Transformación digital en la gestión contractual: automatización inteligente de la identificación de requisitos contractuales mediante ciencia de datos

Mónica Jazmín Antolínez Becerra - Código 8973053
Adriana Marcela Güiza Saavedra - Código 8992246

Tesis de maestría presentada para optar al título de Maestría en Ciencia de datos

Directora: Gloria Inés Álvarez Vargas, Doctor (PhD) en Reconocimiento de Formas e Inteligencia Artificial
Codirector: Diego Luis Linares Ospina, Doctor (PhD) en Ingeniería de la Programación e Inteligencia artificial.

Pontificia Universidad Javeriana de Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Santiago de Cali
2025

Dedicatoria

A nuestras familias, porque sin su apoyo y paciencia, esto no hubiera sido posible.

Agradecimientos

A nuestros directores de proyecto, por su constante disposición, apoyo y motivación a lo largo de todo este proceso.

TABLA DE CONTENIDO

1. DEFINICIÓN DEL PROBLEMA 10

1.1 Planteamiento del problema 10

1.2 Formulación del problema 11

2. OBJETIVOS 12

2.1 Objetivo general 12

2.2 Objetivos específicos 12

3. MARCO TÉORICO Y ANTECEDENTES 13

3.1 Marco Teórico 13

3.1.1 Gestión contractual 13

3.1.2 Procesamiento de lenguaje natural 13

3.1.3 Representación de Textos 15

3.1.4 Aprendizaje automático 16

3.1.5 Aprendizaje supervisado 17

3.2 Antecedentes 19

4. PREPARACIÓN DE DATOS 22

4.1 Alistamiento y Consolidación del Corpus 22

4.2 Preprocesamiento del Texto 23

4.3 Etiquetado Manual de requisitos 24

4.4 Balanceo de clases 26

5. MODELADO 29

5.1 Vectorización 29

5.1.1. Representación TF-IDF 29

5.1.1. Representación Word2Vec 30

5.2 Modelado preliminar con datos etiquetados manualmente 30

5.2.1 División entrenamiento y prueba 30
5.2.2 Entrenamiento de modelos SVC preliminares con TF-IDF y Word2Vec 31
5.2.3 Evaluación y selección preliminar 33
5.3 Etiquetado automático Supervisado 37
5.3.1 Identificación y separación de minutas sin etiquetar 37
5.3.2 Configuración de los modelos de predicción 37
5.3.3 Aplicación de modelos y generación de predicciones 38
5.4 Modelado final sobre el corpus completo etiquetado 39
5.4.1 División del conjunto de datos 39
5.4.2 Vectorización global 39
5.4.3 Balanceo de clases 40
5.4.4 Entrenamiento final con parámetros optimizados previamente 40
5.4.5. Análisis De Los Resultados 43
6. IMPLEMENTACIÓN DE UN PROTOTIPO DE IDENTIFICACIÓN AUTOMÁTICA DE REQUISITOS CONTRACTUALES 46
7. DISCUSIÓN 48
8. CONCLUSIONES 50

LISTA DE TABLAS

Tabla 1. Proporción de clases por requisito. 27
Tabla 2. Resultados de entrenamiento por cada etiqueta 33
Tabla 3. Resultado de las evaluaciones aplicando F1-Score Macro 36
Tabla 4. Modelo por cada Requisito 41
Tabla 5. Métricas de desempeño Final de los Modelos 43

LISTA DE IMÁGENES

Imagen 1. Aprendizaje supervisado 17
Imagen 2.Muestra minutas preprocesadas 24
Imagen 3. Muestra de 100 minutas 24
Imagen 4. Dataset etiquetado manual Excel 25
Imagen 5. Dataset contenido minutas con etiquetado 26
Imagen 6. Resultados aplicando undersampling en el dataframe del requisito Uso de Opción ... 28
Imagen 7. Muestra del contenido de un documento vectorizado 30
Imagen 8. Muestra etiquetado automático 38
Imagen 9. Predicción porcentual de presencia por cada requisito 39
Imagen 10. Visualización del Prototipo de Sistema de Identificación de Requisitos Contractuales en Streamlit 47

RESUMEN

Las entidades que contratan con recursos públicos en Colombia deben velar por la transparencia en el proceso contractual, para ello se tiene un sistema electrónico donde reposa toda la información para dicho fin denominado SECOP. Esta gestión la ejecutan profesionales de gestión contractual asegurando que se cumplan los requisitos acordados en los documentos para el seguimiento periódico. La identificación de los requisitos es una tarea que actualmente se hace manualmente y al no ser una tarea exclusiva, se corre con el riesgo de pasar por alto requisitos que puedan poner en peligro a la empresa en términos económicos, legales entre otros.

Este proyecto consiste en la identificación automatizada de requisitos contractuales, utilizando técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN) y Aprendizaje Automático (AA). El sistema desarrollado toma como entrada documentos contractuales en formato PDF, extrae su contenido textual y lo somete a procesos de limpieza, normalización y vectorización.

A partir de una muestra de minutas etiquetadas manualmente, se entrenaron modelos supervisados de clasificación binaria para cada requisito contractual, utilizando dos técnicas de representación de texto: TF-IDF y Word2Vec. Como algoritmo de clasificación se empleó Support Vector Classifier (SVC), optimizado mediante búsqueda en cuadrícula (GridSearchCV) para maximizar el desempeño de predicción.

Posteriormente, se aplicaron los modelos entrenados para etiquetar automáticamente un corpus más amplio de minutas sin ninguna etiqueta. Con la base de datos consolidada, se realizó un entrenamiento final de los modelos para cada requisito, seleccionando la técnica de vectorización más adecuada según los resultados obtenidos en validaciones anteriores. Las métricas utilizadas para evaluar el desempeño fueron: Accuracy, Precision Macro, Recall Macro, F1-Score Macro, Precision Weighted, Recall Weighted, F1-Score Weighted, esta última priorizada, debido al desbalance de las clases.

La herramienta cuenta con una interfaz de usuario intuitiva y funcional que permite a los profesionales de gestión contractual adjuntar documentos contractuales y recibir un listado de todos los requisitos contractuales identificados. Esta interfaz facilita la carga de documentos y la visualización del listado con los requisitos contractuales identificados.

ABSTRACT

Entities that contract with public resources in Colombia must ensure transparency in the contractual process. To this end, an electronic system called SECOP stores all information for this purpose. This process is carried out by contract management professionals, ensuring that the requirements agreed upon in the documents are met for periodic monitoring. Identifying requirements is currently performed manually, and since it is not exclusively performed, it carries the risk of exceeding high requirements that could jeopardize the company in economic, legal, and other terms.

This project involves the automated identification of contractual requirements using advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques. The developed system takes contractual documents in PDF format as input, extracts their textual content, and subjects it to cleansing, normalization, and vectorization processes.

Using a sample of manually labeled minutes, supervised binary classification models were trained for each contractual requirement, using two text representation techniques: TF-IDF and Word2Vec. The Support Vector Classifier (SVC) was implemented as a classification algorithm, optimized using grid search (GridSearchCV) to maximize prediction performance.

The trained models were then applied to automatically label a larger corpus of unlabeled bills of minutes. With the consolidated database, the models were then trained for each requirement, selecting the most appropriate vectorization technique based on the results obtained in previous validations. The metrics used to evaluate performance were: Accuracy, Precision Macro, Recall Macro, F1-Score Macro, Precision Weighted, Recall Weighted, and F1-Score Weighted, the latter being prioritized due to class imbalance.

The tool features an intuitive and functional user interface that allows contract management professionals to attach contractual documents and receive a list of all identified contractual requirements. This interface facilitates document upload and viewing the list of identified contractual requirements.

INTRODUCCIÓN

El seguimiento a la gestión contractual con recursos públicos en Colombia es fundamental porque permite velar por los mismos asegurando la transparencia y ejecución correcta. Actualmente este seguimiento se hace de forma manual lo que hace que el seguimiento esté propenso a errores y omisiones poniendo en riesgo la integridad del proceso, generando posibles problemas económicos y legales para los que participan en él.

En este trabajo se desarrolló un sistema automatizado que permite la lectura de las minutas e identificación de los requisitos básicos para el respectivo seguimiento. Para ello, se utilizó técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN) y Aprendizaje Automático (AA), metodologías de entrenamiento y validación de modelos para optimizar su rendimiento, ajustando hiperparámetros y utilizando técnicas de regularización.

Para confirmar la efectividad de los resultados Este proyecto también contempla la evaluación del sistema con métricas específicas, casos de estudio utilizando contratos reales y comparaciones con métodos tradicionales, con el objetivo de confirmar su efectividad.

Por último, se desarrolló una interfaz de usuario amigable que permite a los gestores de contratos cargar documentos y ver automáticamente los requisitos identificados. Esto facilitará su trabajo diario y reducirá los riesgos relacionados con errores humanos en la revisión de contratos.

Con todo lo anterior, se busca contribuir a la modernización y mejora de los procesos de gestión contractual pública en Colombia, fomentando la transparencia, la eficiencia y el cumplimiento normativo a través de la innovación tecnológica.

1. DEFINICIÓN DEL PROBLEMA

1.1 Planteamiento del problema

El seguimiento administrativo de contratos es crucial para asegurar el cumplimiento de las obligaciones contractuales establecidas. Actualmente, este proceso se realiza manualmente con documentos almacenados en la plataforma SECOP, donde diferentes roles gestionan y aseguran la ejecución de los controles del proceso. Uno de estos roles es el Profesional de Gestión Contractual, que debe asegurar el cumplimiento estricto del contrato revisando la documentación para identificar los requisitos que deben monitorear periódicamente.

El proceso manual de revisión y extracción de requisitos contractuales es laborioso y propenso a errores debido a la carga de trabajo elevada y la cantidad significativa de contratos a gestionar. Además, estos profesionales a menudo tienen responsabilidades adicionales, lo que aumenta el riesgo de que algunos requisitos de cumplimiento no sean debidamente identificados. El incumplimiento de estos requisitos debido a un monitoreo inadecuado puede acarrear consecuencias económicas, reputacionales y legales significativas para la empresa.

Desde la perspectiva de la ciencia de datos, este problema presenta varios desafíos que se pueden abordar mediante técnicas avanzadas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (AA). Los documentos contractuales, generalmente en formato PDF, contienen información no estructurada que debe ser transformada en datos estructurados para su análisis automático. La extracción precisa y eficiente de los requisitos contractuales implica el desarrollo de modelos de PLN que puedan interpretar el lenguaje jurídico y reconocer patrones relevantes en los textos contractuales.

