Distilling Multilingual Models for Few-Shot Gender-Bias Detection in Spanish Judicial Texts: A Systematic Literature Review

Christian Javier Ratovicius[1], Antonela Tommasel[2]

[1]Universidad Abierta Interamericana, Facultad de Tecnología Informática,
Centro de Altos Estudios en Tecnología Informática, Buenos Aires, Argentina
christianjavier.ratovicius@alumnos.edu.ar,
[2]ISISTAN, CONICET-UNICEN
Tandil, Argentina
antonela.tommasel@isistan.unicen.edu.ar

Abstract. Gender bias in judicial decisions constitutes a critical challenge for the fairness and legitimacy of justice systems. In the field of Natural Language Processing (NLP), automated bias detection typically relies on large volumes of labeled data, which limits its applicability in low-resource settings, such as the Spanish legal domain. Although recent studies have independently addressed knowledge distillation, domain adaptation, and fairness evaluation, their integration in judicial scenarios with limited resources has not yet been systematically examined. This work presents a Systematic Mapping Study (SMS) that analyzes 65 primary studies at the intersection of gender bias detection, efficient NLP models, and legal domain adaptation. The review follows established methodological guidelines to identify, classify, and evaluate existing approaches, with a focus on distillation techniques, domain-specific pretraining, and counterfactual evaluation. The results suggest that compact models such as *DistilBERT* and *TinyLlama* can preserve a substantial portion of the predictive performance of larger architectures, such as *BERT-large* and *LLaMA*, while significantly reducing computational requirements and enabling deployment in local infrastructures with limited resources. Furthermore, domain adaptation improves the detection of implicit biases, while counterfactual evaluation emerges as a relevant, albeit methodologically challenging, approach in languages with grammatical gender such as Spanish. Nevertheless, limitations are identified regarding evaluation consistency, the linguistic validity of counterfactuals, and potential trade-offs between efficiency and fairness.
.

Keywords: Gender bias, Natural Language Processing, Judicial Decisions, Knowledge Distillation, Domain Adaptation

Destilación de Modelos Multilingües para la Detección de Sesgo de Género con Pocos Ejemplos en Textos Judiciales en Español: Una Revisión Sistemática de la Literatura

Resumen. El sesgo de género en las decisiones judiciales constituye un desafío crítico para la equidad y la legitimidad de los sistemas de justicia. En el ámbito del Procesamiento de Lenguaje Natural (NLP), la detección automatizada de sesgos suele depender de grandes volúmenes de datos etiquetados, lo que limita su aplicación en contextos de bajos recursos, como el dominio jurídico en español. Si bien trabajos recientes han abordado de forma independiente la destilación de conocimiento, la adaptación de dominio y la evaluación de equidad, su integración en escenarios judiciales con recursos limitados aún no ha sido sistematizada de manera exhaustiva. Este trabajo presenta un Mapeo Sistemático de la Literatura (MSL) que analiza 65 estudios primarios en la intersección entre detección de sesgo de género, modelos eficientes de NLP y adaptación al dominio legal. La revisión sigue lineamientos metodológicos establecidos para identificar, clasificar y evaluar los enfoques existentes, con foco en técnicas de destilación, preentrenamiento específico de dominio y evaluación contrafáctica.
Los resultados sugieren que los modelos compactos como *DistilBERT* o *TinyLlama* pueden preservar gran parte del desempeño predictivo de arquitecturas de mayor escala como *BERT-large* o *LLaMA*, reduciendo significativamente los requerimientos computacionales y habilitando su despliegue en infraestructuras locales con recursos limitados. Asimismo, la adaptación de dominio mejora la detección de sesgos implícitos, mientras que la evaluación contrafáctica se presenta como una herramienta relevante, aunque metodológicamente desafiante en lenguas con género gramatical como el español. No obstante, se identifican limitaciones vinculadas a la consistencia en la evaluación, la validez lingüística de los contrafácticos y los posibles *trade-offs* entre eficiencia y equidad.

Palabras clave: Sesgo de género, Procesamiento de Lenguaje Natural, Decisiones judiciales, Destilación de conocimiento, Adaptación de dominio

1 Antecedentes

Los avances recientes en Procesamiento de Lenguaje Natural (NLP, por su nombre en inglés) han emergido como un factor clave que influye en la modernización de los sistemas judiciales, particularmente en tareas de análisis documental y apoyo a la toma de decisiones. No obstante, la literatura advierte que la aceptación de estas tecnologías suele ser reactiva al desempeño percibido de las instituciones, fenómeno conocido como aversión al algoritmo (Jing-Huey, Wei-Ting, & Bao-Yu, 2026).
En este escenario, surge la necesidad de extraer conocimiento de manera eficiente para apoyar la toma de decisiones, evolucionando hacia métodos que busquen escalabilidad y rendimiento sin comprometer la equidad. Un desafío crítico es la presencia de sesgos de género latentes en el lenguaje legal, los cuales son particularmente complejos en

español debido a unidades fraseológicas específicas que los modelos genéricos no logran capturar (Gutiérrez Rubio, 2019).

