Universidad Central
Facultad de Ingeniería y Ciencias Básicas
Maestría en Analítica de Datos

Segmentación Automatizada de Incidentes en una Mesa de Servicio para un Fondo de Pensiones mediante Machine Learning, con el diseño de una interfaz interactiva para el usuario

Tesis de Maestría

Raúl Andrés Gamba Hastamorir
1.023.003.554
Diana Carolina Gómez Boada
1.030.583.926
Michael David Gualteros Garcia
1.023.980.438

Trabajo de grado como requisito para optar el título de:

Magíster en Analítica de Datos

Bogotá, Colombia

2025

Universidad Central
Facultad de Ingeniería y Ciencias Básicas
Maestría en Analitica de Datos

Segmentación Automatizada de Incidentes en una Mesa de Servicio para un Fondo de Pensiones mediante Machine Learning, con el diseño de una interfaz interactiva para el usuario

Tesis de Maestría

Raul Andres Gamba Hastamorir
Diana Carolina Gomez Boada
Michael David Gualteros Garcia

Director (a)
Luis Andrés Campos Maldonado

Bogotá, Colombia
2025

Aprobación

Director(a) de Tesis

_____

Firma
Luis Andrés Campos Maldonado
[Títulos]

Codirector(a) de Tesis

_____

Firma
[Nombres Completos]
[Títulos]

Jurados de la Tesis

Jurado

_____

Firma
[Nombres Completos]
[Títulos]

Jurado

_____

Firma
[Nombres Completos]
[Títulos]

Jurado

_____

Firma
[Nombres Completos]
[Títulos]

[Fecha de sustentación]

Resumen.

El sistema pensional en Colombia, inició a partir de 1945 y a la fecha ha presentado dos reformas pensiones que han tenido como objetivo garantizar la sostenibilidad y cobertura del sistema pensional, en virtud a la automatización de procesos y a la vanguardia de las tecnologías de la información, se abre oportunidades para optimizar proceso en este sector.

En este contexto, Colfondos enfrenta un desafío y este se da debido al alto volumen de solicitudes y quejas recibidas en la mesa de servicio, teniendo en cuenta que en 2024 alcanzó 12.741 casos, lo cual representa a nivel de gremio el 17,8%, Gran parte de las solicitudes hacen referencia a incidentes recurrentes que pueden resolverse de forma automatizada, lo cual permitiría disminuir la carga operativa, mejorar los tiempos de respuesta y de paso esto podría llegar a ofrecer un mejor servicio al cliente.

El problema identificado radica en la alta recurrencia de incidentes similares y el tiempo excesivo requerido para su resolución, debido a la dependencia de intervención humana en los procesos de clasificación y respuesta. Esto ha generado reprocesos, insatisfacción del usuario y una distribución ineficiente de los recursos. La solución propuesta se enfoca en reducir los tiempos de atención y elevar la calidad de las respuestas mediante un modelo de PLN que permita clasificar automáticamente los casos y generar respuestas automatizadas de primer nivel, facilitando la autogestión de los usuarios funcionales.

El objetivo general del proyecto es desarrollar una solución basada en PLN y técnicas de aprendizaje automático para segmentar, clasificar y responder de forma automatizada los incidentes reportados por los usuarios internos de Colfondos. Los objetivos específicos incluyen:

(i) estructurar los datos históricos de incidentes, (ii) representar semánticamente los textos mediante modelos de *embeddings*, (iii) aplicar técnicas de *clustering* para identificar patrones y categorías frecuentes, y (iv) desarrollar una interfaz que ofrezca respuestas automáticas a los casos de baja complejidad.

La metodología se basa en el modelo CRISP-DM, que guiará las fases de comprensión del negocio, análisis de los datos, preparación, modelado y evaluación. Se utilizarán técnicas como la vectorización semántica (*sentence embeddings con MiniLM*) y agrupamiento no supervisado (*K-means*, *DBSCAN*), complementadas con análisis de frecuencia y métricas de evaluación como el índice de silueta. La herramienta final será una interfaz interactiva que provea sugerencias automáticas a partir de incidentes previamente resueltos.

Entre los resultados esperados se encuentra la reducción del tiempo promedio de respuesta en al menos un 30%, la disminución de reprocesos, y una mejora sustancial en la satisfacción del usuario interno. Asimismo, se proyecta una redistribución más eficiente de la carga operativa, permitiendo al equipo de soporte enfocarse en incidentes de mayor complejidad.

En conclusión, este anteproyecto busca posicionar a Colfondos como una entidad más eficiente y centrada en el usuario, mediante la implementación de tecnologías inteligentes que optimicen sus procesos críticos. Además, representa un aporte tangible a la transformación digital del sector pensional en Colombia, alineado con las tendencias actuales de automatización y analítica avanzada.

Palabras clave: Procesamiento de lenguaje natural, mesa de servicio, clustering, embeddings, automatización, inteligencia artificial.

Abtract.

The pension system in Colombia began in 1945 and, to date, has undergone two pension reforms aimed at ensuring the sustainability and coverage of the system. In light of process automation and the advancement of information technologies, new opportunities have emerged to optimize processes in this sector.

In this context, Colfondos faces a significant challenge due to the high volume of requests and complaints received by its service desk. In 2024 alone, the organization registered **12,741 cases**, which represents **17.8%** of the total within the industry. A large portion of these requests relate to recurring incidents that could be resolved automatically, which would help reduce operational workload, improve response times, and ultimately enhance customer service.

The core problem lies in the high recurrence of similar incidents and the excessive time required for their resolution, caused by the heavy reliance on human intervention in the classification and response processes. This has resulted in rework, user dissatisfaction, and inefficient resource allocation. The proposed solution focuses on reducing response times and improving the quality of resolutions through a Natural Language Processing (NLP) model that can automatically classify cases and generate first-level automated responses, thus enabling self-management by functional users.

The main objective of the project is to develop a solution based on NLP and machine learning techniques to segment, classify, and automatically respond to incidents reported by internal users at Colfondos. The specific objectives include: (i) structuring historical incident data, (ii) semantically representing texts using embedding models, (iii) applying clustering techniques to

identify common patterns and categories, and (iv) developing an interface that provides automated responses to low-complexity cases.

The methodology is based on the CRISP-DM model, which will guide the phases of business understanding, data analysis, data preparation, modeling, and evaluation. Techniques such as semantic vectorization (sentence embeddings using MiniLM) and unsupervised clustering (K-means, DBSCAN) will be used, along with frequency analysis and evaluation metrics such as the silhouette score. The final tool will be an interactive interface that provides automatic suggestions based on previously resolved incidents.

Expected outcomes include a **reduction of average response time by at least 30%**, decreased rework, and a significant improvement in internal user satisfaction. Additionally, a more efficient distribution of the operational workload is projected, allowing the support team to focus on more complex and strategic incidents.

In conclusion, this pre-project aims to position Colfondos as a more efficient and user-focused organization through the implementation of intelligent technologies that optimize its critical processes. It also represents a tangible contribution to the digital transformation of Colombia's pension sector, in line with current trends in automation and advanced analytics.

Keywords: natural language processing, service desk, clustering, embeddings, automation, artificial intelligence.

Tabla de contenido

Contenido

Lista de Ilustraciones

Lista de Tablas.

Lista de Siglas y Abreviaturas.

Sigla	Ingles	Español
RMSE	Root Mean Squared Error	Raíz del Error Cuadrático Medio
MSE	Mean Squared Error	Error Cuadrático Medio
MAE	Mean Absolute Error	Error absoluto promedio
CRISP-DM	Cross-Industry Standard Process for Data Mining	Proceso Estándar Inter-Industrias para Minería de Datos
XGBoost	Extreme Gradient Boosting	Potenciación de Gradiente Extrema
LightGBM	Light Gradient Boosting Machine.	Máquina de Impulso de Gradiente Ligero.
KPI	Key Performance Indicator.	Indicador Clave de Desempeño
RF	Random Forest	Bosque Aleatorio
Logs	Logs	Registro de eventos que ocurren dentro de una aplicación de software.
KDE	Kernel Density Estimation	Estimación de Densidad con Núcleo
EDA	Exploratory Data Analysis	Análisis exploratorio de los datos

IQR	Interquartile Range	Rango Incuartilico Q3y Q1

Lista de Anexos

1. Introducción

El sistema de pensiones en Colombia ha atravesado diversas transformaciones desde su origen en 1945 con la promulgación de la Ley 1600, que dio lugar a la Caja Nacional de Jubilaciones y Pensiones de los Empleados Públicos, encargada de garantizar una pensión a los trabajadores del Estado (Zúñiga, s.f.). Con la implementación de la Ley 100 en 1993, se reformó el sistema de seguridad social, estableciendo el Sistema General de Pensiones y creando los regímenes de Prima Media (RPM) y Ahorro Individual con Solidaridad (RAIS). Posteriormente, la Ley 797 de 2003 introdujo ajustes enfocados en la sostenibilidad financiera del RPM. Estas reformas responden a la necesidad de brindar protección económica en la vejez, la invalidez o la muerte, bajo criterios de equidad, eficiencia y solidaridad.

En este contexto, el sistema pensional colombiano no solo se enfrenta desafíos de sostenibilidad financiera y cobertura, sino que también tiene como reto el poder adaptarse a la evolución tecnológica que actualmente se está evidenciando en el país, El uso de tecnologías emergentes como la inteligencia artificial (IA), el aprendizaje automático (ML) y el procesamiento de lenguaje natural (PLN) representa una oportunidad clave para modernizar los procesos internos y mejorar la experiencia de los afiliados. En un entorno cada vez más digital, los fondos de pensiones deben incorporar herramientas que permitan extraer valor de grandes volúmenes de datos, mejorar la toma de decisiones y automatizar respuestas a solicitudes recurrentes.

En un mundo donde las actualizaciones tecnológicas son cada vez más intensas y el avance de las inteligencias artificiales es cada vez más impactante, en Colombia se han comenzado a implementar estas herramientas con el propósito de optimizar procesos y permitir el crecimiento de las empresas colombianas. El país ha venido creciendo exponencialmente con el desarrollo y

aplicación de estas tecnologías. Algunas entidades educativas como, la Universidad Central, Universidad de los Andes, Universidad Jorge Tadeo Lozano y Universidad El Bosque, entre otras, optaron por ofrecer carreras que permiten el aprendizaje en este medio, como es la Maestría en Análisis de Datos, que permite generar profesionales más competentes para brindar un uso adecuado de estas herramientas y hacer uso de lo aprendido en diferentes campos empresariales del país. Esto permite a las empresas que se encuentran interesadas en esta tecnología desarrollar capacidades para contar con información a tiempo y mejorar sus decisiones.

En virtud a la automatización de procesos y la mejora en los tiempos de respuesta, muchas empresas han mostrado interés en dar uso a estas tecnologías. Una de ellas es Colfondos, cuyo propósito es brindar respuestas más rápidas a sus clientes, ya que sus operaciones diarias generan diferentes solicitudes que esperan ser atendidas y resueltas en el menor tiempo posible. Para ello, se propone implementar un asistente virtual que facilite a los operadores de Colfondos brindar las respuestas de manera más ágil.

Este objetivo del proyecto se centrará en Desarrollar modelos de Machine Learning para la segmentación de casos en la mesa de servicio interna de Colfondos, con el fin de automatizar y a través de una interfaz interactiva brindar respuestas de primer nivel a casos recurrentes para reducir la carga operativa del equipo de soporte, este proceso contempla la exploración profunda de las incidencias registradas, así como la segmentación automatizada de los casos a través de algoritmos de agrupamiento, lo que posibilita detectar grupos de solicitudes similares con características compartidas. Esta segmentación no solo mejora la comprensión del comportamiento de las incidencias, sino que también sirve como base para construir mecanismos de respuesta automáticos y contextualizados.

El alcance de la investigación abarca el análisis de datos históricos de incidentes del fondo de pensiones Colfondos, partiendo de casos del año 2024 hasta febrero 2025, la evaluación y elección del modelo de PLN que arroje mejor resultado y la validación de su desempeño mediante métricas de clasificación y la propuesta de una arquitectura de integración.

Este estudio aporta de manera significativa al campo de la analítica aplicada al servicio al cliente en el sector pensional, demostrando cómo la inteligencia artificial puede ser una aliada estratégica para optimizar procesos reiterativos, al mejorar la percepción del servicio al cliente. En particular, la aplicación de técnicas de *clustering* como **K-means** y **DBSCAN**, apoyadas en métricas de validación interna que permiten medir la calidad y coherencia de los grupos formados, refuerza la solidez de los resultados al identificar patrones de incidencias y casos recurrentes. Su aplicación tiene implicaciones prácticas no solo para Colfondos, sino también para otras entidades del sector pensional, financiero teniendo en cuenta que enfrentan problemáticas similares y buscan transformar sus procesos de atención mediante analítica avanzada.

2. Planteamiento Del Problema y Justificación

2.1 Contexto Organizacional

Colfondos S.A es una de las administradoras de pensiones y cesantías en Colombia, su misión está orientada en garantizar la administración, sostenibilidad de los recursos de sus afiliados, lo cual contribuye al aseguramiento de un bienestar para su vejez, invalidez o su etapa cesante. En el marco de su gestión operacional, la mesa de servicio hace parte de un área estratégica para la gestión de incidencias internas y solicitudes técnicas, ya que actúa como primer punto de contacto para resolver requerimientos operativos de sus usuarios funcionales.