Además, la gran cantidad de datos y la diversidad de formatos y estructuras en los contratos requieren algoritmos de AA capaces de aprender y generalizar a partir de ejemplos etiquetados. Esto involucra procesos de limpieza y preprocesamiento de datos, selección y entrenamiento de modelos, y validación de su desempeño. La optimización de estos modelos es esencial para garantizar su precisión y eficiencia en un entorno real.

Para abordar este desafío, se desarrolló una herramienta automatizada que utiliza técnicas avanzadas de PLN y AA para leer automáticamente los documentos contractuales y listar los requisitos del contrato. Esta herramienta permite un monitoreo más eficiente y efectivo de las obligaciones contractuales, reduciendo el riesgo de incumplimientos y sus posibles consecuencias negativas. Mediante la implementación de técnicas de ciencia de datos, se espera transformar el proceso de gestión contractual, mejorando la precisión, reduciendo el tiempo de revisión y minimizando los errores humanos.

1.2 Formulación del problema

¿Cómo puede desarrollarse una herramienta automatizada utilizando técnicas avanzadas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (AA) para resolver la problemática de la extracción precisa y eficiente de requisitos en documentos contractuales, mejorando así el monitoreo y cumplimiento de las obligaciones contractuales en organizaciones complejas?

¿Cuáles son las técnicas más efectivas para recolectar y preprocesar los documentos contractuales en formato PDF adjuntados por profesionales en la aplicación, para su posterior análisis mediante técnicas de PLN y AA?

¿Qué algoritmos de PLN y AA son más adecuados para entrenar modelos que identifiquen y clasifiquen de manera precisa los requisitos contractuales en documentos?

¿Qué métricas y criterios son más adecuados para evaluar la efectividad de los modelos entrenados en términos de precisión y utilidad en la gestión contractual?

2. OBJETIVOS

2.1 Objetivo general

Desarrollar una solución automatizada basada en técnicas avanzadas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (AA) para la extracción y clasificación precisa de los requisitos contractuales de contratos.

2.2 Objetivos específicos

- Desarrollar técnicas efectivas para la recolección y preprocesamiento de documentos contractuales en formato PDF.
- Seleccionar y entrenar algoritmos de procesamiento de lenguaje natural (PLN) y aprendizaje automático (AA) para la identificación y clasificación precisa de requisitos contractuales en documentos.
- Definir y aplicar métricas y criterios adecuados para evaluar la efectividad de los modelos entrenados en términos de precisión y utilidad en la gestión contractual.

3. MARCO TÉORICO Y ANTECEDENTES

3.1 Marco Teórico

3.1.1 Gestión contractual

La Gestión de Contratos tiene como objetivo asegurar el inicio, ejecución, cierre y balance de los contratos, realizando una adecuada gestión de eventualidades y realizando la evaluación del desempeño del contratista.

Dentro del aseguramiento de la gestión contractual se tiene la ejecución contractual que es está cargo de un funcionario de seguimiento quien debe velar por que se cumpla y asegure la documentación de los requisitos contractuales previstos en la ejecución del contrato. En cada contrato dependiendo su fin, varían estos requisitos contractuales.

Para este proyecto se han escogido los requisitos contractuales más generales e indispensables para una correcta ejecución contractual que son: Retención en garantía, Gastos reembolsables, Uso de opción, Reajuste salarial, Reajuste de tarifas y precios, Anticipo o pago anticipado, Garantías y seguros, Cláusula de Cesión, Contribución de obra pública, Estampilla Universidad Nacional, Subcontratación, GAB-F-213, GAB-F-214, GAB-F-221, GAB-F-060, GAB-F-105, Reunión de inicio, Socialización.

3.1.2 Procesamiento de lenguaje natural

El procesamiento del lenguaje natural (PLN) combina la lingüística computacional (modelización del lenguaje humano basada en reglas) con modelos estadísticos y de machine learning para que los ordenadores y dispositivos digitales reconozcan, comprendan y generen texto y voz.

El PLN, una rama de la inteligencia artificial (IA), se encuentra en el corazón de las aplicaciones y los dispositivos que pueden:

- Traducir texto de un idioma a otro
- Responder a órdenes escritas u orales

- Reconocer o autenticar usuarios por voz
- Resumir grandes volúmenes de texto
- Evaluar la intención o el sentimiento de un texto o discurso
- Generar texto o gráficos u otros contenidos a petición

Esto se realiza con el objetivo de desarrollar técnicas y herramientas que permitan la implementación de sistemas capaces de interpretar y utilizar el lenguaje natural para desempeñar las tareas deseadas, como, por ejemplo, una identificación de requisito dentro de un documento contractual.

El PNL se basa en una serie de técnicas y enfoques, cada uno diseñado para abordar aspectos específicos del procesamiento del lenguaje humano. A continuación, se presentan algunas de las clases fundamentales de PNL que son ampliamente utilizadas en la investigación y la industria.

- ***Tokenización****:* Es el proceso de dividir un texto en unidades más pequeñas, como palabras o caracteres. La tokenización es el primer paso en muchos sistemas de PNL y es fundamental para analizar y procesar el texto.

- ***Análisis morfológico:*** Se refiere al análisis de la estructura y la forma de las palabras en un texto. Esto incluye identificar la raíz de una palabra, su forma flexionada y otras características morfológicas.

- ***Etiquetado de Partes del Discurso (POS tagging)***: Consiste en asignar a cada palabra en un texto una etiqueta que indica su categoría gramatical, como sustantivo, verbo, adjetivo, etc.

- ***Análisis sintáctico:*** Implica analizar la estructura sintáctica de una oración para comprender las relaciones entre las palabras y la estructura gramatical. Esto puede incluir la identificación de frases, cláusulas y la jerarquía gramatical.

- ***Análisis Semántico:*** Se centra en comprender el significado de las palabras y las relaciones entre ellas en un contexto determinado. Esto puede incluir la identificación de entidades nombradas, la resolución de la correferencia y el análisis de sentimientos.

- ***Generación de lenguaje natural***: Implica generar texto de forma automática a partir de datos o instrucciones dadas. Esto puede incluir la generación de resúmenes automáticos, la traducción automática y la respuesta automática a preguntas.

3.1.3 Representación de Textos

El procesamiento de lenguaje natural requiere transformar los datos textuales en una estructura que sea interpretable por los algoritmos de aprendizaje automático. Este proceso se conoce como **representación vectorial de textos** o **vectorización**, y consiste en convertir documentos en vectores numéricos que capturen información relevante de su contenido lingüístico. A continuación, se describen los principales enfoques utilizados en esta investigación: TF-IDF y Word2Vec.

- **Frecuencia de Término – Frecuencia Inversa de Documento (TF-IDF):** Este modelo pondera cada término de acuerdo con su importancia relativa en el corpus. La frecuencia de término (TF) mide la frecuencia del término en un documento específico, mientras que la frecuencia inversa de documento (IDF) mide la rareza del término en el conjunto total de documentos y no capta relaciones semánticas entre palabras. Este enfoque reduce el peso de palabras comunes y frecuentes en todos los documentos, y destaca los términos más discriminativos.

  La frecuencia se mide como: $TF\text{-}IDF_{i\,j} = TF_{ij} \cdot \log\left(\frac{N}{df_j}\right)$

  Donde: donde:

  - TF_{ij} : frecuencia del término t_j en el documento d_i
  - df_j: número de documentos que contienen el término t_j
  - N: número total de documentos en el corpus.

- **Word2Vec** es un modelo de representación densa (embeddings) que proyecta palabras en un espacio vectorial de menor dimensión, donde las relaciones semánticas entre palabras se preservan en la geometría del espacio y se entrena mediante redes neuronales simples usando dos enfoques principales:
    - **CBOW (Continuous Bag of Words)**: predice una palabra a partir del contexto circundante.
    - **Skip-gram**: predice el contexto a partir de una palabra objetivo.

  El Word2Vec representa cada palabra mediante un vector de características continuas, permitiendo que palabras con significados similares estén cerca en el espacio vectorial. Para representar un documento, se pueden utilizar estrategias como el promedio o la suma de los vectores de palabras que lo componen. [1]

  La principal ventaja de Word2Vec es que preserva la semántica del lenguaje. No obstante, requiere mayor cantidad de datos para su entrenamiento efectivo, y su interpretación puede ser menos intuitiva.

La elección del método de vectorización depende del tipo de tarea, del tamaño y naturaleza del corpus, y de los algoritmos de clasificación a utilizar.

3.1.4 Aprendizaje automático

En el contexto de la automatización inteligente de la identificación de requisitos contractuales, el Aprendizaje Automático (AA) se convierte en una herramienta fundamental. Siguiendo la teoría del aprendizaje supervisado, propuesta por Tom Mitchell en su influyente libro "Machine Learning" [2], podemos comprender cómo los modelos de AA pueden ser entrenados a partir de conjuntos de datos etiquetados de requisitos contractuales previamente identificados. Esta perspectiva teórica subyacente guía el desarrollo de algoritmos y técnicas de AA que tienen como objetivo aprender patrones y estructuras a partir de datos contractuales, lo que permite automatizar el proceso de identificación de requisitos. Al explorar los principios del aprendizaje supervisado

en el contexto específico de la gestión contractual, podemos comprender cómo estas técnicas de AA pueden contribuir a mejorar la eficiencia y precisión del monitoreo contractual, reduciendo el riesgo de incumplimientos y sus consecuencias negativas.

3.1.5 Aprendizaje supervisado

Se define por el uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifican los datos o predicen los resultados con precisión.

A medida que se introducen datos en el modelo, éste ajusta sus ponderaciones hasta que el modelo se ha ajustado adecuadamente, lo que ocurre como parte del proceso de validación cruzada. Se puede utilizar para crear modelos de machine learning de alta precisión. [3]

Imagen 1. Aprendizaje supervisado

Fuente: Tomado de la presentación Ciencia de datos –Retos en Ciencia de datos Pontifica Universidad Javeriana de Cali.

