ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

PROCESAMIENTO DE LENGUAJE NATURAL Y GENERACIÓN AUTOMÁTICA DE ALERTAS DE LAS RESEÑAS DE CLIENTES, EN UNA EMPRESA DE TELECOMUNICACIONES DEL ECUADOR.

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

JIMMY JOEL LANDÍN CASAL

CARLOS ISIDRO REINA CAMPUZANO

GUAYAQUIL - ECUADOR
Año: 2022

DEDICATORIA

Le dedico este proyecto a mi familia y amigos, quienes siempre estuvieron apoyándome en todo mi camino, y que, sin ellos no podría haber llegado hasta aquí. Le dedico especialmente este trabajo a mis padres Petita Casal y Jimmy Landín, quienes me han inculcado siempre disciplina, responsabilidad y dedicación, lo cual es mi pilar para siempre buscar la excelencia académica y superación personal; le dedico este trabajo también a mi esposa Viviana Vanegas y a mis hijos Ethan y Emma, que ellos siempre han sido y serán mi fuente de inspiración y mi motivo de esfuerzo en todo ámbito de mi vida.

Ing. Jimmy Landín

El trabajo realizado en este proyecto lo dedico a mis padres María Campuzano y Carl Reina, por ser los pilares fundamentales en mi vida y siempre creer en mí, por incentivarme a seguir adelante y enseñarme a nunca rendirme ante ninguna circunstancia.

A mi abuela materna y segunda madre Abigail Gurumendi (†), fuente de la máxima expresión de amor conocida.

A mi leal e incondicional compañera y la persona más importante de mi vida Merli Marín, gracias por tanto, amor de mi vida.

Ing. Carlos Reina

AGRADECIMIENTO

Primero agradecemos a Dios por la vida, por siempre estar con nosotros en todo momento, iluminándonos con conocimiento para desarrollar este proyecto.

Se agradece de manera muy afectuosa a los profesores quienes nos guiaron en este camino, el cual no concluye aquí, sino más bien, es el comienzo de un mundo de mejora continua, aprendizaje autónomo y dedicación por la ciencia de datos.

Agradecemos también a nuestra tutora, PhD. Carmen Vaca quien nos ha compartido mucho de su conocimiento, y su aporte en este proyecto ha sido fundamental para llegar a este punto.

Extendemos un agradecimiento muy especial a la empresa en que laboramos, específicamente al Departamento de Calidad, por brindarnos su apoyo y facilitarnos los datos necesarios para que este proyecto fuese posible, además de las herramientas y los recursos informáticos necesarios para su implementación.

A todos aquellos que participaron de manera directa e indirecta en este proyecto, se les agradece infinitamente por su colaboración.

DECLARACION EXPRESA

“La responsabilidad y autoría de esta Tesis de Grado, nos corresponden a nosotros exclusivamente (Jimmy Joel Landín Casal, Carlos Isidro Reina Campuzano); y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL”.

Firmado electrónicamente por:
JIMMY JOEL LANDIN

Jimmy Joel Landín Casal

Firmado electrónicamente por:
CARLOS ISIDRO REINA CAMPUZANO

Carlos Isidro Reina Campuzano

COMITÉ EVALUADOR

El tribunal de sustentación del Trabajo de Titulación del Sr. JIMMY JOEL LANDIN CASAL y del Sr. CARLOS ISIDRO REINA CAMPUZANO, después de ser examinado en su presentación, memoria científica y de defensa oral, da por aprobado el Trabajo de Titulación.

Firmado electrónicamente por:
CARMEN KARINA VACA

Ph.D. Carmen Vaca Ruiz
PROFESOR TUTOR

Firmado electrónicamente por:
EDUARDO SEGUNDO CRUZ RAMIREZ

MSc. Eduardo Cruz Ramírez
PROFESOR REVISOR

RESUMEN

El presente proyecto, analiza el uso de algoritmos de Aprendizaje Automático con Procesamiento de Lenguaje Natural (NLP), para el estudio de las reseñas emitidas por los clientes en una Encuesta de Satisfacción de Calidad, de una reconocida Empresa de Telecomunicaciones en el Ecuador, así mismo, se realiza una clasificación de tipos de comentarios de la encuesta mediante distintos algoritmos de aprendizaje automático supervisado, y posteriormente se desarrolla un estudio de segmentación de los clientes a partir de variables temporales, espaciales y demográficas con algoritmos de aprendizaje automático no supervisado, con la finalidad de proveer una herramienta de generación de alertas automáticas para las partes interesadas de la empresa, que permitan realizar un seguimiento oportuno cuando exista un nueva reseña urgente que requiera ser atendida, repercutiendo así en planes de acción para la experiencia del cliente.

Palabras Claves: Procesamiento de Lenguaje Natural, NLP, Aprendizaje Automático, Alertas, Segmentación, Clasificación.

ABSTRACT

The present project analyzes the use of Automatic Learning algorithms with Natural Language Processing (NLP), for the study of the reviews issued by the clients in a Quality Satisfaction Survey, of a recognized Telecommunications Company in Ecuador, as well as Likewise, a classification of the types of comments in the survey is carried out using different supervised machine learning algorithms, and subsequently a customer segmentation study is developed based on temporal, spatial and demographic variables with unsupervised machine learning algorithms, with the purpose of providing a tool for generating automatic alerts for the interested parties of the company, which allow timely follow-up when there is a new urgent review that requires attention, thus having an impact on action plans for the customer experience.

Key Words: Natural Language Processing, NLP, Machine Learning, Alerts, Segmentation, Classification.

INDICE GENERAL

ABREVIATURAS

NLP	Natural Language Processing
ML	Machine Learning
AI	Artificial Intelligence
MLP	Multilayer Perceptron
SMOTE	Synthetic Minority Over-sampling Technique
PCA	Principal Component Analysis
SVM	Support Vector Machine
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
API	Application Programming Interface
ETL	Extract Transform Load
SUPERCIAS	Superintendencia de Compañías

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

El presente proyecto analiza con mayor profundidad data no estructurada, como lo son, las reseñas de los clientes en una encuesta telefónica de satisfacción de calidad en una empresa de telecomunicaciones, aquellos comentarios que los clientes expresan en cada una de las distintas partes del cuestionario, y que significan una fuente de información muy valiosa para la empresa, pero que hasta el momento no había sido explotada, o por lo menos no había generado un valor significativo para la misma.

Uno de los principales motivadores de este proyecto, es la rápida evolución que existe en la industria por generar al cliente una experiencia grata, no sólo medir un indicador de satisfacción, no basta con que el cliente califique bien o mal un indicador de calidad, sino que al cliente se le genere una experiencia que vaya a recordar, y la cual va a compartir con otras personas; dicha evolución, va mucho más allá de conocer qué producto o servicio prefiere el cliente, sino también lo que este piensa de la empresa en todos sus ámbitos, dichos pensamientos es muy difícil de captar con preguntas cerradas, por lo que es mejor evaluar todo lo que haya dicho el cliente en su reseña o comentario, cómo este se expresó, con que intensidad dijo ciertas palabras, o que palabras dijo para referirse a la empresa; todo esto que representa datos no estructurados en la encuesta, se lo analiza en el presente proyecto con Procesamiento de Lenguaje Natural para obtener una categorización de contenido, así como un análisis de sentimientos; con la finalidad de obtener una herramienta de generación de alertas de reseñas, la cual permita a las distintas partes interesadas, realizar planes de acción que gestionen los principales problemas, y se revaliden las buenas prácticas que se están implementando en la empresa. Para cumplir el objetivo de este proyecto, se recurren a distintos tipos de algoritmos del aprendizaje automático supervisado y no supervisado, así como técnicas de preprocesamiento de texto, para finalmente obtener una visualización en Tableau para analizar las alertas de las reseñas, junto con una integración con Slack para obtener notificaciones de dichas alertas.

CAPÍTULO 1

1 PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del Problema

En la actualidad, en el territorio ecuatoriano la oferta de empresas que proporcionan servicios tecnológicos como conexión a Internet y datos para la sección corporativa es muy limitada. Las diferentes estrategias comerciales de dichas empresas han permitido que pocas de ellas hayan quedado en el mercado, captando cada una un alto porcentaje de los clientes de la industria. Entre dichas empresas que se han manejado muy bien en el mercado está la empresa “AAA”, la cual ha hecho una gestión eficaz e inversión en infraestructura y tecnología, ofreciendo una amplia gama de servicios y soluciones tecnológicas en el Ecuador, siempre estando a la vanguardia en procesos e innovación de sus plataformas.

Una de las áreas que ha tenido una gran evolución al interior de la empresa “AAA”, es la sección de Customer Experience del Departamento de Calidad, la cual realiza planes de acción en conjunto con los demás departamentos de la empresa. A partir de los resultados de una encuesta telefónica, que se realiza a los clientes para medir su satisfacción mediante indicadores ponderados. Con el pasar del tiempo dicha encuesta ha tenido cambios en su forma de medición y el formulario respectivo, para que cada vez más abarque las diferentes interacciones del cliente con la empresa, y, además, se realicen planes de acción con los distintos departamentos que sean más orientadas al cliente, no sólo pensando en la medición, sino también en la experiencia que tiene el cliente en la empresa.

Con el pasar de los años se evidencia un avance progresivo en el indicador de satisfacción de calidad, pero aún existen debilidades en el proceso de evaluación. Estas debilidades pueden ser abordadas usando estrategias y técnicas de analítica de datos. Una de las primeras observaciones hechas se

refiere al hecho de que los resultados de los indicadores pueden ser subjetivos por las calificaciones numéricas dadas por los mismos clientes.

Adicionalmente, no se está analizando una información valiosa, que puede enriquecer el análisis, contenida en las reseñas de los clientes (data no estructurada), que corresponde al mayor porcentaje del total de la llamada de satisfacción de calidad. Analizar solamente el indicador, puede resultar en una visión sesgada, por más representativa que sea la muestra, o por más correcta que sea la redacción de las preguntas del cuestionario; el complementar la información dada por los indicadores, junto con el análisis de las reseñas de los clientes genera un mayor valor hacia una analítica más objetiva, y por ende unos planes de acción más orientados al cliente para mejorar continuamente en su experiencia al usar los servicios que brinda la empresa.

1.2 Justificación

La empresa “AAA” al analizar las reseñas realizadas por los clientes a lo largo de toda la encuesta de Satisfacción de Calidad, podrá identificar de modo más eficaz los problemas que enfrentan los clientes, con el fin de orientar los planes de acción de una forma más completa y precisa; actualmente las acciones se diseñan, basándose solo en indicadores subjetivos, por lo que al cambiar dicho enfoque, resultaría en la posibilidad de identificar grupos con problemas comunes, dado que, al explorar los sentimientos y las categorías de contenido derivados de las reseñas de los clientes, se podría clasificar a los clientes. Por ejemplo aquellos clientes que tienen muchos problemas graves, aquellos clientes que aunque están satisfechos con la empresa, siempre se les presenta un tipo de problema en particular, o incluso aquellos que son leales a la empresa pero que necesitan un requerimiento específico, y así poder lograr un mayor seguimiento y tratamiento a los clientes que en verdad lo necesitan con planes de acción más enfocados, aumentando así la satisfacción del cliente y por ende permitiendo un crecimiento de la empresa.

Adicional a la analítica de datos necesaria para poder llevar a efecto todo lo propuesto, también se requiere para este caso particular, un sistema de alertas que se integre activamente a la plataforma de la empresa, mediante el cual, se van a generar acciones por parte del personal interno, estas alertas serán integradas y configuradas en la herramienta de trabajo colaborativo por medio de canales a los que tendrán acceso los usuarios específicos de cada área, en ella podrán conocer indicadores de su interés y tener un panorama más claro acerca de la situación del cliente, lo que contribuirá para una ágil y oportuna toma de decisiones.

En los últimos años debido a los avances tecnológicos que se han dado y a las facilidades que estos han proporcionado se ha observado que la Transformación Digital debe ser tomada como una obligatoriedad en el ámbito empresarial, es por esto que diversas industrias apuntan sus esfuerzos y recursos en proporcionar infraestructura y tecnología que le ayude a sus empresas a generar un valor agregado que sea reconocido en el mercado por sus clientes. En este contexto el procesamiento de lenguaje natural (NLP por sus siglas en inglés) ha ganado gran protagonismo entre las tecnologías que captan la atención de todo tipo de empresas de diversas industrias siendo cada vez más relevante y generando mayor interés, entre sus ventajas se puede mencionar que permite la comprensión y el análisis de grandes volúmenes de datos de texto el cual puede ser semi-estructurado o no estructurado, también mediante estos algoritmos es posible automatizar procesos de una manera rápida y en tiempo real de tal manera que funcionen prácticamente sin intervención humana o en su defecto con una intervención mínima.

1.3 Solución Propuesta

Una vez explicado el problema y la justificación de este, es de vital importancia tomar un curso de acción hacia alternativas de solución, por lo que en el presente proyecto se propone el siguiente flujo.

En la primera fase, se realiza la recolección de datos no estructurados por parte de los operadores telefónicos de la sección de Customer Experience de la empresa “AAA”, los cuales levantan los datos mediante la encuesta de satisfacción de calidad, estos registros son almacenados de manera simultánea tanto en audio como en texto en los repositorios de la empresa, para luego realizar una depuración, preprocesamiento y etiquetado de los datos, los mismos que servirán para el entrenamiento de un modelo de Clasificación, que permita generar predicciones de las categorías del contenido de las reseñas de los clientes.

En una segunda fase, una vez obtenida la categorización del contenido de las reseñas, se enriquecerá la información con características internas y externas del cliente, con lo cual, se procederá a realizar una segmentación de clientes, que formará parte de los análisis y resultados dentro un dashboard interactivo, que se integrará al flujo de trabajo de la empresa mediante una herramienta de trabajo colaborativo, la cual generará alertas y notificará a las partes interesadas cuando exista un caso de urgencia que atender.

El uso de sistemas de procesamiento de lenguaje natural consolidará a la empresa “AAA” en el camino hacia una implementación eficiente de la transformación digital, y proporcionará al cliente una mejor experiencia, dando la credibilidad a la empresa en temas como optimización de procesos, reducción de tiempos de ejecución y respuesta oportuna.

Con el fin de evaluar la precisión del modelo, se utilizarán métricas de evaluación para escoger el mejor modelo de procesamiento de lenguaje natural (NLP).

1.4 Objetivos (General y Específico)

1.4.1 Objetivo General

Implementar un modelo de aprendizaje automático para clasificar texto de reseñas de clientes, en una Encuesta de Satisfacción de Calidad, y usar los resultados como datos para una herramienta integrada de generación automática de alertas, que genere valor para la empresa.

1.4.2 Objetivos Específicos

- Determinar la polaridad de los textos de las reseñas de los clientes, en la Encuesta de Satisfacción de Calidad.

- Implementar un clasificador multiclase para categorizar las reseñas de los clientes, en la Encuesta de Satisfacción de Calidad.

- Caracterizar a los clientes según variables económicas, financieras y sociodemográficas, para crear un nivel de prioridad / urgencia de atención de alertas.

- Implementar un módulo de generación de alertas que usa el conocimiento inferido de los datos de texto analizados para generar notificaciones automáticas en una herramienta de trabajo colaborativo.

- Desarrollar una herramienta de visualización que proporcione a las partes interesadas un análisis 360 de los distintos aspectos del cliente, obteniendo así un detalle más profundo de su situación y sus perfiles en distintos ámbitos.

1.5 Metodología

Para implementar la solución propuesta y alcanzar los objetivos planteados se aplicará diferentes técnicas en cada una de las etapas del proyecto.

Al comenzar se analizará las reseñas del cliente por cada sección, se realizará una depuración previa para obtener una buena calidad de los datos, además se llevará el texto a la representación de una matriz document-term para ser usada como entrada para algoritmos de NLP que permitan identificar grupos de clientes con problemas comunes (no solo con los sentimientos, o categorías de los comentarios, sino también con características internas y externas del cliente). Se entrenarán diversos modelos de clasificación de contenido para escoger aquel de mejor rendimiento en la categorización de contenido. Luego se usarán algoritmos de análisis de sentimientos para identificar la polaridad de los comentarios de los clientes.

Finalmente se creará alertas en tiempo real en un dashboard dinámico, las mismas que se ejecutarán mediante una herramienta de trabajo colaborativo, estas notificaciones permitirán a los usuarios finales, (los cuales son las partes interesadas en este proyecto, en este caso los distintos departamentos de la empresa “AAA” que participan en la cadena de valor del cliente), generar las mejores estrategias y planes de negocio con los clientes.

El pipeline de la metodología a utilizar es el siguiente:

Figura 1.1 Pipeline del Proyecto
Fuente: Elaboración Propia

En dicha arquitectura, podemos ver que existen varias fases:

Fase 1.- A partir de las llamadas que realizan los operadores telefónicos de la sección de Customer Experience con respecto a la encuesta satisfacción de calidad, se almacena en un repositorio propio de la empresa los audios de las llamadas, y además se registran los resultados tabulados de la encuesta en la plataforma de QuestionPro (donde se encuentran registradas las reseñas de los clientes).

Fase 2.- Se obtiene una base depurada de comentarios históricos con un etiquetado de contenido manual de distintas categorías, por parte de la sección de Customer Experience (Satisfacción de Calidad de la Empresa “AAA”), la cual posteriormente pasa por otro proceso de depuración y de preprocesamiento de texto, para que pueda ser usada en el entrenamiento del modelo.

Fase 3.- Se entrenan los modelos de procesamiento de lenguaje natural con los datos ya depurados (categorización de contenido), mediante métricas de evaluación se determina cuál es el mejor modelo que se obtuvo para poder implementarlo. Además, se realiza el análisis de sentimientos por cada reseña y se obtiene finalmente una segmentación de tipo de cliente, para crear un nivel de prioridad o urgencia de atención.

Fase 4.- Mediante un orquestador de tareas se ejecutará scripts en Python los cuales gestionan todo el proceso del proyecto desde la llamada de API de los registros de QuestionPro para guardarlos en una base de datos analítica propia (EXASOL), después su respectiva transformación necesaria y predicción de los resultados de los algoritmos NLP, hasta finalmente la construcción del dataset final a ser usado para la visualización.

Fase 5.- Se elabora el dashboard de las alertas en Tableau, la cual realiza una conexión en vivo hacia la base de datos analítica (EXASOL) que guarda los resultados de las nuevas encuestas junto con su predicción. Dicho dashboard será cargado a Tableau Server, en el cual lo podrán visualizar sólo las personas pertinentes, con sus credenciales respectivas, y mediante una conexión a la red interna/privada de la empresa.

Fase 6.- Se remiten las alertas mediante Slack, hacia los canales respectivos de comunicación mediante una integración con Tableau, es decir cada vez que llegue una nueva reseña de cierto tipo, será notificado al cliente interno por

Slack, con la opción de que pueda abrir Tableau Server para ver más detalle de esta.

Fase 7.- Se realizará un posterior seguimiento del desempeño de la herramienta y la precisión de este, en una prueba piloto con la sección de Customer Experience.

1.6 Resultados Esperados

En este proyecto se espera finalizar con el desarrollo de un dashboard interactivo en el servidor web propio de la empresa, el cual podrá ser consultado por las partes interesadas, de forma que tengan acceso a los resultados de los modelos de NLP, así como información adicional del cliente, para que puedan complementarla con los resultados de los indicadores de Satisfacción de Calidad y así proceder con planes de acción más precisos. Todo esto va a llevar a que la empresa “AAA” pueda conocer más a sus clientes, gestionar mejor sus problemas, mejorar la satisfacción del cliente, generar mayor captación de clientes nuevos al tener mayor recomendación de los clientes actuales, prevenir y disminuir las deserciones de clientes a otros proveedores, gestionar soportes de manera más efectiva y por último un mayor crecimiento de la empresa de forma global.

El dashboard interactivo será uno de los resultados que genere un valor agregado al proyecto y se desea implementar como una herramienta de visualización que proporcione a las partes interesadas un análisis 360 de los distintos aspectos del cliente, obteniendo así un detalle más profundo de la situación del cliente y sus perfiles en distintos ámbitos que serán de mucha ayuda para la toma de decisiones con respecto al mismo, dicha herramienta constará de distintos módulos los cuales serán desarrollados con el fin de obtener información relevante para la empresa.

Se podrá acceder a esta herramienta mediante las alertas de Slack y una integración con Tableau, a continuación, se muestra un prototipo.

Figura 1.2 Prototipo Herramienta de Visualización
Fuente: Elaboración Propia

1.7 Dataset:

El Dataset para el desarrollo de este proyecto está conformado por varias fuentes de datos, estas fuentes de datos en su gran mayoría son propiedad de la empresa, pero también se hace uso de fuentes de datos externas ya que, para obtener toda la información necesaria, se requiere relacionar varias fuentes de datos simultáneamente.

Las fuentes de datos que han sido usadas para este desarrollo son las siguientes:

Fuente de Datos	Variables	Descripción Campos	Tipo de Datos
Repositorio Interno de Llamadas de Encuestas de Calidad	Audio de la Llamada	Archivos no estructurados que contienen el audio de la llamada de la encuesta de Calidad, dichos archivos son sacados de un sistema externo de llamadas, estos al ser descargados son guardados en el repositorio de la empresa. Dichas llamadas dependiendo de su peso, pueden estar particionadas en dos.	.aac, .mp3, .wav, .m4a, .mpeg, etc.

Repositorio Interno de Llamadas de Encuestas de Calidad	Metadata del Audio de la Llamada	Información adicional de la llamada de la encuesta, como nombre del operador que realizó la llamada, hora de la llamada, cliente al que se contactó, tiempo de duración la llamada, etc.	.csv
Fuente de Datos Generales de Clientes	Variables Demográficas de los Clientes de la empresa "**AAA**"	Se tiene como ejemplo en esta fuente campos como Antigüedad, RUC, Nombre, Ubicación, Fecha de Creación, Tamaño, etc.	varchar (query)
Fuente de Datos Técnicos de Clientes	Variables de Soportes de los Clientes, como los casos generados en el tiempo, los servicios y productos activados actualmente, etc.	Se tiene como ejemplo en esta fuente campos como Numero Caso, Fecha Creación Caso, Estado Caso, Fecha Finalización Caso, Login Afectado, ID Servicio, Nombre Producto, Estado Servicio, Nombre Vertical, etc.	varchar (query)
Fuentes de Datos Comercial de Clientes	Variables de Facturación y Cobranzas de los clientes de la empresa "**AAA**"	Se tiene como ejemplo en esta fuente campos como Numero Factura, Venta, Producto, Id Servicio, Nombre Cliente, Nombre Vendedor, etc.	varchar (query)
Superintendencia de Compañías – Portal Web	Balances Financieros de las Empresas del Ecuador	Cerca de 800 variables financieras que resultan del formulario 101 del SRI, donde se reporta el Balance General, el Estado de Resultados, entre otros. (Ejemplo: Activos, Pasivos, Patrimonio, Ingresos, Gastos, Costos, Utilidad, etc.)	.csv
QuestionPro	Encuestas de Satisfacción del Departamento de Calidad	Respuestas a las preguntas de la Encuesta de Satisfacción y Comentarios de los clientes.	.xlsx

Tabla 1.1 Descripción del Dataset
Fuente: Elaboración Propia

CAPÍTULO 2

2 ESTADO DEL ARTE

2.1 Procesamiento de Lenguaje Natural

La mayor parte del conocimiento que se genera diariamente en el mundo no ha sido explorado por el ser humano, gran cantidad de estudios y análisis en la industria abarcan en su mayoría datos estructurados, los cuales son fácilmente tabulados para su entendimiento, mientras que los datos no estructurados como textos de documentos, correos, audios, videos e imágenes, no han sido explotados en todo su potencial, por el simple hecho de que se necesita un mayor trabajo y conocimiento de pre procesamiento, así como los recursos informáticos para llevar a cabo un trabajo de este tipo.

Una importante empresa de tecnología como ForestRim en colaboración con Databricks indica en un estudio, que entre el 80 y 90% de data que maneja una empresa son datos no estructurados (Inmon, Levins, Kermany, & Sanky, 2021); así mismo, Computer Generated Solutions Inc. (CGS) una empresa norteamericana en sus estadísticas del 2022 encontró que el 80% de consumidores se sienten conectados emocionalmente con una marca, cuando esta entiende sus problemas (Customer Experience Stats, 2022); por estos motivos, dentro del Análisis de la Experiencia de Cliente ha tenido un impacto importante el Procesamiento de Lenguaje Natural, por sus diversas aplicaciones dentro de este campo.

Este estudio trata sobre el procesamiento del lenguaje natural de reseñas de clientes en una encuesta de satisfacción de calidad, en una empresa de telecomunicaciones ecuatoriana, pero para llegar a dicho punto, es necesario una base de fundamentos teóricos que serán expuestos en el presente capítulo.

2.1.1 Definiciones

La definición de Procesamiento de Lenguaje Natural ha existido desde hace muchas décadas, remontándolo desde la década de los 50s, como una intersección entre la inteligencia artificial y la lingüística (Nadkarni, Ohno-Machado, & Chapman, 2011); aunque no con el término de Procesamiento de Lenguaje Natural per sé, dado que dicho término ha tenido mayor relevancia desde aproximadamente el año 2010, abarcando un conjunto mucho más amplio de técnicas y aplicaciones, antes de dicho año cuando se referían al análisis estadístico de texto, el término usado era Minería de Texto, o Análisis de Texto.

