UNIVERSIDAD NACIONAL DE MOQUEGUA

FACULTAD DE INGENIERÍAS Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERIA DE SISTEMAS E INFORMÁTICA

TESIS:

Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú

Presentado por:

Bach. María Del Pilar Laura Ramírez Chipana

Bach. Anthony Roger Flor Velásquez

Asesor:

Mgr. Hugo Euler Tito Chura

Co Asesor:

Ing. Honorio Apaza Alanoca

Moquegua, octubre de 2025

UNAM
UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Folio **Nº 028**

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis : [X]
Trabajo de Suficiencia Profesional : [ ]
Otro: ..... : [ ]

II. Documento de aprobación para la sustentación:

Resolucion Consejo de Facultad 290-2025-CF-FAIN-UNAM

Fecha: 22-12-2025

III. El jurado ha dictaminado la (el) tesis, TITULADO:

Sistema de recomendacion de jurisprudencia basado en embeddings de texto para la identificacion y sugerencia de precedentes relevantes de las cortes superiores de justicia de republica del Perú.

Bachiller: Maria del Pilar Laura Ramirez Chipana

Asesor Principal: Hugo Euler Tito Chora

Co-Asesor: Honorio Apaza Alanoca

IV. Se dictamina:

Como: aprobado por unanimidad la sustentación de manera:

Presencial [X] Virtual [ ]

Con nota: dieciseis (16) por lo que; Sí [X] No [ ]

Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA.

Siendo las 13:00 horas, del 22 de diciembre del año 2025 se procede con la firma en conformidad de lo dictaminado por el Jurado.

Carlos Asselvo Delgado
Presidente

Victor Yanz Mamani
Primer Miembro

Juan Carlos Claros Loza
Segundo Miembro

Maria del Pilar Laura Ramirez Chipana
Bachiller

Folio Nº 027

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis : [X]
Trabajo de Suficiencia Profesional : [ ]
Otro: ..... : [ ]

II. Documento de aprobación para la sustentación:

Resolucion Consejo de Facultad 290-2025-CF-UNAM

Fecha: 22.12.2025

III. El jurado ha dictaminado la (el) Tesis, TITULADO:

Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las cortes superiores de justicia de la republica del Peru

Bachiller: Anthony Roger Flor Velasquez

Asesor Principal: Hugo Euler Tito Chora

Co-Asesor: Honorio Apaza Alanoca

IV. Se dictamina:

Como: aprobado por unanimidad la sustentación de manera:

Presencial [X] Virtual [ ]

Con nota: dieciseis (16) por lo que; Sí [X] No [ ]

Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA .

Siendo las 13:00 horas, del 22 de diciembre del año 2025 se procede con la firma en conformidad de lo dictaminado por el Jurado.

Presidente

Victor Yana Mamani
Primer Miembro

Juan Carlos Clarel Perca
Segundo Miembro

Anthony Roger Flor Velasquez
Bachiller

Agradecimiento

Con profundo respeto y gratitud, extendemos nuestro reconocimiento a todas las personas e instituciones que hicieron posible el desarrollo de esta investigación y la consolidación de nuestra formación académica y profesional.

A la Universidad Nacional de Moquegua, y de manera especial a la Escuela Profesional de Ingeniería de Sistemas e Informática, por brindarnos una formación integral, el acompañamiento constante y los recursos necesarios para la realización de esta investigación.

A nuestros padres, por su amor incondicional, comprensión y apoyo permanente. Su ejemplo, paciencia y sacrificio fueron fundamentales para mantenernos firmes en este proceso. Este logro es el resultado de su esfuerzo tanto como del nuestro.

A nuestros docentes y asesores, por su orientación, compromiso y dedicación, que guiaron nuestro aprendizaje y fortalecieron el desarrollo de esta investigación.

Finalmente, a nuestros amigos y compañeros, por su apoyo, colaboración y palabras de aliento durante todo este camino académico. A todos ellos, nuestra más sincera gratitud.

Resumen

La recuperación eficiente de jurisprudencia relevante es un desafío fundamental en el ámbito legal. Esta investigación tiene como objetivo principal desarrollar y evaluar comparativamente la efectividad de tres enfoques algorítmicos para la recomendación de resoluciones judiciales, aplicados a un corpus de jurisprudencia laboral peruana.

Se utilizó un corpus de 6918 resoluciones de apelación en materia laboral (período 2021-2022) de la Corte Superior de Justicia de Lima, dividiéndolo en 6848 documentos para entrenamiento/indexación y 30 para evaluación. Los modelos implementados y comparados fueron: 1) Búsqueda Léxica (Okapi BM25), 2) Búsqueda Semántica (SentenceTransformer, bert-base-multilingual-cased), y 3) Búsqueda Híbrida (Reciprocal Rank Fusion - RRF).

La evaluación se realizó comparando las 5 mejores recomendaciones (Top-5) de cada modelo contra un "Gold Standard" de relevancia binaria, anotado manualmente por el investigador mediante un sistema de criterios heurísticos estructurados. Se emplearon métricas estándar de recuperación de información, incluyendo Precisión@5 (P@5) y NDCG@5.

Los resultados cuantitativos demuestran una superioridad significativa del modelo léxico. Okapi BM25 alcanzó el rendimiento más alto (P@5=0.8400; NDCG@5=0.8604), superando ampliamente al modelo híbrido (P@5=0.6133; NDCG@5=0.5943) y al modelo semántico puro (P@5=0.3600; NDCG@5=0.4001).

Se concluye que, para la tarea específica de recomendación basada en resúmenes de jurisprudencia laboral peruana, los métodos léxicos tradicionales que ponderan la frecuencia terminológica (BM25) son, en este contexto, más precisos y efectivos que los enfoques semánticos puros pre-entrenados o su combinación híbrida mediante RRF. Finalmente, se implementó un prototipo local funcional con Flask para la demostración interactiva del sistema.

Palabras clave: Sistemas de Recomendación, Recuperación de Información Jurídica, Jurisprudencia, BM25, SentenceTransformer, Búsqueda Híbrida, Procesamiento de Lenguaje Natural (PLN), Derecho Laboral.

Abstract

The efficient retrieval of relevant jurisprudence is a fundamental challenge in the legal field. This research's main objective is to develop and comparatively evaluate the effectiveness of three algorithmic approaches for recommending judicial resolutions, applied to a corpus of Peruvian labor jurisprudence.

A corpus of 6918 labor appeal resolutions (2021-2022) from the Lima Superior Court of Justice was used, divided into 6848 documents for training/indexing and 30 for evaluation. The implemented and compared models were: 1) Lexical Search (Okapi BM25), 2) Semantic Search (SentenceTransformer, bert-base-multilingual-cased), and 3) Hybrid Search (Reciprocal Rank Fusion - RRF).

The evaluation was conducted by comparing the Top-5 recommendations from each model against a binary relevance "Gold Standard," manually annotated by the researcher using a structured heuristic criteria system. Standard information retrieval metrics, including Precision@5 (P@5) and NDCG@5, were employed.

The quantitative results demonstrate a significant superiority of the lexical model. Okapi BM25 achieved the highest performance (P@5=0.8400; NDCG@5=0.8604), far surpassing the hybrid model (P@5=0.6133; NDCG@5=0.5943) and the pure semantic model (P@5=0.3600; NDCG@5=0.4001).

It is concluded that, for the specific task of recommendation based on Peruvian labor jurisprudence summaries, traditional lexical methods that weigh term frequency (BM25) are, in this context, more accurate and effective than pre-trained pure semantic approaches or their hybrid combination via RRF. Finally, a functional local prototype was implemented with Flask for interactive demonstration of the system.

Keywords: Recommender Systems, Legal Information Retrieval, Jurisprudence, BM25, SentenceTransformer, Hybrid Search, Natural Language Processing (NLP), Labor Law.

Tabla de contenido

Introducción 11

I. Planteamiento del problema 13

1.1 Descripción y formulación del problema 13

1.1.1 Interrogante General 16

1.1.2 Interrogantes específicas 16

1.2 Objetivos 17

1.2.1 Objetivo General 17

1.2.2 Objetivos Específicos 17

1.3 Justificación e importancia 18

1.4 Hipótesis 20

1.4.1 Hipótesis General 20

1.4.2 Hipótesis Específicas 20

1.5 Variables 21

1.5.1 Variable independiente 21

1.5.2 Variable dependiente 21

II. Marco Teórico 22

2.1 Antecedentes del Problema 22

2.2 Bases Teóricas 24

2.2.1 Introducción y Contextualización 24

2.2.2 El Sistema Judicial en el Perú y la Corte Superior de Justicia 25

2.2.3 Procesamiento de Lenguaje Natural 37

2.2.4 Modelos de Embeddings de Texto 39

2.2.5 Sistemas de Recomendación en el Ámbito Jurídico 50

2.3 Definición de términos 55

III. Metodología 57
3.1 Tipo y nivel de investigación 57
3.1.1 Tipo de investigación 57
3.1.2 Nivel de investigación 57
3.2 Ámbito temporal y espacial 57
3.2.1 Población 58
3.2.2 Muestra 58
3.3 Instrumentos 59
3.4 Procedimientos 60
3.5 Análisis de datos 63
IV. Resultados y discusión 64
4.1 Construcción de corpus 64
4.1.1 Recuperación de documentos legales 64
4.1.2 Exploración de los datos 67
4.1.3 Implicaciones para el Sistema de Recomendación 70
4.2 Preprocesamiento de los datos 72
4.2.1 Estandarización de metadatos categóricos 72
4.2.2 Limpieza y normalización del contenido textual 73
4.3 Desarrollo de sistema de recomendación 74
4.3.1 Modelo 1: Búsqueda semántica basada en transformers (SentenceTransformer) 74
4.3.2 Modelo 2: Búsqueda léxica con Okapi BM25) 79
4.3.3 Modelo 3: Búsqueda híbrida y fusión de rankings (Reciprocal Rank Fusion) 84
4.4 Evaluación y validación 87

4.4.1 Métricas de evaluación de rendimiento 89
4.4.2 Selección de consultas (Queries) para evaluación 93
4.4.3 Generación de documentos candidatos recomendados 95
4.4.4 Proceso de anotación manual de relevancia (Creación del Gold Standard) 97
4.4.5 Estructura y uso del conjunto de datos de evaluación final 99
4.5 Resultados 100
4.5.1 Modelo léxico (BM25) 101
4.5.2 Modelo semántico (SentenceTransformer) 101
4.5.3 Modelo híbrido (RRF) 102
4.6 Prototipo 103
4.6.1 Arquitectura Tecnológica del Prototipo 103
V. Conclusiones 105
5.1 Superioridad del modelo léxico (BM25) 105
5.2 Limitaciones del modelo semántico puro (SentenceTransformer) 105
5.3 Rendimiento Intermedio del Modelo Híbrido (RRF) 106
5.4 Importancia de la coincidencia léxica en el dominio legal laboral 106
5.5 Viabilidad de prototipo local con Flask 106
VI. Recomendaciones 107
6.1 Priorizar Modelos Léxicos (BM25) para Aplicaciones 107
6.2 Explorar Técnicas Avanzadas para Modelos Híbridos 107
6.3 Investigar el Uso de Texto Completo vs. Resúmenes 108
6.4 Considerar la Incorporación de Metadatos 108
VII. Referencias Bibliográficas 109

Lista de Figuras

Figura 1 (BERT: Preentrenamiento de transformadores bidireccionales profundos para la comprensión del lenguaje)..... 45

Figura 2 (Arquitectura de Modelo Transformer) 47

Figura 3 (Metodología de investigación) 60

Figura 4: Jurisprudencia Nacional Sistematizada 65

Figura 5 Corpus de jurisprudencias..... 68

Figura 6 Distribución de materias 69

Figura 7 Distribución de textos en campos texto_completo y Texto_resumido... 70

Figura 8 Muestra de datos de la prueba (Test Set)..... 95

Figura 9 Características de los datos de prueba 95

Figura 10 Recomendaciones generadas por los 3 modelos..... 96

Figura 11 Características de recomendaciones 97

Figura 12 Resultados de anotación manual de relevancia 99

Figura 13 Comparación de los resultados de precisión de los modelos..... 102

Figura 14 Interfaz del prototipo de recomendación de jurisprudencias..... 104

Lista de Tablas

Tabla 1 (Ejemplo de Matriz de Bag of Words Binaria) 39

Tabla 2 (Ejemplo de cálculo TF, IDF y TF*IDF para dos documentos con términos jurídicos) 41

Tabla 3 Resultados de precisión de los modelos..... 101

Introducción

En los últimos años, el acelerado proceso de digitalización en el ámbito judicial ha generado un volumen creciente de información jurisprudencial que, si bien amplía el acceso al conocimiento jurídico, también plantea nuevos desafíos en cuanto a su organización, búsqueda y aprovechamiento. Los tribunales de justicia producen miles de resoluciones cada año, muchas de las cuales contienen fundamentos relevantes que podrían servir como precedentes en la resolución de casos similares. Sin embargo, la ausencia de herramientas tecnológicas que permitan identificar y recomendar eficientemente dichos precedentes limita la eficacia y uniformidad de las decisiones judiciales.

En este contexto, la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN) y de aprendizaje automático se ha convertido en una alternativa prometedora para optimizar la gestión y recuperación de información jurídica. En particular, los modelos de embeddings de texto permiten representar documentos legales en vectores numéricos capaces de capturar su significado semántico, facilitando la búsqueda de similitudes entre resoluciones judiciales. Esta línea de investigación ha mostrado resultados alentadores en diversos países, donde se han desarrollado sistemas de recomendación y clasificación de sentencias basados en modelos como Word2Vec, BERT o Doc2Vec, con el fin de mejorar el acceso a la jurisprudencia y la coherencia de las decisiones judiciales.

No obstante, en el contexto peruano, el desarrollo de este tipo de soluciones tecnológicas sigue siendo incipiente. Aunque el Poder Judicial ha impulsado iniciativas de digitalización, como el Expediente Judicial Electrónico (EJE) o la plataforma de Jurisprudencia Nacional Sistematizada, aún no se cuenta con sistemas automatizados que integren técnicas avanzadas de inteligencia artificial para la recomendación de precedentes judiciales. Esta carencia tecnológica no solo dificulta el trabajo de magistrados y asistentes judiciales, sino que también retrasa la modernización del sistema de justicia y el aprovechamiento del conocimiento jurisprudencial acumulado.

En respuesta a esta problemática, la presente investigación propone el desarrollo de un sistema de recomendación de jurisprudencia basado en

embeddings de texto, orientado a la identificación y sugerencia de precedentes relevantes a partir de un corpus de resoluciones emitidas por las Cortes Superiores de Justicia de la República del Perú. Además, se plantea la evaluación comparativa del desempeño de distintos enfoques algorítmicos —léxicos, semánticos e híbridos— con el propósito de determinar cuál de ellos ofrece los mejores resultados en términos de precisión y pertinencia de las recomendaciones.

La investigación se estructura en cuatro etapas principales: construcción del corpus jurisprudencial, preprocesamiento de datos textuales, desarrollo del sistema de recomendación mediante distintos modelos de embeddings, y evaluación de su desempeño mediante métricas de recuperación de información. Finalmente, se implementa un prototipo funcional que permite demostrar la aplicabilidad del modelo más eficiente en un entorno web desarrollado con Flask.

De esta manera, el estudio busca contribuir al fortalecimiento del sistema judicial peruano mediante el uso de tecnologías de inteligencia artificial, promoviendo una gestión más eficiente del conocimiento jurídico y una administración de justicia más coherente, accesible y basada en evidencia.

I. Planteamiento del problema

1.1 Descripción y formulación del problema

A nivel global se ha evidenciado una creciente digitalización de los sistemas jurídicos, motivada por la necesidad de mejorar el acceso a la justicia y a la información legal. Esta transformación ha generado una acumulación masiva de jurisprudencia y decisiones judiciales, lo que ha derivado en un vasto volumen de información, cada vez más complejo y heterogéneo. Este exceso de datos plantea desafíos importantes en términos de búsqueda, recuperación y análisis de información jurídica, afectando la eficiencia y la equidad en la administración de justicia. En ese sentido, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) advierte que la gobernanza efectiva de los datos judiciales enfrenta obstáculos estructurales derivados de la sobrecarga informativa y de la falta de mecanismos adecuados para procesar y aprovechar el conocimiento jurídico disponible (Byrom et al., 2024).

A nivel internacional, diversos países enfrentan retos comunes en el acceso eficiente a jurisprudencia relevante debido al volumen creciente de decisiones judiciales. En Brasil, la sobrecarga del sistema judicial ha impulsado investigaciones centradas en la automatización del análisis de sentencias. Mentzingen et al. (2024) demostraron que, ante la dificultad de extraer conceptos relevantes de grandes volúmenes de datos legales, las representaciones textuales granulares resultan más eficaces. Otro estudio en el Tribunal de Justicia de Rio Grande do Norte, empleó un corpus de 30,000 documentos y evidenció que técnicas tradicionales como TF-IDF pueden superar a modelos complejos como BERT en ciertas tareas de clasificación, superando incluso todas las métricas consideradas (Dantas, 2023).

En Colombia, la ineficiencia del sistema judicial caracterizada por la alta congestión de procesos y la escasez de funcionarios ha sido abordada mediante técnicas de transfer learning y BERT para la clasificación y búsqueda de textos legales, utilizando un corpus de 23,750 documentos de la “Corte Constitucional” y métodos como FAIIS, distancia euclidiana y KNN, lo que ha permitido mejorar la calidad de las búsquedas y recomendaciones, logrando mejores resultados que los

enfoques tradicionales (Martínez, 2022). De igual forma, en el contexto de la ley filipina se han comparado técnicas de embedding como Doc2Vec, TF-IDF y OpenAI embedding (textembedding-ada-002) utilizando un conjunto de 4,400 decisiones del Tribunal Supremo, lo que ha permitido obtener precisiones destacadas en la clasificación de similitud y resaltar el potencial de estas técnicas para optimizar la investigación legal (Ruiz, 2021)

En países de habla árabe, se han empleado arquitecturas basadas en doc2vec para la extracción de principios de jurisprudencia (JPs) evidenciando que el enfoque PV-DBOW es especialmente eficaz (Alrumayyan & Al-Yahya, 2022), así mismo que la aplicación de técnicas de embeddings tiene gran potencial aún en distintos contextos geográficos y lingüísticos.

Estas experiencias internacionales evidencian que la integración de técnicas de IA y NLP en el ámbito jurídico ha permitido alcanzar logros notables como la automatización de la identificación de precedentes, la mejora en la clasificación de textos y la reducción del tiempo de búsqueda, aunque aún persisten desafíos técnicos y prácticos relacionados con la adaptación de modelos a dominios específicos, la interpretabilidad de los resultados y la gestión de sesgos en los datos.

A nivel nacional, el sistema judicial peruano enfrenta dificultades estructurales en el acceso oportuno y eficiente a la jurisprudencia relevante. A pesar de los esfuerzos iniciales hacia la digitalización, como la implementación del expediente judicial electrónico (EJE), aún no se ha desarrollado una infraestructura tecnológica adecuada que permita automatizar la identificación de precedentes judiciales. La aplicación de técnicas de inteligencia artificial como BERT, redes siamesas o modelos de recomendación en este ámbito sigue siendo inexistente en investigaciones aplicadas en el contexto peruano, lo que evidencia una brecha importante entre el avance tecnológico global y las capacidades locales. Esta situación se ve agravada por la falta de articulación normativa sobre el uso ético y técnico de estas herramientas, así como por la escasa interoperabilidad entre bases de datos jurídicas y por las desigualdades de acceso en zonas rurales. En conjunto, estas limitaciones obstaculizan la modernización del sistema judicial y restringen el aprovechamiento del conocimiento jurisprudencial acumulado.

A nivel local, el problema se manifiesta en el trabajo diario de las Cortes Superiores de Justicia del país, cuyas unidades jurisdiccionales deben revisar un elevado volumen de jurisprudencia sin contar con herramientas automatizadas que faciliten la identificación de precedentes relevantes. Esta situación genera sobrecarga laboral, inconsistencias en la argumentación jurídica y dificultades para garantizar decisiones fundamentadas de manera uniforme. A pesar de los esfuerzos del Poder Judicial en digitalizar expedientes y promover la innovación, ninguna corte dispone actualmente de un sistema que utilice técnicas como embeddings de texto para recomendar jurisprudencia. Esta carencia tecnológica repercute directamente en la eficiencia y calidad del servicio judicial ofrecido en el territorio nacional.

Frente a lo expuesto, la falta de herramientas basadas en modelos de lenguaje y embedding de texto limita significativamente la eficiencia en la búsqueda de precedentes judiciales en el contexto peruano. Ante esta situación, se formula la siguiente pregunta de investigación.

1.1.1 Interrogante General

¿Qué modelo basado en embeddings de texto presenta el mejor rendimiento en un sistema de recomendación para la identificación y sugerencia de precedentes relevantes, a partir de un corpus de jurisprudencias de las Cortes Superiores de Justicia de la República del Perú?

1.1.2 Interrogantes específicas

- ¿Cómo crear y consolidar un corpus de jurisprudencias para el entrenamiento de un sistema de recomendación basado en embeddings textuales, aplicado a jurisprudencias de las Cortes Superiores de Justicia de la República del Perú?
- ¿Cómo influye el proceso de preprocesamiento en el rendimiento de un sistema de recomendación de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú?
- ¿Qué modelos de recomendación basados en embeddings son más eficaces para identificar casos similares y precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú?
- ¿Qué métricas permiten evaluar el rendimiento de un sistema de recomendación de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú?
- ¿Cómo construir un prototipo para un sistema de recomendación que incorpore el modelo más eficiente, utilizando frameworks web en Python?

1.2 Objetivos

1.2.1 Objetivo General

Desarrollar un sistema de recomendación de jurisprudencia basado en embeddings de texto, que permita identificar y sugerir precedentes relevantes a partir de un corpus de jurisprudencias de las Cortes Superiores de Justicia de la República del Perú.

1.2.2 Objetivos Específicos

- Crear y consolidar un corpus de jurisprudencias para el entrenamiento de un sistema de recomendación basado en embeddings textuales, aplicado a jurisprudencias de las Cortes Superiores de Justicia de la República del Perú.
- Preprocesar y normalizar el corpus de jurisprudencias de las Cortes Superiores de Justicia de la República del Perú mediante técnicas de limpieza, tokenización, lematización y otras transformaciones necesarias para su análisis.
- Desarrollar e implementar modelos de recomendación basados en embeddings de texto para identificar y sugerir precedentes relevantes de las jurisprudencias de las Cortes Superiores de Justicia de la República del Perú.
- Evaluar el rendimiento del sistema de recomendación basados en embeddings de texto para identificar y sugerir precedentes relevantes de las jurisprudencias de las Cortes Superiores de Justicia de la República del Perú mediante métricas de evaluación.
- Construir un prototipo de Sistema de Recomendación, integrando el modelo de más alto rendimiento, sobre una base tecnológica de web frameworks para python.

1.3 Justificación e importancia

La creciente digitalización de los sistemas jurídicos a nivel global ha resultado en una acumulación masiva de jurisprudencia y decisiones judiciales. Este vasto y complejo volumen de información plantea desafíos significativos en la búsqueda, recuperación y análisis de datos legales. Históricamente, la identificación manual de precedentes, que es crucial para garantizar una jurisprudencia coherente, se ha convertido en un cuello de botella debido al creciente número de casos y decisiones. Esta situación impulsa la necesidad imperante de automatizar este proceso para mejorar la eficiencia y la uniformidad en la aplicación de la ley.

En este contexto, la integración de técnicas avanzadas de Inteligencia Artificial (IA) y Procesamiento de Lenguaje Natural (PLN) ha demostrado ser una solución prometedora. Modelos preentrenados como BERT y sus variantes como RoBERTa, han revolucionado la extracción y representación de información semántica, permitiendo obtener representaciones vectoriales ricas en significado. Diversos estudios han explorado la automatización para identificar casos similares que puedan interpretarse como precedentes, mostrando que los modelos con representaciones textuales granulares son óptimos para extraer conceptos y relaciones.

La aplicación de técnicas de embeddings ha demostrado su potencial en varios contextos geográficos y lingüísticos dentro del ámbito jurídico. Estos avances permiten transformar resoluciones judiciales y otros documentos legales en representaciones numéricas, facilitando la medición de la similitud semántica entre textos. La irrupción del PLN también ha propiciado la proliferación de herramientas que automatizan desde la redacción de contratos hasta la búsqueda de jurisprudencia y el desarrollo de chatbots legales, potenciando significativamente el servicio jurídico.

Aunque la integración de IA y PLN ha logrado avances notables en la automatización de la identificación de precedentes y la clasificación de textos, aún persisten desafíos técnicos y prácticos relacionados con la adaptación de modelos a dominios específicos, la interpretabilidad de los resultados y la gestión de sesgos en los datos.

En este marco, la presente tesis tiene como objetivo desarrollar un sistema de recomendación de jurisprudencia basado en embeddings de texto. Este sistema permitirá identificar y sugerir precedentes legales relevantes a partir de un corpus de resoluciones judiciales de la Cortes Superiores del Perú. Al reducir el tiempo invertido en la búsqueda de información y mejorar la calidad de las decisiones judiciales mediante el apoyo de herramientas tecnológicas innovadoras, este proyecto busca facilitar el trabajo de los profesionales del derecho.

Esta investigación se enmarca en el ámbito de la inteligencia artificial aplicada al derecho y se justifica por la necesidad apremiante de modernizar y optimizar los procesos de análisis de información legal. Al integrar técnicas de PLN y modelos de embeddings en la práctica jurídica, se espera contribuir a la construcción de herramientas que apoyen la toma de decisiones y promuevan una jurisprudencia más coherente y eficiente en el sistema judicial peruano.

1.4 Hipótesis

1.4.1 Hipótesis General

Un sistema de recomendación basado en embeddings de texto permite identificar y sugerir precedentes jurídicos relevantes a partir de un corpus de jurisprudencias de las Cortes Superiores de Justicia de la República del Perú.