Dentro de las técnicas de aprendizaje se cuenta con:

El **k-Nearest Neighbors (k-NN)** es un algoritmo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Su fundamento teórico, originado en los trabajos de Cover y

Hart (1967), se basa en la noción de que instancias cercanas en el espacio de características tienden a compartir etiquetas similares. [3]

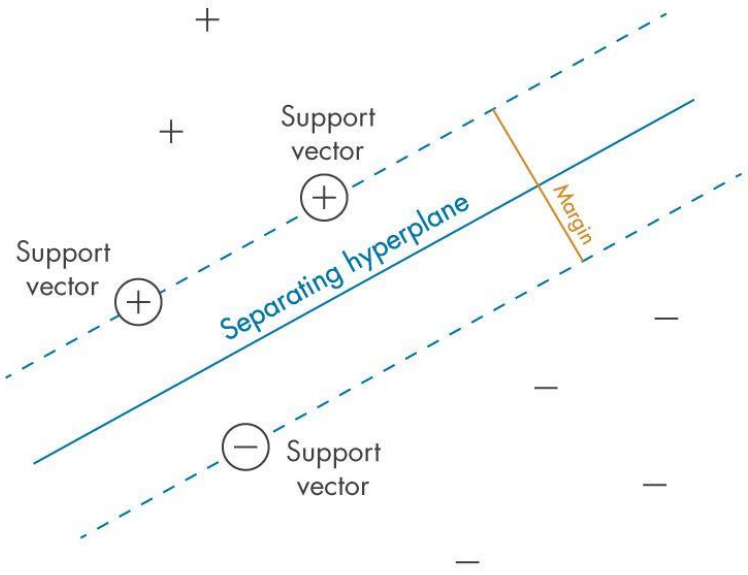
Los pasos de este algoritmo se pueden resumir en 6 etapas:

1. Seleccionar el número de K vecino
2. Calcular la distancia. Las mas utilizadas son la Eucladiana y Manhattan
3. Toma las k vecinos más cercanas según la distancia calculada
4. Entre los k vecinos, cuenta el número de puntos en cada categoría
5. Atribuye un nuevo punto a la categoría mas presente entre los k vecinos
6. El modelo está listo.

SVM: es una clasificación y regresión robustas Técnica que maximiza la precisión predictiva de un modelo sin sobreajustar los datos de entrenamiento. SVM es especialmente adecuado para analizar datos con un número muy grande (por ejemplo, miles) de campos de predicción.

SVM funciona mediante la asignación de datos a un espacio de entidades de alta dimensión para que los datos Los puntos se pueden categorizar, incluso cuando los datos no se pueden separar linealmente. Un separador entre las categorías, entonces los datos se transforman de tal manera que el separador podría dibujarse como un hiperplano. A continuación, las características de los nuevos datos se pueden utilizar para predecir el grupo al que debe pertenecer un nuevo registro.

El algoritmo SVM se utiliza ampliamente en el machine learning, ya que puede manejar tareas de clasificación tanto lineales como no lineales. Sin embargo, cuando los datos no son separables linealmente, se utilizan funciones de núcleo para transformar los datos en un espacio de mayor dimensión y permitir la separación lineal. Esta aplicación de las funciones de kernel puede conocerse como el "truco de kernel", y la elección de la función de kernel, como los kernels lineales, los kernels polinómicos, los kernels de función de base radial (RBF) o los kernels sigmoides, depende de las características de los datos y del caso de uso específico. [3]



Imágen 1. Explicación funcionamiento algoritmos SVM

Figura tomada de: https://www.ibm.com/es-es/think/topics/support-vector-machine

3.2 Antecedentes

La automatización en la gestión contractual ha despertado un creciente interés en la literatura académica. Investigaciones recientes han explorado la aplicación de técnicas avanzadas de Procesamiento de Lenguaje Natural (PNL) y Aprendizaje Automático (AA) para mejorar la eficiencia y precisión en la identificación de requisitos contractuales. A continuación, se presentan estudios previos que han investigado la aplicación de tecnologías de transformación digital en este ámbito, proporcionando un marco teórico robusto para el desarrollo de soluciones innovadoras en la gestión contractual.

Ashley, examina cómo las tecnologías de inteligencia artificial, y en particular el procesamiento de lenguaje natural (PLN), están revolucionando la práctica del derecho. Este libro, titulado "Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age" [4], explora en profundidad las aplicaciones del PLN en el análisis de documentos legales complejos. Ashley argumenta que las herramientas basadas en inteligencia artificial están proporcionando a los profesionales del derecho capacidades avanzadas para analizar y entender grandes volúmenes

de texto legal, identificando patrones, conceptos y relaciones que serían extremadamente difíciles de detectar manualmente.

El autor describe cómo los modelos de PLN pueden extraer información relevante de contratos, legislación, casos judiciales y otros documentos legales, facilitando tareas como la revisión de documentos, la búsqueda de precedentes y la predicción de resultados legales. Además, Ashley discute el impacto de estas tecnologías en la eficiencia y precisión del trabajo legal, señalando que la automatización de procesos rutinarios permite a los abogados concentrarse en actividades de mayor valor añadido, como la estrategia legal y el asesoramiento a cliente.

En la investigación de Chalkidis examinan el uso de técnicas de procesamiento de lenguaje natural (PLN) para la extracción de elementos específicos en contratos legales. En el trabajo titulado "Extracting Contract Elements", los autores demuestran la viabilidad de automatizar la identificación de cláusulas contractuales mediante algoritmos de clasificación y modelos de aprendizaje profundo. El estudio se centra en cómo estas técnicas pueden ser aplicadas para analizar y procesar grandes volúmenes de contratos, identificando automáticamente cláusulas relevantes que suelen requerir una revisión minuciosa por parte de los profesionales del derecho.

Chalkidis y su equipo implementaron una serie de modelos de PLN, incluyendo técnicas de tokenización, etiquetado de partes del discurso (POS tagging) y análisis sintáctico, combinadas con algoritmos de aprendizaje profundo como redes neuronales recurrentes (RNN) y redes neuronales convolucionales (CNN). Estos modelos fueron entrenados con un gran conjunto de datos de contratos previamente anotados para aprender a identificar y clasificar diferentes tipos de cláusulas contractuales, tales como cláusulas de indemnización, términos de pago y condiciones de rescisión.

El estudio destacó varios aspectos clave para la efectividad de estas técnicas. Primero, la importancia de un preprocesamiento adecuado de los textos contractuales para normalizar el lenguaje y estructurar los datos de manera que los modelos de PLN puedan procesarlos de forma eficiente. Segundo, la implementación de algoritmos de clasificación supervisada que puedan

aprender de ejemplos etiquetados y generalizar el conocimiento adquirido a nuevos documentos no vistos anteriormente.

Además Chalkidis discute los desafíos técnicos asociados con la automatización de la extracción de cláusulas, incluyendo la variabilidad en el lenguaje legal, las diferentes formas en que las cláusulas pueden ser redactadas y la necesidad de grandes conjuntos de datos etiquetados para entrenar los modelos de aprendizaje profundo. A pesar de estos desafíos, los resultados de su investigación demostraron que los modelos de PLN y aprendizaje profundo pueden lograr altos niveles de precisión en la identificación de cláusulas contractuales, reduciendo significativamente el tiempo y esfuerzo necesarios para la revisión manual de contratos. [5]

En el ámbito nacional, se encontró una tesis de maestría que utilizó el Procesamiento de Lenguaje Natural (PNL). El estudio, titulado "Procesamiento del lenguaje natural y análisis de redes sociales para comprender las renegociaciones de APP en medio de la pandemia de Covid-19" [3], se centra en identificar los mecanismos contractuales implementados o afectados durante la pandemia y en determinar los impactos de la transformación digital en la gestión contractual de proyectos APP en Colombia. Con el análisis de contenido, el procesamiento de lenguaje natural y el análisis de redes sociales, esta investigación busca comprender los procesos de renegociación y establecer una metodología que facilite el análisis contractual y los procesos futuros de renegociación.
Estos estudios subrayan la viabilidad y efectividad de utilizar técnicas avanzadas de PNL y AA para automatizar tareas complejas en la gestión de contratos.

4. PREPARACIÓN DE DATOS

Este capítulo describe de manera detallada el proceso de preparación de datos, que comprendió tanto la recopilación y preprocesamiento del corpus documental como la construcción de un conjunto de datos etiquetados necesario para el entrenamiento supervisado de los modelos de clasificación. Dado que no se disponía de una base de datos que identificara de manera explícita los requisitos contractuales contenidos en cada minuta.

La elaboración del corpus se hizo en dos fases:

1. **Etiquetado manual:** Se seleccionó una muestra representativa de minutas y se anotaron de forma explícita los requisitos contractuales presentes en cada documento.
2. **Etiquetado automático supervisado:** Con los datos manualmente etiquetados se entrenaron modelos preliminares, que luego se aplicaron al resto del corpus para generar etiquetas de manera automática.

De este modo, la fase de preprocesamiento no solo contempló la limpieza y transformación de los textos, sino también la generación iterativa de etiquetas: primero a mano, y luego mediante los modelos entrenados, ampliando y refinando el conjunto de datos de referencia para las etapas de modelado y evaluación.

4.1 Alistamiento y Consolidación del Corpus

Para el desarrollo de las técnicas de recolección y preprocesamiento de documentos contractuales, se seleccionó una muestra de 1.001 minutas contractuales en formato PDF, obtenidas de la plataforma de datos abiertos del Gobierno de Colombia, la cual permite acceder a información contractual de manera pública. Los documentos abarcan los años 2020 a 2023, con el propósito de incluir contratos recientes que permitieran analizar tendencias antes, durante y después de la pandemia de COVID-19. Esta selección buscó reflejar potenciales cambios en los requisitos contractuales a lo largo de dicho período.

Durante esta etapa se identificó que no existía un repositorio unificado que consolidara todas las minutas. Por este motivo, fue necesario construir manualmente la base de datos descargando una a una cada minuta hasta completar la muestra final. Este proceso demandó un esfuerzo considerable en términos de tiempo y validación.

Una vez almacenadas en un repositorio en la nube, las minutas fueron procesadas de forma automatizada mediante herramientas de procesamiento en Python, utilizando las librerías PyPDF2 y PdfReadError para extraer su contenido textual. La información resultante se organizó en un DataFrame con las siguientes columnas:

- nombre_archivo: nombre del documento.
- contenido: texto extraído.
- tamaño_contenido: cantidad de caracteres del contenido.

En esta etapa se identificaron 63 minutas cuyo contenido resultó vacío (cero caracteres), por lo que fueron excluidas del análisis. Como resultado, la base consolidada quedó conformada por 938 minutas válidas, que constituyeron el conjunto de datos definitivo utilizado en los siguientes procedimientos de preprocesamiento y modelado.

4.2 Preprocesamiento del Texto

A partir del corpus extraído, se llevó a cabo un proceso de limpieza de los datos orientado a reducir el ruido lingüístico y facilitar la posterior aplicación de modelos de procesamiento de lenguaje natural (PLN).
El preprocesamiento consistió en:

- Conversión de todo el texto a minúsculas.
- Eliminación de caracteres especiales.
- Remoción de palabras vacías (stopwords) que no aportan información relevante.

- Steemming para reducir palabras a su raíz.
- Eliminación de espacios vacíos y normalización de los textos.

