Automatización de los procesos de extracción, transformación, carga de sucursales y la categorización de los comentarios de los clientes del área de Inteligencia Experiencia del Cliente de Bancolombia S. A

Stiven Agudelo Tabares

Informe de práctica para optar al título de Ingeniero Industrial

Asesor
Olga Cecilia Úsuga Manco, PhD en Ciencias - Estadística

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería Industrial
Medellín
2023

Cita	(Agudelo Tabares, 2023)
Referencia **Estilo APA 7 (2020)**	Agudelo Tabares, S. (2023). *Automatización de los procesos de extracción, transformación, carga de sucursales y la categorización de los comentarios de los clientes del área de Inteligencia Experiencia del Cliente de Bancolombia S. A* [Semestre de industria]. Universidad de Antioquia, Medellín.

Créditos a escenario de prácticas, personas, proyectos que aportaron al desarrollo de la práctica (interna y externamente: empresa y área de la empresa, grupo de investigación, proyecto, organización)

Dedicatoria

Dedico este proyecto a mi amada madre en reconocimiento por todo el apoyo incansable que me brindó durante mi proceso educativo para alcanzar este título y quien ha sido mi apoyo incondicional en cada uno de mis logros. Gracias por haberme enseñado la importancia del trabajo duro, la perseverancia y la dedicación, valores que han guiado cada uno de mis pasos. Todo lo que he logrado es gracias a su dedicación y amor incondicional, y por eso, este logro es para ti. Espero que este proyecto sea el reflejo de todo el cariño, esfuerzo y aprendizaje que me has brindado. Agradezco a Dios por haberme permitido recibir de ti los valores que me han hecho crecer como ser humano y por inspirarme a ser cada día mejor. Te amo con todo mi corazón y sin ti, nada de esto habría sido posible.

Quiero también dedicar este proyecto a mi novia. Ella ha sido un gran apoyo y motivación en todo momento, siempre ha creído en mí y me ha animado a seguir adelante en cada paso que he dado. Agradezco su amor, su paciencia y su incondicionalidad. Gracias por ser mi roca en los momentos difíciles y por ser una persona tan maravillosa.

Agradecimientos

Quiero primeramente expresar mi más sincero agradecimiento a Bancolombia por concederme la oportunidad de realizar mis prácticas profesionales y desarrollar mi proyecto en su prestigiosa institución líder del sector financiero. A la Universidad de Antioquia, por brindarme la oportunidad de aprender y desarrollarme en un ambiente académico de calidad. El haber sido estudiante de esta institución ha significado mucho para mi formación profesional y personal. Además, quiero agradecer a Diego Osorio que me acompañó en la ejecución del proyecto, quien desde el principio me brindó su apoyo y guía, su conocimiento y dedicación fueron fundamentales para culminar este proyecto de manera satisfactoria. También quiero hacer extensivo mi agradecimiento a mi asesora Olga Úsuga, quien, desde su experiencia en el área de investigación, me brindó su asesoría y orientación en el desarrollo del proyecto, permitiéndome así obtener los resultados esperados. Por último, quiero agradecer al asesor externo Daniel Zuluaga, quien con una visión fresca y experta hizo que este proyecto tuviera un enfoque más amplio y una perspectiva innovadora que permitió el éxito de este.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

ETL	Extract, Transform, Load
PLN	Procesamiento de Lenguajes Naturales
NLP	Natural Language Processing
S. A	Sociedad Anonima
APA	American Psychological Association
PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia
NLTK	Natural Language Toolkit
DW	Data Warehouse
RE	Regular Expression Operations
LZ	Landing Zone

Resumen

El presente proyecto se llevó a cabo en la Gerencia de Inteligencia Experiencia del Cliente de la compañía Bancolombia S.A, donde se identificaron dos desafíos principales. En primer lugar, el análisis de los datos de las sucursales se encontraba limitado por el método utilizado, el cual no permitía un control efectivo del número de personas que acudían a las sucursales, la cantidad de encuestados y las respuestas obtenidas. En segundo lugar, era necesario categorizar los comentarios de los clientes con el objetivo de detectar oportunidades de mejora en la experiencia del usuario, siendo un reto identificar múltiples categorías en un solo comentario. Con el objetivo de automatizar los procesos de extracción, transformación y carga (ETL) para la experiencia de sucursales y categorización de los comentarios de los usuarios, se planteó este proyecto. Los resultados obtenidos fueron beneficiosos tanto en términos de eficiencia como de productividad, ya que se logró la automatización de procesos y la reducción del tiempo de ejecución de 9 a 1.5 minutos. En cuanto a la categorización de los comentarios, el 80% de los mismos fue clasificado en alguna categoría. En conclusión, este proyecto permitió mejorar el control y análisis de los datos, lo que a su vez facilitará la toma de decisiones en el futuro.