En la actualidad, la compañía, a pesar de contar con herramientas de gestión de casos en la mesa de servicio que cuentan con nuevas tecnologías, se evidencia que los indicadores de escalamiento hacia la Superintendencia no han mostrado mejoras significativas. Esta situación impacta de manera directa en la percepción de los afiliados, quienes se enfrentan a tiempos prolongados de resolución de sus solicitudes y mayores niveles de insatisfacción, lo que a su vez se traduce en un incremento de las quejas formales y en una afectación reputacional hacia la organización.

2.2 Problemática Identificada

Actualmente, Colfondos enfrenta un alto volumen de solicitudes que ingresan a través de su mesa de servicio, muchas de ellas asociadas a fallas en aplicaciones, incidentes técnicos recurrentes o requerimientos de información que no logran ser resueltos en un primer contacto. En 2024, la entidad recibió 12.741 quejas, lo que representó el 17.8% del total gremial, posicionándose como el segundo fondo con mayor número de quejas

escaladas. Este dato refleja no sólo la magnitud del problema, sino también la necesidad de implementar mecanismos que optimicen la gestión de incidencias.

La principal dificultad que enfrenta la organización radica en la sobrecarga operativa de la mesa de servicio, que debe atender un gran número de solicitudes reiterativas sin contar con un sistema de clasificación automatizado. Esto genera tiempos de respuesta más largos, reprocesos innecesarios y una disminución en la eficiencia del área encargada, afectando tanto la percepción de los usuarios internos como la satisfacción de los afiliados.

Las principales problemáticas identificadas fueron:

- No existe segmentación y priorización de casos.
- Ausencia de usos de nuevas tecnologías, que permitan identificar patrones y segmentar incidencias recurrentes.
- Limitada explotación de los datos históricos de la mesa de servicio.
- Falta de herramientas interactivas que permitan a los usuarios funcionales autogestionar solicitudes simples.

2.3 Impacto Financiero y Estratégico.

La ineficiencia en la gestión de casos de la mesa de servicio también presenta un impacto financiero y estratégico para Colfondos. Al revisar detalladamente el impacto financiero, se puede ver que el exceso de tiempo invertido en la atención de incidencias incrementa los costos en horas-hombre, mientras que el uso inadecuado de la mesa de servicio destina recursos a solicitudes de bajo impacto, reduciendo la capacidad de respuesta ante incidentes críticos. A esto se suma el almacenamiento de información sin segmentación, lo cual genera gastos en infraestructura

tecnológica y, al mismo tiempo, limita la posibilidad de aprovechar los datos como un activo estratégico. En el plano organizacional, estas deficiencias afectan la reputación institucional al aumentar la percepción de insatisfacción de los afiliados, lo que conlleva un mayor número de quejas y un deterioro de la confianza hacia la compañía y esto puede dar resultado deserción de afiliados. Además, la falta de mecanismos analíticos adecuados debilita la capacidad de Colfondos para anticiparse a tendencias, cumplir con los requerimientos regulatorios y consolidar ventajas competitivas en un entorno cada vez más digitalizado.

2.4 Propuesta de Solución.

De acuerdo con lo anterior se hace necesario diseñar e implementar soluciones tecnológicas que permitan mejorar la gestión de incidencias haciendo uso de técnicas de procesamiento de lenguaje natural (PLN) y de algoritmos de agrupamiento como K-means y DBSCAN, junto con métricas de validación de clústeres, lo cual permitirá facilitar la identificación de patrones en los casos reportados, habilitar la segmentación automática de solicitudes y apoyar la construcción de respuestas de primer nivel, junto con la propuesta de una interfaz interactiva la cual permitiría a los usuarios obtener soluciones inmediatas a casos recurrentes, mejorando así los tiempos de respuesta y liberando recursos para atender situaciones de mayor complejidad.

2.5 Pregunta de Investigación

¿Cómo se puede segmentar de manera efectiva los casos recurrentes en la mesa de servicio de Colfondos, con el fin de ofrecer una respuesta automatizada de primer nivel?

2.6 Justificación

La justificación de la elaboración de este proyecto radica inicialmente en su impacto organizacional. en la segmentación y atención de incidencias mediante técnicas de procesamiento de lenguaje natural (PLN) la cual permitirá segmentar adecuadamente los casos, reducir los tiempos de respuesta y optimizar los recursos de la mesa de servicio. Al implementar algoritmos de *clustering* como K-means y DBSCAN, validados con métricas como el índice de silueta y el coeficiente de Dunn, se podrán identificar patrones de incidencias recurrentes y diseñar respuestas de primer nivel que faciliten la autogestión por parte de los usuarios funcionales.

En segundo lugar, se justifica desde la parte financiera, esto teniendo en cuenta que la reducción en el tiempo de atención de casos puede llegar a disminuir los costos asociados a la operación y libera capacidad para atender incidentes.

De la misma forma se tiene impacto estratégico para Colfondos. La implementación de soluciones basadas en inteligencia artificial fortalece la reputación institucional al mejorar la percepción de los afiliados, garantiza un mejor cumplimiento frente a los entes de control y contribuye a la construcción de una ventaja competitiva en un sector cada vez más digitalizado. De igual manera, esta iniciativa se alinea con las tendencias globales de transformación digital y con los objetivos de sostenibilidad y eficiencia que las administradoras de pensiones deben perseguir en el mediano y largo plazo.

Académicamente el estudio se fundamenta en la aplicación de herramientas avanzadas de analítica de datos al sector pensional, un campo en el que la investigación es aún incipiente. Este proyecto demuestra cómo la combinación de minería de texto, aprendizaje automático y visualización interactiva puede transformar procesos críticos de servicio al cliente, aportando un modelo

replicable en otras entidades del sector financiero y asegurador que enfrentan problemáticas similares.

3. Objetivos.

3.1 Objetivo General

Desarrollar modelos de Machine Learning para la segmentación de casos en la mesa de servicio interna de Colfondos, con el fin de automatizar y a través de una interfaz interactiva brindar respuestas de primer nivel a casos recurrentes para reducir la carga operativa del equipo de soporte.

3.2 Objetivos Específicos

- Realizar la recolección, limpieza y análisis exploratorio de la base de datos de casos generados en la mesa de servicio, con el fin de identificar patrones y tendencias relevantes que permitan optimizar su gestión.

- Implementar técnicas de clustering para agrupar casos recurrentes según patrones comunes, asegurando una segmentación adecuada y evaluando su desempeño mediante métricas como el índice de silueta y el coeficiente de Dunn.

- Desplegar los resultados obtenidos a través de una interfaz interactiva, que facilite al usuario escribir un caso y le sugiera soluciones de primer nivel basadas en los grupos definidos

4. Antecedentes

4.1. Aplicaciones y enfoques actuales en la gestión inteligente de incidentes

El crecimiento acelerado de las Tecnologías de la Información ha transformado la manera en que las empresas gestionan la atención al cliente y la resolución de incidencias. La gestión de servicios ha evolucionado para garantizar la continuidad operativa y mejorar la eficiencia de los procesos de soporte (Zuev et al., 2018). En este contexto, la gestión de incidentes se ha convertido en una práctica clave, ya que permite registrar, clasificar y dar respuestas de primer nivel a distintos problemas.

Las mesas de servicio o service desks en la gestión de casos juegan un papel importante en el registro de incidencias, sin embargo, su dependencia histórica con la intervención manual en el procesamiento de tickets o casos de soporte, ha generado retos significativos como, demoras en la atención de errores, clasificación y asignación ineficiente de recursos. En este punto, la implementación de la inteligencia artificial y el aprendizaje automatizado pueden optimizar el proceso de atención a incidentes, siendo más exactos con la respuesta al usuario, disminuyendo tiempos de respuesta y minimizando reprocesos

Dado que el número de solicitudes en las empresas sigue aumentando debido a los esfuerzos de digitalización, los errores en la clasificación y asignación de casos pueden incrementar significativamente los costos operativos y los tiempos de resolución. Esto, a su vez, afecta negativamente la satisfacción del cliente y perjudica la experiencia del usuario final (Fuchs et al., 2022).

El avance de la inteligencia artificial (IA) y el procesamiento de lenguaje natural (PLN) ha permitido la implementación de soluciones más eficientes para la gestión de incidentes, y esta tendencia se refleja en el creciente número de publicaciones sobre el tema.

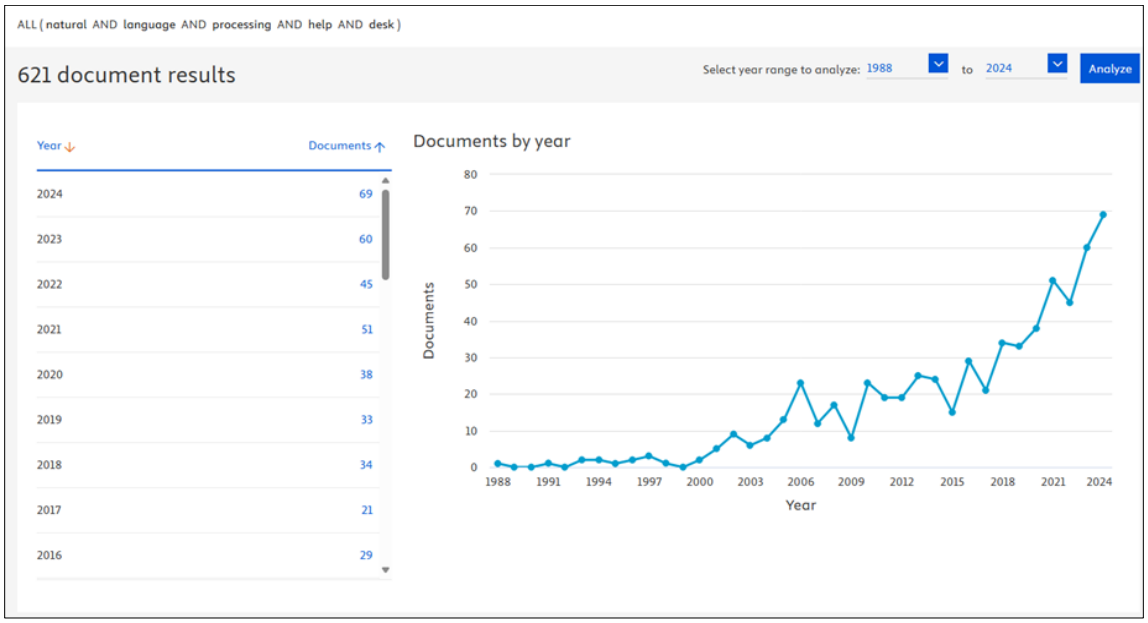


Ilustración 1 . Documentos publicados anualmente relacionados con lenguaje natural y sistemas de mesa de ayuda, según datos obtenidos de Scopus. Elaboración propia

En la actualidad, se utilizan modelos basados en aprendizaje automático para perfeccionar la categorización automática de casos, disminuir el tiempo de respuesta y mejorar la distribución de recursos en las mesas de servicio (Vital et al., 2024).

Dentro de los enfoques más empleados para mejorar la eficiencia en la gestión de incidentes se encuentran:

- Modelos de clasificación supervisada: Algoritmos como random forest y redes neuronales profundas han sido utilizados para la categorización de incidentes.
- Embeddings y representación semántica: Técnicas como embeddings basados en transformers han mejorado la comprensión contextual de los incidentes reportados.
- Modelos avanzados de PLN: Arquitecturas como BERT, T5 y GPT han demostrado gran capacidad para analizar y generar respuestas a partir de texto libre, facilitando la automatización de procesos en las mesas de servicio.

Estos avances han permitido reducir la carga operativa del equipo de asistencia de primer nivel, mejorar la categorización de incidentes y aumentar la exactitud en la asignación de casos a los grupos de solución pertinentes. En muchos casos, el uso de machine learning no sólo apoya, sino que puede incluso reemplazar algunas tareas realizadas por los operadores de primer nivel, mejorando la eficacia del proceso (Venegas Villarreal, Villar García & Mendoza De Los Santos, 2022).

Diversos estudios han aplicado modelos de machine learning para mejorar la gestión de incidentes en mesas de servicio. Un ejemplo es el trabajo de Qamili et al. (2018), quienes utilizaron aprendizaje automático para proponer un marco inteligente que optimiza la gestión de sistemas de tickets. Este marco aborda tres desafíos principales: la detección de spam, la asignación automática de tickets y el análisis de sentimientos. El estudio empleó un conjunto de datos de 18,917 registros provenientes de un sistema de tickets de una empresa de desarrollo de software. El proceso incluyó la aplicación de técnicas de limpieza de texto para eliminar signos de puntuación y palabras irrelevantes, seguido de la representación de los datos mediante un enfoque de "bag-of-words", donde cada palabra individual se convirtió en una característica. Los modelos entrenados, como SVM, Random Forest y SGD, mostraron un desempeño superior en términos de precisión y consistencia, mientras que los Decision Trees presentaron menores niveles de precisión. Este enfoque permitió mejorar la eficiencia en la asignación de tickets a los departamentos correspondientes y minimizar los falsos positivos en la clasificación de spam.

A nivel local, un ejemplo de implementación exitosa lo encontramos en la investigación de Ramírez Devia (2021), quien desarrolló un modelo de clasificación y priorización de gestión de PQRS en Colsubsidio, basado en procesamiento de lenguaje natural y aprendizaje automático. En

su estudio, se utilizó Naïve Bayes, árboles de decisión y máquinas de vectores de soporte (SVM) para vectorizar los textos de las PQRS y clasificarlos de manera eficiente. El modelo alcanzó un 94.5% de equilibrio constante entre las peticiones de los clientes, demostrando que este tipo de sistemas puede optimizar la gestión de solicitudes en organizaciones como Colsubsidio, con la posibilidad de seguir entrenando el modelo para mejorar la precisión y el recall.