El resultado se almacenó en el archivo t_minutas_preprocesadas.csv, que contiene el corpus limpio, listo para las siguientes etapas (véase la Imagen 3).

	nombre_archivo	contenido	tamaño_contenido
0	CONTRATO 3042986.pdf	gabf220 versin 4 127 contrat n 3042986 contrat...	99042
1	00.ORDEN_DE_SERVICIOS_No._2562413-FIRMA FA.pdf	docusign envelop id df464f38 87d3 4395 beb2 17...	153714
2	3035552.doc.pdf	gabf220 versin 6 fech 21122018 142 contrat n 3...	148437
3	CONTRATO 3021581.pdf	contrat n 3021581 gabf220 versin 6 fech 211220...	171622
4	CONTRATO 3044364.pdf	contrat n 304436 4 gabf220 versin 6 fech 21122...	147118

Imagen 2.Muestra minutas preprocesadas

4.3 Etiquetado Manual de requisitos

Para construir un conjunto de entrenamiento supervisado, se realizó un etiquetado manual de una muestra de 100 minutas. La selección de estas minutas se realizó de manera aleatoria, tomando 25 documentos por cada año (2020, 2021, 2022 y 2023).

Posteriormente, a estas mismas minutas se les aplicó nuevamente el proceso de preprocesamiento previamente descrito, con el fin de asegurar la homogeneidad del texto antes del análisis. Esta versión final consolidada de los datos se almacenó en un archivo que sirvió como insumo principal para el modelado supervisado

	nombre_archivo	tamaño_contenido
0	CONTRATO 3042986.pdf	99042
1	00.ORDEN_DE_SERVICIOS_No._2562413-FIRMA FA.pdf	153714
2	3035552.doc.pdf	148437
3	CONTRATO 3021581.pdf	171622
4	CONTRATO 3044364.pdf	147118
..	...	...
95	CONTRATO 3044390.pdf	91940
96	CONTRATO No. 3044344.docx.pdf	131436
97	CONTRATO_3044156.pdf	141648
98	CONTRATO 3044094.doc.pdf	12416
99	ORDEN DE SERVICIO No. 3043802.doc.pdf	116885

[100 rows x 2 columns]

Imagen 3. Muestra de 100 minutas

Cada minuta fue revisada de forma manual, marcando con “sí” o “no” la presencia de los requisitos contractuales identificados. La Imagen 5 muestra un ejemplo de la matriz de etiquetado.

4	ítem	Contrato	Retención en garantía	Gastos reembolsables	Uso de opción	Reajuste salarial	Reajuste de tarifas y precios	Anticipo o pago anticipado	Garantías y seguros	Cl
5	1	CONTRATO 3042986	No	Si	No	Si	No	Si	Si	
6	2	00.ORDEN_DE_SERVICIOS_No._	No	No	No	No	No	No	Si	
7	3	3035552.doc	No	No	No	Si	Si	No	Si	
8	4	CONTRATO 3021581	Si	No	No	Si	Si	No	Si	
9	5	CONTRATO 3044364	Si	No	Si	Si	Si	Si	Si	
10	6	CONTRATO 3037546	No	No	Si	Si	No	Si	Si	
11	7	CONTRATO 3038340	No	No	No	No	No	No	Si	
12	8	CONTRATO 3043915	No	No	No	No	No	No	Si	
13	9	CONTRATO 3045120	No	No	No	No	No	No	No	
14	10	CONTRATO 3045301	Si	Si	No	Si	Si	No	No	
15	11	CONTRATO 3045305	Si	Si	No	Si	Si	No	Si	
16	12	CONTRATO 3045357	Si	Si	No	Si	Si	No	Si	
17	13	CONTRATO 3045574	Si	Si	Si	Si	Si	No	Si	
18	14	CONTRATO 3045833	No	No	Si	Si	Si	No	Si	
19	15	CONTRATO 3052847	No	No	No	No	No	No	No	

Imagen 4. Dataset etiquetado manual Excel

La información final, que integra contenido preprocesado y etiquetas, se consolidó en un solo dataset como se puede ver en la Imagen 6.

```
                          nombre_archivo  \
0                       CONTRATO  3042986
1  00.ORDEN_DE_SERVICIOS_No._2562413-FIRMA FA
2                              3035552.doc
3                       CONTRATO 3021581
4                       CONTRATO 3044364

                                          contenido  tamaño_contenido  \
0  gabf220 versin 4 127 contrat n 3042986 contrat...             99042
1  docusign envelop id df464f38 87d3 4395 beb2 17...            153714
2  gabf220 versin 6 fech 21122018 142 contrat n 3...            148437
3  contrat n 3021581 gabf220 versin 6 fech 211220...            171622
4  contrat n 304436 4 gabf220 versin 6 fech 21122...            147118

  Retención en garantía Gastos reembolsables Uso de opción Reajuste salarial  \
0                    No                   Si           No                Si
1                    No                   No           No                No
2                    No                   No           No                Si
3                    Si                   No           No                Si
4                    Si                   No           Si                Si

  Reajuste de tarifas y precios Anticipo o pago anticipado  \
0                            No                         Si
1                            No                         No
2                            Si                         No
3                            Si                         No
4                            Si                         Si

  Garantías y seguros Cláusula de Cesión Subcontratación GAB-F-213 GAB-F-214  \
0                  Si                 Si                Si       Si       Si
1                  Si                 Si                Si       Si       Si
2                  Si                 Si                Si       Si       Si
3                  Si                 Si                Si       Si       Si
4                  Si                 Si                Si       Si       Si
```

Imagen 5. Dataset contenido minutas con etiquetado

Los requisitos definidos inicialmente para el proceso de etiquetado fueron:

1. Retención en garantía
2. Gastos reembolsables
3. Uso de opción
4. Reajuste salarial
5. Reajuste de tarifas y precios
6. Anticipo o pago anticipado
7. Garantías y seguros
8. Cláusula de Cesión
9. Contribución de obra pública
10. Estampilla Universidad Nacional
11. Subcontratación
12. GAB-F-213
13. GAB-F-214
14. GAB-F-221
15. GAB-F-060
16. GAB-F-105
17. Reunión de inicio
18. Socialización

Durante la fase de etiquetado manual se identificó que los requisitos: Contribución de obra pública y Estampilla Universidad Nacional no estaban presentes en la muestra, por lo que fueron descartados del análisis fin

4.4 Balanceo de clases

Para cada requisito se generó un dataset específico que contiene su respectiva etiqueta binaria, lo que resultó en 16 datasets independientes. A las clases “sí” y “no” se les asignaron los valores 1 y 0, respectivamente.

Se calculó la cantidad de observaciones por clase, permitiendo identificar la distribución de cada variable, lo que permitió identificar cuál era la clase mayoritaria y minoritaria como se muestra en la siguiente tabla:

Tabla 1. Proporción de clases por requisito.

Variable	Presencia del requisito: Sí (1)	Ausencia del requisito: No (0)	Balance/desbalanceo
Retención en garantía	15	84	Desbalance
Gastos reembolsables	24	75	Desbalance
Uso de opción	24	75	Desbalance
Reajuste salarial	42	57	Balanceo
Reajuste de tarifas y precios	36	63	Desbalance
Anticipo o pago anticipado	5	94	Desbalance
Garantía y seguros	83	16	Desbalance
Cláusula de cesión	93	6	Desbalance
Subcontratación	71	28	Desbalance
GAB-F-213	55	44	Balanceo
GAB-F-214	53	46	Balanceo
GAB-F-221	50	49	Balanceo
GAB-F-060	3	96	Desbalance
GAB-F105	28	71	Desbalance
Reunión de inicio	45	54	Balanceo
Socialización	22	77	Desbalance

El desbalance de clases puede afectar el desempeño de los modelos de clasificación, en particular su capacidad de identificar correctamente la clase minoritaria. Para mitigar este problema, se aplicó una estrategia de undersampling de la clase mayoritaria, buscando construir un conjunto de datos más equilibrado para favorecer un aprendizaje representativo de cambas clases. Este procedimiento se llevó a cabo de la siguiente manera en cada uno de los dataset construidos por requisito:

- **Separación de datos por clase:** Se dividió el dataset original en dos subconjuntos: uno con las observaciones de la clase positiva (1) y otro con los de la clase negativa (0).
- **Definición de proporciones objetivo:** Se definieron dos posibles proporciones para balancear las clases 60/40 y 70/30 (mayoritaria/minoritaria). Representando así un compromiso entre balance y preservación.
- **Submuestreo condicional de la clase mayoritaria:** Por cada proporción objetivo, se calculó el número de muestras necesarias de la clase mayoritaria que permitiría alcanzar dicha proporción respecto a la clase minoritaria. Si la clase mayoritaria tenía suficientes muestras, se realizaba un submuestreo aleatorio con semilla fija para garantizar la reproducibilidad. Luego, las muestras balanceadas eran combinadas y reordenadas aleatoriamente.

- **Respaldo ante fallos:** Al intentar equilibrar clases mediante submuestreo, se definió proporciones objetivo (por ejemplo 60/40 y 70/30 entre mayoritaria/minoritaria). Para cada proporción, se calculó cuántas muestras de la clase mayoritaria se necesitaría eliminar para alcanzarla. Si la clase mayoritaria no tenía suficientes instancias para cumplir alguna de esas proporciones (por ejemplo, cuando hay muy pocas muestras en la clase minoritaria y, por tanto, no se podía extraer una muestra mayoritaria que respetara la ratio deseada), no se aplicó submuestreo y se conservó el conjunto original sin cambios. De esta manera, se evitó dejar al modelo con muy poca información de la clase mayoritaria y garantizando un conjunto de entrenamiento completo.
- **Verificación de la nueva distribución:** Finalmente, se imprimió la distribución de las clases resultantes, tanto en valores absolutos como proporcionales como se muestra en la Imagen 7, para confirma que el balanceo fue exitoso.

```
Distribución original:
Uso de opción
0    75
1    24
Name: count, dtype: int64

Distribución final de clases:
Uso de opción
0    0.59322
1    0.40678
Name: Proporción, dtype: float64

Distribución después del balanceo:
Uso de opción
0    35
1    24
Name: count, dtype: int64
```

Imagen 6. Resultados aplicando undersampling en el dataframe del requisito Uso de Opción

5. MODELADO

Este capítulo describe el ciclo completo de modelado supervisado, que se dividió en dos etapas principales. La primera consistió en el entrenamiento y validación preliminar de modelos SVC empleando un subconjunto reducido de minutas etiquetadas manualmente. Estos modelos se utilizaron para realizar un etiquetado automático supervisado sobre el corpus restante. Posteriormente, se consolidó el dataset completo y se procedió al entrenamiento final de los clasificadores, con el objetivo de optimizar su desempeño predictivo sobre un conjunto de datos más amplio y representativo.