La motivación de este trabajo surge de la necesidad crítica de garantizar la imparcialidad en el Poder Judicial, donde el sesgo de género puede comprometer el derecho judicial efectivo. A pesar de los esfuerzos por digitalizar la justicia, la evaluación de la equidad en las sentencias sigue siendo un proceso manual, costoso y propenso a errores humanos. La literatura reciente indica que la falta de herramientas de auditoría automatizadas permite que los estereotipos de género se perpetúen de forma invisible en el lenguaje jurídico (John, Aiswarya, & Panachakel, 2025); (Derner, Fuente, & Gutiérrez, 2024). Desde una perspectiva técnica, esto se agrava porque los sistemas judiciales operan bajo estrictas restricciones de confidencialidad que limitan el uso de infraestructura computacional externa o servicios en la nube, convirtiéndolos en entornos de bajos recursos por definición. En consecuencia, los modelos de lenguaje de gran escala resultan inviables tanto por su costo computacional como por los riesgos que implica el procesamiento externo de datos sensibles, por lo que investigar técnicas de destilación de conocimiento (Xinyin, Gongfan, & Xinchao, 2023; Chalkidis, Androutsopoulos, & Aletraa, 2019) y poda estructural (Xinyin, Gongfan, & Xinchao, 2023) no es solo un desafío académico, sino una necesidad operativa. En este marco, la destilación de conocimiento (Ma, Xiaowei, & Jiazhang, 2026) se presenta como una solución estratégica, ya que permite transferir la capacidad de razonamiento de modelos computacionalmente costosos hacia modelos *student* más compactos y eficientes —como DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020) o TinyLlama (Zhang, Zeng, Wang, & Lu, 2024)— habilitando su despliegue en infraestructuras locales y garantizando así tanto la soberanía tecnológica como la privacidad de los datos judiciales. Finalmente, la escasez de recursos específicos para el español en el dominio legal motiva el desarrollo de un *framework* que integre la fraseología técnica (Gutiérrez Rubio, 2019) y la adaptación de dominio (Chalkidis, Malakasiotis, Aletras, & Fergadiotis, 2020), asegurando que la tecnología sea capaz de entender las sutilezas del derecho hispanohablante y reduciendo la dependencia de soluciones genéricas diseñadas para el inglés.

En este contexto, el presente trabajo propone realizar un Mapeo Sistemático de la Literatura (MSL) con el objetivo de analizar, clasificar y sintetizar la evidencia existente en la intersección entre detección de sesgo de género, modelos eficientes de NLP y adaptación al dominio legal. En particular, se examina el rol de la destilación de conocimiento, el preentrenamiento específico de dominio y las técnicas de evaluación contrafáctica en escenarios con recursos limitados.

En adelante, este MSL se organiza de la siguiente manera: en la sección 2 se describen las preguntas de investigación que guían el proceso de búsqueda y selección de estudios; en la sección 3 se explica el método de revisión con las fuentes de búsqueda, los criterios y la extracción de datos; en la sección 4 se presentan los estudios incluidos y excluidos con su síntesis; en la sección 5 se discuten los resultados y se responden las preguntas de investigación; y finalmente en la sección 6 las recomendaciones para futuras investigaciones y las conclusiones del trabajo.

2 Preguntas de investigación

La presente sección detalla el marco metodológico utilizado para la realización del Mapeo Sistemático de la Literatura (MSL). Se sigue un proceso estructurado para la identificación, selección y evaluación crítica de los 65 artículos que componen el estado del arte de esta investigación, asegurando la reproducibilidad y el rigor académico del estudio.

2.1 Preguntas de Contexto

Con el fin de obtener un panorama general y estadístico de la producción científica en el área de estudio, se han definido cuatro Preguntas de Contexto (PC). Estas preguntas permiten clasificar la literatura seleccionada según su origen, cronología y naturaleza, proporcionando una base sólida para el análisis de tendencias:

PC1: ¿Cuál es la distribución cronológica de las publicaciones identificadas? Esta pregunta busca identificar el nivel de actualidad del tema y si existe un crecimiento exponencial en el interés por la detección de sesgos y la eficiencia en los últimos años.
PC2: ¿Cuáles son los principales canales de publicación (revistas, conferencias, repositorios)? Esta pregunta busca determinar si el conocimiento sobre IA judicial y destilación se está consolidando en ámbitos puramente técnicos (*Computer Science*) o en espacios interdisciplinarios (*AI & Law*).
PC3: ¿Cuál es el origen geográfico de las investigaciones y los corpus utilizados? Esta pregunta busca contrastar la predominancia de estudios en inglés frente a la emergencia de investigaciones situadas en el ámbito hispanohablante y europeo.
PC4: ¿Qué tipos de arquitecturas de modelos de lenguaje (LLM, *Transformers*, *Small Models*) son las más referenciadas? Esta pregunta busca mapear la transición desde modelos masivos hacia arquitecturas eficientes, destiladas y especializadas que sustentan el marco teórico de este MSL.

