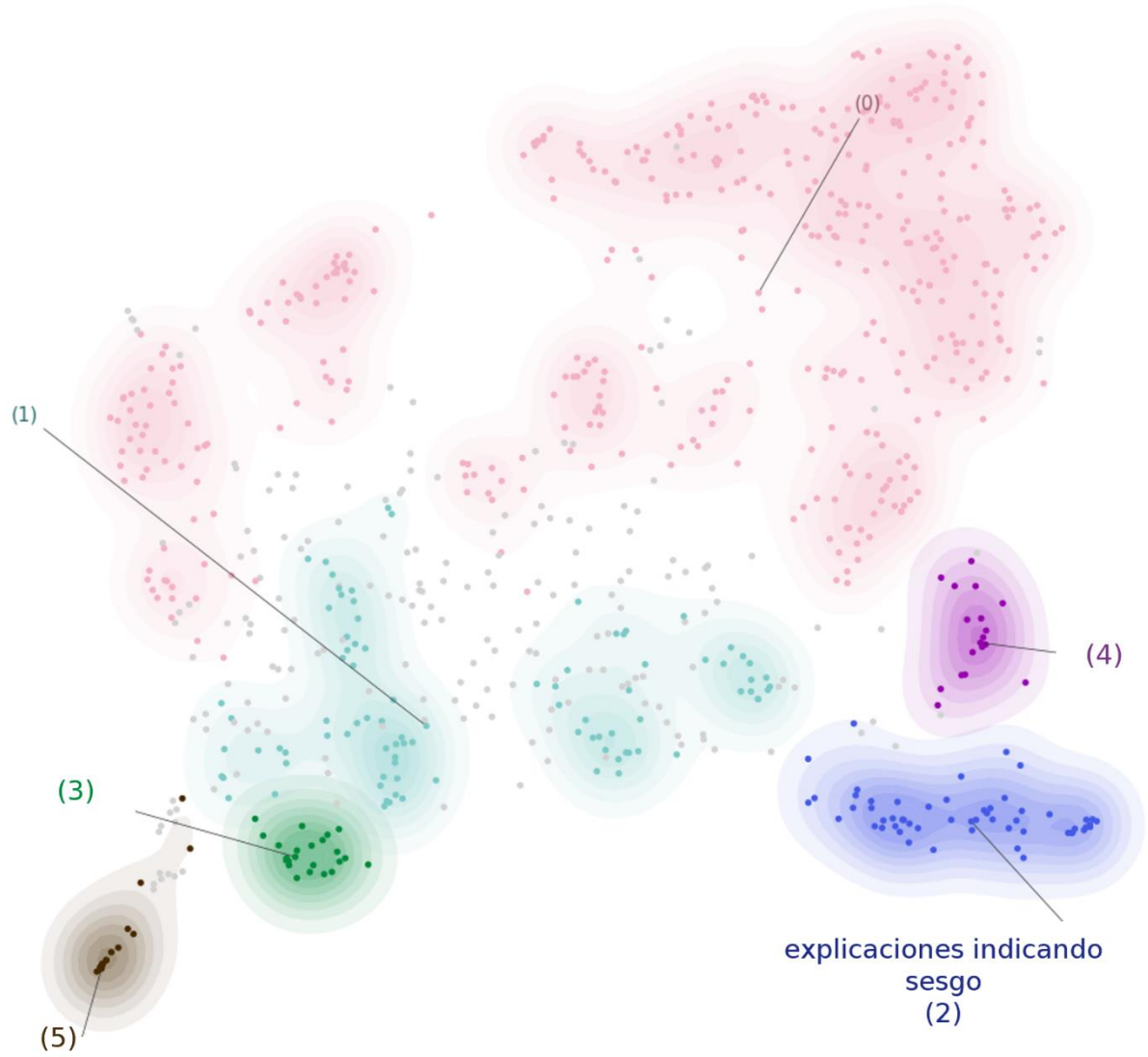

Figure 29Analisis de tópicos generado por BERTopic para las explicaciones. Proyección UMAP sobre embeddings BETO