5.1 Vectorización

Para convertir los textos contractuales en una forma numérica susceptible de ser procesada por algoritmos de aprendizaje automático, se generaron dos representaciones vectoriales del corpus completo: una basada en frecuencias ponderadas (TF-IDF) y otra basada en embeddings semánticos (Word2Vec). Este proceso constituye la base para todas las etapas posteriores de entrenamiento y predicción.

5.1.1. Representación TF-IDF

Previo al entrenamiento del modelo de clasificación fue necesario convertir los documentos en una representación numérica que pudiera ser interpretada por los algoritmos de aprendizaje automático. Para ello, se utilizó la técnica de vectorización TF-IDF (Term Frequency-Inverse Document Frequency) ampliamente empleada en tareas de minería de texto y procesamiento de lenguaje natural. [6] [7] Esta técnica de vectorización asigna un peso a cada término de un documento en función de su frecuencia en ese documento y su rareza en el conjunto completo de documentos. De esta forma, se penalizan los términos demasiado frecuentes (que suelen ser poco informativos) y se da mayor peso a aquellos que son más distintivos.

5.1.1. Representación Word2Vec

Con el fin de capturar relaciones semánticas más profundas entre las palabras y superar las limitaciones del TF-IDF, se aplicó la técnica de embeddings Word2Vec para capturar relaciones semánticas entre palabras. Cada documento se representó como el promedio de los vectores de sus palabras.

Para el entrenamiento del modelo Word2Vec, fue necesario transformar los documentos del corpus en secuencias de términos (tokens). Esta tokenización se realizó sobre los textos previamente normalizados y limpios descritos en el Capítulo 4, permitiendo que el algoritmo aprendiera representaciones semánticas de cada palabra en función de su contexto. Posteriormente, cada documento se representó como el promedio de los vectores de sus palabras, generando una matriz de características de dimensión fija. Como se muestra en el Imagen 8, el resultado fue una estructura de datos en la que cada elemento corresponde a una lista de palabras, en lugar de una cadena de texto.

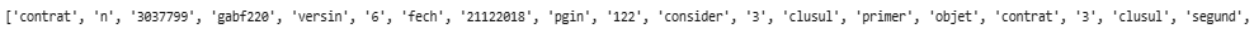


Imagen 7. Muestra del contenido de un documento vectorizado

5.2 Modelado preliminar con datos etiquetados manualmente

Como fase preliminar, se implementó un proceso de entrenamiento y evaluación sobre la muestra de minutas etiquetadas manualmente, con el propósito de comprobar la viabilidad de identificar de manera automática los requisitos contractuales mediante clasificadores SVC. Este procedimiento permitió, además, seleccionar los hiperparámetros iniciales más adecuados para su posterior aplicación en el proceso de etiquetado automático supervisado

5.2.1 División entrenamiento y prueba

Previo al entrenamiento de los modelos de clasificación, se realizó la partición del conjunto de datos etiquetados manualmente en dos subconjuntos: un 70% destinado al entrenamiento y un 30% reservado para la fase de prueba. Esta división se aplicó de manera global sobre el dataset,

manteniendo la proporción original de clases mediante un esquema de estratificación. Este enfoque garantizó que tanto el conjunto de entrenamiento como el de evaluación conservaran la distribución representativa de las etiquetas de interés.

Teniendo un dataset por cada requisito, se procedió a dividir el conjunto de datos en dos subconjuntos: uno para entrenamiento (70%) y otro para prueba (30%). aplicando estratificación para mantener la proporción original de clases como se muestra en la siguiente figura:

5.2.2 Entrenamiento de modelos SVC preliminares con TF-IDF y Word2Vec

Con el objetivo de construir clasificadores robustos capaces de identificar cláusulas contractuales asociadas a cada requisito, se implementaron modelos basados en Support Vector Classifier (SVC), empleando dos representaciones vectoriales diferentes: TF-IDF y Word2Vec. La elección de este modelo se fundamentó en un análisis teórico y una revisión de literatura especializada, que evidencian que el SVC es especialmente adecuado en contextos de clasificación binaria sobre datos textuales de alta dimensionalidad y naturaleza dispersa [8][9][10].

Entre sus principales ventajas destacan su capacidad para encontrar hiperplanos óptimos de separación, incluso en espacios no lineales mediante el uso de funciones kernel; su robustez frente al overfitting, especialmente en conjuntos moderadamente balanceados con un número limitado de muestras; y su desempeño consistente en problemas donde la separación entre clases no es evidente o donde existe ruido en los datos textuales. Por estas razones, se consideró que el SVC constituía la alternativa más adecuada para abordar la clasificación de minutas contractuales, en las que la presencia o ausencia de las cláusulas objeto de estudio puede expresarse con gran diversidad léxica y sintáctica.

Para cada requisito contractual se diseñó un flujo experimental sistemático compuesto por cuatro etapas de modelado, que combinaron las dos técnicas de vectorización (TF-IDF y Word2Vec) y la optimización de hiperparámetros mediante búsqueda en cuadrícula (GridSearchCV). En cada caso, el procedimiento se desarrolló de la siguiente forma:

- **Modelo SVC con vectorización TF-IDF y kernel lineal:**

Se entrenó un modelo inicial de SVC utilizando la representación TF-IDF de los documentos y un kernel lineal, con el fin de obtener una primera aproximación al rendimiento predictivo bajo una configuración sencilla y computacionalmente eficiente.

- **Modelo SVC con vectorización TF-IDF optimizado mediante búsqueda en cuadrícula:**

A partir de este primer modelo, se implementó un proceso de optimización de hiperparámetros mediante GridSearchCV. Se definió un espacio de búsqueda que incluyó distintos valores de regularización (C = 0.1, 1, 10), el tipo de kernel (lineal, radial basis function – RBF– y polinómico) y el grado del polinomio (grados 2, 3 y 4). Este procedimiento permitió identificar la combinación de hiperparámetros con mejor desempeño validado mediante validación cruzada de 5 pliegues (cv=5).

- **Modelo SVC con vectorización Word2Vec y kernel lineal:**

De manera análoga, se entrenó un modelo SVC empleando la representación semántica Word2Vec promedio de cada documento, utilizando inicialmente un kernel lineal como punto de partida.

- **Modelo SVC con vectorización Word2Vec optimizado mediante búsqueda en cuadrícula:**

Finalmente, se aplicó nuevamente GridSearchCV sobre la representación Word2Vec, explorando el mismo espacio de hiperparámetros que en el caso de TF-IDF. Esta etapa permitió evaluar la capacidad de la representación densa de Word2Vec para mejorar el desempeño predictivo tras la optimización.

De esta manera, para cada requisito contractual se construyeron cuatro modelos: un SVC con vectorización TF-IDF y configuración inicial, un SVC con vectorización TF-IDF optimizado, un SVC con vectorización Word2Vec y configuración inicial, y un SVC con vectorización Word2Vec

optimizado. Este enfoque sistemático facilitó la comparación rigurosa del rendimiento de cada combinación de técnicas de representación y configuración del clasificador, identificando la estrategia más adecuada para su posterior implementación en el sistema de clasificación automática de minutas contractuales.

5.2.3 Evaluación y selección preliminar

En esta etapa, se presentan los resultados obtenidos tras la ejecución sistemática de los cuatro experimentos de clasificación por cada requisito contractual. La imagen 10 resume de forma comparativa el desempeño de los modelos entrenados, considerando tanto el uso de diferentes técnicas de vectorización (TF-IDF y Word2Vec) como el impacto de la optimización de hiperparámetros mediante búsqueda en cuadrícula.

Las métricas reportadas incluyen las métricas de precisión, exhaustividad (recall) y F1-score (en su versión macro y ponderada), como el rendimiento general en términos de exactitud. Estas visualizaciones permiten identificar con claridad el impacto de las diferentes técnicas de vectorización y de la optimización de hiperparámetros sobre la capacidad predictiva del clasificador.

En particular, se destacan las diferencias de comportamiento entre los modelos basados en TF-IDF y aquellos que emplean representaciones Word2Vec, así como la mejora obtenida tras la aplicación de la búsqueda en cuadrícula respecto a los modelos iniciales con configuración lineal. Esta evidencia comparativa constituye un insumo clave para la selección fundamentada de los modelos finales que serán integrados en el sistema de identificación automática de requisitos contractuales.

Tabla 2. Resultados de entrenamiento por cada etiqueta

Etiqueta	**Vectorización**	**Modelo**	**Accuracy**	**Precision Macro**	**Recall Macro**	**F1-Score Macro**	**Precision Weighted**	**Recall Weighted**	**F1-Score Weighted**
Retención en Garantía	TF-IDF	SVC	0,625	0,313	0,500	0,385	0,391	0,625	0,481
Retención en Garantía	TF-IDF	SVC Mejorado	0,750	0,733	0,733	0,733	0,750	0,750	0,750