1.4.2 Hipótesis Específicas

- La creación y consolidación de un corpus de jurisprudencias de las Cortes Superiores de Justicia de la República del Perú proporciona una base sólida para el entrenamiento de modelos de recomendación basados en embeddings textuales.
- El preprocesamiento y la normalización del corpus, mediante técnicas como limpieza, tokenización y lematización, mejoran significativamente el rendimiento del sistema de recomendación.
- Los modelos de recomendación basados en embeddings de texto permiten identificar y sugerir precedentes jurídicos relevantes de manera eficaz, en comparación con enfoques tradicionales
- El rendimiento de los sistemas de recomendación de jurisprudencia puede ser evaluado objetivamente mediante métricas de evaluación específicas
- La construcción de un prototipo funcional utilizando frameworks web para Python permite integrar eficazmente el modelo de mayor rendimiento en un entorno accesible y aplicable para consultas de jurisprudencia.

1.5 Variables

1.5.1 Variable independiente

Embeddings de Texto

1.5.2 Variable dependiente

Identificación y sugerencia de Precedentes relevantes

II. Marco Teórico

2.1 Antecedentes del Problema

El desarrollo de sistemas automáticos para la identificación de precedentes legales se sustenta en décadas de avances en el análisis de textos y la inteligencia artificial. Inicialmente, los métodos tradicionales basados en técnicas de extracción de características, como TF-IDF y modelos de bolsa de palabras, fueron aplicados al análisis de documentos legales. Sin embargo, estos enfoques se mostraron limitados al no capturar de manera adecuada la semántica y las complejidades inherentes al lenguaje jurídico.

La irrupción de los modelos de Deep learning, en particular las arquitecturas basadas en Transformers como BERT (Devlin et al., 2018) y sus optimizaciones como RoBERTa (Liu et al., 2019), marcó un hito en el procesamiento del lenguaje natural. Estos modelos permitieron obtener representaciones vectoriales ricas en información semántica y sintáctica, posibilitando aplicaciones más precisas en ámbitos tan exigentes como el jurídico (Getman Oleg M. Yaroshenko, 2021) (González-Moreno, 2023).

Posteriores investigaciones se centraron en adaptar estos avances a contextos legales específicos. Se exploraron técnicas de embeddings a nivel de documento, como *doc2vec*, para la elicitación de principios jurídicos en textos árabes, así como la aplicación de métodos de embedding en decisiones de tribunales, evidenciando resultados prometedores en jurisdicciones tan diversas como Filipinas (Ruiz, 2021).

Paralelamente, se han desarrollado enfoques que combinan modelos clásicos con arquitecturas neuronales avanzadas, tales como redes neuronales siamesas y modelos basados en grafos de conocimiento, orientados a medir la similitud entre descripciones de hechos y a reducir la carga de trabajo de los profesionales del derecho (Ruiz, 2021)(Shao, 2024) . Andrade & Becker (2023) proponen un enfoque híbrido que combina BM25 y embeddings basados en BERT para la sumarización de documentos legales, demostrando una mejora significativa en la identificación de información relevante y en la calidad de los resúmenes jurídicos, especialmente en términos de precisión y cobertura. La integración de

técnicas de visual analytics, como se observa en el sistema LegalVis, y la aplicación de estrategias de interpretabilidad en ML han abierto nuevos caminos para una toma de decisiones más informada en entornos judiciales (Resck et al., 2023).

Adicionalmente, el creciente interés en el uso de IA en el sector legal ha impulsado a la comunidad investigadora a abordar desafíos de eficiencia, coherencia y accesibilidad en la búsqueda de precedentes, al mismo tiempo que se reconocen las limitaciones actuales en la adaptación de modelos genéricos a dominios especializados y la necesidad de supervisión experta en la validación de los resultados(Fuentes González, 2021),(Martínez, 2022). Estos antecedentes configuran el estado actual de la investigación en sistemas de recomendación de jurisprudencia y orientan las líneas futuras de desarrollo en la intersección de la tecnología y el derecho.

2.2 Bases Teóricas

2.2.1 Introducción y Contextualización

La transformación digital ha revolucionado múltiples ámbitos del conocimiento, y el sector jurídico no es la excepción. En los últimos años, la creciente digitalización de los sistemas judiciales ha generado un volumen sin precedentes de resoluciones y jurisprudencias, lo que, si bien facilita el acceso a la información, también plantea desafíos significativos en términos de búsqueda, análisis y utilización de datos legales. En particular, la identificación de precedentes relevantes se ha convertido en un proceso laborioso y dependiente de la experiencia del profesional, lo que ralentiza la toma de decisiones y afecta la eficiencia de la administración de justicia.

En este contexto, el uso de técnicas avanzadas de *Procesamiento de Lenguaje Natural* (PLN) e *Inteligencia Artificial* (IA) ofrece nuevas oportunidades para automatizar y optimizar el análisis de grandes corpus textuales. Los modelos de embeddings de texto, que permiten representar documentos en espacios vectoriales donde se capturan relaciones semánticas, han demostrado su potencial en diversas aplicaciones, desde la búsqueda de información hasta la clasificación y recomendación de contenidos. Sin embargo, su aplicación en el ámbito jurídico, y en particular en la identificación de precedentes a partir de resoluciones judiciales, aún presenta desafíos que requieren un análisis profundo y adaptaciones específicas.

La presente tesis tiene como objetivo desarrollar un sistema de recomendación de jurisprudencia que utilice embeddings de texto para identificar y sugerir precedentes legales relevantes a partir de un corpus de resoluciones judiciales de la Cortes Superiores. Este sistema pretende facilitar el trabajo de los profesionales del derecho, reduciendo el tiempo invertido en la búsqueda de información y mejorando la calidad de las decisiones judiciales mediante el apoyo de herramientas tecnológicas innovadoras.

Para alcanzar este objetivo, se ha diseñado una metodología que abarca la recolección y preprocesamiento de datos, el desarrollo e implementación de

modelos de recomendación, y la evaluación del desempeño del sistema utilizando métricas específicas.

Esta investigación se enmarca en el ámbito de la inteligencia artificial aplicada al derecho, y se justifica por la necesidad de modernizar y optimizar los procesos de análisis de información legal. Al integrar técnicas de PLN y modelos de embeddings en la práctica jurídica, se espera contribuir a la construcción de herramientas que apoyen la toma de decisiones y promuevan una jurisprudencia más coherente y eficiente.

En las siguientes secciones se detallarán los fundamentos teóricos, la metodología y los resultados obtenidos, permitiendo evaluar el potencial de esta propuesta en el contexto de la administración de justicia en Perú.

2.2.2 El Sistema Judicial en el Perú y la Corte Superior de Justicia

El sistema judicial peruano es un sistema jerárquico, organizado y diseñado para garantizar la administración de justicia de manera eficiente, objetiva e imparcial. En este sistema, la Corte Suprema de Justicia es la máxima autoridad judicial en el país, seguida por las Cortes Superiores de Justicia, que se encargan de administrar justicia en las diversas regiones del Perú.

El acceso a la justicia es un derecho fundamental en el país, y el sistema judicial está regido por la Constitución Política del Perú, así como por un conjunto de leyes y normativas que aseguran la igualdad ante la ley y el derecho a la defensa.

Este marco legal busca establecer un equilibrio entre el respeto a los derechos fundamentales y la eficiencia en la resolución de conflictos. En este contexto, la Corte Superior de Justicia juega un papel crucial, ya que es la encargada de la resolución de causas en primera instancia dentro de cada jurisdicción regional.

2.2.2.1 El Sistema Judicial Peruano

El sistema judicial del Perú se basa en una estructura jerárquica que tiene como máxima autoridad a la Corte Suprema de Justicia, que es la instancia judicial más alta del país. A nivel regional, las Cortes Superiores de Justicia operan en las

principales ciudades del Perú y están encargadas de resolver los casos judiciales en primera instancia.

El sistema judicial peruano está compuesto por los siguientes órganos principales:

- **La Corte Suprema de Justicia:**

La Corte Suprema de Justicia es la máxima instancia judicial en el Perú y tiene la responsabilidad de garantizar la correcta aplicación de las leyes a nivel nacional. Esta corte está compuesta por 12 jueces supremos, distribuidos en diferentes salas especializadas, como la Sala Penal, Civil, Constitucional, y otras. La Corte Suprema actúa como un tribunal de casación, revisando las sentencias emitidas por las Cortes Superiores de Justicia y otros tribunales inferiores. Esta instancia también se encarga de resolver los casos más complejos o trascendentales para la sociedad, que pueden tener implicancias constitucionales o derechos fundamentales en juego.

Además de su función judicial, la Corte Suprema tiene un papel de supervisión y coordinación de las Cortes Superiores de Justicia, asegurando que los criterios jurídicos y las interpretaciones legales sean coherentes en todo el país. En casos excepcionales, también puede intervenir en la interpretación de la Constitución y garantizar su cumplimiento. La Corte Suprema también tiene la capacidad de emitir jurisprudencia vinculante, que debe ser seguida por otras instancias judiciales del país.

- **Las Cortes Superiores de Justicia:**

Las Cortes Superiores de Justicia son órganos judiciales de segunda instancia, encargados de resolver los casos que hayan sido apelados desde los juzgados de primera instancia dentro de una región geográfica determinada. Cada Corte Superior se organiza en salas especializadas, las cuales se encargan de revisar casos en diferentes ramas del derecho. Las salas más comunes son:

    - **Sala Penal:**

La Sala Penal es una de las más relevantes dentro de las Cortes Superiores de Justicia, ya que se encarga de revisar casos relacionados con delitos y crímenes cometidos dentro de su jurisdicción. Estos casos incluyen desde delitos menores hasta crímenes graves, como homicidios, violaciones, delitos económicos, y crímenes organizados. La función principal de esta sala es garantizar la legalidad y justicia en el sistema penal, asegurando que las sentencias emitidas en primera instancia sean justas, fundamentadas en pruebas legales y en el debido proceso.

En su labor, los jueces de la Sala Penal no solo deben revisar la validez de las pruebas y los argumentos presentados por ambas partes, sino también examinar la correcta aplicación de la ley y los derechos fundamentales del acusado. En muchos casos, la Sala Penal también actúa como un tribunal de apelación, revisando las decisiones tomadas en los juzgados de primera instancia para asegurar que no haya errores procesales ni violaciones a los derechos de los involucrados.

Además, en este tipo de casos es común la aplicación de principios constitucionales e internacionales relacionados con la protección de los derechos humanos, por lo que los jueces deben tener un conocimiento exhaustivo de las normativas nacionales e internacionales que afectan al sistema penal.

- **Sala Civil:**

La Sala Civil es encargada de resolver los casos relacionados con el derecho civil, que incluyen una amplia gama de asuntos legales que afectan a las personas y sus propiedades. Entre los casos más comunes que se abordan en esta sala se encuentran los relacionados con contratos, herencias, divorcios, obligaciones y derechos civiles. Su función es asegurar que las disputas entre personas sean resueltas de manera equitativa y conforme a la ley.

Esta sala tiene una especial relevancia en la vida cotidiana de los ciudadanos, ya que muchas de las disputas legales involucran derechos

fundamentales como la propiedad, la familia, los contratos, y la responsabilidad civil. Los jueces de la Sala Civil deben revisar los hechos, la interpretación de las leyes y la aplicación de principios como la buena fe en los contratos y la equidad en los litigios familiares. Además, también son responsables de la ejecución de sentencias civiles, como el pago de deudas, la restitución de bienes, o la compensación por daños.

La Sala Civil puede intervenir en casos de desalojo, reclamaciones por daños y perjuicios, nulidad de contratos, y conflictos entre herederos, y su resolución puede tener un impacto directo en la vida personal y financiera de los ciudadanos.

- **Sala Comercial:**

La Sala Comercial se ocupa de resolver disputas comerciales y societarias entre empresas y personas, que suelen involucrar la interpretación de contratos comerciales, quiebras, reclamaciones de propiedad intelectual, disputas de competencia y conflictos derivados de la actividad empresarial. Esta sala es clave para el adecuado funcionamiento de la economía y para garantizar que las relaciones comerciales se desarrollen bajo principios de legalidad y transparencia.

Los casos más frecuentes que se abordan incluyen disputas sobre la interpretación de contratos comerciales, como acuerdos de compraventa, distribución, asociaciones comerciales, entre otros. Además, la Sala Comercial también se encarga de los casos relacionados con quiebras y concursos de empresas, donde se debe determinar cómo se distribuirán los activos de una empresa que ha dejado de ser viable económicamente. Los jueces de esta sala son fundamentales para asegurar que las actividades empresariales se desarrollen bajo un marco jurídico claro y predecible, permitiendo así el crecimiento del sector privado y la confianza en el sistema económico.

Además, la Sala Comercial se involucra en casos de propiedad intelectual, tales como disputas por derechos de autor, patentes y marcas, que son esenciales en un mundo globalizado y digitalizado

donde los derechos sobre la propiedad intelectual son cada vez más importantes.

- **Sala de Familia:**

La Sala de Familia tiene la responsabilidad de resolver conflictos familiares que involucran derechos fundamentales como el matrimonio, la tenencia de hijos, herencias familiares y otros aspectos relacionados con la vida familiar. Este tipo de disputas afecta profundamente la vida personal de los individuos involucrados, por lo que los jueces en esta sala deben tener una alta capacidad de empatía y de comprensión de las complejidades emocionales que a menudo acompañan a estos casos.

Entre los casos más comunes que se tratan en la Sala de Familia se incluyen los divorcios, las disputas sobre la custodia de menores, las adopciones, y los regímenes patrimoniales en el matrimonio. La Sala de Familia también juega un papel clave en la resolución de conflictos hereditarios, garantizando que las disposiciones de un testamento o el reparto de bienes familiares se realicen conforme a la ley.

Esta sala se rige por principios como el interés superior del niño, la protección de los derechos fundamentales de los miembros de la familia, y la equidad en la distribución de bienes. Los jueces de la Sala de Familia son responsables de tomar decisiones que impactan directamente en la vida de los involucrados, por lo que se espera de ellos un enfoque sensible y justo en cada caso.

- **Sala Laboral:**

La Sala Laboral es la encargada de resolver los conflictos que surgen entre empleadores y trabajadores. Este tipo de disputas puede involucrar cuestiones como el despido injustificado, condiciones laborales inapropiadas, salarios impagos, acuerdos de indemnización y otros derechos laborales. La función principal de esta sala es asegurar que los trabajadores reciban un trato justo y que sus derechos laborales sean respetados conforme a las normativas vigentes en el país.

La Sala Laboral también interviene en casos de seguridad social y en disputas que involucran la afiliación o pago de pensiones por parte de los empleadores. En muchos casos, se abordan conflictos relacionados con acuerdos colectivos entre sindicatos y empresas, especialmente en sectores estratégicos de la economía. Los jueces de la Sala Laboral deben asegurarse de que la ley se aplique equitativamente, protegiendo los derechos de los trabajadores mientras se resguardan los intereses de las empresas.

Además, la Sala Laboral también se encarga de resolver las reclamaciones de acoso laboral, discriminación en el trabajo, y otros conflictos que afectan el ambiente laboral y las relaciones entre empleadores y empleados. En este sentido, la Sala Laboral juega un papel crucial en la promoción de un trabajo decente y en la creación de un ambiente de trabajo justo y equitativo.

Cada una de estas salas está compuesta por un panel de jueces, quienes son responsables de analizar los casos y dictar sentencias. La Corte Superior de Justicia también tiene la facultad de emitir resoluciones en procesos de amparo, hábeas corpus, y otras acciones constitucionales en los casos que sean de su competencia.

- **Juzgados de primera instancia**

Los juzgados de primera instancia son los encargados de conocer los casos judiciales desde su inicio, cuando las partes presentan su demanda y se inicia el proceso legal. Los jueces de primera instancia son responsables de examinar los hechos, las pruebas y las leyes aplicables para emitir una sentencia inicial sobre el caso. En este nivel, se resuelven una gran variedad de casos, desde conflictos civiles hasta penales, y se abordan tanto cuestiones simples como complejas.

Los jueces de primera instancia tienen la responsabilidad de actuar de manera imparcial, siguiendo el debido proceso legal. Su labor es crucial porque, dependiendo de su decisión, los litigantes pueden optar por apelar ante una Corte Superior de Justicia. Por tanto, los juzgados de primera instancia son el primer filtro en la administración de justicia, y sus decisiones son fundamentales en la resolución de conflictos legales.

- **Juzgados de paz**

Los juzgados de paz son órganos judiciales que se encargan de resolver conflictos menores, aquellos que no afectan gravemente los derechos fundamentales de las personas ni tienen implicaciones de alto impacto social. Los casos que suelen tratar estos juzgados incluyen conflictos de vecindad, divorcios de mutuo acuerdo, reclamaciones de pequeñas deudas, entre otros asuntos menos complejos que los que son tratados en los juzgados de primera instancia.

A diferencia de los juzgados de primera instancia, los jueces de paz no requieren de una formación tan especializada, y suelen ser personas con formación básica en derecho o profesionales con experiencia en la comunidad. El objetivo de los juzgados de paz es proporcionar una respuesta rápida y accesible a las personas, permitiendo la resolución de conflictos sin necesidad de acudir a instancias judiciales más formales. Estos jueces actúan bajo el principio de conciliación y fomentan la resolución pacífica de disputas dentro de la comunidad. Además, los juzgados de paz suelen desempeñar un papel fundamental en la democratización del acceso a la justicia, al acercar los mecanismos legales a las zonas rurales y de difícil acceso.

El sistema judicial está diseñado para garantizar la imparcialidad de los jueces y tribunales y para asegurar que todas las personas puedan acceder a un juicio justo y equitativo, tal como lo establece la Constitución y las leyes nacionales.

2.2.2.2 Estructura y Funciones de las Cortes Superiores de Justicia

La Corte Superior de Justicia es un órgano jurisdiccional especializado encargado de impartir justicia en las regiones del Perú. Su estructura está organizada en diversas salas, cada una especializada en áreas específicas del derecho. Las principales funciones de las Cortes Superiores de Justicia son:

- **Impartir justicia en primera instancia**

La **Corte Superior de Justicia** tiene la responsabilidad de conocer los casos judiciales en su fase inicial, lo que implica que los jueces de esta corte son los encargados de escuchar las demandas presentadas por las partes y de analizar

los hechos y pruebas. Los casos que llegan a la Corte Superior son de gran importancia para la administración de justicia, y los jueces deben aplicar las leyes pertinentes de manera imparcial y objetiva.

Los casos pueden ser de diferentes índoles, incluyendo causas civiles, penales, comerciales, laborales, entre otras. Esto implica que la Corte Superior tiene un amplio campo de acción, con salas especializadas para cada tipo de litigio. La función de los jueces de primera instancia de la Corte Superior es esencial, pues es en este nivel donde se establecen las bases de la resolución de los conflictos legales.

En los casos civiles, los jueces deben abordar temas complejos relacionados con propiedad, contratos, herencias, etc., mientras que, en los casos penales, se deben tomar decisiones sobre la culpabilidad o inocencia de los acusados, siempre garantizando el derecho a la defensa. Además, los jueces deben asegurar que los procesos sean equilibrados y que ambas partes puedan presentar sus argumentos y evidencias.

- **Emitir decisiones de apelación**

La Corte Superior de Justicia también tiene la función de revisar las apelaciones presentadas contra las decisiones emitidas por los juzgados de primera instancia. En este contexto, la Corte Superior actúa como un tribunal de segunda instancia, en el cual se revisa si la decisión tomada en primera instancia fue correcta, tanto en cuanto a la aplicación de la ley como a la interpretación de los hechos.

En muchos casos, los fallos de los jueces de primera instancia pueden ser apelados por las partes que no estén de acuerdo con la decisión. Esto permite que la Corte Superior tenga la última palabra en cuanto a la corrección y equidad de las resoluciones judiciales, lo que garantiza que las decisiones sean consistentes y justas. La Corte Superior puede, en consecuencia, reformar, confirmar o anular las decisiones de los jueces inferiores, y en casos excepcionales, puede volver a examinar los hechos del caso, si lo considera necesario.

Además, los jueces de la Corte Superior tienen un papel clave en el control de la legalidad del proceso, ya que deben verificar que se haya seguido el debido

proceso y que se hayan respetado los derechos fundamentales de los involucrados, especialmente en causas penales.

- **Interpretar y aplicar las leyes**

Una de las funciones fundamentales de los jueces de las Cortes Superiores de Justicia es la interpretación y aplicación correcta de las leyes. Cuando se presentan conflictos legales complejos, los jueces deben interpretar las normas jurídicas de manera que se ajusten a los principios constitucionales y al contexto social y económico del país. La jurisprudencia de la Corte Superior, por tanto, no solo resuelve casos particulares, sino que también contribuye al desarrollo y evolución del derecho en el Perú.

Esta función es especialmente importante cuando se enfrentan situaciones que no están claramente definidas por la ley o cuando las leyes existentes deben ser interpretadas en función de nuevas circunstancias sociales o económicas. En estos casos, la Corte Superior de Justicia tiene el poder de adaptarse a los cambios y ofrecer interpretaciones innovadoras que resuelvan las disputas sin dejar de respetar los principios fundamentales del derecho.

Los jueces deben basarse en los antecedentes legales, las normas internacionales, y los principios constitucionales para tomar decisiones que sean tanto justas como consistentes con el marco normativo del país. En muchos casos, esto puede implicar un análisis profundo de leyes complejas y una interpretación de la ley que contemple los derechos fundamentales, el acceso a la justicia y la equidad.

- **Desarrollar jurisprudencia vinculante**

Las decisiones tomadas por la Corte Superior de Justicia no solo resuelven los casos que se presentan, sino que también tienen un impacto duradero en el sistema judicial del país, ya que establecen precedentes jurídicos que serán seguidos por otros tribunales. Este proceso de desarrollo de jurisprudencia vinculante es fundamental para asegurar la coherencia y la uniformidad en las decisiones judiciales, lo que garantiza la certeza jurídica para todos los ciudadanos.

La Corte Superior, al emitir sus decisiones, puede establecer principios jurídicos clave que guiarán la interpretación y aplicación de la ley en casos futuros. Este proceso es vital, ya que las decisiones judiciales se convierten en normas indirectas que influencian cómo se resuelven los casos similares en el futuro. De esta manera, la jurisprudencia vinculante asegura que los jueces y tribunales inferiores sigan un criterio uniforme a la hora de tomar decisiones, lo que evita contradicciones y asegura que se haga justicia de manera consistente.

La jurisprudencia vinculante también juega un rol importante en la evolución del derecho en el país, ya que permite a los tribunales adaptar la ley a los cambios sociales y económicos, sin necesidad de reformas legislativas.

- **Garantizar el acceso a la justicia**

Uno de los roles más importantes de las Cortes Superiores de Justicia es el de garantizar el acceso a la justicia para todas las personas, independientemente de su situación económica, social o geográfica. El acceso a la justicia es un derecho fundamental que asegura que todas las personas, sin distinción alguna, puedan hacer valer sus derechos a través del sistema judicial.

La Corte Superior desempeña un papel clave en este sentido, promoviendo una administración de justicia accesible y eficiente. Esto incluye la reducción de barreras que puedan existir para que las personas accedan a los tribunales, tales como las barreras económicas, la complejidad de los procedimientos legales o las barreras geográficas en zonas rurales o de difícil acceso.

Además, la Corte Superior debe garantizar que los procesos judiciales sean transparentes y justos, asegurando que todas las partes involucradas en un caso tengan las mismas oportunidades de presentar su caso y defender sus intereses. Esto también implica el uso de tecnologías para facilitar el acceso, como la implementación de expedientes electrónicos, el uso de videoconferencias para audiencias a distancia, y la digitalización de documentos judiciales, lo que permite a los ciudadanos acceder a la información y a los procesos sin necesidad de desplazarse físicamente a los tribunales.

2.2.2.3 Desafíos en la Gestión de la Jurisprudencia

Uno de los principales desafíos que enfrenta las Cortes Superiores de Justicia es la gestión de la vasta cantidad de jurisprudencia generada por los jueces y tribunales en el país. Con el creciente volumen de casos judiciales, la cantidad de resoluciones, sentencias y fallos emitidos por los tribunales aumenta cada año, lo que genera una carga administrativa significativa.

Algunos de los desafíos específicos en la gestión de la jurisprudencia son los siguientes:

- **Acceso limitado a la información**

  Aunque la Corte Suprema ha implementado iniciativas para digitalizar las resoluciones judiciales, muchas veces los documentos no están completamente centralizados o son de difícil acceso para los abogados y jueces, lo que dificulta la búsqueda de precedentes relevantes.

- **Desorden en la organización de la jurisprudencia**

  La falta de estandarización en la redacción de los fallos puede generar dificultades para encontrar sentencias similares o relacionadas. Además, el uso de diferentes terminologías en las sentencias puede hacer que la búsqueda de precedentes sea ineficiente.

- **Tiempo y recursos limitados**

  Los jueces y abogados suelen enfrentar dificultades para acceder a jurisprudencia relevante debido a la falta de tiempo y recursos para revisar manualmente miles de documentos judiciales.

- **Dificultades en la identificación de precedentes relevantes**

  Debido a la gran cantidad de documentos, es común que se presenten dificultades en la identificación de precedentes clave que puedan influir en la resolución de un caso.

Para enfrentar estos desafíos, se ha propuesto la implementación de tecnologías avanzadas como los sistemas de recomendación basados en inteligencia artificial (IA) y el procesamiento de lenguaje natural (PLN), que permiten realizar búsquedas automatizadas de jurisprudencia relevante de manera eficiente.

2.2.2.4 Tecnología en el Sistema Judicial

La tecnología ha comenzado a desempeñar un papel fundamental en la mejora del sistema judicial del Perú, especialmente en lo que respecta al acceso a la justicia y la gestión de la información legal. Entre las principales iniciativas tecnológicas implementadas se incluyen:

- **Digitalización de expedientes judiciales**

  La Corte Suprema y las Cortes Superiores de Justicia están trabajando para digitalizar los expedientes judiciales, lo que facilita el acceso y la gestión de los casos. La digitalización también mejora la seguridad de los documentos, evitando pérdidas y facilitando el seguimiento de los casos.

- **Sistemas de gestión judicial**

  Estas plataformas permiten a los abogados, jueces y otros actores del sistema judicial seguir el progreso de los casos judiciales de manera electrónica, lo que agiliza el proceso y facilita el intercambio de información.

