Detection of Gender Bias in Legal Texts Using Classification and LLMs

Christian Javier Ratovicius
Universidad Abierta Interamericana
Facultad de Tecnología Informática
Centro de Altos Estudios en Tecnología Informática
Buenos Aires, Argentina
cratovicius@gmail.com

J. Andrés Diaz-Pace
ISISTAN, CONICET-UNICEN
Tandil, Buenos Aires, Argentina
andres.diazpace@isistan.unicen.edu.ar

Antonela Tommasel
ISISTAN, CONICET-UNICEN
Tandil, Buenos Aires, Argentina
antonela.tommasel@isistan.unicen.edu.ar

***Abstract*—Gender bias is a common and often neglected issue in legal documents. It arises from perceptions or prejudices about the characteristics of a group, or the roles individuals should play in society. This bias can significantly impact the reasoning or outcomes of legal processes, such as judicial rulings. To ensure equal treatment for all individuals, it is crucial to effectively reduce this bias. The first step to reduce bias is to define approaches that can identify manifestations of gender bias. However, these manifestations are not usually easily detectable in text (e.g., through keywords) as they often require detailed contextual analysis, typically done manually by experts. This paper addresses this issue by leveraging natural language processing and machine learning techniques to automate parts of the analysis. Specifically, it proposes a processing pipeline based on text embeddings, binary classification, and the use of large language models (LLMs) to explain classification results. An initial evaluation on a set of judicial rulings shows promising results in terms of precision and recall, along with qualitative insights into the potential of these techniques in the legal domain.**

***Index Terms*—Natural Language Processing, Text Classification, Judicial Rulings, Gender bias, LLMs.**

I. Introducción

El sesgo puede definirse como una predisposición o inclinación hacia ciertos puntos de vista, opiniones o acciones, generalmente basada en percepciones preconcebidas o experiencias pasadas. Una de las formas en que se manifiesta el sesgo es a través de estereotipos. Un estereotipo refiere a una idea o imagen generalizada (y, probablemente, simplista) de un grupo específico de personas, y se relaciona con la expectativa que se puede tener sobre todos los individuos que pertenecen a un grupo en particular o al comportamiento que se espera que adopten [1], independientemente de la situación real de dicho grupo. En particular, los estereotipos de género abarcan creencias, actitudes o expectativas de cómo deben verse, pensar o comportarse las personas en relación con su género [2]. Estos estereotipos alimentan el sesgo al proporcionar una base para actitudes y percepciones discriminatorias.

Si bien la perspectiva de género no es un concepto nuevo (su uso data de los 70' en la Organización de las Naciones Unidas[1]), no fue hasta los últimos años que cobró importancia en la administración pública, incluyendo al Poder Judicial, sus procesos y actores. En Latinoamérica existen iniciativas relacionadas en Chile, Colombia, Uruguay, México y Argentina. Un ejemplo en Argentina para el abordaje de esta problemática es la capacitación conocida como "Ley Micaela"[2].

Las relaciones de poder juegan un rol fundamental en la creación del Derecho, y esto puede traer aparejado que el derecho que se establece como neutral promueva una posición dominante, pudiendo así (inadvertidamente o no) incluir y perpetuar sesgos (y consiguientes estereotipos) en razón de género y propiciar un trato desigual [3]. Es por esto que es deseable que en los tribunales, así como en otros ambientes normativos, se incorporen estrategias y mecanismos para detectar situaciones de sesgo de género en pos de un trato igualitario [1]. Varios trabajos han analizado cómo los estereotipos afectan los razonamientos judiciales (por ejemplo, [3, 4]). Sin embargo, dado que estos análisis normalmente se realizan en forma manual por expertos y sobre un conjunto reducido de documentos, resulta difícil escalar y generalizar estas observaciones y guías a otros documentos. En este trabajo, se argumenta que el análisis puede mejorarse mediante el procesamiento semi-automatizado de documentos textuales, en base a técnicas de *Machine Learning* (ML) y *Natural Language Processing* (NLP). Es importante destacar que estas técnicas no tienen como objetivo reemplazar el factor humano en la toma de decisiones judiciales o en la creación de un sistema propio de resolución, sino que cumplen un rol de asistencia a la labor humana [5].

El objetivo de de este trabajo es explorar el uso de NLP, ML supervisado, y avances en *Large Language Models* (LLMs) para detectar sesgos en sentencias judiciales. La hipótesis de trabajo es que ciertos estereotipos o sesgos (en particular los relacionados al género) pueden identificarse a partir de los contextos gramaticales, sintácticos y semánticos de los párrafos que componen un documento. Estos sesgos no solo se manifiestan en el lenguaje de forma explícita, sino también

[1]https://www.un.org/es/global-issues/gender-equality
[2]https://www.argentina.gob.ar/generos/ley-micaela

979-8-3315-4097-5/24/$31.00 ©2024 IEEE

de forma implícita mediante percepciones subconscientes que no se expresan de forma directa en el lenguaje compartido [6], debiendo ser interpretados en función del conocimiento del contexto y cultura donde se expresa [7]. 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 [8].

Para abordar la problemática, se propone una cadena de procesamiento (*pipeline*) que parte de representar el texto con *embeddings*, luego aplicar técnicas de clasificación binaria, y por último generar explicaciones textuales para dicha clasificación. Para evaluar experimentalmente el *pipeline* propuesto, se utilizó un conjunto de sentencias judiciales con anotaciones respecto a la existencia (o no) de sesgo de género en sus párrafos. Con este aporte, se espera contribuir al desarrollo de herramientas que permitan analizar indicadores de sesgo en documentos legales, y también lograr una mayor concientización del problema en la formación del personal judicial.

II. Trabajos Relacionados

En la literatura existen trabajos que han abordado la problemática de sesgo de género en idioma inglés [9, 10, 11]. Estos trabajos se enfocan mayoritariamente en la definición de léxicos que describen profesiones, pronombres, comportamiento y aspectos sociales que distinguen a los géneros (usualmente, masculino y femenino) y la utilización de representaciones semánticas (por ejemplo, *embeddings*). Sobre esta base, se calculan métricas o se aplican técnicas de ML no supervisado (por ejemplo, *clustering*) en función de las representaciones de los términos en los léxicos para determinar así los espacios de los géneros y expresiones que recurrentemente aparezcan más cerca de un género que del otro. Si bien estas técnicas han reportado una cierta efectividad para el inglés, un idioma que no tiene género, no son directamente aplicables al español, que sí lo tiene.