Pero, en definitiva, ¿Qué es el Procesamiento de Lenguaje Natural o NLP (por sus siglas en Inglés)?, Chowdhary lo define en su libro de Fundamentos de Inteligencia Artificial como, “Una colección de técnicas computacionales para el análisis automático y representación del lenguaje humano, motivado por la teoría" (Chowdhary, 2020), pero deja muy en claro, que dicho trabajo de realizar un entendimiento profundo del lenguaje por parte de las máquinas, a la par con los seres humanos, está lejos de la realidad, existen tareas como la extracción de información, el análisis de bajo nivel de las palabras, entre otras, en las cuales el NLP es muy eficiente, pudiéndose ubicar a la par de un ser humano o incluso mayor, pero hay otras tareas como el entendimiento de oraciones o el significado a un bajo nivel de un texto, en las cuales el ser humano está por encima de la máquina en la actualidad.

La complejidad de entender el lenguaje natural radica en su mayor parte en la longitud de los textos a analizar, dado que entender el contexto de un documento o alguna conversación, implica acumular muchas oraciones que guardan relación con la oración actual, donde también puede haber ambigüedad en el lenguaje, tener palabras que pueden significar la misma cosa por los distintos modismos del lenguaje a analizar, o por lo contrario tener ciertas palabras que puedan tener muchos significados. Teniendo en cuenta todos estos tipos de problemas que existen es importante entonces tener alguna forma de poder representar un texto, y que una máquina lo pueda

entender en una forma más estructurada, y a su vez reducir el ruido generado por palabras que no proporcionan un gran significado al texto o que corresponden a palabras no representativas del contexto del contenido, por lo que es necesario un preprocesamiento de la data.

Preprocesamiento de Textos para NLP

El preprocesamiento del texto es un proceso de múltiples pasos, el cual tiene como objetivo transformar el texto original del documento o reseña, a una representación de datos que pueda usarse como entrada de un algoritmo de ML, dicho preprocesamiento consta de pasos que se realizan de forma secuencial, la cantidad de pasos y el orden de estos podría variar según el autor a consultar, pero en su gran mayoría describen los siguientes:

Figura 2.1 Preprocesamiento de Textos para NLP
Fuente: Elaboración Propia

Tokenización

La tokenización permite representar largas cadenas de texto, los cuales son datos no estructurados, en unidades más simples llamados tokens, estos pueden ser oraciones simples de todo un texto, también pueden ser palabras de un texto, entre otras. El token formalmente se define como un grupo de caracteres que tienen un significado (Torres, 2013), es decir que representan unas características dentro de los datos.

Eliminación de Caracteres Inválidos

Este proceso consta de eliminar los caracteres especiales innecesarios dentro del lenguaje que estamos analizando, con la finalidad de que pueda servir como input para una máquina, y no tenga problemas al leer los datos (por problemas de codificación), como por ejemplo dentro del idioma español, podría ser los signos de puntuación de las palabras (tildes), la virgulilla, diéresis, o caracteres de admiración, entre otros.

Eliminación de Stopwords

Dentro del análisis de lenguaje natural, existen palabras que pueden causar un ruido innecesario dentro del modelo NLP, o que en general no aportan un valor al estudio del texto, esas son las stopwords. Las stopwords son un listado de verbos auxiliares, artículos, conjunciones y términos que pueden obviarse dentro del estudio. Dichas palabras son las que aparecen con una mayor frecuencia en los datos de texto que se analizan, como por ejemplo, “la”, “los”, “el”, “que”, o a su vez, términos que no queremos que se analicen como por ejemplo, el nombre de la empresa en la cual se está analizando las reseñas de los clientes, dado que es una de las palabras que más utilizan los clientes en sus expresiones, y que generalmente no agrega un valor relevante al estudio. Decidir qué palabra es incluida dentro de dicho listado no es tan trivial como parece, dado que depende mucho del dominio del problema a analizar.

Stemming

El término stemming hace referencia a un proceso de reducción de las palabras a su forma raíz o canónica llamado stem por su término en inglés, dicha forma es la base para la construcción de la palabra, por ejemplo, si se habla de consultoría, consultor, consulta, el stem o base canónica del mismo resultaría en consult. Uno de los algoritmos más utilizados para este proceso es el Algoritmo de Porter. (M.F.Porter, 1980)

La finalidad de este proceso es reducir / agrupar el listado inicial de tokens a una forma en la cual, distintos términos expresen el mismo significado, logrando eliminar hasta palabras mal escritas y también reducir el tiempo de procesamiento cuando se requiera entrenar un modelo de NLP, al tener un listado mucho menor.

Representación Vectorial

Una vez obtenido un listado depurado de características de un texto en particular, es necesario que dicho conjunto de palabras tenga una estructura adecuada para que una computadora lo pueda leer como entrada de un algoritmo de aprendizaje de máquina. Para este fin, se utiliza una representación vectorial de los tokens.

Un documento, texto o reseña es representado por el vector r_j , donde j va desde 1 hasta n, siendo n la cantidad total de reseñas en el estudio, dicho vector es una colección de pesos por cada término en la reseña o texto a analizar, por lo cual sería:

$$r_j = [\, \omega_{1j}, \omega_{2j}, \dots, \omega_{mj}]$$

Siendo m la cantidad total de términos o características que aparecen en los documentos, que en este caso particular representan las palabras de las reseñas con toda la depuración previa, que por lo menos haya aparecido una vez; además ω_{ij}, representa el peso de contribución que tiene la palabra i en la reseña j, dicho peso puede ser calculado de distintas formas, por ejemplo una de las formas de realizar esta tarea es dado un peso binario de 0 o 1, dependiendo si la palabra i se encuentra dentro de la reseña j, independientemente cuentas veces haya aparecido; otra forma de realizar esta tarea es creando pesos que representen la frecuencia nominal u absoluta de la palabra i dentro de la reseña j. En definitiva, en este paso se tiene una gran parte del preprocesamiento de los datos, el cual podría estar listo como input para un modelo de aprendizaje automático, pero se podría mejorar el rendimiento de este, al evitar un gran problema en muchos de los modelos ML en la vida real, como lo es el Overfitting.

Smoothing

Dentro del procesamiento de lenguaje natural, así como en cualquier otro modelo de aprendizaje automático existe un problema que puede aparecer, como lo es el sobreajuste (overfitting) en los datos, que en términos generales es que el modelo aprende mucho de los patrones de los datos de entrenamiento, y por lo tanto se hace muy bueno para realizar predicciones del training set, pero no de datos nuevos, es decir, no es bueno generalizando. En el caso del procesamiento de lenguaje natural, al tener un vector de características, el cual puede ser muy grande por cada uno de los textos, puede existir un sobreajuste por parte de los modelos, al aprender patrones específicos que al final de cuentas son solo ruido, por parte de las expresiones que utilizan la personas en su diario vivir. Para evitar este tipo de problemas y

que el modelo tenga un mayor rendimiento, por ende, pueda generalizar con nuevos registros en los distintos modelos entrenados, se utiliza una técnica de smoothing, el cual, como su nombre lo dice, permite suavizar ese vector de características, en este proyecto, se recorrerá dos técnicas en particular:

- Selección de Características
- Transformación de Características

Selección de Características

La selección de características hace referencia a una afectación directa a los pesos de los términos ω_{ij}, en los cuales, al realizar ciertas operaciones matemáticas a dichos pesos, puede mejorar que un modelo discrimine mejor entre términos que realmente son importantes en la reseña, de aquellos términos irrelevantes que aparecen en la reseña, pero no guardan un sentido semántico que ayude en el objetivo final del estudio, al ponderar con un valor mayor a aquellas palabras relevantes, el modelo se enfocará en estas. Dentro de las principales técnicas de este proceso, se encuentran los siguientes:

1. **DF (Document Frequency)**

   Siendo esta la más sencilla de las nombradas en el presente documento, el DF corresponde al número de reseñas (documentos) que contienen una palabra en particular, por ejemplo, si de 100 reseñas, en 20 de ellas está el término “problema”, entonces el DF de problema es 20 de forma nominal, o 20/100 de forma porcentual.

2. **IDF (Inverse Document Frequency)**

   A diferencia del primero, este indicador permite crear una mayor puntuación a los términos o palabras que no tienen una gran frecuencia de ocurrencia, mientras que las palabras que aparecen en muchas de las reseñas o documentos tendrán una puntuación baja, la forma de cálculo es la siguiente:

$$IDF_i = \log(\frac{n}{1 + df_i})$$

Siendo n la cantidad total de reseñas a entrenar y df_i la cantidad de documentos donde aparece la palabra i. Para reducir la cantidad de características, se suele ordenar de manera global la puntuación obtenida por el IDF, y se selecciona los términos que tengan una mayor puntuación, con la finalidad de encontrar patrones en aquellas palabras que no son tan frecuentes.

3. **TF (Term Frequency)**

   TF es la más conocida de todas las nombradas, la cual corresponde al número de apariciones de una palabra dentro de una reseña o documento en particular, siendo el cálculo el siguiente:

$$TF_{ij} = \frac{k_{ij}}{m_j}$$

   Siendo k_{ij} el número de veces que aparece la palabra i en la reseña j, y m_j la cantidad total de palabras que existen dentro de la reseña j. Dicho indicador nos representa una frecuencia relativa de las palabras, donde si la puntuación es más alta, entonces dicha palabra ha tenido mayor ocurrencia y además es más relevante para el modelo.

4. **TFxIDF (Term Frequency x Inverse Document Frequency)**

   Este método realiza el producto de dos métodos antes vistos como lo es TF e IDF, resultando en lo siguiente:

$$TF - IDF = TF_{ij} * IDF_i$$

   Esto representa un punto de vista distinto a los comunes, en la cual no sólo ve la frecuencia que tiene las palabras dentro de un documento, sino que también realiza un ponderación por las puntuaciones de las palabras más “raras” o menos frecuentes entre todos los documentos, esto quiere decir que el modelo no se enfocará netamente en las palabras que representen una mayor frecuencia, que a veces esto puede ser que no contenga tanto valor informativo, sino también de aquellas palabras especiales que se

encuentran dentro de todo el documento, con el objetivo de determinar mejores patrones en las reseñas de los clientes.

Transformación de Características (Reducción de Dimensionalidad)

Otra de las metodologías que permiten una reducción de las características del vector de tokens, es la transformación de estas, con el objetivo de no afectar a los pesos, sino más a la dimensionalidad de los vectores. Un claro ejemplo de esto es el análisis de componentes principales o PCA por sus siglas en inglés, la cual permite reducir la dimensión de los datos a través de todo un proceso que parte de la matriz de varianzas y covarianzas, hasta llegar a los componentes principales con los eigenvalores y eigenvectores, a coste de perder un poco de la variabilidad de los datos originales, con el objetivo de tener una buena representación de los mismos pero con una dimensión menor. A su vez cuando el vector transformado pase como input en el modelo de ML, este podrá generalizar mejor, dado que no contiene la data tan específica del conjunto de entrenamiento, sino una representación buena de los mismos.

En el presente estudio no se incluye literatura sobre PCA, dado que no se utilizará en el proyecto, pero si hay que tener en cuenta la importancia de este, a la hora de depurar el vector de características y tener un mejor rendimiento en el modelo.

2.1.2 Beneficios del NLP

El procesamiento de lenguaje natural tiene un amplio rango de aplicaciones y usos dentro de la industria y la academia, dependiendo de cuál se explore, el beneficio puede variar, pero de manera general se han enlistado los siguientes:

- Automatización de procesos que antes se realizaban de forma manual por operadores, lo que ahorra tiempo y costos para la empresa, si antes se destinaba mucho tiempo analizando ciertos textos o categorizando a los mismos, esta tarea podría ser automatizada.

- Añade un valor significativo a los datos que antes no se utilizaban, como lo son los textos de documentos, correos, reseñas, entre otras. Los datos son un activo para la empresa, y como activo, tiene un valor financiero cuantificable, si el dato no se está utilizando, no está generando un valor para la empresa; por lo tanto, al comenzar a analizar los textos de los documentos, no es un dato en desuso, no sólo se está guardando en un repositorio, sino que está generando información para el mismo funcionamiento de la empresa.
- Ayuda en la toma de decisiones por parte de la alta dirección de la empresa, dado que cómo se ha descrito en el comienzo del proyecto, por ejemplo, a veces analizar indicadores que están sujetos a la percepción de la persona, podría llevar a resultados sesgados. Las técnicas de NLP permiten tener un poco más de contexto, e incluso, a partir de indicadores cuantitativos, inferir, las circunstancias por las cuales una persona dijo o calificó cierto tema.
- Es un gran paso para la investigación, dado que cada vez más la inteligencia artificial trata de recrear las capacidades del ser humano en su diario vivir, el lenguaje es una parte fundamental de este objetivo, mientras más avance la investigación de NLP, mejor serán los chatbots creados, los robots interactivos, entre otras. Se espera un futuro brillante en esta área, pero que claramente no va a reemplazar al ser humano en sus capacidades lingüísticas y cognitivas en su totalidad.

2.1.3 Evolución del NLP

Muchos investigadores datan el origen del procesamiento de lenguaje natural en la década de los 50, cuando se creaban los primeros trabajos e investigaciones de análisis de texto basado en reglas explícitas, el más famoso de ellos, la prueba de ensayo denominada “Computing Machinery and Intelligence” (Turing, 1950), en la cual se proponía evaluar conversaciones entre un humano y una máquina. Uno de los primeros hitos reconocidos fue la traducción automática de más de 60 oraciones rusas al inglés en el experimento de Georgetown en 1954, época donde se esperaba que evolucione rápidamente este campo, lo cual no fue así, pero igualmente se hicieron

grandes investigaciones que propusieron un gran avance en el lenguaje natural. Después de la década de los 60, el NLP se enfocó en su mayor parte por el análisis de la sintaxis de los textos, lo cual era la mayor necesidad en la mayoría de las aplicaciones en su momento, posteriormente con el tiempo tomó un rumbo hacia la semántica del texto. Dentro de las primeras tareas del NLP, estaban la traducción, la recuperación de información, resumen de textos, entre otros; una de las técnicas más usadas de NLP era el FOPL (First Order Predicate Logic), el cual usa reglas predefinidas para construir predicados en las oraciones, y así nacieron con el tiempo varias técnicas basadas en reglas como Default Logic (Reiter, 1980), Production Rules (Chomsky, 1956), que trataban al lenguaje como un conjunto de reglas a seguir, pero que guardaban sentido con la sintaxis, semántica, incluso la pragmática.

Con el pasar de los años, llegando a la década de los 80 y con la evolución de los algoritmos de aprendizaje automático, se dio un enfoque distinto al análisis de textos. Este enfoque no es basado en reglas, no está programado explícitamente, sino que basado en una representación de los textos, donde la máquina pueda aprender de los patrones de estos, y utilizarlos posteriormente para predicciones, segmentaciones, etc. Una vez llegado a este punto el procesamiento del lenguaje natural logra obtener una gran relevancia en el mundo, por sus distintas aplicaciones y beneficios que puede otorgar. Además de ampliar el interés de este campo en la comunidad científica, generando mayor investigación y aportes en el NLP.

2.1.4 Componentes del NLP

Dentro de la lingüística hay dos grandes áreas, las cuales son la lingüística computacional y la lingüística teórica, donde la primera trata de los distintos algoritmos para el manejo del lenguaje natural, mientras que la segunda, trata como su nombre lo dice los aspectos teóricos de la lingüística, como el desempeño del lenguaje, como las personas aceptan ciertas expresiones, siguiendo ciertas reglas gramaticales. A su vez existen otras subdivisiones que se van a tratar a continuación:

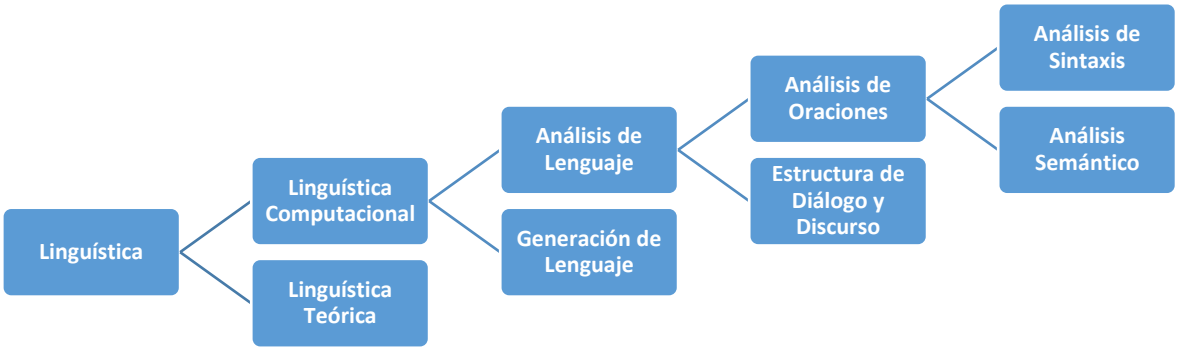


Figura 2.2 Componentes del NLP
Fuente: Elaboración Propia

Análisis de Lenguaje

Dado un texto, reseña o texto, analizar el contenido de dicha data estructurada en específico, esta se puede separar en dos, análisis de oraciones y estructura de diálogo.

Análisis de Oraciones

El análisis de oraciones se enfoca como su nombre lo dice en entidades individuales de un texto completo, como una oración, sin tener una relación semántica de lo que tiene que ver con una oración del algún párrafo pasado. Dicho análisis se puede separar en dos, el análisis de la sintaxis, y el análisis semántico de las oraciones.

Análisis de Sintaxis

Análisis orientado a examinar la estructura de las oraciones, como determinar el sujeto, predicado o verbo de una oración, y cómo puede afectar el cambio de una palabra a la estructura de esta.

Análisis Semántico

Análisis orientado a examinar la interpretación de lo que significa la oración, quitando las posibles ambigüedades, proporciona una inferencia lógica del tema guiado por la forma en la cual está escrita la oración.

Estructura de Diálogo y Discurso

Análisis orientado al análisis de un texto completo y no por partes (oraciones), encuentra el sentido y relación que existe entre distintas oraciones de un texto

completo, es un análisis mucho más complejo debido a que se guarda una relación secuencial del lenguaje natural.

Generación de Lenguaje

Análisis orientado a la creación de texto nuevo a partir de distintos textos explorados, este tipo de área ha tenido mucha relevancia en los últimos años, al tener inteligencias artificiales que crean libros de cierta temática desde cero, o predicciones de texto muy precisas a partir de un buen conjunto de entrenamiento.

2.1.5 Modelos de Clasificación

El aprendizaje automático supervisado es una rama del aprendizaje automático que tiene como objetivo que la máquina aprenda los patrones ocultos de la data para obtener predicciones de resultados previamente etiquetados, todo esto sin ser programados explícitamente. Entre los distintos algoritmos de aprendizaje automático que existen, se implementarán alternativas usando los siguientes:

Máquina de Vectores de Soporte o Support Vector Machine

La máquina de vectores de soporte o SVM por sus siglas en inglés, es uno de los algoritmos más famosos que existe dentro del aprendizaje automático supervisado (Cortes & Vapnik, 1995), el cual tiene como objetivo separar distintos grupos de puntos de datos a través de un hiperplano de dimensión N, dicho hiperplano que encuentra la mejor separación entre dos o más grupos distintos de datos, junto con los diferentes puntos de datos más cercanos al hiperplano conforma lo que se llama vectores de soporte, todo esto minimizando una función convexa de costo, medido por el margen de una mejor clasificación (Chowdhary, 2020)

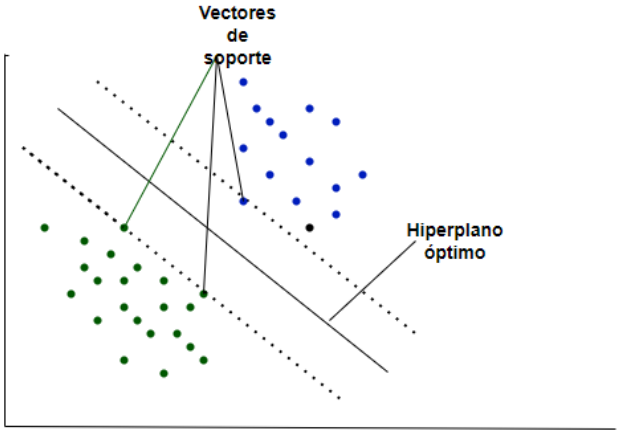


Figura 2.3 Máquina de Vectores de Soporte o Support Vector Machine
Fuente: Elaboración Propia

Originalmente se concibió al SVM como un clasificador binario, es decir, aquellos puntos de datos que están de un lado del hiperplano se le asignaba un valor 0 y los que estaban del otro lado, un valor 1, pero con el tiempo se creó un clasificador múltiple que genere varios hiperplanos a la vez en la nube de datos, para tener distintas categorías, o a su vez crear categorías binarias anidadas con un solo hiperplano. Una de las ventajas que se puede obtener de este tipo de modelo, es la aplicación de funciones de kernel dentro de la nube de datos, dado que, en la vida real, la mayor parte de las nubes de datos no son lineales, no se comportan de dicha forma, por lo que es necesario realizar una transformación primero y luego el hiperplano que los separe de forma óptima. Dentro de las funciones kernel que existen y son famosas por su aplicación están:

- Lineal
- Polinomial
- Función de Base Radial (RBF) o gaussiano
- Sigmoidal

Random Forest

El Random Forest es una técnica del aprendizaje automático supervisado (Breiman, 2001), el cual utiliza múltiples árboles de decisión, con el objetivo de obtener la predicción con mayor frecuencia. El valor de la cantidad de árboles a utilizar queda a elección del investigador. Los árboles de decisión de forma individual tienen como objetivo obtener reglas basadas en las características propias de los datos, en el cual divide de forma binaria ramas lógicas en cada paso o iteración, partiendo de alguna variable y siguiendo con scores definidos como la entropía del modelo, entre muchas otras. En el caso de Random Forest, cada árbol parte de un subconjunto de características, con el fin de combinar los resultados de todos los árboles en uno solo (en el caso de clasificación se tiene la clase con mayor frecuencia), siendo este tipo de modelo más robusto, a diferencia de tener solamente un árbol de decisión, el cual puede sobreajustar mucho a los datos de entrenamiento y no generalizar bien los patrones.

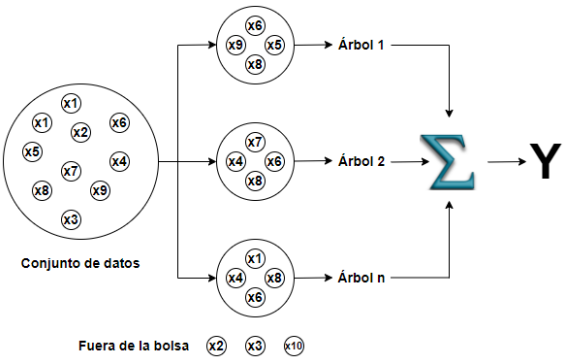


Figura 2.4 Random Forest
Fuente: Elaboración Propia

Este tipo de modelo fue considerado por muchos años un gran avance en el ámbito del aprendizaje automático, pero con el tiempo se hicieron algunas variaciones y se consiguió mejorar aún más el rendimiento de los árboles de decisión, uno de esos avances fue el algoritmo de XGBoost.

XGBoost

El Extreme Gradient Boosting es una técnica de aprendizaje automático supervisado (Chen & Guestrin, 2016), el cual, así como el Random Forest utiliza árboles de decisión en su estructura principal, pero lo hace de una forma distinta, de manera secuencial, es decir el resultado de un árbol de decisión es el input para el siguiente árbol de decisión, tiene en cuenta también el error generado al final de cada árbol, con el objetivo de minimizar el error en cada paso por el gradiente generado. A diferencia del Random Forest donde uno tenía control de la cantidad de árboles generados, en este caso la extensión de árboles no está a control del usuario, sino que se extiende hasta que el error sea el mínimo posible o a su vez haya pasado un umbral. Dicho algoritmo es uno de los principales estados del arte actuales, y que se está trabajando mucho en la industria.

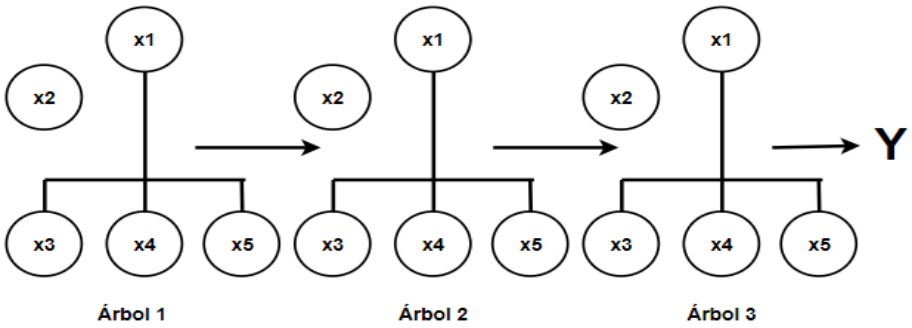


Figura 2.5 XGBoost
Fuente: Elaboración Propia

Multilayer Perceptron (Redes Neuronales)

Las redes neuronales artificiales forman parte del campo de la inteligencia artificial, estas redes tratan de simular la estructura y el funcionamiento de un cerebro y son utilizadas para dar solución a todo tipo de tareas que van desde las más sencillas hasta las más complejas. El Perceptron multicapa forma parte del conjunto de redes neuronales artificiales, fue desarrollado en la década de los 60 para resolver problemas de patrones e interpolación.