- **Plataformas de consulta de jurisprudencia**

  Las plataformas electrónicas de consulta de jurisprudencia permiten que los profesionales del derecho accedan rápidamente a sentencias previas, facilitando la identificación de precedentes y la toma de decisiones.

- **Implementación de IA y PLN**

  Estos avances tecnológicos tienen el potencial de transformar la manera en que se gestionan y procesan los documentos legales. Los sistemas de recomendación pueden ser utilizados para sugerir jurisprudencia relevante de manera automática, basándose en los antecedentes de los casos previos y los textos de las resoluciones judiciales.

La implementación de estas tecnologías sigue siendo un desafío, pero representa una oportunidad clave para mejorar la eficiencia del sistema judicial y facilitar el acceso a la justicia para todos los ciudadanos.

2.2.3 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) es una rama de la inteligencia artificial que se encarga de la interacción entre las computadoras y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten y generen texto de manera que resulte útil para tareas como la traducción automática, el análisis de sentimientos, la extracción de información y la generación de recomendaciones (Jurafsky & Martin, 2020).

Para lograr estos objetivos, el PLN combina técnicas de lingüística computacional, métodos estadísticos y algoritmos de aprendizaje automático. A continuación, se describen algunas de las etapas fundamentales en el procesamiento de textos:

- **Tokenización**

  Es el proceso de dividir el texto en unidades básicas llamadas tokens, que pueden ser palabras, frases o incluso caracteres. Por ejemplo, la frase "El sistema de recomendación es efectivo" se divide en tokens: *[El, sistema, de, recomendación, es, efectivo]*.

- **Lematización y Stemming**

  Estas técnicas reducen las palabras a su forma base o raíz. La lematización transforma las palabras según reglas lingüísticas (por ejemplo, "corriendo" se convierte en "correr"), mientras que el stemming utiliza algoritmos para eliminar sufijos y prefijos.

- **Análisis Sintáctico**

  Consiste en identificar la estructura gramatical de las oraciones. Se construyen árboles sintácticos que muestran las relaciones entre las palabras, lo

que permite comprender la función que desempeña cada término dentro de la oración.

- **Análisis Semántico**

  Tiene como objetivo extraer el significado de los textos, identificando relaciones entre conceptos y palabras. Esto es crucial para tareas como la identificación de similitud entre documentos o la extracción de información relevante.

Un concepto central en el PLN es la representación vectorial del texto. En lugar de trabajar con cadenas de caracteres, se transforman las palabras y documentos en vectores numéricos que capturan sus características semánticas. Una de las medidas más comunes para evaluar la similitud entre estos vectores es la **similitud coseno**, definida por la siguiente ecuación:

$$\text{cosine}(x, y) = \frac{x \cdot y}{||x||\,||y||}$$

donde (x) y (y) son los vectores correspondientes a dos textos. $(\text{x} \cdot \text{y})$ es su producto punto y $||x||$ y $||y||$ son sus respectivas normas euclidianas. Esta métrica es esencial para determinar cuán similares son dos textos en el espacio vectorial.

En los últimos años, el avance del aprendizaje profundo ha revolucionado el campo del PLN. Modelos preentrenados como BERT (Devlin et al., 2018) y sus variantes, como RoBERTa (Liu et al., 2019), han demostrado una capacidad sobresaliente para capturar la complejidad del lenguaje. Estos modelos se entrenan con enormes cantidades de datos y pueden luego ser afinados para tareas específicas, ofreciendo mejoras significativas en la interpretación y generación de texto.

La aplicación de estas técnicas en el ámbito jurídico, por ejemplo, en sistemas de recomendación de jurisprudencia, permite transformar resoluciones judiciales en representaciones vectoriales. Esto facilita la identificación de casos similares al medir la distancia (o similitud) entre vectores, ayudando a los profesionales del derecho a encontrar precedentes relevantes de manera eficiente.

2.2.4 Modelos de Embeddings de Texto

Los embeddings de texto son representaciones vectoriales densas que transforman palabras, frases o documentos en vectores numéricos en un espacio de alta dimensión. La idea central es que la proximidad en este espacio refleja similitudes semánticas; es decir, términos o documentos con significados similares estarán cercanos entre sí. Esta propiedad resulta esencial en aplicaciones como la recomendación de jurisprudencia, donde es crucial identificar precedentes legales relevantes basados en el contenido semántico de las resoluciones.

2.2.4.1 Modelos Tradicionales

Los modelos tradicionales de embeddings de texto se basan en la idea de que las palabras con contextos similares tienen significados similares, y se entrenan para capturar esta información a partir de grandes corpus de texto.

2.2.4.1.1 Word2Vec

Introducido por (Mikolov et al., 2013), Word2Vec se compone de dos arquitecturas principales:

- **Continuous Bag-of-Words (CBOW)**

  En este enfoque, el modelo predice la palabra central a partir de un conjunto de palabras de contexto. Es decir, dada una ventana de contexto de tamaño (c), el modelo estima la probabilidad de la palabra objetivo basándose en las palabras circundantes.

  En la Tabla 1 se ilustra una matriz de Bag-of-Words en la que las palabras se representan con valores binarios (0 y 1). En este ejemplo, se muestran tres documentos (o frases) y se indica la presencia (1) o ausencia (0) de cada palabra del vocabulario en dichos documentos.

Tabla 1 ***(Ejemplo de Matriz de Bag of Words Binaria)***

Ejemplo de Matriz de Bag of Words Binaria

Documento	**El**	**gato**	**negro**	**duerme**	**Plácidamente**

Doc 1	1	1	1	1	1
Doc 2	1	0	0	1	0
Doc 3	1	1	0	0	0

Nota: Elaboración propia

En este ejemplo:

- Se ha definido un vocabulario compuesto por las palabras: *El, gato, negro, duerme, plácidamente*.
- Cada fila de la tabla representa un documento (o una oración). Por ejemplo, el **Doc 1** contiene todas las palabras del vocabulario (por ello se asigna 1 en cada celda), mientras que en el **Doc 2** solo aparecen las palabras "El" y "duerme" (por ello se asigna 1 a esas columnas y 0 a las demás).
- Esta representación binaria permite indicar de forma simple si cada palabra está presente o no en cada documento, lo cual es una forma básica de representar la información para modelos de PLN.

Este tipo de matriz es fundamental en tareas de clasificación y búsqueda de similitud, ya que convierte los textos en vectores que pueden ser comparados utilizando, por ejemplo, la similitud coseno.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**

TF-IDF es una técnica de ponderación que mide la importancia de un término en un documento dentro de un corpus. A diferencia del enfoque binario, TF-IDF considera no solo la presencia de una palabra, sino también su frecuencia y la rareza en el conjunto de documentos. La fórmula general es:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right),$$

Donde:

- $(\mathrm{TF}(\mathrm{t}, \mathrm{d}))$ es la frecuencia del término (t) en el documento (d).
- (N) es el número total de documentos en el corpus.
- $(\mathrm{DF}(\mathrm{t}))$ es el número de documentos en los que aparece el término (t).

Esta técnica, descrita en (Manning et al., 2008), ayuda a resaltar términos distintivos y reduce el impacto de palabras muy comunes.

En la Tabla 2 se ilustra el cálculo de TF, IDF y TF*IDF para dos documentos (A y B) con términos jurídicos. Se asume que el vocabulario está compuesto por las palabras: "el," "tribunal," "resuelve," "la," "sentencia," "crea," "un," y "precedente."

- Documento A: "El tribunal resuelve la sentencia."
- Documento B: "El tribunal crea un precedente."

Tabla 2 ***(Ejemplo de cálculo TF, IDF y TF*IDF para dos documentos con términos jurídicos)***

Ejemplo de cálculo TF, IDF y TF*IDF para dos documentos con términos jurídicos

Word	**TF (A)**	**TF (B)**	**IDF**	**TF*IDF (A)**	**TF*IDF (B)**
El	0.20	0.20	0.000	0.000	0.000
Tribunal	0.20	0.20	0.000	0.000	0.000
Resuelve	0.20	0.00	0.6931	0.1386	0.000
La	0.20	0.00	0.6931	0.1386	0.000
Sentencia	0.20	0.00	0.6931	0.1386	0.000
Crea	0.00	0.20	0.6931	0.000	0.1386
Un	0.00	0.20	0.6931	0.000	0.1386

precedente	0.00	0.20	0.6931	0.000	0.1386

Nota: Elaboración propia

Esta tabla muestra de forma clara la relevancia de cada término en cada documento, destacando aquellos que son más distintivos (IDF elevado) y que aparecen con mayor frecuencia (TF elevado). De este modo, TF-IDF ofrece una representación numérica útil para tareas de recuperación de información y análisis de similitud en el ámbito jurídico.

- **Skip-Gram**

Este modelo, utilizado comúnmente en aplicaciones donde se dispone de corpus de gran tamaño, realiza la tarea inversa: dada una palabra central, predice las palabras del contexto. La función objetivo para el modelo Skip-Gram se define como:

$$J = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \le j \le c j \neq 0} \log p\left(w_{t+j} \mid w_t\right),$$

Donde:

- T es el número total de palabras en el corpus
- c representa el tamaño de la ventana de contexto

- **Otras Consideraciones en la Representación de Texto**

Además del modelo binario de Bag-of-Words, es importante considerar otras técnicas que permiten captar de manera más rica y precisa la información semántica de los textos, especialmente en contextos complejos como el jurídico. Entre estas técnicas se destacan:

- **Modelos de N-gramas**

Los N-gramas extienden la idea de Bag-of-Words al considerar secuencias de N palabras consecutivas. Esto permite capturar cierta información de orden y contexto que se pierde cuando se tratan las palabras de forma independiente. Por ejemplo, en la frase "la

sentencia fue unánime", un modelo de bigramas podría capturar las relaciones "la sentencia" y "sentencia fue", proporcionando una visión más completa del contexto.

- **Reducción de Dimensionalidad**

  Las representaciones basadas en Bag-of-Words y TF-IDF suelen ser de alta dimensión y muy dispersas. Técnicas de reducción de dimensionalidad, como el Análisis Semántico Latente (LSA) mediante la descomposición en valores singulares (SVD), permiten obtener una representación más compacta. Esto no solo reduce la complejidad computacional, sino que también ayuda a capturar relaciones latentes entre términos y documentos, facilitando la identificación de patrones semánticos.

- **Embeddings de Palabras**

  El uso de modelos de embeddings de palabras, como Word2Vec, GloVe y los modelos contextuales (BERT, RoBERTa), ha permitido superar las limitaciones del modelo Bag-of-Words. Estos métodos generan representaciones densas y continuas que capturan relaciones semánticas y contextuales de manera más precisa. Por ejemplo, en un contexto jurídico, los embeddings pueden distinguir entre diferentes usos de términos que, en un modelo binario, se tratarían de forma idéntica.

Ambas arquitecturas utilizan una red neuronal de una sola capa sin función de activación en la capa oculta, cuyo objetivo es aprender una representación densa y de baja dimensión de cada palabra. Para reducir la complejidad computacional, se emplean técnicas como el **negative sampling**, que consiste en actualizar únicamente una pequeña muestra de los parámetros en cada iteración de entrenamiento, en lugar de todas las palabras del vocabulario.

2.2.4.1.2 GloVe (Global Vectors for Word Representation)

GloVe, propuesto por (Pennington et al., 2014), adopta un enfoque distinto al de Word2Vec, ya que aprovecha estadísticas globales de coocurrencia de

palabras en el corpus. La idea es construir una matriz de coocurrencia *X*, donde cada elemento (X_{ij}) indica cuántas veces aparece la palabra *j* en el contexto de la palabra *i*.

El objetivo de GloVe es que el producto punto de los vectores de dos palabras se aproxime al logaritmo del número de veces que ambas palabras coocurren, es decir:

$$w_i^{\top}\widetilde{w_j} + b_i + \widetilde{b_j} \approx \log(X_{ij}),$$

Donde: (w_i) y ($\widetilde{w_j}$) son los vectores de las palabras *i* y *j* respectivamente, y (b_i)

y ($\widetilde{b_j}$) son los sesgos asociados. Para alcanzar este objetivo, GloVe minimiza la siguiente función de costo ponderada:

$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^{\top}\widetilde{w_j} + b_i + \widetilde{b_j} - \log(X_{ij})\right)^2,$$

Donde: *V* es el tamaño del vocabulario y ($f(X_{ij})$) es una función de ponderación diseñada para limitar la influencia de coocurrencias extremadamente frecuentes. Una elección común para *f* es:

$$f(x) = \begin{cases} \left(\frac{x}{x_{máx}}\right)^{\alpha} & si\ x\ < x_{máx}, \\ 1 & si\ x\ \geq x_{máx}, \end{cases}$$

con parámetros $x_{máx}$ y α determinados empíricamente.

Mientras que Word2Vec se centra en capturar relaciones semánticas a nivel local a través de ventanas de contexto, GloVe incorpora información global del corpus a través de la matriz de coocurrencia. Ambos enfoques han demostrado ser efectivos para tareas de similitud semántica y han sentado las bases para el desarrollo de modelos de embeddings más avanzados.

En aplicaciones jurídicas, estos modelos permiten transformar resoluciones y otros documentos legales en representaciones numéricas, facilitando la medición

de similitud semántica entre textos. Esto resulta fundamental en sistemas de recomendación de jurisprudencia, donde identificar precedentes legales relevantes se traduce en comparar vectores en un espacio semántico de alta dimensión.

2.2.4.2 Modelos Contextuales

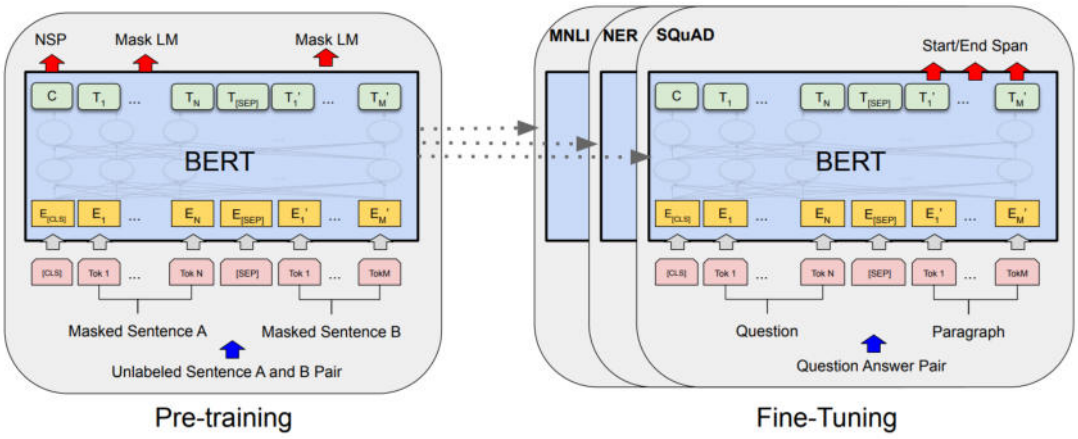
A diferencia de los modelos tradicionales que generan representaciones estáticas donde una misma palabra tiene siempre el mismo vector sin importar su contexto, los modelos contextuales generan embeddings dinámicos. Esto significa que el vector de una palabra varía según el contexto en el que aparece, permitiendo capturar matices semánticos que son esenciales en aplicaciones donde el significado depende fuertemente del entorno, como en textos jurídicos.

2.2.4.2.1 BERT:

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) es uno de los modelos contextuales más influyentes. Utiliza la arquitectura Transformer para procesar el texto de manera bidireccional, es decir, considerando simultáneamente el contexto anterior y posterior de cada palabra. Para lograr esto, BERT se entrena con dos tareas principales:

Figura 1 ***(BERT: Preentrenamiento de transformadores bidireccionales profundos para la comprensión del lenguaje)***

Arquitectura BERT



Nota: La figura fue extraída del artículo (Devlin et al., 2018).

- **Masked Language Modeling (MLM)**

Durante el preentrenamiento, se enmascaran aleatoriamente ciertos tokens del texto y el modelo se entrena para predecir los tokens enmascarados utilizando el contexto completo. La función de pérdida para esta tarea se expresa como:

$$\mathcal{L}_{\mathcal{MLM}} = -\sum_{i \in \mathcal{M}} \log P\left(x_i \mid x_{\setminus M}\right),$$

Donde: **M** es el conjunto de posiciones de los tokens enmascarados y $\boldsymbol{x_{\setminus M}}$ representa el resto del contexto.

- **Next Sentence Prediction (NSP)**

Además, BERT se entrena para predecir si una oración sigue a otra, lo que ayuda a capturar relaciones inter-sentenciales y mejora la comprensión del texto en secuencias largas.

La Figura 1 ilustra el procedimiento de preentrenamiento y ajuste fino de BERT. La parte izquierda de la figura muestra que BERT se pre entrena en frases enmascaradas y predice si la frase B sigue a la frase A. La parte derecha de la figura muestra que BERT se ajusta para tareas específicas como la inferencia del lenguaje natural, el reconocimiento de entidades nombradas y la respuesta a preguntas.

Tras el preentrenamiento, BERT se puede afinar (fine-tune) para tareas específicas, permitiendo adaptar las representaciones contextuales a dominios particulares, como el análisis de resoluciones judiciales.

2.2.4.2.2 RoBERTa

BERT es un nuevo modelo de representación del lenguaje que pre entrena representaciones bidireccionales profundas a partir de texto no etiquetado al condicionar conjuntamente el contexto izquierdo y derecho en todas las capas. Como resultado, el modelo BERT preentrenado se puede ajustar con solo una capa de salida adicional para crear modelos de última generación para una amplia gama de tareas, como respuesta a preguntas e inferencia del lenguaje, sin modificaciones

sustanciales de la arquitectura específica de la tarea. BERT es conceptualmente simple y empíricamente poderoso. Obtiene nuevos resultados de última generación en once tareas de procesamiento del lenguaje natural, incluida el aumento de la puntuación GLUE a 80,5% (mejora absoluta de 7,7 puntos), la precisión de MultiNLI a 86,7% (mejora absoluta de 4,6 puntos), SQUAD v1.1 respuesta a preguntas Test F1 a 93,2 (mejora absoluta de 1,5 puntos) y SQUAD v2.0 Test F1 a 83,1 (mejora absoluta de 5,1 puntos).

RoBERTa (Liu et al., 2019), es una variante optimizada de BERT que introduce varias mejoras en el proceso de entrenamiento:

- **Eliminación de NSP:** RoBERTa descarta la tarea de Next Sentence Prediction, centrándose exclusivamente en el MLM, lo que permite un entrenamiento más eficiente y enfocado.
- **Dynamic Masking:** Emplea técnicas de enmascarado dinámico, de manera que en cada época se generan nuevas máscaras, lo que enriquece el entrenamiento.
- **Ajuste de Hiperparámetros:** Utiliza mayores tamaños de lote (batch size), más iteraciones y conjuntos de datos más extensos, lo que contribuye a generar representaciones más robustas.

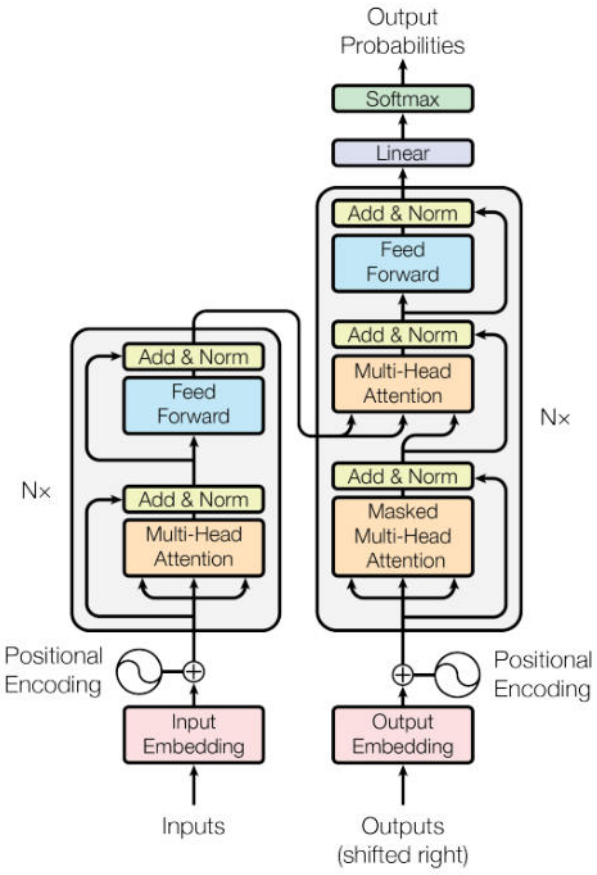
Ambos modelos se basan en la arquitectura Transformer, cuyo mecanismo de atención escalada se define en la siguiente sección.

2.2.4.3 Arquitectura Transformer

La arquitectura Transformer fue introducida por Vaswani et al. en 2017 en el artículo "Attention Is All You Need" (Vaswani et al., 2023). Este modelo revolucionó el campo del Procesamiento de Lenguaje Natural al abandonar las estructuras secuenciales tradicionales (como las RNNs y LSTMs) en favor de mecanismos de atención, permitiendo procesar secuencias de forma paralela y capturar relaciones de largo alcance de manera más eficiente.

Figura 2 ***(Arquitectura de Modelo Transformer)***

Arquitectura Transformer



Nota: La figura fue extraída del artículo (Vaswani et al., 2023)

La Figura 2 ilustra la arquitectura del modelo Transformer, un modelo de aprendizaje profundo revolucionario en el campo del procesamiento del lenguaje natural, particularmente para tareas de traducción automática. El modelo se compone de dos componentes principales: el codificador (a la izquierda) y el decodificador (a la derecha).

2.2.4.3.1 Mecanismo de Atención Escalada (Scaled Dot-Product Attention)

El núcleo del Transformer es el mecanismo de atención, el cual permite que cada palabra en una secuencia preste atención a todas las demás. Dado un conjunto de *queries Q, keys K* y *values V*, la atención se calcula mediante:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

Donde: (d_k) es la dimensión de las *keys*. La división por $(\sqrt{\mathrm{d_k}})$ se utiliza para estabilizar los gradientes y evitar que los valores del producto punto sean demasiado grandes.

2.2.4.3.2 Atención Multi-Cabeza (Multi-Head Attention)

En lugar de aplicar una única función de atención, el Transformer utiliza múltiples "cabezas" de atención, lo que permite al modelo capturar diferentes aspectos de la información. Cada cabeza realiza su propia operación de atención de forma independiente:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O,$$

Donde cada cabeza se define como:

$$texthead_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right).$$

Aquí, QW_i^Q, W_i^K y W_i^V son matrices de proyección para la i-ésima cabeza, y $(\mathrm{W^O})$ es la matriz de proyección final que combina las cabezas.

2.2.4.3.3 Codificación Posicional (Positional Encoding)

A diferencia de las RNNs, el Transformer no procesa las secuencias en orden. Para inyectar información de la posición de las palabras, se añaden vectores de codificación posicional a las entradas. Una formulación común para estas codificaciones es:

$$\mathrm{PE}_{(pos,2i+1)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

Donde: *pos* es la posición de la palabra en la secuencia, *i* es la dimensión y $(\mathrm{d_{model}})$ es la dimensión del modelo. Estas funciones sinusoidales permiten que el modelo aprenda relaciones de posición de forma continua.

2.2.4.3.4 Estructura del Modelo

La arquitectura Transformer se divide en dos bloques principales:

- **Encoder:** Compuesto por una serie de capas idénticas, donde cada capa consta de un mecanismo de atención multi-cabeza seguido de una red neuronal feed-forward posicional. Se utilizan conexiones residuales y normalización de capas (layer normalization) en cada subcapa para facilitar el entrenamiento.
- **Decoder:** También conformado por capas idénticas, pero con una estructura ligeramente modificada. Cada capa del decoder incluye:
    - Una capa de atención multi-cabeza enmascarada para asegurar que el modelo no "vea" información futura.
    - Una capa de atención multi-cabeza que opera sobre la salida del encoder, permitiendo al decoder enfocarse en partes relevantes de la secuencia de entrada.
    - Una red neuronal feed-forward posicional.

Esta estructura permite que el Transformer procese información de manera paralela, superando las limitaciones de los modelos secuenciales tradicionales, y ha demostrado un rendimiento superior en tareas de traducción, resumen de textos, y otras aplicaciones del Procesamiento de Lenguaje Natural.

2.2.4.3.5 Importancia en Aplicaciones de PLN:

La arquitectura Transformer ha marcado un antes y un después en el campo del PLN, ya que sus mecanismos de atención permiten capturar dependencias complejas y relaciones semánticas de largo alcance sin recurrir a estructuras recurrentes. Esto es particularmente relevante en aplicaciones jurídicas, donde la interpretación contextual de las resoluciones judiciales es crucial para la identificación de precedentes y el desarrollo de sistemas de recomendación de jurisprudencia.

2.2.5 Sistemas de Recomendación en el Ámbito Jurídico

Los sistemas de recomendación son herramientas diseñadas para filtrar y sugerir información relevante a partir de grandes volúmenes de datos. En el ámbito

jurídico, estos sistemas tienen como objetivo facilitar la identificación de precedentes legales y casos similares, lo que resulta fundamental para apoyar la toma de decisiones en la práctica profesional y la investigación legal.

2.2.5.1 Fundamentos y Enfoques

Los sistemas de recomendación en el contexto jurídico se basan en métodos de recuperación de información y técnicas de filtrado, que pueden clasificarse en tres enfoques principales:

- **Filtrado basado en contenido**

  Este enfoque se centra en analizar las características intrínsecas de los documentos legales, como resoluciones y sentencias, para generar representaciones vectoriales mediante técnicas de Procesamiento de Lenguaje Natural (PLN). La transformación de documentos en vectores se fundamenta en conceptos presentados en (Manning et al., 2008), donde se destaca la importancia de la tokenización, la ponderación de términos y la representación de textos en espacios vectoriales. Además, la medición de similitud semántica comúnmente realizada mediante la similitud coseno permite identificar documentos que tratan temas o conceptos similares. La utilización de embeddings de texto, tanto estáticos (como Word2Vec y GloVe) como contextuales (como BERT y RoBERTa), amplía las capacidades de este enfoque al capturar matices semánticos en función del contexto, como se detalla en (Aggarwal, 2016). Los modelos utilizados pueden ser:

    - **Estáticos:** Como Word2Vec y GloVe, que asignan un vector fijo a cada palabra, sin importar su contexto.
    - **Contextuales:** Como BERT y RoBERTa, que generan representaciones dinámicas en función del contexto en el que aparece la palabra.