Dentro del idioma español, algunos trabajos han analizado manualmente tanto textos de leyes como sentencias judiciales [12, 13]. Existen trabajos que abordan tareas similares, como la detección de misoginia en redes sociales basados principalmente en entrenar clasificadores con características léxicas o *embeddings*. Sin embargo, considerando las diferencias de medio, estilo y vocabulario, no hay garantías de que dichas técnicas puedan obtener buenos resultados respecto al sesgo de género en el dominio legal.

En lo que respecta a modelos de lenguaje, se destacan *Legal-BETO*[3][4] y *Flair-NER-spanish-Judicial*[5]. El primer modelo permite generar representaciones de *embeddings*. Fue entrenado con un corpus legal y administrativo de España, lo que no necesariamente posee correspondencia con el contexto cultural y legal de Argentina. El segundo modelo es para la detección de entidades en sentencias judiciales y fue entrenado con causas de un único juzgado de Argentina. Si bien algunas entidades se encuentran relacionadas con la temática, es preciso evaluar su pertinencia para la clasificación de sentencias. Finalmente, la herramienta comercial Themis[6] se publicita como el primer corrector de lenguaje inclusivo, apuntando a detectar expresiones con sesgo y sugerir expresiones alternativas.

En los últimos años, los LLMs [14] han transformado el estado del arte en tareas de NLP mediante su pre-entrenamiento en grandes conjuntos de datos y su posterior *fine-tuning* de acuerdo a instrucciones humanas. En este sentido, los LLMs tienen un gran potencial para varios usos en el dominio legal, ya que posibilitarían aumentar las capacidades de los profesionales del derecho y mejorar la eficiencia de diversos procesos. Por ejemplo, Kang et al. [15] y Tan et al. [16] estudiaron la capacidad de LLMs para analizar casos legales.

III. Enfoque metodológico

Uno de los desafíos para identificar sesgos de género en textos legales es que estos no son siempre explícitos, sino que pueden manifestarse en el uso de ciertas construcciones que referencian estereotipos en contextos particulares, y que pueden llevar luego a deducciones o conclusiones sesgadas. Un ejemplo explícito puede ser "*La sentencia considera que la víctima, al ser mujer, probablemente exageró la gravedad del incidente debido a su naturaleza emocional y propensión a dramatizar situaciones*", donde un analizador podría detectar el patrón (de sesgo) en función del uso de palabras claves. Un ejemplo menos obvio es "*Se presume que la madre tendrá una mayor capacidad para cuidar y proveer un entorno estable para los hijos, por lo que se le otorga la custodia principal en el caso de divorcio*", ya que asume que por el hecho de ser madres, las mujeres son inherentemente mejores cuidadoras y proveedoras de un entorno estable para los hijos, atribuyendo roles y responsabilidades basados en el género, en lugar de evaluar las capacidades parentales de cada individuo de manera objetiva e imparcial. En este segundo ejemplo, la tarea para un analizador basado en NLP es más compleja.

El procesamiento de textos legales y la identificación de patrones relacionados con sesgo requieren un análisis del contexto en que aparecen ciertas palabras. Asimismo, se observa que la cantidad de texto que puede aludir a un sesgo de género en un documento suele ser una porción relativamente pequeña en relación a todo el contenido del documento. Para esta problemática, se diseñó un *pipeline* que combina técnicas de NLP, ML supervisado (clasificación) y LLMs. La Figura 1 presenta los principales componentes de este *pipeline*.

III-A. Extracción de texto

El punto de partida es un conjunto de documentos legales (por ejemplo, sentencias judiciales) en los cuales se han etiquetado los fragmentos de texto que se considera que poseen sesgo de género. Las sentencias usualmente se encuentran en formato PDF o documentos de texto (por ejemplo, *Office*), lo

[3]https://www.iic.uam.es/procesamiento-del-lenguaje-natural/primer-modelo-lenguaje-en-espanol-adaptado-sector-legal/

[4]https://github.com/PlanTL-GOB-ES/lm-legal-es

[5]https://huggingface.co/aymurai/flair-ner-spanish-judicial

[6]https://themis.es/

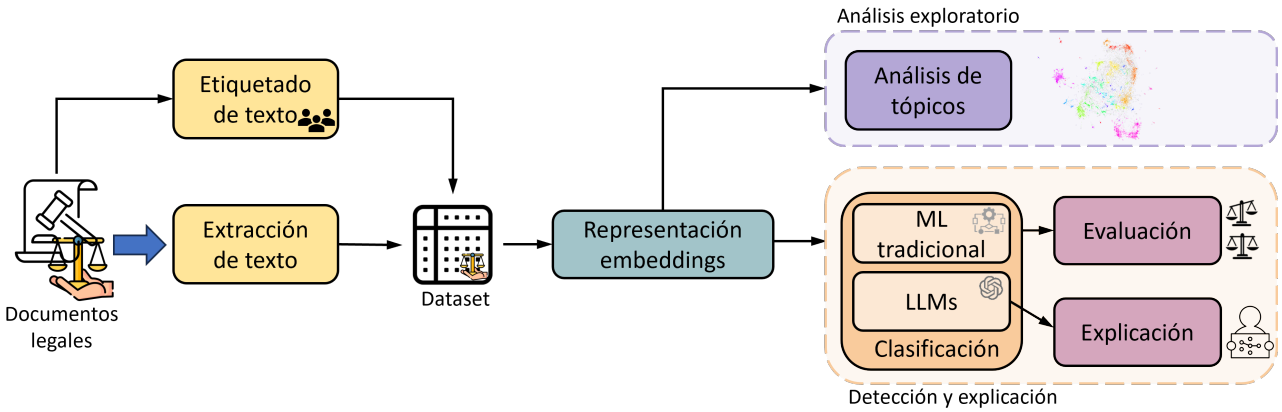

Figura 1: Esquema del pipeline de procesamiento propuesto.

que requiere automatizar la extracción y limpieza de texto (por ejemplo, eliminar caracteres no deseados o frases recurrentes en los encabezados o pies de página). En lugar de tratar cada sentencia como un único documento, se consideró el párrafo como unidad de procesamiento para permitir individualizar de forma más precisa las expresiones que podrían indicar la existencia de sesgos. Luego, cada párrafo se asocia con una etiqueta binaria para indicar la presencia o ausencia de sesgo.