2.2 Preguntas de Investigación específicas

Siguiendo el estándar de los mapeos sistemáticos en el área de Ciencias de la Computación, el objetivo de esta revisión se descompone en preguntas de investigación (PI) específicas. Estas preguntas han sido diseñadas para cubrir tanto la dimensión ética (sesgo) como la dimensión técnica (eficiencia y destilación) del *framework* propuesto:

PI1: ¿Es viable detectar sesgos de género en sentencias en español con un modelo destilado y adaptado al dominio con pocas etiquetas? Esta pregunta busca identificar si la literatura académica valida el uso de modelos compactos frente a grandes modelos generales en contextos de escasez de datos. Es crucial determinar si la adaptación *in-domain* permite capturar las sutilezas lingüísticas y jurídicas del español necesarias para una auditoría de sesgos efectiva**.**

PI2: ¿En qué medida la destilación preserva la precisión del *teacher* reduciendo costo y tamaño? La literatura reporta un *trade-off* entre tamaño y rendimiento. Esta pregunta tiene como fin analizar la viabilidad técnica de transferir capacidades cogni-

tivas de modelos masivos hacia modelos más eficientes (Chalkidis, Malakasiotis, Aletras, & Fergadiotis, 2020) (Gu, Dong, Wei, & Huang, 2026), permitiendo su despliegue en infraestructuras judiciales locales con recursos limitados.

PI3: ¿Cuánto aporta la adaptación de dominio (frente a no adaptarla) en el desempeño *in-domain*? Se pretende cuantificar el beneficio de especializar modelos genéricos mediante el uso de corpus legales. La literatura sugiere que la fraseología del español jurídico requiere herramientas de auditoría específicas, como la intervención de nombres y roles (Ribeiro, Wu, & Guestrin, 2020) (Derner, Fuente, & Gutiérrez, 2024), para garantizar la robustez del sistema.

PI4: ¿Qué impacto tiene cada componente (destilación, adaptación, contrafactuales) en eficacia y equidad? Esta pregunta busca desglosar el aporte individual de cada técnica en la mitigación del sesgo. Interesa identificar si el uso de modelos especializados como BETO (Cañete, Ho, Kang, & Perez, 2023) o MEL (Betancur Sánchez, Aldama García, Barbero Jiménez, & Guerrero Nieto, 2025) ofrece una ventaja competitiva frente a soluciones globales, asegurando que la búsqueda de eficiencia no comprometa la integridad ética de las decisiones judiciales.

3 Métodos de revisión

Para garantizar la validez y el rigor científico de este MSL, se ha implementado un protocolo de revisión basado en las directrices de Kitchenham (d'Amato, Rubini, & Didio, 2025) para la ingeniería de software. Este proceso se divide en la identificación de fuentes, la definición de estrategias de búsqueda y la aplicación de criterios de selección para consolidar el corpus de estudio.

3.1 Fuentes de datos

La estrategia de búsqueda se diseñó para capturar la literatura más relevante y reciente en la intersección del Derecho, la Inteligencia Artificial y la eficiencia computacional. Se seleccionaron bases de datos académicas y motores de búsqueda de alto impacto que garantizan la cobertura tanto de artículos de revistas indexadas como de actas de conferencias internacionales de primer nivel.

Las fuentes de datos utilizadas para la extracción de los artículos analizados son:

- **Scopus**. Utilizadas para identificar artículos con revisión por pares en revistas de alto factor de impacto y conferencias consolidadas en IA y Derecho.
- **arXiv**. Fundamental para capturar el "estado del arte" más reciente en Grandes Modelos de Lenguaje (LLMs, por su nombre en inglés) y técnicas de destilación de conocimiento (*knowledge distillation*), dado que el ciclo de publicación en esta área es extremadamente rápido.
- **ACL**. Seleccionadas específicamente en NLP la misión es promover el estudio científico del lenguaje desde una perspectiva computacional. Funciona como un centro neurálgico para la investigación que conecta la lingüística, la informática, la inteligencia artificial y las ciencias cognitivas.

- **Google Scholar**. Empleado como motor de búsqueda complementario para la validación de citas y la identificación de literatura gris relevante (informes técnicos de organismos judiciales).
- **IEEE Xplore.** Biblioteca digital que reúne publicaciones científicas y técnicas de alta calidad en áreas como ingeniería, computación e inteligencia artificial, incluyendo revistas y conferencias con revisión por pares ampliamente reconocidas en la comunidad académica.