Etiqueta	Vectorización	Modelo	Accuracy	Precision Macro	Recall Macro	F1-Score Macro	Precision Weighted	Recall Weighted	F1-Score Weighted
Retención en Garantía	Word2Vec	SVC	0,625	0,583	0,567	0,564	0,604	0,625	0,605
Retención en Garantía	Word2Vec	SVC Mejorado	0,625	0,583	0,567	0,564	0,604	0,625	0,605
Gastos Reembolsables	TF-IDF	SVC	0,583	0,292	0,500	0,368	0,340	0,583	0,430
Gastos Reembolsables	TF-IDF	SVC Mejorado	0,667	0,657	0,657	0,657	0,667	0,667	0,667
Gastos Reembolsables	Word2Vec	SVC	0,667	0,667	0,629	0,625	0,667	0,667	0,646
Gastos Reembolsables	Word2Vec	SVC Mejorado	0,667	0,667	0,629	0,625	0,667	0,667	0,646
Uso de Opción	TF-IDF	SVC	0,583	0,292	0,500	0,368	0,340	0,583	0,430
Uso de Opción	TF-IDF	SVC Mejorado	0,833	0,889	0,800	0,813	0,870	0,833	0,823
Uso de Opción	Word2Vec	SVC	0,583	0,292	0,500	0,368	0,340	0,583	0,430
Uso de Opción	Word2Vec	SVC Mejorado	0,583	0,292	0,500	0,368	0,340	0,583	0,430
Reajuste salarial	TF-IDF	SVC	0,800	0,810	0,771	0,780	0,805	0,800	0,793
Reajuste salarial	TF-IDF	SVC Mejorado	0,900	0,900	0,917	0,899	0,920	0,900	0,901
Reajuste salarial	Word2Vec	SVC	0,800	0,800	0,813	0,798	0,820	0,800	0,802
Reajuste salarial	Word2Vec	SVC Mejorado	0,850	0,843	0,854	0,847	0,857	0,850	0,851
Reajuste de tarifas y precios	TF-IDF	SVC	0,667	0,824	0,571	0,518	0,784	0,667	0,577
Reajuste de tarifas y precios	TF-IDF	SVC Mejorado	0,667	0,646	0,623	0,625	0,656	0,667	0,653
Reajuste de tarifas y precios	Word2Vec	SVC	0,722	0,708	0,695	0,699	0,718	0,722	0,718
Reajuste de tarifas y precios	Word2Vec	SVC Mejorado	0,833	0,825	0,838	0,829	0,842	0,833	0,835
Anticipo o pago anticipado	TF-IDF	SVC	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Anticipo o pago anticipado	TF-IDF	SVC Mejorado	0,333	0,167	0,500	0,250	0,111	0,333	0,167
Anticipo o pago anticipado	Word2Vec	SVC	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Anticipo o pago anticipado	Word2Vec	SVC Mejorado	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Garantías y seguros	TF-IDF	SVC	0,625	0,313	0,500	0,385	0,391	0,625	0,481
Garantías y seguros	TF-IDF	SVC Mejorado	0,750	0,800	0,800	0,750	0,850	0,750	0,750
Garantías y seguros	Word2Vec	SVC	0,875	0,917	0,833	0,855	0,896	0,875	0,868
Garantías y seguros	Word2Vec	SVC Mejorado	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Etiqueta	Vectorización	Modelo	Accuracy	Precision Macro	Recall Macro	F1-Score Macro	Precision Weighted	Recall Weighted	F1-Score Weighted
Cláusula de Cesión	TF-IDF	SVC	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Cláusula de Cesión	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Cláusula de Cesión	Word2Vec	SVC	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Cláusula de Cesión	Word2Vec	SVC Mejorado	0,667	0,333	0,500	0,400	0,444	0,667	0,533
Subcontratación	TF-IDF	SVC	0,857	0,900	0,833	0,844	0,886	0,857	0,851
Subcontratación	TF-IDF	SVC Mejorado	0,929	0,944	0,917	0,925	0,937	0,929	0,927
Subcontratación	Word2Vec	SVC	0,857	0,900	0,833	0,844	0,886	0,857	0,851
Subcontratación	Word2Vec	SVC Mejorado	0,857	0,900	0,833	0,844	0,886	0,857	0,851
GAB-F-213	TF-IDF	SVC	0,750	0,844	0,722	0,715	0,828	0,750	0,725
GAB-F-213	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GAB-F-213	Word2Vec	SVC	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GAB-F-213	Word2Vec	SVC Mejorado	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GAB-F-214	TF-IDF	SVC	0,750	0,844	0,722	0,715	0,828	0,750	0,725
GAB-F-214	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GAB-F-214	Word2Vec	SVC	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GAB-F-214	Word2Vec	SVC Mejorado	0,950	0,958	0,944	0,949	0,954	0,950	0,950
GAB-F-221	TF-IDF	SVC	0,600	0,778	0,600	0,524	0,778	0,600	0,524
GAB-F-221	TF-IDF	SVC Mejorado	0,950	0,955	0,950	0,950	0,955	0,950	0,950
GAB-F-221	Word2Vec	SVC	0,950	0,955	0,950	0,950	0,955	0,950	0,950
GAB-F-221	Word2Vec	SVC Mejorado	0,950	0,955	0,950	0,950	0,955	0,950	0,950
GAB-F-060	TF-IDF	SVC	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GAB-F-060	TF-IDF	SVC Mejorado	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GAB-F-060	Word2Vec	SVC	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GAB-F-060	Word2Vec	SVC Mejorado	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GAB-F-105	TF-IDF	SVC	0,571	0,286	0,500	0,364	0,327	0,571	0,416
GAB-F-105	TF-IDF	SVC Mejorado	0,714	0,708	0,708	0,708	0,714	0,714	0,714
GAB-F-105	Word2Vec	SVC	0,786	0,864	0,750	0,754	0,844	0,786	0,767
GAB-F-105	Word2Vec	SVC Mejorado	0,857	0,854	0,854	0,854	0,857	0,857	0,857
Socialización	TF-IDF	SVC	0,636	0,318	0,500	0,389	0,405	0,636	0,495
Socialización	TF-IDF	SVC Mejorado	0,818	0,804	0,804	0,804	0,818	0,818	0,818

Etiqueta	Vectorización	Modelo	Accuracy	Precision Macro	Recall Macro	F1-Score Macro	Precision Weighted	Recall Weighted	F1-Score Weighted
Socialización	Word2Vec	SVC	0,727	0,708	0,679	0,686	0,720	0,727	0,717
Socialización	Word2Vec	SVC Mejorado	0,818	0,804	0,804	0,804	0,818	0,818	0,818
Reunión de Inicio	TF-IDF	SVC	0,545	0,273	0,500	0,353	0,298	0,545	0,385
Reunión de Inicio	TF-IDF	SVC Mejorado	0,818	0,817	0,817	0,817	0,818	0,818	0,818
Reunión de Inicio	Word2Vec	SVC	0,545	0,528	0,517	0,476	0,530	0,545	0,494
Reunión de Inicio	Word2Vec	SVC Mejorado	0,636	0,646	0,617	0,607	0,644	0,636	0,617

Para determinar el mejor modelo por cada etiqueta, se construyó una tabla consolidada con los resultados de las evaluaciones y se aplicó un criterio de selección basado en la métrica F1-Score Macro, este es espacialmente adecuado en escenarios de clases con desbalanceo, ya que calcula la media aritmética del F1-Score para cada clase sin ponderar por frecuencia, asegurando así una evaluación equilibrada del rendimiento del modelo en todas las clases. [8] [9] [10]

Tabla 3. Resultado de las evaluaciones aplicando F1-Score Macro

Etiqueta	Vectorización	Modelo	Accuracy	Precision Macro	Recall Macro	F1-Score Macro
Anticipo o pago anticipado	TF-IDF	SVC	0,667	0,333	0,500	0,400
Cláusula de Cesión	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000
GAB-F-060	TF-IDF	SVC	0,000	0,000	0,000	0,000
GAB-F-105	Word2Vec	SVC Mejorado	0,857	0,854	0,854	0,854
GAB-F-213	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000
GAB-F-214	TF-IDF	SVC Mejorado	1,000	1,000	1,000	1,000
GAB-F-221	TF-IDF	SVC Mejorado	0,950	0,955	0,950	0,950
Garantías y seguros	Word2Vec	SVC Mejorado	1,000	1,000	1,000	1,000
Gastos Reembolsables	TF-IDF	SVC Mejorado	0,667	0,657	0,657	0,657
Reajuste de tarifas y precios	Word2Vec	SVC Mejorado	0,833	0,825	0,838	0,829
Reajuste salarial	TF-IDF	SVC Mejorado	0,900	0,900	0,917	0,899
Retención en Garantía	TF-IDF	SVC Mejorado	0,750	0,733	0,733	0,733
Reunión de Inicio	TF-IDF	SVC Mejorado	0,818	0,817	0,817	0,817
Socialización	TF-IDF	SVC Mejorado	0,818	0,804	0,804	0,804
Subcontratación	TF-IDF	SVC Mejorado	0,929	0,944	0,917	0,925
Uso de Opción	TF-IDF	SVC Mejorado	0,833	0,889	0,800	0,813

Este enfoque facilitó la comparación, selección final de modelos óptimos y los requisitos finales que serán identificados para el sistema de monitoreo automático de cláusulas contractuales, priorizando no solo el rendimiento global sino también la equidad entre clases. Los requisitos que se descartaron son GAB-F-060 y "Anticipo o pago por anticipado" quedando en total 14 requisitos.

Este procedimiento permitió realizar una comparación rigurosa y fundamentada de los distintos modelos evaluados, garantizando criterios homogéneos en la selección de la alternativa óptima para cada requisito contractual. La priorización se centró tanto en el desempeño global como en la capacidad de mantener un equilibrio en la clasificación de ambas clases, aspecto crítico en contextos de desbalance. Como resultado de este análisis, se determinó descartar los requisitos *GAB-F-060* y *Anticipo o pago anticipado*, dado que no alcanzaron un nivel de desempeño satisfactorio que justificara su inclusión. De este modo, se consolidó un conjunto final de 14 requisitos que fueron seleccionados para integrar el sistema de identificación automática de requisitos contractuales desarrollado en este proyecto.

5.3 Etiquetado automático Supervisado

En esta sección se describe el procedimiento de clasificación supervisada automática aplicado al conjunto de minutas no etiquetadas, con el fin de ampliar el volumen de datos etiquetados y disponer de un corpus que integre predicciones generadas por los modelos entrenados previamente.

5.3.1 Identificación y separación de minutas sin etiquetar

Se cargó el conjunto completo de minutas preprocesadas con el objetivo de identificar las minutas que no contaban con etiquetado. Para ello, se compararon los identificadores de archivo presentes en el corpus total con aquellos del dataset etiquetado manualmente. Esta verificación permitió aislar 839 minutas que permanecían sin etiqueta, constituyendo el conjunto sobre el cual se aplicó el procedimiento automático de clasificación.

5.3.2 Configuración de los modelos de predicción

Para cada requisito contractual se definió una configuración específica que incluía: el nombre del requisito, el modelo previamente entrenado y optimizado, el tipo de vectorización utilizado (TF-IDF o Word2Vec) y el objeto vectorizador correspondiente. Esta estructura permitió organizar de manera sistemática la aplicación de los clasificadores, considerando tanto los requisitos que emplean representaciones basadas en frecuencias como aquellos que utilizan embeddings semánticos.

5.3.3 Aplicación de modelos y generación de predicciones

Se implementó una función que transforma el texto según la técnica de vectorización especificada y aplica el modelo de clasificación para obtener predicciones binarias. En el caso de TF-IDF, los documentos fueron convertidos en matrices de características mediante el vectorizador entrenado. Para Word2Vec, cada documento fue tokenizado y representado como el promedio de los vectores correspondientes a las palabras incluidas en el vocabulario del modelo.