Se puede definir a las redes neuronales artificiales como conjuntos de neuronas organizadas en capas. Esta arquitectura recibe una entrada, y a la través de las capas intermedias y pesos que se aprenden en el proceso, generan una salida. Estas redes aprenden encontrando los valores que serán asignados como pesos para cada capa correspondiente. (Noriega, 2005)

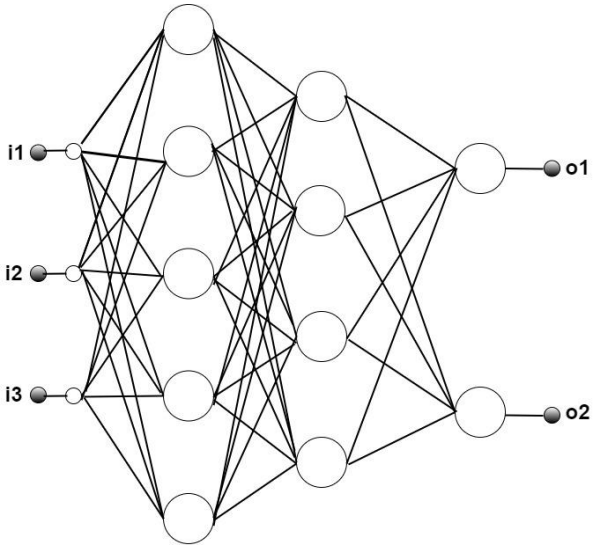


Figura 2.6 Multilayer Perceptron (Redes Neuronales)
Fuente: Elaboración Propia

Ensemble Models

Este es un enfoque del aprendizaje automático, el cual tiene como objetivo combinar distintos resultados de diferentes algoritmos de aprendizaje automático, para que la precisión de los resultados sea mayor a tenerlos por separado. Adicionalmente este enfoque realiza un modelo mucho más robusto y que puede generalizar mejor. En la última etapa de un Ensemble Model, se realiza un proceso de agregación de resultados, que en el caso de regresión podría ser el promedio de las predicciones, o en el caso de clasificación la clase de mayor frecuencia. Entre las principales técnicas de este tipo se encuentran las siguientes:

Bagging

Dicha técnica utiliza conjuntos de datos ligeramente distintos que son entrenados en distintos modelos, con el fin de obtener resultados que al final se agreguen y generen una sola predicción, para que generalice mejor y reduzca el error general; uno de los modelos vistos que siguen esta línea, es el Random Forest.

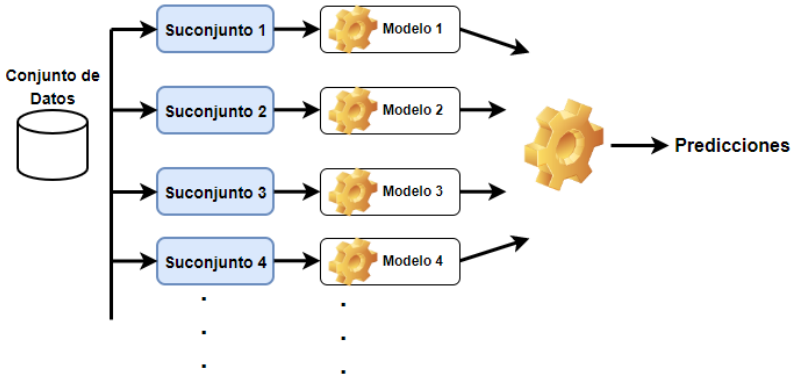


Figura 2.7 Bagging
Fuente: Elaboración Propia

Boosting

Este tipo de metodología sigue un orden secuencial donde cada modelo genera un resultado con un error asociado y este servirá como input para otro modelo para reducir el error, así mismo hace al modelo más robusto y con una mayor generalización; uno de los modelos de este tipo es el XGBoost.

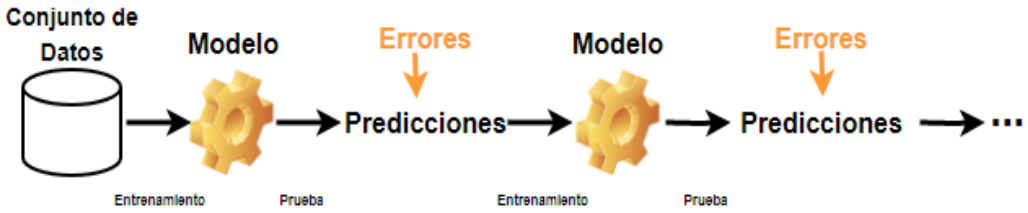


Figura 2.8 Boosting
Fuente: Elaboración Propia

Stacking

Esta técnica se enfoca así mismo en la combinación de modelos, pero que a diferencia de Bagging, no usa subconjuntos del dataset principal. En stacking, los modelos que contribuyen pueden ser distintos, por ejemplo, no todos van a ser árboles de decisión, pueden existir modelos de SVM, modelos de K-NN, etc. Adicionalmente, a diferencia de Boosting, no será de manera secuencial, sino que sólo será un modelo final que realice la combinación de distintos modelos mediante pesos, por ejemplo, pueden existir tres modelos, en los cuales la predicción de una regresión del SVM represente el 60% de la predicción final, mientras que un modelo de XGBoost represente el 20% y finalmente un modelo de regresión lineal múltiple represente el 20% restante de la predicción final.

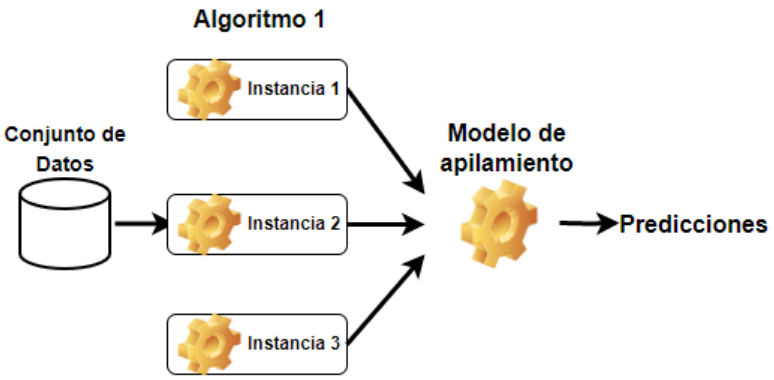


Figura 2.9 Stacking
Fuente: Elaboración Propia

2.1.6 Métricas de Evaluación

Conforme se va entrenando un modelo de clasificación va siendo necesario evaluar su desempeño y determinar qué tan bueno es. Existen diversas métricas para evaluar el desempeño de un modelo de Clasificación entre las cuales tenemos las siguientes:

Accuracy

Accuracy (exactitud) del modelo de clasificación es la tasa que existe entre las predicciones realizadas correctamente y el total de predicciones, es decir que tan seguido el clasificador lo hace correctamente, es posible calcular el accuracy mediante una matriz de confusión. Su fórmula viene dada por la siguiente ecuación:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision

Precision (la precisión) del modelo de clasificación es la tasa que existe entre las predicciones realizadas correctamente y el total de predicciones que han sido correctas. Por medio de esta métrica se obtiene la precisión del clasificador en los casos positivos. Su fórmula viene dada por la siguiente ecuación:

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall (Sensibilidad) del modelo de clasificación es la tasa que existe entre las predicciones positivas correctas y el número total de predicciones positivas, es decir por medio de esta métrica se obtiene la precisión del clasificador respecto de instancias positivas a lo cual se lo conoce como tasa de verdaderos positivos. Su fórmula viene dada por la siguiente ecuación:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

F1 Score (Puntaje F1) es una métrica de evaluación el cual combina dos métricas en específico como el recall y la precisión, con la finalidad de llevar un balance de ambas para escoger el mejor modelo. Su fórmula viene dada por la siguiente ecuación:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2.1.7 Aplicaciones del NLP

Debido a la gran importancia que ha tenido la Big Data dentro de las empresas y de la sociedad en general, cada vez hay mayor cantidad de datos en el mundo, y por ende mucha mayor información accesible que se puede obtener de la misma, se han realizado distintas técnicas que permiten explotar las ventajas de las soluciones relacionada al lenguaje natural, dentro de las muchas aplicaciones que existen en el NLP, podemos nombrar algunas como:

- Categorización de Contenido
- Reconocimiento de Entidades Nombradas
- Anonimización de Documentos
- Chatbots
- Traducción de Texto
- Análisis de Sentimientos y Opinión

2.1.7.1 Categorización de contenido

La categorización de contenido es una aplicación del procesamiento de lenguaje natural que tiene como objetivo realizar una clasificación de los textos o reseñas previamente etiquetadas, en base a la estructura de las oraciones representadas por el vector de tokens. Dicha aplicación tiene un gran rango de uso, desde la automatización de distintos procesos manuales, hasta la creación de sistemas nuevos que utilicen inputs como texto, por ejemplo, en el presente proyecto se desea crear una categorización automática de las reseñas de clientes de una encuesta de calidad, la cual permita analizar de forma oportuna dichas categorías como alertas. Una reseña nueva podría resultar ser una queja grave por parte del cliente (categoría de contenido), por lo que obtener esa categoría del texto lo más rápido posible, es fundamental para una empresa, dado que se podría separar los problemas importantes de los requerimientos comunes.

2.1.7.2 Reconocimiento de Entidades Nombradas (NER)

El reconocimiento de entidades es una aplicación de NLP, que permite detectar ciertas categorías en particular llamadas Entidades dentro de un texto, por ejemplo, si se tiene distintos textos de publicaciones de blogs, tweets o estados de alguna red social en particular, se podría detectar cuales están relacionadas con mascotas, o cuales están relacionadas con la actualidad política del país, entre otras; con la finalidad de que se pueda crear un etiquetado, que permitan ayudar por ejemplo a búsqueda personalizadas, o creación de tags automáticos, entre otras.

2.1.7.3 Anonimización de Documentos

Dentro de la seguridad de la información es de vital importancia asegurar la confidencialidad de datos personales de los usuarios, detalles privados de las empresas, secretos comerciales, entre muchas otras; por lo que el NLP puede llevar el reconocimiento de entidades nombradas a un siguiente nivel, con el objetivo de reconocer este tipo de tokens que necesitan mayor seguridad, y aplicarles un algoritmo de Anonimización, con la finalidad de cumplir con distintas regulaciones, proteger los intereses de la empresa, limitar las posibles

amenazas disminuyendo las posibles vulnerabilidades en los documentos de la empresa, etc.

2.1.7.4 Chatbots

Los chatbots es una aplicación dentro del NLP, que permite crear conversaciones virtuales con un bot de forma escrita o con un asistente de voz virtual, el cual ha aprendido de conversaciones anteriores, para poder remitir una respuesta que tenga sentido semántico según las preguntas que existan; también existen chatbots con reglas predefinidas en cuanto a las conversaciones, pero que no explotan las bondades que tiene los algoritmos de ML en la actualidad.

2.1.7.5 Traducción de Texto

Una de las aplicaciones más comunes que existe dentro del procesamiento de lenguaje natural es la traducción de textos, el cual consiste pasar el dominio, lenguaje nativo o idioma de un texto hacia otro tipo de dominio o lenguaje, por ejemplo pasar un texto de español a inglés; dicha tarea podría parecer que puede ser programado explícitamente con reglas predefinidas de relaciones entre palabras de distinto dominio, pero no es del todo correcto, dado que no sólo se traduce una palabra, sino una frase, un texto o un documento, en el cual las expresiones tienen que guardar sentido tanto de sintaxis como semántico en los distintos dominios que se utilizan, en otras palabras, el texto tiene que significar lo mismo en cualquier idioma que se traduzca, dicha tarea no es sencilla, por los distintos modismos que existen en las diferentes lenguas, por lo que es necesario una herramienta que aprenda a como se expresa ciertas ideas en los distintos idiomas, para que genere un resultado muy preciso.

2.1.7.6 Análisis de sentimientos y opinión

El análisis de sentimientos nace a comienzos del nuevo milenio como una solución básica, para determinar si un texto es positivo o negativo (Pang, 2002).

Para toda empresa, independientemente del sector al que pertenezca es indispensable conocer la opinión de sus clientes con respecto a los productos o servicios que estos ofrecen. En general, con lo que se cuenta es con opiniones escritas o reseñas de los productos y debido a ello, nace una aplicación dentro del NLP, que busca establecer la polaridad de los textos: positivo o negativo. Así mismo con el pasar del tiempo esta aplicación fue mejorando con nuevas técnicas y metodologías que aprovechaban los beneficios de los algoritmos de Aprendizaje Automático (Pang, 2008). En definitiva, este tipo de aplicación es muy necesaria cuando se tienen reseñas escritas de clientes, los cuales pueden dar un contexto general a la empresa si se está haciendo un buen trabajo o no.

2.2 Aprendizaje Automático No Supervisado

El aprendizaje no supervisado también denominado análisis de conglomerados y en otros casos descubrimiento de tareas, son algoritmos que trabajan de manera similar a los supervisados, pero los modelos no supervisados a diferencia del aprendizaje automático supervisado no reciben una variable de objetivo, ya sea numérica, binaria o categórica, es decir, no se les proporciona una salida a la cual deban ajustarse, por lo que no es posible realizar validación cruzada (Gentleman & Carey, 2008).

Este tipo de aprendizaje automático más bien busca realzar el valor de la información existente, y de los datos futuros provenientes de una situación similar, es decir solo se toman en cuenta los datos de entrada sin dar importancia a los datos de salida, dando más relevancia a las similitudes de las características entre los mismos datos existentes.

Usualmente en las técnicas de aprendizaje supervisado se presentan una serie de características y etiquetas para los datos, pero en caso de las técnicas de aprendizaje no supervisado se carece de las mismas, sin embargo, sin poseer etiquetas previas va a surgir la necesidad de segmentar de manera adecuada estos datos, y en este proceso se van a presentar similitudes en ciertas características, las cuales serán aprovechadas para poder predecir la clasificación de los nuevos datos. (Román, 2019)

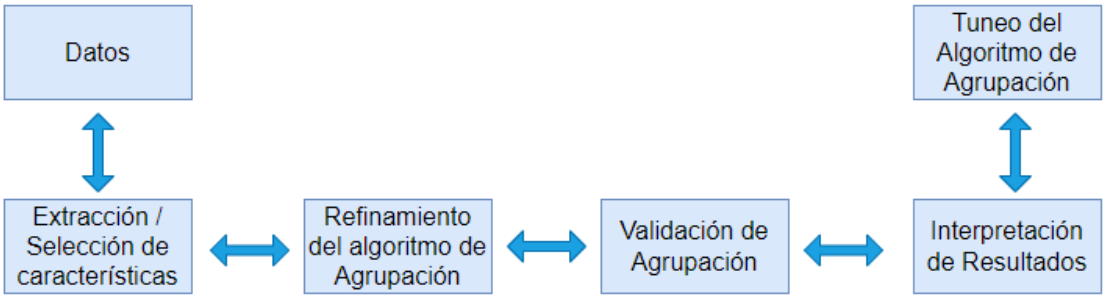


Figura 2.10 Aprendizaje Automático No Supervisado
Fuente: Elaboración Propia

2.2.1 K-Means

K-Means es un algoritmo dentro del aprendizaje automático no supervisado, es el método más sencillo para encontrar grupos de puntos relacionados y a su vez es el más conocido. Sin embargo, es un algoritmo cuyo rendimiento decrece cuando se busca identificar grupos en un dataset que carece de una distribución esférica (Román, 2019).

K-Means es un algoritmo iterativo cuya salida es un conjunto de grupos de los puntos del dataset en k grupos, de manera que cada punto i pertenezca a un grupo, específicamente a aquel cuyo centroide se encuentre más cerca del punto i. El valor óptimo de k para que el rendimiento del algoritmo sea más óptimo no es posible obtenerlo a priori, por lo que depende en su totalidad del conjunto de datos con el que se trabaje. Debido a que es un algoritmo no supervisado, no posee un conocimiento previo que permita saber la mejor manera de agrupar los datos disponibles, es decir que no existe la posibilidad de saber si la tarea va a ser o no bien ejecutada. Sin embargo, se pueden establecer parámetros subjetivos.

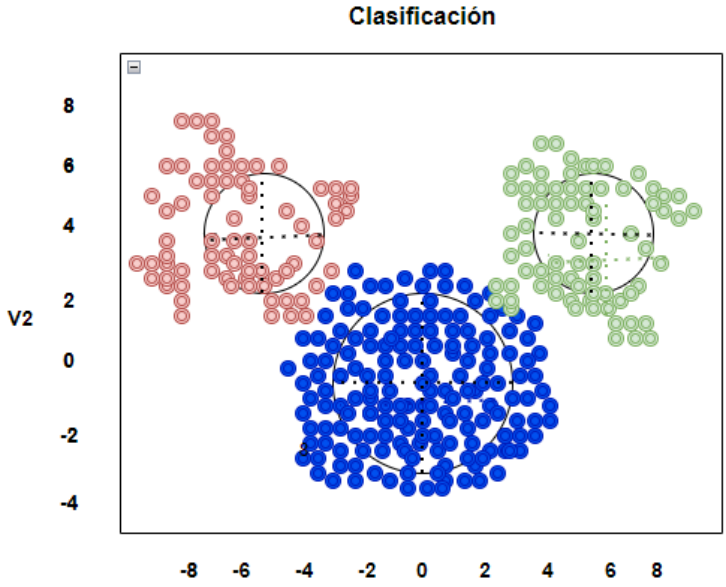


Figura 2.11 K-Means
Fuente: Elaboración Propia

2.2.2 Density-based spatial clustering of applications with noise (DBSCAN)

En el contexto del análisis de datos un tema muy importante es el análisis de agrupamiento o análisis de clústeres, entre las técnicas de análisis de clústeres DBSCAN es una de las técnicas con mayor popularidad.

DBSCAN (Agrupación espacial basada en la densidad de aplicaciones con ruido) es un algoritmo de agrupación de datos que se basa en la densidad de los datos. Este algoritmo tiene la habilidad de determinar la existencia de grupos que posean arbitrariamente cualquier forma y tamaño dentro de una base de datos incluso si la misma posee ruido y valores atípicos. (Khan, Rehman, Aziz, Fong, & Sarasvady, 2014)

El agrupamiento basado en densidad proviene de los métodos intuitivos del ser humano para agrupar, es decir se deriva de la identificación visual de grupos y el ruido alrededor de estos donde la densidad disminuye.
Los clústeres son regiones determinadas por la gran densidad de datos y que se encuentran separadas por regiones cuya densidad es menor.

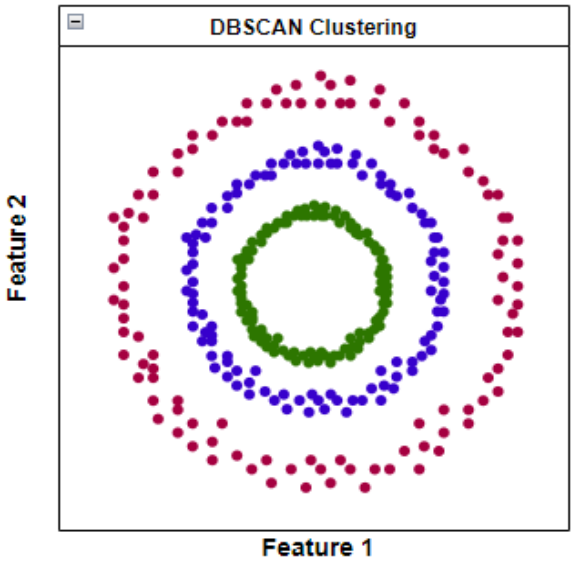


Figura 2.12 DBSCAN
Fuente: Elaboración Propia

2.3 Soluciones de analítica relacionadas al problema

Es evidente que la tecnología no se detiene y tiene avances precipitados y vertiginosos en diferentes ámbitos a diario, por lo que soluciones de analítica y aprendizaje relacionadas con el problema están a disposición frecuentemente como es el caso de “Rankings sociales: análisis de sentimiento visual en redes sociales” (Fernández, Gutiérrez, Gómez, & Martínez-Barco, 2015), el cual es una aplicación web que realiza en tiempo real seguimiento a entidades en las redes sociales. Mediante técnicas de análisis de sentimientos. Genera un informe de valoración en el que se puede observar su evolución en el tiempo de manera visual.

Las redes sociales en la actualidad y ya desde hace varios años han venido convirtiéndose en enormes repositorios de datos no estructurados, en estas plataformas las personas comparten frecuentemente sentimientos y opiniones sobre múltiples temas, productos y servicios. Debido a esto las organizaciones, el gobierno y las personas cada día aumentan su interés por analizar y estudiar estos sentimientos y mediante la investigación obtener información comercial valiosa para ellos, la cual les permita tomar decisiones basadas en datos de opinión y preferencias, (Wankhade, Rao, & Kulkarni, 2022) en su artículo “A survey on sentiment analysis methods, applications, and challenges”, analizan

un poco del estado del arte de las aplicaciones del análisis de sentimientos, comparando ventajas y desventajas de distintas metodologías.

Es conocido que para analizar estos sentimientos se han realizado anteriormente estudios con un enfoque en algoritmos de aprendizaje automático y procesamiento de lenguaje natural. Sin embargo últimamente han ganado popularidad los algoritmos basados en aprendizaje profundo por su capacidad y alto desempeño. En el estudio denominado “Sentiment analysis using deep learning architectures: a review” (Agarwal, Yadav, & Vishwakarma, 2019), se enfoca en el uso de las técnicas Aprendizaje Profundo con el fin de resaltar la habilidad y fortaleza que poseen estos algoritmos para resolver problemas de análisis de sentimientos.

Una de las aplicaciones que lleva el análisis de sentimientos hacia otras industrias se encuentra el estudio denominado “Sentiment Analysis of Customers Review Using Hybrid Approach”, en dicho estudio se realiza un análisis de sentimientos sobre un conjunto de datos de Amazon, más específico sobre las reseñas de los productos, al momento que los clientes lo usan, tratando de dar un resultado más preciso y fiable, comparándolo con las valoraciones puntuales de dichos clientes en su página. (Budwar & Singh, 2022).

Como un enfoque del análisis de sentimientos relacionado con la satisfacción del cliente (Ulloa Sepulveda, Rosales Ferreira, & Bermúdez Navarrete, 2018) realizaron el trabajo denominado “Propuesta de sistema de fidelización de clientes en la empresa de telecomunicaciones Avantel a través de indicador NPS (net promotor score) usando análisis de sentimientos en Facebook”, este trabajo consistió en generar un indicador mediante un estudio de fidelización de clientes, el cual sirvió para la obtener un indicador que se basa en un análisis de sentimientos de los comentarios de clientes y de terceras personas que siguen a la empresa y están suscritos en su página de Facebook, se usó este método con el fin de evitar por varios motivos el uso de formularios tradicionales. Para este desarrollo se implementó un modelo NLP usando técnicas probabilísticas que se basan en corpus para poder analizar los

sentimientos expresados en los comentarios de los usuarios.

La empresa “AAA” en la actualidad no realiza un análisis exhaustivo de las reseñas de los clientes, enfocándose en alertas generadas de manera subjetiva por los operadores telefónicos en la encuesta de satisfacción de calidad, cuando estos consideran que las reseñas que están expresando los clientes es algo de gravedad o de urgencia en atender, generando así una alerta mediante correo electrónico a las áreas pertinentes de la empresa que tienen que ver con la reseña; dicha gestión ha dado buenos resultados con el tiempo, pero se considera un poco obsoleta, porque depende mucho de la decisión del operador, además que no toma en cuenta otros factores o características del cliente en el momento de la alerta.

2.4 Herramienta para la generación de modelos (NLP)

Para este proyecto se ha decido hacer uso de lenguajes de programación de código abierto, la principal herramienta de desarrollo para este proyecto será Python con el intérprete de Jupyter Notebook (Anaconda), dado que al hacer uso de este tipo de lenguaje se tiene acceso a una gran comunidad que se encuentra siempre activa, y que constantemente está contribuyendo al desarrollo del código fuente a la par de las demandas y necesidades de los usuarios. En la actualidad y desde hace ya algunos años, Python se ha convertido en una herramienta muy demandada por muchas industrias debido a su versatilidad, su lenguaje sencillo y potente. Se utilizará para el proceso de preprocesamiento de los datos, entrenamiento de modelos, así como la automatización de las tareas, por ejemplo, el flujo automático de las predicciones de las nuevas encuestas.

Con respecto a las librerías a utilizar en este proyecto, se nombran a continuación las más relevantes a utilizar:

Scikit-Learn, es una librería dedicada al aprendizaje automático, esta librería ofrece Algoritmos de Clasificación, Regresión, Agrupación, Reducción de Dimensionalidad, Selección de Modelo, Preprocesamiento. (scikit-learn_ machine learning in Python — scikit-learn 1.0.2 documentation, s.f.)

NLTK, es una librería que permite el desarrollo de programas en Python para trabajar con datos del lenguaje humano. Esta librería proporciona interfaces de fácil uso y un conjunto de bibliotecas para procesamiento de texto para clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico. (NLTK __ Natural Language Toolkit, s.f.)

TextBlob, es fácil de usar se desarrolló encima de NYLK. Posee funcionalidades adicionales como por ejemplo análisis de sentimientos y corrector ortográfico. (TextBlob_ Simplified Text Processing — TextBlob 0.16.0 documentation, 2020)

Pandas, es una librería de alto nivel que sirve para realizar análisis prácticos de datos. (pandas - Python Data Analysis Library, s.f.).

2.5 Herramientas de Visualización de Inteligencia de Negocios

La inteligencia de negocios trata sobre el uso de metodologías y herramientas que ayudan a darle valor agregado a los datos y convertirlos en conocimiento de tal manera que sirva para que las organizaciones puedan tomar decisiones. En el proceso de generación de conocimiento se pueden considerar fuentes de datos tanto internas como externas.

Mediante las herramientas de inteligencia de negocios es posible extraer información, realizar análisis, procesamiento y reportería, de esta manera los datos servirán para alinear los esfuerzos corporativos hacia un mejor desempeño, determinar la existencia de deficiencias y poder obtener un factor diferenciador en la industria. Con este fin existen diversas aplicaciones y plataformas, los tipos principales son:

Gestión de datos: Su enfoque está dado en obtener información, procesarla y exportarla hacia otros sistemas.

Descubrimiento: Este tipo de aplicaciones están más enfocadas en la obtención de nuevos datos para su análisis y por medio de algoritmos de aprendizaje automático realizar pronósticos.