3.2 Estrategia de búsqueda

La estrategia de búsqueda fue elaborada con el objetivo de optimizar la identificación de estudios relevantes, mediante el empleo de operadores booleanos y términos técnicos específicos. En la Tabla 1 se presentan los términos principales y sus variantes, así como las cadenas de búsqueda utilizadas en el proceso de revisión.

Tabla 1: Listado de términos principales y alternativos

Términos principales	Términos alternativos
Legal NLP	Minería de textos legales, Procesamiento de lenguaje jurídico.
Sesgo de género	Equidad (Fairness), Sesgo de genero
Knowledge Distillation	Compresión de modelos, Aprendizaje Maestro-Estudiante, Destilación de conocimiento.
Sistema Judicial	Sentencias judiciales, Dominio legal.

Además, se realizó una búsqueda revisando las referencias de los artículos más citados como (Bolukbasi, Chan, Zou, & Saligrama, 2016) y (Ma, Xiaowei, & Jiazhang, 2026) para identificar estudios fundacionales que no hubieran sido capturados por la cadena de búsqueda inicial.

La misma se definió utilizando cadenas en inglés, dado que la mayor parte de la producción científica en el área de Procesamiento de Lenguaje Natural se publica en este idioma. No obstante, con el objetivo de capturar estudios relevantes en el contexto hispanohablante, se incorporaron **cadenas de búsqueda** complementarias en español. La misma en inglés es *("natural language processing" OR NLP OR "text mining" OR "language models") AND ("gender bias" OR "bias detection" OR fairness) AND (legal OR judicial OR "legal text") AND ("knowledge distillation")* y su contraparte en español *("procesamiento de lenguaje natural" OR NLP OR "minería de texto" OR "modelos de lenguaje") AND ("sesgo de género" OR "detección de sesgo" OR equidad) AND (legal OR judicial OR "texto legal") AND ("destilación de conocimiento").*

3.3 Criterios de selección de estudios

Para refinar el corpus inicial y llegar a los 65 artículos finales, se aplicaron criterios de inclusión **(CI)** y exclusión **(CE)** rigurosos, garantizando que solo la evidencia de mayor calidad y relevancia fuera analizada:

Criterios de Inclusión (**CI**):

CI1. Artículos que propongan o evalúen métodos de detección de sesgo de género en textos escritos.
CI2. Investigaciones que utilicen técnicas de compresión de modelos, específicamente destilación de conocimiento, aplicada a modelos *Transformer*.
CI3. Estudios centrados en el dominio legal o judicial, o que presenten metodologías transferibles a este ámbito.
CI4. Trabajos que incluyan experimentación o validación en idioma español o modelos multilingües con soporte para español.

Criterios de Exclusión (**CE**):

CE1. Artículos que no hayan pasado por un proceso de revisión por pares, exceptuando *pre-prints* de alto impacto con métricas de citación verificables.
CE2. Estudios que no estén disponibles (o traducidos) en idioma inglés o español.
CE3. Trabajos cuya antigüedad sea superior a 10 años, priorizando la literatura publicada entre 2016 y 2026 para reflejar el estado del arte de los modelos basados en *Transformers*.
CE4. Estudios cuyo texto no este completo.

3.4 Evaluación de la calidad de los estudios

Como resultado de la ejecución de la cadena de búsqueda en las bibliotecas, se obtiene la siguiente cantidad de estudios que han sido publicados, que figuran en la Tabla 2 Para garantizar que los resultados del mapeo se basen en evidencia científica robusta, cada uno de los 65 artículos seleccionados fue sometido a un proceso de evaluación de calidad. Se definieron criterios de puntuación basados en 4 dimensiones críticas.

1. *Rigor Metodológico*. ¿El estudio describe claramente el diseño experimental, el corpus utilizado y las métricas de evaluación?
2. *Relevancia Temática*. ¿El artículo aborda de manera directa el sesgo de género o técnicas de eficiencia (KD/Compresión) aplicables al lenguaje?
3. *Contribución al Dominio*. ¿Los hallazgos son transferibles o específicos para el ámbito legal o judicial?
4. *Validación de Resultados*. ¿Se presentan pruebas de significancia estadística o comparaciones con modelos de referencia?

Tabla 2: Estudios resultantes totales en cada biblioteca digital

Biblioteca digital	Estudios Resultantes	Biblioteca digital	Estudios Resultantes
Scopus	43	IEEE Xplore	18
arXiv	55	Google Académico	12
ACL	32		

3.5 Extracción de datos

Con el propósito de recolectar la información necesaria en función de las preguntas de investigación, se realiza una segunda etapa de evaluación que consiste en la definición de un proceso sistemático aplicado a cada uno de los estudios relevantes. En esta instancia, se extraen los datos principales y se explicitan las preguntas de investigación,

para luego identificar, dentro de cada estudio preseleccionado, citas o fragmentos que den respuesta a dichas preguntas. Los datos considerados incluyen autores, título, año de publicación, tipo de documento, país, fuente de origen y un breve resumen. Como resultado, se seleccionaron 65 estudios primarios que fueron considerados pertinentes y que contribuyen de manera significativa al MSL.