III-B. Representación de embeddings

Los modelos tradicionales de representación de texto se basan principalmente en el léxico, es decir, en las palabras que componen el texto y su frecuencia, a veces teniendo en cuenta algunas palabras cercanas [17]. Al centrarse únicamente en aspectos léxicos, esta representación pierde información crucial relacionada con el orden y la secuencia de las palabras en el texto. Estos modelos también está afectados por el "*curse of dimensionality*", lo que significa que se vuelven menos eficientes a medida que aumenta la dimensionalidad de los datos. Aunque los modelos tradicionales son útiles para una variedad de tareas, carecen de la capacidad necesaria para abordar cuestiones semánticas del análisis de texto, como se motivó anteriormente para sentencias con sesgo implícito.

En contraposición a las representaciones léxicas, las representaciones semánticas basadas en *embeddings* permiten capturar mejor el significado, relaciones y contextos en los que las palabras son utilizadas. Considerando que existe una delgada línea entre el uso de estereotipos positivos, el refuerzo de estereotipos negativos y la consideración de las realidades para lograr la igualdad [1], los *embeddings* capturan el contexto para determinar si los patrones presentes sugieren sesgos.

En el *pipeline* propuesto, para cada párrafo se crea una representación semántica basada en modelos contextuales de *embeddings* pre-entrenados. Para esto, se consideraron diferentes variantes tanto muti-linguales como específicas de español, por ejemplo *BETO*[7], la versión en español de *BERT*.

III-C. Modelado de tópicos

El modelado de tópicos es una técnica de análisis de texto que busca identificar los temas principales latentes y

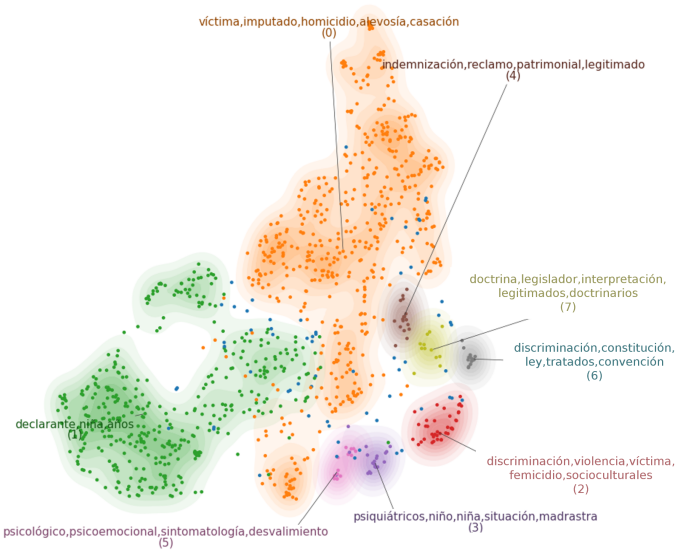

Figura 2: Análisis de tópicos generado por *BERTopic* sobre un cuerpo de sentencias. Proyección 2D con *UMAP* sobre embeddings de *BETO*.

transversales en un conjunto de documentos, permitiendo así descubrir patrones ocultos y extraer información significativa sobre los temas abordados. Un tópico a menudo se describe por un conjunto de palabras o frases que aluden a su temática.

Los tópicos pueden estar distribuidos en los diferentes documentos (por ejemplo, una sentencia, o un párrafo de una sentencia), y por otra parte, un documento puede involucrar uno o más tópicos. Por ejemplo, la Figura 2 muestra los principales tópicos identificados por *BERTopic*[8] en los párrafos de un cuerpo de sentencias. En este caso, cada tópico se describe resumidamente por 5 palabras. Cada punto en la Figura representa un párrafo, los cuales fueron representados con *BETO*. La organización espacial (2D) de estos *embeddings* refleja (aproximadamente) la distancia entre los mismos. Este espacio 2D es de utilidad para entender la distribución y densidad de los diferentes tópicos, y particularmente previsualizar aquellos que puedan referir a estereotipos de sesgo

[7]https://huggingface.co/espejelomar/sentece-embeddings-BETO

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

de género de manera mas o menos explícita. En cierto modo, esta técnica funciona como un diagnóstico (visual) del cuerpo de sentencias, preparatorio para la fase de clasificación.

III-D. Clasificación basada en modelos tradicionales

Tomando la representación de cada párrafo y la etiqueta asignada al mismo, es posible formular un problema de clasificación binaria. Por las características del dominio, donde existe una mayor prevalencia de párrafos sin sesgo (clase negativa), respecto a aquellos con sesgo (clase positiva), se contará con una distribución desbalanceada de las clases.

En un conjunto de datos desbalanceado, la clasificación puede sesgarse hacia la clase dominante y perder la capacidad de identificar correctamente las instancias de la clase positiva. 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. En este caso, se trabajó con algoritmos del tipo *gradient boosting*, que son conocidos por ser robustos para el problema y proveer estrategias de ajuste de pesos para conjuntos de datos desbalanceados [18].

III-E. Clasificación basada en LLMs

Una alternativa para la clasificación es utilizar LLMs. Los LLMs constituyen un cambio de paradigma en NLP ya que facilitan el aprendizaje simplemente definiendo instrucciones en lenguaje natural [14, 19], promoviendo así su aplicación en diversos dominios. En este contexto, la entrada que se le brinda al LLM para resolver una tarea es de vital importancia. Esta entrada, a menudo de caracter instruccional, se conoce como *prompt*. El *prompting* es lo que permite a cualquier usuario (experto o no) interactuar con un LLM utilizando lenguaje natural de manera controlada, aprovechando su capacidad para comprender texto y generar una salida [20].