Otro ejemplo claro se presenta en el repositorio de la Universidad de Antioquia, que centró su investigación en validar la información manejada a nivel interno por la empresa Bancolombia. La compañía utilizó Microsoft Teams para conectar a sus colaboradores independientemente de su ubicación geográfica, con el fin de optimizar la gestión de comunicaciones internas. Dado el aumento en el uso de esta herramienta, se propuso un proyecto para validar, mediante modelos de machine learning, si los mensajes intercambiados tenían un propósito laboral o no. Para ello, se empleó la técnica GloVe, que utiliza vectores de 500 dimensiones para facilitar la clasificación de los mensajes en tres categorías: Negativo, Neutral y Positivo.

Dentro del análisis que se realiza en este repositorio se puede entender que los diferentes métodos de vectorización de palabras (KNN, SVM, RF y XGBoost), ajustaban a la clasificación de palabras y que en general 12 de las palabras ajustaban a una conversación de manera laboral y cuando una conversación contiene más de 40 palabras es porque no se trataba de algo laboral. El porcentaje de conversaciones no laborales equivale aproximadamente a un 25% de las conversaciones obtenidas.

Para las conclusiones generales de este proyecto se pudo indicar que las metodologías más efectivas fueron las XG Boost y RF sobresalieron con una efectividad del 96% y de las demás técnicas utilizadas, pudieron brindar respuestas claras para lo que se estaba solicitando. finalmente son metodologías aplicables para resolver problemáticas relacionadas con el lenguaje natural.

4.2. Modelos de representación semántica utilizados

En los últimos años, los avances en modelos de lenguaje han permitido desarrollar representaciones vectoriales altamente efectivas para capturar la semántica de los textos. Estas representaciones, conocidas como embeddings, convierten oraciones o documentos en vectores numéricos de alta dimensión que preservan relaciones semánticas y contextuales. Su uso ha transformado tareas tradicionales del procesamiento de lenguaje natural (PLN), como la clasificación de textos, la detección de similitud semántica, la recuperación de información y la agrupación automática de documentos.

Para el desarrollo del presente trabajo, se utilizarán diferentes modelos preentrenados de embeddings, seleccionados por su capacidad de generalización, eficiencia computacional y precisión en contextos reales. A continuación, se describen los principales modelos que serán empleados.

4.2.1. Modelo all-MiniLM-L6-v2

El modelo all-MiniLM-L6-v2 es parte de la familia Sentence-Transformers y ha sido diseñado para generar embeddings de alta calidad a partir de oraciones y fragmentos de texto. Su arquitectura se basa en un transformer con seis capas, lo que proporciona un equilibrio adecuado entre rendimiento computacional y precisión semántica. Este modelo es particularmente útil en escenarios donde se requiere procesar grandes volúmenes de texto con eficiencia, sin comprometer la capacidad del modelo para capturar el contexto y el significado de las oraciones.

Su aplicación es común en tareas como:

- Clasificación automática de texto
- Recuperación semántica de información
- Análisis de similitud textual

- Agrupamiento temático de documentos (clustering)

Tokenización con WordPiece

El proceso de Tokenización representa una etapa crítica en los modelos de lenguaje, ya que determina cómo se fragmentan los textos antes de ser procesados. En el caso de all-MiniLM-L6-v2, se emplea la técnica WordPiece, que divide las palabras en subunidades denominadas tokens. Esto permite manejar palabras poco frecuentes o fuera del vocabulario original, reduciendo la complejidad del modelo y aumentando su capacidad de generalización.

Por ejemplo, la palabra "descomunalmente" puede dividirse en los tokens "des", "##comunal" y "##mente". Esta estrategia permite representar tanto el significado individual como el contexto completo de las palabras compuestas o derivadas. El modelo puede procesar hasta un máximo de 512 tokens por secuencia, lo que lo hace adecuado para textos de longitud media.

4.2.2. Modelo sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

El modelo sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 es parte de la familia de modelos desarrollados por Sentence-Transformers, diseñado específicamente para generar representaciones semánticas (embeddings) de frases o textos completos. Su funcionamiento se basa en convertir entradas de texto en vectores numéricos de tamaño fijo que conservan el significado semántico, permitiendo comparaciones, agrupamientos o clasificaciones posteriores. Este modelo en particular está optimizado para la detección de similitud entre oraciones (paraphrase mining) y soporta más de 50 idiomas, lo que lo hace altamente útil en contextos multilingües.

Este modelo está basado en la arquitectura de **MiniLM (Miniature Language Model)**,

que reduce significativamente el tamaño y los tiempos de inferencia sin sacrificar mucho rendimiento. Tiene 12 capas (de ahí su nombre L12), 384 dimensiones en el embedding resultante, y fue entrenado con técnicas de distillation a partir de modelos más grandes como BERT y XLM-RoBERTa. Utiliza la técnica de pooling mean para condensar la información de los tokens de la oración en un único vector. Gracias a su estructura ligera y multilingüe, es ideal para tareas de clasificación de texto, búsqueda semántica, agrupamiento de casos similares o recomendación automática en sistemas de soporte como mesas de ayuda.

(Reimers, 2019) “Fue entrenado usando la librería sentence-transformers, lo que permite usarlo directamente con SentenceTransformer, soportando más de **50 idiomas**, este ellos están Español, inglés, portugués, francés, alemán, italiano, árabe, ruso, chino, japonés, turco”

Por ejemplo, si se alimentan las frases:

- *"El sistema no permite el ingreso al portal"*
- *"No puedo acceder a la plataforma de servicios"*

el modelo generará embeddings similares para ambas frases, reflejando que tienen un significado similar, aunque las palabras sean distintas. Esto es clave en aplicaciones como segmentación de incidentes, ya que permite agrupar tickets escritos de forma diferente, pero con el mismo problema de fondo, facilitando respuestas automáticas y clasificación eficiente.

4.2.3. Modelo Qwen3-Embedding-0.6B

El modelo Qwen3-Embedding-0.6B es parte de la serie Qwen3, diseñada específicamente para tareas de embebido y reordenamiento de texto. Basado en una arquitectura *dual-encoder*, procesa

textos individuales para extraer representaciones semánticas densas utilizando el estado oculto final del token [EOS] arXiv. Gracias a un pipeline de entrenamiento en tres fases — preentrenamiento contrastivo no supervisado, ajuste supervisado con datos etiquetados y técnicas de fusión de modelos— el modelo logra equilibrar generalización, robustez y adaptabilidad a múltiples tareas

Las características más relevantes de este modelo es que está entrenado con 0.6 mil millones de parámetros, lo que lo hace ligero y eficiente.

Admite hasta 32,000 tokens de contexto y vectores de salida personalizables de hasta 1024 dimensiones (MRL support), siendo *instruction-aware*, es decir, puede ajustar su comportamiento según indicaciones específicas Hugging FaceSiliconFlow.

Posee capacidad multilingüe para más de 100 idiomas, incluso lenguajes de programación, y ha mostrado rendimiento competitivo en tareas como recuperación de texto, clustering, clasificación y búsqueda semántica

Qwen3-Embedding-0.6B es un modelo adecuado para nuestro proyecto porque ofrece embeddings multilingües optimizados para recuperación de información y agrupamiento de texto en lenguaje natural, justo lo que se requiere para automatizar y escalar la mesa de servicio de Colfondos. (Group, 2024)

1.1. Marco Teórico

4.3.1 Definición de chatbot

Un **chatbot** es un programa informático diseñado para simular una conversación con usuarios humanos, ya sea por texto o voz. Estos sistemas pueden funcionar basados en reglas predefinidas o utilizar inteligencia artificial, especialmente técnicas de Procesamiento de Lenguaje Natural (PLN), para entender la intención del usuario y generar respuestas.

4.3.2 Definición de mesa de servicio (Help Desk) y mesa de ayuda

La *mesa de servicio* o *help desk* es un servicio que centraliza la gestión de solicitudes, incidentes y consultas técnicas de los usuarios dentro de una organización. Su objetivo es proveer soporte técnico de primer nivel, resolver problemas comunes, orientar al usuario, y derivar los casos más complejos a niveles especializados, asegurando que exista un punto de contacto eficiente para atender fallas, requerimientos o quejas.

4.3.3 Definición PLN

El **Procesamiento de Lenguaje Natural (PLN)** es una disciplina de la inteligencia artificial que estudia la interacción entre computadoras y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten y generen lenguaje natural, para tareas como clasificación de texto, análisis de sentimiento, generación automática de lenguaje, traducción, extracción de información, entre otras.

4.3.4 Tokenización

La *tokenización* es la técnica que divide un texto en unidades más pequeñas llamadas *tokens*, que pueden ser palabras, subpalabras, caracteres o signos de puntuación, dependiendo del nivel de granularidad que se desea. Es uno de los primeros

pasos en los procesos de PLN, ya que prepara el texto para aplicar otras transformaciones como lematización o eliminación de stopwords.

4.3.5 Lematización y stemming

Stemming es un proceso de normalización de texto que consiste en recortar los morfemas (afijos) de las palabras para reducirlas a su raíz ("stem"), que puede no corresponder a una palabra real del idioma

Lematización es un procedimiento más preciso que utiliza diccionarios y análisis morfológico para convertir formas flexionadas de palabras en su forma base o lema, asegurando que sea una palabra válida en el idioma y considerando el contexto gramatical.

4.3.6 Eliminación de stopwords

La eliminación de *stopwords* consiste en suprimir palabras comunes que aparecen con mucha frecuencia en un idioma (como “de”, “la”, “el”, “y”, etc.), que tienen poco valor semántico para tareas analíticas de texto. Su eliminación reduce el ruido y mejora la eficacia de algoritmos de PLN al centrarse en términos más informativos

4.3.7 Análisis de frecuencia de términos

El análisis de frecuencia de términos es una técnica que consiste en contar cuántas veces aparece cada término (o token) dentro de un documento o corpus de documentos. Este análisis permite identificar las palabras más comunes, patrones, temas frecuentes y posibles sesgos en el texto. Es útil en etapas exploratorias de PLN para entendimiento del contenido

4.3.8 Bag of Words

El modelo *Bag of Words* (BoW) es una representación de texto en la que cada documento se describe por la frecuencia de aparición de cada palabra (token) en un vocabulario predefinido, sin tener en cuenta el orden de las palabras. Cada documento se convierte en un vector numérico donde cada dimensión corresponde a un término del vocabulario.

4.3.9 TF-IDF

TF-IDF es una técnica para ponderar términos en documentos del corpus. Combina dos métricas: *frecuencia de término (TF)*, que mide cuántas veces aparece un término en un documento, y *frecuencia inversa de documento (IDF)*, que penaliza términos muy comunes en muchos documentos. El resultado es un valor que indica la importancia de un término en un documento relativo al corpus.

4.3.10 Embeddings

Los *embeddings* son representaciones vectoriales densas de palabras o tokens en un espacio numérico de múltiples dimensiones, donde términos semánticamente similares quedan cerca entre sí. Estas representaciones permiten capturar relaciones semánticas (sinónimos, contextos similares) y se emplean en modelos modernos de PLN para superar limitaciones de modelos basados sólo en frecuencia

4.3.11 Regresión logística

La regresión logística es un modelo estadístico de clasificación supervisada que estima la probabilidad de que una observación pertenezca a una de dos clases posibles (o múltiples clases si se extiende), usando una función logística. En PLN sirve para clasificar textos, por ejemplo, determinar si una descripción pertenece a

categoría X o Y, basándose en características extraídas como TF-IDF o embeddings. (Para definiciones más formales puedes consultar textos de Machine Learning/clasificación de texto).

4.3.12 K-Means

Es un algoritmo de agrupamiento (*clustering*) no supervisado que particiona un conjunto de datos en *K* clusters, de modo que cada punto pertenece al cluster con la media más cercana (centroide). Se usa frecuentemente en PLN para agrupar documentos o términos similares en clusters semánticos.

4.3.13 DBSCAN

Es un algoritmo de clustering basado en densidad que identifica clusters de puntos densamente conectados, separándolos del ruido. No requiere que se especifique el número de clusters de antemano, y puede detectar clusters de forma arbitraria, siendo robusto ante datos atípicos.

4.3.14 UMAP

Es una técnica de reducción de dimensionalidad que permite proyectar datos de alta dimensión en espacios de menor dimensión (por ejemplo, 2D o 3D) conservando tanto la estructura global como local de los datos. Se utiliza en técnicas de visualización y para facilitar clustering sobre representaciones complejas como embeddings.

4.3.15 Nubes de palabras

Una *nube de palabras* es una representación visual de términos en la que el tamaño de cada palabra indica su frecuencia o importancia dentro del corpus. Es una

forma intuitiva de mostrar rápidamente cuáles son los términos más frecuentes o más relevantes en un conjunto de textos

4.3. Fuente de los datos

Los datos utilizados en este estudio provienen de Colfondos, específicamente de su mesa de servicio interna. La información corresponde a un histórico de 45,000 casos registrados entre agosto de 2023 a febrero de 2025, los cuales contienen descripciones de problemas y resolución en texto libre. El acceso a estos datos fue otorgado como parte de una solicitud formal, en la que se especificó que la información sería utilizada exclusivamente con fines académicos para el desarrollo de este trabajo de grado.