En la Figura 1 se observa el crecimiento de artículos desde 2021 hasta 2026. Se observa un pico en 2023-2024 debido al auge de los LLMs y técnicas de destilación de conocimiento.

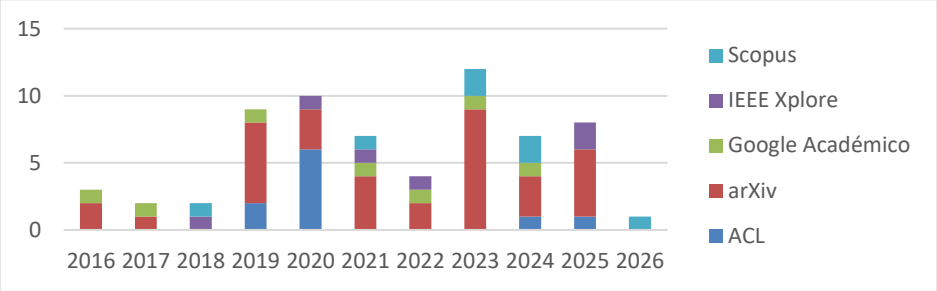

Fig 1 :Evolución de la cantidad de estudios publicados en cada biblioteca digital

En cuanto a los canales de publicación, los 65 estudios primarios se distribuyen principalmente entre conferencias como ACL, EMNLP y JAIIO y revistas indexadas como *Information* o *Computer Law & Security Review*, con una proporción menor de *preprints*. Respecto a la naturaleza de los trabajos, predominan los estudios de orientación teórica, que representan aproximadamente el 85% del total, frente a los de carácter práctico, enfocados en herramientas, aplicaciones y procedimientos concretos.

En la Figura 2, se muestra una fuerte concentración de estudios en Estados Unidos (32), seguido por China (8) y España (4), mientras que el resto de los países presenta una participación marginal, evidenciando una distribución geográfica desigual.

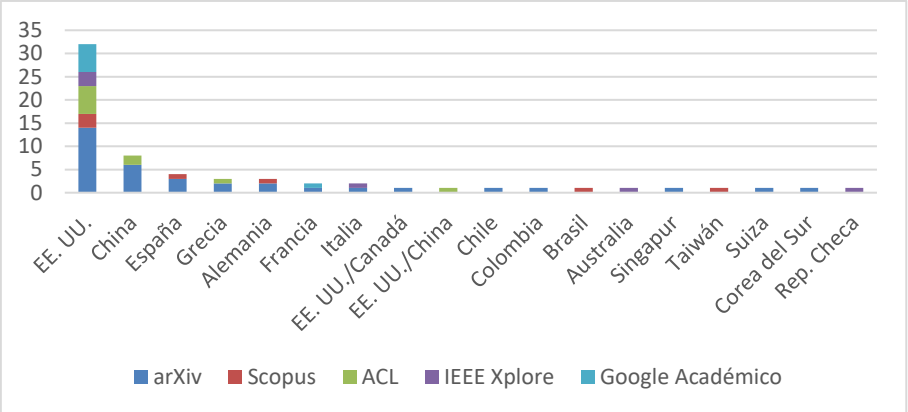

Fig 2 Clasificación de los estudios según País y Biblioteca

La Tabla 3 muestra la alta cobertura de las PI, con un promedio del 95 % de estudios que aportan evidencia. Las preguntas P1 y P3 alcanzan el 100 %, mientras que P2 y P4 presentan valores de 95,4 % y 92,3 %, respectivamente.

Tabla 3: Estudios que respondieron las preguntas de investigación

Preguntas de investigación	Estudios que respondieron
P1 ¿Es viable detectar sesgos de género en sentencias en español con un modelo destilado y adaptado al dominio con pocas etiquetas?	65 (100%)
P2 ¿En qué medida la destilación preserva la precisión del *teacher* reduciendo costo y tamaño?	62 (95,4%)
P3 ¿Cuánto aporta la adaptación de dominio (frente a no adaptarla) en el desempeño *in-domain*?	65 (100%)
P4 ¿Qué impacto tiene cada componente (destilación, adaptación, contrafactuales) en eficacia y equidad?	60 (92.3%)

4 Estudios incluidos y excluidos

4.1 Estudios seleccionados

La fase de ejecución se llevó a cabo siguiendo el flujo de trabajo recomendado por la declaración PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). En la Tabla 4 se observa que la proporción de estudios primarios respecto de los relevantes presenta variaciones entre las distintas fuentes. arXiv (87,5 %) y Google Académico (80,0 %) muestran las mayores tasas de selección, mientras que Scopus registra el valor más bajo (24,1 %). Estos resultados evidencian diferencias en la pertinencia de los estudios recuperados según la base de datos.