Al diseñar un *prompt*, uno de los elementos a definir es el rol del LLM, es decir, qué función cumple el LLM en la tarea a desarrollar o a qué perfil está representando. La definición del rol tiene un impacto directo en las respuestas, influyendo no solo en el contenido generado sino también en el nivel de experiencia reflejado en dichas respuestas y límites éticos [21]. En el contexto del problema, se le asignó al LLM el rol de abogado, 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:

> *[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. *Retorná la respuesta en formato JSON conteniendo la clave 'label' con el valor correspondiente a la categoría asignada.*

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.

III-E1. Zero-shot prompting: La estrategia *zero-shot* refiere a la capacidad de realizar una tarea sin haber visto ningún ejemplo de entrenamiento relacionado. De esta forma, el *prompt* definido incluye únicamente el texto a clasificar:

> *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.*
>
> *[DESCRIPCIÓN TAREA]*
>
> *Lista de categorías: {labels}*
> *Texto: "'{text}"'*
> *Tu respuesta en formato JSON:*

La Figura 3 presenta un ejemplo de clasificación utilizando el *prompt* anterior, en el que el LLM recibe el texto a analizar, las potenciales categorías (en este caso, "sesgo" y "no sesgo") e instrucciones respecto al formato para la respuesta.

III-E2. Few-shot prompting: Por su parte, *few-shot* implica el conocimiento de un número limitado de ejemplos relacionados, los cuales son utilizados a modo de "entrenamiento", aunque siguen siendo un número muy acotado de ejemplos en comparación a los necesarios para realizar un entrenamiento completo del modelo. Es interesante destacar que el uso de estrategias *few-shot* puede ser una opción mas confiable que estrategias de *fine-tuning* para la incorporación de nuevo conocimiento en el LLM [22].

Para el caso de *few-shot*, el *prompt* se modifica para incorporar los ejemplos a utilizar como entrenamiento:

> *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.*

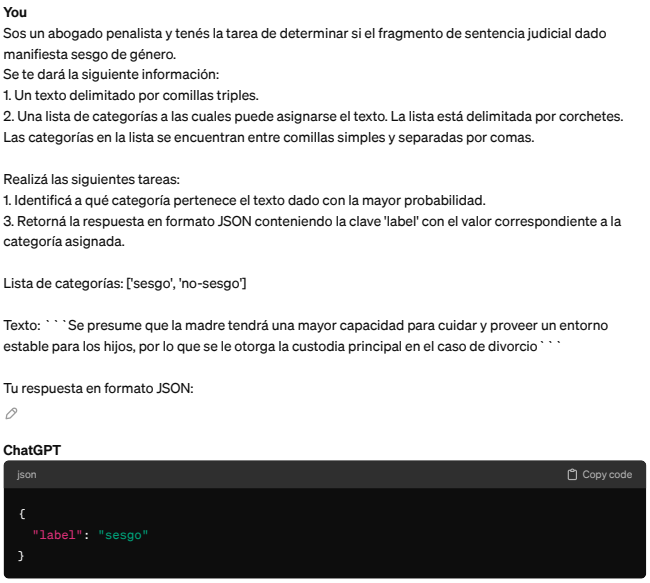

Figura 3: Ejemplo de clasificación de un párrafo siguiendo la estrategia *zero-shot* con *GPT-3.5*.

> *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.*
>
> *[DESCRIPCIÓN TAREA]*
>
> *Lista de categorías:* {*labels*}
> *Textos de entrenamiento:* {*training_data*}
> *Texto:* “‘{*text*}“‘
> *Tu respuesta en formato JSON:*

Considerando la variabilidad de respuestas que pueden obtenerse de acuerdo a la selección de ejemplos de entrenamiento [23], se consideraron dos escenarios:

• `Estático`. Los ejemplos son fijos para todos los textos a analizar. La selección de los ejemplos fue realizada por los autores de común acuerdo. Se definió una cantidad balanceada de ejemplos de ambas clases.

• `Dinámico`. Los ejemplos se eligen de acuerdo al texto a analizar [24]. Para cada texto se eligen sus N vecinos más cercanos[9], de acuerdo a su representación de *embeddings* ya descripta [23]. La cercanía (o semejanza) fue calculada mediante la distancia Euclídea. Este escenario asume que cuanto más parecidos los ejemplos, más información aportan al modelo para lograr respuestas satisfactorias [25].

III-F. Generación de explicaciones

Como los LLMs se pueden ajustar a instrucciones en lenguaje natural para proveer respuestas “útiles”, también pueden generar explicaciones sobre dichas respuestas [26]. 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 asignación a la categoría.*

IV. Configuración experimental

El objetivo del trabajo es determinar un modelo de clasificación binaria adecuado para el problema de sesgo de género. Para ello, se realizó un estudio comparativo de distintas representaciones de texto y estrategias de clasificación[10]. Dado el uso de LLMs para clasificación, se incluyó también el aspecto de explicaciones de los resultados.

Las preguntas de investigación son:

- **PI-1:** ¿Qué representación de texto es más efectiva para una clasificación tradicional?
- **PI-2:** ¿Es la clasificación basada en LLMs comparable a la clasificación tradicional? ¿Hay diferencias en los resultados de los distintos LLMs?
- **PI-3:** ¿Pueden los LLMs proveer explicaciones coherentes y entendibles por humanos de la presencia de sesgo?

IV-A. Datos utilizados

El análisis fue realizado sobre sentencias judiciales de Argentina recopiladas por el “*Observatorio de Sentencias Judiciales*” de “*Articulación Regional Feminista*”[11]. Este Observatorio releva decisiones judiciales de los países involucrados (Argentina, Bolivia, Chile, Colombia, Ecuador, México y Perú),respecto al cumplimiento con la *Convención para la Eliminación de toda forma de Discriminación contra la Mujer*.