Para garantizar el cumplimiento de las políticas internas de seguridad y privacidad de Colfondos, la empresa realizó un proceso de enmascaramiento de datos antes de la entrega de los datos, asegurando que no contengan información sensible o identificable; el tratamiento de los datos incluirá pre procesamiento y limpieza, aplicando técnicas de procesamiento de lenguaje natural para estructurar la información y facilitar su análisis. Se garantizará el cumplimiento de normativas de privacidad y uso ético de los datos a lo largo de todo el proyecto.

4.4. Aplicación y/o aporte específico al campo

Este trabajo se enmarca principalmente en el campo de la analítica de datos, poniendo especial atención en el procesamiento de lenguaje natural y el aprendizaje automático. La implementación de estas técnicas permite convertir datos textuales sin estructura en información valiosa para la toma de decisiones, lo cual es crucial en la mejora de procesos en ambientes corporativos.

Bajo la perspectiva de la analítica de datos, este análisis ayuda a segmentar y organizar grandes cantidades de datos históricos de la mesa de servicio a través de técnicas de embeddings y clustering. Al agrupar incidentes parecidos y generar grupos estandarizados, el sistema puede identificar de una mejor forma patrones de recurrencia, lo que permite la automatización de respuestas a consultas frecuentes o casos similares.

Además de su impacto en la analítica de datos, este trabajo también realiza aportes significativos al campo de la ingeniería de software, específicamente en el desarrollo de asistentes virtuales inteligentes o chatbots para la gestión de soporte técnico. La implementación de un modelo de segmentación basado en NLP dentro de la mesa de servicio de Colfondos representa un avance hacia la automatización de procesos en las áreas de tecnología y servicio al cliente, lo que optimiza la asignación de recursos y mejora la experiencia del usuario.

Así mismo, desde una perspectiva organizacional, este proyecto tiene implicaciones en el ámbito de la gestión empresarial y la optimización de procesos operativos, al reducir la carga de trabajo manual en la mesa de servicio y mejorar los tiempos de respuesta a incidentes recurrentes.

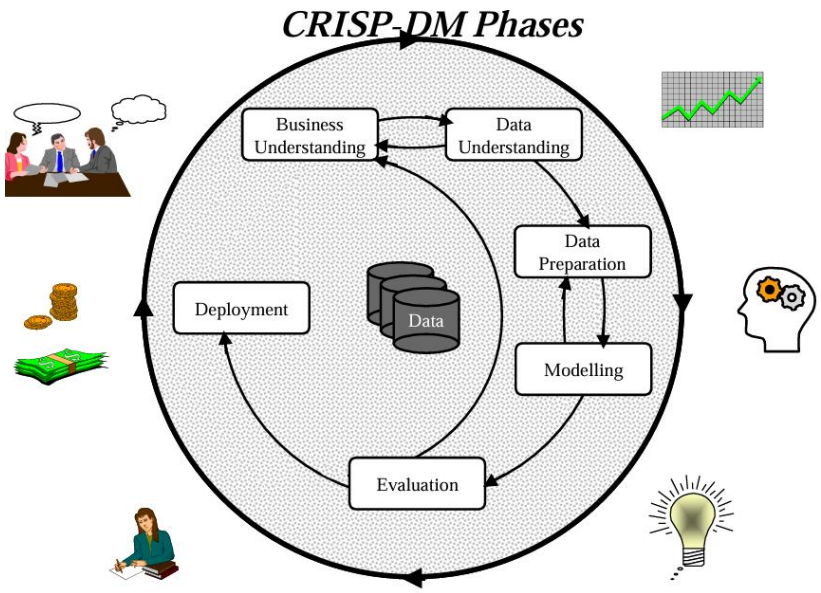
5. Metodología

La metodología que será usada en el proyecto será la CRISP-DM (Cross Industry Standard Process for Data Mining) ya que es uno de los métodos más usados en los proyectos de analítica de datos

por su propuesta de trabajo estructurado en fases y etapas que permiten llevar un orden en la ejecución y brinda la posibilidad de identificar oportunidades de mejora en el ciclo del proyecto.

Esta metodología consta de 6 etapas que abarcan desde el entendimiento de las necesidades del negocio y de los datos que se usarán en el proyecto, hasta el despliegue de la solución que cumple con la necesidad planteada.

“CRISP-DM ayuda a las organizaciones a comprender el proceso de minería de datos y proporcionan una hoja de ruta a seguir mientras se planifica y lleva a cabo un proyecto de minería de datos” (The CRISP-DM model: The New Blueprint for Data Mining, 2000).



.

Ilustración 2 Esquema de las fases del modelo CRISP-DM tomado de Chapman, P. (1999). The CRISP-DM User Guide

El diseño del modelo CRISP-DM muestra de manera evidente su carácter cíclico y organizado, resaltando cómo las etapas interrelacionadas facilitan una implementación organizada y adaptable de proyectos de minería de datos. Este ciclo permite el intercambio

constante entre fases como la comprensión del negocio, el entendimiento de los datos, su preparación, el modelado, la evaluación y la implementación. Además, su diseño visual resalta la relevancia de la iteración, dado que en numerosas situaciones se requiere retornar a etapas anteriores para efectuar modificaciones y mejoras, garantizando de esta manera la calidad y la capacidad de adaptación de las soluciones sugeridas. Esta perspectiva asegura que cada etapa del proyecto se encuentre en sintonía con los objetivos iniciales y con las demandas empresariales.

A continuación, se describe cada una de las fases del ciclo de vida en la metodología CRISP-DM:

5.1. Fase 1: Comprensión del Negocio

Esta fase es la parte inicial y fundamental para comprender el comportamiento actual del negocio y de la misma forma poder identificar la necesidad que motiva la ejecución del proyecto planteado en este anteproyecto. A partir de lo anterior, se identificó que una de las problemáticas radica en los largo tiempos de respuesta que se generan en la atención de casos de la mesa de servicio y el uso de intervención humano requerida para la gestión de solicitudes, lo cual incrementa la carga operativa, genera reprocesos y afecta negativamente

la satisfacción del usuario interno.

- Iniciamos la primera actividad mediante sesiones de trabajo con la persona que contaba con el conocimiento completo del funcionamiento operativo de la mesa de servicio de Colfondos. Estas reuniones fueron clave para comprender el flujo real de las solicitudes, los tipos de incidentes que pueden recibirse y el proceso de gestión para la resolución de los casos. Asimismo, se obtuvo una visión general de la base de datos entregada, lo que permitió identificar patrones comunes, estructuras recurrentes en los textos, tiempos de

respuesta dentro del flujo de atención y las tareas manuales involucradas. Esta comprensión contextual resulta esencial para diseñar un modelo de procesamiento de lenguaje natural (PLN) que sea útil y aplicable a la dinámica operativa real.

- De acuerdo a los datos históricos y el conocimiento adquirido en la etapa anterior, se procedió a realizar un análisis exploratorio sobre la base de datos de incidencias identificando en resumen las siguiente variables

TIPO DE VARIABLE	CANTIDAD	VARIABLES
Numéricas	1	ID de la solicitud
Categóricas nominales	8	Tipo de solicitud, Asunto, Nombre de la plantilla, Categoría, Subcategoría, Artículo, Resolución, Grupo
Categóricas ordinales	1	Prioridad
Texto libre	1	Descripción
Fecha/Hora	3	Hora de creación, Hora de vencimiento, Hora de resolución

Tabla 1 Clasificación de las variables según su tipo, cantidad y nombre

Posterior se tomó la muestra de la estructura de la base de datos con el fin de conocer el contexto del contenido que viene en cada una de las variables

Tipo de solicitud	ID de la solicitud	Asunto	Nombre de la plantilla	Categoría	Subcategoría	Artículo	Descripción	Resolución	Grupo	Hora de creación	Hora de vencimiento	Hora de resolución	Prioridad
Requerimiento	520	Habilitación - Futura	Default Request	REQUERIMIENTO - APLICATIVOS ADMINISTRATIVOS	Am	Default Request	Datos Nombre Usuario Yuly Grecia DIAZ DIAZ Teléfono de contacto: Sede: Nombre de la aplicación/producto y No. de versión: Futura Descripción de la Solicitud/ Error/ Inconveniente:	habilitado	**Control de Accesos**	2023-08-18 15:32	2023-08-22 15:32	2023-08-22 11:57	Prioritario

Ilustración 3 Ejemplo estructura contenido base de datos mesa de servicio Colfondos.

De la misma forma se realizó un análisis de las 15 categorías de incidentes más representativas dentro de la base de datos (Ilustración 3).

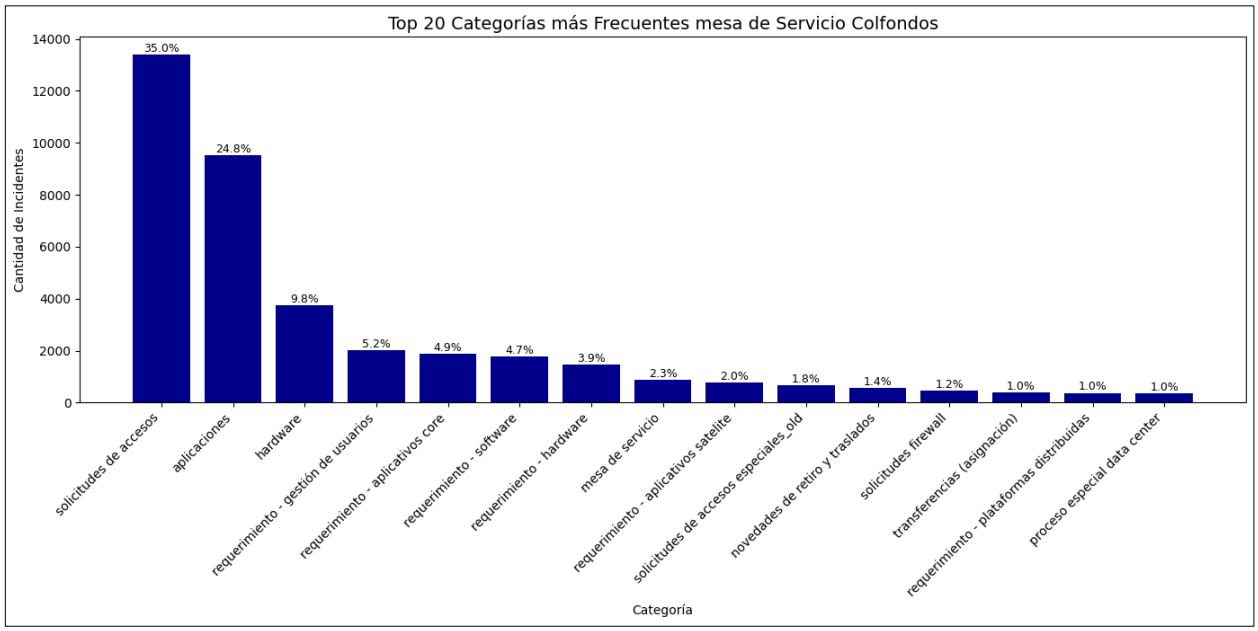


Ilustración 4 Top 20 categorías más frecuentes en la Mesa de Servicio de Colfondos. Este gráfico de barras representa la distribución porcentual de las categorías con mayor número de incidentes reportados.

Encontrando que estas categorías representan el 90% del total de solicitudes registradas, como se puede observar la categoría “solicitudes de accesos” es la más recurrente con

una participación del 33.5% del total seguida por “aplicaciones” que representa un 23.8%. En este conjunto estas dos categorías agrupan el 57.3% del total de incidencias.

A pesar de contar con una distribución clara de variables, durante el desarrollo de la primera actividad de esta fase se evidenció que, en múltiples ocasiones, las categorías presentan errores de clasificación. Esta situación compromete su confiabilidad como variables predictoras. En consecuencia, se definió que la variable objetivo será la columna *Descripción*, dado que contiene el relato real del caso y aporta información sustancial. Esta columna se considera clave para el cumplimiento del objetivo del proyecto, ya que permite enfocar los esfuerzos iniciales del modelo de procesamiento de lenguaje natural (PLN) en los aspectos más representativos y relevantes del incidente reportado.

- Finalmente, se definieron métricas cuantitativas y cualitativas que permitirán evaluar la efectividad del modelo propuesto. Entre los criterios de éxito establecidos se encuentran:
    - Precisión mínima esperada del modelo en la clasificación de textos, mayor al 75 % como primera entrega del ejercicio educativo,
    - Cobertura de respuestas automatizadas, casos resueltos sin intervención humana, y una reducción medible en los tiempos promedio de respuesta del 15%

5.2. Fase 2: Comprensión de los Datos

Después de tener claros los objetivos y el contexto desde la perspectiva de la organización.; se obtiene un conjunto de datos del fondo de pensiones el cual contiene el tipo de solicitud escalada por el usuario interno, donde se evidencia si es un requerimiento o un incidente. Dentro de la información, también se ve un ID específico para cada uno de los casos, el cual nos permite trabajar de manera más específica y ordenada permitiendo generar una trazabilidad de cada caso escalado dentro de la mesa de servicio. para

posteriormente segmentar los datos por el asunto que conlleva cada uno de los casos teniendo en cuenta diferentes variables.

La información obtenida dentro de la base de datos pasó por un proceso de enmascaramiento con el propósito de proteger la información confidencial del fondo de pensiones y se está manejando con fines académicos.

El objetivo principal de esta fase fue comprender en detalle las características de la base de datos proveniente de la mesa de servicio, con el fin de garantizar que los insumos utilizados en el modelado posterior fueran confiables, representativos y adecuados para la construcción de un modelo de clasificación de incidentes.