La similitud entre documentos se mide comúnmente mediante la **similitud coseno**, definida como:

$$\text{sim}(x, y) = \frac{x \cdot y}{||x||\,||y||}$$

lo que permite identificar documentos jurídicos que traten temas o conceptos similares, facilitando la búsqueda de precedentes relevantes.

- **Filtrado colaborativo**

Este método aprovecha la información derivada de las interacciones y preferencias de los usuarios, tales como abogados o jueces, para recomendar documentos que hayan sido valorados positivamente por otros usuarios con perfiles similares. La idea de que usuarios con comportamientos parecidos compartirán intereses comunes se fundamenta en los principios expuestos en (Ricci et al., 2011), donde se exploran técnicas basadas en la factorización matricial y en algoritmos de aprendizaje colaborativo para modelar dichas interacciones. Para ello, se emplean técnicas como:

  - **Filtrado colaborativo basado en usuarios:** Donde se comparan perfiles de usuarios para recomendar documentos que hayan sido valorados de manera similar por usuarios con intereses comunes.
  - **Filtrado colaborativo basado en ítems:** Que analiza las interacciones entre documentos para identificar aquellos que son comúnmente utilizados o valorados juntos.

Métodos de factorización matricial, como la descomposición en valores singulares (SVD), se utilizan para descubrir patrones latentes en la matriz de interacciones. Aunque en el ámbito jurídico puede haber limitaciones en la disponibilidad de datos explícitos de interacción, cuando se dispone de ellos, este enfoque ayuda a generar recomendaciones basadas en la experiencia colectiva.

- **Enfoques híbridos**

Los métodos híbridos combinan las ventajas del filtrado basado en contenido y del filtrado colaborativo, integrando tanto el análisis del contenido intrínseco de los documentos como la información derivada de las interacciones del usuario. Según (Aggarwal, 2016), la combinación de ambos métodos permite mitigar las limitaciones individuales como la escasez de datos en el filtrado colaborativo o la dependencia exclusiva del contenido en el filtrado

basado en contenido ofreciendo recomendaciones más precisas y personalizadas, especialmente en dominios complejos como el jurídico. por ejemplo:

- **Combinación lineal:** Se ponderan y combinan las puntuaciones obtenidas por cada método.
- **Modelos de aprendizaje conjunto:** Que optimizan simultáneamente ambos componentes para ajustar la recomendación final.

En contextos jurídicos, donde la precisión y la relevancia son críticas, los métodos híbridos permiten compensar las limitaciones individuales de cada técnica, asegurando una recomendación que refleje tanto el contenido intrínseco de los documentos como la experiencia y las preferencias de los usuarios.

2.2.5.2 Aplicaciones y Beneficios en el Ámbito Jurídico:

La implementación de sistemas de recomendación en el sector jurídico puede transformar radicalmente la forma en que se realiza la búsqueda y el análisis de información legal. Algunos de los beneficios son:

- **Optimización de la Búsqueda de Precedentes:** Automatizando la identificación de resoluciones y casos similares, se reduce el tiempo y esfuerzo requerido para la investigación jurídica, permitiendo a los profesionales centrarse en la interpretación y aplicación del derecho.
- **Mejora en la Coherencia Jurisprudencial:** Al facilitar el acceso a precedentes relevantes, se contribuye a una aplicación más uniforme y coherente de la ley.
- **Soporte en la Toma de Decisiones:** La integración de técnicas de PLN y modelos de recomendación permite sugerir de manera proactiva casos y argumentos jurídicos, lo que puede ser de gran ayuda durante la elaboración de estrategias legales.

2.2.5.3 Ejemplos de Implementación

Diversos estudios han abordado el desarrollo de sistemas de recomendación en el ámbito jurídico. Por ejemplo, se han utilizado técnicas de embeddings combinadas con redes neuronales para clasificar la similitud entre descripciones de hechos y extraer relaciones contextuales en resoluciones judiciales (Ruiz, 2021). Asimismo, otros trabajos han explorado el uso de modelos que integran representaciones granulares de texto para identificar casos similares y extraer conceptos relevantes, contribuyendo a la automatización de la identificación de precedentes legales (Hugo Mentzingen Nuno António, 2023) Adicionalmente, enfoques basados en grafos de conocimiento han sido aplicados para comparar puntuaciones de similitud derivadas de expertos con aquellas obtenidas de análisis automatizados, mejorando la precisión del sistema de recomendación (Shao, 2024).

Las Cortes Superiores de Justicia del Perú juegan un rol esencial en el sistema judicial al resolver los casos judiciales en su fase inicial, revisar las apelaciones, interpretar y aplicar las leyes, desarrollar jurisprudencia vinculante, y garantizar el acceso a la justicia para todos los ciudadanos. Sin embargo, debido a la creciente complejidad y volumen de las resoluciones judiciales, se han presentado importantes desafíos en la gestión eficiente de la jurisprudencia. En este contexto, los sistemas de recomendación basados en inteligencia artificial y el procesamiento de lenguaje natural ofrecen una solución innovadora. Estos sistemas tienen la capacidad de analizar grandes volúmenes de textos legales, identificar precedentes relevantes y proporcionar recomendaciones personalizadas a los jueces, abogados y otros actores del sistema judicial. Al utilizar modelos de embeddings de texto, los sistemas de recomendación pueden mejorar la eficiencia en la búsqueda de jurisprudencia, facilitando el acceso a información relevante sin la necesidad de realizar búsquedas manuales extensas. La implementación de estos sistemas en la Cortes Superiores de Justicia permitiría optimizar el proceso de toma de decisiones, reducir los tiempos de resolución de casos y garantizar que las decisiones judiciales sean más rápidas, coherentes y basadas en precedentes jurídicos actualizados. Así, la combinación de la experiencia humana en la interpretación de la ley con las capacidades de los sistemas de recomendación

basados en IA puede transformar la administración de justicia, promoviendo una justicia más accesible y eficiente para todos los ciudadanos.

2.3 Definición de términos

A continuación, se presentan definiciones sencillas de algunos términos técnicos utilizados en este proyecto, con el objetivo de facilitar su comprensión para quienes no son especialistas en el tema:

- **Procesamiento de Lenguaje Natural (PLN):** Conjunto de técnicas y herramientas que permiten a las computadoras entender, interpretar y generar el lenguaje humano. Esto es fundamental para tareas como la traducción, el análisis de sentimientos y la búsqueda de información en grandes volúmenes de texto (Jurafsky & Martin, 2020).
- **Embeddings de Texto:** Representaciones numéricas (vectores) que convierten palabras, frases o documentos en puntos dentro de un espacio de alta dimensión. La idea es que palabras o textos con significados similares estén cerca entre sí, facilitando así la comparación y el análisis semántico
- **Modelo Transformer:** Arquitectura de redes neuronales que utiliza mecanismos de atención para procesar secuencias de texto de forma paralela. Este modelo es capaz de capturar relaciones complejas entre palabras a lo largo de todo el documento, lo que mejora significativamente el rendimiento en tareas de PLN (Vaswani et al., 2023).
- **Word2Vec:** Modelo que aprende representaciones vectoriales de palabras a partir de grandes corpus de texto. Emplea dos enfoques principales: el Continuous Bag-of-Words (CBOW), que predice una palabra a partir de su contexto, y el modelo Skip-Gram, que utiliza una palabra central para predecir sus palabras de contexto (Mikolov et al., 2013).
- **TF-IDF:** Siglas de Term Frequency-Inverse Document Frequency. Es una técnica que combina la frecuencia de un término en un documento (TF) con la rareza del término en todo el corpus (IDF), permitiendo resaltar palabras que son informativas y distintivas en cada documento (Manning et al., 2008)

- **BERT:** Acrónimo de Bidirectional Encoder Representations from Transformers. Es un modelo de lenguaje basado en la arquitectura Transformer que procesa el texto de manera bidireccional, es decir, tiene en cuenta tanto el contexto anterior como el posterior de cada palabra. Esto permite generar representaciones más precisas y contextualmente relevantes (Devlin et al., 2018).
- **Filtrado basado en contenido:** Método de sistemas de recomendación que analiza las características intrínsecas de los documentos (por ejemplo, resoluciones judiciales) para identificar y sugerir aquellos que sean similares en contenido y significado. Este enfoque utiliza técnicas de PLN, como la generación de embeddings, para comparar la similitud semántica entre textos.
- **Filtrado colaborativo:** Enfoque que utiliza las interacciones y preferencias de los usuarios para recomendar documentos que hayan sido valorados positivamente por otros usuarios con perfiles similares. Este método se basa en el análisis de patrones de comportamiento y en técnicas de factorización matricial (Ricci et al., 2011).
- **Enfoques híbridos:** Combinan el filtrado basado en contenido y el filtrado colaborativo para aprovechar lo mejor de ambos métodos. La integración de ambas fuentes de información (características del contenido y datos de interacción del usuario) permite generar recomendaciones más precisas y personalizadas, especialmente en contextos complejos como el jurídico (Aggarwal, 2016).

III. Metodología

3.1 Tipo y nivel de investigación

3.1.1 Tipo de investigación

La investigación es de tipo aplicada, ya que tiene como objetivo desarrollar un sistema de recomendación de jurisprudencia basado en embeddings de texto, con el propósito de resolver un problema específico en el ámbito jurídico: la identificación y sugerencia de precedentes legales relevantes.

Y cuantitativo según su enfoque metodológico, ya que se basa en la recolección de datos textuales estructurados, la transformación de estos en vectores numéricos mediante embeddings**, y** la evaluación del desempeño del modelo mediante métricas estadísticas como precisión, recall y F1-score.

3.1.2 Nivel de investigación

De acuerdo con (Sampieri et al., 2014), el presente trabajo corresponde al nivel de investigación correlacional, ya que se analiza la relación entre los modelos de embeddings de texto utilizados (Word2Vec, RoBERTa, etc.) y el rendimiento del sistema de recomendación propuesto.

3.2 Ámbito temporal y espacial

La delimitación del estudio es un componente metodológico esencial, ya que define el alcance del corpus documental, el contexto de entrenamiento de los modelos y el marco en el cual los resultados pueden ser interpretados. El ámbito de la presente investigación se define por los siguientes criterios espaciales y temáticos.

El ámbito espacial del estudio se circunscribe a una única entidad judicial, el corpus documental íntegro, que comprende 6918 resoluciones , fue extraído de los repositorios de la Corte Superior de Justicia de Lima. Por lo tanto, tanto el

conjunto de entrenamiento (6848 documentos) como el conjunto de evaluación (30 documentos) provienen de esta misma fuente jurisdiccional.

Temáticamente, la investigación se enfoca de manera exclusiva en la materia laboral esta delimitación es crucial, ya que los modelos (léxicos, semánticos e híbridos) fueron entrenados y evaluados utilizando la terminología, los fundamentos de derecho y las estructuras argumentativas propias de las resoluciones de apelación en este campo específico del derecho.

El ámbito temporal de la investigación abarca un período de dos años, todas las resoluciones de apelación que componen el corpus documental (tanto de entrenamiento como de evaluación) fueron emitidas entre el año 2021 y el año 2022. Esta acotación temporal asegura que los modelos sean entrenados y evaluados sobre jurisprudencia laboral reciente.

3.2.1 Población

La población de esta investigación se compone de las resoluciones judiciales emitidas por las Cortes Superiores del Perú. Estas resoluciones contienen información esencial sobre casos previos y sus fundamentos legales, constituyéndose en una fuente valiosa de precedentes para la toma de decisiones jurídicas con un total aproximado de 56,000. Para fines de desarrollo y validación, la población efectiva se delimitó a todas las resoluciones de la Sala Laboral de la Corte Superior de Justicia de Lima correspondientes al período 2021–2022, compiladas por los investigadores para la creación del corpus.

3.2.2 Muestra

La selección de la muestra se realizó teniendo en cuenta criterios técnicos y experimentales que son fundamentales para el entrenamiento y la evaluación de modelos de procesamiento de lenguaje natural. El propósito de esta investigación es analizar el rendimiento de un sistema de recomendación que utiliza embeddings de texto para identificar y sugerir precedentes relevantes.

En ese sentido, se adoptó un enfoque de censo, utilizando la totalidad de las resoluciones judiciales de apelación en materia laboral correspondientes a la población de estudio delimitada, conformando un corpus de 6,918 resoluciones

emitidas por la Sala Laboral de la Corte Superior de Justicia de Lima durante el período 2021–2022. La utilización de la totalidad del corpus permitió garantizar una adecuada cobertura temática, homogeneidad textual y volumen suficiente de datos para el entrenamiento de modelos basados en embeddings.

Diversos estudios en el ámbito del procesamiento de lenguaje natural señalan que modelos basados en arquitecturas tipo BERT pueden entrenarse de manera eficaz utilizando corpus de tamaño moderado, siempre que estos presenten calidad y coherencia temática. En este contexto, el tamaño del corpus empleado resulta suficiente para desarrollar y validar un prototipo funcional del sistema de recomendación propuesto.

3.3 Instrumentos

La recolección de datos en esta investigación se llevó a cabo mediante técnicas automatizadas de extracción y procesamiento de información contenida en resoluciones judiciales. A continuación, se detallan las técnicas e instrumentos utilizados:

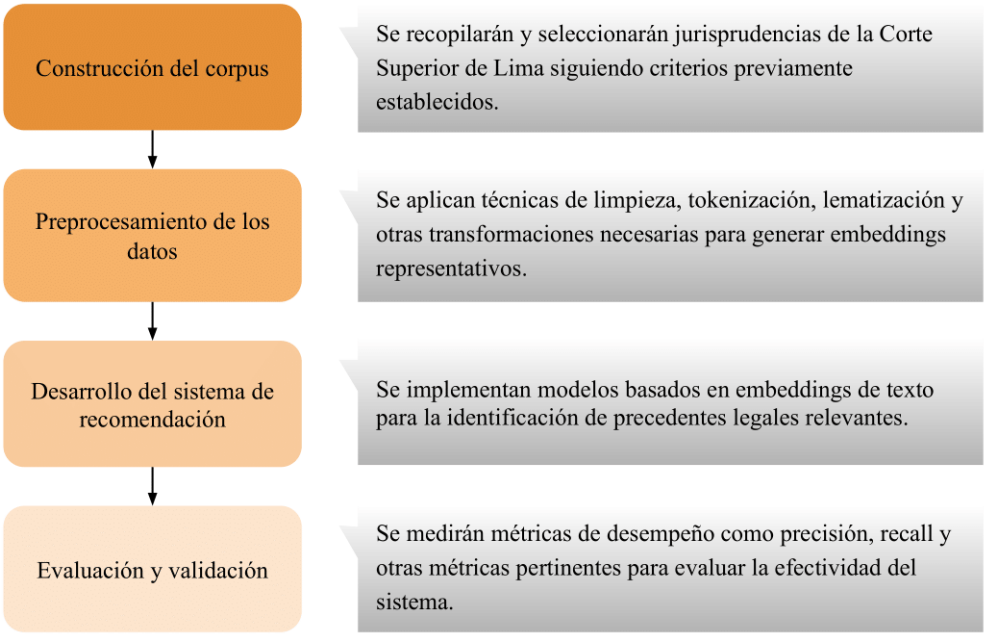
- **Web Scraping y Extracción de Datos:** Se emplearán herramientas como Scrapy y BeautifulSoup para la extracción automatizada de resoluciones judiciales desde fuentes oficiales en línea.
- **Procesamiento de Documentos PDF:** Se implementarán algoritmos en Python utilizando bibliotecas como PyPDF2 y PDFMiner para la extracción de texto estructurado a partir de documentos en formato PDF.
- **Preprocesamiento de Texto:** Se aplicarán técnicas de limpieza, tokenización y lematización para normalizar los textos jurídicos, asegurando su calidad y consistencia en el análisis.
- **Evaluación del Sistema:** Se medirán el desempeño y la efectividad del sistema de recomendación mediante métricas como Precisión, Recall, F1-Score y NDCG.

3.4 Procedimientos

El diseño de la investigación es experimental, ya que se desarrollará un sistema de recomendación basado en embeddings de texto y se evaluará su desempeño a través de pruebas controladas. Se empleará un enfoque cuantitativo, dado que se medirán métricas de rendimiento y precisión del sistema. Se ilustra mejor en la Figura 3.

Figura 3 ***(Metodología de investigación)***

Metodología de investigación



Nota: Elaboración propia

La metodología de esta investigación se estructura en cuatro etapas principales, las cuales se describen a continuación:

- **Construcción del Corpus**

  En esta etapa se recopilaron y seleccionaron las resoluciones judiciales emitidas por la Sala Laboral de la Corte Superior de Justicia de Lima durante el período 2021–2022, las cuales conformaron el corpus de estudio utilizado para el entrenamiento, indexación y evaluación del sistema de recomendación de jurisprudencia propuesto.

- **Definición de criterios:** Se establecerán criterios de inclusión y exclusión para asegurar la calidad y relevancia de las resoluciones. Se priorizarán aquellos documentos que contengan fundamentos legales completos y que hayan sido citados como precedentes en otros casos.
- **Recopilación:** Los datos serán extraídos del Sistema: **Jurisprudencia Nacional Sistematizada** https://jurisprudencia.pj.gob.pe/jurisprudenciaweb/ en formato HTML y PDF. Se utilizarán técnicas de web scraping (empleando herramientas como Scrapy y BeautifulSoup) para la obtención automatizada de las resoluciones desde la fuente en línea.
- **Verificación y Validación:** Se revisará la integridad y calidad de cada documento, descartando aquellos que presenten errores o información incompleta, con el fin de conformar un corpus robusto y representativo.

- **Preprocesamiento de los Datos**

Una vez construido el corpus, es esencial transformarlo en datos adecuados para el análisis y modelado. Las principales actividades en esta etapa son:

- **Extracción de Texto:** Se realizará la extracción del contenido textual de los documentos, especialmente de los archivos en formato PDF, utilizando algoritmos específicos.
- **Limpieza y Normalización:** Se eliminarán caracteres especiales, se corregirán errores tipográficos y se normalizará el texto (por ejemplo, convirtiendo todas las palabras a minúsculas) para garantizar la uniformidad.
- **Tokenización:** El texto se dividirá en unidades básicas (tokens), que pueden ser palabras o frases, utilizando herramientas de procesamiento de lenguaje natural.
- **Lematización y Stemming:** Se reducirán las palabras a su forma base para minimizar la variabilidad del vocabulario y capturar correctamente los conceptos subyacentes.

- **Eliminación de Stopwords:** Se descartarán palabras comunes que aporten poco valor semántico (como artículos y preposiciones) para centrarse en los términos jurídicos relevantes.

- **Desarrollo del Sistema de Recomendación**

Esta fase consiste en la implementación del sistema de recomendación, que se basa en técnicas de embeddings y algoritmos de similitud. Las actividades incluyen:

- **Generación de Embeddings:** Se utilizarán modelos preentrenados (como BERT, Word2Vec o GloVe) para transformar cada resolución judicial en un vector en un espacio semántico. Estos embeddings capturan la información semántica y contextual del texto.
- **Diseño del Algoritmo de Recomendación:** Se desarrollará un algoritmo que compare los vectores generados (usando, por ejemplo, la similitud coseno) para identificar documentos similares y sugerir precedentes legales relevantes.
- **Optimización y Ajuste:** Se experimentará con diferentes configuraciones y parámetros del modelo para optimizar el desempeño del sistema, evaluando diversas métricas y ajustando según los resultados.

- **Evaluación y Validación**

La última etapa se centra en medir y validar la efectividad del sistema desarrollado. Esta fase incluye:

- **Métricas de Desempeño:** Se evaluará el sistema utilizando indicadores cuantitativos como precisión, recall, F1-score y NDCG, que permitirán determinar cuán efectivas son las recomendaciones en la identificación de precedentes relevantes.
- **Análisis e Interpretación de Resultados:** Se analizarán los resultados obtenidos para identificar fortalezas y debilidades, y se propondrán mejoras para futuras iteraciones del sistema.

3.5 Análisis de datos

Para el análisis de los resultados obtenidos del sistema de recomendación de jurisprudencia, se emplearán diversas técnicas que permitan la representación e interpretación clara de los datos. Entre ellas se destacan:

- **Análisis Estadístico Descriptivo:** Se utilizarán medidas de tendencia central, dispersión y frecuencias para resumir y describir las métricas de evaluación del sistema (por ejemplo, precisión, recall, F1-Score y NDCG).
- **Visualización de Datos:** Se aplicarán herramientas de visualización (como Matplotlib y Seaborn en Python) para generar gráficos de barras, líneas y diagramas de dispersión que faciliten la comparación del desempeño del sistema bajo diferentes configuraciones y técnicas de embeddings.
- **Tabulación de Resultados:** Se presentarán tablas resumen que contengan los valores obtenidos en las evaluaciones, permitiendo una comparación directa entre los distintos modelos y técnicas de preprocesamiento aplicadas.

IV. Resultados y discusión

4.1 Construcción de corpus

A continuación, se desarrolla la documentación de todas las etapas experimentadas según la metodología planteando en el proyecto.

4.1.1 Recuperación de documentos legales

La construcción de un corpus de alta calidad es el pilar fundamental para el desarrollo de cualquier sistema de Procesamiento de Lenguaje Natural, y en el dominio jurídico, esta etapa adquiere una relevancia crítica. El objetivo de este procedimiento fue recopilar un conjunto de datos representativo, limpio y completo que sirviera como base para el entrenamiento y la evaluación de los modelos de recomendación.

Fuente de Datos y Delimitación del Alcance

La fuente primaria para la adquisición de datos fue el portal oficial de Jurisprudencia Nacional Sistematizada del Poder Judicial del Perú, accesible a través de su plataforma web (https://jurisprudencia.pj.gob.pe/jurisprudenciaweb/). Este repositorio centraliza las resoluciones emitidas por las distintas cortes del país, constituyendo la fuente de verdad más fiable para la presente investigación.

Para garantizar la viabilidad del estudio y la coherencia semántica del corpus, se tomó la decisión metodológica de delimitar el alcance a una única Especialidad: Laboral. La justificación para esta selección se basa en tres criterios principales:

1. Alto Volumen y Relevancia Social: El derecho laboral es una de las áreas con mayor carga procesal, generando un volumen significativo de jurisprudencia. Esta abundancia de datos es ideal para el entrenamiento de modelos de embeddings, que se benefician de corpus extensos.

Figura 4: Jurisprudencia Nacional Sistematizada



Fuente: portada de Jurisprudencia Nacional Sistematizada

2. Complejidad y Repetitividad: Los casos laborales, aunque diversos, a menudo giran en torno a un conjunto definido de conceptos (despidos, beneficios sociales, indemnizaciones), lo que presenta un escenario idóneo para que un sistema de IA aprenda a identificar patrones y precedentes relevantes.
3. Control del Dominio Lingüístico: Enfocarse en una sola especialidad permite trabajar con un vocabulario y un contexto jurídico homogéneo, reduciendo el ruido semántico que se produciría al mezclar terminología de áreas tan dispares como el derecho penal o civil.

Dentro de la especialidad laboral, se procedió a recopilar documentos de todas las materias disponibles para asegurar que el corpus final fuera diverso y representativo de la casuística de dicha área.

Metodología de Extracción Automatizada en Dos Fases

Dada la estructura de la plataforma web, se diseñó un proceso de extracción automatizada (web scraping) en dos fases distintas para asegurar la integridad y la correcta vinculación de los datos.

Fase 1: Extracción de Metadatos Estructurados

En esta primera fase, se desarrolló un script utilizando las librerías de Python Scrapy y BeautifulSoup para navegar sistemáticamente por el portal de jurisprudencia. El script automatizó los siguientes pasos:

1. Filtrado de la Búsqueda: Se aplicó el filtro de "Especialidad: Laboral" en la interfaz del portal.
2. Iteración y Recopilación: El script navegó a través de las páginas de resultados, extrayendo de cada entrada la información estructurada visible, o metadatos. Los campos clave recopilados en esta etapa fueron: Apelación (utilizado como identificador único), Especialidad y Materia.
3. Captura del Enlace al Contenido: De manera crucial, junto con los metadatos, se extrajo y almacenó el enlace directo al documento completo, que en la mayoría de los casos se encontraba en formato PDF.

El resultado de esta fase fue un archivo CSV preliminar que actuó como un índice maestro, conteniendo todos los metadatos de los casos y la ruta para acceder a su contenido textual completo.

Fase 2: Descarga y Procesamiento de Documentos PDF

Con el índice de metadatos ya consolidado, se ejecutó una segunda fase automatizada, cuyo objetivo era obtener el texto íntegro de cada resolución:

1. Lectura del Índice: Un segundo script leyó el archivo CSV generado en la fase anterior.
2. Descarga Programática: Por cada fila del índice, el script utilizó el enlace almacenado para descargar el archivo PDF correspondiente y lo guardó localmente, asociándolo a su ID (Apelación).
3. Extracción del Texto Crudo: Utilizando la librería PyPDF2, se procesó cada archivo PDF descargado para extraer su contenido textual completo. Este texto, sin procesar, se guardó en memoria.

El paso final consistió en fusionar los resultados de ambas fases. El texto extraído de cada PDF se vinculó con su respectiva fila de metadatos utilizando el ID (Apelación) como clave de unión. Este proceso dio como resultado la creación del

corpus final: un único archivo CSV (corpus_final_resumido.csv) donde cada fila representa una resolución judicial completa, conteniendo tanto sus metadatos (Apelación, Especialidad, Materia) como su contenido textual íntegro (texto_completo). Esta estructura de datos unificada y enriquecida se convirtió en la base para las subsiguientes etapas de preprocesamiento, modelado y evaluación del sistema de recomendación.

4.1.2 Exploración de los datos

Una vez compilado el corpus documental, se procedió a una fase inicial de Análisis Exploratorio de Datos (EDA) con el objetivo de comprender sus características fundamentales, verificar su calidad y prepararlo para las etapas subsecuentes de modelado y evaluación.