Se seleccionaron aquellas sentencias judiciales (documentos) donde se determinó la existencia de sesgo. Cada sentencia está acompañada por un texto breve que describe los párrafos afectados por el sesgo detectado. En función de dichos textos, dos de los autores revisaron y anotaron las sentencias seleccionadas para construir la solución de referencia (*ground truth*). En este análisis se revisaron 12 sentencias (documentos), con

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

[10]En este repositorio se comparte la implementación y los resultados: https://github.com/tommantonela/gender-bias-text

[11]http://www.articulacionfeminista.org/

un total de 1094 párrafos, y para 41 (3,74 %) de ellos se encontraron indicadores de sesgo de género. Esta distribución de las clases es altamente desbalanceada, con una proporción $1:26$ para las clases positiva (sesgo) y negativa.

IV-B. Representaciones textuales

Para los párrafos de las sentencias analizadas, se evaluaron distintos tipos de *embeddings*, según se resume en la Tabla I. A modo de comparación, también se consideró una representación tradicional basada en *TF-IDF*[12].

De acuerdo al *pipeline* propuesto, se realizó un análisis exploratorio de estas representaciones con modelado de tópicos.

TABLA I: Representaciones de texto utilizadas.

	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

IV-C. Modelo de clasificación tradicional

Dentro de la familia de algoritmos de *gradient boosting*, se seleccionó *CatBoost*[13] [18] debido a su robustez y versatilidad. Este algoritmo se corrió sobre cada una de las 4 representaciones. No se realizó ajuste de hiper-parámetros, con la excepción de un ajuste de pesos en la relación entre las clases negativa y positiva. Esto se debe a que el desempeño del algoritmo es sensible al valor de dicho hiper-parámetro. Para cada representación, se aplicó un esquema estandard de división en conjuntos de entrenamiento y testeo, con validación cruzada (particionamiento en 5 folds).

IV-D. Selección de LLMs

Los modelos de LLM utilizados, tanto para clasificación como para explicaciones, se resumen en la Tabla II. Los LLMs fueron accedidos mediante el *framework LangChain*[14], configurando su temperatura en 0 para reducir el azar en las respuestas[15]. Cada texto fue procesado en una interacción separada con *Langchain*, y no se utilizó memoria conversacional.

TABLA II: LLMs utilizados.

	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

IV-E. Otras herramientas

Los modelos entrenados fueron complementados con dos herramientas de terceros: `Themis` y `GenBit` [27][16]. `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 realizar sugerencias de adaptación del texto. Estos términos con sesgo son aquellos morfológicamente masculinos, ya que se aclara que los términos en femenino no se consideran términos excluyentes[17]. Asimismo, `Themis` no parece considerar el contexto de uso, lo que no siempre resulta adecuado[18]. Se consideró que un párrafo presentaba sesgo si contenía alguno de los términos del diccionario de `Themis`. `GenBit` es una herramienta de Microsoft dedicada a medir el sesgo en texto para determinar si el género está distribuido de manera uniforme en los datos. Para esto, analiza la asociación entre una lista predefinida de palabras que definen el género y otras palabras en el corpus a través de estadísticas de co-ocurrencia. Para determinar la presencia de sesgo, se utilizó el *threshold* predefinido por la herramienta.

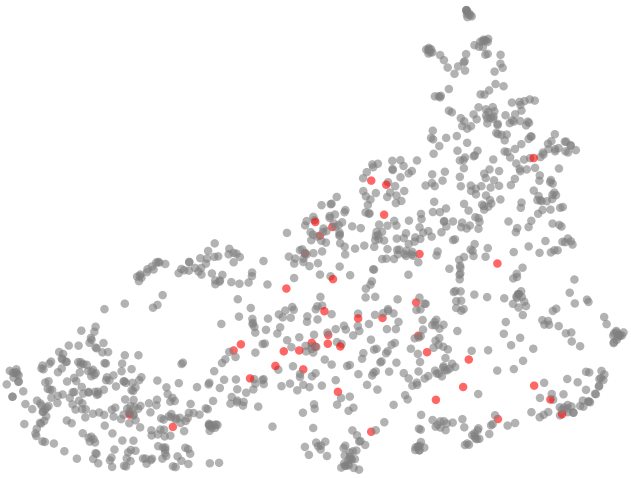

Figura 4: Cuerpo de párrafos con sesgo (rojo) y sin sesgo (gris) de género. Proyección *UMAP* sobre embeddings *BETO*.

IV-F. Métricas de evaluación

La evaluación se basó en precisión y *recall* tanto para la clase positiva (sesgo) y su versión ponderada de acuerdo al peso de las clases[19]. De acuerdo con la naturaleza de la tarea, se considera que el *recall* es más relevante que la precisión. También se incluyó el Coeficiente de Matthews [28], el cual suele ser más confiable que AUC-ROC en conjuntos de datos desbalanceados. Sobre el conjunto de métricas de clasificación se realizó un análisis estadístico por pares ($\alpha = 0{,}01$).

V. Evaluación

Para contextualizar la tarea de clasificación, la Figura 4 ejemplifica la distribución de las instancias de la clase positiva (en rojo), para los *embeddings* de *BETO*, donde puede apreciarse el desbalanceo entre las clases. El análisis exploratorio de tópicos para esos párrafos se presentó en la Figura 2. En particular, se identificaron los siguientes 7 tópicos:

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

[13]https://catboost.ai/

[14]https://python.langchain.com

[15]https://platform.openai.com/docs/api-reference/chat

[16]https://github.com/microsoft/responsible-ai-toolbox-genbit

[17]https://themis.es/faq/

[18]Por ejemplo, ante la expresión "aspectos físicos", sugiere el reemplazo de "físicos" por "físicas".

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

0) Refiere a testimonios de víctimas en delitos contra la integridad sexual, y a la relevancia de las declaraciones de los testigos, relacionados con las medidas judiciales tomadas respecto a situaciones de violencia.

1) Refiere a la presentación de pruebas durante un juicio, tanto de la fiscalía como de la defensa.

2) Refiere a hechos de violencia física explícita a razón de disputas por roles y reclamos de cuidado.

3) Hace referencias explícitas a la violencia de género, y destaca la importancia de no ignorar las manifestaciones de violencia de género, y de reconocer las condiciones que constituyen violencia de género.

4) Hace referencia a situaciones de violencia psicológica y sus efectos, y a los roles ocupados por víctimas y victimarios.

5) Refiere a reclamos de igualad de trato ante reclamos por daño material y moral para ambos géneros.