5.2.1. Longitud de los textos

En primera instancia, se llevó a cabo un análisis exploratorio de los datos (EDA), en el que se examinaron las principales variables disponibles, identificando la frecuencia de los diferentes tipos de incidentes reportados y la distribución general de los registros en el tiempo.

Columna	Tipo de Dato	Valores No Nulos	Valores Nulos	Porcentaje Nulos
Tipo de solicitud	object	42.845	0	0%
ID de la solicitud	int64	42.845	0	0%
Asunto	object	42.845	1	0%
Nombre de la plantilla	object	42.845	0	0%
Categoría	object	42.845	0	0%
Subcategoría	object	42.845	0	0%
Artículo	object	42.845	0	0%
Descripción	object	42.845	39	9%
Resolución	object	42.845	0	0%
Grupo	object	42.845	0	0%
Estado de solicitud	object	42.845	0	0%
Departamento	object	42.845	0	0%
Hora de creación	object	42.845	0	0%
Hora de vencimiento	object	42.845	0	0%

Hora de resolución	object	42.845	0	0%
Impacto	object	42.845	0	0%
Urgencia	object	42.845	0	0%
Prioridad	object	42.845	0	0%

Tabla 2 Análisis exploratorio de los datos (EDA)

La base de datos cuenta con un total de **42.845 registros** y **18 variables** asociadas a solicitudes de la mesa de servicio. Si bien el dataset incluye información diversa (tipo de solicitud, categoría, prioridad, estado, entre otros metadatos), para el desarrollo del presente proyecto únicamente se consideró el campo **"Descripción"**, dado que constituye la fuente de texto que será transformada en representaciones vectoriales (*embeddings*) y utilizada en el modelo de clasificación.

En la revisión inicial se identificó que la variable *"Descripción"* presenta un total de **42.806 registros completos**, mientras que **42 entradas se encuentran vacías**. Esto representa aproximadamente un **0,10% aprox. de valores faltantes**, lo cual, aunque es un porcentaje bajo, debe ser tratado para garantizar la consistencia del corpus textual. Para esta variable en especial se realizó un análisis exploratorio específico con el fin de evaluar la completitud, la extensión y la variabilidad del texto, así como de identificar posibles problemas de calidad que pudieran afectar la posterior construcción de embeddings.

Se calculó la longitud de cada descripción en caracteres y en número de palabras, con el fin de identificar registros demasiado cortos (poco informativos) o demasiado extensos.

Variable	Caracteres	Palabras
Cantidad	42.806	42.806
Media	563,37	68,8
Desviación Estándar	1248,08	131,13
Mínimo	1	1
Percentil 25	143	20
Mediana	264	37
Percentil 75	459	64
Máximo	32767	3862

Tabla 3 Resumen Estadístico de la Longitud de las Descripciones (Caracteres y Palabras)

Posterior se realizó el histograma de caracteres con KDE esto con el fin de poder entender cómo esta distribución de los caracteres de la columna descripción encontrando lo siguiente

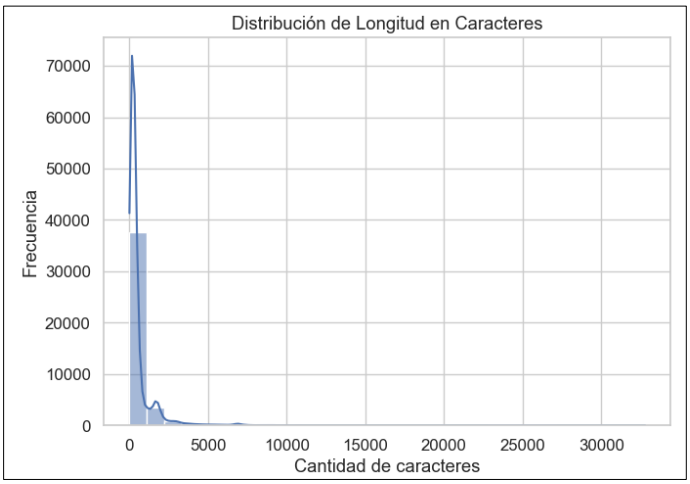


Ilustración 5 Distribución de la longitud de las descripciones en caracteres. ***Nota.*** *La figura muestra la densidad y frecuencia de aparición de textos según su longitud en caracteres*

En la ilustración 5 se observa una distribución sesgada a la derecha, con presencia de valores atípicos, lo cual puede ser generado producto de casos específicos como adjuntos legales, logs técnicos, mensajes automáticos, que requieren preprocesamiento especial o ser analizadas por separado como casos atípicos.

Así mismo, se interpreta que la variable descripción entre 0 y 1000 caracteres, con una alta concentración de registros por debajo de los 500, no obstante, se identifica una cola larga que se extiende hacia la derecha hasta más de 30.000 caracteres, lo que indica la presencia de valores atípicos o extremos que pueden afectar el modelo.

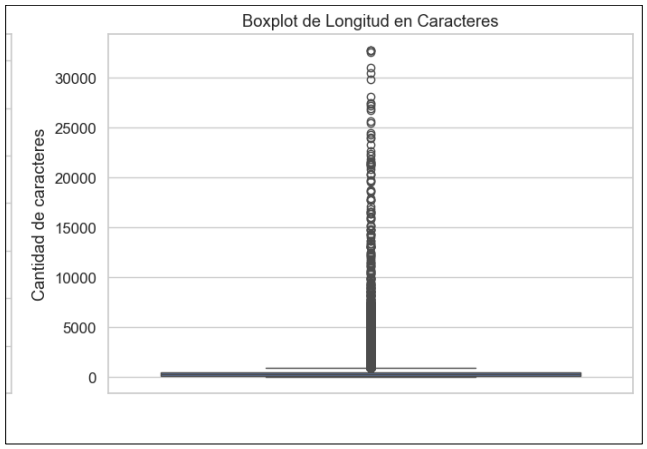


Ilustración 6 Boxplot de la longitud de las descripciones en caracteres.

De la ilustración 6 se evidencia una gran cantidad de valores atípicos (outliers) por encima del tercer cuartil (Percentil 75 = 459 caracteres), lo cual indica una alta variabilidad en la longitud de los textos analizados. El rango IQR es relativamente estrecho (entre 143 y 459 caracteres), lo que indica que el 50% de los datos entre el cuartil 3 y 1 está bien contenido.

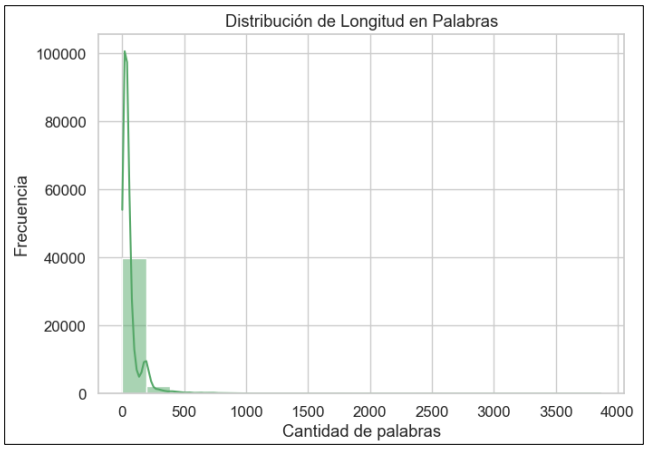


Ilustración 7 Distribución de la longitud de las descripciones en palabras

La imagen 7 muestra una distribución asimétrica hacia la derecha, donde la mayoría de las descripciones contienen una cantidad baja de palabras menor a 100, adicional se observa una cola larga hacia la derecha**,** con descripciones que alcanzan hasta aproximadamente 3.800 palabras,

aunque estos casos son muy poco frecuentes. Finalmente se crea una curva lo cual muestra una alta concentración de casos breves, posiblemente correspondientes a tickets, incidentes o mensajes simples, mientras que los textos más extensos podrían estar relacionados con reportes, mensajes concatenados o copias de correos automáticos.

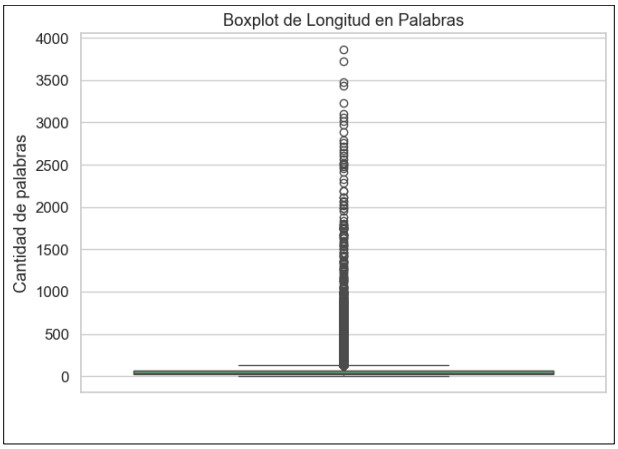


Ilustración 8 boxplot de longitud en palabras

En la ilustración 8 se puede evidenciar que la mayor parte de las descripciones se concentran en rangos bajos de palabras menores a 100 aprox., lo que concuerda con el histograma de la ilustración 7; se observan numerosos puntos por encima de los bigotes del boxplot, que corresponden a descripciones excepcionalmente largas, llegando a superar las 3.500 palabras.

Finalmente, el análisis completo mostró que la mayoría de los casos se encuentran en un rango manejable (20–64 palabras), lo cual resulta adecuado para su posterior representación en embeddings. Sin embargo, se identificaron registros problemáticos:

- Textos extremadamente cortos (1–2 palabras), con baja carga semántica y poco valor informativo para el modelo.

- Textos atípicamente largos (miles de palabras o caracteres), que podrían corresponder a errores de carga o inclusión de información irrelevante (ej. copias de logs, trazas técnicas, etc.).
- Estos hallazgos sugieren la necesidad de implementar reglas de filtrado o normalización durante el pre procesamiento.

5.2.2. Análisis de faltantes

En esta etapa se verificó la completitud de la columna *"Descripción"*. Los resultados evidenciaron que, de un total de 42.845 registros, únicamente 42 se encuentran vacíos, lo que equivale a un 0,10% del total de la base de datos.

Este hallazgo permite concluir que la proporción de valores faltantes es mínima y, en consecuencia, no afecta de manera significativa la representatividad del corpus textual. Sin embargo, al tratarse de nuestra variable objetivo para la generación de vectores mediante modelos de embeddings, incluso los registros vacíos adquieren relevancia práctica no aportan al entrenamiento ni a la evaluación del modelo, ya que las técnicas de procesamiento de lenguaje natural (PLN) requieren entradas textuales explícitas para la construcción de representaciones vectoriales. Por esta razón, dichos casos serán descartados en las fases de pre procesamiento, asegurando la coherencia del corpus y evitando inconsistencias en la tokenización y posterior cálculo de embeddings.

5.2.3. Registros Duplicados

El análisis permitió identificar que un total de 1391 descripciones duplicadas, lo que corresponde al 3.25% de la base de datos

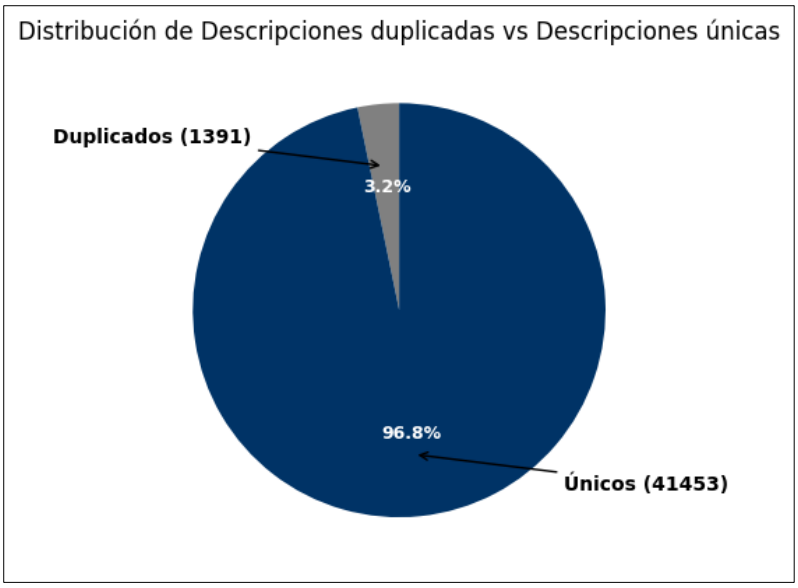


Ilustración 9 Distribución de participación de duplicados en la base de datos.

. Aunque el porcentaje de duplicados es relativamente bajo, es relevante para el modelado de embeddings, ya que:

- Algunos duplicados reflejan incidentes genuinos recurrentes (por ejemplo, solicitudes de restablecimiento de contraseña), lo que es consistente con la operación de la mesa de servicio.
- Otros duplicados podrían corresponder a errores de registro, introduciendo ruido y sesgando al modelo hacia incidentes frecuentes sin aportar información adicional.

Como estrategia, se plantea eliminarlas en el pre procesamiento para garantizar que el modelo de embeddings se entrene sobre casos únicos y no sobre repeticiones textuales.

5.2.4. Variabilidad léxica

El análisis de la variabilidad léxica de los casos se hizo con el objetivo de evaluar la riqueza del vocabulario, identificar términos frecuentes y detectar posibles problemas de redundancia o ruido que podrían afectar la construcción de embeddings.