Previo al desarrollo y entrenamiento de los modelos, se realizó una verificación exhaustiva de la calidad e integridad del corpus documental compilado. Este corpus, compuesto por 6,848 registros correspondientes a resoluciones de apelación en materia laboral destinados al entrenamiento, fue sometido a un análisis preliminar para identificar posibles inconsistencias o datos faltantes.

Un hallazgo fundamental de esta verificación fue la completitud total del conjunto de datos. Se constató la ausencia absoluta de valores nulos (null) en la totalidad de las columnas que conforman cada registro. Esta característica representa una ventaja metodológica significativa, ya que elimina la necesidad de aplicar técnicas de imputación para estimar datos faltantes o de recurrir a la eliminación de filas incompletas, procedimientos que podrían introducir sesgos o reducir el tamaño efectivo del corpus. Por consiguiente, se garantiza que todos los análisis y modelos subsecuentes se benefician del 100% de la información originalmente recopilado.

Figura 5 Corpus de jurisprudencias

[7]:

	Apelación	Especialidad	Materia	texto_completo	Texto_resumido
0	Apelación 000011-2021	Especialidad: Laboral	Materia: Contratos	CORTE SUPERIOR DE JUSTICIA DE LIMA\nCUARTA SAL...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**CUAR...
1	Apelación 000016-2022	Especialidad: Laboral	Materia: Reposición	8° SALA LABORAL\nEXPEDIENTE : 00016-2022-0-180...	Este documento corresponde al encabezado (cará...
2	Apelación 000030-2020	Especialidad: Laboral	Materia: Indemnización Laboral	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...
3	Apelación 000031-2019	Especialidad: Laboral	Materia: Pago de Beneficios Sociales	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...
4	Apelación 000037-2018	Especialidad: Laboral	Materia: Pago de Beneficios Sociales	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...
...	...	...	...	...	...
6843	Queja 000223-2019	Especialidad: Laboral	Materia: Incumplimiento de Disposiciones y Nor...	1° SALA LABORAL PERMANENTE\nEXPEDIENTE : 00223...	**RESUMEN DE RESOLUCIÓN JUDICIAL**\n\n**EXPEDI...
6844	Queja 000233-2019	Especialidad: Laboral	Materia: Pago de Beneficios Sociales	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...
6845	Queja 000376-2019	Especialidad: Laboral	Materia: Ofrecimiento de Pago	1° SALA LABORAL PERMANENTE\nEXPEDIENTE : 00376...	**EXPEDIENTE:** 00376-2019-0-1801-SP-LA-01\n**...
6846	Queja 000381-2019	Especialidad: Laboral	Materia: Contratos	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...
6847	Queja 000408-2019	Especialidad: Laboral	Materia: Incumplimiento de Disposiciones y Nor...	CORTE SUPERIOR DE JUSTICIA DE LIMA\nSÉPTIMA SA...	**EXPEDIENTE:** N° 00408-2019-0-1801-JR-LA-60 ...

6848 rows × 5 columns

[8]:
```
df.to_csv('corpus_final_resumido.csv', index=False)
```

```
--- Información General del DataFrame ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6848 entries, 0 to 6847
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Apelación       6848 non-null   object
 1   Especialidad    6848 non-null   object
 2   Materia         6848 non-null   object
 3   texto_completo  6848 non-null   object
 4   Texto_resumido  6848 non-null   object
dtypes: object(5)
memory usage: 267.6+ KB

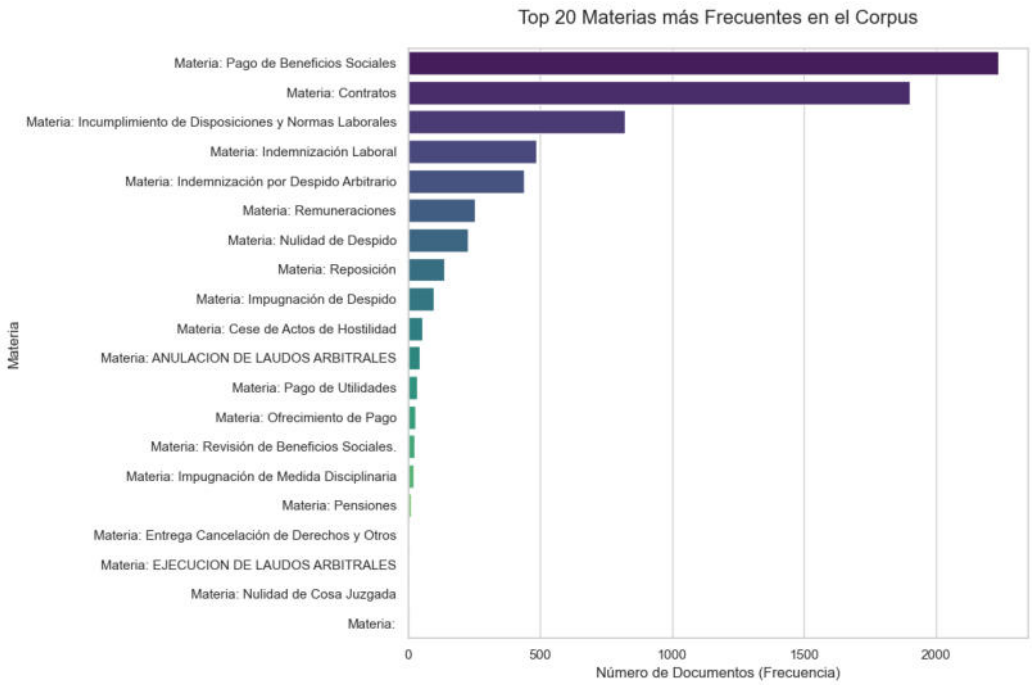
--- Conteo de Valores Nulos por Columna ---
Apelación         0
Especialidad      0
Materia           0
texto_completo    0
Texto_resumido    0
dtype: int64
```

Adicionalmente, se validó la correcta asignación de los tipos de datos para cada columna. Las columnas que contienen información textual tales como query_apelacion, query_resumen, candidate_apelacion, candidate_resumen y Materia fueron adecuadamente representadas con el tipo de dato object (cadena de texto) dentro del entorno de análisis de datos. Esta correspondencia asegura el manejo apropiado de la información textual en las fases posteriores. Dada la adecuación de los tipos de datos asignados, no se requirieron conversiones adicionales en esta etapa preparatoria.

Habiendo confirmado la integridad estructural, la completitud y la correcta tipificación de los datos del corpus, se concluye esta fase de verificación de calidad.

El siguiente paso dentro del Análisis Exploratorio de Datos se enfoca en la comprensión del contenido semántico y la distribución de las características documentales. Este análisis se iniciará con un examen detallado de la columna Materia, identificada como un metadato clave para caracterizar la diversidad temática del corpus.

Figura 6 Distribución de materias



Distribución Altamente Desbalanceada

El corpus presenta una distribución de "ley de potencias" (o *long tail*), lo cual es muy común en datos del mundo real. Las dos materias principales, "Pago de Beneficios Sociales" y "Contratos", suman 4,136 casos, lo que representa aproximadamente el **60.4%** del total de tu dataset. El resto de las materias se distribuyen en una "larga cola" con frecuencias mucho menores.

Inconsistencias en el Etiquetado

Se observan problemas de formato que deben ser estandarizados para que la Materia sea un metadato fiable:

a) **Prefijo Redundante:** Muchas etiquetas comienzan con "Materia: ".

b) **Capitalización:** Hay inconsistencias como "ANULACION DE LAUDOS ARBITRALES" (mayúsculas) frente a otras en formato de título.

c) **Puntuación:** Algunas etiquetas terminan con un punto, como "Revisión de Beneficios Sociales.".

d) **Etiquetas Vacías:** Existe una categoría sin nombre (' ') que agrupa 3 documentos.

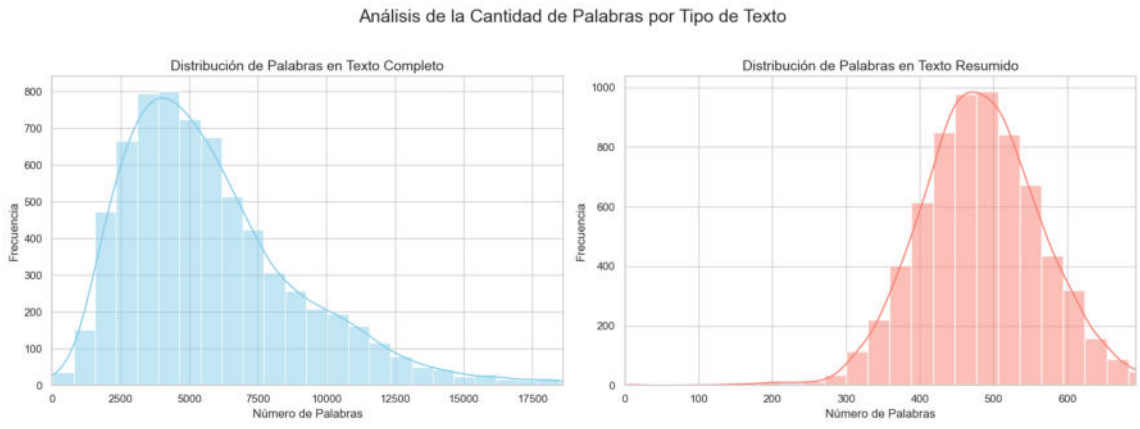
4.1.3 Implicaciones para el Sistema de Recomendación

a) **Rendimiento del Modelo:** El sistema será naturalmente más preciso y tendrá más "experiencia" en las materias mayoritarias. Para las materias de baja frecuencia, la relevancia de las recomendaciones podría ser menor, ya que el modelo tiene menos ejemplos de los que aprender.

b) **Necesidad de Limpieza:** Es **imperativo** limpiar y estandarizar las etiquetas de Materia. Si no lo hacemos, el sistema podría tratar "Materia: Contratos" y "Contratos" como dos categorías distintas, lo cual afectaría negativamente cualquier lógica de filtrado o re-clasificación que implementemos en el futuro.

Ahora que entendemos la distribución de las categorías, el siguiente paso lógico es analizar el contenido de las columnas de texto. Necesitamos saber la longitud y la variabilidad de texto_completo y Texto_resumido.

Figura 7 Distribución de textos en campos texto_completo y Texto_resumido



Texto_completo (Gráfico Izquierdo):

Distribución: Como esperábamos, la longitud de los documentos originales es muy variable. La mayoría se concentra entre las 2,500 y 7,500 palabras, pero existe una "larga cola" de documentos mucho más extensos.

Estadísticas: El promedio es de 5,955 palabras, pero la desviación estándar es muy alta (3,648), lo que confirma esta variabilidad. El documento más largo tiene más de 38,000 palabras, normal en textos legales.

Texto_resumido (Gráfico Derecho):

Distribución: Este es el resultado más revelador e importante. La distribución de la longitud de los resúmenes generados por Gemini es casi una distribución normal perfecta. Está claramente centrada alrededor de las 480 palabras.

Estadísticas: La desviación estándar es muy baja (85 palabras), lo que demuestra una notable consistencia. Gemini ha generado resúmenes de una longitud muy predecible, independientemente de la longitud del documento original.

Calidad del Proceso de Resumen: Podemos confirmar que el proceso de resumen fue de alta calidad y muy consistente. Esto es ideal, ya que significa que las consultas de los usuarios (basadas en resúmenes) tendrán una longitud y un nivel de detalle similares, lo que facilita que el modelo de IA las compare.

Viabilidad del Enfoque: El plan de usar texto_completo para entrenar (crear la base de conocimiento) y Texto_resumido para las pruebas (simular consultas) es perfecto. Tenemos un corpus rico en detalles y consultas consistentes y bien formadas.

Hemos completado el análisis exploratorio. Ahora estamos listos para actuar sobre estos hallazgos y preparar los datos para el modelo.

4.2 Preprocesamiento de los datos

Una vez construido y explorado el corpus inicial, la siguiente fase crucial es el preprocesamiento de los datos. El objetivo de esta etapa es transformar el texto crudo y los metadatos en un formato limpio, estandarizado y optimizado para ser procesado por los modelos de embeddings. El análisis exploratorio previo reveló la necesidad de abordar dos áreas principales: la estandarización de los metadatos categóricos y la normalización del contenido textual para reducir el ruido semántico.

El proceso se estructuró en los siguientes pasos metodológicos:

4.2.1 Estandarización de metadatos categóricos

El análisis de la columna Materia evidenció inconsistencias en el formato de las etiquetas, tales como la inclusión del prefijo "Materia: ", variaciones en el uso de mayúsculas y minúsculas, y la presencia de puntuación superflua. Para rectificar esto y asegurar que la columna sea un metadato fiable para futuras tareas de filtrado y análisis, se aplicó una función de limpieza que realizó las siguientes transformaciones:

1) **Eliminación de prefijos:** Se removió el prefijo "Materia: " de todas las etiquetas para aislar el concepto principal.
2) **Normalización de mayúsculas:** Se estandarizó el texto a formato de título (ej. "Pago De Beneficios Sociales"), unificando las diferentes capitalizaciones.
3) **Eliminación de puntuación y espacios:** Se suprimieron los puntos al final de las etiquetas y se eliminaron los espacios en blanco innecesarios al inicio o al final.

El resultado de este paso fue la creación de una nueva columna, materia_limpia, que contiene etiquetas categóricas consistentes y listas para ser utilizadas.

4.2.2 Limpieza y normalización del contenido textual

El contenido de las columnas texto_completo y Texto_resumido presentaba ruido inherente al proceso de extracción, como caracteres especiales (ej. asteriscos), saltos de línea y espaciado inconsistente. Para preparar estos textos para la vectorización semántica, se implementó una función de limpieza con los siguientes objetivos:

i) **Conversión a minúsculas:** Todo el texto fue convertido a minúsculas para asegurar que el modelo no trate la misma palabra con diferente capitalización (ej. "ley" y "Ley") como dos conceptos distintos.
ii) **Eliminación de caracteres especiales:** Se eliminaron caracteres no alfanuméricos, como los asteriscos (*), que no aportan valor semántico y pueden ser interpretados como ruido por el modelo.
iii) **Normalización de espacios en blanco:** Se reemplazaron todos los saltos de línea (\n), tabulaciones (\t) y secuencias de múltiples espacios con un único espacio en blanco. Esto garantiza un texto fluido y estructuralmente homogéneo.

Estas transformaciones se aplicaron para crear dos nuevas columnas: texto_completo_limpio y texto_resumido_limpio. Al finalizar esta etapa, el DataFrame se encontraba completamente preprocesado, con un corpus de texto normalizado y metadatos estandarizados, constituyendo la base fundamental para la siguiente fase de generación de embeddings.

Código:

```
import re

def limpiar_materia(materia_texto: str) -> str:
    """Estandariza las etiquetas de la columna 'Materia'."""
    if not isinstance(materia_texto, str): return ""
    texto = re.sub(r'^materia:\s*', '', materia_texto, flags=re.IGNORECASE)
    texto = texto.title()
    texto = texto.rstrip('.')
    return texto.strip()

def limpiar_texto(texto: str) -> str:
```

```
    """Limpia y estandariza el contenido de los textos para el modelo."""
    if not isinstance(texto, str): return ""
    texto = texto.lower()
    texto = texto.replace('*', '')
    texto = re.sub(r'[\n\t]', ' ', texto)
    texto = re.sub(r'\s+', ' ', texto).strip()
    return texto

# Aplicación sobre el DataFrame 'df'
# df['materia_limpia'] = df['Materia'].apply(limpiar_materia)
# df['texto_completo_limpio'] = df['texto_completo'].apply(limpiar_texto)
# df['texto_resumido_limpio'] = df['Texto_resumido'].apply(limpiar_texto)
```

4.3 Desarrollo de sistema de recomendación

Esta fase, central en la presente investigación, aborda la implementación técnica del sistema de recomendación, de acuerdo con el tercer bloque de la metodología propuesta. El objetivo es traducir el corpus preprocesado en un sistema funcional capaz de identificar y sugerir precedentes legales relevantes. El desarrollo se fundamenta en la generación de embeddings de texto y el cálculo de similitud semántica.

El proceso se estructuró en torno a la implementación de un primer modelo de referencia, basado en una arquitectura Transformer, que servirá como base para comparaciones posteriores con otros enfoques del estado del arte.

4.3.1 Modelo 1: Búsqueda semántica basada en transformers (SentenceTransformer)

Para la implementación inicial, se seleccionó un modelo basado en la arquitectura Transformer, específicamente ***paraphrase-multilingual-MiniLM-L12-v2***. La elección de este modelo se justifica por varios factores clave:

- **Eficiencia computacional:** Es un modelo ligero y rápido, lo que permite un procesamiento ágil del corpus completo sin requerir hardware especializado de alto rendimiento.
- **Rendimiento en tareas de similitud:** Ha sido pre-entrenado y afinado (fine-tuned) específicamente para tareas de **Sentence Similarity**, es decir,

para determinar cuán semánticamente parecidas son dos piezas de texto. Esto lo hace ideal para la tarea de encontrar jurisprudencia similar.

- **Capacidad multilingüe:** Aunque nuestro corpus está en español, la naturaleza multilingüe del modelo asegura una comprensión robusta de las estructuras lingüísticas, beneficiando la calidad de las representaciones vectoriales generadas.

La implementación de este modelo se dividió en dos etapas principales: la generación de la base de datos vectorial y el diseño del algoritmo de recomendación.

Generación de la base de datos vectorial (Embeddings)

El núcleo del sistema reside en la transformación de cada documento legal del corpus en una representación numérica que capture su significado. Para este fin, se utilizó el modelo SentenceTransformer cargado para procesar la columna texto_completo_limpio.

El uso del texto completo, en lugar del resumido, es una decisión metodológica deliberada. El texto completo contiene la totalidad de los argumentos, el razonamiento jurídico y los matices contextuales, lo que permite la creación de vectores de alta fidelidad. Esta riqueza semántica es fundamental para construir una base de conocimiento robusta, capaz de identificar similitudes sutiles que un resumen podría omitir.

El proceso se ejecutó de la siguiente manera:

- Se tomó la lista de todos los documentos de la columna **texto_completo_limpio**.
- Se utilizó la función **model.encode()** para convertir cada documento en un vector de 384 dimensiones.
- El resultado fue una matriz NumPy, denominada **document_vectors**, con dimensiones (6848, 384), donde cada fila corresponde a un documento del corpus.

Esta matriz constituye la base de conocimiento estática del sistema. Para asegurar la eficiencia y la reproducibilidad, tanto el modelo inicializado como la matriz de vectores generada fueron guardados en disco como "artefactos", eliminando la necesidad de recalcularlos en cada sesión de uso.

Diseño del algoritmo de recomendación

Con la base de conocimiento vectorial establecida, se diseñó una función (encontrar_similares) que encapsula la lógica para procesar una nueva consulta y devolver los precedentes más relevantes. El algoritmo opera de la siguiente manera:

1. **Recepción y limpieza de la consulta:** La función recibe como entrada un texto de consulta en crudo (**simulando el Texto_resumido** de un caso). A este texto se le aplica la misma función **limpiar_texto** utilizada durante el preprocesamiento del corpus, garantizando la consistencia entre la consulta y los datos almacenados.
2. **Vectorización de la consulta:** La consulta limpia es procesada por el mismo modelo SentenceTransformer cargado para convertirla en un vector de 384 dimensiones. Esto posiciona la consulta del usuario dentro del mismo espacio semántico que los documentos del corpus.
3. **Cálculo de similitud semántica:** Para medir la relevancia entre la consulta y cada documento del corpus, se utiliza la métrica de similitud coseno. Esta medida calcula el coseno del ángulo entre el vector de la consulta y cada uno de los 6,848 vectores de la base de datos, arrojando una puntuación entre -1 y 1 (o 0 y 1 en este contexto) donde un valor más cercano a 1 indica una mayor similitud semántica.

   En el Procesamiento de Lenguaje Natural, la similitud coseno es la métrica preferida sobre otras, como la distancia euclidiana, por una razón fundamental: es insensible a la magnitud de los vectores y se centra exclusivamente en su dirección.

   En el contexto de textos legales, la magnitud de un vector a menudo se correlaciona con la longitud del documento. Un documento de 10,000 palabras tendrá un vector con una magnitud mayor que un resumen de 500 palabras, incluso si ambos discuten exactamente los mismos conceptos legales. Si se

utilizara la distancia euclidiana, la diferencia de longitud los consideraría muy distantes.

La similitud coseno resuelve este problema midiendo el coseno del ángulo (θ) entre dos vectores. Si dos vectores apuntan en la misma dirección, el ángulo entre ellos es 0°, y el coseno es 1 (máxima similitud), sin importar si un vector es más "largo" que el otro. Si son ortogonales (90°), su similitud es 0, indicando que no comparten contexto semántico.

La similitud coseno entre dos vectores, A y B, se calcula mediante la siguiente fórmula:

$$\text{similitud}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Donde:

- A·B es el producto punto de los vectores A y B.
- ‖A‖ y ‖B‖ son las normas euclidianas (o magnitudes) de los vectores A y B, respectivamente.
- A i y B i son los componentes de los vectores en la dimensión i.

El resultado de esta operación es un valor en el rango de [-1, 1]. En el contexto de los embeddings generados por SentenceTransformer, que son no negativos, el rango práctico es de [0, 1].

- Una puntuación de 1 indica que los vectores son idénticos en orientación (máxima similitud semántica).
- Una puntuación de 0 indica que los vectores son ortogonales (no hay similitud)

4. **Ranking y selección de resultados:** Los documentos del corpus se ordenan de mayor a menor según su puntuación de similitud. El sistema está configurado para retornar las N recomendaciones principales (por defecto, N=5). Se implementó una lógica de auto-exclusión para que, si el texto de consulta es idéntico al de un documento en el corpus (por ejemplo, al usar el resumen de un caso para buscar), ese mismo caso no se recomiende a sí mismo.
5. **Presentación de resultados:** Finalmente, el sistema devuelve un listado de los casos recomendados, presentando su ID (Apelación), la materia_limpia, el

Texto_resumido y el porcentaje de similitud, facilitando al usuario la interpretación de la relevancia de cada resultado.

Esta implementación inicial del sistema de recomendación semántico establece una base sólida y un punto de referencia (baseline) de alto rendimiento. Los siguientes pasos de esta investigación se centrarán en implementar y evaluar dos enfoques adicionales un modelo de búsqueda léxica (BM25) y un sistema de búsqueda híbrida que combine ambos para determinar, tal como se plantea en los objetivos, qué arquitectura ofrece el mayor rendimiento en la identificación de precedentes relevantes en el contexto de la jurisprudencia peruana.

Código:

```
# -- Fragmento conceptual de la implementación --

from sklearn.metrics.pairwise import cosine_similarity

# vector_consulta: Un array de NumPy de forma (1, 384)
# document_vectors: La matriz de NumPy de forma (6848, 384)

# La función calcula la similitud de la consulta contra todos los documentos
similitudes = cosine_similarity(vector_consulta, document_vectors)

# El resultado es una matriz de (1, 6848), que se aplana (flatten)
# para obtener un array simple de puntuaciones.
puntuaciones_de_similitud = similitudes.flatten()
```

Para la implementación de este cálculo, se utilizó la función **cosine_similarity** de la librería de Python scikit-learn, la cual está altamente optimizada para operaciones matriciales. El proceso en el sistema es el siguiente:

- Se toma el vector_consulta (dimensiones: 1x384).
- Se toma la matriz completa document_vectors (dimensiones: 6848x384).
- La función cosine_similarity calcula de manera eficiente el producto punto entre el vector de consulta y cada uno de los 6,848 vectores del corpus, y lo normaliza por sus magnitudes.

El resultado es un array de 6,848 puntuaciones de similitud, una por cada documento del corpus. Este array es la base sobre la cual se realiza el ranking para extraer las recomendaciones más relevantes.

4.3.2 Modelo 2: Búsqueda léxica con Okapi BM25)

Si bien el modelo basado en Transformers es altamente eficaz para capturar la similitud semántica y conceptual, presenta una limitación inherente: su capacidad para garantizar la recuperación de documentos que contengan términos clave específicos, como números de artículos, nombres propios o terminología técnica precisa, puede ser subóptima. Para abordar esta brecha y complementar la búsqueda semántica, se implementó un segundo modelo basado en un enfoque léxico: Okapi BM25.

BM25 (Best Matching 25) es un algoritmo de ranking de la familia TF-IDF, considerado el estado del arte en la Recuperación de Información para la búsqueda por palabras clave. A diferencia de la búsqueda semántica que mide la similitud de significado, BM25 calcula una puntuación de relevancia para cada documento basándose en los términos de la consulta. Su fortaleza radica en su capacidad para priorizar documentos que contienen las palabras exactas buscadas, ponderando inteligentemente su importancia.

1. Justificación del modelo y preparación de datos

La implementación de BM25 requiere que el corpus textual esté tokenizado, es decir, dividido en una lista de palabras individuales. Para este modelo, se tomó la decisión metodológica de utilizar la columna texto_resumido_limpio. Esta elección se justifica por dos razones:

- Eficiencia Computacional: Indexar los resúmenes es significativamente más rápido y consume menos memoria que procesar los textos completos.
- Alineación con el Caso de Uso: Las consultas de los usuarios suelen ser cortas y directas. La búsqueda por palabras clave en los resúmenes se alinea mejor con este tipo de interacción, ya que los términos más importantes del documento suelen estar presentes en su resumen.

El primer paso consistió en tokenizar cada documento de la columna texto_resumido_limpio, dividiéndolo en una lista de palabras. El resultado fue una lista de listas que sirvió como entrada para el modelo.

2. Indexación del corpus y guardado del artefacto

Una vez tokenizado el corpus, se procedió a "entrenar" el modelo BM25Okapi utilizando la librería rank-bm25. Es importante destacar que el "entrenamiento" en este contexto no implica un proceso de aprendizaje profundo. En cambio, consiste en una fase de indexación, durante la cual el algoritmo calcula las estadísticas necesarias sobre el corpus, principalmente:

- Frecuencia de Término (TF): Cuántas veces aparece cada palabra en cada documento.
- Frecuencia Inversa de Documento (IDF): Qué tan rara o común es cada palabra en la totalidad del corpus. Las palabras raras reciben un mayor peso.

TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF, acrónimo de Term Frequency-Inverse Document Frequency (Frecuencia de Término–Frecuencia Inversa de Documento), es una de las técnicas de ponderación más fundamentales y efectivas en la Recuperación de Información y el Procesamiento de Lenguaje Natural. Su objetivo es medir la importancia de un término no solo por su frecuencia dentro de un documento específico, sino también por su relevancia en el contexto de un corpus completo de documentos.

La intuición detrás de TF-IDF es que los términos más significativos para un documento son aquellos que aparecen frecuentemente en él (alta Frecuencia de Término), pero que son, a su vez, raros en el resto de los documentos del corpus (alta Frecuencia Inversa de Documento). Esto permite mitigar el peso de palabras muy comunes (como artículos o preposiciones) y dar mayor importancia a los términos que son verdaderamente distintivos del documento en cuestión.

El cálculo de la puntuación TF-IDF se descompone en dos partes:

a. Frecuencia de término (TF)

La Frecuencia de Término mide la relevancia de un término dentro de un documento específico. Representa qué tan a menudo aparece una palabra en un texto. Aunque puede calcularse como un conteo bruto, la práctica estándar es normalizar esta frecuencia para evitar un sesgo hacia los documentos más largos. La fórmula normalizada es:

$$\mathrm{TF}(t,d) = \frac{\text{Número de veces que el término } t \text{ aparece en el documento } d}{\text{Número total de términos en el documento } d}$$

Donde:

- t es el término de interés.
- d es el documento específico que se está analizando.

Por ejemplo, si el término "indemnización" aparece 3 veces en una resolución judicial (d) de 100 palabras, su TF sería 3/100=0.03.

b. Frecuencia inversa de documento (IDF)

La Frecuencia Inversa de Documento mide qué tan informativa o rara es una palabra en todo el corpus. Si un término aparece en muchos documentos, su valor IDF será bajo, indicando que no es un buen discriminador. Por el contrario, si un término aparece en muy pocos documentos, su valor IDF será alto, señalando que es un término más específico y relevante.

La fórmula para calcular el IDF es:

$$\mathrm{IDF}(t,D) = \log\left(\frac{N}{df_t}\right)$$

Donde:

- t es el término de interés.
- D es el corpus completo de documentos.
- N es el número total de documentos en el corpus.
- df t es el número de documentos del corpus en los que aparece el término t (del inglés, document frequency).

El uso del logaritmo ayuda a atenuar el efecto de la escala; de lo contrario, la diferencia de IDF entre términos muy raros sería desproporcionadamente grande. Es común aplicar una variante suavizada de la fórmula para evitar divisiones por cero si un término no aparece en ningún documento (df t=0) y para dar un peso ligeramente mayor a la escala, conocida como "IDF Smooth":

$$\text{IDF Smooth}(t, D) = \log\left(\frac{N+1}{df_t+1}\right) + 1$$

La puntuación final TF-IDF para un término en un documento es simplemente el producto de su TF y su IDF. Esta puntuación final refleja tanto la importancia del término en su documento local como su rareza en el corpus global.

La fórmula completa es:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Un término tendrá una puntuación TF-IDF alta si:

- Tiene una alta frecuencia en el documento d (TF alto).
- Es raro en el corpus D (IDF alto).

Por el contrario, la puntuación será baja si el término es muy común en todo el corpus (como "el", "la", "de"), resultando en un IDF cercano a cero, o si el término es raro en el documento mismo (TF bajo).

Al final, cada documento del corpus puede ser representado como un vector, donde cada dimensión corresponde a un término del vocabulario y el valor en esa dimensión es la puntuación TF-IDF de ese término en el documento.

El objeto bm25 resultante de este proceso contiene el índice completo y está listo para realizar búsquedas de manera eficiente. Siguiendo la metodología de creación de artefactos desplegables, este objeto fue serializado utilizando la librería pickle y guardado en el disco con el nombre bm25_model.pkl dentro del directorio modelo_artefactos.

3. Algoritmo de búsqueda y cálculo de relevancia

Se diseñó una función de búsqueda (encontrar_similares_bm25) que opera de la siguiente manera:

- Recepción y Tokenización de la Consulta: La función recibe un texto de consulta en crudo, le aplica la misma función limpiar_texto y lo tokeniza dividiéndolo en una lista de palabras.
- Cálculo de Puntuaciones: Se utiliza el método .get_scores() del objeto bm25 para calcular una puntuación de relevancia para cada uno de los 6,848 documentos del corpus en relación a la consulta tokenizada.
- Ranking de Resultados: Los documentos se ordenan de mayor a menor según su puntuación BM25, y se seleccionan los N mejores.

Es fundamental señalar que la puntuación generada por BM25 no es una probabilidad ni un porcentaje de similitud, sino una medida de relevancia aditiva y no acotada, cuyo único propósito es el ranking relativo de los documentos para una consulta específica.

La implementación de este modelo léxico establece el segundo pilar del sistema de recomendación. Con un motor de búsqueda semántico y otro léxico ya construidos y guardados como artefactos, el siguiente paso lógico en la investigación es desarrollar un sistema de búsqueda híbrida que combine las fortalezas de ambos enfoques para lograr una relevancia superior.

Código:

```
# -- Fragmento conceptual de la implementación --

from rank_bm25 import BM25Okapi
import pickle

# 1. Tokenización (se asume que df['texto_resumido_limpio'] existe)
tokenized_corpus = [doc.split(" ") for doc in df['texto_resumido_limpio'].tolist()]

# 2. Indexación ("Entrenamiento")
bm25 = BM25Okapi(tokenized_corpus)

# 3. Guardado del artefacto
```

```
# with open('modelo_artefactos/bm25_model.pkl', 'wb') as f:
#     pickle.dump(bm25, f)

# 4. Búsqueda
# query = "indemnización por despido arbitrario"
# tokenized_query = query.split(" ")
# doc_scores = bm25.get_scores(tokenized_query)
# top_n_indices = doc_scores.argsort()[::-1][:5]
```

4.3.3 Modelo 3: Búsqueda híbrida y fusión de rankings (Reciprocal Rank Fusion)

Tras la implementación de los modelos de búsqueda semántica (basado en Transformers) y léxica (BM25), se evidenció la complementariedad de sus fortalezas y debilidades. Mientras el modelo semántico sobresale en la comprensión del significado contextual y las relaciones conceptuales, el modelo léxico garantiza una alta precisión en la recuperación de documentos que contienen palabras clave específicas. Sin embargo, depender de un solo enfoque introduce limitaciones inherentes.

Para superar estas limitaciones y construir un sistema de recomendación más robusto y de mayor rendimiento, se desarrolló un tercer modelo basado en una arquitectura de búsqueda híbrida. El objetivo de este enfoque es combinar los resultados de ambos motores de búsqueda (semántico y léxico) para producir un único ranking final que sea superior a la suma de sus partes.

1. Fundamento de la búsqueda híbrida

La búsqueda híbrida opera bajo el principio de que los mejores resultados se obtienen al fusionar las señales de relevancia de diferentes paradigmas de búsqueda. Una consulta puede tener componentes tanto conceptuales como literales. Al ejecutar la búsqueda en paralelo, el sistema se asegura de capturar ambos aspectos:

- La búsqueda semántica aporta documentos que son contextualmente relevantes, incluso si no comparten las mismas palabras clave.

- La búsqueda léxica aporta documentos que contienen los términos exactos de la consulta, garantizando la precisión literal.

Para fusionar los dos rankings generados por estos sistemas, se seleccionó el algoritmo Reciprocal Rank Fusion (RRF).

2. Reciprocal rank fusion (RRF)

RRF es un método de fusión de rankings simple pero extraordinariamente efectivo que no requiere entrenamiento ni ajuste de parámetros complejos. Su lógica se basa en la idea de que la importancia de un documento aumenta cuantas más listas de resultados lo contengan y cuanto más alta sea su posición en ellas.

La puntuación RRF para un documento d se calcula de la siguiente manera:

$$\mathrm{RRF}(d) = \sum_{i=1}^{m} \frac{1}{k + \mathrm{rank}_i(d)}$$

Donde:

- m es el número de sistemas de ranking que se están fusionando (en nuestro caso, m=2).
- rank i (d) es la posición (rango) del documento d en la lista de resultados del sistema i. Si el documento no aparece en la lista, su contribución es cero.
- k es una constante utilizada para disminuir el impacto de los documentos que aparecen en posiciones muy bajas en los rankings. Un valor comúnmente utilizado en la literatura es k=60.

Los documentos que aparecen en posiciones altas (ej. rank=1, rank=2) en múltiples listas reciben una puntuación RRF significativamente mayor, lo que los impulsa a la cima del ranking final fusionado.

3. Implementación del algoritmo de búsqueda híbrida

Se diseñó una función (busqueda_hibrida) que orquesta todo el proceso en los siguientes pasos:

a. Recuperación Paralela de Candidatos: Al recibir un texto de consulta, el sistema lo procesa y lo envía simultáneamente a los dos motores de búsqueda:
   - El modelo SentenceTransformer devuelve una lista de los k_candidates (ej. 50) documentos semánticamente más similares.
   - El modelo BM25 devuelve una lista de los k_candidates documentos léxicamente más relevantes. Se recupera un número mayor de candidatos que los 5 finales para proporcionar al algoritmo de fusión un conjunto rico de documentos para re-clasificar.

b. Aplicación de Reciprocal Rank Fusion (RRF): El sistema itera sobre las dos listas de resultados. Para cada documento único que aparece en cualquiera de las listas, se calcula su puntuación RRF sumando el inverso de su rango en cada lista, ponderado por la constante k.

c. Re-clasificación y Presentación Final: Los documentos se ordenan de mayor a menor según su puntuación RRF final. El sistema devuelve al usuario las N recomendaciones principales de esta nueva lista fusionada, junto con su puntuación RRF, que ahora representa una medida de relevancia combinada.

La implementación de este sistema híbrido representa la culminación del desarrollo de los modelos de recomendación. Se espera que este enfoque ofrezca el rendimiento más alto y robusto, ya que mitiga las debilidades individuales de los sistemas puramente semánticos o léxicos y responde de manera efectiva a una gama más amplia de consultas de los usuarios. La evaluación de este modelo, en comparación con los dos anteriores, será fundamental para validar la hipótesis central de esta investigación.

Código:

```
# -- Fragmento conceptual de la implementación --
def busqueda_hibrida(texto_consulta, model, document_vectors, bm25_model, df):
```

```
# 1. Búsqueda Semántica
vector_consulta = model.encode([limpiar_texto(texto_consulta)])
similitudes = cosine_similarity(vector_consulta, document_vectors).flatten()
semantic_indices = similitudes.argsort()[::-1][:50]

# 2. Búsqueda Léxica
tokenized_query = limpiar_texto(texto_consulta).split(" ")
doc_scores_bm25 = bm25_model.get_scores(tokenized_query)
lexical_indices = doc_scores_bm25.argsort()[::-1][:50]

# 3. Fusión con RRF
rrf_scores = {}
k = 60

# Puntuar resultados semánticos
for rank, doc_index in enumerate(semantic_indices):
    if doc_index not in rrf_scores: rrf_scores[doc_index] = 0
    rrf_scores[doc_index] += 1 / (k + rank + 1)

# Puntuar resultados léxicos
for rank, doc_index in enumerate(lexical_indices):
    if doc_index not in rrf_scores: rrf_scores[doc_index] = 0
    rrf_scores[doc_index] += 1 / (k + rank + 1)

# 4. Obtener ranking final
sorted_fused_indices = sorted(rrf_scores.keys(), key=lambda x: rrf_scores[x], reverse=True)

# Devolver los N mejores resultados
return df.iloc[sorted_fused_indices[:5]]
```

4.4 Evaluación y validación

La evaluación rigurosa del desempeño de los modelos de recomendación desarrollados en esta tesis específicamente, el modelo de Búsqueda Léxica con **Okapi BM25**, el modelo de Búsqueda Semántica Basada en **Transformers** (SentenceTransformer) utilizando bert-base-multilingual-cased, y el modelo de Búsqueda **Híbrida** y Fusión de Rankings (Reciprocal Rank Fusion) exige la construcción de un conjunto de datos de referencia fiable y representativo. Este conjunto, conocido en la literatura especializada como "Gold Standard" o "Ground

Truth", actúa como el estándar de verdad objetiva indispensable para contrastar cuantitativamente las predicciones generadas por cada sistema algorítmico.

Antes de proceder a la selección y anotación de los datos, resulta fundamental definir los componentes conceptuales clave que estructuran el proceso evaluativo:

- **Corpus de evaluación (Test Set):** Constituye un subconjunto específico de los datos (en esta investigación, compuesto por 30 documentos) que los modelos bajo escrutinio no han encontrado durante ninguna fase de su entrenamiento o ajuste. El propósito primordial de este conjunto es simular de manera fidedigna el comportamiento esperado del sistema de recomendación frente a casos nuevos, permitiendo así medir su rendimiento de forma objetiva y generalizable.
- **"Gold Standard" o "Ground Truth":** Se refiere al conjunto de datos de evaluación (Test Set) una vez que ha sido meticulosamente etiquetado con las respuestas consideradas "correctas" o "ideales" desde una perspectiva experta. En el contexto de esta tesis, para cada documento de consulta (query), se establece una correspondencia con una lista de documentos candidatos (candidates) recomendados por un modelo. Cada uno de estos candidatos es anotado manualmente con una etiqueta de relevancia (pertinente o no pertinente) asignada por un experto humano. Este conjunto anotado representa el estándar de oro contra el cual se comparan las puntuaciones y rankings producidos por los modelos evaluados.
- **Métricas de evaluación:** Corresponden a las formulaciones matemáticas y estadísticas que se emplearán para cuantificar la disparidad o similitud entre los resultados (predicciones, rankings) de los modelos y las etiquetas de verdad establecidas en el "Gold Standard". Ejemplos de métricas relevantes incluyen Precisión@K, Recall@K, F1-Score@K, NDCG@K (Normalized Discounted Cumulative Gain), y MAP (Mean Average Precision). La validez y fiabilidad interpretativa de estas métricas dependen intrínsecamente de la calidad, consistencia y representatividad del "Gold Standard" construido.

4.4.1 Métricas de evaluación de rendimiento

Para cuantificar y comparar objetivamente el rendimiento de los tres modelos de recomendación implementados (Búsqueda Léxica con Okapi BM25, Búsqueda Semántica Basada en Transformers, y Búsqueda Híbrida y Fusión de Rankings), se seleccionó un conjunto de métricas estándar ampliamente utilizadas en el campo de la Recuperación de Información (IR) y Sistemas de Recomendación. Estas métricas permiten evaluar diferentes facetas de la calidad de las listas de documentos recomendados, utilizando como referencia el "Gold Standard" de relevancia manual descrito en la sección anterior.

La evaluación se centró en los primeros 5 resultados (K=5) proporcionados por cada modelo para cada una de las 30 consultas del conjunto de prueba. Esta elección de K=5 se justifica por la presunción de que los usuarios, en un escenario real de búsqueda o recomendación de jurisprudencia, tienden a concentrar su atención en los primeros resultados presentados. Las métricas seleccionadas y aplicadas en esta investigación son las siguientes:

Precisión en K (Precision@K)

La Precisión en K (P@K) mide la proporción de documentos recomendados dentro de los primeros K resultados que son efectivamente relevantes. Es una métrica centrada en el usuario, ya que refleja la calidad de los resultados que el usuario ve inmediatamente. En este estudio, se calculó Precision@5 (P@5).

La fórmula para P@5 para una única consulta es:

$$P@5 = \frac{\text{Número de documentos relevantes en el top 5}}{\text{Número total de documentos en el top 5}(= 5)}$$

Un valor alto de P@5 indica que una gran proporción de las recomendaciones iniciales son útiles para el caso consultado.

Recall en K (Recall@K)

El Recall en K (R@K), también conocido como Exhaustividad o Sensibilidad, mide la proporción de todos los documentos relevantes disponibles para esa consulta en el conjunto anotado que fueron recuperados dentro de los primeros K resultados. Evalúa la capacidad del modelo para encontrar la mayor cantidad posible de información pertinente. Se calculó Recall@5 (R@5). La fórmula para R@5 para una única consulta es:

$$R@5 = \frac{\text{Número de documentos relevantes en el top 5}}{\text{Número total de documentos relevantes para esa consulta (en el Gold Standa}}$$

Un valor alto de R@5 sugiere que el modelo es eficaz encontrando una buena parte de los documentos pertinentes identificados por el experto. Es importante notar que el denominador es el número total de documentos relevantes identificados en el proceso de anotación para esa consulta específica, no el total de documentos relevantes en todo el corpus (que es desconocido).

Puntuación F1 en K (F1-Score@K)

La Puntuación F1 en K (F1@K) es la media armónica de Precision@K y Recall@K. Proporciona una medida única que balancea ambas métricas, siendo útil cuando se desea considerar tanto la precisión como la exhaustividad de las recomendaciones. Se calculó F1-Score@5 (F1@5). La fórmula para F1@5 es:

$$F1@5 = 2 \cdot \frac{P@5 \cdot R@5}{P@5 + R@5}$$

Un F1@5 alto indica un buen equilibrio entre la precisión (no recomendar irrelevantes) y el recall (encontrar los relevantes). Es particularmente útil si existe un desbalance en el número de documentos relevantes.

Precisión Media Promedio (Mean Average Precision - MAP)

La Precisión Media Promedio (MAP) es una métrica global que evalúa la calidad del ranking a lo largo de toda la lista de resultados (o hasta un K profundo), promediada sobre todas las consultas. Se basa en la Precisión Promedio (Average Precision - AP) calculada para cada consulta individual. AP considera la precisión en cada posición donde se encuentra un documento relevante, recompensando a los modelos que posicionan los documentos relevantes más arriba en el ranking. La fórmula de AP para una consulta es:

$$AP = \frac{\sum_{k=1}^{N} (P@k \times \text{rel}(k))}{\text{Número total de documentos relevantes para esa consulta (en el Gold Standar}}$$

donde N es el número total de documentos recomendados (en nuestro caso, 5), P@k es la precisión calculada considerando solo los resultados hasta la posición k, y rel(k) es una función binaria que vale 1 si el documento en la posición k es relevante y 0 si no lo es.

El MAP se calcula como la media aritmética de los valores de AP obtenidos para cada una de las 30 consultas:

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

donde Q es el número total de consultas (30). MAP es una métrica robusta que proporciona una visión general de la calidad del ranking del modelo a través de todo el conjunto de evaluación.

Ganancia acumulada descontada normalizada en K (Normalized Discounted Cumulative Gain - NDCG@K)

La Ganancia Acumulada Descontada Normalizada en K (NDCG@K) es una métrica sofisticada que evalúa la calidad del ranking basándose en la utilidad (o

ganancia) de los documentos recomendados y su posición en la lista. Asigna mayor importancia a los documentos relevantes que aparecen en las primeras posiciones, aplicando un descuento logarítmico a la relevancia de los documentos a medida que descienden en el ranking. Se calculó NDCG@5.

Se calcula en tres pasos para cada consulta:

- Ganancia Acumulada (Cumulative Gain - CG@5): Suma de las relevancias (0 o 1 en nuestro caso) de los primeros 5 resultados.
- Ganancia Acumulada Descontada (Discounted Cumulative Gain - DCG@5): Suma de las relevancias divididas por un factor de descuento basado en el logaritmo de la posición.

$$DCG@5 = \sum_{i=1}^{5} \frac{rel_i}{\log_2(i+1)}$$

  donde rel_i es la relevancia (1 si es relevante, 0 si no) del documento en la posición i.
- Ganancia Acumulada Descontada Normalizada (NDCG@5): Se normaliza el DCG dividiéndolo por el DCG Ideal (IDCG@5), que representa el DCG máximo posible si los documentos relevantes estuvieran perfectamente ordenados en las primeras posiciones.

$$NDCG@5 = \frac{DCG@5}{IDCG@5}$$

El IDCG@5 se calcula ordenando idealmente los documentos relevantes (hasta un máximo de 5) y aplicando la fórmula de DCG. El valor de NDCG@5 varía entre 0 y 1, donde 1 representa un ranking perfecto. Esta métrica es especialmente valiosa porque penaliza fuertemente a los modelos que colocan documentos relevantes en posiciones bajas.

4.4.2 Selección de consultas (Queries) para evaluación

La muestra del corpus se seleccionó bajo un criterio riguroso de muestreo estratificado no probabilístico para garantizar la diversidad temática y la representatividad de la casuística judicial. Este enfoque es crucial para la evaluación objetiva de la eficiencia de los modelos, ya que permite probar el sistema en distintos niveles de dificultad y volumen de datos.

El tamaño final de la muestra para la etapa de validación se fijó en 30 resoluciones judiciales. Esta cantidad cumple con la referencia metodológica mínima de 385 documentos sugerida por Sampieri et al. para garantizar la estabilidad estadística, a la vez que permite un proceso de anotación manual (etiquetado) riguroso por parte de un experto.

La selección fue estructurada en dos estratos principales, combinando la rareza de los temas con la densidad de los datos, tal como se detalla a continuación:

Estrategia de selección de la muestra (30 Casos)

La selección se realizó directamente desde el corpus total (6,848 documentos) y se dividió en los siguientes estratos:

Estrato 1: Amplitud y diversidad (La "Larga Cola")

Este estrato se enfoca en la capacidad de generalización del sistema en el dominio completo.

El objetivo es evaluar el rendimiento del modelo en temas poco frecuentes, donde la identificación de precedentes es inherentemente más difícil. Para este propósito se seleccionó un caso único de cada una de las 20 materias menos frecuentes del corpus. Como resultado se obtuvieron 20 casos de consulta que prueban la robustez del sistema en el conocimiento de nicho y en temas con baja representación en los datos de entrenamiento.

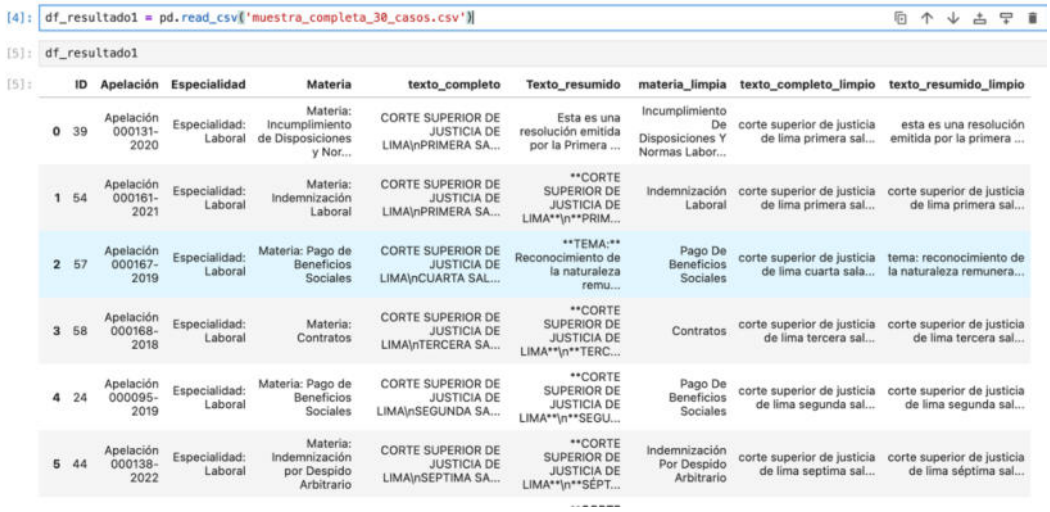
Estrato 2: Densidad y representatividad (La "Cabeza")

Este estrato asegura que el sistema sea probado en el escenario de uso más común, reflejando la distribución real de los expedientes. El objetico es validar la precisión del modelo en las categorías con mayor volumen de precedentes, que constituyen más del 60% del corpus total. Por tanto, se seleccionaron los 10 casos restantes de forma aleatoria de entre las 3 materias más comunes ("Pago de Beneficios Sociales", "Contratos", e "Incumplimiento de Disposiciones"). Se obtuvieron 10 casos de consulta para validar el rendimiento en entornos densos, donde la mayoría de las consultas reales se concentran.

La muestra final de 30 Casos resultante es el conjunto que se utiliza en la etapa de Evaluación y Validación para medir métricas de desempeño como Precision, Recall, MAP y NDCG.

Esta separación y reserva exclusiva del conjunto de evaluación antes de cualquier entrenamiento o ajuste es una práctica estándar que asegura una estimación insesgada de la capacidad de generalización de los modelos. Cada documento designado como consulta fue identificado mediante un identificador único (query_id). El query_resumen correspondiente a cada query_id sirvió como el punto de partida (input) para que los modelos generaran sus listas de recomendaciones y, simultáneamente, como el texto de referencia principal para el posterior juicio de relevancia manual. Como se puede apreciar en el siguiente imagen.

Figura 8 Muestra de datos de la prueba (Test Set)



```
[4]: df_resultado1 = pd.read_csv('muestra_completa_30_casos.csv')
[5]: df_resultado1
```

[5]:

	ID	Apelación	Especialidad	Materia	texto_completo	Texto_resumido	materia_limpia	texto_completo_limpio	texto_resumido_limpio
0	39	Apelación 000131-2020	Especialidad: Laboral	Materia: Incumplimiento de Disposiciones y Nor...	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	Esta es una resolución emitida por la Primera ...	Incumplimiento De Disposiciones Y Normas Labor...	corte superior de justicia de lima primera sal...	esta es una resolución emitida por la primera ...
1	54	Apelación 000161-2021	Especialidad: Laboral	Materia: Indemnización Laboral	CORTE SUPERIOR DE JUSTICIA DE LIMA\nPRIMERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...	Indemnización Laboral	corte superior de justicia de lima primera sal...	corte superior de justicia de lima primera sal...
2	57	Apelación 000167-2019	Especialidad: Laboral	Materia: Pago de Beneficios Sociales	CORTE SUPERIOR DE JUSTICIA DE LIMA\nCUARTA SAL...	**TEMA:** Reconocimiento de la naturaleza remu...	Pago De Beneficios Sociales	corte superior de justicia de lima cuarta sala...	tema: reconocimiento de la naturaleza remunera...
3	58	Apelación 000168-2018	Especialidad: Laboral	Materia: Contratos	CORTE SUPERIOR DE JUSTICIA DE LIMA\nTERCERA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**TERC...	Contratos	corte superior de justicia de lima tercera sal...	corte superior de justicia de lima tercera sal...
4	24	Apelación 000095-2019	Especialidad: Laboral	Materia: Pago de Beneficios Sociales	CORTE SUPERIOR DE JUSTICIA DE LIMA\nSEGUNDA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**SEGU...	Pago De Beneficios Sociales	corte superior de justicia de lima segunda sal...	corte superior de justicia de lima segunda sal...
5	44	Apelación 000138-2022	Especialidad: Laboral	Materia: Indemnización por Despido Arbitrario	CORTE SUPERIOR DE JUSTICIA DE LIMA\nSEPTIMA SA...	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**SÉPT...	Indemnización Por Despido Arbitrario	corte superior de justicia de lima septima sal...	corte superior de justicia de lima séptima sal...