6) Refiere a efectos psicológicos y sus consecuencias sobre la vulnerabilidad personal y social.

A partir de la Figura 4, puede hacerse una correlación entre las instancias con sesgo y la distribución de tópicos (e instancias) de la Figura 2. 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. Asimismo, en ambos tópicos se observaron manifestaciones implicitas de sesgos, como parte de los alegatos de cada caso. De forma complementaria, los tópicos 2 y 4 refieren de forma explícita a indicadores de sesgo de género, por ejemplo, mediante estereotipos de roles de género como el rol de cuidadora de una madre respecto a un menor a cargo. Estos escenarios motivan la realización de un análisis semántico (más que basado en palabras claves) del texto para la clasificación. En los restantes cuatro tópicos no se evidencia la existencia de párrafos con sesgo.

TABLA III: Resultados de clasificación de sesgo de género

	Binaria (clase sesgo)		Ponderada		Coef.
	Precision	Recall	Precision	Recall	Matthews
Mayoría	0.0	0.0	0.926	0.963	0.0
TF-IDF	0.4	0.5	0.96	0.954	0.424
GenBit	0.047	0.146	0.929	0.857	0.019
Themis	0.029	0.098	0.926	0.843	-0.017
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
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-estático	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	0.001	0.037	0.0
gpt-dinámico	0.093	0.732	0.952	0.722	0.189
Llama2-dinámico	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

V-A. PI-1. Clasificación tradicional

La Tabla III presenta los mejores resultados obtenidos para las alternativas de clasificación tradicional[20]. Se incluye también una clasificación en base a la clase mayoritaria como referencia de los resultados más bajos que pueden obtenerse.

Como se observa, los resultados más bajos fueron obtenidos por Themis y Genbit. En Themis, esto puede deberse a la fuerte relación que establece entre el género masculino de las palabras y la existencia de sesgo, resultando en que cualquier párrafo con una palabra masculina sea etiquetado como con sesgo. Si bien esto le permite encontrar algunos de los párrafos con sesgo, esta estrategia detecta una gran cantidad de falsos positivos (evidenciado por la baja precisión) y no permite identificar aquellos párrafos que incluyen demostraciones sutiles de sesgo. Por otra parte, en Genbit, si bien mejora el *recall*, sigue teniendo una gran cantidad de falsos positivos. Considerando que el *threshold* por defecto para detectar sesgo es alto, esta situación puede deberse a diferencias entre el dominio de los textos utilizados en su entrenamiento y los párrafos evaluados, que puede dar lugar a diferencias entre el vocabulario y registro utilizados. En ambos casos se observaron diferencias estadísticamente significativas a favor de las otras alternativas.

Tanto la alternativa basada en *TF-IDF* como las de *embeddings* mejoraron los resultados previos. En particular, el *recall* fue similar, pero se observaron diferencias para precisión. Esto indica que las alternativas difieren en la cantidad de falsos positivos que encuentran. Dado que la precisión más alta se obtuvo para las representaciones de *embeddings*, es posible inferir que analizar de forma individual el vocabulario utilizado (como hace TF-IDF) no es suficiente para detectar el sesgo, sino que se requiere de un análisis (semántico) que reconozca las expresiones implícitas de sesgo. Se observaron diferencias estadísticamente significativas entre RobertaLex, BETO y TF-IDF, 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 importante 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 [29]. Adicionalmente, se observó que las diferencias entre RobertaLex y BETO son estadísticamente significativas con un tamaño de efecto pequeño, lo cuál muestra la utilidad de contar con *embeddings* entrenados en textos del dominio de la tarea.

> En resumen, tanto *BETO* como *RobertaLex* resultaron representaciones efectivas para la clasificación, con una leve ventaja de *RobertaLex*.

V-B. PI-2. Clasificación con LLMs

La Tabla III presenta los resultados de las alternativas de clasificación *zero-shot* (zero) y *few-shot* (estático y dinámico) para los LLMs evaluados. En el caso dinámico, se reportan los resultados seleccionando los vecinos utilizando BETO, dado que fue la que obtuvo el mejor desempeño.

[20]El resto de los resultados están disponibles en el repositorio.

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 Gemma y Llama2. Si bien Llama2 alcanzó el *recall* más alto, fue a expensas de determinar que casi todos los párrafos contenían sesgo (mostrando así una alta sensibilidad al lenguaje utilizado [30]), generando muchos falsos positivos, y en consecuencia el peor valor del Coeficiente de Matthews.

Comparando *zero-shot* con *few-shot* estático, excepto para Gemma, las alternativas *few-shot* obtuvieron un mejor valor para el Coeficiente de Matthews. Para Llama2, si bien el *few-shot* disminuyó su *recall* para la clase positiva en 66 %, introdujo también una mejora en precisión de 100 %. Para Gemma, el incremento de *recall* fue acompañado por una disminución de la precisión, dado que todos los textos fueron clasificados como sesgo. La superioridad de las alternativas *few-shot* se observa también para el Coeficiente de Matthews. 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*, dinámica mejoró el balance de precisión y *recall* para todos los LLMs. Si bien, excepto para Llama2, se observan valores de *recall* más bajos (con una disminución promedio de 22 %), se logró mejorar la precisión en todos los casos (con un aumento promedio de 84 %). Este mejor balance de la clasificación se ve acompañado por mejoras en el Coeficiente de Matthews. Para GPT y Gemma las diferencias con estática fueron significativas. Asimismo, todos los modelos mostraron diferencias significativas respecto a las alternativas *zero-shot* correspondientes. En lo que respecta a las métricas ponderadas, en promedio, el *recall* se incrementó en un 143 %. Si bien la disminución en *recall* de la clase positiva implica que queda sin identificar una mayor cantidad de textos, al incrementar la precisión se disminuyen los falsos positivos, potencialmente permitiendo reducir los esfuerzos de una tarea manual posterior.