5.2.4.1 Análisis cuantitativo del corpus

- El corpus contiene un total de 2.825.017 palabras distribuidas en 40.607 palabras únicas, lo que evidencia una diversidad léxica significativa.
- Las 20 palabras más frecuentes se presentan a continuación, reflejando la concentración de ciertos términos en las descripciones de los casos:

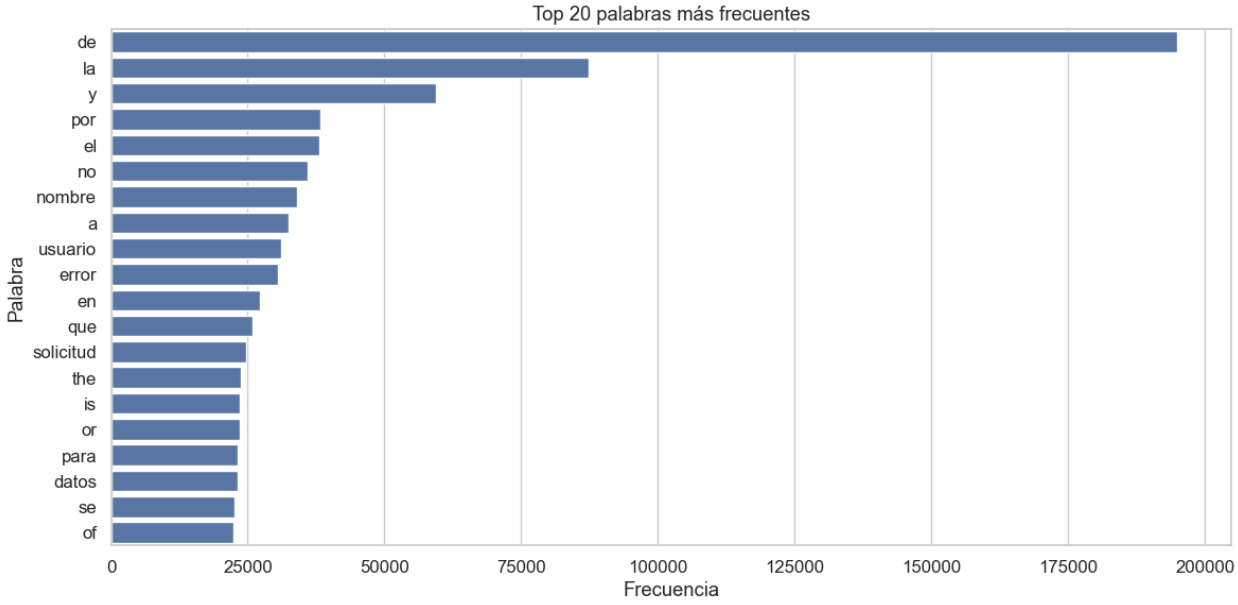


Ilustración 10 Distribución de palabras más frecuentes sin limpieza de base de datos

Una gran parte de los términos más frecuentes son palabras funcionales (stopwords), como “de”, “la”, “y”, “por”, “el”, “no”, “en”, “que”, “a”, “se”, “the”, “or”, “is”, “of”, que no aportan contenido semántico para la clasificación.

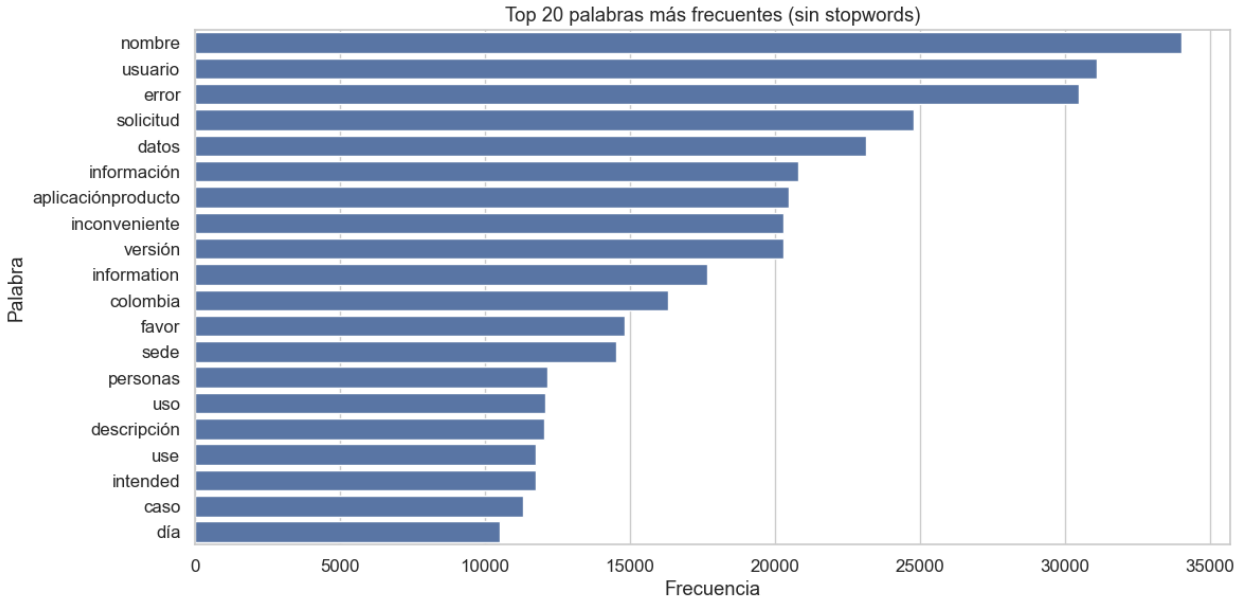


Ilustración 11 Distribución de palabras más frecuentes sin stopwords

Sin embargo, al no tener en cuenta esas palabras se destacan palabras clave relacionadas con el dominio, como **“usuario”**, **“error”**, **“solicitud”**, **“nombre”** y **“datos”**, que reflejan la naturaleza de los incidentes en la mesa de servicio, particularmente problemas de acceso y errores en los sistemas.

Adicionalmente se observa la presencia de palabras en inglés (“the”, “or”, “is”, “of”) indica que algunas descripciones contienen términos técnicos o copias de mensajes del sistema, lo que será importante considerar en el preprocesamiento.

Finalmente, para complementar el análisis estadístico, se generó una nube de palabras que permite observar gráficamente los términos más frecuentes en las descripciones. La nube de palabras refleja visualmente la concentración de incidentes recurrentes y la predominancia de ciertos conceptos en el corpus.

Ilustración 12 Nube de palabras con mayor frecuencia

5.2.5. Sesgos en la información

El análisis de sesgos se centró en identificar si ciertos términos o temas predominantes en las descripciones podrían afectar la representatividad del modelo de segmentación. Esto es relevante, ya que los embeddings captaron patrones semánticos del texto, y si algunos incidentes están sobrerrepresentados, el modelo podría sesgar hacia esos casos.

5.2.5.1 Frecuencia de términos clave

Se analizaron las palabras más frecuentes ya identificadas en la sección 2.2.4. Entre las palabras clave más comunes destacan:

- **“usuario”**, **“error”**, **“nombre”**, **“datos”**, **“solicitud”**

Estas palabras reflejan un predominio de incidentes relacionados con problemas de **acceso al sistema, errores de autenticación y manejo de datos**.

5.2.5.2 Frecuencia de términos clave

- La alta frecuencia de estos términos indica que la base está concentrada en incidentes recurrentes específicos, lo que puede provocar que el modelo tenga un sesgo hacia estas clases de incidentes al generar embeddings.
- Otros temas menos frecuentes podrían quedar subrepresentados, haciendo que el modelo tenga menor capacidad de generalización en incidentes atípicos o menos comunes.
- Además, la presencia de palabras en inglés (“the”, “or”, “is”, “of”) sugiere que ciertos textos técnicos de sistemas se incluyen de manera sistemática, lo que podría introducir un sesgo adicional hacia términos no traducidos o log de sistemas.

5.3 Fase 3: Preparación de los Datos

En esta fase construimos el conjunto de datos final que será utilizado para el modelado. Aunque en la fase anterior ya se haya realizado una comprensión inicial de los datos, en esta etapa se realiza un trabajo más específico para limpiar, transformar, seleccionar y estructurar los datos de manera adecuada. El objetivo es garantizar que la calidad y formato de los datos permiten extraer patrones significativos y obtener resultados confiables.

En el caso de nuestro proyecto, la preparación de los datos será fundamental para lograr modelos precisos y robustos. Basándonos en antecedentes como el estudio de Qamili et al. (2018), donde se aplicaron técnicas de limpieza de texto y representación mediante bag-of-words, y en la investigación de Ramírez Devia (2021) en PQRS de Colsubsidio utilizando técnicas de vectorización y clasificación de textos, adoptaremos estrategias similares. Realizaremos en primera instancia y para pruebas una limpieza profunda de los textos, normalización, eliminación de ruido (stopwords, puntuaciones), y posteriormente aplicaremos técnicas modernas de

representación semántica, como la generación de embeddings a partir de modelos de lenguaje, para mejorar la capacidad de clasificación.

Este procedimiento busca maximizar la calidad de las entradas al modelo de machine learning, asegurando un mejor desempeño en la segmentación de incidentes y priorización de casos.

Durante el proceso de limpieza del corpus textual se identificó la presencia recurrente de un mensaje automático corporativo (disclaimer legal y de confidencialidad) al final de múltiples descripciones. Este contenido se repetía hasta tres veces en casos reiterativos y no aportaba información relevante para la caracterización semántica de los incidentes.

Dado que dichos mensajes constituyen ruido textual y podrían distorsionar la representación vectorial de los embeddings, se procedió a eliminarlos de forma sistemática en la etapa de preprocesamiento. Esta decisión permitió depurar el conjunto de datos y garantizar que el modelo se entrenará exclusivamente sobre información significativa vinculada al problema de la mesa de servicio, mejorando la calidad de las representaciones y reduciendo el riesgo de sesgos hacia patrones irrelevantes.

5.3.1 Generación de embeddings con múltiples modelos

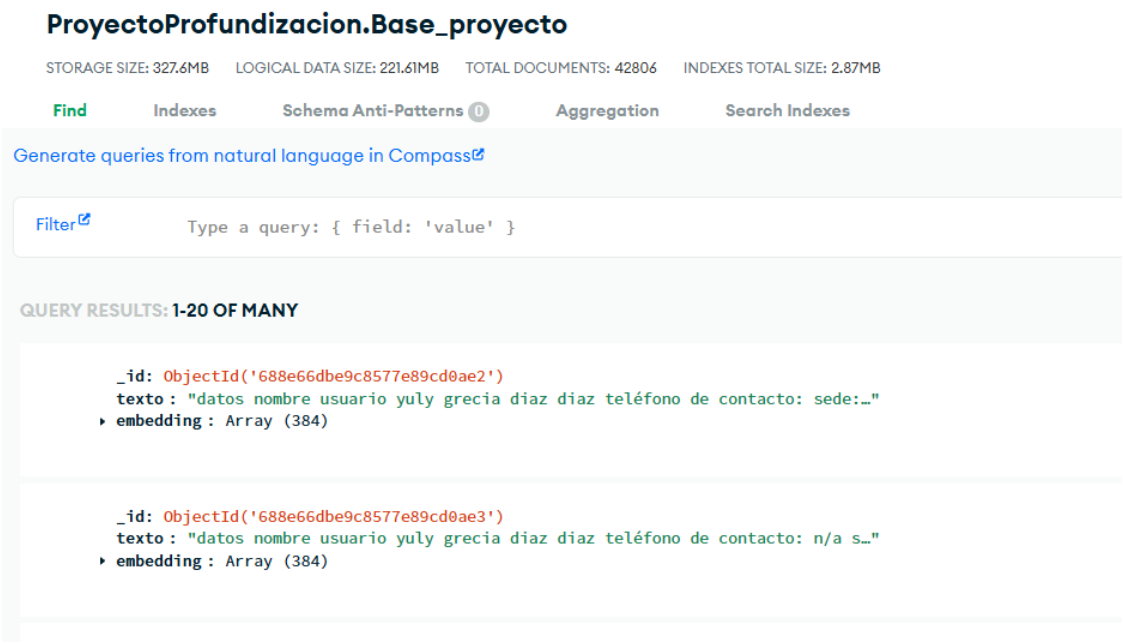
En esta fase también se generaron las representaciones vectoriales de las descripciones y segmentar los incidentes de la mesa de servicio utilizando técnicas de clustering.

Para capturar distintas perspectivas semánticas, se utilizaron tres modelos de embeddings pre entrenados de la librería sentence-transformers:

1. paraphrase-multilingual-MiniLM-L12-v2
2. Qwen3-Embedding-0.6B
3. paraphrase-multilingual-MiniLM-L6-

Cada modelo transforma la columna Descripcion_limpia en un vector denso de características, preservando la similitud semántica entre descripciones.

Estos embeddings se fueron almacenando en MongoAtlas para que pudieran ser usados en los pasos posteriores.



Cada uno de estos modelos generó embeddings distintos para el mismo conjunto de casos, almacenados en la base de datos MongoDB.

5.4 Fase 4: Modelado y segmentación de casos

5.4.1. Aplicación de algoritmos de clustering

El objetivo de esta fase fue realizar la segmentación de los casos mediante técnicas de aprendizaje no supervisado, con el fin de identificar grupos homogéneos y patrones subyacentes en los datos. Para ello, se utilizaron las representaciones vectoriales (embeddings) obtenidas previamente a partir de tres modelos distintos de lenguaje, lo que permitió evaluar el impacto del modelo de representación en la calidad de los clusters.