Se realizó una inspección manual de las explicaciones de GPT. Un ejemplo de explicación de sesgo es "*El texto presenta un sesgo de genero al cuestionar la credibilidad y la actitud de la menor que denuncia abuso sexual, basándose en estereotipos y prejuicios sobre cómo debería comportarse una víctima. Se enfoca en desacreditar el testimonio de la niña, minimizando su relato y destacando aspectos irrelevantes como su rendimiento académico o la falta de detalles emocionales, lo cual refleja una visión sesgada y poco empática hacia las víctimas de abuso*.", lo cual se condice con lo observado en el párrafo clasificado. Un ejemplo de explicación para no sesgo es *"El texto presenta argumentos legales y razonamientos sobre la falta de fundamentación en un fallo judicial, sin hacer referencia a cuestiones de género o discriminación. Se centra en la evaluación de pruebas y la aplicación correcta del*

derecho vigente en el caso.", donde se observa un planteo neutral, o bien "*El texto hace referencia a la jurisprudencia interamericana en casos de violencia sexual, destacando la importancia de la declaración de la víctima como prueba fundamental, así como la consideración de las imprecisiones en dichas declaraciones debido al impacto traumático de los hechos. No se observa un sesgo de género en la argumentación presentada, sino más bien una consideración objetiva de la complejidad de los casos de violencia sexual.*", donde se expone la necesidad de un análisis cuidadoso de las declaraciones de las víctimas.

En general, para las explicaciones agrupadas en el **tópico 2** (Figure 29), se observaron referencias a situaciones de violencia, abuso y manipulación, donde el sesgo se vincula al cuestionamiento a la veracidad de la víctima (mujer), minimizando la gravedad de la situación sufrida, y reflejando un cierto prejuicio hacia las mujeres en casos de abuso sexual, mediante la desacreditación de los testimonios. Este análisis cualitativo muestra, en principio, que las explicaciones generadas por *GPT* son razonables para las etiquetas asignadas durante la clasificación. No obstante, debe realizarse un análisis más extensivo (y sobre un conjunto de sentencias mayor) para incrementar la confianza sobre la calidad de las explicaciones de los LLMs.

5.5 Resumen a las preguntas planteadas

La siguiente tabla presenta un resumen de las principales preguntas de investigación (PI) abordadas en el estudio, junto con los resultados obtenidos con relación a la clasificación de textos mediante representaciones tradicionales, *embeddings* y modelos de lenguaje de gran escala (LLMs).

Pregunta de Investigación (PI)	Descripción	Resumen de Resultados
PI-1: ¿Qué representación de texto es más efectiva para una clasificación tradicional?	**Representaciones Textuales**: Se evaluaron diferentes representaciones textuales tradicionales.	No se observó un modelo con una diferencia significativa sobre los demás, lo que llevó a explorar otras herramientas.

	Representaciones de Embeddings: Se evaluaron modelos basados en embeddings como BETO y RobertaLex.	BETO y RobertaLex fueron efectivos, con una leve ventaja para RobertaLex.
PI-2: ¿Es la clasificación basada en LLMs comparable a la clasificación tradicional? ¿Hay diferencias en los resultados de los distintos LLMs?	Comparación de la clasificación con LLMs y métodos tradicionales.	Los métodos tradicionales obtuvieron un mejor desempeño que los LLMs, aunque no todas las diferencias fueron estadísticamente significativas.
PI-3: ¿Pueden los LLMs proveer explicaciones coherentes y entendibles por humanos respecto a la presencia de sesgo?	Evaluación de la capacidad de los LLMs para generar explicaciones comprensibles sobre sesgo.	Algunos LLMs, como GPT, pueden generar explicaciones satisfactorias y comprensibles para humanos sobre la existencia de sesgo.

Tabla 23: Resumen preguntas

Capítulo 6

Conclusiones

En este capítulo final, se presenta un resumen general del trabajo realizado, seguido de un análisis de las principales contribuciones y, finalmente, una proyección hacia futuras líneas de investigación. Se ofrece una síntesis que integra los resultados obtenidos y se relaciona nuestra propuesta con los trabajos previos en la literatura.

6.1 Contribuciones Principales

Este trabajo ha demostrado la aplicabilidad del procesamiento del lenguaje natural (PLN) y modelos de aprendizaje automático para detectar y analizar sesgos en sentencias judiciales. Al proponer un enfoque que considera características demográficas y contextuales de los autores de las sentencias, se logró una perspectiva más rica sobre cómo los sesgos pueden manifestarse de manera sutil pero sistemática en el ámbito legal. La integración de técnicas de análisis de tópicos, reducción de dimensionalidad y clasificación tradicional permitió no solo evaluar la presencia de sesgos, sino también aportar evidencia sobre la utilidad de estos métodos en entornos reales de justicia.