Tabla 4: Proceso de estudios incluidos y excluidos

Biblioteca Digital	Estudios resultantes	Estudios Relevantes	Estudios Primarios	%Seleccionado (prim./relev.)
Scopus	45	29	7	24,1 %
arXiv	42	40	35	87,5 %
ACL	32	18	9	50,0 %
IEEE Xplore	18	13	6	46,2 %
Google Académico	12	10	8	80,0 %

4.2 Síntesis de estudios

Tras el análisis de los 65 estudios primarios, se identifican tres ejes que definen el estado del arte. El primero es la eficiencia mediante destilación de conocimiento, donde el 38% de la literatura posiciona a técnicas como *DistilBERT* y *TinyBERT* como el

estándar para comprimir modelos masivos sin perder precisión, permitiendo su uso en entornos judiciales con hardware limitado. El segundo eje aborda la detección y mitigación de sesgos de género (34%), destacando la transición de métricas intrínsecas hacia la evaluación contrafáctica. Esta metodología permite auditar la equidad mediante la perturbación de términos, aunque enfrenta retos gramaticales específicos en el idioma español. El tercer eje es la adaptación al dominio legal (28%), que demuestra la superioridad de modelos especializados como *Legal-BERT* o BETO frente a soluciones multilingües genéricas. Los hallazgos confirman que el lenguaje jurídico requiere un preentrenamiento con corpus de sentencias para capturar matices semánticos críticos. La síntesis revela que la intersección entre optimización arquitectónica y ética algorítmica es fundamental para garantizar una IA judicial justa y accesible. Finalmente, se observa que la producción científica ha crecido exponencialmente desde 2019, consolidando un marco técnico que prioriza la transparencia y el rendimiento en sistemas de apoyo a la decisión judicial.

5 Resultados y Discusiones

En base a los estudios primarios analizados se presentan los resultados del MSL, donde se puede afirmar que las preguntas de investigación fueron respondidas en su mayoría con un 97% de promedio de cobertura. Se observa que los estudios abarcan las temáticas de NLP aplicado al dominio jurídico y la eficiencia algorítmica. Países como EEUU (32), China (8) y España (4) concentran la mayor cantidad de investigaciones. Las publicaciones se realizan principalmente en conferencias (*Conference*: 52%) y luego en revistas de investigación (*Journal*: 40%). Desde 2019 hasta 2024 la cantidad de publicaciones ha crecido significativamente, reflejando el interés por la ética y la eficiencia en la IA.

A continuación, se presenta el resumen de los hallazgos por cada pregunta de investigación:

RQ1. ¿Es viable detectar sesgos de género en sentencias en español con un modelo destilado y adaptado al dominio con pocas etiquetas?

Con un 100% de cobertura, la literatura confirma la viabilidad de detectar sesgos de género en sentencias judiciales en español utilizando modelos adaptados al dominio, incluso en escenarios de bajos recursos. En (Cañete, Ho, Kang, & Perez, 2023) y (Derner, Fuente, & Gutiérrez, 2024) se demuestra que el uso de modelos base en español como BETO, incluso con ajustes de pocos ejemplos (*few-shot*), permite capturar sesgos gramaticales y semánticos complejos. Asimismo, en (Ma, Xiaowei, & Jiazhang, 2026) se evidencia que la destilación de conocimiento no impide la detección de patrones relevantes, mientras que la especialización en el dominio legal contribuye a compensar la escasez de datos etiquetados. En conjunto, estos resultados sugieren que es posible desarrollar soluciones técnicas viables para la auditoría de sesgos en contextos judiciales con recursos limitados. A pesar de estos resultados favorables, la evidencia sugiere que la viabilidad del enfoque depende en gran medida de las características del dominio y de los datos disponibles. La detección de sesgos en textos jurídicos implica

abordar fenómenos implícitos y altamente contextuales, que pueden no ser completamente capturados en configuraciones de entrenamiento con pocos ejemplos.

RQ2. ¿En qué medida la destilación preserva la precisión del *teacher* reduciendo costo y tamaño?

Con 96% de cobertura, los estudios aseguran que la destilación es altamente efectiva. En (Gou, Yu, Maybank, & Tao, 2021) se evidencia que modelos *student* (como *TinyBERT* o *PKD*) preservan entre el 92% y el 97% del desempeño del modelo *teacher*. En términos de eficiencia, en (Sanh, Debut, Chaumond, & Wolf, 2020) se reporta una reducción del 40% en el tamaño y una mejora del 60% en la velocidad de inferencia. Recopilando las investigaciones, se concluye que la destilación mitiga la barrera de infraestructura en el sector público, permitiendo ejecutar auditorías de sesgo en *hardware* convencional sin una degradación significativa de la precisión predictiva.