5.4.2. Definición de evaluador de clustering

Para estandarizar el procedimiento se diseñó la clase ClusterEvaluator, que integra distintos algoritmos de segmentación y sus métricas de evaluación. Esta clase permitió reutilizar código y comparar de manera sistemática el rendimiento de los modelos de embeddings bajo los mismos criterios.

Los algoritmos implementados fueron:

- K-Means: se evaluó un rango de valores de k entre 2 y 30.
- DBSCAN: se analizaron distintos valores del parámetro eps y min_samples.

Las métricas calculadas fueron:

- Inertia: como medida de compacidad de los clusters.
- Silhouette score: como indicador de separación y coherencia de los grupos.
- Calinski-Harabasz index: para medir la dispersión entre e intra-clusters.

(Diagrama ilustrando K-Means y DBSCAN, con ejemplos de partición esférica vs densidad de puntos)(RESULTADO FINAL DE LOS EXPERIMENTOS)

5.4.3. Protocolo experimental - Clusterización modelos no supervisados

Se diseñaron dos pipelines experimentales para cada modelo de embeddings (paraphrase-multilingual-MiniLM-L12-v2, distiluse-base-multilingual-cased-v2, paraphrase-multilingual-MiniLM-L6-v2):

- Pipeline A — Sin reducción de dimensionalidad (directo sobre embeddings originales)

- Pipeline B — Con reducción de dimensionalidad (UMAP a 2D) y posterior clustering sobre la proyección)

El propósito de comparar ambos pipelines fue:

1. Cuantificar la calidad de los clusters en el espacio original de alta dimensión.
2. Verificar el efecto de una proyección no lineal (UMAP) sobre la interpretabilidad y sobre las métricas de clusterización.
3. Comparar resultados entre los tres modelos de embeddings.

5.4.3.1 Preparación de datos (pasos comunes)

Se realizó un proceso de extracción por lotes desde MongoDB, con el propósito de cargar de manera eficiente los embeddings de cada modelo y consolidarlos en matrices separadas. Cada vector fue asociado a metadatos básicos del caso (identificador y texto original), lo que permitió rastrear los resultados del clustering hasta la fuente de datos.

Este paso garantizó que la información estuviera lista para ser utilizada para los algoritmos de agrupamiento y reducciones de dimensionalidad posteriores.

(Imagen de diagrama del flujo de extracción desde MongoDB)

(Diagrama esquemático que muestre: consulta a colección de Mongo → extracción por lotes → creación de embeddings_final y metadata_total. Para obtenerlo: crear un diagrama en draw.io o similar, o capturar el diagrama de flujo del script de extracción.)

5.4.3.2 Pipeline A — Clustering sobre embeddings originales (alto dimensional)

En la primera fase del modelado se evaluó el desempeño de los algoritmos de clustering directamente sobre los embeddings generados por los tres modelos seleccionado; el objetivo de este pipeline consistió en identificar si, a partir de los vectores en su forma original (sin aplicar técnicas de reducción de dimensionalidad), era posible obtener una partición clara de los datos mediante el algoritmo K-Means.

Evaluación de métricas de clustering

Se exploraron distintos valores de k (número de clústeres) y se calcularon indicadores internos de validación, tales como:

- Coeficiente de silueta, que midió la coherencia interna de los clústeres.
- Índice de Calinski-Harabasz, que valoró la separación y compacidad de los grupos.
- Índice de Davies-Bouldin, que reflejó el grado de solapamiento entre clústeres.

RESULTADOS = QWEN(GUIA MICHAEL)

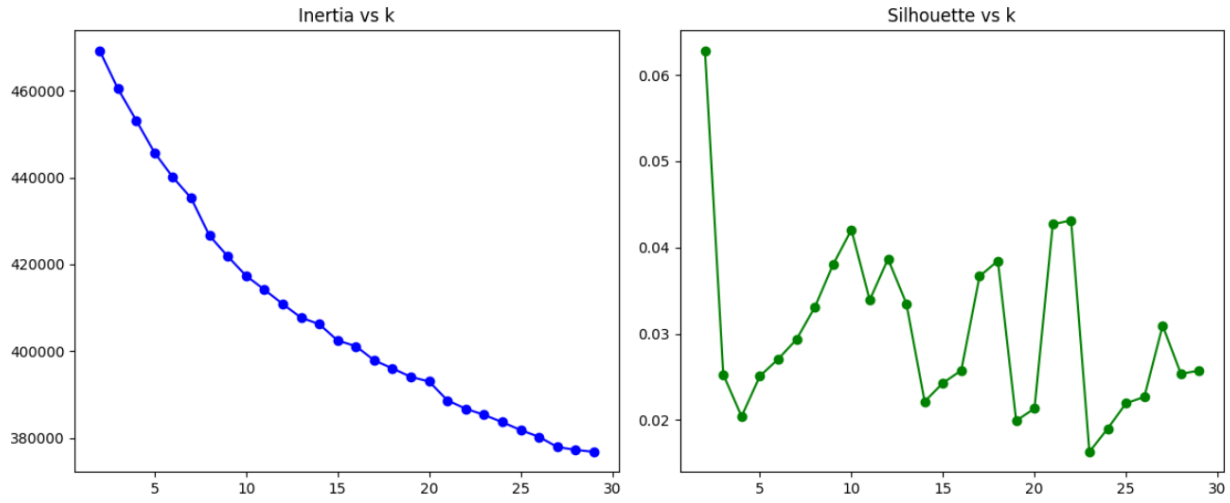
RESULTADOS

Para el modelo paraphrase-multilingual-MiniLM-L6-v2 se realizo la evaluación de diferentes configuraciones de k, teniendo como resultado

k	Inertia	Silhouette	Calinski-Harabasz
2	469,287.19	0.063	1,039.7
3	460,498.99	0.025	938.0
4	453,190.02	0.020	865.5
5	445,759.27	0.025	838.2
6	440,094.49	0.027	789.3
7	435,275.29	0.029	744.0
8	426,640.17	0.033	774.3
9	421,853.38	0.038	745.9
10	417,387.15	0.042	721.0

k	Inertia	Silhouette	Calinski-Harabasz
11	414,183.20	0.034	687.0
12	410,883.61	0.039	660.7
13	407,739.63	0.034	637.8
14	406,220.48	0.022	603.3
15	402,462.02	0.024	593.9
16	401,155.28	0.026	565.4
17	397,887.56	0.037	556.4
18	396,022.66	0.038	537.9
19	394,084.60	0.020	522.2

k	Inertia	Silhouette	Calinski-Harabasz
20	393,007.15	0.021	502.3
21	388,661.40	0.043	506.4
22	386,737.26	0.043	494.8
23	385,334.22	0.016	481.1
24	383,640.76	0.019	470.4
25	381,820.24	0.022	461.4
26	380,259.23	0.023	451.8
27	377,926.66	0.031	447.3
28	377,250.43	0.025	434.3
29	376,796.46	0.026	421.1



Al observar la tabla y la gráfica, se nota que:

- La inercia disminuye de manera gradual al aumentar k, lo que es esperado ya que más clusters permiten ajustar mejor los datos.
- El coeficiente de Silhouette es muy bajo en todos los casos (máximo 0.063 para k=2), indicando que los clusters no están claramente separados y hay considerable solapamiento.
- El índice de Calinski-Harabasz también disminuye con k, sugiriendo que la densidad relativa de los clusters se degrada al aumentar el número de grupos.

Entre los valores evaluados, el modelo con k=2 fue seleccionado como el de mejor desempeño relativo, presentando la inercia más alta, el coeficiente de Silhouette más alto (0.063) y el mayor

índice de Calinski-Harabasz (1,039.7). Los demás valores k mostraron disminuciones progresivas en estos indicadores.

(Tabla resumen con los valores de silueta, Calinski-Harabasz y Davies-Bouldin obtenidos para cada modelo de embeddings, destacando a MiniLM-L12-v2 como el mejor relativo)

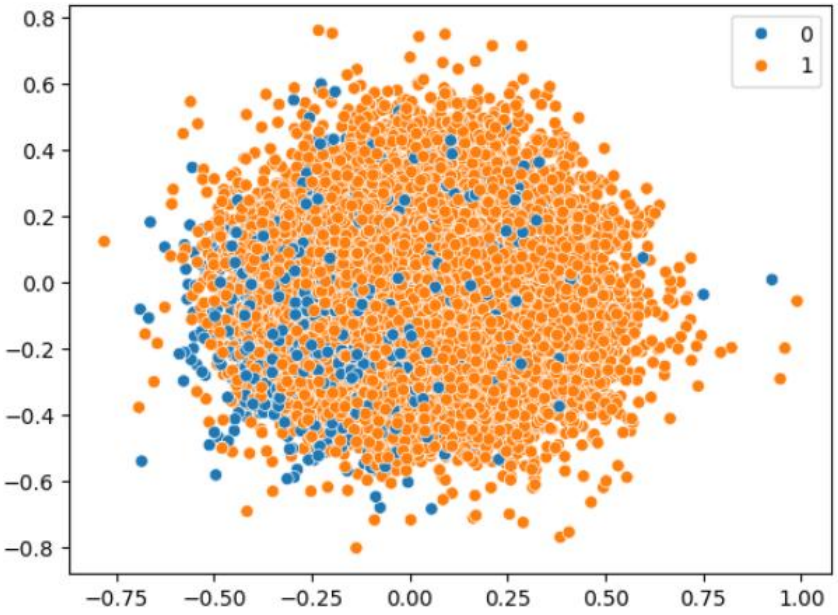
Los resultados obtenidos fueron bajos en todos los modelos, lo cual evidenció que los embeddings no generaron separaciones marcadas entre los datos. Sin embargo, las métricas fueron consistentes, mostrando que los tres modelos de embeddings capturaron patrones de forma similar.

Entre ellos, el modelo MiniLM-L12-v2 se destacó con un desempeño ligeramente superior, aunque la diferencia respecto a los demás no fue significativa.

Con el fin de complementar las métricas, se generaron gráficos de dispersión en dos dimensiones a partir de las primeras componentes de los embeddings. Estos gráficos representaron los datos coloreados según los clústeres asignados por K-Means.

Grafica mas analisis.qwen

Para paraphrase-multilingual-MiniLM-L6-v2 .



(Gráfico de dispersión de los embeddings sin reducción de dimensionalidad, mostrando la asignación de clústeres en colores para cada modelo)

La inspección visual mostró que los grupos formados presentaban un alto nivel de solapamiento, lo que coincidió con lo evidenciado por los indicadores cuantitativos. En particular, los límites entre clústeres fueron difusos y no se observó una segmentación clara.

Hallazgos del Pipeline 1

En este primer pipeline se observó que los embeddings lograron capturar cierta estructura latente en los datos, aunque dicha representación no resultó suficiente para conformar clústeres claramente diferenciados. Los tres modelos evaluados mostraron un comportamiento muy similar en términos de desempeño, destacándose de manera relativa el modelo MiniLM-L12-v2, que presentó métricas ligeramente superiores. Al contrastar los indicadores de validación con las

visualizaciones generadas, se evidenció que el clustering aplicado directamente sobre los embeddings, sin reducción de dimensionalidad, produjo resultados limitados en cuanto a la separación y definición de los grupos.

(Tabla que muestre en una sola vista la comparación de las métricas para los tres modelos)

5.4.3.3 Pipeline B — Clustering con reducción de dimensionalidad

En un segundo experimento se implementó un pipeline que incorporó una etapa de reducción de dimensionalidad mediante la técnica UMAP (Uniform Manifold Approximation and Projection) antes de aplicar los algoritmos de clustering. Esta decisión se tomó debido a que, en el pipeline anterior, las métricas obtenidas fueron bajas y las visualizaciones mostraron un alto nivel de solapamiento entre grupos, lo que sugirió la necesidad de proyectar los embeddings a un espacio más compacto que facilitara la identificación de patrones.

La reducción se realizó configurando UMAP con parámetros que priorizaron tanto la preservación de la estructura local de los datos como la separación global. El resultado fue una proyección bidimensional de los embeddings, lo que permitió representar los casos en un plano de fácil interpretación.

((crear codigo)Gráfico de dispersión de los embeddings reducidos con UMAP, sin aplicar aún clustering, mostrando la distribución general de los datos en dos dimensiones)

Aplicación de clustering sobre el espacio reducido

Una vez obtenida la proyección, se aplicaron nuevamente algoritmos de clustering, principalmente K-Means y DBSCAN, siguiendo la misma lógica de evaluación del pipeline

anterior. En este caso, el espacio reducido permitió una mayor claridad en la separación de ciertos grupos, aunque las métricas siguieron siendo moderadas.

El análisis de K-Means mostró que los valores de silueta, Calinski-Harabasz y Davies-Bouldin mejoraron en comparación con los obtenidos sobre los embeddings originales. Aun así, los resultados no alcanzaron niveles altos, lo que indicó que la segmentación de los casos seguía siendo un desafío.

((lineas)Tabla resumen comparativa entre las métricas de K-Means con reducción mediante UMAP, evidenciando las mejoras relativas alcanzadas)

Por su parte, el uso de DBSCAN en el espacio reducido permitió detectar posibles agrupaciones densas en subconjuntos de los datos, aunque la sensibilidad a los parámetros (particularmente el valor de ε) generó variaciones importantes en el número de clústeres detectados.