Las principales contribuciones del estudio incluyen el diseño de un pipeline de PLN adaptable para analizar textos legales y la identificación de patrones específicos que sugieren la existencia de sesgos en función de variables como el género y la procedencia geográfica de los jueces. Además, este trabajo resalta la importancia de considerar factores temporales y socioculturales, lo que abre nuevas vías para el análisis longitudinal de los sesgos en textos legales.

Si bien el tamaño del corpus y el desbalance en los datos limitan la capacidad de generalización de los resultados, este estudio establece una base sólida para futuras investigaciones que busquen una mejor comprensión del impacto de los sesgos en el proceso de toma de decisiones judiciales. La combinación de enfoques cuantitativos y cualitativos en el análisis de sentencias ofrece un marco robusto que puede ser perfeccionado para abordar los desafíos inherentes a la detección de sesgos en dominios legales.

6.2 Trabajos Futuros

Existen múltiples áreas en las que este trabajo puede ser mejorado y extendido, y estas se detallan en la presente sección.

Entre las limitaciones a abordar en estudios futuros se destaca el tamaño reducido del conjunto de sentencias utilizado en los experimentos, lo que limita la capacidad de generalización de los resultados. Sería necesario disponer de un corpus más amplio que incluya sentencias de diferentes provincias y épocas, lo que permitiría evaluar si las manifestaciones de sesgo varían en función de los contextos socioculturales y, además, analizar posibles cambios en las tendencias a lo largo del tiempo.

Dado el desbalance en el corpus utilizado, sería beneficioso explorar técnicas de aumentación de datos, que generen instancias sintéticas para la clase menos representada e incorporen variantes léxicas y sintácticas. Asimismo, se propone refinar la definición de sesgos mediante la operacionalización de estereotipos concretos[45] (Cook & Cusack, 2010), lo que podría facilitar un análisis más detallado de cada caso.

[45] 21Algunos gobiernos ya han comenzado a trabajar con estereotipos, por ejemplo: https://www.gub.uy/sites/gubuy/files/inline-files/Gu%C3%ADa%20para%20el%20Poder%20Judicial.pdf

Se podrían considerar documentos escritos antes y después de 2014/2015, años que marcan la modificación del Código Civil Argentino con la incorporación de perspectiva de género y el surgimiento del movimiento 'Ni una menos' en Argentina, con el fin de analizar el impacto de dichos eventos. Además, sería relevante explorar la capacidad de generalización de las técnicas y modelos generados para el análisis de textos de otros países latinoamericanos, como Bolivia, Chile y Colombia que, aunque escritos en español, pueden presentar variantes lingüísticas y culturales."

Un enfoque novedoso para explorar en investigaciones futuras sería el estudio de técnicas de regeneración de párrafos o sentencias con sesgo. La idea sería no solo identificar los fragmentos sesgados, sino también utilizar modelos generativos para reescribir dichos fragmentos eliminando el sesgo o modificando su redacción para que sea más neutral. Esto permitiría no solo detectar sesgos, sino también proponer soluciones automáticas y evaluar cómo afectan las correcciones en la interpretación general del texto. Este tipo de abordaje podría contribuir significativamente a mejorar la calidad de los documentos legales y reducir la influencia de estereotipos en la redacción del mismo.

Considerando las limitaciones de los LLMs, cuyo conocimiento es estático y generalmente enfocado en dominios amplios, sería valioso explorar enfoques como *fine-tuning* (Ovadia, Brief, Mishaeli, & Elisha) o el esquema RAG (*Retrieval Augmented Generation*) (Lewis, Perez, Piktus, & Petroni, 2020) que permitan incorporar conocimiento más específico del dominio legal.

6.3 Consideraciones Éticas

Dado que este trabajo trata temas sensibles relacionados con el género, se ha optado por compartir públicamente[46] los *embeddings*, el código y los modelos desarrollados, con el objetivo de fomentar la transparencia metodológica y facilitar la

[46] bonzokimba/deteccionsesgo: Detección de sesgos debido al género en decisiones judiciales utilizando PNL

reproducibilidad de los resultados. No obstante, con el fin de preservar la identidad y la integridad de las partes involucradas, el conjunto de sentencias utilizado no se publicará de manera abierta. Los análisis realizados no deben ser utilizados para identificar ni estigmatizar a individuos, ni para introducir, aunque sea de manera inadvertida, sesgos en los documentos. En cambio, deben ser entendidos como una oportunidad para fomentar un diálogo constructivo sobre el sesgo y sus consecuencias.