Reporte: Mediante este tipo de aplicaciones es posible desarrollar tableros de control para poder representar la información que pueda organizarse de presentarse de una manera clara y objetiva.

La extracción de información relevante para el negocio antes era exclusiva de los especialistas en analítica, sin embargo, mediante el uso de herramientas de visualización en la actualidad se ha democratizado esta importante actividad. Además, este tipo de herramientas ponen a disposición del usuario toda una gama de oportunidades, para desarrollar análisis profesionales y extraer conocimiento que contribuya al desarrollo, resolución de situaciones y recopilación de información.

Entre las herramientas de visualización más importantes se encuentran: Microsoft Power BI, Tableau, QlikView, SAP BI, Cognos Analytics, entre otras.

La herramienta de Visualización e Inteligencia de Negocios escogida para la implementación de la solución propuesta es Tableau, la cual es una de las herramientas más importantes del mercado.

Tableau

Tableau es una herramienta de Inteligencia de Negocios que fue creada como parte de un proyecto de estudiantes de Stanford, con el objetivo de mejorar el flujo de los análisis, y democratizar para los usuarios el manejo de datos mediante visualizaciones, esta plataforma permite comprender mejor los datos. (Tableau Software, 2003-2022)

Es una aplicación que soporta múltiples fuentes de datos y facilita la exportación y administración de estos, y a su vez es de gran ayuda para compartir de manera ágil información, que ayude en el giro del negocio y permita generar cambios. El diseño es muy versátil y está enfocado en los usuarios por lo que permite una fácil comprensión de los datos.

2.6 Herramientas de Trabajo Colaborativo

Las herramientas de trabajo colaborativo nacen de la necesidad que tienen dos o más personas de abordar una problemática con el fin de cumplir un objetivo en común.

Actualmente para poder hablar de trabajo colaborativo, indudablemente se debe tener en consideración una serie de herramientas tecnológicas que hace varios años no se encontraban disponibles, ya que, la sociedad se está desarrollando entre una amplia oferta de plataformas digitales, las cuales a medida que ganan popularidad aumentan su número de usuarios y a su vez el tráfico de información.

El trabajo colaborativo puede desarrollarse de forma presencial en espacios que presten las comodidades necesarias para que los miembros de un equipo puedan interactuar naturalmente y sin restricciones con el fin de proporcionar un ambiente en que puedan efectuarse actividades grupales que permitan la colaboración y ayuda mutuas.
También existen los entornos colaborativos virtuales por medio de herramientas tecnológicas que facilitan la interacción y permiten reunir a todo el equipo de involucrados en alguna actividad de tal modo que puedan comunicarse efectivamente y ser productivos en su trabajo desde cualquier lugar en el que se encuentren.

Hoy en día existen una serie de plataformas y aplicaciones con compatibilidad tanto móvil como de escritorio que son de mucha ayuda para el trabajo colaborativo, ya que, facilitan comunicaciones, ayudan a establecer y asignar responsabilidades en las actividades a efectuarse, calendarizan las fechas de entrega y facilitan la trazabilidad conjuntamente con otro tipo de acciones necesarias en esta modalidad de trabajo.

Para poder escoger una herramienta colaborativa adecuada es necesario considerar los factores clave en el trabajo y el tipo de producto o servicio ofertado por la empresa que va a implementar la solución.

Entre las herramientas de trabajo colaborativo más importantes se encuentran: Chanty, Fleep, Slack, Microsoft Teams, Google Chat, entre otras. Entre todo este conjunto de herramientas de trabajo colaborativo, para este caso particular se va a usar Slack.

Slack

Es una herramienta de mensajería corporativa mediante la cual se conectan las personas con la información que necesitan, es decir funciona como una sede digital, además facilita los entornos de trabajo virtual organizados por medio de la utilización de canales, los cuales se establecen de acuerdo al tema o tópico a tratar, esta herramienta proporciona una facilidad de interacción y comunicación por medio de chat, también permite la creación de grupos con niveles de acceso tanto públicos como privados, facilita definir fechas de entrega y la toma de decisiones.

2.6.1 Integración de Herramientas de Visualización con Herramientas de Trabajo Colaborativo

Tableau a partir de su versión 2021.3 puso a disposición de sus usuarios la integración con Slack, esto permite que para los usuarios las notificaciones de Tableau se encuentren disponibles en su espacio de trabajo de Slack. A partir de su activación los usuarios de Tableau pueden recibir notificaciones mediante Slack cuando un miembro del equipo les ha compartido contenido, al ser etiquetados o en el momento que, en base a los datos se genera una alerta predefinida. Cuando un administrador de Tableau activa las notificaciones los usuarios tienen la facilidad de configurar las notificaciones que desea recibir en Slack.

Esta funcionalidad permitirá al proyecto generar la herramienta de alertas, que sea en tiempo real, y permita gestionar de manera oportuna, un caso urgente generado por las reseñas o características propias del cliente.

CAPÍTULO 3

3 DISEÑO E IMPLEMENTACIÓN

En este capítulo, se presentará el diseño de la arquitectura de la solución incluyendo los requerimientos de la misma, el pipeline del procesamiento de texto y del modelamiento usando técnicas de NLP.

El diseño de la solución propuesta busca aprovechar las ventajas de las técnicas NLP para el procesamiento, generando datos de calidad, y a la vez integrar diferentes herramientas usadas a diario en la empresa “AAA”.

3.1 Esquema General de Implementación

Con respecto al diseño del proyecto, este se ha dividido en dos grandes partes que son:

1. Desarrollo del Pipeline del Entrenamiento del Modelo
2. Desarrollo del Pipeline de la Implementación

Cada una de estas partes es notablemente relevante para el desarrollo del proyecto, y consta de varios pasos a seguir para poder llevar a cabo la ejecución del mismo.

3.1.1 Desarrollo del Pipeline del Entrenamiento del Modelo

El Pipeline del Entrenamiento del Modelo consta de varias etapas secuenciales, cada tarea a realizar depende de la anterior a medida que se va avanzando en el desarrollo del proyecto, las etapas son las siguientes:

1. Obtención de datos tanto de fuentes internas (características del cliente, encuesta de satisfacción), como externas, en este caso de la base de datos de la Superintendencia de Compañías.
2. Etiquetado manual de las reseñas de los clientes realizado por personal del

área de Customer Experience.

3. Consolidar los datos de las diferentes fuentes y procesarlos para realizar agregaciones requeridas.
4. Aplicar procesos de transformación y depuración de datos (Pivoteo de columnas, eliminar datos duplicados e incoherentes, entre otras)
5. Aplicar tareas de preprocesamiento de Texto:
   - Transformación de todo el texto de las reseñas a letras minúsculas.
   - Eliminación de caracteres especiales.
   - Eliminación de Stopwords.
   - Stemming.
   - Tokenización.
   - Generación de la matriz document-term aplicando TF-IDF
   - Se realiza un Análisis de Componentes Principales PCA con la finalidad de realizar una reducción de la dimensión de las variables.
6. Aplicar el algoritmo de Análisis de Sentimientos (Cálculo de polaridad y Subjetividad) al dataset final, es decir se define si el sentimiento expresado en las reseñas es positivo, negativo o neutro.
7. Obtener los conjuntos de datos training y testing.
8. Aplicar Upsampling (sobremuestreo) usando SMOTE dado que se prevén clases desbalanceadas.
9. Entrenar los modelos y aplicar técnicas de optimización de Hiperparámetros considerando los siguientes algoritmos:
   - Modelo SVM
   - Modelo Random Forest
   - Modelo XGBoost
   - Modelo MLP
10. Entrenar un Ensemble Model usando la técnica de Stacking: combinación de los modelos entrenados en el paso anterior.
11. Segmentar los clientes por tipo cliente. El dataset incluye todas las características internas y externas del cliente, además de variables calculadas tales como la cantidad de comentarios positivos, negativos del cliente en el último año. Los algoritmos a usar para obtener los clústers de clientes son:

- K means
- DBSCAN

12. Escoger el mejor modelo de Segmentación y caracterizar los grupos obtenidos a partir de la ejecución del modelo seleccionado.

Se guardan los mejores modelos de clasificación y a su vez el dataset con los respectivos grupos en el repositorio, para la posterior predicción en las alertas. A continuación, se presenta una representación gráfica de este proceso.

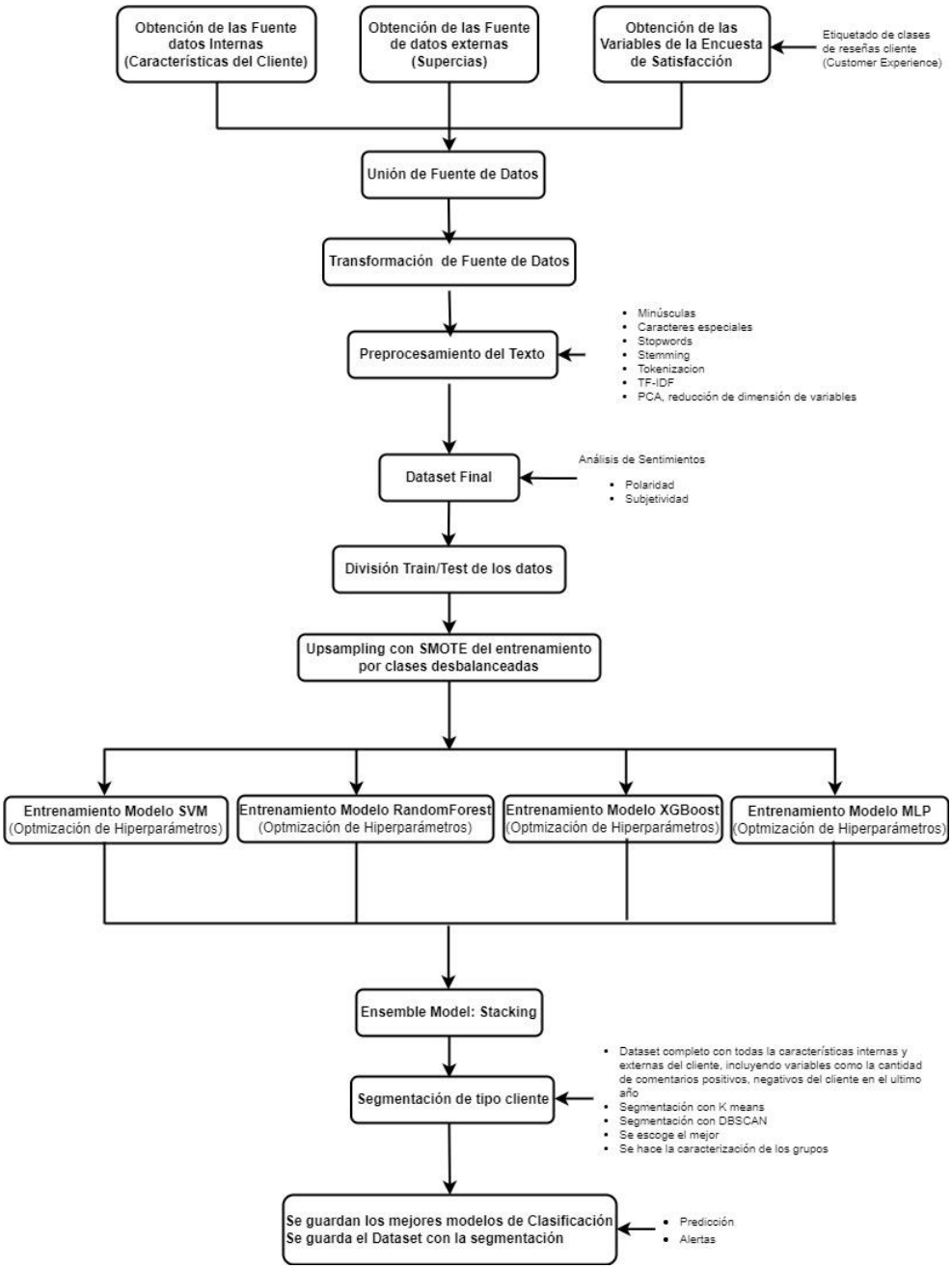


Figura 3.1 Pipeline del Entrenamiento del Modelo
Fuente: Elaboración Propia

3.1.2 Desarrollo del Pipeline de la Implementación

Una vez definido la guía de acción para la obtención de los modelos de clasificación de contenido de las reseñas de clientes, es necesario también crear una metodología que permita usar los modelos antes creados, para la predicción de nuevas reseñas en una herramienta de alertas para el proyecto. A continuación, se menciona en qué consiste cada una de las etapas de este pipeline de implementación:

1. Los operadores telefónicos de Customer Experience realizan las llamadas para aplicar la encuesta de satisfacción de calidad.
2. Se realiza la recolección de los datos de la encuesta y los registros se almacenan simultáneamente tanto en texto como en audio, el texto de la encuesta se recoge y almacena en la plataforma QuestionPro mientras que el audio se almacena en un repositorio propio de la empresa.
3. Mediante un orquestador de tareas se corren códigos en Python, el cual gestiona todo el proceso hasta la construcción final del dataset, luego se almacena en una base de datos de Exasol.
4. Implementar la conexión automática a Exasol para extraer los comentarios nuevos de la base de datos.
5. Se añade el análisis de sentimientos (polaridad y subjetividad).
6. Se realiza el Procesamiento de texto de reseñas nuevas.
7. y 8. Predecir las categorías binarias de las nuevas reseñas usando los algoritmos de clasificación previamente cargadas.
9. Se crea un dataset final.
10. Se guarda el dataset final en la misma base de datos de Exasol.
11. y 12. Alimentar Tableau con las predicciones generadas que se encuentran en EXASOL, y generar, desde Tableau, alertas automáticas para Slack con las nuevas encuestas de cierta categoría de contenido.
13. Slack manda una notificación al área de customer experience, con una captura, el texto del problema (categoría de contenido de la reseña) y un enlace a la plataforma de Tableau en la web, para ver los resultados con mayor detalle.

14. El usuario entra al enlace, y puede visualizar la situación del cliente en el dashboard dinámico, que contiene tanto características del cliente internas, externas, además de la categorización del contenido de la reseña, así como el segmento del cliente, realizado por el modelo de clustering, entre otras cosas.

A continuación, se presenta una representación gráfica de este proceso.

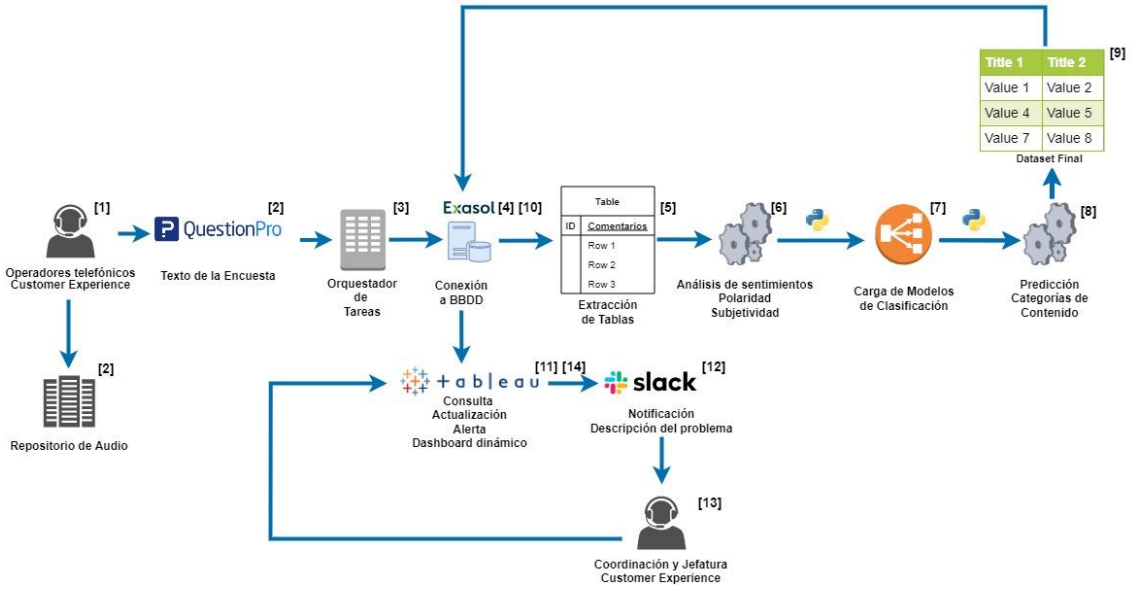


Figura 3.2 Pipeline de la Implementación del Modelo
Fuente: Elaboración Propia

3.1.3 Infraestructura necesaria

Es necesario mencionar que para poder llevar a efecto un producto de datos son necesarios recursos más allá de los modelos de machine Learning y es importante conocer el modelo de negocio del proceso que se está optimizando. Es importante entender cuál es la estructura de la empresa y cómo trabajan en conjunto sus departamentos, para tener claro cuál será el flujo de información, adicional a esto es necesario mencionar que para ciertos desarrollos se puede llegar a incurrir en gastos bastante elevados en el área de infraestructura.

Sin embargo, al ser la empresa “AAA” una empresa proveedora de servicios de internet y datos, posee la infraestructura necesaria para poder desarrollar este

proyecto, por lo que tanto los datos como los modelos generados se almacenarán y procesarán en un Servidor propio de la empresa que contiene una base de datos **EXASOL** (Base de Datos Analítica), un servidor con un **orquestador de tareas** instalado para la automatización de tareas, a su vez se tiene un servidor que contiene **Tableau Server** para las visualizaciones finales, y además que se cuenta con licencias corporativas de **Slack** para las notificaciones, y por último, el modelamiento se realizó con una computadora portátil que cumple con las siguientes características:

	Requisitos Mínimos Necesarios	Especificaciones Sugeridas
Sistema Operativo	Windows 7 de 64 bits Windows 10 de 64 bits	Windows 10 de 64 bits
Procesador	Intel Core I5 – 6 núcleos	Intel Core I7 – 8 núcleos
RAM	16 GB	32 GB
GPU	No Necesaria	NVIDIA GeForce GTX 970 -
HDD	1 TB	1 TB
Red	Conexión a internet de banda ancha	

Tabla 3.1 Infraestructura necesaria
Fuente: Elaboración Propia

La instalación del software y los paquetes necesarios para el tratamiento y procesamiento de los datos se realizó por parte del departamento de calidad de la empresa “AAA”. Se tendrá acceso a este servidor por medio de la red interna de la empresa y podrá hacerse consultas al mismo por medio de una VPN.

3.1.4 Restricciones / Limitaciones

En la implementación del proyecto, no siempre será factible realizar todo lo propuesto en el diseño de la arquitectura del mismo. Existen limitantes y se expone a continuación las más importantes:

- Debido a la mala calidad de los audios históricos de las llamadas telefónicas de la encuesta de satisfacción de calidad, así como el hecho de que no se tiene muchos registros históricos de los mismos, no se utilizó esta fuente de datos, como recurso principal para el posterior procesamiento de texto (con un proceso de speech to text), por lo que se optó por hacer uso de los comentarios escritos por parte de los operadores en un software web de encuestas; dichos operadores están entrenados para expresar las palabras del cliente con exactitud.
- Debido al mismo hecho de que depende de un intermediario, por más bueno que este sea al escribir las reseñas de los clientes, siempre habrá un poco de sesgo indirecto, algo de lo cual se está asumiendo en el mismo proyecto.
- Debido a la cantidad de reseñas para el entrenamiento posterior, se procede a dejar fuera del entrenamiento a ciertas categorías dadas por la parte interesada, porque no cumple con un mínimo de registros para encontrar patrones, aunque en fases posteriores se podrían incluir cada vez más categorías de contenido (fuera del alcance del documento).
- El etiquetado de las distintas categorías de reseñas fue realizado por la sección de customer experience, en el cual también existe un poco de sesgo, al ser una tarea subjetiva, por interpretación de un grupo de personas.
- El proceso de reentrenamiento de los modelos utilizados no está contemplado en el presente proyecto, dado que implica un nuevo proceso de etiquetado de nuevas reseñas por parte de la sección de customer experience (dicho proceso queda pendiente de definir los periodos), aunque el algoritmo utilizado para el entrenamiento está estructurado para ser fácilmente automatizado por un orquestador de tareas.

3.2 Obtención y Preprocesamiento de los datos

Como bien se ha explicado en el pipeline del proyecto, el mismo comienza con la obtención de las distintas fuentes de datos que intervienen en el proyecto, para posteriormente realizar una depuración de las fuentes de datos, y luego el preprocesamiento del texto, que sirva como input para el efectivo entrenamiento del modelo, dicho proceso será explicado con mayor detalle en el presente capítulo.

3.2.1 Obtención de las Fuentes de Datos

Una de las partes fundamentales del proyecto, es el primer paso de este, el cual consiste en la obtención de fuentes de datos que cuenten con integridad y veracidad de la información, sobre todo de las reseñas de los clientes en la encuesta de calidad; en el presente capítulo se divide entre las fuentes de datos internas y externas.

3.2.1.1 Fuentes de Datos Internas

Dentro de los procesos de obtención de las fuentes de datos internas que se ha utilizado en el proyecto, se puede acotar lo siguiente:

- La fuente de datos más importante del proyecto es la Base de Respuestas Históricas de los Indicadores y Reseñas de Clientes, obtenida de una herramienta de gestión de encuestas vía web, a la cual se tiene acceso. Cabe destacar que el proceso de llenado de las respuestas en la herramienta web, es mediante la tabulación de la información obtenida por el operador telefónico, es decir, que mientras se está preguntando las distintas dimensiones a ser evaluadas en la encuesta, el operador llena tanto los comentarios que dice el cliente como las respectivas calificaciones, al finalizar la misma se revisa toda la información llenada y se envía la encuesta, a su vez, también se guarda los audios de la encuesta telefónica con la respectiva autorización del cliente, para este proyecto en particular no se usarán los audios de las encuestas, debido a dos motivos en particular, primero

por la calidad del audio que se tiene de forma histórica, en el cual, tratar de mejorar la calidad de los audios está fuera del alcance del presente proyecto y llevaría un proceso mucho más largo, y segundo por permisos con distintas áreas de la empresa, por lo cual se optó por usar solamente las reseñas tabuladas por los mismos operadores, los cuales están entrenados para realizar un trabajo preciso de las acotaciones y comentarios del cliente durante la llamada. Posteriormente se descargó en forma de archivo plano, las respuestas tabuladas y completas de la encuesta de satisfacción de calidad (desde enero 2020 hasta julio 2022), con la finalidad de que la sección de Customer Experience de la Empresa “AAA” categorice las reseñas de los clientes, con las clases que ellos consideren relevantes para una gestión más efectiva, ya sean estas categorías positivas o negativas; después de muchas reuniones y trabajos en conjunto se obtuvieron un total de 25 clases, las cuales forman parte del ”Ground Truth” del proyecto, y que permitirán al modelo de clasificación tener un objetivo al cual ajustarse (Aprendizaje Automático Supervisado). Las 25 clases originales de las categorías de contenido de las reseñas se conforman por:

Canales de contacto	Cobertura	Comentario Bueno	Comentario Excelente	Contacta directo a su asesor
Demora en contestar (PBX - Correos)	Demora la atención - Análisis - Solución del caso	El PBX no contesta	Facturación	Falta de información - capacitación
Inconforme con la atención	Información de productos	Intermitencias en el servicio	Me contacto directo a L2 - VIP	Me explicó el asesor anterior
No califica - No contacto	No conoce el portafolio de productos	No es cliente VIP	No ha sido necesario	Otros
Precios	Problemas Con La Escalabilidad	Seguimiento	Si le han explicado los tiempos y proceso más relevantes	Tiempos de respuesta

Tabla 3.2 Categorías de Contenido de las Reseñas
Fuente: Elaboración Propia

- Otra de las fuentes de datos importantes del proyecto, son las características de los clientes, dentro de la empresa “AAA”, como por ejemplo la antigüedad que estas tienen, la cantidad de servicios, la cantidad de cancelaciones, la facturación, entre otras cosas; dichas fuentes se componen de distintas tablas y vistas de distintos sistemas de información de la empresa “AAA”, afortunadamente el Departamento de Calidad tiene acceso a dichas fuentes de datos, y además ha realizado un Datawarehouse que combina todas esas características en una sola tabla/vista en una base de datos analítica llamada EXASOL, a la cual se tiene acceso para este proyecto con las respectivas credenciales de lectura. Estas fuentes de datos permitirán realizar una posterior segmentación de tipo de clientes, analizando no sólo las características resultantes del procesamiento de lenguaje natural, sino también de las características internas del cliente con la empresa, a su vez, permitirán a las partes interesadas, visualizar información más detallada en la herramienta de alertas.

3.2.1.2 Fuentes de Datos Externas

El presente proyecto no está enfocado en el uso de muchas fuentes de datos externas, pero sí usa una en particular, que son las cuentas de los balances generales cargados en la web de la Superintendencia de Compañías, la mayoría tiene un formato de formulario 101 del SRI, donde cada una de las empresas registradas tienen la obligación de subir sus respectivos documentos año tras año, y la SUPERCIAS pone a disposición dicha información de manera pública. El departamento de Calidad ha formado un repositorio histórico de dicha información, teniendo en el Datawarehouse de EXASOL, una tabla con la información de las cuentas financieras e indicadores desde el 2015 hasta el 2021. Estos datos permitirán enriquecer a las fuentes de datos internas de las características de los clientes, dándole una perspectiva financiera que antes no se tenía, y servirá además para el modelo de segmentación de clientes.