((`result_dbscan`)Tabla resumen comparativa entre las métricas de DBSCAN sin reducción de dimensionalidad y con reducción mediante UMAP, evidenciando las mejoras relativas alcanzadas)(entre mas alto el eps mas bajo podria ser el indice de sillueta)

Observaciones generales del Pipeline 2

El segundo pipeline permitió apreciar que la reducción de dimensionalidad con UMAP ofreció ventajas tanto en la interpretación visual como en la ligera mejora de los indicadores de validación de clustering. En términos de desempeño, los tres modelos de embeddings mantuvieron un comportamiento similar, y nuevamente el modelo MiniLM-L12-v2 se ubicó como la alternativa con el rendimiento relativo más alto. Sin embargo, el comportamiento de las métricas evidenció que, si bien UMAP contribuyó a una mejor organización del espacio de

representación, los casos analizados continuaron mostrando una estructura difusa que dificultó la segmentación precisa en grupos bien definidos.

(Gráfico resultados de clustering con reducción de dimensionalidad)(realizar analisis)

5.4.3.4 Comparación de pipelines (revisar y complementar en caso de ser necesario)

Tras la ejecución de los dos pipelines diseñados, se llevó a cabo una comparación sistemática entre los resultados obtenidos sin reducción de dimensionalidad y aquellos logrados aplicando previamente UMAP.

En el Pipeline 1, donde los algoritmos de clustering se aplicaron directamente sobre los embeddings originales, se evidenció que los modelos lograron capturar cierta estructura latente, pero no alcanzaron a generar clústeres bien diferenciados. Las métricas de validación presentaron valores bajos en todos los casos y las visualizaciones mostraron un marcado solapamiento entre los grupos, lo que dificultó la interpretación y segmentación de los datos.

Por otro lado, en el Pipeline 2, al incorporar UMAP como etapa intermedia, se observó una ligera mejora en los indicadores de validación y, sobre todo, una representación más clara en términos visuales. La reducción de dimensionalidad permitió proyectar los embeddings en un espacio bidimensional donde las agrupaciones fueron más fáciles de identificar, aunque sin llegar a niveles de separación óptimos.

Al comparar los tres modelos de embeddings a lo largo de ambos pipelines, se observó que su desempeño fue muy similar, pero el modelo MiniLM-L12-v2 mantuvo un rendimiento relativo superior en ambos enfoques. Este comportamiento se tradujo en métricas consistentemente mejores, aunque aún moderadas, y en una organización más coherente de los clústeres. Cabe

resaltar que el modelo QWEN requirió tiempos de procesamiento considerablemente mayores en comparación con los otros dos, lo que limitó su eficiencia práctica.

(Tabla comparativa consolidada entre Pipeline 1 y Pipeline 2)(tomar el mejor resultado de cada modelo)

Finalmente, la comparación entre pipelines permitió establecer que la inclusión de UMAP fue beneficiosa para el proceso de clustering, tanto por la ganancia visual como por la ligera mejora en las métricas de validación. Sin embargo, los resultados continuaron reflejando que la estructura de los datos no era fácilmente separable en clústeres bien definidos, lo que constituye un hallazgo clave en esta fase de modelado y segmentación.

5.4.4 Experimento con embeddings sin limpieza robusta

Después de haber evaluado los distintos modelos de embeddings y los dos pipelines de clustering, se determinó que el modelo MiniLM-L12-v2 ofreció el mejor desempeño relativo y que la reducción de dimensionalidad mediante UMAP aportó mejoras significativas en la organización de los datos. Con base en este aprendizaje, se diseñó un nuevo experimento cuyo objetivo fue explorar cómo variaba el desempeño del clustering al modificar la estrategia de preprocesamiento de los textos.

En este caso, se decidió no aplicar una limpieza robusta al texto, es decir, no se eliminaron palabras claves, stopwords ni se realizó lematización. Esta decisión se justificó en función del tipo de tarea que se estaba desarrollando: la segmentación de casos en una mesa de servicio. En este contexto, mantener palabras funcionales y de uso frecuente resultó valioso, ya que muchas de ellas pueden aportar matices relevantes en la forma en que los usuarios formulan sus solicitudes. A diferencia de tareas clásicas de clasificación de texto donde la eliminación de

stopwords y la lematización suelen mejorar la generalización, en este escenario se buscaba capturar con mayor fidelidad el lenguaje natural de los casos reportados.

Con estos nuevos embeddings generados bajo un preprocesamiento más flexible, se repitió el pipeline de reducción con UMAP y posterior aplicación de K-Means. Los resultados obtenidos mostraron una mejora clara en las métricas de validación interna, evidenciando que la preservación de más información textual favoreció la formación de grupos más consistentes. El modelo identificó un total de xxxx clústeres, cada uno con características lingüísticas propias.

(Gráfico de dispersión en dos dimensiones con UMAP mostrando la distribución de los ocho clústeres resultantes, cada uno representado con un color diferente)(pendiente del nuevo proceso)

5.5 Fase 5: Evaluación

Una vez definido el modelo de embeddings más adecuado (MiniLM-L12-v2), aplicado el procedimiento de reducción de dimensionalidad mediante UMAP y determinado que la mejor configuración técnica correspondía a la generación de 8 clústeres, fue necesario pasar a una etapa de evaluación con el fin de validar la utilidad práctica y la coherencia semántica de los grupos obtenidos.

En esta fase no se buscó ajustar parámetros ni mejorar métricas, sino interpretar y comprender los resultados del modelado, verificando si los clústeres encontrados reflejaban patrones significativos en los textos analizados y si resultaban consistentes con los objetivos del proyecto.

5.5.1. Identificación de textos representativos por clúster

Con el fin de interpretar los clústeres, se calculó para cada grupo un texto representativo, definido como aquel más cercano al centroide del clúster en el espacio reducido por UMAP. Esta

elección se fundamentó en que el centroide constituye una “media” del comportamiento de los puntos en el espacio vectorial, y por tanto el texto más cercano a este puede considerarse como el ejemplo más característico de todo el grupo.

Adicionalmente, se listaron ejemplos adicionales de textos presentes en cada clúster. El propósito de esta estrategia fue ofrecer material concreto que permitiera ilustrar la naturaleza y diversidad interna de los datos agrupados. De esta manera, los clústeres no solo fueron definidos en términos numéricos, sino también a partir de contenidos comprensibles para expertos del dominio.

Cabe resaltar que esta fase de interpretación fue clave, ya que al tratarse de un proyecto basado en aprendizaje no supervisado, los algoritmos por sí solos no asignan significados a los clústeres. La identificación de representativos y ejemplos se convierte en el puente entre la salida técnica del modelo y la validación por parte del negocio o área de aplicación.

5.5.2. Identificación de textos representativos por clúster

Una vez seleccionados los textos representativos, se revisó manualmente cada clúster con el fin de analizar si los elementos agrupados compartían un hilo conductor en cuanto a temática, palabras clave o intención. Este proceso permitió identificar la consistencia interna de los clústeres y evaluar en qué medida los grupos respondían a categorías lógicas para el problema de negocio.

Se observó que, si bien existía variabilidad dentro de algunos clústeres, en general cada grupo logró capturar un eje común de significado. Esto reforzó la pertinencia del número de clústeres seleccionado en la fase de modelado.

5.5.3. Resultados consolidados

Finalmente, se elaboró una tabla con los resultados obtenidos en esta fase de evaluación, en la que se organizaron los ocho clústeres identificados. Para cada uno de ellos se incluyó un texto representativo, calculado como el caso más cercano al centroide del grupo en el espacio reducido, así como una selección de ejemplos adicionales que permiten ilustrar la naturaleza de las solicitudes que lo componen.

Esta tabla sirvió como un insumo central para la interpretación de los resultados, ya que ofrece una visión sintética pero al mismo tiempo concreta de los patrones detectados por el modelo. Con base en ella, el equipo de trabajo contó con evidencia clara para iniciar la discusión con los expertos del dominio, quienes son los encargados de validar la pertinencia de cada agrupamiento, asignar significados a los clústeres y determinar su posible valor para la organización.

En este sentido, la construcción de los clústeres no se entendió como un fin en sí mismo, sino como un punto de partida para el diálogo con el negocio, garantizando que los resultados técnicos puedan transformarse en conocimiento útil y accionable.

(Tabla resumen con los ocho clústeres, mostrando para cada uno el texto representativo y ejemplos asociados que evidencian su contenido semántico)

5.6. Fase 6: Despliegue

Una vez validados los resultados del modelado y consolidada la interpretación preliminar de los clústeres, se dio inicio a la fase de despliegue. En esta etapa, el propósito fue trasladar los hallazgos técnicos hacia un entorno de aplicación práctica, asegurando que los XXXX clústeres identificados no quedarán únicamente como un resultado exploratorio, sino que se integraran a un proceso concreto de apoyo al negocio.

El escenario de despliegue definido consistió en la implementación de un chatbot de primer nivel, en el cual los clústeres actuarían como categorías base para la clasificación automática de las consultas de los usuarios. Sin embargo, el despliegue no se concibió únicamente como una herramienta de etiquetado o enrutamiento: el chatbot fue diseñado para ofrecer a cada consulta una respuesta de primer nivel, es decir, una solución a priori basada en el clúster al que pertenece el caso.

IMAGEN INTERFAZ(raul)

De esta forma, si una solicitud coincidía con un clúster específico (por ejemplo, relacionado con trámites de afiliación, aportes, retiro de cesantías, etc.), el chatbot no solo la clasificaba, sino que proporcionaba inmediatamente una respuesta inicial orientativa. Esta respuesta podía incluir información general, pasos básicos de resolución o enlaces a recursos pertinentes, reduciendo así la necesidad de escalar automáticamente cada caso al equipo humano y optimizando el tiempo de atención.

EJEMPLO CHATBOT(cuando funcione)

El despliegue, por lo tanto, representó un salto cualitativo en el uso de los clústeres: se pasó de una agrupación técnica de casos a un sistema interactivo que entrega valor directo a los usuarios desde el primer contacto.

De esta forma, esta fase no consiste en trasladar los resultados a un ambiente productivo, sino que los transformó en capacidades operativas concretas, permitiendo que se pueda ofrecer a los usuarios un nivel de atención más ágil y eficiente desde el primer contacto.

(Esquema ilustrativo del funcionamiento del chatbot: consulta del usuario → identificación del clúster → generación de respuesta de primer nivel → enrutamiento posterior si es necesario)(diagrama de flujo con el funcionamiento de la interfaz interactiva)

5.6.1. Beneficios esperados

Aunque este trabajo tiene un carácter académico y no implica una implementación directa en Colfondos, los resultados alcanzados permiten vislumbrar beneficios potenciales si en el futuro se llegaran a aplicar. Entre ellos se destacan:

- **Optimización en la atención al usuario:** la agrupación de solicitudes facilita la identificación de patrones frecuentes, lo que permitiría ofrecer respuestas de "primer nivel" en un chatbot o asistente virtual. Estas respuestas funcionarían como soluciones preliminares que orientan al usuario de manera inmediata.
- **Reducción de carga operativa:** al filtrar y resolver consultas simples desde el inicio, se liberaría tiempo del personal especializado, concentrándolo en casos complejos que realmente requieren intervención humana.
- **Estandarización de la comunicación:** los clústeres posibilitan la construcción de un repositorio de respuestas coherentes y alineadas con la naturaleza de cada grupo de solicitudes.
- **Escalabilidad del servicio:** con una base de clústeres ya definida, se podrían integrar nuevas solicitudes de manera más sencilla, ajustando o ampliando los grupos identificados.
- **Valor agregado para la analítica organizacional**: más allá de la atención, la clasificación automática de casos aporta insumos para la toma de decisiones estratégicas, al permitir detectar tendencias y problemáticas recurrentes.

En conjunto, estos beneficios refuerzan la relevancia académica del estudio, mostrando cómo el análisis de lenguaje natural y la agrupación de casos pueden sentar las bases para soluciones innovadoras en la gestión de la mesa de servicio.

6. Lista de referencias

(Reimers, 2019)

(Group, 2024)

- Amazon Web Services (AWS). (s. f.). *¿Qué es un chatbot?* AWS. https://aws.amazon.com/es/what-is/chatbot/

- Zendesk. (s. f.). *¿Cómo funciona un help desk?* Zendesk México. https://www.zendesk.com.mx/blog/como-funciona-un-help-desk/

- OpenWebinars. (2022). *Técnicas clave para procesamiento de texto en NLP*. OpenWebinars. https://openwebinars.net/blog/tecnicas-clave-para-procesamiento-texto-nlp/

- DataCamp. (2023). *What is tokenization in NLP?* DataCamp. https://www.datacamp.com/es/blog/what-is-tokenization

- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. (Referencia sobre stemming y lematización:

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html)

- Analytics Vidhya. (2022, junio). *Stemming vs Lemmatization in NLP – Must Know Differences*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv. https://arxiv.org/abs/1301.3781 (Referencia para embeddings).

- Sidorov, G. (2019). *TF-IDF, Bag of Words and Beyond*. arXiv. https://arxiv.org/abs/1802.00400

- Wikipedia. (s. f.). *DBSCAN*. En *Wikipedia*. https://es.wikipedia.org/wiki/DBSCAN

- McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861. joss.theoj.org

- Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer: Text analytics based on word clouds. *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, 1833–1842. https://doi.org/10.1109/HICSS.2014.231. ACM Digital Library

- Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

- Salton, G., Wu, H., & Yu, C. T. (1981). The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science*, *32*(3), 175–186. https://doi.org/10.1002/asi.4630320304