Figura 9 Características de los datos de prueba

```
df_completo.to_csv('muestra_completa_30_casos.csv', index=False)

print(f"Archivo 'muestra_completa_30_casos.csv' creado con {len(df_completo)} filas.")
```

```
Archivo 'muestra_completa_30_casos.csv' creado con 30 filas.
```

```
[7]: df_completo.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   ID                    30 non-null     int64 
 1   Apelación             30 non-null     object
 2   Especialidad          30 non-null     object
 3   Materia               30 non-null     object
 4   texto_completo        30 non-null     object
 5   Texto_resumido        30 non-null     object
 6   materia_limpia        30 non-null     object
 7   texto_completo_limpio 30 non-null     object
 8   texto_resumido_limpio 30 non-null     object
dtypes: int64(1), object(8)
memory usage: 2.2+ KB
```

4.4.3 Generación de documentos candidatos recomendados

Para cada uno de los 30 documentos de consulta (query) definidos en el conjunto de evaluación, se ejecutó cada uno de los tres modelos bajo estudio: Búsqueda Léxica con Okapi BM25, Búsqueda Semántica Basada en Transformers (SentenceTransformer), y Búsqueda Híbrida y Fusión de Rankings (Reciprocal Rank Fusion). A cada modelo se le solicitó que procesara la consulta (basada en el query_resumen) y generara una lista ordenada de documentos recomendados, seleccionados del corpus de entrenamiento (compuesto por los 6848 documentos

restantes). De las listas completas de recomendaciones generadas por cada modelo para cada consulta, se procedió a seleccionar únicamente los 5 documentos candidatos (candidates) que obtuvieron la puntuación de similitud o relevancia más alta (model_score) asignada por el modelo respectivo.

Este procedimiento estandarizado de seleccionar el top-5 de recomendaciones resultó en la creación de tres conjuntos de datos de evaluación distintos, uno específico para cada modelo (Léxico, Semántico, Híbrido). Cada uno de estos conjuntos contiene exactamente 150 pares (query, candidate) listos para ser anotados manualmente (calculado como 30 consultas × 5 candidatos por consulta). Es importante notar que, aunque existe la posibilidad de que diferentes modelos recomienden los mismos documentos candidatos para una misma consulta (solapamiento), cada conjunto de 150 pares fue tratado y anotado de forma independiente, permitiendo así una evaluación aislada y justa del rendimiento particular de cada modelo. La elección del valor K=5 se alinea con prácticas comunes en la evaluación de sistemas de recomendación y recuperación de información, ya que permite valorar eficazmente la habilidad del sistema para posicionar los resultados más pertinentes en las primeras posiciones del ranking, que son, en la práctica, las que reciben la mayor atención por parte del usuario final.

Figura 10 Recomendaciones generadas por los 3 modelos

```
[13]: df_resultado1 = pd.read_csv('plantilla_evaluacion_semantico.csv')
[14]: df_resultado1
```

[14]:

	query_id	query_apelacion	query_resumen	candidate_id	candidate_apelacion	candidate_resumen	model_score	relevancia_manual
0	39	Apelación 000131-2020	Esta es una resolución emitida por la Primera ...	5087	Apelación 020221-2019	**ASUNTO:** Reintegro de beneficios económicos...	89.03%	NaN
1	39	Apelación 000131-2020	Esta es una resolución emitida por la Primera ...	5164	Apelación 020692-2019	**EXPEDIENTE N°:** 20692-2019-0-1801-JR-LA-06\...	88.63%	NaN
2	39	Apelación 000131-2020	Esta es una resolución emitida por la Primera ...	5081	Apelación 020212-2019	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**SEGU...	88.55%	NaN
3	39	Apelación 000131-2020	Esta es una resolución emitida por la Primera ...	6394	Apelación 026995-2018	Este es un resumen conciso de la Sentencia de ...	88.50%	NaN
4	39	Apelación 000131-2020	Esta es una resolución emitida por la Primera ...	4800	Apelación 018601-2017	**RESUMEN DE SENTENCIA DE VISTA**\n**Expedient...	88.46%	NaN
...	...	...	...	...	...	...	...	...
145	55	Apelación 000163-2020	**ASUNTO:** Recurso de Apelación interpuesto p...	2511	Apelación 008544-2020	**ASUNTO:** Ineficacia de Contrato Administrat...	81.82%	NaN
146	55	Apelación 000163-2020	**ASUNTO:** Recurso de Apelación interpuesto p...	190	Apelación 000570-2021	**Julián Ismael Gutiérrez Prado c/ Dalkomhan T...	80.64%	NaN
147	55	Apelación 000163-2020	**ASUNTO:** Recurso de Apelación interpuesto p...	3261	Apelación 011523-2018	**CORTE SUPERIOR DE JUSTICIA DE LIMA**\n**PRIM...	79.72%	NaN
148	55	Apelación 000163-2020	**ASUNTO:** Recurso de Apelación interpuesto p...	1067	Apelación 003729-2019	Este resumen se basa en la Sentencia de Vista,...	78.22%	NaN
149	55	Apelación 000163-2020	**ASUNTO:** Recurso de Apelación interpuesto p...	4937	Apelación 019406-2018	**EXPEDIENTE N.° 19406-2018-0-1801-JR-LA-06**\...	77.85%	NaN

Figura 11 Características de recomendaciones

```
[15]: df_resultado1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   query_id             150 non-null    int64
 1   query_apelacion      150 non-null    object
 2   query_resumen        150 non-null    object
 3   candidate_id         150 non-null    int64
 4   candidate_apelacion  150 non-null    object
 5   candidate_resumen    150 non-null    object
 6   model_score          150 non-null    object
 7   relevancia_manual    0 non-null      float64
dtypes: float64(1), int64(2), object(5)
memory usage: 9.5+ KB
```

4.4.4 Proceso de anotación manual de relevancia (Creación del Gold Standard)

La fase culminante y más crítica en la construcción del "Gold Standard" consistió en la anotación manual exhaustiva de la relevancia para cada uno de los pares (query, candidate) generados por los modelos en la etapa anterior. Este proceso de juicio experto fue llevado a cabo íntegramente por el investigador principal. La decisión de centralizar la anotación se basó en la necesidad de aplicar criterios consistentes y uniformes a lo largo de todo el conjunto de datos, aprovechando el conocimiento especializado del anotador en el dominio específico del derecho laboral peruano y los objetivos particulares de la investigación.

Para cada par (query, candidate) presentado (potencialmente hasta 150 pares por cada modelo), el anotador realizó una evaluación contextual comparativa detallada entre el contenido del query_resumen (representando el caso de entrada o la necesidad de información) y el contenido del candidate_resumen (representando la recomendación proporcionada por el sistema). La metodología de anotación adoptada se basó en una evaluación heurística estructurada, diseñada para emular el proceso cognitivo que seguiría un experto legal al determinar la pertinencia de una jurisprudencia recomendada para un caso dado. Esta evaluación consideró múltiples facetas de la similitud, ponderadas según su importancia estimada en el dominio legal:

a) Similitud semántica (Ponderación: 40%): Se evaluó la correspondencia conceptual profunda entre los resúmenes. El análisis se centró en verificar si ambos documentos abordaban los mismos hechos sustanciales, aplicaban los mismos fundamentos de derecho clave o llegaban a decisiones sobre pretensiones análogas, priorizando la coincidencia del sentido jurídico global por sobre la mera coincidencia léxica superficial.

b) Coincidencia textual o léxica (Ponderación: 25%): Se cuantificó la presencia y frecuencia de vocabulario jurídico técnico común y relevante en ambos textos (e.g., términos como apelación, sentencia, nulidad, despido, beneficios sociales, indemnización, prescripción), considerando su pertinencia contextual dentro de la materia laboral.

c) Coherencia redaccional y argumentativa (Ponderación: 15%): Se valoró la similitud en la estructura narrativa y la organización argumentativa de los resúmenes. Se observó si ambos seguían una secuencia lógica comparable en la exposición de los antecedentes, los puntos controvertidos, el razonamiento judicial y la parte resolutiva.

d) Coincidencia jurídica o materia legal (Ponderación: 10%): Se verificó explícitamente que ambos documentos pertenecieran a la misma rama específica del derecho (en este caso, derecho laboral), asignando la máxima puntuación en este criterio si la coincidencia era clara.

e) Concordancia entre actos procesales (Ponderación: 10%): Se analizaron los campos query_apelacion y candidate_apelacion para determinar si existía una relación lógica o de similitud entre los tipos de recurso, actos procesales o instancias judiciales involucradas en ambos casos.

A cada par (query, candidate) se le asignó una puntuación global resultante de la suma ponderada de las valoraciones obtenidas en los cinco criterios anteriores. Con base en esta puntuación global, se aplicó un umbral predefinido para asignar la etiqueta binaria final de relevancia_manual:

1. **(Relevante = 1):** Se asignó esta etiqueta si la puntuación de similitud total ponderada alcanzaba o superaba el umbral de 0.6. Este valor se estableció considerando que representa una indicación de una relación temática, jurídica o contextual fuerte y útil entre la consulta y el candidato.

2. **(No Relevante = 0):** Se asignó esta etiqueta si la puntuación ponderada resultaba inferior a 0.6, interpretándose como una indicación de escasa relación sustancial, tratamiento de materias distintas o una conexión meramente tangencial o superficial.

Este método estructurado y transparente de anotación buscó minimizar la subjetividad y asegurar la consistencia, generando así una base de verdad terreno ("Ground Truth") fiable y de alta calidad para la subsiguiente evaluación cuantitativa del rendimiento de los modelos de recomendación. Los resultados se muestran en la Figura 12.

Figura 12 Resultados de anotación manual de relevancia

[3]:

	query_id	query_apelacion	candidate_id	candidate_apelacion	model_score	relevancia_manual	modelo
0	39	Apelación 000131-2020	5087	Apelación 020221-2019	89.03%	1	Semántico
1	39	Apelación 000131-2020	5164	Apelación 020692-2019	88.63%	0	Semántico
2	39	Apelación 000131-2020	5081	Apelación 020212-2019	88.55%	0	Semántico
3	39	Apelación 000131-2020	6394	Apelación 026995-2018	88.50%	1	Semántico
4	39	Apelación 000131-2020	4800	Apelación 018601-2017	88.46%	0	Semántico
...	...	...	...	...	...	...	...
445	55	Apelación 000163-2020	1067	Apelación 003729-2019	0.0281	0	Híbrido (RRF)
446	55	Apelación 000163-2020	3046	Apelación 010679-2019	0.0238	1	Híbrido (RRF)
447	55	Apelación 000163-2020	6297	Apelación 026573-2018	0.0217	0	Híbrido (RRF)
448	55	Apelación 000163-2020	2511	Apelación 008544-2020	0.0164	1	Híbrido (RRF)
449	55	Apelación 000163-2020	190	Apelación 000570-2021	0.0161	0	Híbrido (RRF)

4.4.5 Estructura y uso del conjunto de datos de evaluación final

El producto final de la metodología descrita consiste en tres conjuntos de datos distintos, cada uno asociado a uno de los modelos evaluados (Búsqueda Léxica con Okapi BM25, Búsqueda Semántica Basada en Transformers, y Búsqueda Híbrida y Fusión de Rankings). Estos conjuntos fueron consolidados y almacenados en formato CSV (Comma-Separated Values) para facilitar su procesamiento computacional. Cada archivo CSV mantiene la estructura fundamental de pares (query, candidate) y contiene, para cada par, un conjunto de columnas esenciales que registran la información relevante del proceso:

- query_id: Identificador único del documento de consulta.
- query_apelacion: Texto o descripción del acto procesal del documento de consulta.

- candidate_id: Identificador único del documento candidato recomendado.
- candidate_apelacion: Texto o descripción del acto procesal del documento candidato.
- model_score: La puntuación numérica original de similitud o relevancia asignada por el modelo específico a ese candidate para esa query.
- relevancia_manual: La etiqueta binaria (0 o 1) resultante del proceso de anotación manual descrito en la sección anterior.

La columna relevancia_manual es el componente crítico que materializa el "Ground Truth". Es esta columna la que sirve como referencia indiscutible para llevar a cabo la evaluación cuantitativa objetiva del rendimiento de cada uno de los modelos de recomendación. Los conjuntos de datos anotados constituyen la entrada directa y necesaria para el cálculo de un abanico de métricas estándar en los campos de Recuperación de Información (IR) y Sistemas de Recomendación. Las métricas calculadas y analizadas en esta tesis incluyen, entre otras: Precisión@K, Recall@K, F1-Score@K, NDCG@K (Normalized Discounted Cumulative Gain) y MAP (Mean Average Precision).

4.5 Resultados

Esta sección presenta los resultados cuantitativos obtenidos al evaluar el rendimiento de los tres modelos de recomendación implementados Búsqueda Léxico con Okapi BM25, Búsqueda Semántica Basada en Transformers (SentenceTransformer), y Búsqueda Híbrida y Fusión de Rankings (Reciprocal Rank Fusion) sobre el conjunto de prueba de 30 consultas, utilizando el "Gold Standard" de relevancia manual descrito previamente. El rendimiento se midió utilizando las métricas estándar de Precisión@5 (P@5), Recall@5 (R@5), Precisión Media Promedio (MAP@5) y Ganancia Acumulada Descontada Normalizada (NDCG@5).

La Tabla 03 resume los valores promedio obtenidos para cada métrica y cada modelo, calculados sobre el total de las 30 consultas.

Tabla 3 Resultados de precisión de los modelos

Modelo	Precision@5	Recall@5	MAP@5	NDCG@5
Semántico	0.3600	0.1863	0.1517	0.4001
Léxico (BM25)	**0.8400**	**0.5060**	**0.4877**	**0.8604**
Híbrido (RRF)	0.6133	0.3554	0.2584	0.5943

Del análisis de los resultados presentados en la Tabla 03, se desprenden las siguientes observaciones sobre el desempeño de cada modelo:

4.5.1 Modelo léxico (BM25)

Este modelo demostró ser el más eficaz en la tarea de recomendación según todas las métricas evaluadas. Obtuvo una Precisión@5 de 0.8400, lo que indica que, en promedio, el 84% de los 5 primeros documentos recomendados fueron considerados relevantes por el anotador humano. Su Recall@5 de 0.5060 sugiere que logró recuperar aproximadamente la mitad de los documentos relevantes identificados para cada consulta dentro de sus 5 primeras recomendaciones. Las métricas de ranking, MAP@5 (0.4877) y NDCG@5 (0.8604), también fueron significativamente más altas, reflejando una superior capacidad para ordenar los documentos relevantes en las posiciones más altas de la lista de resultados.

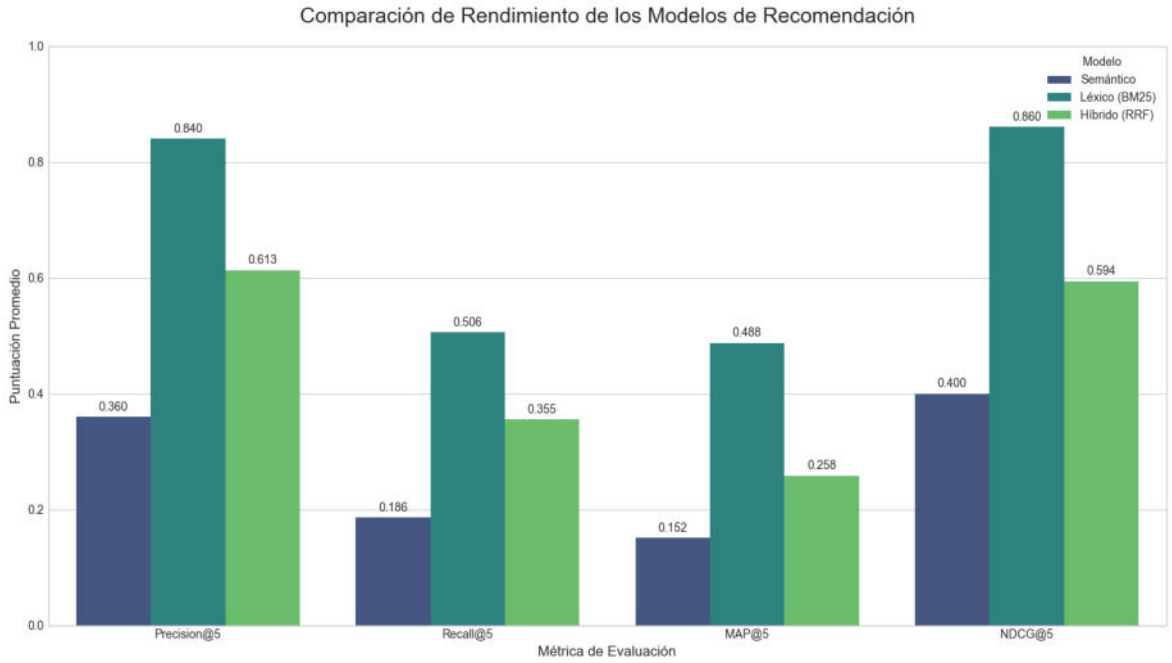
4.5.2 Modelo semántico (SentenceTransformer)

El modelo basado exclusivamente en embeddings semánticos obtuvo el rendimiento más bajo en comparación con los otros enfoques. Su Precisión@5 fue de 0.3600 y su Recall@5 de 0.1863, indicando que una menor proporción de sus recomendaciones iniciales fueron relevantes y que recuperó una fracción menor del total de documentos pertinentes. Las métricas de ranking MAP@5 (0.1517) y NDCG@5 (0.4001) confirman esta tendencia, sugiriendo dificultades para priorizar adecuadamente los resultados más relevantes.

4.5.3 Modelo híbrido (RRF)

El enfoque híbrido, que combina los rankings de los modelos léxico y semántico mediante Reciprocal Rank Fusion, mostró un rendimiento intermedio. Con una Precisión@5 de 0.6133 y un Recall@5 de 0.3554, superó claramente al modelo semántico puro, pero no alcanzó la efectividad del modelo léxico BM25. Sus puntuaciones de MAP@5 (0.2584) y NDCG@5 (0.5943) se situaron también entre los otros dos modelos, indicando una calidad de ranking moderada.

Figura 13 Comparación de los resultados de precisión de los modelos



Los resultados cuantitativos demuestran una clara ventaja del modelo léxico BM25 sobre los enfoques semántico e híbrido en este conjunto de datos y tarea específica. La diferencia de rendimiento es notable en todas las métricas, especialmente en Precision@5 y NDCG@5, lo que sugiere que para este corpus de resúmenes de jurisprudencia laboral, las señales basadas en la frecuencia y especificidad de los términos (capturadas por BM25) fueron más efectivas para identificar y priorizar documentos relevantes que las representaciones semánticas densas (capturadas por SentenceTransformer). El modelo híbrido, si bien mejoró respecto al semántico, no logró capitalizar completamente las fortalezas del BM25, posiblemente debido a la influencia del ranking semántico de menor calidad en el proceso de fusión RRF.

Estos hallazgos cuantitativos proporcionan una base sólida para la discusión sobre la efectividad relativa de los diferentes enfoques de recomendación en el dominio legal estudiado, la cual se desarrollará en secciones posteriores.

4.6 Prototipo

Con el fin de proporcionar una demostración funcional e interactiva de los modelos de recomendación desarrollados y facilitar su evaluación cualitativa, se implementó un prototipo de aplicación web ejecutado en un entorno local. Esta sección describe la arquitectura y las tecnologías empleadas para la construcción de dicho prototipo.

4.6.1 Arquitectura Tecnológica del Prototipo

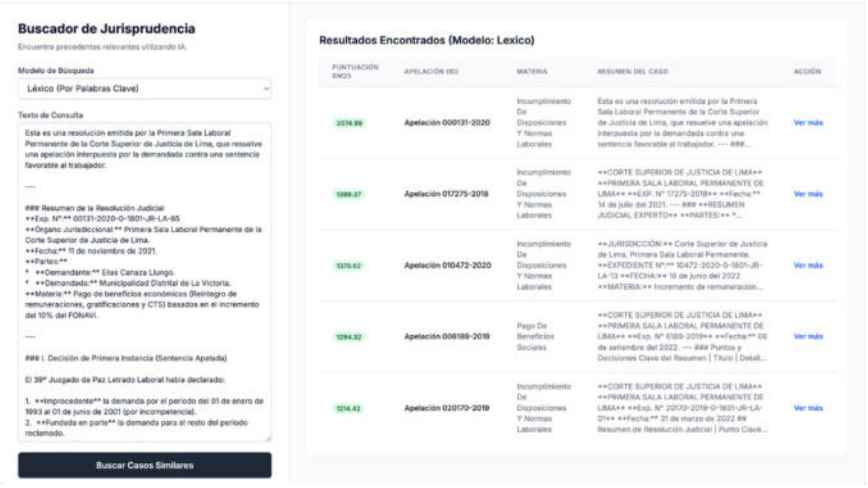
El prototipo se desarrolló como una aplicación web utilizando el microframework Flask en Python. Flask fue seleccionado por su simplicidad, flexibilidad y facilidad para crear rápidamente interfaces web funcionales que interactúen con lógica de backend escrita en Python.

La aplicación integra los tres modelos de recomendación evaluados en esta tesis:

- Búsqueda Léxica con Okapi BM25
- Búsqueda Semántica Basada en Transformers (SentenceTransformer)
- Búsqueda Híbrida y Fusión de Rankings (Reciprocal Rank Fusion)

Los componentes precalculados necesarios para la ejecución de los modelos (como los índices TF-IDF/BM25, el modelo SentenceTransformer pre-entrenado y los embeddings si fueron precalculados) fueron cargados por la aplicación Flask al iniciarse.

Figura 14 Interfaz del prototipo de recomendación de jurisprudencias



Se diseñó una interfaz de usuario web básica, renderizada a través de plantillas HTML servidas por Flask Como se observa en la Figura 15, la interfaz permite al usuario:

- Ingresar el nuevo caso de consulta.
- Activar la Búsqueda: Al enviar el ID, la aplicación Flask ejecuta en el backend los tres modelos de recomendación utilizando el query_resumen asociado a ese ID.
- Visualizar Resultados: La interfaz presenta de manera organizada las 5 recomendaciones principales (top-5) generadas por cada uno de los tres modelos (Léxico, Semántico e Híbrido), mostrando información relevante de cada documento recomendado (como su ID, texto de apelación y, posiblemente, el model_score).

La implementación de este prototipo local cumplió varios propósitos clave dentro de la investigación. Validación Funcional, Permitió verificar que los modelos implementados podían ser integrados correctamente en una aplicación y generar recomendaciones en respuesta a una entrada de usuario. Demostración Interactiva, Facilitó la demostración práctica del sistema a terceros (como asesores o jurados de tesis), permitiendo una comprensión más clara de su funcionamiento y resultados. Evaluación Cualitativa, Aunque la evaluación principal fue cuantitativa (basada en métricas), la interfaz permitió una inspección cualitativa rápida de las recomendaciones generadas por los diferentes modelos para consultas específicas.

V. Conclusiones

La presente investigación se propuso desarrollar y evaluar comparativamente diferentes enfoques para la recomendación de jurisprudencia relevante en el dominio del derecho laboral peruano, utilizando resúmenes de resoluciones de apelación de la Corte Superior de Justicia de Lima. Se implementaron y compararon tres modelos: un enfoque léxico tradicional (Búsqueda Léxica con Okapi BM25), un enfoque semántico basado en Transformers (Búsqueda Semántica con SentenceTransformer) y un enfoque Híbrido combinando ambos mediante Reciprocal Rank Fusion (RRF). La evaluación se realizó sobre un conjunto de 30 consultas (queries) con 5 recomendaciones (candidates) por consulta, utilizando un "Gold Standard" de relevancia binaria anotado manualmente por el investigador mediante criterios heurísticos estructurados. El rendimiento se midió con métricas estándar @5 (Precision, Recall, MAP, NDCG).

A partir de los resultados cuantitativos y el análisis realizado, se derivan las siguientes conclusiones principales:

5.1 Superioridad del modelo léxico (BM25)

El modelo BM25 demostró consistentemente el mejor rendimiento en todas las métricas evaluadas (P@5=0.8400, R@5=0.5060, MAP@5=0.4877, NDCG@5=0.8604). Este hallazgo subraya la alta efectividad de las señales léxicas (frecuencia de términos, especificidad IDF) para identificar y clasificar jurisprudencia relevante dentro de este corpus específico de resúmenes legales en materia laboral. La precisión terminológica parece ser un factor crucial en este dominio.

5.2 Limitaciones del modelo semántico puro (SentenceTransformer)

El modelo semántico basado en SentenceTransformer (bert-base-multilingual-cased) obtuvo el rendimiento más bajo (P@5=0.3600, R@5=0.1863, MAP@5=0.1517, NDCG@5=0.4001). Si bien los embeddings contextuales capturan relaciones semánticas generales, este modelo pre-entrenado multipropósito no fue suficiente por sí solo para discernir con alta precisión las

sutilezas de la relevancia jurídica en los resúmenes. Esto sugiere que, sin un ajuste fino específico para el dominio legal o sin información léxica complementaria, la semántica pura puede tener dificultades para priorizar adecuadamente los casos más pertinentes.