3.2.2 Depuración de las Fuentes de Datos

Una vez obtenidos todas las fuentes de datos para el proyecto, es de suma importancia realizar una validación y depuración de estos, la calidad de los resultados del modelo será en la misma proporción que los datos proporcionados. Dentro de las principales tareas de depuración realizadas de las fuentes de datos obtenidas, se nombran las siguientes:

- Debido a que la tabla de las reseñas de los clientes está de forma tabular/columnar, existen distintos comentarios en distintas partes de la encuesta, es decir, existe un comentario para un área o dimensión en particular a evaluar al comienzo de la encuesta, después existe otro campo de comentario para otra área o dimensión, y así sucesivamente, por lo cual se realizó un pivoteo de los distintos campos relacionados a las reseñas y a su vez a las categorías o clases asociadas a estas, con la finalidad de tener un solo campo de texto, que permita categorizar el contenido, además se puede obtener varias categorías relacionadas a distintas reseñas de un mismo cliente, analizar cuáles están asociadas a un área en específico, y optimizar recursos informáticos al recorrer el proceso una sola vez por categoría. Se eliminan los registros, donde la reseña esté en nulo.

- Posteriormente, se realizó un One-Hot Encoding del campo de Categoría de la Reseña, con la finalidad de tener campos binarios, que representen presencia o ausencia de la categoría por cada reseña. Esto permitirá tener un valor objetivo por cada uno de los modelos de clasificación de contenido que se realicen. Este proceso es importante dado que se parte de la hipótesis que una reseña nueva, puede contener más de una categoría, por lo que es necesario hacer “n” modelos de clasificación binaria, y no un modelo de clasificación multiclase.

Cliente	Área 1 - Dimensión 1	Reseña 1	Área 1 - Dimensión 2	Reseña 2	Área 2 - Dimensión 1	Reseña 3
ABC	5	….	3	….	4	....
CDE	4	….	4	….	1	….
FGH	4	….	1	….	5	….

Tabla 3.3 Depuración de las Fuentes de Datos - Parte 1
Fuente: Elaboración Propia

Cliente	Área	Dimensión	Calificación	Reseña	Categoría Reseña
ABC	Área 1	Dimensión 1	5	….	A
ABC	Área 1	Dimensión 2	3	….	B
ABC	Área 2	Dimensión 1	4	….	C
CDE	Área 1	Dimensión 1	4	….	B
CDE	Área 1	Dimensión 2	4	….	B
CDE	Área 2	Dimensión 1	1	….	C
FGH	Área 1	Dimensión 1	4	….	A
FGH	Área 1	Dimensión 2	1	….	A
FGH	Área 2	Dimensión 1	5	….	B

Tabla 3.4 Depuración de las Fuentes de Datos - Parte 2
Fuente: Elaboración Propia

Cliente	Área	Dimensión	Calificación	Reseña	Categoría A	Categoría B	Categoría C
ABC	Área 1	Dimensión 1	5	….	1	0	0
ABC	Área 1	Dimensión 2	3	….	0	1	0
ABC	Área 2	Dimensión 1	4	….	0	0	1
CDE	Área 1	Dimensión 1	4	….	0	1	0
CDE	Área 1	Dimensión 2	4	….	0	1	0
CDE	Área 2	Dimensión 1	1	….	0	0	1
FGH	Área 1	Dimensión 1	4	….	1	0	0
FGH	Área 1	Dimensión 2	1	….	1	0	0
FGH	Área 2	Dimensión 1	5	….	0	1	0

Tabla 3.5 Depuración de las Fuentes de Datos - Parte 3
Fuente: Elaboración Propia

- Adicionalmente, con las distintas fuentes de datos internas y externas que contienen las características del cliente, se realizó una agregación a nivel de cliente de los campos a utilizar o relevantes para el proyecto, por ejemplo, en el caso de soportes generados, tener una fuente de datos a nivel de clientes que contenga la cantidad de soportes, o en el caso de facturación, no tener un nivel de granularidad a nivel de factura, sino la suma de la facturación a nivel de cliente, entre muchos otros casos. Posteriormente se hizo un merge / combinación de las tablas agregadas, para tener una sola fuente centralizada de información, que esté a nivel de cliente.

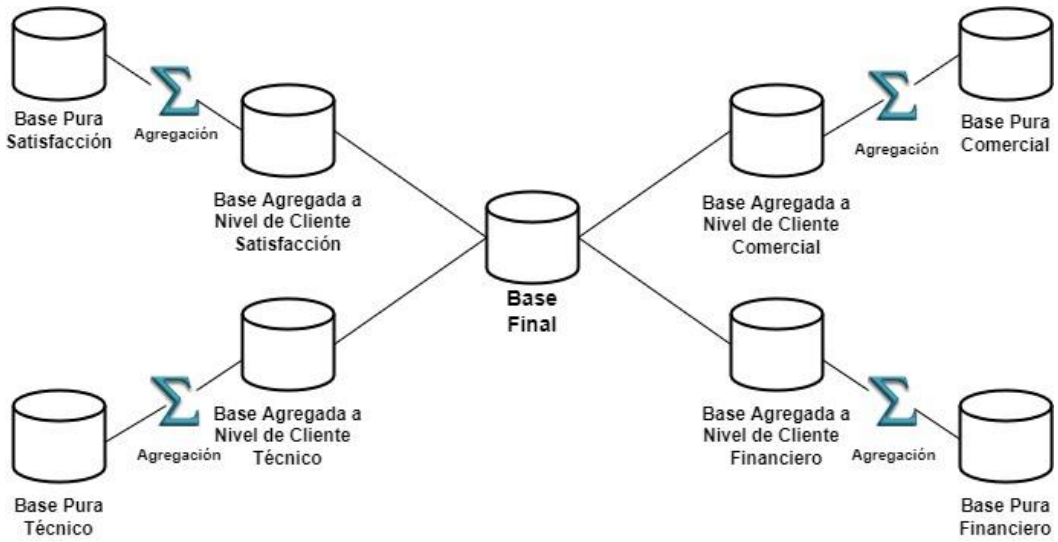


Figura 3.3 Agregación a Nivel de Cliente
Fuente: Elaboración Propia

- Finalmente, en todas las fuentes de datos se realizan trabajos básicos de depuración, como por ejemplo eliminar datos duplicados, validar datos vacíos, datos inconsistentes, categorías de texto mal escritas, integridad de los datos, entre otras.

3.2.3 Preprocesamiento del texto

Una vez obtenido una estructura adecuada en las fuentes de datos depuradas, para el modelo de clasificación de contenido de las reseñas es necesario realizar un preprocesamiento, con el fin de que la máquina pueda interpretar un texto de la forma adecuada, dado que si tiene solo el dato no estructurado, los diferentes algoritmos de clasificación, no podrán encontrar los patrones necesarios para realizar una buena generalización para las predicciones, por lo tanto, el proceso que se realizó al campo de reseña, fue el siguiente:

- En primer lugar, se redujo todo el texto a minúsculas **(Dimensión Original: 5040 reseñas, 1 columna de texto).**
- Se eliminaron caracteres especiales del texto como $, %, &, ‘, /, -, entre otras.
- Se reemplazaron caracteres que tienen signos de puntuación, con aquellos que no lo tienen como, por ejemplo, á con a, é con e; o a su vez otros caracteres como ñ por n, ü por u, entre otros.
- Se eliminaron stopwords definidas por una librería de textos en español, así como también se añadieron stopwords adicionales, por conocimiento del negocio, definidos tanto por el grupo del proyecto como por la sección de customer experience.
- Se realizó stemming a las reseñas, con la finalidad de reducirlos a su forma canónica, para que el modelo generalice términos comunes, o que tienen significados parecidos.
- Se procede a realizar una tokenización de cada una de las reseñas, al tener un vector de distinta dimensión por cada reseña generada, donde cada token es una palabra del texto.
- Una vez obtenidos los tokens, se realiza una representación matricial de los tokens por cada una de las reseñas, en la cual cada fila es una reseña / comentario en la encuesta, y cada columna es una palabra en específico, que puede o no aparecer en el texto de la reseña **(Dimensión: 5040 reseñas, 2980 palabras).**
- Se aplica el proceso de selección de características (smoothing del texto) mediante TF-IDF, el cual permite tener en cuenta tanto la

frecuencia de las palabras más frecuentes, así como de aquellas palabras “raras” o poco frecuentes dentro del total de reseñas generadas.

- Se aplica Análisis de Componentes Principales (PCA), con la finalidad de reducir la dimensionalidad de los términos generados, manteniendo gran parte de la variabilidad de los datos; mediante el gráfico de sedimentación se escoge las primeras 1000 componentes principales, el cual mantiene una proporción de varianza acumulada explicada del 0.9432 (perdiendo menos del 6% del total de la variabilidad, pero reduciendo más de la mitad de la dimensión original). A su vez, se conoce que dicho proceso ayuda a tener un mejor rendimiento en la generalización de modelos de clasificación **(Dimensión: 5040 reseñas, 1000 componentes).**

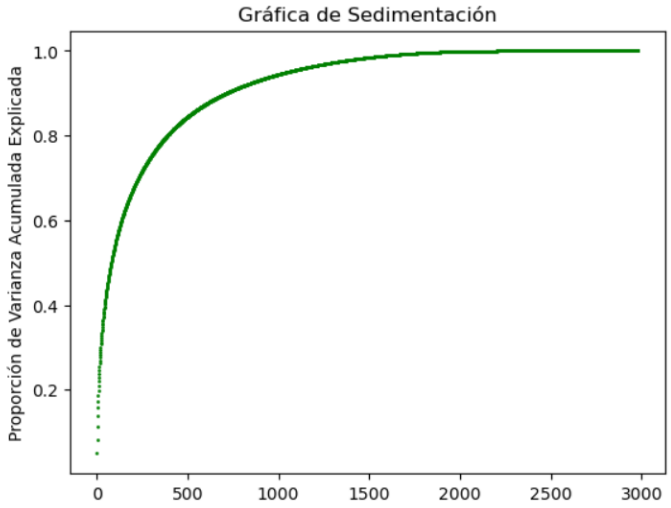


Figura 3.4 Gráfica de Sedimentación Análisis de Componentes Principales (PCA)
Fuente: Elaboración Propia

Finalmente, se obtiene una matriz de datos la cual tiene las 5040 reseñas originales, pero con sólo 1000 variables que representan las componentes principales de los términos utilizados con TF-IDF, dicha matriz representa el input para los distintos modelos de clasificación que se verán más adelante junto con los campos binarios de las distintas clases de contenido de reseñas. Por otro lado, la matriz de características de los clientes junto con los resultados de la clasificación de contenido de las reseñas, serán el input para el modelo de segmentación de tipos de clientes. Posterior a este proceso, viene un paso

importante para el proyecto, el cual es, el entrenamiento de los distintos modelos de clasificación, optimización de Hiperparámetros, modelo de segmentación y, por último, la herramienta de generación de alertas de las reseñas, para las partes interesadas.

3.3 Exploración de Algoritmos para Análisis de Sentimientos y Clasificación de Categoría de las Reseñas

Para esta sección del proyecto, se realizaron dos tareas, la primera es utilizar una herramienta, de la librería de TextBlob (Al igual que NLTK, tiene una amplia gama de funciones para NLP), para calcular la polaridad y subjetividad de cada una de las reseñas de los clientes en la encuesta, es decir, no se va a entrenar un modelo, sólo se va a "predecir" los resultados de dichos indicadores, con un paquete ya constituido por muchas reglas léxicas y semánticas, que miden la forma en que está estructurada la oración, además de la intensidad de cada una de las palabras; la segunda tarea de este capítulo es entrenar varios modelos de clasificación binarios, para la respectiva categorización de contenido, en base al "Ground Truth" definido por la sección de Customer Experience.

3.3.1 Análisis de Sentimientos

Como se había descrito anteriormente, en esta parte del proyecto no se va a entrenar algún modelo de clasificación, sólo se va a utilizar una herramienta muy potente en el mercado como lo es el paquete de TextBlob, para obtener la polaridad y subjetividad del texto, pero antes de continuar, ¿Qué es la polaridad y subjetividad?, la polaridad es el grado de sentimiento positivo-negativo que tiene un texto en base a su contenido, la librería TextBlob devuelve como respuesta un rango de -1 a 1, siendo -1 un sentimiento muy negativo y +1 un sentimiento muy positivo; mientras que la subjetividad, como su nombre lo dice mide el grado de aspectos subjetivos que impone la persona al escribir una reseña o comentario, así mismo TextBlob devuelve un rango de valores, pero en este caso de 0 a 1, siendo 0 un grado donde no hay subjetividad impuesta por la persona, y 1 un grado alto de subjetividad por parte de la persona. En el presente proyecto se analizará en mayor proporción la polaridad, más que la

subjetividad, debido a que el cliente como tal en el proyecto, no escribe una reseña, sino que pasa por un intermediario, en este caso el operador, el cual no puede expresar en un 100% la totalidad de las expresiones del cliente, ni tampoco la intensidad de las palabras, por lo cual se va perdiendo un poco la subjetividad por parte del cliente. En el caso de polaridad, para el presente proyecto se definieron categorías con rangos de valores, los cuales categorizan lo que es un comentario positivo, negativo o neutro, el cual es el siguiente:

Rango de Polaridad	Tipo de Sentimiento
-1 < Polaridad <=-0.2	Comentario Negativo
-0.2 < Polaridad < 0.2	Comentario Neutro
0.2 < Polaridad < 1	Comentario Positivo

Tabla 3.6 Análisis de Sentimientos
Fuente: Elaboración Propia

El objetivo principal de este proceso de análisis de sentimientos es darle un mayor contexto del problema a analizar a las partes interesadas, no es lo mismo que la reseña sea de un problema en específico, pero con un comentario neutro, a que sea el mismo problema, pero con un comentario muy negativo, dado que, por diversas razones, el cliente puede expresarse con palabras más fuertes por los sentimientos de enojo o ira, por lo cual sea más de urgencia que atender.

Este proceso de análisis de sentimientos se lo realizará tanto para las nuevas reseñas de la encuesta de satisfacción de calidad (herramienta de generación de alertas), como para la data histórica de reseñas, con la finalidad de analizar descriptivamente la evolución y la actualidad de los comentarios positivos y negativos en la empresa “**AAA**”, los resultados se expondrán en el capítulo 4.

3.3.2 Modelo de Clasificación de Categorías de Reseñas

Uno de los procesos claves para lograr la herramienta de generación de alertas automáticas en base al contenido de las reseñas, es el modelo de clasificación, dicho componente en todo este proceso permitirá realizar las predicciones de las reseñas de los clientes, cuando llegue una nueva encuesta de calidad a la

base, por lo que es de suma importancia realizar un buen diseño e implementación del mismo.

En primer lugar, hay que recalcar que no es un solo modelo de clasificación multiclase, sino **n** modelos de clasificación binaria, porque como antes se había descrito, una reseña puede tener distintas categorías o problemas en específico, en la sección 3.2.1.1 Fuentes de Datos Internas, se indicó que se habían encontrado 25 clases particulares, pero no todas van a ser posible entrenar un modelo, debido a la cantidad de datos contenida en cada una de las categorías.

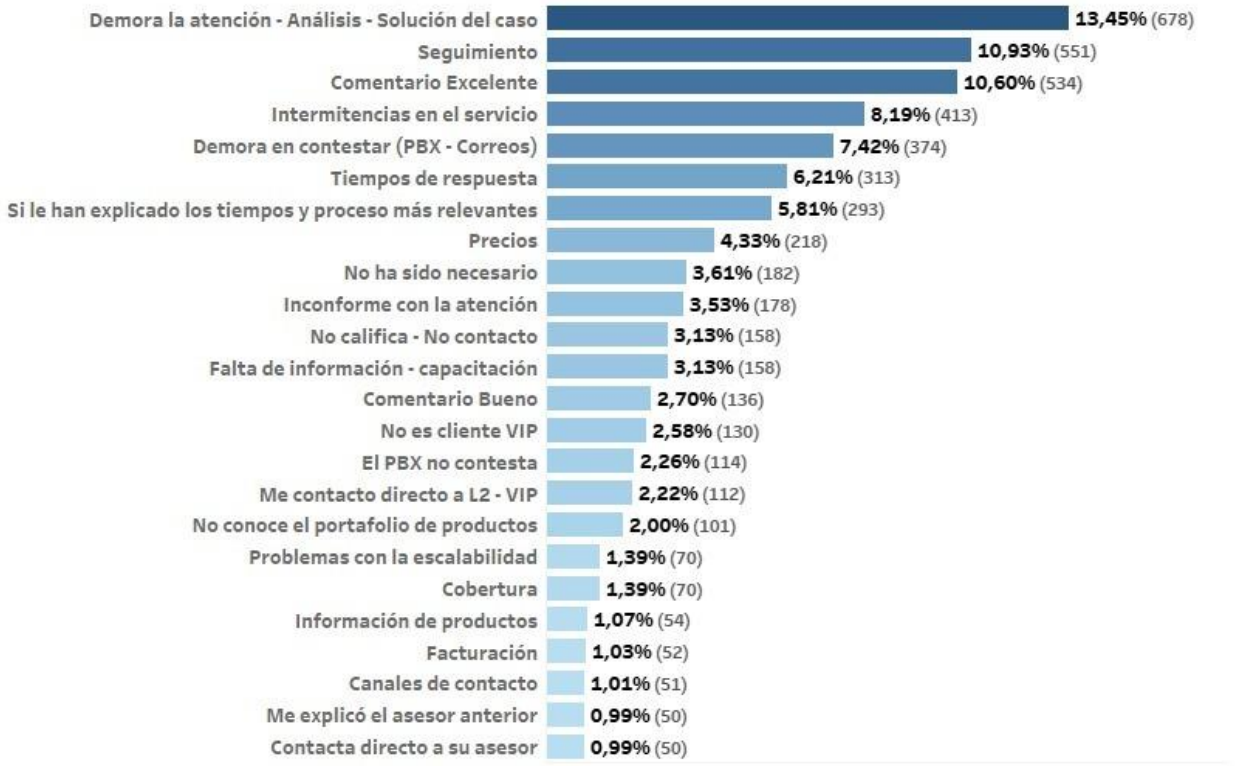


Figura 3.5 Modelo de Clasificación de Categorías de Reseñas
Fuente: Elaboración Propia

Tomando en cuenta las categorías más importantes, y la cantidad de registros que existen en la base depurada de reseñas que hacen posible el entrenamiento de este, se decidió como fase 1 del proyecto, comenzar sólo con las siguientes 10 categorías (10 modelos de clasificación binaria):

Intermitencias en el servicio	Demora en contestar (PBX - Correos)	Tiempos de respuesta	Demora la atención Solución del caso	Seguimiento
Comentario Excelente	Comentario Bueno	No conoce el portafolio de productos	Problemas con los Precios	Inconforme con la atención

Tabla 3.7 Categorías Finales Modelos de Clasificación de Categorías de Reseñas
Fuente: Elaboración Propia

Posteriormente a definir que categorías se van a predecir como modelos de clasificación binaria, se realizaron ciertos procedimientos previos al entrenamiento de los distintos modelos de clasificación como:

- Se realizó una división del training set con el testing set de 80-20, tomando en cuenta la matriz de 5040 reseñas con 1000 componentes principales (features del modelo) y con la variable objetivo de la clase/categoría que se está modelando (10 en total), con una estratificación por la proporción de la variable objetivo, dado que son distintos datos que tienen clases desbalanceadas, viéndolo como una variable binaria, existen más 0 que 1, más ausencia de una categoría en específico en comparación a las demás, por lo que la separación de training y testing de la variable objetivo dispone de la misma proporción, para una mayor representación del problema cuando se evalúe.
- Previo al entrenamiento de los modelos, se realizó upsampling usando SMOTE para evitar tener clases desbalanceadas.

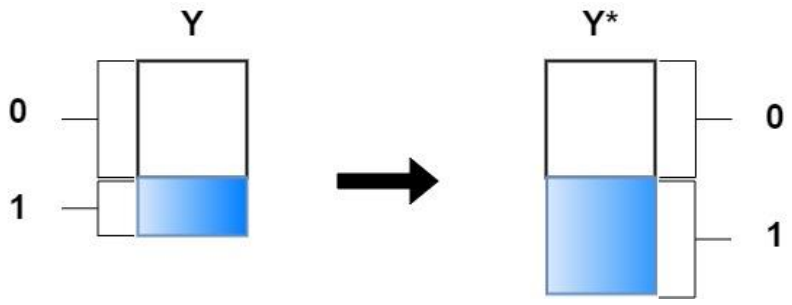


Figura 3.6 Upsampling con algoritmo de SMOTE
Fuente: Elaboración Propia

- Una vez obtenidos los conjuntos de datos de training y de testing definitivos, se procede a entrenar los distintos modelos de clasificación para el presente proyecto.

Al iniciar el proyecto y consultar el estado del arte dentro de la categorización de contenido, se ha podido evidenciar un notable uso por ciertos algoritmos dentro del aprendizaje automático supervisado, los cuales tienen un buen rendimiento al usar matrices que contienen puntuaciones de textos, por lo que se optó por los siguientes algoritmos:

- Support Vector Machine
- Random Forest
- XGBoost
- Multilayer Perceptron

Finalmente se entrenarán 4 modelos por cada una de las 10 categorías binarias que existen, pero no termina ahí, dado que cuando se tengan los mejores 4 modelos por cada categoría (optimización de Hiperparámetros), se realizará un voting classifier (Clasificador por votos), el cual es una metodología de Ensemble Models (Stacking), que permite usar las predicciones de los 4 modelos, al crear una predicción final por mayoría de votos de los modelos utilizados (Metodología Hard).

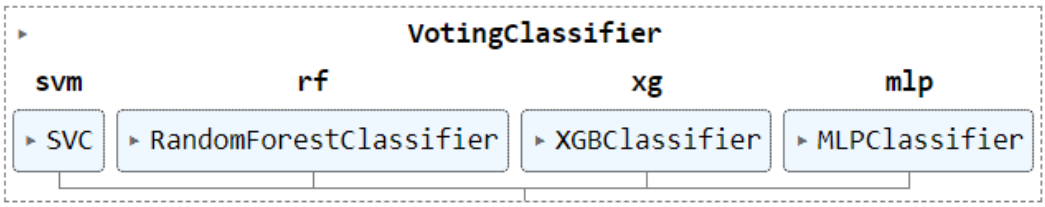


Figura 3.7 Voting Classifier (Clasificador por votos)
Fuente: Scikit-learn

3.3.3 Elección de Hiperparámetros y Evaluación de Métricas

Antes de comenzar a entrenar y escoger los mejores hiperparámetros de los modelos, se tiene que definir la métrica a usar para evaluar dichos hiperparámetros, y algo que hay que tener muy en cuenta en este problema, es que cada una de las 10 categorías objetivo tienen clases desbalanceadas, por lo que las métricas de evaluación como el accuracy pueden ser muy engañosas. En este caso particular, se tomó la decisión de evaluar una métrica distinta, creada a partir de una ponderación entre el recall y la precisión:

$$score\_pro = 0.75 * recall + 0.25 * precision$$

Dicha ponderación 0.75 y 0.25 fue dada por la sección de customer experience de la empresa “AAA”, al evaluar los posibles riesgos o consecuencias de falsos positivos y falsos negativos del modelo para la gestión de alertas (proporción 3:1), dado que en primer lugar no se desea que exista una gran cantidad de falsos negativos. Un falso negativo corresponde a una alerta que es importante analizar, pero que el modelo indica que la reseña del cliente no es importante (nunca se alertó dicha encuesta), este caso es un tema muy delicado para la empresa, porque pierde una gestión importante para el cliente, por lo cual esto representa mucho más para la sección de customer experience que los falsos positivos, lo cual son falsas alarmas, es decir reseñas que el modelo alertó, pero que no necesariamente son verdaderas; dicha métrica creada para la evaluación de los modelos se la nombró como score_pro y representa la forma en la cual se escogerá el mejor modelo por cada tipo de alerta. Como línea base de métrica del proyecto, se tomó en cuenta una regresión logística sin regularizar por cada uno de los 10 modelos de clasificación binaria, donde el score_pro promedio de dichos modelos fue de aproximadamente **0.55**, cualquier modelo que esté por encima de dicha métrica será adecuado para el proyecto.

La metodología para la elección de los mejores hiperparámetros de los distintos modelos, es creando una elección aleatoria de valores iniciales de estos hiperparámetros, en los cuales, mediante validación cruzada, va a evaluar la

métrica escogida, en este caso el score_pro, este proceso no se realiza una sola vez, sino en muchas iteraciones, guardando en memoria los mejores hiperparámetros que resulten en el mejor score_pro, a todo este proceso se le ha denominado **primera aproximación**. Una vez escogido los valores iniciales de los hiperparámetros, se procede a realizar un rango de posibles valores más ajustados, con la finalidad de ser más precisos, en la elección de hiperparámetros, a este proceso se le denomina en este proyecto la **segunda aproximación**. Dicho proceso se repite para cada uno de los 4 modelos principales: SVM, Random Forest, XGBoost y MLP, y para cada una de las 10 categorías binarias a realizar.