El proceso se ejecutó iterativamente sobre cada requisito, generando una predicción de presencia o ausencia del requisito por cada minuta. Las salidas se consolidaron en un único conjunto de datos que almacenó las etiquetas binarias asociadas a cada minuta. Finalmente, se elaboró un resumen cuantitativo de los resultados, que muestra la proporción de documentos clasificados positivamente por requisito.

nombre_archivo	contenido	Retención en garantía	Gastos reembolsables	Uso de opción	Reajuste salarial	Reajuste de tarifas y precios	Cláusula de Cesión	GAB-F-105	GAB-F-213	GAB-F-214	GAB-F-221	Garantías y seguros	Reunión de inicio	Socialización	Subcontratación
Orden_de_Servicios_No._2559725.firmafa	gabf220 versin 3 fech 07042017 14 orden servic...	0	1	0	0	1	0	0	0	1	1	0	0	0	0
Carta de Adscripción Dr. ALEXANDER GONZÁLEZ....	gabf259 versin 4 fech 21102020 12 derech reser...	0	0	0	0	0	1	0	0	0	0	0	0	1	0
Anexo 05 Minuta Contrato firmado	contrat anticipacin demand ande contrat n 3035...	0	1	1	1	1	0	0	1	1	1	1	0	1	1
CONTRATO No. 3035784.doc	contrat n 3035 784 gabf220 versin 6 fech 21122...	0	0	0	0	0	0	0	0	0	0	1	1	1	1
ORDEN DE SERVICIOS No. 2291563.doc	gabf220 versin 6 fech 21122018 15 orden servic...	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Imagen 8. Muestra etiquetado automático

Con el fin de cuantificar la proporción de minutas en los que se identificó o no la presencia de cada cláusula se hizo un resumen estadístico de las predicciones realizadas para cada tipo de requisito contractual.

	Etiqueta	No (%)	Sí (%)
0	Cláusula de Cesión	28.6	71.4
1	GAB-F-105	85.8	14.2
2	GAB-F-213	52.0	48.0
3	GAB-F-214	51.5	48.5
4	GAB-F-221	52.6	47.4
5	Garantías y seguros	38.7	61.3
6	Gastos reembolsables	79.1	20.9
7	Reajuste de tarifas y precios	62.5	37.5
8	Reajuste salarial	64.6	35.4
9	Retención en garantía	90.4	9.6
10	Reunión de inicio	70.2	29.8
11	Socialización	68.3	31.7
12	Subcontratación	39.9	60.1
13	Uso de opción	79.4	20.6

Imagen 9. Predicción porcentual de presencia por cada requisito

5.4 Modelado final sobre el corpus completo etiquetado

En esta etapa se desarrolló el modelado final utilizando el conjunto consolidado de minutas, que integró tanto las muestras etiquetadas manualmente como las etiquetada mediante clasificación automática supervisada. Este proceso tuvo como propósito construir los modelos definitivos de predicción para cada requisito contractual.

5.4.1 División del conjunto de datos

Inicialmente, se eliminaron las etiquetas correspondientes a requisitos descartados (GAB-F-060 y Anticipo o pago anticipado), conservando un total de 14 requisitos. El corpus fue dividido en dos subconjuntos de manera estratificada, manteniendo la distribución proporcional de clases: el 70% de los registros se utilizó para entrenamiento y el 30% para evaluación de desempeño.

5.4.2 Vectorización global

Para garantizar coherencia y consistencia en la representación de los documentos, se aplicaron dos técnicas de vectorización de forma global sobre el corpus de entrenamiento:

- **TF-IDF:** Se entrenó un modelo TfidfVectorizer sobre el conjunto completo de entrenamiento, transformando cada documento en una matriz dispersa de pesos de términos.

- **Word2Vec:** Cada documento fue tokenizado y representado como el promedio de los embeddings de sus palabras. El modelo Word2Vec se entrenó sobre los tokens de todos los documentos de entrenamiento, generando vectores de dimensión 100.

5.4.3 Balanceo de clases

Para abordar el desbalance de clases presente en varias etiquetas, se definieron proporciones objetivo personalizadas (por ejemplo, 40% de clase positiva), aplicando submuestreo aleatorio de la clase mayoritaria. Este procedimiento se ejecutó de forma independiente para cada requisito. Cuando la proporción objetivo no podía alcanzarse debido al volumen reducido de la clase minoritaria, se conservó la distribución original.

5.4.4 Entrenamiento final con parámetros optimizados previamente

Se desarrollaron funciones específicas de entrenamiento y evaluación para cada técnica de vectorización:

- **Entrenamiento con TF-IDF**: La función entrenar_y_evaluar_tfidf_personalizado recibió como parámetros la etiqueta, la matriz de características, la proporción deseada de la clase positiva, y los hiperparámetros óptimos previamente seleccionados durante la etapa de validación con datos etiquetados manualmente.

- **Entrenamiento conWord2Vec:** La función entrenar_y_evaluar_word2vec_personalizado aplicó la misma lógica sobre las representaciones densas generadas con Word2Vec, utilizando también la configuración de hiperparámetros determinada con anterioridad.

Tabla 4. Modelo por cada Requisito

Requisito	Modelo
Retención en garantía	TF-IDF SVC Mejorado
Gastos Reembolsables	TF-IDF SVC Mejorado
Uso de opción	TF-IDF SVC Mejorado
Reajuste salarial	TF-IDF SVC Mejorado
Reajuste de tarifas y precios	Word2VEC SVC Mejorado
Garantías y seguros	Word2VEC SVC Mejorado
GAB-F-105	Word2VEC SVC Mejorado
GAB-F-213	TF-IDF SVC Mejorado
GAB-F-214	TF-IDF SVC Mejorado
GAB-F-221	TF-IDF SVC Mejorado
Reunión de inicio	TF-IDF SVC Mejorado
Cláusula de Cesión	TF-IDF SVC Mejorado
Socialización	TF-IDF SVC Mejorado
Subcontratación	TF-IDF SVC Mejorado

En esta fase no se realizó una búsqueda adicional de hiperparámetros, dado que cada requisito contaba con una configuración óptima previamente identificada, tanto en lo relativo al modelo (SVC) como a la técnica de vectorización (TF-IDF o Word2Vec). Esta decisión tuvo como propósito asegurar consistencia y reproducibilidad respecto al desempeño observado durante la etapa de entrenamiento preliminar.

Cada función realizó de manera sistemática:

- Submuestreo condicional de la clase mayoritaria en el conjunto de entrenamiento.
- Entrenamiento del modelo SVC con los hiperparámetros específicos ya optimizados.
- Predicción sobre el conjunto de prueba.
- Cálculo de métricas de desempeño (accuracy, precision, recall, F1-score macro y ponderado).
- Almacenamiento de los resultados en un DataFrame centralizado y serialización del modelo final en formato .pkl.

Para cada requisito contractual se entrenó únicamente el modelo correspondiente a la técnica de vectorización que mostró mejor desempeño en la fase previa. Así, no se generaron dos modelos alternativos por etiqueta, sino que se aplicó la configuración más favorable identificada en el proceso de validación inicial.

Una vez seleccionada la configuración más adecuada por cada requisito contractual, se procedió a evaluar de manera sistemática su desempeño sobre el conjunto de prueba independiente. Para este propósito, se calcularon métricas de clasificación, incluyendo exactitud (accuracy), precisión, recall y F1-score, reportadas en su versión macro y ponderada (weighted). Estas métricas permiten valorar tanto el rendimiento global de cada clasificador como su capacidad de identificar correctamente la clase minoritaria, aspecto especialmente relevante dado el desbalance observado en algunas etiquetas. A continuación, se presentan los resultados obtenidos para cada modelo entrenado.

Tabla 5. Métricas de desempeño Final de los Modelos

Etiqueta	Vectorización	Modelo	Accuracy	Precision Macro	Recall Macro	F1-Score Macro	Precision Weighted	Recall Weighted	F1-Score Weighted
Retención en garantía	TF-IDF	SVC Mejorado	0,913	0,768	0,936	0,821	0,949	0,913	0,923
Gastos reembolsables	TF-IDF	SVC Mejorado	0,899	0,821	0,868	0,841	0,909	0,899	0,902
Uso de opción	TF-IDF	SVC Mejorado	0,928	0,861	0,955	0,896	0,948	0,928	0,932
Reajuste salarial	TF-IDF	SVC Mejorado	0,964	0,953	0,970	0,960	0,966	0,964	0,964
Reajuste de tarifas y precios	Word2Vec	SVC Mejorado	0,957	0,955	0,951	0,953	0,956	0,957	0,956
Garantías y seguros	Word2Vec	SVC Mejorado	0,946	0,939	0,953	0,944	0,951	0,946	0,946
GAB-F-105	Word2Vec	SVC Mejorado	0,888	0,791	0,826	0,806	0,896	0,888	0,891
GAB-F-213	TF-IDF	SVC Mejorado	0,975	0,978	0,972	0,974	0,976	0,975	0,975
GAB-F-214	TF-IDF	SVC Mejorado	0,975	0,976	0,973	0,974	0,975	0,975	0,975
GAB-F-221	TF-IDF	SVC Mejorado	0,978	0,979	0,977	0,978	0,978	0,978	0,978
Reunión de inicio	TF-IDF	SVC Mejorado	0,909	0,893	0,900	0,897	0,910	0,909	0,910
Cláusula de Cesión	TF-IDF	SVC Mejorado	0,917	0,875	0,914	0,891	0,924	0,917	0,919
Socialización	TF-IDF	SVC Mejorado	0,917	0,893	0,908	0,900	0,919	0,917	0,917
Subcontratación	TF-IDF	SVC Mejorado	0,978	0,980	0,976	0,978	0,979	0,978	0,978

5.4.5. Análisis De Los Resultados

Los resultados obtenidos con los modelos finales entrenados sobre el corpus completo permitieron evaluar de manera detallada la capacidad predictiva del sistema para cada uno de los 14 requisitos contractuales identificados como relevantes.

Para cada uno de los 14 requisitos contractuales, se calcularon las siguientes métricas en el conjunto de prueba:

Precisión (Precision): proporción de verdaderos positivos sobre todas las predicciones positivas.

Exhaustividad (Recall): proporción de verdaderos positivos sobre todas las muestras positivas reales.

F1-Score ponderado: media armónica de precisión y exhaustividad, ponderada según la proporción de cada clase en el conjunto de prueba.

En términos generales, los clasificadores presentaron un desempeño robusto, con valores de F1-Score ponderado comprendidos entre el 89 % y el 97,8 %. Las etiquetas que alcanzaron los mejores niveles de desempeño fueron Subcontratación y GAB-F-221, ambas con F1-Score ponderados de 97,8 %, mientras que el requisito con menor rendimiento correspondió a GAB-F-105, que obtuvo un F1-Score ponderado de 89,1 %. Estas diferencias pueden atribuirse principalmente a la menor cantidad de ejemplos positivos en el conjunto de entrenamiento de algunos requisitos y a la elevada variabilidad semántica y léxica utilizada en la redacción de ciertas cláusulas.