5.3 Rendimiento Intermedio del Modelo Híbrido (RRF)

El modelo híbrido, utilizando RRF para fusionar los rankings de BM25 y SentenceTransformer, logró un rendimiento intermedio (P@5=0.6133, R@5=0.3554, MAP@5=0.2584, NDCG@5=0.5943). Si bien mejoró significativamente respecto al modelo semántico puro, no consiguió superar al modelo léxico BM25. Esto podría indicar que la fusión mediante RRF, aunque beneficiosa, se vio limitada por la menor calidad del ranking semántico de entrada, o que la combinación lineal simple de rankings no fue la estrategia óptima para este escenario. No obstante, su mejora sobre el modelo semántico valida el potencial de combinar señales léxicas y semánticas.

5.4 Importancia de la coincidencia léxica en el dominio legal laboral

Los resultados comparativos sugieren fuertemente que, para la tarea de recomendación basada en resúmenes de jurisprudencia laboral peruana, los métodos que ponderan adecuadamente la presencia de términos clave específicos (como BM25) son altamente efectivos. La naturaleza técnica y precisa del lenguaje jurídico parece favorecer estos enfoques, al menos en el contexto de los resúmenes analizados.

5.5 Viabilidad de prototipo local con Flask

La implementación de un prototipo funcional utilizando Flask demostró la viabilidad de integrar los modelos desarrollados en una aplicación interactiva, facilitando la demostración y la evaluación cualitativa de las recomendaciones en un entorno local controlado, sirviendo como una herramienta útil para la validación del sistema.

VI. Recomendaciones

A partir de las conclusiones obtenidas en la presente investigación sobre la efectividad comparativa de modelos léxicos, semánticos e híbridos para la recomendación de jurisprudencia laboral peruana, se formulan las siguientes recomendaciones orientadas tanto a la aplicación práctica como a futuras líneas de investigación en el área de recuperación de información jurídica:

6.1 Priorizar Modelos Léxicos (BM25) para Aplicaciones

Dado el rendimiento superior y consistente del modelo Okapi BM25 en todas las métricas evaluadas sobre el corpus de resúmenes, se recomienda su consideración prioritaria para el desarrollo de herramientas de búsqueda o recomendación de jurisprudencia en entornos donde la precisión terminológica y la eficiencia computacional sean factores críticos. Su implementación es relativamente sencilla y demostró alta efectividad en este contexto específico.

6.2 Explorar Técnicas Avanzadas para Modelos Híbridos

Aunque el modelo híbrido basado en RRF mejoró el rendimiento del modelo semántico puro, no superó al BM25. Se recomienda investigar estrategias de fusión más sofisticadas que RRF, como métodos ponderados basados en aprendizaje (learning-to-rank) o técnicas que consideren la confianza de cada modelo. Asimismo, antes de la fusión, explorar el ajuste fino (fine-tuning) del modelo SentenceTransformer en un corpus jurídico peruano podría mejorar significativamente la calidad de las representaciones semánticas y, por ende, el potencial del enfoque híbrido.

Ampliar y Diversificar el Conjunto de Evaluación: La evaluación se basó en 30 consultas y 5 candidatos por consulta, anotados por un único experto. Para futuras validaciones y una mayor generalización de los resultados, se recomienda construir un "Gold Standard" más extenso, que incluya un mayor número de consultas, cubra una gama más amplia de subtemas laborales, y considere un mayor número de candidatos (K más grande). Idealmente, este conjunto debería ser anotado por múltiples expertos legales para calcular el acuerdo inter-anotador y asegurar una mayor objetividad y fiabilidad de las etiquetas de relevancia.

6.3 Investigar el Uso de Texto Completo vs. Resúmenes

La presente tesis se centró en el análisis de resúmenes. Si bien son informativos, pueden omitir detalles cruciales. Se recomienda como línea de trabajo futuro investigar la aplicación de estos modelos (especialmente los semánticos y híbridos) sobre el texto completo de las resoluciones. Esto podría permitir capturar matices contextuales y relaciones más complejas que no están presentes en los resúmenes, aunque implicaría mayores desafíos computacionales y de modelado.

6.4 Considerar la Incorporación de Metadatos

Explorar la integración de metadatos disponibles en las resoluciones (como fecha, sala emisora, partes, tipo de fallo específico, leyes citadas) como características adicionales en los modelos de recomendación, especialmente en enfoques híbridos o basados en aprendizaje automático, podría mejorar la contextualización y la precisión de las recomendaciones.

La adopción de estas recomendaciones podría contribuir significativamente al avance de herramientas más efectivas y útiles para la gestión del conocimiento jurídico en el contexto peruano, beneficiando tanto a profesionales del derecho como a académicos e investigadores.

VII.Referencias Bibliográficas

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.

Alrumayyan, N., & Al-Yahya, M. (2022). Neural Embeddings for the Elicitation of Jurisprudence Principles: The Case of Arabic Legal Texts. *Applied Sciences*, *12*(9), 4188. https://doi.org/10.3390/app12094188

Andrade, L. de, & Becker, K. (2023). *BB25HLegalSum: Leveraging BM25 and BERT-Based Clustering for the Summarization of Legal Documents* (pp. 255–263). https://doi.org/10.26615/978-954-452-092-2_029

Byrom, N. , Piccinin-Barbieri, & Wells, P. (2024). *Towards effective governance of justice data*. *74*, 7. https://doi.org/10.1787/D2950E02-EN

Dantas, N. C. D. and S. E. D. S. A. C. J. A. F. and. (2023). Evaluating Text Classification in the Legal Domain Using BERT Embeddings. En D. and Y. H. and G. T. and J. V. and T.-B. A. J. Q. P. and Camacho (Ed.), *Intelligent Data Engineering and Automated Learning – IDEAL 2023* (pp. 51–63). Springer Nature Switzerland.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Fuentes González, A. (2021). *EL PROCESAMIENTO DE TEXTOS JURÍDICOS: NATURAL LANGUAGE PROCESSING & LEGAL TECH*.

González-Moreno, L. (2023). *INTELIGENCIA ARTIFICIAL Y DERECHO: RETOS JURÍDICOS*.

Hugo Mentzingen Nuno António, F. B. M. C. (2023). *Using textual similarity to identify legal precedents: appraising machine learning models for administrative courts — researchsquare.com*.

Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd draft).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martínez, M. F. (2022). *Búsqueda semántica de sentencias judiciales usando técnicas actuales del NLP y métodos de recuperación de información*. https://doi.org/10.11144/JAVERIANA.10554.62083<!---->

Mentzingen, H., António, N., Bacao, F., & Cunha, M. (2024). Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court. *International Journal of Information Management Data Insights*, *4*(2), 100247. https://doi.org/10.1016/j.jjimei.2024.100247

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.

n Getman Oleg M. Yaroshenko, R. V. S. R. Ye. P. M. I. D. (2021). The Impact of Artificial Intelligence on Legal Decision Making . *International Comparative Jurisprudence*, *15*(2), 123–145.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Resck, L. E., Ponciano, J. R., Nonato, L. G., & Poco, J. (2023). LegalVis: Exploring and Inferring Precedent Citations in Legal Documents. *IEEE Transactions on Visualization and Computer Graphics*, *29*(6), 3105–3120. https://doi.org/10.1109/TVCG.2022.3152450

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender Systems Handbook*. Springer.

Ruiz, R. A. (2021). *Jurisprudence search based on facts similarity using NLP and ML techniques.* Universidade de São Paulo.

Sampieri, R. H., Collado, C. F., & Lucio, P. B. (2014). *Metodología de la investigación* (6a ed.). McGraw-Hill Education.

Shao, H.-L. (2024). *Judgment2vec: Apply Graph Analytics to Searching and Recommendation of Similar Judgments*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*. https://arxiv.org/abs/1706.03762

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos	: Anthony Roger Flor Velásquez
DNI/Carnet de Extranjería/Pasaporte N°	: 47071758
Correo electrónico	: antony9015@gmail.com
Teléfono/Celular	: 935123932

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis **(X)**
b) Trabajo de Investigación **( )**
c) Trabajo de suficiencia profesional **( )**
d) Trabajo académico **( )**

5. DATOS DE LA PUBLICACION

a) Título : Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú
b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra : https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html) : **1.02.01**
c) Fecha de publicación : **2026**

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : Hugo Euler Tito Chura
b) DNI/Pasaporte/Cédula : 01315850
c) ORCID : 0000-0001-8269-0655

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 47071758
NOMBRES Y APELLIDOS : Anthony Roger Flor Velásquez

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos	: María del Pilar Laura Ramírez Chipana
DNI/Carnet de Extranjería/Pasaporte N°	: 72659077
Correo electrónico	: mariaramireschipana@gmail.com
Teléfono/Celular	: 995252993

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis **(X)**
b) Trabajo de Investigación **( )**
c) Trabajo de suficiencia profesional **( )**
d) Trabajo académico **( )**

5. DATOS DE LA PUBLICACION

a) Título : SISTEMA DE RECOMENDACIÓN DE JURISPRUDENCIA BASADO EN EMBEDDINGS DE TEXTO PARA LA IDENTIFICACIÓN Y SUGERENCIA DE PRECEDENTES RELEVANTES DE LAS CORTES SUPERIORES DE JUSTICIA DE LA REPÚBLICA DEL PERÚ
b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra : https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html) : 1.02.01
c) Fecha de publicación : 2026

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : Hugo Euler Tito Chura
b) DNI/Pasaporte/Cédula : 01315850
c) ORCID : 0000-0001-8269-0655

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 72659077
NOMBRES Y APELLIDOS : María Del Pilar Laura Ramírez Chipana

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAM.

Con la autorización de depósito da mi trabajo de investigación, otorgo a la Universidad Nacional de Moquegua, una Licencia No Exclusiva para reproducir. distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAM, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias, y libre de remuneraciones.

En virtud de dicha Licencia, la UNIVERSIDAD NACIONAL DE MOQUEGUA, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La UNIVERSIDAD NACIONAL DE MOQUEGUA consignara el nombre del/los autor/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

(X) **SI autorizo que se deposite inmediatamente.**

FIRMA : .....



DOCUMENTO DE IDENTIDAD : 47071758
NOMBRES Y APELLIDOS : Anthony Roger Flor Velásquez
FECHA : **enero del 2026**

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAM.

Con la autorización de depósito da mi trabajo de investigación, otorgo a la Universidad Nacional de Moquegua, una Licencia No Exclusiva para reproducir. distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAM, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias, y libre de remuneraciones.

En virtud de dicha Licencia, la UNIVERSIDAD NACIONAL DE MOQUEGUA, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La UNIVERSIDAD NACIONAL DE MOQUEGUA consignara el nombre del/los autor/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

(X) **SI autorizo que se deposite inmediatamente.**

FIRMA : .....



DOCUMENTO DE IDENTIDAD : 72659077
NOMBRES Y APELLIDOS : María Del Pilar Laura Ramírez Chipana
FECHA : 15 de enero del 2026

ANEXO VIII

DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN / TESIS / TRABAJO DE SUFICIENCIA PROFESIONAL

Yo, Anthony Roger Flor Velásquez , bachiller de la Escuela Profesional de:
Ingeniería de Sistemas e Informática ,
de la Universidad Nacional de Moquegua, identificado(a) con DNI Nº 47071758 , autor(a) del Trabajo de Investigación / Tesis / Trabajo de suficiencia profesional denominado:
Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú.

Declaro bajo juramento que:

1. El trabajo de investigación es auténtico, siendo resultado de mi trabajo personal, que no se ha copiado, que no se ha utilizado ideas, formulaciones, citas integrales e ilustraciones diversas, sacadas de cualquier tesis, obra, artículo, memoria, etc., (en versión digital o impresa), sin mencionar de forma clara y exacta su origen o autor, tanto en el cuerpo del texto, figuras, cuadros, tablas u otros que tengan derechos de autor. Por tanto, el trabajo de investigación no ha sido plagiado total ni parcialmente.
2. El trabajo de investigación no ha sido autoplagiada, es decir, no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
3. Los datos presentados en los resultados son reales, no han sido falseados, ni duplicados, ni copiados y por tanto los resultados que se presentan en el trabajo de investigación constituirán aporte a la realidad investigada.

Soy consciente de que, el hecho de no respetar los derechos de autor y hacer plagio, es objeto de sanciones universitarias y/o legales, por lo que asumo cualquier responsabilidad que pudiera derivarse de irregularidades en el trabajo de investigación, así como de los derechos sobre la obra presentada.

Moquegua, 15 enero del 2026



Huella digital

Nombres y Apellidos: Anthony Roger Flor Velásquez
DNI: 47071758

ANEXO VIII

DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN / TESIS / TRABAJO DE SUFICIENCIA PROFESIONAL

Yo, María Del Pilar Laura Ramírez Chipana ,
bachiller de la Escuela Profesional de:
Ingeniería de Sistemas e Informática ,
de la Universidad Nacional de Moquegua, identificado(a) con DNI Nº... 72659077 , autor(a) del Trabajo de Investigación / Tesis / Trabajo de suficiencia profesional denominado:
SISTEMA DE RECOMENDACIÓN DE JURISPRUDENCIA BASADO EN EMBEDDINGS DE TEXTO PARA LA IDENTIFICACIÓN Y SUGERENCIA DE PRECEDENTES RELEVANTES DE LAS CORTES SUPERIORES DE JUSTICIA DE LA REPÚBLICA DEL PERÚ

Declaro bajo juramento que:

1. El trabajo de investigación es auténtico, siendo resultado de mi trabajo personal, que no se ha copiado, que no se ha utilizado ideas, formulaciones, citas integrales e ilustraciones diversas, sacadas de cualquier tesis, obra, artículo, memoria, etc., (en versión digital o impresa), sin mencionar de forma clara y exacta su origen o autor, tanto en el cuerpo del texto, figuras, cuadros, tablas u otros que tengan derechos de autor. Por tanto, el trabajo de investigación no ha sido plagiado total ni parcialmente.
2. El trabajo de investigación no ha sido autoplagiada, es decir, no ha sido publicada ni presentada anteriormente para obtener algún grado académico previo o título profesional.
3. Los datos presentados en los resultados son reales, no han sido falseados, ni duplicados, ni copiados y por tanto los resultados que se presentan en el trabajo de investigación constituirán aporte a la realidad investigada.

Soy consciente de que, el hecho de no respetar los derechos de autor y hacer plagio, es objeto de sanciones universitarias y/o legales, por lo que asumo cualquier responsabilidad que pudiera derivarse de irregularidades en el trabajo de investigación, así como de los derechos sobre la obra presentada.

Moquegua, 15 de enero del 2026



Huella digital

Nombres y Apellidos: María Del Pilar L. Ramírez Chipana
DNI: 72659077

UNAM

Firmado digitalmente por TITO CHURA Hugo Euler FAU 20449347448 soft
Motivo: Soy el autor del documento
Fecha: 03.12.2025 15:12:11 -05:00

Hugo Euler TITO CHURA

INFORME TESIS Sistema de recomendación de jurisprudencia basado en embeddings de texto.pdf

Tesis de María

Tesis 2025

Universidad Nacional de Moquegua

Detalles del documento

Identificador de la entrega
trn:oid:::1:3408465716

Fecha de entrega
12 nov 2025, 4:18 p.m. GMT-5

Fecha de descarga
12 nov 2025, 4:21 p.m. GMT-5

Nombre del archivo
INFORME_TESIS_Sistema_de_recomendación_de_jurisprudencia_basado_en_embeddings_de_texto.pdf

Tamaño del archivo
3.0 MB

113 páginas

25.749 palabras

152.123 caracteres

21

3% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 15 palabras)

Fuentes principales

2% Fuentes de Internet
1% Publicaciones
3% Trabajos entregados (trabajos del estudiante)

turnitin Página 2 de 118 - Descripción general de integridad Identificador de la entrega trn:oid:::1:3408465716

Fuentes principales

2% Fuentes de Internet
1% Publicaciones
3% Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1	Trabajos del estudiante Universidad Europea de Madrid	<1%
2	Trabajos del estudiante Universidad Nacional de Moquegua	<1%
3	Trabajos del estudiante Universidad Loyola Andalucia	<1%
4	Internet www.collegesidekick.com	<1%
5	Internet repositorio.unjbg.edu.pe	<1%
6	Internet repositorio.continental.edu.pe	<1%
7	Trabajos del estudiante Universidad Autónoma de Bucaramanga,UNAB	<1%
8	Trabajos del estudiante Corporación Universitaria Minuto de Dios, UNIMINUTO	<1%
9	Internet repositorio.upse.edu.ec	<1%
10	Trabajos del estudiante Pontificia Universidad Catolica del Peru	<1%
11	Publicación Choque Castro, Marcos Denys. "Modelo de análisis de tópicos en ordenanzas regi...	<1%

12 Trabajos del estudiante
UNIBA <1%

13 Trabajos del estudiante
Universidad Católica del Norte <1%

14 Trabajos del estudiante
Universidad Internacional de la Rioja <1%

15 Trabajos del estudiante
Universidad del País Vasco <1%

16 Internet
computerhoy.com <1%

17 Internet
peteringwersen.info <1%

18 Internet
www.elsevier.es <1%

19 Internet
www.repositorio.unam.edu.pe <1%

20 Internet
github.com <1%

21 Trabajos del estudiante
Instituto Superior de Artes, Ciencias y Comunicación IACC <1%

22 Trabajos del estudiante
Universitat Politècnica de València <1%

23 Internet
dspace.unitru.edu.pe <1%

24 Internet
dspace.ups.edu.ec <1%

25 Internet
hdl.handle.net <1%

26 Trabajos del estudiante
ufidelitas <1%

27 Internet
www2.ing.puc.cl <1%

RESOLUCIÓN DE CONSEJO UNIVERSITARIO
N° 039-2026-CU

Moquegua, 28 de enero de 2026

VISTOS:

El Proveído N° 248-2026-R de fecha 22 de enero de 2026 emitido por el Rectorado, Informe N° D0023-2026-UNAM-R-SG de fecha 22 de enero de 2026 emitido por el Secretario General, Informe Técnico N° D0002-2026-SEGE-GT de Grados y Títulos, el Acuerdo del Pleno del Consejo Universitario de Sesión Extraordinaria N° 05-2026-SE-CU-UNAM, de fecha 28 de enero de 2026, y;

CONSIDERANDO:

Que, conforme al 4to párrafo del artículo 18, de la Constitución Política del Perú, dispone que cada Universidad es autónoma en su régimen normativo, de gobierno, académico, administrativo y económico. Y el artículo 8º de la Ley Nº 30220-Ley Universitaria, establece que la referida autonomía inherente a las Universidades, se ejerce de conformidad con la Constitución y demás normativa aplicable. Concordante con el artículo 1, de la Ley Nº 31520, Ley que restablece la autonomía y la institucionalidad de las universidades peruanas, así como la Ley N° 32141, Ley que modifica la Ley N° 30220-Ley Universitaria, a fin de fortalecer la institucionalidad de las Universidades Públicas licenciadas;

Que, el Artículo 58, de la Ley N° 30220-Ley Universitaria, establece que el Consejo Universitario es el máximo órgano de gestión, dirección y ejecución académica y administrativa de la Universidad. Concordante con el artículo 18° del Estatuto de la Universidad Nacional de Moquegua, aprobado bajo Resolución de Asamblea Universitaria N° 015-2025-AU-UNAM;

Que, el artículo 59 de la Ley Universitaria-Ley N° 30220, establece las atribuciones del Consejo Universitario, y específicamente el numeral 59.9, establece; Conferir los grados académicos y los títulos profesionales aprobados por las Facultades y Escuela de Posgrado, así como otorgar distinciones honoríficas y reconocer y revalidar los estudios, grados y títulos de universidades extranjeras, cuando la universidad está autorizada por la Superintendencia Nacional de Educación Superior Universitaria. Concordante con el numeral 14, del artículo 20° del Estatuto de la Universidad Nacional de Moquegua;

Que, el Sr. Rector es el personero y representante legal de la Universidad Nacional de Moquegua, conforme a lo dispuesto por la Ley Nº 30220-Ley Universitaria, tiene a su cargo la Dirección, Conducción y Gestión del Gobierno Universitario en todos sus ámbitos, concordante con el artículo 24° y numeral 2 del artículo 26° del Estatuto de la Universidad Nacional de Moquegua;

Que, mediante Informe Técnico N° D0002-2026-SEGE-GT de fecha 21 de enero de 2026, e Informe N° D0023-2026-UNAM-R-SG de fecha 22 de enero de 2026, el Secretario General de la UNAM, remite al Rectorado los expedientes para el trámite de Título Profesional de Ingeniero de Sistemas e Informática a los Bachilleres: RAMÍREZ CHIPANA MARÍA DEL PILAR LAURA, FLOR VELASQUEZ ANTHONY ROGER y RAMOS CALLA PEDRO LUIS de la Escuela Profesional de Ingeniería de Sistemas e Informática, quienes se encuentran aptos para la expedición del Título Profesional de Ingeniero de Sistemas e Informática, al haber cumplido satisfactoriamente con los requisitos legales y reglamentarios, exigidos en el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, conforme lo señala la Dirección de Actividades y Servicios Académicos y la Dirección de la Escuela Profesional de Ingeniería de Sistemas e Informática, razón por la cual, el señor Secretario General incluye dichos expedientes como punto de agenda para ser tratado en Sesión Extraordinaria de Consejo Universitario de fecha 28 de enero de 2026;

Que, el artículo 44° de la Ley N° 30220, Ley Universitaria señala que las universidades otorgan los grados académicos de Bachiller, Maestro, Doctor y los Títulos Profesionales que correspondan, a nombre de la Nación (...). Asimismo, la Ley N° 31803, modifica el artículo en mención adicionando; La Superintendencia Nacional de Educación Superior Universitaria (SUNEDU) determina los criterios, así como establece y supervisa el procedimiento administrativo para el reconocimiento de grados y títulos otorgados en el extranjero, en el marco de la normativa vigente;

Que, el numeral 45.2 del artículo 45° de la Ley Nº 30220, Ley Universitaria, precisa los requisitos mínimos para la obtención del título profesional, como son: el grado de Bachiller y la aprobación de una tesis o trabajo de suficiencia profesional. Asimismo, establece que el título profesional solo se puede obtener en la universidad en la cual se haya obtenido el grado de bachiller;

Que, mediante Resolución de Comisión Organizadora N° 1153-2021-UNAM, se aprobó el Reglamento de Grados y Títulos de la Universidad Nacional de Moquegua, el mismo que fue modificado a través de las Resoluciones de Comisión Organizadora N° 410; 453 y 890-2022, N° 880-2023-UNAM y N° 081, 484-2024-UNAM el cual establece los requisitos y norma

Domicilio Legal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe

RESOLUCIÓN DE CONSEJO UNIVERSITARIO
N° 039-2026-CU

Moquegua, 28 de enero de 2026

los procedimientos para la obtención de los grados académicos y títulos profesionales conferidos por esta Casa Superior de Estudios;

Que, bajo el amparo de los considerandos precedentes, el Pleno del Consejo Universitario en Sesión Extraordinaria N° 05-2026-SE-CU-UNAM, de fecha 28 de enero de 2026, emitió, el siguiente, Acuerdo, POR UNANIMIDAD: CONFERIR, a nombre de la Nación el Título Profesional de Ingeniero de Sistemas e Informática por la modalidad de sustentación de tesis a los Bachilleres: RAMÍREZ CHIPANA MARÍA DEL PILAR LAURA, FLOR VELASQUEZ ANTHONY ROGER y RAMOS CALLA PEDRO LUIS, de la Escuela Profesional de Ingeniería de Sistemas e Informática, de la Facultad de Ingenierías de la Universidad Nacional de Moquegua;

En uso de las facultades y atribuciones conferidas por el Art. 18, de la Constitución Política del Perú, la Ley N° 30220 Ley Universitaria, el Estatuto de la Universidad Nacional de Moquegua, y a lo acordado por el pleno del Consejo Universitario:

SE RESUELVE:

ARTÍCULO 1°: CONFERIR, A NOMBRE DE LA NACIÓN a los Bachilleres

- **RAMÍREZ CHIPANA MARÍA DEL PILAR LAURA**
  TESIS : "Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú", cuyo Asesor Principal fue el Mgr. Hugo Euler Tito Chura, y el Co - asesor fue el Ing. Honorio Apaza Alanoca.
- **FLOR VELASQUEZ ANTHONY ROGER**
  TESIS : "Sistema de recomendación de jurisprudencia basado en embeddings de texto para la identificación y sugerencia de precedentes relevantes de las Cortes Superiores de Justicia de la República del Perú", cuyo Asesor Principal fue el Mgr. Hugo Euler Tito Chura, y el Co - asesor fue el Ing. Honorio Apaza Alanoca.
- **RAMOS CALLA PEDRO LUIS**
  TESIS : "Modelos de Deep learning para la detección de noticias falsas en español sobre COVID-19", cuyo Asesor Principal fue el Dr. Anibal Fernando Flores García.

El **TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS E INFORMÁTICA.**

Por la modalidad de SUSTENTACIÓN DE TESIS, de la Escuela Profesional de Ingeniería de Sistemas e Informática, de la Facultad de Ingenierías de la Universidad Nacional de Moquegua.

ARTÍCULO 2°: ENCARGAR, al Secretario General a través de quien corresponda, para que realice las acciones pertinentes relacionadas con la inscripción correspondiente en el Registro de Grados y Títulos de la Superintendencia Nacional de Educación Superior Universitaria - SUNEDU.

ARTÍCULO 3°: DISPONER, el cumplimiento de la presente Resolución a todas las dependencias de la Universidad Nacional de Moquegua.

ARTÍCULO 4°: DISPONER la publicación de la presente resolución en el Portal Institucional de la Universidad Nacional de Moquegua www.unam.edu.pe

REGÍSTRESE, PUBLÍQUESE Y CÚMPLASE;

UNIVERSIDAD NACIONAL DE MOQUEGUA SECRETARIA GENERAL

ABOG. EDGAR PINTO HUANCAVILCA
SECRETARIO GENERAL

UNIVERSIDAD NACIONAL DE MOQUEGUA CONSEJO UNIVERSITARIO

DR. HUGO RUBEN MARCA MAQUERA
RECTOR
PRESIDENTE DEL CONSEJO UNIVERSITARIO

DISTRIBUCIÓN:
VRA VRI FACS EPGPDS GyT DASA OTI Interesado c.c. Arch.

Domicilio Legal: Prolongación Av., Ancash S/N - Moquegua
RUC: 20449347448 - Teléfono: (053) 670168
Página Web: www.unam.edu.pe