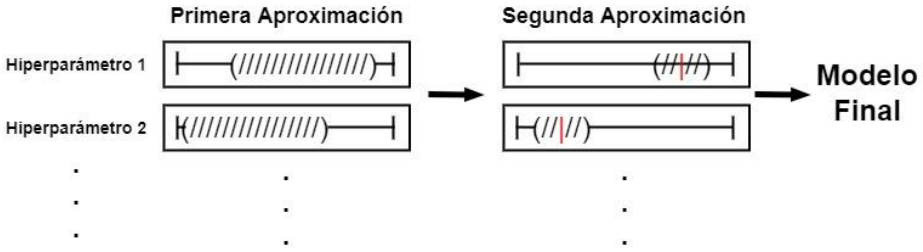


Figura 3.8 Metodología para elección de hiperparámetros
Fuente: Elaboración Propia

Finalmente teniendo los mejores Hiperparámetros de la segunda aproximación, se procede a realizar un modelo final de cada uno, y posteriormente se realiza el voting classifier, que tome en cuenta los 4 modelos, para cada una de las 10 categorías de contenido en este proyecto. A continuación, se exponen las métricas de evaluación, así como las matrices de confusión de cada modelo final:

		SVM				
Categoría	**Descripción**	**Accuracy**	**Precision**	**Recall**	**f1-score**	**score_pro**
Categoría 1	Intermitencias en el servicio	0,972	0,867	0,783	0,823	0,804
Categoría 2	Demora en contestar	0,954	0,796	0,520	0,629	0,589
Categoría 3	Tiempos de respuesta	0,964	0,765	0,619	0,684	0,655
Categoría 4	Demora la Solución del caso	0,928	0,818	0,629	0,711	0,677
Categoría 5	Seguimiento	0,947	0,806	0,682	0,739	0,713
Categoría 6	Comentario Excelente	0,955	0,878	0,673	0,762	0,724
Categoría 7	Comentario Bueno	0,986	0,760	0,704	0,731	0,718
Categoría 8	No conoce el portafolio de productos	0,995	1,000	0,750	0,857	0,813
Categoría 9	Precios	0,985	0,914	0,727	0,810	0,774
Categoría 10	Inconforme con la atención	0,986	0,839	0,743	0,788	0,767

Tabla 3.8 Métricas de Evaluación SVM
Fuente: Elaboración Propia

		Random Forest				
Categoría	**Descripción**	**Accuracy**	**Precision**	**Recall**	**f1-score**	**score_pro**
Categoría 1	Intermitencias en el servicio	0,970	0,921	0,699	0,795	0,754
Categoría 2	Demora en contestar	0,966	0,902	0,613	0,730	0,685
Categoría 3	Tiempos de respuesta	0,964	0,829	0,540	0,654	0,612
Categoría 4	Demora la Solución del caso	0,910	0,724	0,587	0,649	0,622
Categoría 5	Seguimiento	0,944	0,821	0,627	0,711	0,676
Categoría 6	Comentario Excelente	0,950	0,806	0,701	0,750	0,727
Categoría 7	Comentario Bueno	0,985	0,731	0,704	0,717	0,710
Categoría 8	No conoce el portafolio de productos	0,992	0,875	0,700	0,778	0,744
Categoría 9	Precios	0,990	0,947	0,818	0,878	0,850
Categoría 10	Inconforme con la atención	0,985	0,885	0,657	0,754	0,714

Tabla 3.9 Métricas de Evaluación Random Forest
Fuente: Elaboración Propia

		XGBoost				
Categoría	**Descripción**	**Accuracy**	**Precision**	**Recall**	**f1-score**	**score_pro**
Categoría 1	Intermitencias en el servicio	0,979	0,956	0,783	0,861	0,826
Categoría 2	Demora en contestar	0,960	0,857	0,560	0,677	0,634
Categoría 3	Tiempos de respuesta	0,966	0,822	0,587	0,685	0,646
Categoría 4	Demora la Solución del caso	0,920	0,754	0,643	0,694	0,671
Categoría 5	Seguimiento	0,937	0,740	0,645	0,689	0,669
Categoría 6	Comentario Excelente	0,967	0,911	0,766	0,832	0,803
Categoría 7	Comentario Bueno	0,985	0,692	0,720	0,706	0,713
Categoría 8	No conoce el portafolio de productos	0,989	0,909	0,500	0,645	0,602
Categoría 9	Precios	0,997	0,977	0,955	0,966	0,960
Categoría 10	Inconforme con la atención	0,988	0,871	0,771	0,818	0,796

Tabla 3.10 Métricas de Evaluación XGBoost
Fuente: Elaboración Propia

		MLP				
Categoría	**Descripción**	**Accuracy**	**Precision**	**Recall**	**f1-score**	**score_pro**
Categoría 1	Intermitencias en el servicio	0,954	0,776	0,627	0,693	0,664
Categoría 2	Demora en contestar	0,954	0,774	0,547	0,641	0,603
Categoría 3	Tiempos de respuesta	0,964	0,755	0,635	0,690	0,665
Categoría 4	Demora la Solución del caso	0,897	0,694	0,476	0,564	0,530
Categoría 5	Seguimiento	0,933	0,698	0,673	0,685	0,679
Categoría 6	Comentario Excelente	0,938	0,759	0,617	0,680	0,652
Categoría 7	Comentario Bueno	0,978	0,586	0,630	0,607	0,619
Categoría 8	No conoce el portafolio de productos	0,996	0,944	0,850	0,895	0,874
Categoría 9	Precios	0,989	0,902	0,841	0,871	0,856
Categoría 10	Inconforme con la atención	0,977	0,700	0,600	0,646	0,625

Tabla 3.11 Métricas de Evaluación MLP
Fuente: Elaboración Propia

		Stacking Model				
Categoría	**Descripción**	**Accuracy**	**Precision**	**Recall**	**f1-score**	**score_pro**
Categoría 1	Intermitencias en el servicio	0,983	0,958	0,831	0,890	0,863
Categoría 2	Demora en contestar	0,966	0,887	0,627	0,734	0,692
Categoría 3	Tiempos de respuesta	0,974	0,878	0,683	0,768	0,731
Categoría 4	Demora la Solución del caso	0,921	0,770	0,617	0,685	0,655
Categoría 5	Seguimiento	0,959	0,888	0,718	0,794	0,761
Categoría 6	Comentario Excelente	0,956	0,854	0,710	0,776	0,746
Categoría 7	Comentario Bueno	0,990	0,840	0,778	0,808	0,793
Categoría 8	No conoce el portafolio de productos	0,995	0,941	0,800	0,865	0,835
Categoría 9	Precios	0,991	0,973	0,818	0,889	0,857
Categoría 10	Inconforme con la atención	0,986	0,839	0,743	0,788	0,767

Tabla 3.12 Métricas de Evaluación Stacking Model
Fuente: Elaboración Propia

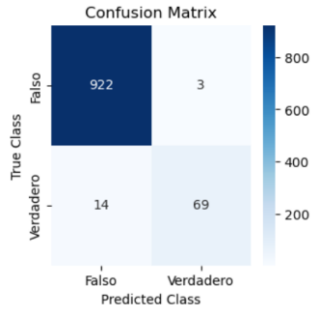
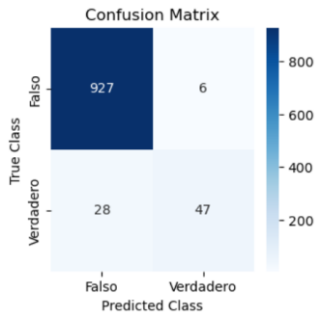
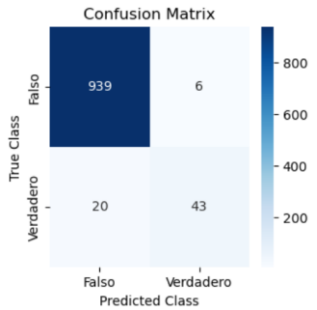
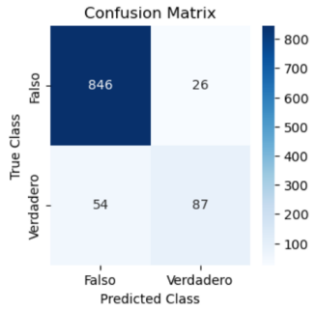
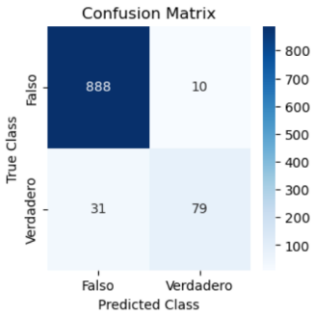
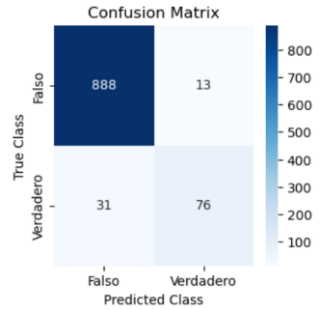
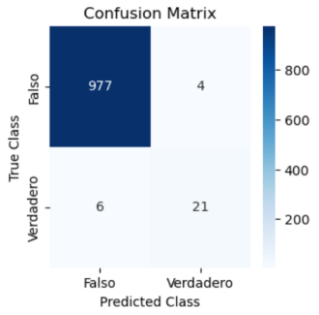
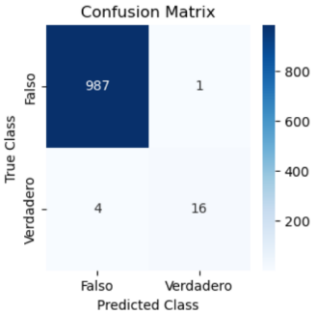
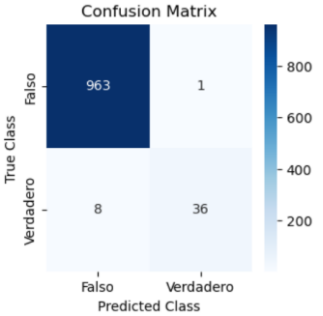
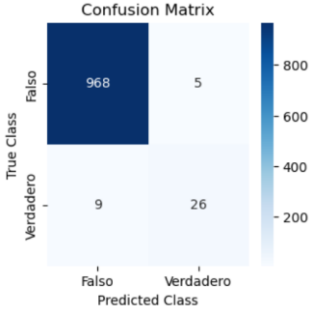
Intermitencias en el Servicio	Demora en contestar	Tiempos de respuesta	Demora en solución del caso	Problemas de Seguimiento
				
Comentario Excelente	**Comentario Bueno**	**No conoce los productos**	**Problemas con los Precios**	**Inconforme con la atención**
				

Tabla 3.13 Matrices de Confusión de Modelos Evaluados
Fuente: Elaboración Propia

3.4 Exploración de Algoritmos para Segmentación de Tipo de Clientes

Una vez realizado, el proceso de Clasificación de Reseñas de Clientes se procede a crear un valor añadido al proyecto. Se va a indicar el tipo de cliente según características, técnicas, comerciales, entre otras, con la finalidad de darle una prioridad a las alertas generadas. Esta tarea, no puede ser sencillamente por rangos de facturación, o por el tamaño de la empresa, como se suele hacer en las mayorías de segmentos, para darle una prioridad o no a un cliente, sino que se tomarán en cuenta campos tanto internos como externos, para segmentar a clientes importantes o potenciales dentro del contexto de la empresa “AAA”. Una vez definida la importancia del cliente, se le deba gestionar con mayor prontitud alguna solicitud o requerimiento del mismo, sin tampoco descuidar aquellas empresas menos importantes. Los campos que se tomaron en cuenta para la segmentación de clientes fueron los siguientes:

Nombre Campo	Descripción
CANT_SERV	Cantidad de Servicios Activos que tiene el cliente.
ANT_ANIOS	Cantidad de Años de antigüedad como cliente de la empresa “AAA”
CANT_EMPLEADOS_2021	Cantidad de Empleados registrados en la SUPERCIAS, para el periodo fiscal 2021.
FACT_ULT_12_MESES	Facturación Total de los últimos 12 meses.
SOP_1T	Cantidad de Soportes Generados en el último trimestre.
SOP_2T	Cantidad de Soportes Generados hace un trimestre.
SOP_3T	Cantidad de Soportes Generados hace dos trimestres.
SOP_4T	Cantidad de Soportes Generados hace tres trimestres.
SERV_1T	Cantidad de Servicios Nuevos Activados en el último trimestre.
SERV_2T	Cantidad de Servicios Nuevos Activados hace un trimestre.
SERV_3T	Cantidad de Servicios Nuevos Activados hace dos trimestres.

SERV_4T	Cantidad de Servicios Nuevos Activados hace tres trimestres.
CAN_1T	Cantidad de Servicios Cancelados en el último trimestre.
CAN_2T	Cantidad de Servicios Cancelados hace un trimestre.
CAN_3T	Cantidad de Servicios Cancelados hace dos trimestres.
CAN_4T	Cantidad de Servicios Cancelados hace tres trimestres.
TOTAL_ACTIVOS_2021	Total de Activos Registrados en la SUPERCIAS, para el periodo fiscal 2021.
TOTAL_PASIVOS_2021	Total de Pasivos Registrados en la SUPERCIAS, para el periodo fiscal 2021.
TOTAL_PATRIMONIO_2021	Total de Patrimonio Registrados en la SUPERCIAS, para el periodo fiscal 2021.
TOTAL_INGRESOS_2021	Total de Ingresos Registrados en la SUPERCIAS, para el periodo fiscal 2021.
TOTAL_COSTOS_GASTOS_2021	Total de Costos + Gastos Registrados en la SUPERCIAS, para el periodo fiscal 2021.
TAMANO_MEDIANA*	Variable Booleana, que indica si la empresa es Mediana o No, registrado en la SUPERCIAS 2021.
TAMANO_MICROEMPRESA*	Variable Booleana, que indica si la empresa es Microempresa o No, registrado en la SUPERCIAS 2021.
TAMANO_PEQUENA*	Variable Booleana, que indica si la empresa es Pequeña o No, registrado en la SUPERCIAS 2021.

Tabla 3.14 Campos Segmentación de Tipo de Clientes
Fuente: Elaboración Propia

**Cabe destacar que en la tres últimas variable, si todas son 0, entonces el cliente es GRANDE.*

Para la segmentación del tipo de cliente, se recurrió al estado del arte para obtener los algoritmos más usados dentro de este proceso, y se decidió utilizar los siguientes dos modelos, por la simplicidad y a la vez el rendimiento obtenido en escenarios similares al considerado en este proyecto:

- K means
- DBSCAN

Una vez que se realicen las agrupaciones, se escogerá el mejor modelo de segmentación, que represente los objetivos del proyecto, teniendo así un valor añadido al proyecto, que se presente en la herramienta de generación de alertas y permita dar prioridades a la gestión.

En primer lugar, se procedió a ejecutar el algoritmo de k-medias con los datos antes descritos con una distancia euclidiana por defecto. Como paso inicial, se determinó el valor de k en 5 clústers como resultado del diagrama de codo. Este método, usa una métrica de distancia media de los puntos hacia su centroide, variando el número de clústers a elegir, teniendo como elección aquel que represente un mayor cambio en su valor; en este caso particular, no se vio un punto de inflexión notorio, por lo que se eligió mantener en pocos grupos (5 clústers); dicho diagrama se presenta a continuación:

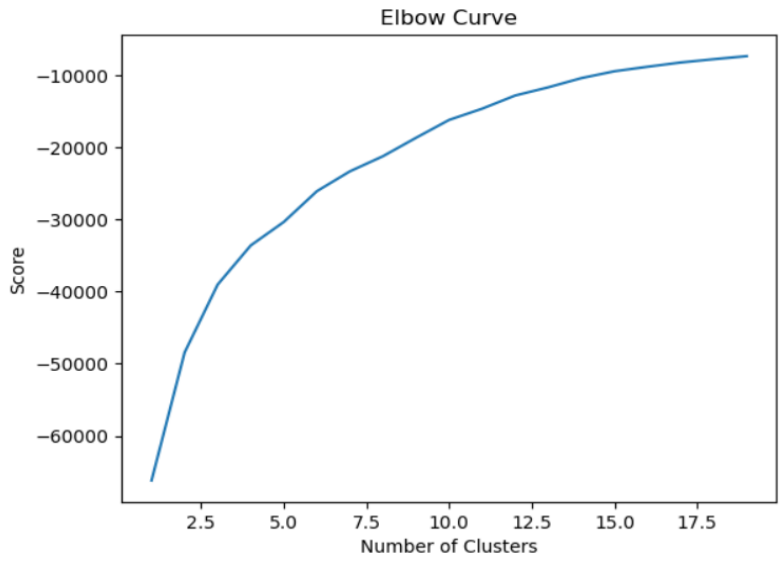


Figura 3.9 K-Medias - Diagrama de codo
Fuente: Elaboración Propia

Los resultados fueron buenos a nivel general, pero tiende a agrupar demasiados clientes promedio en un solo grupo, en vez de segmentarlos más equitativamente por sus características. Luego se procedió a correr el algoritmo de DBSCAN con distintos hiperparámetros. DBSCAN generó 5 grupos al igual que k-medias (por decisión del algoritmo y los parámetros de distancia). Los grupos obtenidos son más dispersos que el de k-medias, y además los outliers son colocados en un grupo separado, es decir más robusto por el tema de las densidades de los datos.

Agrupación K-medias	Cantidad de Clientes
Clúster 1	1
Clúster 2	16
Clúster 3	17
Clúster 4	602
Clúster 5	2012

Tabla 3.16 Agrupación K-medias
Fuente: Elaboración Propia

Agrupación DBSCAN*	Cantidad de Clientes
Clúster 1	106
Clúster 2	173
Clúster 3	599
Clúster 4	810
Clúster 5	838

Tabla 3.15 Agrupación DBSCAN
Fuente: Elaboración Propia

**Se descartaron 122 clientes de los grupos en DBSCAN por ser outliers.*

3.4.1 Selección del Modelo

Finalmente, al analizar las diferentes segmentaciones realizadas por los dos modelos, se decidió tomar como modelo final la segmentación de DBSCAN, por dos razones: K medias es muy sensible a los outliers a diferencia de DBSCAN que los identifica automáticamente, y segundo, porque al analizar el tipo de segmentación que estaba realizando el DBSCAN está cualitativamente más alineado a los objetivos del proyecto de definir una prioridad a las alertas.

3.4.2 Caracterización de los Grupos

Una vez seleccionado el modelo, en este caso particular el resultado de segmentación por DBCAN, se procede a realizar un etiquetado de los grupos por cada uno de los clientes, para finalmente poder caracterizar dichos grupos con medidas de tendencia central, dispersión y posición, como lo son el promedio, la mediana, la desviación estándar, cuantiles, etc. Dicho proceso dio como resultado una división de 5 grupos, los cuales están caracterizados de la siguiente forma:

Clúster	Descripción de Grupo	Nivel de Urgencia
1	Grupo de Clientes de tamaño grande con muchos empleados, con una gran cantidad de servicios activos (el mayor de todos los grupos), que disponen de la mayor tasa de activaciones nuevas y así mismo cancelaciones y soportes. Empresas que tienen una distribución grande en dólares en la parte financiera.	URGENCIA MAYOR
2	Grupo de Clientes así mismo de tamaño grande, y que tienen una gran distribución en dólares en la parte financiera (Clientes Potenciales), pero que no representa mucho en la parte de Facturación, Cantidad de Servicios Activos, Servicios Nuevos, Soportes y Cancelaciones, con respecto al primer grupo.	URGENCIA MENOR
3	Grupo de Clientes de tamaño mediano, el cual abarca la mayoría de los clientes de la empresa “AAA”, contiene valores promedio en cuanto a Facturación, Cantidad de Servicios, así como en la parte financiera. Toma	IMPORTANTE
4	Grupo de Clientes de tamaño Pequeño, los cuales no tienen tantos servicios activos, ni tampoco generan tantos soportes o activaciones nuevas. Tienen valores reducidos en cuanto a la parte financiera.	REVISION
5	Grupo de Clientes de tamaño de Microempresa, los cuales representan la menor distribución de valores como Facturación, o cuentas contables financiera. Tienen pocos servicios, llegando a 1 o 2, por ser algún local en específico, y no generan por lo tanto muchos soportes o cancelaciones.	PENDIENTE

Tabla 3.17 Caracterización de los Grupos
Fuente: Elaboración Propia

3.5 Diseño del Prototipo de la Herramienta de Generación de Alertas

Una vez definidos todos los procesos hasta llegar al punto de tener la predicción tanto de la clasificación de contenido de las reseñas, como el análisis de sentimientos y la segmentación de los tipos de clientes, se procede a crear el flujo de automatización de la generación de las nuevas alertas, el cual ya se explicó en la sección 3.1.1, donde se obtiene una notificación en Slack, en la cual puede ver un vistazo general a la herramienta de visualización, así como el URL para observar el dashboard con más detalle en Tableau Server, el cual es la parte culminante de todo el proceso.

Dicha visualización consta de un resumen general de los resultados de todas las predicciones de una nueva reseña, como por ejemplo cada uno de los comentarios que ha dicho el cliente, la categoría de contenido del comentario, la polaridad y subjetividad del texto, un wordcloud de los principales 25 términos utilizados por el cliente en toda la llamada, además de información adicional básica, como el tamaño de la empresa, el nivel de facturación, el tipo de industria, el subtipo de industria, la antigüedad en la empresa, y por último, el segmento de tipo de cliente que es, originado por el algoritmo de segmentación. El prototipo de la herramienta se puede observar a continuación:

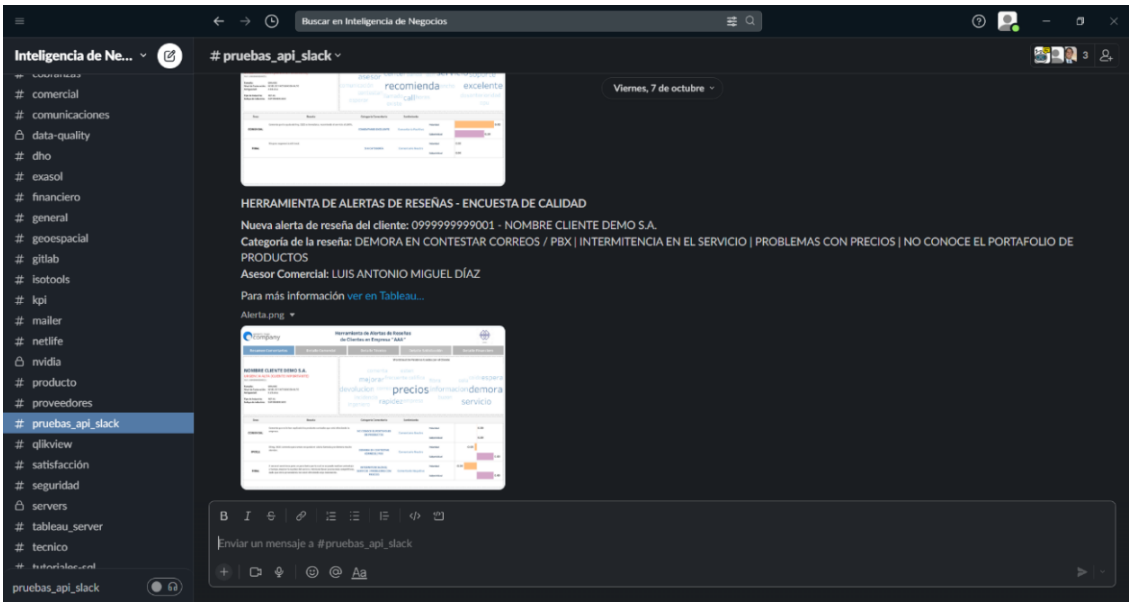


Figura 3.10 Prototipo de la Herramienta de Generación de Alertas
Fuente: Elaboración Propia

Figura 3.11 Prototipo de la Herramienta de Generación de Alertas – Resumen
Fuente: Elaboración Propia

Después del resumen de todo el análisis de Procesamiento de Lenguaje Natural, el usuario final podrá moverse por cada uno de los tabs/botones superiores, con la finalidad de obtener un detalle de cada una de las fuentes de información que tienen que ver con el cliente; en este caso, por alcance del proyecto, sólo se definieron el detalle comercial, el detalle técnico, el detalle de indicadores de Satisfacción y el detalle financiero, aunque no se descarta que en fases futuras del proyecto en la empresa "AAA", se incluyan más fuentes de información; todo con la finalidad de darle un contexto, y perspectiva más detallada de la situación del cliente. A continuación, se presentan estos detalles de información:

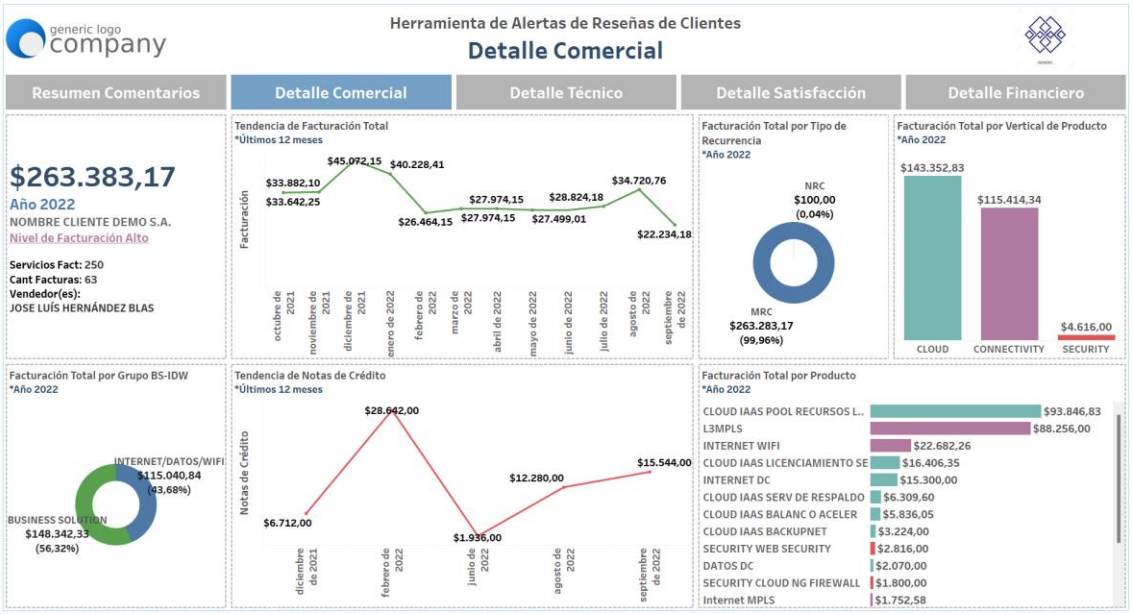


Figura 3.12 Prototipo de la Herramienta de Generación de Alertas - Detalle Comercial
Fuente: Elaboración Propia