Es importante considerar también la existencia de posibles sesgos introducidos en el análisis, por ejemplo, en el proceso de etiquetado de datos o en el estudio de resultados. Asimismo, el uso de LLMs en el dominio legal puede impactar sobre aspectos de privacidad de los datos, confidencialidad y responsabilidad. Los LLMs podrían incluso ser una nueva fuente de sesgos de género en las respuestas generadas, por ejemplo, al reforzar estereotipos de ocupaciones por genero tomadas de los conjuntos de datos con los que fueron entrenados (Salinas, Shah, Huang, McCormack, & Morstatter, 2023) (Kotek, Dockum, & Sun, 2023)

Referencias

A. Deshpande, V. M. (s.f.). Toxicity in chatgpt: Analyzing personaassigned language models,. *arXiv:2304.05335*, 2023.

al., A. S. (2016). Gender stereotyping in the case law of the eu court of justice. *European Equality Law Review, no. 1*, 37-46.

B. Pecher, I. S. (2024). Automatic combination of sample selection strategies for few-shot learning. *arXiv:2402.03038*.

Blei, D., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*.

Bojanowski, P., Grave, E., Joulin, A., & T., M. (2017). Enriching Word Vectors with Subword Information. 12.

Bonina, N. (2020). *Inteligencia artificial y derecho: ¿las máquinas van a reemplazar a los abogados?* Thomson Reuters La Ley, AR/DOC/3809/2020.

Bordia, S., & Bowman, S. (2019). Identifying and reducing gender bias in word-level language models. *NAACL*.

Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. *arXiv:1904.03035*.

Breiman, L. (2001). Random Forests.

Brown, T., Mann, B., & et, a. (2020). Language Models are Few-Shot Learners.

Calderon Suarez, R., Ortega Mendoza, R., Montes, M., Toxqui Quitl, C., & Marquez Vera, M. (2023). Enhancing the Detection of Misogynistic Content in social media by Transferring Knowledge From Song Phrases.

Cañete J, C. G. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR 2020*.

Cañete, J., Chaperon, G., Fuentes, R., & Ho, J.-H. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR 202*.

Chang, K., Xu, S., & Luo, Y. (2024). Efficient Prompting Methods for Large Language Models: A Survey. *arxiv.org*.

Chen, H., Xu, A., Lui, Z., & Guo, Y. (2020). A General Methodology to Quantify Biases in Natural Language Data.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics, vol. 21, no. 1*, 1-13.

Chuang, Y., & Tang, R. (2023). Spec:A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. *arXiv preprint*.

Chuang, Y.-N., Tang, R., Jiang, X., & Hu, X. (2023). Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization,. *arXiv preprint arXiv:2303.13035*.

Cook, R. J., & Cusack, S. (2010). Gender Stereotyping: Transnational Legal Perspectives. *University of Pennsylvania Press.*

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., & Zhao, B. (2020). Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. 11.

Cusack, R. J. (2010). *ESTEREOTIPOS DEGÉNERO ,Perspectivas Legales Transnacionales.* Pennsylvania : University of Pennsylvania Press.

Dacon, J., & Liu, H. (2021). Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles.
Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 16.
Elliott, A., Chen, A., & Ornaghi, A. (2020). Stereotypes in High-Stakes Decisions: Evidence from U.S. Circuit Courts.
Fagbohun, O., Harrison, R., & Dereventsov, A. (2024). An Empirical Categorization of Prompting Techniques for Large Language Models: A Practitioner's Guide. *https://arxiv.org/*.
Ferrer, X., van Nuenen, T., Such, J., & Criado, N. (2021). Discovering and Categorising Language Biases in Reddit.
Frenda, S., Ghanem, B., Montes y Gomez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter.
García Díaz, J., Jiménez Zafra, S., García Cumbreras, M., & Valencia García, R. (2022). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers.
Gillis, N. (August de 2021). *Sexism in the judiciary: The importance of bias definition in NLP and in our courts.* Obtenido de Association for Computational Linguistics: https://aclanthology.org/2021.gebnlp-1.6/
González Gavaldón, B. (1999). Los estereotipos como factor de socialización en el género. *Coomunicar*.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol Rev*.
J. Liu, D. S. (May 2022). What makes good in-context examples for GPT-3?" in Proceedings of Deep Learning Inside Out (DeeLIO 2022). *Dublin, Ireland and Online: ACL*, 100–114.
J. Zamfirescu-Pereira, R. Y. (2023). Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. *in Proceedings of the 2023 CHI Conference*, 1-21.
Joshi, R., Velankar, H., & Patil, H. (2022). Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. *Artificial Neural Networks in Pattern Recognition*, págs. 121-128.
Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. *in Proceedings of The ACM Collective Intelligence Conference*, 12-24.
L. Prokhorenkova, G. G. (2018). Catboost: unbiased boosting with categorical features.
Lewis, P., Perez, E., Piktus, A., & Petroni, F. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems,*, 9459–9474.
Lippmann, W. (1922). Public Opinion. Adansonia Publishing.
Liu, Y., Ott, M., Goyal, N., & Du, J. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.
Mackie, M. M. (1973). Arriving at Truth by Definition Case of Stereotype Innacuracy. *Social Problems, 20*, 431-447.
Melucci, M., & Baeza-Yates, R. (2011). Advanced topics in information retrieval. *Springer Science & Business Media*, Vol. 33.
Mikolov, T., G., C., K., C., & J., D. (2013). Efficient Estimation of Word Representations in Vector Space. 12.
Miller, A. (2018). Gender roles highlight gender bias in judicial decisions.
Molina, M. L. (2022). Legal Decision-Making and Implicit Bias.Brief, 52(1).
Nishtha, M., Sameep, M., Taneea, S., & Y. G. (2018). Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies.