En general, para *few-shot*, se observaron mejoras promedio en precisión de 87 % (mínimo de 19 % para GPT, y máximo de 155 % para Gemma) y en Coeficiente de Matthews de 178 % (mínimo de 15 % para GPT y máximo de 845 % para Gemma). Asimismo, definiendo el balance entre precisión y *recall* como *f-measure*, se observaron mejoras de 70 % (mínimo de 16 % para GPT y máximo 128 % para Gemma). Estos resultados muestran el efecto diferenciado que el contexto de la tarea tiene sobre cada LLM, siendo importante no solo agregar los ejemplos, sino también cómo se definen, ya que elegir los textos más cercanos a aquel a clasificar permitió mejorar la precisión, el balance entre precisión y *recall* y el Coeficiente de Matthews 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 si bien 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 [31, 32].

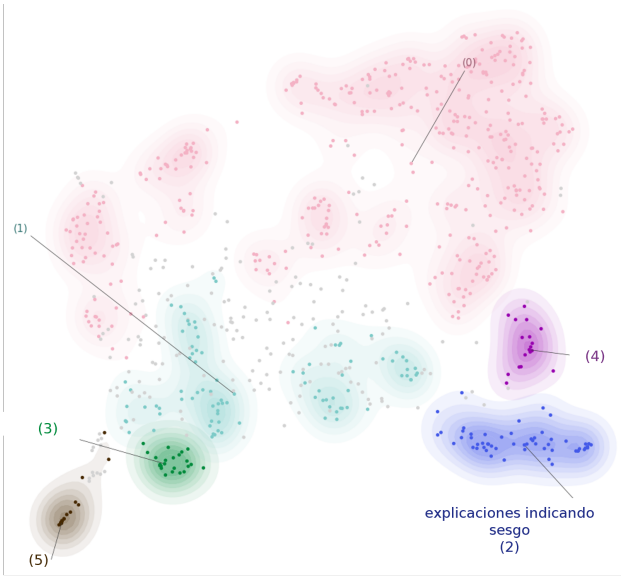

Figura 5: Análisis de tópicos generado por *BERTopic* para las explicaciones. Proyección *UMAP* sobre embeddings *BETO*.

Finalmente, en relación a la valoración cualitativa de las respuestas en términos de cuán 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 devolvían respuestas cambiando los nombres de las claves solicitadas en el formato JSON, y Llama2 alternaba entre español e inglés, entre otras incidencias.

> En resumen, la clasificación basada en LLMs obtuvo un desempeño inferior al de la tradicional, aunque no todas las diferencias fueron estadísticamente significativas.

V-C. PI-3. Explicaciones provistas por los LLMs

Más allá de su capacidad de clasificación de textos, los LLMs pueden generar explicaciones textuales que acompañen el resultado de su clasificación. La Figura 5 presenta la separación en tópicos para las explicaciones dadas por GPT, considerando aquellos parrafos 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. 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.

Se realizó una inspección manual de las explicaciones de GPT. Un ejemplo de explicación de sesgo es "*El texto presenta un sesgo de género 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 cuál 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 (Figura 5), 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.

En resumen, ciertos LLMs, como `GPT`, pueden generar explicaciones satisfactorias de los resultados de la clasificación desde el punto de vista de un humano, especialmente respecto a la existencia de sesgo. No obstante, debe realizarse una validación más exhaustiva de los resultados por parte de un experto legal, y asegurar que el LLM no introdujo alucinaciones.

VI. Conclusiones

Este trabajo aborda la tarea de detección (automatizada) de sesgo de género en textos legales mediante el uso de técnicas de NLP, ML y LLMs. En esta línea, se desarrolló un *pipeline* de procesamiento, cuyas etapas principales involucran la representación semántica (via *embeddings*) del texto y su clasificación binaria. Para esta clasificación se compararon estrategias tradicionales y la capacidad de inferencia de distintos LLMs. Los resultados iniciales mostraron un buen desempeño de la clasificación tradicional en términos de precisión y *recall*. Las representaciones semánticas elegidas resultaron adecuadas para la tarea. Sin embargo, resulta necesario realizar una evaluación en mayor profundidad de representaciones específicas al lenguaje legal en español. Por su parte, si bien la clasificación basada en LLMs obtuvo resultados inferiores, se considera que su capacidad para generar explicaciones textuales es auspiciosa. En este sentido, sería factible combinar una clasificación basada en *gradient boosting* con LLMs que generen las explicaciones correspondientes.

Desde un punto de vista de aplicación social, este trabajo tiene potencial para asistir a actores del dominio jurídico, contribuyendo a su accionar con una revisión y análisis de documentos que ayuden a la detección y prevención de posibles sesgos de género. Las técnicas y modelos desarrollados pueden ser utilizados no solo para analizar documentos ya escritos, sino también durante su proceso de escritura, permitiendo una reducción efectiva de la problemática. Por ejemplo, podría usarse el enfoque propuesto para computar un puntaje (*score*) para un documento legal en función de la clasificación de los párrafos que lo componen, y usar este puntaje para alertar sobre documentos con una alta probabilidad de contener sesgo.

Como limitaciones a abordar en estudios futuros pueden mencionarse el tamaño acotado del conjunto de sentencias utilizado en los experimentos, y también los desafíos de confiar en los resultados generados por los LLMs. En primer lugar, se debiera contar con una colección de datos más amplia, que incorpore sentencias de distintas provincias y años que permita evaluar si las manifestaciones de sesgo varían en relación a diferentes contextos socio-culturales, y si es posible identificar cambios en las características y/o tendencias a lo largo de los años. Segundo, podría considerarse un refinamiento de las manifestaciones de sesgo en términos de operacionalizar estereotipos concretos[21] [33], que ayuden a un mejor análisis de cada caso. Tercero, teniendo en cuanta que el conocimiento de los LLMs es estático y orientado a dominios generales, lo que limita su capacidad de generalización a dominios específicos, se podrían extender la alternativa *few-shot* para incorporar más conocimiento específico del dominio, por ejemplo, mediante un esquema RAG (*Retrieval Augmented Generation*) o una técnica de *fine-tuning* [22, 34]. Finalmente, podrían explorarse otras alternativas de *prompting*, por ejemplo, variantes de lenguaje, estilo o registro.

Consideraciones éticas

Dada la sensibilidad de los temas relacionados al género, se comparten públicamente los *embeddings*, código y modelos derivados para fomentar la transparencia de la metodología empleada y los resulados obtenidos. Sin embargo, para preservar la identidad e integridad de las partes involucradas, no se comparte de forma directa el conjunto de sentencias originales.