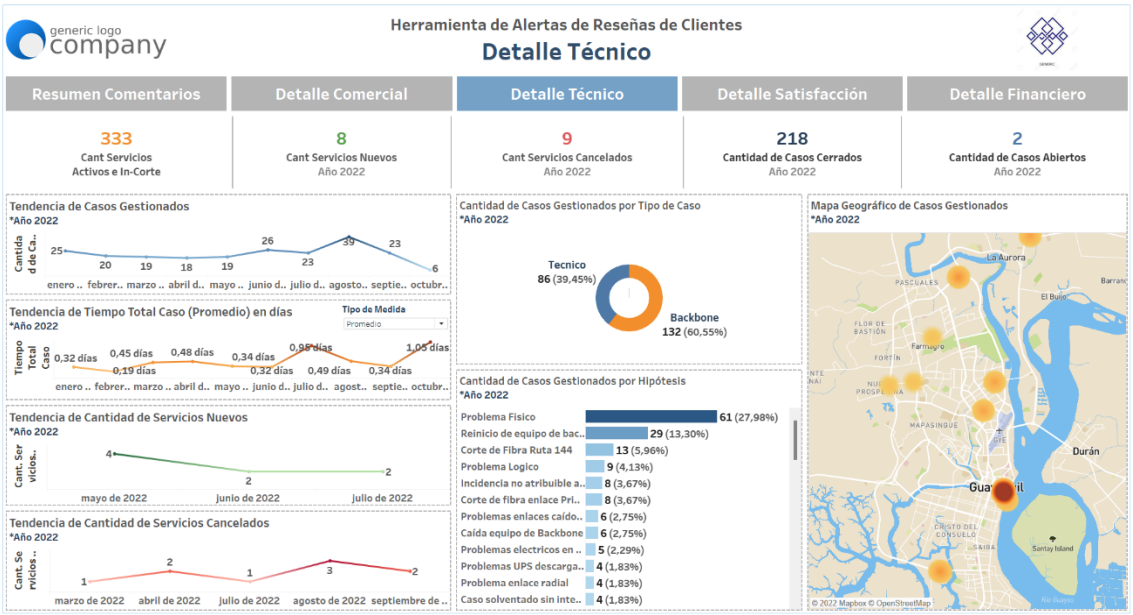


Figura 3.13 Prototipo de la Herramienta de Generación de Alertas - Detalle Técnico
Fuente: Elaboración Propia

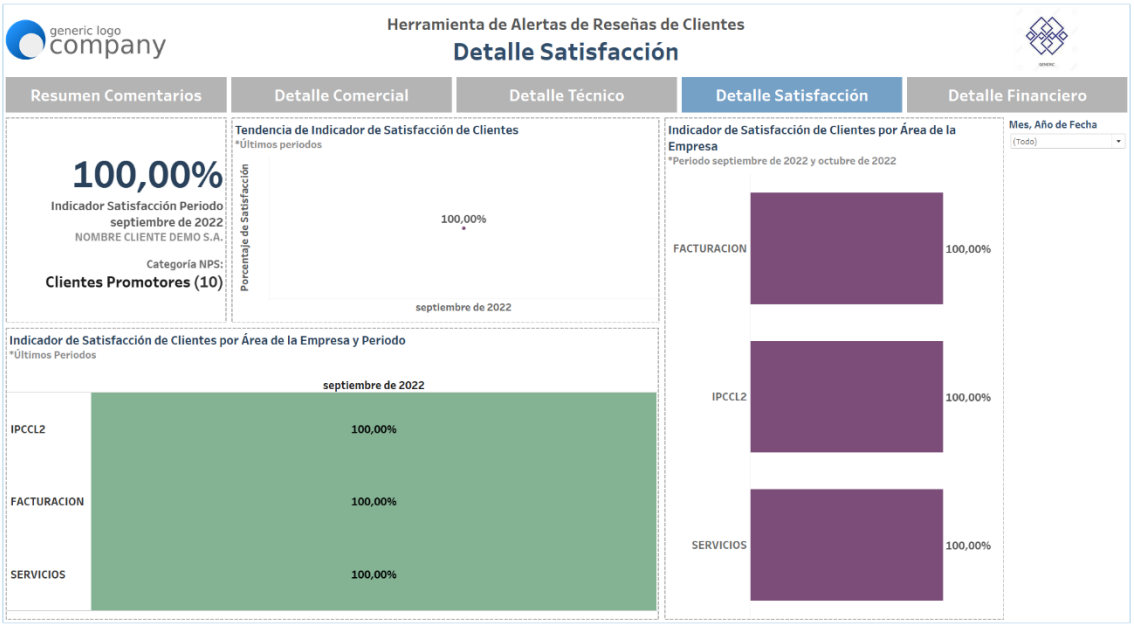


Figura 3.14 Prototipo de la Herramienta de Generación de Alertas - Detalle Satisfacción
Fuente: Elaboración Propia

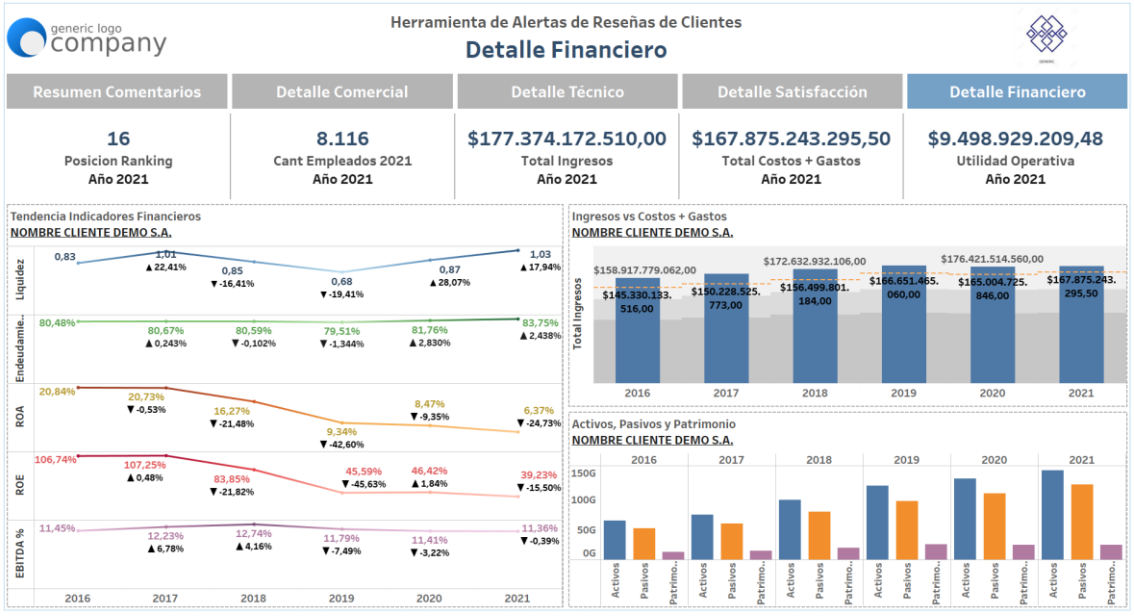


Figura 3.15 Prototipo de la Herramienta de Generación de Alertas - Detalle Financiero
Fuente: Elaboración Propia

CAPÍTULO 4

4 ANÁLISIS DE RESULTADOS

En el presente capítulo se exponen los principales resultados obtenidos de todo el proyecto, desde el análisis exploratorio de las reseñas, hasta el funcionamiento de la herramienta de generación de alertas en producción con las partes interesadas.

4.1 Análisis Exploratorio de las Reseñas de los Clientes

En primer lugar, se comienza por un análisis exploratorio de las reseñas de los clientes en la base de datos final, en este caso se puede observar un wordcloud de los principales términos utilizados en todas las reseñas de manera histórica, donde se puede destacar términos como: servicio, problema, contacto, tiempo, respuesta y demora. Estos términos permiten deducir que la mayor parte de las reseñas históricas de los clientes se refieren al campo temporal, es decir a los tiempos de respuesta, tiempos de atención, entre otras.

Figura 4.1 Wordcloud Histórico de las reseñas
Fuente: Elaboración Propia

Dividiéndolo un poco más por periodo de la reseña podemos ver lo siguiente:

Periodo 1	Periodo 2	Periodo 3
		
Figura 4.2 Wordcloud Reseñas Periodo 1 ***Fuente: Elaboración Propia***	***Figura 4.3 Wordcloud Reseñas Periodo 2*** ***Fuente: Elaboración Propia***	***Figura 4.4 Wordcloud Reseñas Periodo 3*** ***Fuente: Elaboración Propia***

Periodo 1: Comprende las reseñas de clientes que se obtuvieron entre Enero 2021 y Junio 2021, en este caso se puede observar un wordcloud de los principales términos utilizados en todas las reseñas dentro de este periodo, donde se puede destacar términos como VIP, contacto, servicio, demoran, tiempo, califica y soporte. Esto permite deducir que la mayor parte de las reseñas de los clientes dentro de este periodo se refieren al campo temporal y al servicio como tal, es decir a los tiempos de respuesta, tiempos de atención, entre otras.

Periodo 2: Comprende las reseñas de clientes que se obtuvieron entre Julio 2021 y Diciembre 2021, en este caso se puede observar un wordcloud de los principales términos utilizados en todas las reseñas dentro de este periodo. Los términos a destacar son servicio, problema, soporte, comenta, tiempo y respuesta. Esto permite deducir que la mayor parte de las reseñas de los clientes dentro de este periodo se refieren al servicio y al campo temporal, es decir a la solución de problemas, soportes, tiempos de respuesta, tiempos de atención, entre otras.

Periodo 3: Comprende las reseñas de clientes que se obtuvieron entre Enero 2022 y Junio 2022, en este caso se puede observar un wordcloud de los principales términos utilizados en todas las reseñas dentro de este periodo, donde se puede destacar términos como servicio, problema, comenta, asesora, internet y llamada, por lo cual, se podría deducir que la mayor parte de las

reseñas de los clientes dentro de este periodo se refieren al servicio y la atención por parte de los operadores, es decir a la solución de problemas, soportes, entre otras.

Análisis de Sentimientos

El análisis de sentimientos hace posible conocer el tipo de reseña que emitió el cliente, así se puede conocer mejor la actitud que tiene con respecto al servicio y la atención que recibe por parte de los asesores.

Para poder clasificar por tipo de sentimiento las reseñas emitidas por los clientes primero es necesario conocer la distribución de polaridad de las reseñas, al analizar el histograma de Polaridad se observa que la mayor parte de las polaridades se encuentran entre -0.2 y 0.2, de acuerdo a esto se puede mencionar que en su gran mayoría los comentarios se clasifican como neutros, es decir existe cierta indiferencia entre los clientes con respecto a la atención recibida por parte de la empresa, sin embargo, es posible visualizar que en hay un grupo considerable de clientes que emiten reseñas positivas.

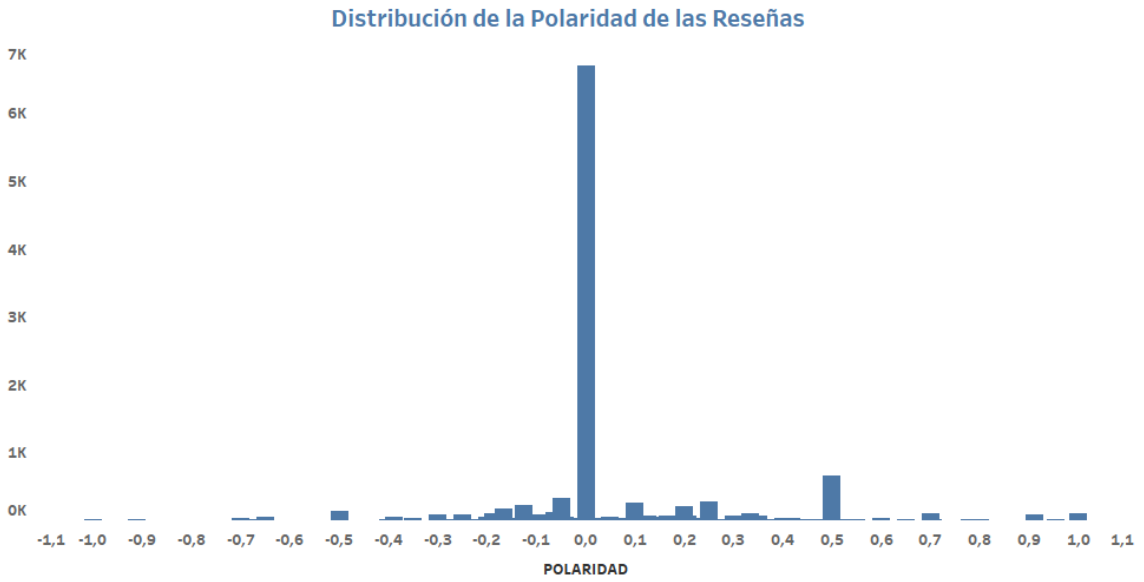


Figura 4.5 Distribución de la Polaridad de las Reseñas
Fuente: Elaboración Propia

El gráfico de proporción determinó que la clasificación de sentimientos es la siguiente: 75,67% de comentarios fueron de carácter neutral es decir la gran mayoría, seguidos por 18.38% de comentarios positivos y un 5,95% de comentarios negativos.

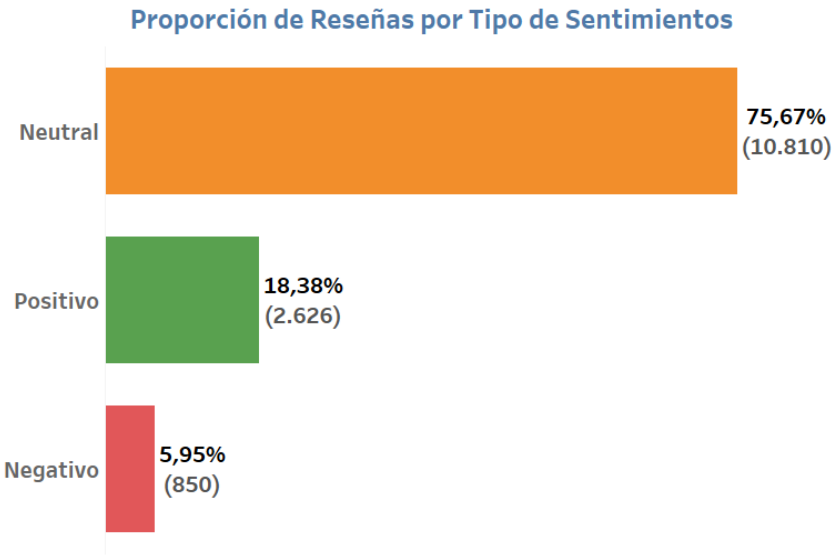


Figura 4.6 Proporción de Sentimientos por Tipo
Fuente: Elaboración Propia

Así mismo es necesario conocer la manera en que las reseñas se comportan en el tiempo según su tipo, por lo cual se analizó la evolución de la proporción de reseñas según el tipo de sentimiento en que se clasifican, es notable que la gran mayoría de reseñas tuvieron la característica de ser neutrales, siendo Abril 2021 y octubre de 2021 los meses en que se presentaron más reseñas de este tipo, así mismo, las reseñas positivas tuvieron un ligero incremento en Noviembre de 2021, Enero del 2022 y Marzo de 2022, mientras que la evolución de casos de reseñas negativas se ha mantenido en valores bajos a lo largo del tiempo.

.

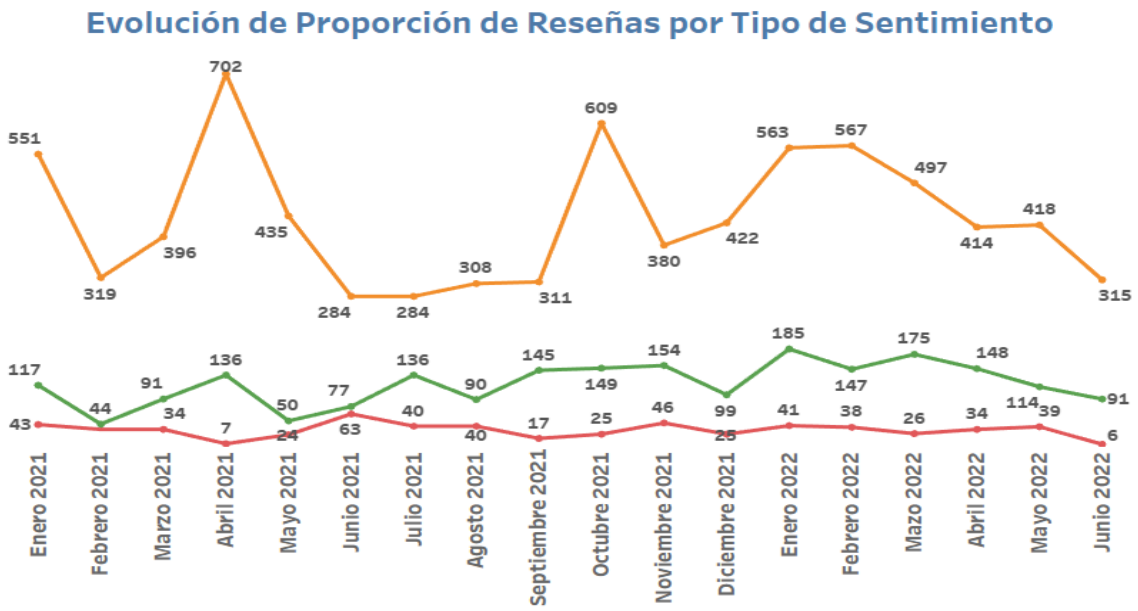


Figura 4.7 Evolución de Proporción de Reseñas por Tipo de Sentimientos
Fuente: Elaboración Propia

Dentro de las reseñas analizadas, la gran parte son de empresas que tienen sus puntos de conexión en Guayaquil y Quito, por lo que es importante analizar la densidad geográfica en dichas ciudades de aquellas reseñas de forma histórica con categorías de sentimientos negativos / categorías de quejas y aquellas con comentarios de sentimientos positivos / categorías buenas.

Guayaquil

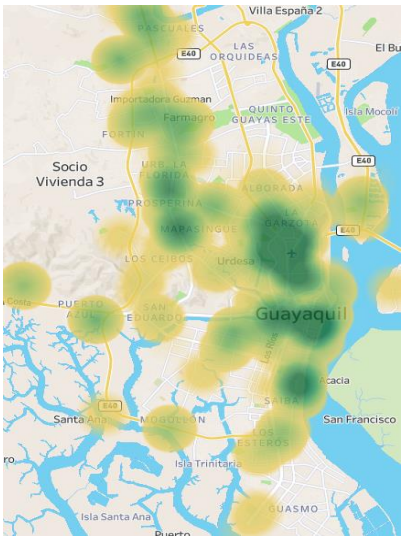


Figura 4.8 Guayaquil Categorías Buenas
Fuente: Elaboración Propia

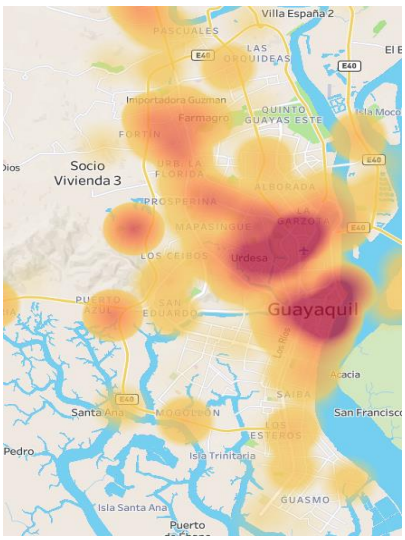


Figura 4.9 Guayaquil Categorías Malas
Fuente: Elaboración Propia

Quito

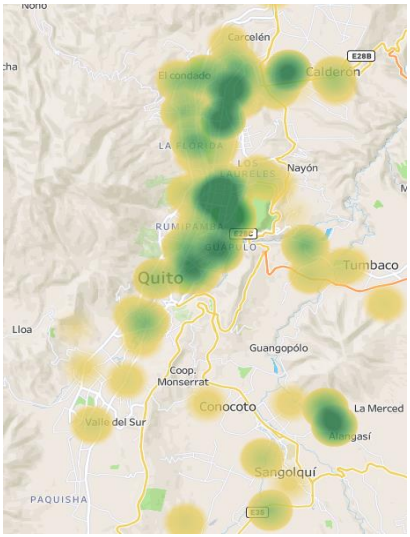


Figura 4.10 Quito Categorías Buenas
Fuente: Elaboración Propia

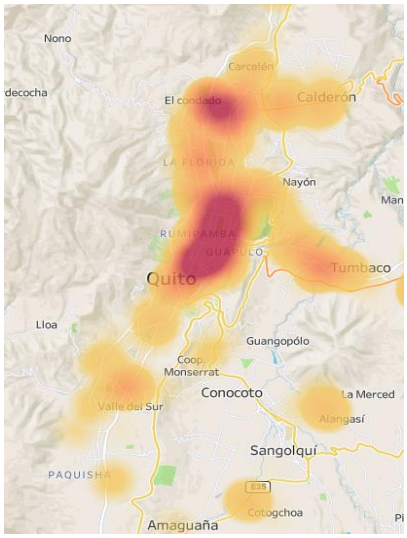


Figura 4.11 Quito Categorías Malas
Fuente: Elaboración Propia

Se puede observar que existe un patrón definido en las ciudades, y es que las reseñas se generan en las zonas industriales / empresariales donde están las principales compañías, en este caso, en el centro y norte de cada ciudad, pero no existe un patrón definido que diferencie, entre que zonas alojan más comentarios positivos en contraste a los negativos, dado que una empresa que ha dejado una reseña positiva en una encuesta, en el siguiente periodo puede dejar una reseña negativa por un problema puntual, no parece existir un problema de zonificación, como antes se había analizado, los problemas más recurrentes tienen que ver con los tiempos de atención, y eso es indistinto a la ubicación geográfica.

4.2 Pruebas de Funcionalidad

Como uno de los pasos finales del proyecto, se procedió a entregar la herramienta de generación de alertas a la sección de customer experience de la empresa “**AAA**”, instalando en sus dispositivos de trabajo y móviles, el software de slack para las notificaciones, a su vez se le dieron los permisos necesarios para entrar a un canal de comunicación de un espacio de trabajo de slack, en el cual se envían las alertas de las reseñas; así mismo, tienen a su disposición las credenciales con las licencias de lectura, del servidor de Tableau que se tiene a disposición para la visualización del detalle de las predicciones e información

general de los clientes.

Una vez automatizado todo el proceso de predicciones con un orquestador de tareas, que se encarga desde el proceso de extracción de la data de encuestas nuevas, mediante una API desde Question Pro hacia la base de datos analítica EXASOL, hasta la depuración, preprocesamiento y predicción de la data transformada del texto, para la creación de una nueva tabla, que pueda ser consultado en tiempo real por Tableau, y este manda una alerta hacia slack con las encuestas que hayan generado una categoría de contenido requerida por la sección de customer experience. Todo este proceso de ETL que se realiza con un orquestador de tareas se ejecuta en un tiempo máximo de 5 minutos en total. La frecuencia de ejecución se estableció en periodos de 2 horas, no por limitaciones en cuanto a recursos informáticos, sino por restricciones del número de llamadas a la API de Question Pro, las cuales generan costos adicionales si se excede un número límite de llamadas al mes (No se quiere generar más costos al proyecto).

La forma en la cual se interactúa con la herramienta es la siguiente, primero al usuario de customer experience le llega una notificación de alerta en slack, por alguna reseña en particular, el mensaje que le llega tiene información del cliente como el nombre, la categoría de la reseña que se ha expuesto, así como una captura del dashboard principal, la cual puede agrandar para obtener más detalle, y por último, un link hacia Tableau server, en el cual podrá ver la información respectiva de dicho cliente, tanto de las reseñas como de aspectos importantes del cliente en distintos ámbitos.

Figura 4.12 Notificación de Alerta en Slack
Fuente: Elaboración Propia

Una vez hecho clic en el enlace del dashboard, se puede visualizar lo siguiente:

Figura 4.13 Dashboard Principal
Fuente: Elaboración Propia

En dicha visualización se puede ver la siguiente información: el nombre del cliente, el RUC del mismo, el nivel de urgencia / prioridad dado por la segmentación, el tamaño de la empresa, el nivel de facturación, la antigüedad del cliente con la empresa, el tipo y subtipo de industria. El componente principal de la visualización, resultado de la analítica de texto, es el wordcloud de los principales términos usados por el cliente durante toda la encuesta, los comentarios dados por el cliente en cada una de las secciones, junto con la categoría de contenido (la cual generó alerta), también el tipo de sentimiento, la polaridad y la subjetividad de la reseña. Los componentes de la visualización fueron expuestos mediante reuniones con el personal del área de customer experience. Los usuarios se mostraron visiblemente satisfechos con la herramienta y cada uno de sus componentes. Como valor añadido al proyecto, se puede observar a su vez dentro de las mismas visualizaciones ciertos botones que redirigen a otras visualizaciones de información general del cliente, como el Detalle Comercial, Detalle Financiero, Detalle Técnico y el Detalle de Satisfacción.

Figura 4.14 Flujo del Dashboard
Fuente: Elaboración Propia

Cada una de las distintas partes proporciona información fundamental. Por ejemplo en la parte comercial indica la facturación en el último año, así como una tendencia de la facturación y notas de crédito de los últimos doce meses, los productos facturados, vendedores, entre otras cosas; la parte técnica indica la cantidad de servicios activos del cliente, la cantidad de cancelaciones, la cantidad de servicios nuevos, la categoría de los soportes generados, y así mismo un mapa de los soportes en sus distintas ubicaciones; en la parte de satisfacción expone la satisfacción en los últimos periodos, un mapa de calor del indicador por área y periodo, entre otras cosas; y por último, la parte financiera, presenta los indicadores financieros principales; EBITDA, Liquidez, ROE, ROA, endeudamiento del cliente, así como los ingresos, costos, gastos, activos, pasivos y patrimonio del mismo. Cada una de las distintas partes tiene su respectiva base relacionada, la cual se puede descargar en formato Excel.