Respecto a las técnicas de representación de texto, la mayor parte de los requisitos alcanzaron mejores métricas con la vectorización TF-IDF. En concreto, once de las catorce etiquetas fueron modeladas con esta técnica, mientras que Word2Vec se empleó en los casos de Reajuste de tarifas y precios, Garantías y seguros y GAB-F-105, donde demostró un rendimiento superior durante la fase de validación. Este hallazgo confirma que, aunque Word2Vec aporta un nivel adicional de representación semántica, la técnica TF-IDF continúa siendo especialmente eficaz en entornos donde la presencia de términos específicos resulta determinante para la identificación de los requisitos.

Por otra parte, se evidenció una mejora sustancial en los indicadores de desempeño frente al entrenamiento preliminar realizado únicamente con los documentos etiquetados de manera manual. En esta etapa inicial, el promedio del F1-Score ponderado se situó aproximadamente en 82 %. La incorporación del proceso de etiquetado automático supervisado sobre el corpus adicional permitió elevar dicho promedio hasta el 93 %, lo que representa un incremento de alrededor de 11 puntos porcentuales en la capacidad de generalización de los modelos.

Finalmente, la elección del algoritmo Support Vector Classifier (SVC) como núcleo de los clasificadores se respaldó en la solidez metodológica ampliamente documentada y validada por la

literatura científica en el campo del procesamiento de lenguaje natural y el aprendizaje automático, que resalta su idoneidad para tareas de clasificación binaria en contextos de alta dimensionalidad y vocabulario disperso, característicos de los documentos contractuales. Esta consideración técnica justificó su adopción como aproximación principal en el presente estudio.

6. IMPLEMENTACIÓN DE UN PROTOTIPO DE IDENTIFICACIÓN AUTOMÁTICA DE REQUISITOS CONTRACTUALES

Este capítulo presenta la construcción e integración de un aplicativo interactivo orientado a la identificación automática de requisitos contractuales en minutas en formato PDF. El sistema fue desarrollado con el fin de ofrecer una herramienta práctica que permita aplicar los modelos de clasificación entrenados, facilitando su uso por parte de profesionales y organizaciones interesadas en el monitoreo y análisis de cláusulas contractuales.

El prototipo cuenta con las siguientes funcionalidades principales:

- Carga de minutas en formato PDF.
- Extracción automática de texto mediante PyPDF2.
- Procesamiento lingüístico que incluye normalización, limpieza y, en su caso, tokenización.
- Vectorización específica por requisito: para cada requisito contractual, el sistema emplea la técnica de representación de texto (TF-IDF o Word2Vec) que fue determinada como la más adecuada durante el proceso de entrenamiento del modelo correspondiente. Así, cada modelo de clasificación aplica su propia estrategia de vectorización sobre el documento analizado, y en conjunto producen las predicciones finales.
- Predicción de la presencia o ausencia de cada requisito contractual mediante los modelos SVC almacenados en formato .pkl.
- Visualización consolidada de los resultados, presentando de forma clara qué requisitos se identifican en el documento analizado.

El flujo de trabajo se compone de seis etapas:

- **Etapa 1:** Carga del documento.
- **Etapa 2:** Extracción y preprocesamiento del contenido.
- **Etapa 3:** Transformación del texto en vectores mediante los modelos de representación entrenados.
- **Etapa 4:** Clasificación del documento respecto a cada requisito.
- **Etapa 5:** Presentación de resultados en pantalla.

- **Etapa 6:** Opcionalmente, exportación de predicciones para su registro.

La aplicación fue validada de manera local, verificando su consistencia con el desempeño observado en la etapa de modelado y su capacidad de procesar nuevos documentos no incluidos en el entrenamiento.

Imagen 10. Visualización del Prototipo de Sistema de Identificación de Requisitos Contractuales en Streamlit

7. DISCUSIÓN

En este trabajo se diseñó e implementó un sistema integral que combina modelos de PLN y AA entrenados sobre un corpus representativo de minutas contractuales. La arquitectura de la solución contempla un flujo completo que inicia con la carga de documentos en formato PDF, continúa con su preprocesamiento y vectorización, y culmina con la predicción de la presencia o ausencia de requisitos mediante modelos SVC optimizados. Este enfoque permite reducir de manera significativa el tiempo requerido para la revisión documental y, al mismo tiempo, incrementar la precisión y consistencia en el monitoreo de cláusulas contractuales, mitigando el riesgo de omisiones humanas. Adicionalmente, la herramienta fue diseñada con una estructura modular que facilita su extensión progresiva a otros tipos de contratos y requisitos, aunque en esta fase del proyecto se enfocó específicamente en un conjunto determinado de minutas y cláusulas de interés. Este enfoque permitió comprobar la viabilidad técnica de la solución en un dominio controlado, sentando las bases para una futura ampliación hacia otros dominios documentales.

La recolección de los documentos contractuales se realizó a partir de la plataforma de datos abiertos. Si bien en un escenario productivo lo más eficiente sería contar con repositorios unificados que consoliden todas las minutas de interés, en este proyecto fue necesario construir de manera manual una base de datos con los archivos a procesar. Este procedimiento evidenció la importancia de disponer de fuentes confiables y organizadas que permitan agilizar el proceso de ingestión documental previo a su análisis automatizado.

Los algoritmos seleccionados fueron los modelos Support Vector Classifier (SVC), combinados con dos técnicas de vectorización: TF-IDF y Word2Vec. Esta elección se fundamentó en su capacidad para procesar datos textuales de alta dimensionalidad y en su robustez frente al sobreajuste, especialmente en contextos con desbalance de clases.
La vectorización TF-IDF resultó eficaz para capturar la relevancia relativa de términos en cada documento.

La técnica Word2Vec permitió incorporar relaciones semánticas más profundas entre palabras. La combinación de estos enfoques con SVC demostró ser efectiva para la clasificación binaria de requisitos contractuales, alcanzando métricas de desempeño satisfactorias en términos de precisión y F1-Score.

La métrica principal utilizada fue el F1-Score Macro, dado que proporciona una evaluación equilibrada del rendimiento del modelo en contextos de clases desbalanceadas. Esta métrica calcula la media aritmética del F1-Score por clase sin ponderación por frecuencia, asegurando así una valoración justa tanto para la clase minoritaria como para la mayoritaria. Adicionalmente, se emplearon métricas complementarias, incluyendo Precisión (Precision), Exhaustividad (Recall) y Exactitud (Accuracy), calculadas en sus versiones macro y ponderada. Este conjunto de métricas permitió realizar una evaluación integral del comportamiento de los modelos, garantizando que su rendimiento fuera no solo estadísticamente adecuado, sino también relevante y aplicable en escenarios reales de gestión contractual.

8. CONCLUSIONES

El presente proyecto aplicado demostró la viabilidad de una solución automatizada para la identificación de requisitos contractuales en documentos legales, mediante la aplicación de técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN) y Aprendizaje Automático (AA).
Entre las principales conclusiones se destacan:

Automatización del análisis contractual: El prototipo desarrollado identificó 14 requisitos contractuales clave, obteniendo un F1-score ponderado entre 89 % y el 98 % para cada uno de ellos.

Estrategia de modelado robusta y adaptable: La combinación de diferentes técnicas de vectorización (TF-IDF y Word2Vec) junto con modelos de clasificación basados en Support Vector Classifier (SVC) permitió construir clasificadores robustos y con capacidad de generalización sobre nuevos documentos. La selección de la técnica de representación más adecuada para cada requisito optimizó el rendimiento predictivo global del sistema.

Importancia del preprocesamiento y del etiquetado inicial: El desempeño alcanzado por los modelos dependió, en gran medida, de la implementación de un preprocesamiento adecuado del texto contractual y de un proceso de etiquetado manual inicial de alta calidad. Este enfoque aseguró la disponibilidad de datos representativos y balanceados para el entrenamiento supervisado, incrementando la precisión y confiabilidad de las predicciones.

Evaluación con métricas adecuadas a escenarios desbalanceados: La utilización de la métrica F1-Score Macro resultó fundamental para evaluar el desempeño de los modelos en contextos con desequilibrios de clase. Este enfoque permitió una valoración equitativa entre la capacidad de detección de la presencia y ausencia de los requisitos contractuales, fortaleciendo la confiabilidad de los resultados obtenidos.

REFERENCIAS

[1] T. C. K. C. G. &. D. J. Mikolov, «Efficient Estimation of Word Representations in Vector Space,» arXiv preprint, 2013. [En línea]. Available: https://arxiv.org/abs/1301.3781.

[2] T. Mitchell, Machine Learning, New York, NY, USA: McGraw-Hill, 1997.

[3] IBM, «Aprendizaje Supervisado,» [En línea]. Available: www.ibm.com/ es-es/topics/supervised-learning .

[4] D. Ashley, «ARtifical Intelligence anda Legal Analytics: New Tools for Law Practice in the Digital,» 2017.

[5] I. A. a. M. M. I. Chalkidis, «, Extracting Contract Elements,» 2017.

[6] A. C. y. R. Patasik, «"Comparison of TF-IDF and Word2Vec for Twitter emotion classification using Support Vector Machine,",» *Procedia Computer Science* .

[7] D. Trușcă, «"Comparative analysis of word embeddings and document representations in text classification",» *Romanian Journal of Information Science and Technology*..

[8] J. T. e. al, «"Selecting and Interpreting Multiclass Classification Metrics in Remote Sensing with Class Imbalance",» *Remote Sensing*, vol. 16, nº 1, 2024.

[9] B. Chandra, «F1 Score: Balancing Precision and Recall in AI Evaluation,» *Medium*, 2023.

[10] A. Dutta, «"Micro-average vs Macro-average vs Weighted-average vs Samples Average",» Vitalflux.com, 2021. [En línea]. Available: https://vitalflux.com/micro-macro-weighted-average-ml/. [Último acceso: 2024].

[11] M.Collins, Head-Driven Statistical Models for Natural Language Parsing," Computational Linguistics, 2003.

[12] D. Rojas Marinkelle, «Procesamiento del lenguaje natural y análisis de redes sociales para comprender las renegociaciones de APP en medio de la pandemia de Covid-19,» Universidad de los Andes, 2023. [En línea]. Available: http://hdl.handle.net/1992/64669.

[13] IBM, «Aprendizaje no supervisado,» [En línea]. Available: www.ibm.com/ mx-es/topics/unsupervised-learning .

[14] A. S. y. A. F. S. S. R. A. Kurniawan, «“Comparison of Sentiment Analysis on Indonesian Social Media Using Support Vector Machine with TF-IDF and Word2Vec,”,» *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi).*


Pontificia Universidad
JAVERIANA
Cali