Los análisis mostrados no deben ser utilizados para identificar ni estigmatizar a individuos introduciendo (aunque sea de forma inadvertida) sesgo en los documentos. Por el contrario, deben ser tomados como una oportunidad para promover un diálogo constructivo sobre el sesgo y sus consecuencias. Es

[21] Algunos gobiernos ya han comenzado a trabajar con estereotipos, por ejemplo, Uruguay definió: https://www.gub.uy/sites/gubuy/files/inline-files/GuaparaelPoderJudicial.pdf

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 género tomadas de los conjuntos de datos con los que fueron entrenados [35, 36].

REFERENCIAS

[1] A. S. Timmer *et al.*, "Gender stereotyping in the case law of the eu court of justice," *European Equality Law Review*, no. 1, pp. 37–46, 2016.

[2] N. Ellemers, "Gender stereotypes," *Annual review of psychology*, vol. 69, pp. 275–298, 2018.

[3] L. Casas and J. P. G. Jansana, "Estereotipos de género en sentencias del tribunal constitucional," *Anuario de Derecho Público*, no. 1, pp. 250–272, 2012.

[4] S. R. Rodríguez, M. J. Reinuaba, L. P. Decap, and J. Z. Yong, "La perspectiva de género en la valoración de la prueba en procedimientos laborales por acoso," *Revista de Derecho*, no. 40, pp. 103–134, 2022.

[5] M. J. Cepeda, G. Otálora *et al.*, "Modernización de la administración de justicia a través de la inteligencia artificial," *Bogotá: Fedesarrollo*, 2020.

[6] A. G. Greenwald and M. R. Banaji, "Implicit social cognition: attitudes, self-esteem, and stereotypes." *Psychological review*, vol. 102, no. 1, p. 4, 1995.

[7] W. Schmeisser-Nieto, M. Nofre, and M. Taulé, "Criteria for the annotation of implicit stereotypes," in *Proceedings of the 13th LREC Conference*. France: ELRA, 2022, pp. 753–762.

[8] T. Huang, F. Brahman, V. Shwartz, and S. Chaturvedi, "Uncovering implicit gender bias in narratives through commonsense inference," in *Findings of the ACL: EMNLP 2021*. Dominican Republic: ACL, Nov. 2021, pp. 3866–3873.

[9] E. Ash, D. L. Chen, and A. Ornaghi, "Gender attitudes in the judiciary: Evidence from us circuit courts," *American Economic Journal: Applied Economics*, vol. 16, no. 1, pp. 314–350, 2024.

[10] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, "Detecting gender stereotypes: Lexicon vs. supervised learning methods," in *Proceedings of the 2020 CHI conference*.

[11] N. Sevim, F. Şahinuç, and A. Koç, "Gender bias in legal corpora and debiasing it," *Nat. Lang. Eng.*, vol. 29, pp. 449–482, 2023.

[12] A. M. Aranguren, "El peso de los estereotipos de género en las decisiones judiciales. una aproximación desde la psicología jurídica," in *Revisión de las políticas y prácticas ante la violencia de género*, 2020, pp. 37–79.

[13] A. d. Castro-Rodrigues and A. Sacau, "La influencia del género en las decisiones de los tribunales: del paternalismo judicial a los papeles familiares," *Estudos Feministas*, vol. 20, pp. 119–132, 2012.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[15] X. Kang, L. Qu, L.-K. Soon, A. Trakic, T. Y. Zhuo, P. C. Emerton, and G. Grant, "Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer?" *arXiv:2310.14880*, 2023.

[16] J. Tan, H. Westermann, and K. Benyekhlef, "Chatgpt as an artificial lawyer?" *AI4AJ 2023*, 2023.

[17] M. Melucci and R. Baeza-Yates, *Advanced topics in information retrieval*. Springer Science & Business Media, 2011, vol. 33.

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," vol. 31, 2018.

[19] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, "Generate rather than retrieve: Large language models are strong context generators," *arXiv:2209.10063*, 2022.

[20] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference*, pp. 1–21.

[21] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," *arXiv:2304.05335*, 2023.

[22] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? comparing knowledge injection in llms," *arXiv:2312.05934*, 2023.

[23] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in *Proceedings of Deep Learning Inside Out (DeeLIO 2022)*. Dublin, Ireland and Online: ACL, May 2022, pp. 100–114.

[24] P. Pezeshkpour, Z. Zhao, and S. Singh, "On the utility of active instance selection for few-shot learning," *NeurIPS HAMLETS*, vol. 3, 2020.

[25] B. Pecher, I. Srba, M. Bielikova, and J. Vanschoren, "Automatic combination of sample selection strategies for few-shot learning," *arXiv:2402.03038*, 2024.

[26] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, and L. H. Gilpin, "Can large language models explain themselves? a study of llm-generated self-explanations," *arXiv:2310.11207*, 2023.

[27] S. Bordia and S. R. Bowman, "Identifying and reducing gender bias in word-level language models," *arXiv:1904.03035*, 2019.

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

[29] A. Velankar, H. Patil, and R. Joshi, "Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi," in *Artificial Neural Networks in Pattern Recognition*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 121–128.

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

[31] Y. Qiu and Y. Jin, "Chatgpt and finetuned bert: A comparative study for developing intelligent design support systems," *Intelligent Systems with Applications*, vol. 21, p. 200308, 2024.

[32] Y.-N. Chuang, R. Tang, X. Jiang, and X. Hu, "Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization," *arXiv preprint arXiv:2303.13035*, 2023.

[33] R. J. Cook and S. Cusack, *Gender Stereotyping: Transnational Legal Perspectives*. University of Pennsylvania Press, 2010. [Online]. Available: http://www.jstor.org/stable/j.ctt3fhmhd

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[35] A. Salinas, P. Shah, Y. Huang, R. McCormack, and F. Morstatter, "The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama," in *Proceedings of the 3rd ACM Conference on EAAMO*. NY, USA: ACM, 2023.

[36] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proceedings of The ACM Collective Intelligence Conference*, 2023, pp. 12–24.