Toda la herramienta en general fue diseñada teniendo en cuenta las necesidades del cliente interno del proyecto: generar valor usando datos no explotados anteriormente como las reseñas, crear una visión 360 de los distintos ámbitos del cliente para complementar los planes de acción; todo el proyecto fue presentado al cliente, y se sintieron muy a gusto con la herramienta. A la vez como resultado de las reuniones con los usuarios se identificaron propuestas de mejoras puntuales y nuevos requisitos funcionales a implementar en el futuro. Actualmente se continúa usando la versión beta de la herramienta en la sección de customer experience de la empresa “AAA”.

4.3 Evaluación de la solución propuesta

Una vez realizado la herramienta de generación de alertas y poniéndola a prueba en una versión beta con los usuarios finales, de la sección de customer experience de la empresa “AAA”, es de vital importancia evaluar el funcionamiento de la misma con reseñas nuevas de los clientes. Para la evaluación, se decidió usar la técnica A/B Testing, la cual tendrá como objetivo determinar si existe una diferencia estadísticamente significativa entre el número de alertas generadas para las 10 categorías de contenido / alertas por parte de los operarios telefónicos, en contraste con las alertas generadas por la herramienta, siendo cada una de las alertas revisadas por la coordinación de la sección de customer experience. Se realizaron las pruebas durante un periodo de dos semanas, con un total de 155 encuestas telefónicas realizadas por los operarios, y que a su vez la herramienta analiza mientras son ingresadas a la base, los operarios no son avisados que la herramienta está en funcionamiento, para evitar posibles sesgos de comportamiento, las alertas generadas por ambos fueron analizadas por la coordinación de la sección de customer experience para evitar posibles falsos negativos, y los resultados fueron los siguientes:

Recurso	Encuestas Analizadas	Alertas Generadas
Operarios Telefónicos	155	26
Herramienta de Alertas	155	38

Tabla 4.1 Resultados de la Evaluación
Fuente: Elaboración Propia

**Total de Alertas Reales Validadas: 41 (Coordinación de CX)*

Con dichos resultados, se puede realizar una prueba de hipótesis de diferencia de proporciones de una sola cola con el objetivo de determinar que no existe un efecto aleatorio de la herramienta la cual pudo verse afectada; con una confianza del 95%, se tiene que la diferencia es estadísticamente significativa (con un valor p de 0.045), por lo tanto, la herramienta está generando alertas que no habían sido consideradas por los operarios telefónicos, es decir, tiene un mayor rendimiento.

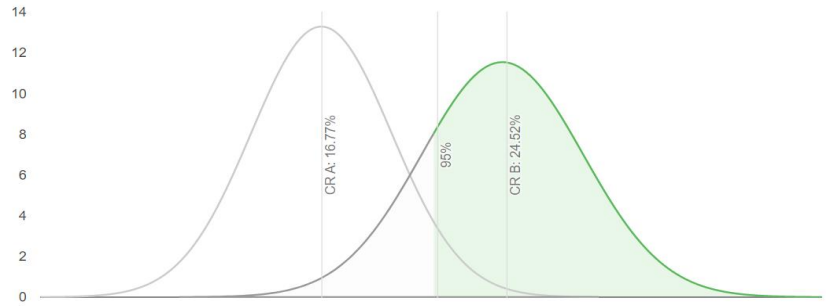


Figura 4.15 Prueba de hipótesis de diferencia de proporciones
Fuente: Calculadora de prueba A/B gratuita de ABTestGuide.com

Adicional a esto, para el proyecto se realizó un análisis costo-beneficio de este, teniendo en cuenta las siguientes estimaciones de Beneficios y Costos:

Beneficios

Para cuantificar los beneficios del uso de la herramienta, se midió el ahorro considerando el tiempo ahorrado por el Coordinador estableciendo que este beneficio monetario es de $176 al mes.

- Antes un coordinador de customer experience dedicaba de 6 a 8 horas a la semana a analizar las alertas generadas por los operadores de una forma manual, con correos enviados a los operadores.
- Ahora el coordinador le toma la mitad de ese tiempo en analizar las alertas de forma más sencilla y amigable al usuario con una herramienta de gestión (total de tiempo 3 a 4 horas a la semana).
- Coordinador de CX tiene un salario de $11 por hora considerando 16 horas ahorradas al mes, resulta en un ahorro de $176 al mes ahorrado.

Beneficios por Tiempo Ahorrado de Operadores Telefónicos: $860 al mes

- Antes un operador telefónico dedicaba dos horas al día generando alertas de forma manual, mediante correos a la coordinación de customer experience, tomando en cuenta todas las encuestas del día.
- Ahora un operador no tiene que dedicarse a realizar alertas manuales, ahorrando ese tiempo, en realizar más encuestas al día.
- Operador Telefónico > $4,30 por Hora de Salario > 40 horas Ahorradas al mes por operador > 5 Operadores Telefónicos Activos en la empresa > $860 al mes ahorrado.

Beneficios por evitar posibles soportes generados por una oportuna gestión: $2400 al mes

- La sección de customer experience informó para el proyecto que el costo promedio de un soporte para la empresa es de $20.
- Con la nueva gestión de alertas automatizadas se espera evitar 120 soportes nuevos por mes (estimación calculada por la sección de customer experience), y a su vez cancelaciones de servicios por una oportuna gestión.

Existen otras fuentes de beneficios que no se pueden cuantificar y a las que por lo tanto no se puede asociar un valor monetario fijo, pero es válido comentarlas:

- Una mejor experiencia del cliente, la cual se puede medir mediante el NPS, pero que se podría ver el cambio a mediano o largo plazo en cada uno de los clientes, en un próximo periodo de encuestas.
- Posibles deserciones evitadas por una buena gestión a tiempo, dado que habría que realizar un análisis de Churn mucho más específico que permita evaluar el impacto de la herramienta, lo cual, está fuera del alcance del proyecto.
- El cambio de opinión del cliente hacia la marca o producto relacionado, así como una posible venta cruzada de otros productos, lo cual no es inmediato.

Costos

Recursos Informáticos: $0

- Dado que la empresa “AAA”, es una empresa de telecomunicaciones que cuenta con los suficientes servicios informáticos, como servidores, licencias de software (Tableau, Slack), bases, etc., no se lo incluye como costo directo del presente proyecto.

Recursos Humanos: $7200

- Ingeniero de Datos > Salario de $1800 mensual > 2 meses de trabajo > $3600 en total.

- Científico de Datos > Salario de $1800 mensual > 2 meses de trabajo > $3600 en total.

Costo de una Alerta No Generada: $800 ($100 por alerta / mes)

- Se definió junto con la sección de customer experience un valor de $100 por una reseña de categoría negativa no alertada, que toma en cuenta posibles soportes generados, cancelación de algún servicio, o afectación a la marca. No se toma en cuenta la deserción completa del cliente, el cual puede sesgar el costo, y no siempre se puede comprobar la posible deserción.
- Se toma en cuenta que según la categoría de alerta pueden existir distintas cantidades de alertas no generadas, pero de manera global no pasan las 8 alertas no generadas por mes (en el peor de los casos), en la prueba piloto con la herramienta.

Costo / Beneficio y Retorno de la Inversión

Tomando en cuenta todos los beneficios y costos antes expuestos, se puede observar en la siguiente tabla, que el retorno de la inversión se generará en el 6to mes de funcionamiento del proyecto (No se toma en cuenta la inflación mes a mes).

	1er Mes	2do Mes	3er Mes	4to Mes	5to Mes	6to Mes	7to Mes
Beneficio Acumulado	$3436	$3436	$3436	$3436	$3436	$3436	$3436+ $1416
Costo Acumulado	$8000	$8000 + $4564	$800 + $9128	$800 + $6492	$800 + $3856	$800 +$1220	$800
Diferencia	-$4564	-$9128	-$6492	-$3856	-$1220	$1416	$4052
Proporción	0.43	0.27	0.35	0.47	0.74	1.7	6.07
Retorno	No	No	No	No	No	Si	Si

Tabla 4.2 Costo / Beneficio y Retorno de la Inversión
Fuente: Elaboración Propia

Gráficamente se puede observar lo antes mencionado, entre el 5to y 6to mes se igualan los costos con los beneficios del proyecto, y a partir del 6to mes se empieza a generar el retorno de la inversión, obteniendo ganancias conforme avance el tiempo del proyecto, el cual está previsto para funcionar por lo menos un año.

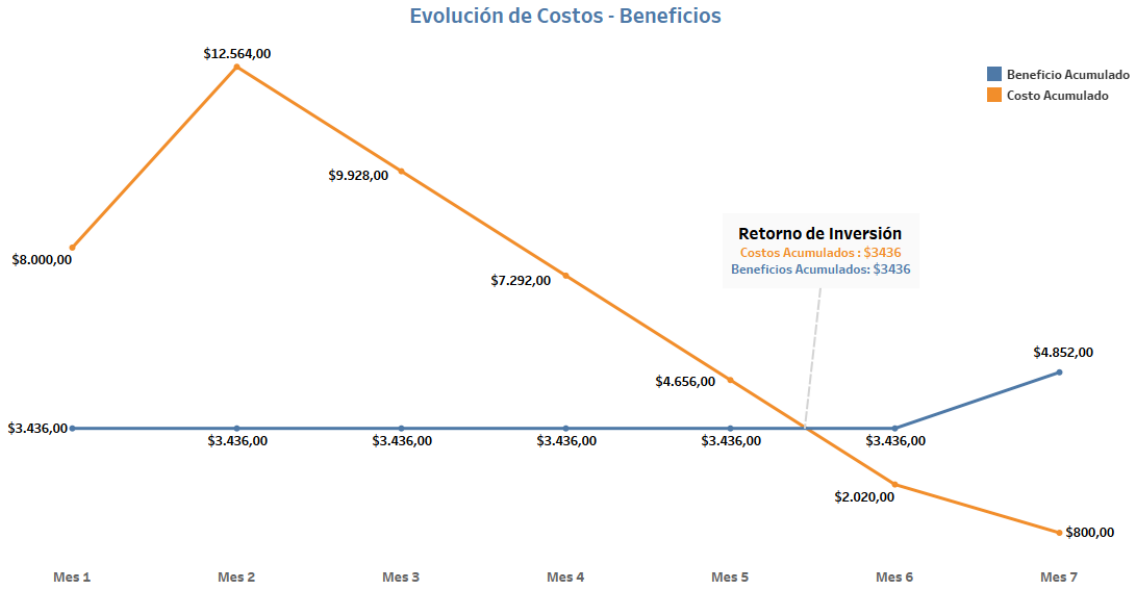


Figura 4.16 Evolución de Costo - Beneficio
Fuente: Elaboración Propia

5 DISCUSIÓN

Una de las principales problemáticas del proyecto al cual se orientó a solucionar, fue el análisis de datos antes no explorados como lo son las reseñas de los clientes para obtener predicciones, que aunque se sabe que existe información valiosa ahí, ¿Qué tanto más puede aportar a los indicadores de satisfacción este análisis de reseñas, o a su vez qué tanto puede contribuir con la mejora de la experiencia del cliente?, pero algo incluso más importante que se ha cuestionado en el transcurso del proyecto es ¿A qué debo darle mayor ponderación en el análisis posterior, al indicador registrado en la encuesta de satisfacción, o lo que dicen los clientes de la empresa en las reseñas, es decir sus sentimientos, sus palabras?, y la respuesta en este proyecto es que, a ambos, el uno se complementa con el otro, es como comparar el uso de la media en un análisis estadístico descriptivo, sin tener la varianza como soporte; el analizar un aspecto general de la empresa con el indicador está bien, da un gran resumen general de la actualidad, pero esto no explora todo el problema, si es que no lo complementas con el análisis de los sentimientos / el NLP de las reseñas, no sólo por las oportunidades de mejora o quejas que los clientes puedan dar, sino porque también al analizar las reseñas se pueden dar cuenta si en verdad están hablando cosas positivas que se vean reflejadas en su calificación, o por el contrario, su calificación es muy arbitraria a lo que dicen, teniendo comentarios muy buenos pero calificaciones regulares, o a su vez, comentarios neutros con calificaciones malas en los indicadores; y este es el punto de todo ese asunto, la subjetividad puede conllevar a un sesgo en el análisis, y esto tiene una gran relevancia dentro de las empresas, dado que con esto la alta dirección o las partes interesadas generan los planes de acción para mejorar la experiencia del cliente, y si los resultados están sesgados, el plan también lo estará.

Por último, vale la pena discutir un poco acerca de la calidad de los datos, y el sesgo indirecto que se puede introducir en este tipo de estudios; una de las limitaciones que se asumieron en el proyecto, fue no trabajar con los audios de las llamadas de la encuesta de calidad, por el gran trabajo de depuración que estos debían tener para el análisis y que el tiempo salía del alcance del presente

proyecto, por lo que finalmente se optó por desarrollar el proyecto con las reseñas escritas por los operadores telefónicos dentro de un software de encuestas, este proceso aunque se lo realice de la forma más precisa posible, siempre se introducirá un sesgo por parte del intermediario, dado que no siempre se podría capturar la intensidad y la forma en que el cliente está diciendo las reseñas, y por lo tanto puede perder un poco de la semántica original de los comentarios, proceso que se puede ver reflejado en todo el pipeline del proyecto, y por lo cual los resultados serían una estimación, es decir, tendrían una incertidumbre de lo que realmente quiere expresar el cliente en su reseña; en un mundo idílico el cliente siempre tendría que responder de forma escrita la encuesta de satisfacción, pero esta al ser demasiado extensa y además que son a clientes corporativos (los cuales no tienen mucho tiempo), tiene que ser de manera telefónica, con el llenado de información de manera manual por el operador en el software. El objetivo siempre estará en tratar de reducir esa incertidumbre de la cual se habló, a través de buenas prácticas en la obtención de datos, como la escritura precisa de los comentarios de los clientes, sin "suavizar" expresiones o alterar la realidad, como la grabación de las llamadas y corrección de datos tabulados posterior a escuchar nuevamente la llamada, entre otras prácticas; dichas buenas prácticas se están realizando en la sección de customer experience de la empresa "AAA", por lo que se tuvo la suficiente confianza de utilizar dichos datos para este proyecto.

6 CONCLUSIONES Y RECOMENDACIONES

Entre las principales conclusiones que tenemos del proyecto están:

- La integración de distintas herramientas, y el uso de algoritmos de aprendizaje automático derivaron en un producto de ciencia de datos, el cual obtiene provecho y valor de datos que antes no se habían utilizado, por lo que queda pendiente analizar en qué otras áreas de la empresa se podría aplicar algo parecido a este proyecto, con la finalidad de agregar valor a los datos no estructurados, y que permitan generar información valiosa para beneficio de la empresa.
- La herramienta obtuvo un mejor rendimiento que los mismos operadores telefónicos identificando posibles alertas de categorías de contenido, debido a que cada operador tiene una subjetividad a la hora de generar una posible alerta.
- El retorno de la inversión generada por el proyecto solamente será de 6 meses, pero no se toman en cuenta otros beneficios subyacentes, como la posible mejora en la experiencia del cliente, lo cual es complicado de medir y por lo tanto monetizar, pero que representa para la empresa un valor muy grande para la marca y que es igual de importante que su posible facturación.
- Los principales problemas relatados por los clientes en sus reseñas, y que se puede evidenciar por los términos usados en sus comentarios, es acerca de los tiempos, ya sea de solución, atención o respuesta, por lo que la empresa “AAA”, tiene que enfocar sus esfuerzos en tratar de resolver este tipo de problema recurrente en sus clientes, con el objetivo de mejorar la experiencia de estos.

Entre las principales recomendaciones que se tiene del proyecto están:

- Se recomienda para futuros trabajos utilizar los audios de la encuesta de satisfacción de calidad, como fuente de datos para el proyecto, dado que, con más tiempo de depuración de los audios, y así mismo con una mayor cantidad de audios, se podría reducir esa incertidumbre que existe entre lo real y lo estimado de las reseñas.
- Se recomienda crear modelos de clasificación de contenido para las demás categorías que se etiquetaron, pero que se dejaron fuera del alcance del proyecto, por la poca cantidad de historia que existe para el entrenamiento del modelo, en una fase posterior, se podría tener adicionalmente categorías que antes ni siquiera se habían contemplado.
- Se recomienda crear más canales de comunicación a las distintas áreas que conforman el customer journey, para notificar las alertas de manera más especializada, para no tener que pasar por un intermediario, que en este caso es la sección de customer experience; todo esto cuando la herramienta tenga un grado de madurez y validez adecuado.
- Se recomienda evaluar la eficacia del proyecto en un mediano plazo, dado que recién se está utilizando en la empresa “AAA” como una prueba beta para la creación de planes de acción mejor direccionados, y que repercutan de mejor manera en la experiencia del cliente.

7 REFERENCIAS BIBLIOGRÁFICAS

[1] Agarwal, A., Yadav, A., & Vishwakarma, D. (2019). Multimodal sentiment analysis via RNN variants. *IEEE international conference on big data, cloud computing, data science and engineering (BCD), pp 19–23*, págs. 19–23.

[2] Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

[3] *Bienvenido a PyCaret - PyCaret Oficial*. (s.f.). Obtenido de https://pycaret.gitbook.io/docs/

[4] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

[6] Budhwar, J., & Singh, S. (2022). Sentiment Analysis of Customers Review Using Hybrid Approach (Vol. 1528). Springer.

[7] Budwar, J., & Singh, S. (8 de Febrero de 2022). Sentiment Analysis of Customers Review Using Hybrid Approach. Communications in Computer and Information Science (CCIS, volumen 1528).

[8] Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine, 9(2), 48–57.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data, 785-794.

[10] Chomsky, N. (1956). Three models for the description of language. IRE Trans Inform Theory, 2(3), 113-124.

[11] Chowdhary, K. (2020). Fundamentals of Artificial Intelligence.

[12] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. J. Mach. Learn. Res., 12, 2493–2537.

[13] Computer Generated Solutions Inc. (CGS). (28 de junio de 2022). 22 Customer Experience Stats for 2022. https://www.cgsinc.com/blog/22-customer-experience-stats-2022

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[15] Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery. Proceedings of the Twelfth International Conference on World Wide Web - WWW, 3.

[16] Fernández, J., Gutiérrez, Y., Gómez, J. M., & Martínez-Barco, P. (Septiembre de 2015). Social Rankings: an´alisis visual de sentimientos en redes sociales. Procesamiento de Lenguaje NaturalTomo 55, páginas 199 - 2021 de septiembre de 2015.

[17] Fix, E., & Hodges, J. (1951). An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation. International Statistical Review, 57(3), 233-238.

[18] Gentleman, R., & Carey, V. J. (2008). Bioconductor case studies. En R. &. Gentleman, Bioconductor case studies (págs. 137-157). New York, NY.: Springer.

[19] Hirschberg, J., & Manning, C. (2015). Advances in natural language processing. Science, 349(6245), 261–266.

[20] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, 328–339.

[21] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification using Machine Learning Techniques. WSEAS Transactions on Computers.

[22] Inmon, B; Levins, M; Kermany, A; Sanky, M. (22 de julio de 2021). Improving Patient Insights With Textual ETL in the Lakehouse Paradigm. https://www.databricks.com/blog/2021/07/22/improving-patient-insights-with-textual-etl-in-the-lakehouse-paradigm.html

[23] Khan, K., Rehman, S., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. The Fifth International Conference on the Applications of Digital Information and Web Technologies, (págs. 232-238).

[24] Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data, 415–463.

[25] MF Porter, 1980, An algorithm for suffix stripping, Program, 14 (3) pp 130-137.

[26] Nadkarni, P., Ohno-Machado, L., & Chapman, W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544–551.

[27] NLTK __ Natural Language Toolkit. (s.f.). Obtenido de https://www.nltk.org/

[28] Noriega, L. (17 de November de 2005). Multilayer Perceptron Tutorial. *School of Computing Staffordshire University*.

[29] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2, 1-135.

[30] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. EMNLP, 2.

[31] pandas - Python Data Analysis Library. (s.f.). Obtenido de https://pandas.pydata.org/about/index.html

[32] Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*, 81-132.

[33] Román, V. (12 de Junio de 2019). Aprendizaje No Supervisado en Machine Learning: Agrupación.

[34] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47.

[35] scikit-learn_ machine learning in Python — scikit-learn 1.0.2 documentation. (s.f.). Obtenido de https://scikit-learn.org/stable/

[36] Tableau Software, L. E. (2003-2022). ¿Qué es Tableau? Obtenido de https://www.tableau.com/es-es/why-tableau/what-is-tableau#video

[37] TextBlob_ Simplified Text Processing — TextBlob 0.16.0 documentation. (2020). Obtenido de https://textblob.readthedocs.io/en/dev/

[38] Torres, J. (2013). Clasificación Automática de Respuestas a foros de discusión de acuerdo con el dominio cognitivo de la taxonomía de Bloom en SIDWEB 4, empleando minería de texto y algoritmos de aprendizaje. Guayaquil: Informe de Proyecto de Graduación ESPOL.

[39] Turing, A. (1950). Computing machinery and intelligence. Obtenido de The Alan Turing Internet Srapbook.

[40] Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

[41] Ulloa Sepulveda, M. A., Rosales Ferreira, K. A., & Bermúdez Navarrete, D. A. (2018). Propuesta de sistema de fidelización de clientes en la empresa de telecomunicaciones Avantel a través de indicador NPS (net promotor score) usando análisis de sentimientos en Facebook.

[42] Wankhade, M., Rao, A. C., & Kulkarni, C. (07 de Febrero de 2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review.

[43] Zu, G., Ohyama, W., Wakabayashi, T., & Kimura, F. (2003). Accuracy improvement of automatic text classification based on feature transformation. Proceedings of the 2003 ACM Symposium on Document Engineering.

8 GLOSARIO

Matriz Document-Term.- La matriz término-documento es una matriz descriptiva que permite medir la frecuencia con la que los términos aparecen en una colección de documentos, en esta matriz las filas son los documentos y las columnas son los términos.

Ground Truth.- Algo cierto o válido, una realidad que requiere ser modelada mediante un algoritmo.

Customer Experience.- Se define como la percepción del cliente al interactuar de cualquier manera con una empresa, producto o servicio y expresar su sentir con respecto a dicha interacción.

Input.- Es un conjunto de datos que es proporcionado como la entrada para una aplicación informática.

Datawarehouse.- Un almacén de datos es una estructura física que permite recolectar datos que pueden provenir de varias fuentes simultáneas, los datos se almacenan con un objetivo específico y facilitan la toma de decisiones dentro de una empresa.

Merge.- Es una sentencia usada en base de datos relacionales la cual facilita agregar nuevos registros o unir entre distintas entidades de datos.

Features.- Son características o atributos los cuales se asocian a los datos.

One-Hot-Encoding.- Es la representación en forma de vector de un grupo de datos, en el cual se asignan valores booleanos. En este tipo de representación solo un elemento es asignando con el valor de 1 y el resto son 0. El elemento al que se le asigna el valor de 1 representa un verdadero.

SMOTE.- Técnica de manejo de datos que se usa con el fin de realizar sobremuestreo y ayuda a aumentar el número de los casos poco comunes.

Upsampling.- Técnica de manejo de datos utilizada para tratar el desequilibrio de clases de los mismos, de esta manera se les asigna pesos similares.

Outlier.- Se define como outlier a un dato u observación anómala o poco común debido a que se encuentra separada o muy distante de las demás observaciones.

Wordcloud.- Es una técnica de representación visual de texto la cual asigna un tamaño mayor a las palabras que son utilizadas dentro del texto con mayor frecuencia. Mediante esta herramienta es posible determinar el contexto real de un texto.

Dashboard.- Es una herramienta de visualización que proporciona una forma ágil de gestionar la información estratégica de indicadores y métricas que se manejan dentro de una empresa, de esta manera es posible realizar seguimiento a casos y facilitar la toma de decisiones estratégicas.

Validación cruzada.- Técnica usada frecuentemente en la evaluación y validación de los resultados de un análisis estadístico con el fin de determinar si existe o no independencia entre los datos de entrenamiento y prueba.

Voting Classifier.- Clasificador de Votaciones es una técnica de aprendizaje automático mediante la cual se entrenan simultáneamente un conjunto de modelos y se pronostica una salida mediante la probabilidad más alta de la clase elegida como salida.

A/B Testing.- Es una herramienta mediante la cual se evalúa dos o más versiones de una aplicación/producto y permite determinar cuál de ellas es la que arroja mejores resultados, se usa comúnmente en marketing digital

API.- Acrónimo de Application Programming Interface o en español Interfaz de Programación de Aplicaciones, se define como un conjunto de reglas y protocolos utilizados para poder generar comunicación entre dos aplicaciones de software y de esta manera se puedan ejecutar diversas funciones.

Pipeline.- Término informático en inglés que puede ser traducido al español como tubería, se usa para definir un flujo de trabajo, este término hace referencia a la manera de organizar y administrar la información dentro de un proyecto.

ETL.- Siglas en inglés de las palabras Extraer, Transformar y Cargar, las cuales forman parte del proceso mediante el cual se mueven datos desde múltiples fuentes para procesarlos y cargarlos en otra fuente de datos que permita su análisis.

Regresión Logística.- Es un tipo de regresión que permite pronosticar el resultado de una variable categórica (en la mayoría de los casos, una variable binaria) a partir de variables independientes, mediante esta regresión es posible determinar la probabilidad de un evento a partir de otras variables.