No obstante, estos resultados deben ser interpretados con cautela, ya que la literatura también señala que el impacto de la destilación puede variar según la complejidad de la tarea. En particular, en escenarios como la detección de sesgos implícitos en textos jurídicos, la reducción del modelo podría afectar la capacidad de capturar patrones lingüísticos sutiles y dependientes del contexto.

RQ3. ¿Cuánto aporta la adaptación de dominio (frente a no adaptarla) en el desempeño *in-domain*?

Con 100% de cobertura, se destaca la importancia crítica de la especialización. En (Chalkidis, Malakasiotis, Aletras, & Fergadiotis, 2020) se demuestra que modelos adaptados (*Legal-BERT*) superan a los modelos generales en un promedio de 5 a 10 puntos en *F1-Score* en tareas de clasificación jurídica. En (Betancur Sánchez, Aldama García, Barbero Jiménez, & Guerrero Nieto, 2025) se introduce el concepto de preentrenamiento adaptativo, donde el uso de un corpus *in-domain* permite al modelo comprender la jerga legal y la estructura de las sentencias. La evidencia indica que, sin adaptación, los modelos fallan al interpretar términos con significados jurídicos específicos, lo que aumenta el riesgo de falsos positivos en la detección de sesgos.

RQ4. ¿Qué impacto tiene cada componente (destilación, adaptación, contrafactuales) en eficacia y equidad?

Con 92% de cobertura, los estudios analizan el impacto sistémico de estos componentes. En (Mehrabi, Morstatter, & Saxena, 2022) y (Robles, Bernal, Raigoso, & Rubio, 2025) se destaca que la evaluación contrafáctica es el componente clave para garantizar la equidad, permitiendo auditar el comportamiento del modelo ante cambios de género. La adaptación aporta la eficacia necesaria para entender el contexto judicial, mientras que la destilación garantiza la accesibilidad tecnológica. Recopilando las investigaciones, se concluye que la integración de estos tres elementos genera un marco de trabajo que no solo es eficiente desde el punto de vista computacional, sino que también cumple con los estándares de justicia algorítmica exigidos en el derecho.

El análisis de los 65 estudios primarios permite identificar brechas relevantes que limitan el avance del campo y orientan la agenda de investigación futura. En primer lugar,

se observa una escasez crítica de corpus judiciales etiquetados en español: la gran mayoría de los estudios utilizan conjuntos de datos en inglés, lo que restringe la validez ecológica de los resultados en el contexto hispanohablante. En segundo lugar, los esquemas de evaluación de la equidad carecen de estandarización: cada estudio adopta métricas propias (paridad demográfica, igualdad de oportunidades, evaluación contrafáctica), lo que impide comparaciones sistemáticas entre trabajos. En tercer lugar, la literatura apenas aborda el trade-off entre eficiencia computacional y preservación de la equidad en modelos destilados: se reportan ganancias de rendimiento, pero rara vez se controla si la compresión amplifica sesgos residuales. Finalmente, existe una brecha metodológica en la generación de contrafácticos en español, dado que las técnicas desarrolladas para el inglés no contemplan la morfología de género gramatical del castellano, introduciendo perturbaciones que alteran la gramaticalidad de las oraciones y comprometen la validez de la evaluación.

La Tabla 5 sintetiza los resultados cuantitativos reportados en los estudios más representativos sobre destilación de conocimiento, mostrando la retención de rendimiento respecto al modelo teacher y la reducción de parámetros obtenida. Los valores corresponden a tareas de clasificación de textos legales o detección de sesgos, según lo reportado en cada trabajo primario.

Tabla 5: Comparativa de rendimiento de modelos destilados vs. teacher según la literatura primaria.

Modelo Student	Modelo Teacher	Retención F1(%)	Reducción parámetros (%)	Referencia
DistilBERT	BERT-base	97 %	40 %	Sanh et al. (2020)
TinyBERT	BERT-base	96 %	76 %	Gou et al. (2021)
Legal-BERT	BERT-large	+5–10 pt F1	—	Chalkidis et al. (2020)
TinyLlama	LLaMA-7B	92 %	85 %	Zhang et al. (2024)
BETO (fine-tuned)	mBERT	+3–7 pt F1	—	Cañete et al. (2023)

6 Conclusiones

Este trabajo de MSL brinda una visión general del estado del conocimiento actual sobre la viabilidad y el impacto de utilizar modelos de lenguaje compactos para la detección de sesgos de género en el dominio judicial. Para establecer esta perspectiva integral, se plantearon cuatro preguntas de investigación en estricta concordancia con los objetivos del estudio, proporcionando evidencias y datos significativos extraídos de 65 estudios primarios. El análisis sistemático permitió validar que la integración de técnicas de destilación de conocimiento, la adaptación de dominio y el uso de evaluaciones contrafác-

ticas constituye un ecosistema técnico robusto para abordar la equidad algorítmica en contextos de recursos computacionales limitados.