O. Ovadia, M. B. (2023). Finetuning or retrieval? comparing knowledge injection in llms. *arXiv:2312.05934*.

Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (s.f.). Finetuning or retrieval? comparing knowledge injection in llms. *arXiv:2312.05934*, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., & et. Al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Peters, M., Neumann, M., & Iyyer, M. :. (2018). Deep contextualized word representations. 15.

Pezeshkpour P., Z. Z. (2020). On the utility of active instance selection for few-shot learning. *NeurIPS HAMLETS*.

Plaza del Arco, F., Molina Gonzalez, M., Ureña Lopez, L., & Valdivia, M. (2020). Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies.

Prokhorenkova, L., Gusev, G., & Vorobev, A. (2019). CatBoost: unbiased boosting with categorical features. *https://arxiv.org/abs/1706.09516*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. Vol. 31.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. vol. 31.

Qiu, Y., & Jin, Y. (2024). Chatgpt and fine tuned bert: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, pág. 200308.

Qiu, Y., & Jin, Y. (2024). Chatgpt and finetuned bert: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications, vol. 21*, 200308.

Ratovicius, C., Diaz-Pace, J. A., & Tommasel, A. (2024). Detección de sesgo de género en textos legales. *CLEI*, 10.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

S. Huang, S. M. (2023). Can large language models explain themselves? a study of llmgenerated self-explanations. *arXiv:2310.11207*.

Salinas, A., Shah, P., Huang, Y., McCormack, R., & Morstatter, F. (2023). The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama.

Schmeisser-Nieto, W., Montserrat, N., & Mariona, T. (2022). Criteria for the Annotation of Implicit Stereotypes. *13th Conference on Language Resources and Evaluation*.

Sevim, N., Sahinuç, F., & Koç, A. (2022). Gender bias in legal corpora and debiasing it.

Stanczak, K., & Augestein, I. (2021). A Survey on Gender Bias in Natural Language Processing.

T. Brown, B. M. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901.

Tenghao, H., Faeze, B., Vered, S., & Snigdha, C. (2021). Uncovering Implicit Gender Bias in Narratives through Commonsense Inference. *EMNLP 2021*, págs. 3866–3873.

Timmer, A. S. (2016). Gender stereotyping in the case law of the eu court of justice. *European Equality Law Review*, págs. 37–46.

Vapnik, V., & Cortes, C. (1995). Support-vector networks.

Vaswani, A., Shazeer, N., Parmar, N., & Uszkoreit, J. (2017). Attention Is All You Need. 11.

Velankar, A., Patil, H., & Joshi, R. (2022). Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. *in Artificial Neural Networks in Pattern Recognition*, 121–128.

W. Yu, D. I. (2022). Generate rather than retrieve: Large language models are strong context generators. *arXiv:2209.10063*.

Zhang, M., He, J., Ji, T., & Lu, C.-T. (s.f.).

Zhang, M., He, J., Ji, T., & Lu, C.-T. (2024). Don’t go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection,. *arXiv preprint arXiv:2402.11406*.

Zhang, M., He, J., Ji, T., & Lu, C.-T. (2024). Don’t go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. *arXiv preprint arXiv:2402.11406*.