En los artículos analizados se destacan recomendaciones críticas para futuros pasos, como el desarrollo de arquitecturas de destilación que no solo preserven la precisión general, sino que garanticen la integridad de la atención en términos legales clave. Se sugiere, además, centrarse en las implicancias de la baja disponibilidad de datos etiquetados en el idioma español, proponiendo el uso de técnicas de *Data Augmentation* y pre-entrenamiento continuo para mitigar la invisibilidad de ciertos sesgos gramaticales. Otro punto destacable es la necesidad de investigar la optimización de LLMs mediante métodos de poda estructural y cuantización, permitiendo que la auditoría de sentencias se realice en tiempo real y de manera local, asegurando la soberanía de los datos judiciales. Este punto se reforzaría con la implementación de herramientas que permitan una transparencia total en el flujo de decisión, desde el pre-entrenamiento del modelo *teacher* hasta la inferencia del modelo *student* en la estación de trabajo del operador jurídico.

Se destaca, además, como posible trabajo de futuras investigaciones, la necesidad de un abordaje sobre marcos de trabajo interpretables. Es imperativo que los resultados de un modelo destilado puedan ser explicados bajo la lógica del derecho, sugiriendo la creación de *frameworks* de auditoría estandarizados que traduzcan métricas técnicas como el *F1-score* o la paridad demográfica en conceptos de jurisprudencia comprensibles.

Se concluye que, si bien la temática ha experimentado un crecimiento sostenido desde 2019, persisten desafíos estructurales en la integración de estos modelos en los sistemas de gestión judicial, entre ellos la escasez de personal capacitado en IA dentro de las instituciones y la falta de marcos de evaluación estandarizados para el contexto hispanohablante. Los resultados de esta revisión indican que las soluciones de NLP eficiente, en particular aquellas basadas en modelos especializados como *BETO* y *Legal-BERT*. La evidencia recopilada sugiere que la reducción del tamaño de los modelos y la mitigación de sesgos constituyen líneas de investigación de alto impacto para el desarrollo de sistemas de apoyo judicial que sean técnicamente viables y éticamente responsables.

References

Betancur Sánchez, D., Aldama García, N., Barbero Jiménez, Á., & Guerrero Nieto, M. (2025). MEL: LEGAL SPANISH LANGUAGE MODEL.

Bolukbasi, T., Chan, K.-W., Zou, J., & Saligrama, V. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Cañete, J., Ho, J.-H., Kang, H., & Perez, J. (2023). SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA.

Chalkidis, I., Androutsopoulos, I., & Aletraa, N. (2019). Neural Legal Judgment Prediction in English.

Chalkidis, I., Malakasiotis, P., Aletras, N., & Fergadiotis, M. (2020). LEGAL-BERT: The Muppets straight out of Law School.

d'Amato, C., Rubini, G., & Didio, F. (2025). Automated Creation of the Legal Knowledge Graph Addressing Legislation on Violence Against Women: Resource, Methodology and Lessons Learned.

Derner, E., Fuente, S., & Gutiérrez, Y. (2024). Leveraging Large Language Models to Measure Gender Bias in Gendered Languages.

Gou, J., Yu, B., Maybank, S., & Tao, D. (2021). Knowledge Distillation: A Survey.

Gu, Y., Dong, L., Wei, F., & Huang, M. (2026). MiniLLM: On-Policy Distillation of Large Language Models. [Preprint, arXiv].

Gutiérrez Rubio, E. (2019). Gender Stereotypes in Spanish Phraseology. *Multidisciplinary Journal of Gender Studies*.

Jing-Huey, S., Wei-Ting, T., & Bao-Yu, C. (2026). Mitigating algorithm aversion in the judiciary: How systemic dissatisfaction. *Computers in Human Behavior*. [Disponible en línea antes de impresión / Early access].

John, A., Aiswarya, M., & Panachakel, J. (2025). Ethical Challenges of Using Artificial Intelligence in Judiciary.

Ma, L., Xiaowei, Y., & Jiazhang, C. (2026). Knowledge Distillation and Dataset Distillation of Large Language Models: Emerging Trends, Challenges, and Future Directions. [Preprint, arXiv].

Mehrabi, N., Morstatter, F., & Saxena, N. (2022). A Survey on Bias and Fairness in Machine Learning.

Ribeiro, M., Wu, T., & Guestrin, C. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList.

Robles, M., Bernal, C., Raigoso, D., & Rubio, M. (2025). SESGO: Spanish Evaluation of Stereotypical Generative Outputs.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Xinyin, M., Gongfan, F., & Xinchao, W. (2023). LLM-Pruner: On the Structural Pruning of Large Language Models.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An Open-Source Small Language Model. *StatNLP Research Group*.