Palabras claves: modelo de categorización, modelo multi tópico, modelo multi categoría, procesamiento de lenguaje natural, expresiones regulares, automatización ETL, python.

Abstract

This project was conducted in the Customer Experience Intelligence Management of Bancolombia S.A company, where two main challenges were identified. Firstly, the analysis of branch data was limited by the method used, which did not allow for effective control of the number of people attending the branches, the number of respondents, and the responses obtained. Secondly, there was a need to categorize customer comments to identify opportunities for improving the user experience, which posed challenges in identifying multiple categories within a single comment. To address these challenges, this project proposed the automation of extraction, transformation, and load (ETL) processes for branch experience data and the categorization of user comments. The results were beneficial in terms of efficiency and productivity, achieving process automation and reducing execution time from 9 to 1.5 minutes. In terms of comment categorization, 80% of the comments were successfully classified into different categories. In conclusion, this project enabled improved control and analysis of data, which will facilitate informed decision-making in the future.

Keywords: categorization model, multi-topic model, multi-category model, natural language processing, regular expressions, regular expressions model, ETL automatization, ETL, python.

Introducción

Este documento se centra en la gerencia de la Inteligencia Experiencia del cliente en la empresa Bancolombia S.A. Sus objetivos principales son escuchar al cliente a través de diversas herramientas electrónicas y telefónicas, con el propósito de obtener múltiples percepciones del cliente durante las interacciones con los servicios y productos ofrecidos por la compañía. Además, a partir de estas respuestas, se identifican las distintas problemáticas que los clientes experimentan, para así, informar al responsable de la gestión de la experiencia y proporcionarle oportunidades de mejora pertinentes.

En los procesos de extracción, transformación y carga de datos, es común hacer uso del software SAS GUIDE, que opera a través de flujos y nodos, con el fin de realizar diversas transformaciones a los datos. Este software ofrece la ventaja de brindar una visualización detallada del entorno de ejecución, aunque presenta algunas desventajas como la necesidad de adquirir su licencia y su ejecución lenta. Por otra parte, en el área se presentan dos necesidades específicas: En primer lugar, se requiere la automatización del ETL para la experiencia de sucursales para mejorar la eficiencia del proceso y obtener mayor cantidad de información con mayor facilidad de entendimiento. Esto se debe a que los datos se utilizan para llevar un control de la cantidad de encuestas enviadas, la cantidad de clientes que acudieron a las sucursales y la cantidad de respuestas obtenidas, habiéndose presentado problemas recurrentes en el análisis de los datos y la disminución de la cantidad de respuestas. En segundo lugar, se requiere el análisis de los comentarios de los clientes, con el fin de detectar los aspectos negativos o positivos que perciben en los productos o servicios ofrecidos. Para ello, se plantea la definición de categorías específicas y la observación de la polaridad de los comentarios.

Considerando lo anterior expuesto, el propósito fundamental del presente proyecto consiste en la automatización de los procedimientos correspondientes al ETL y categorización de los comentarios vertidos por los clientes. Para lograr tal objetivo, se deberán abordar de manera específica los siguientes objetivos: 1) diagnosticar los procesos relacionados al ETL de los datos,

así como a la categorización de los mismos, 2) proponer y desarrollar un proceso de ETL automatizado que integre las experiencias de las sucursales, 3) elaborar un modelo de categorización acorde a las necesidades del proyecto, 4) implementar el proceso automatizado de extracción, transformación y carga de los datos, 5) poner en práctica el modelo de categorización diseñado, y 6) llevar a cabo una validación exhaustiva tanto de la automatización como del modelo de categorización implementado.

En consecuencia, se llevará a cabo una metodología que consistirá en realizar un diagnóstico minucioso del método actualmente utilizado para la automatización del proceso ETL de la experiencia de sucursales. Una vez establecido el nuevo método, se procederá a su automatización mediante el uso del lenguaje de programación Python y se verificará su validez. Por otro lado, se documentará la metodología actual utilizada en el modelo de categorización, proponiendo una mejora que permita cumplir con las necesidades del área correspondiente. Para lograr esto, se utilizará una metodología de procesamiento de lenguaje natural y también se empleará expresiones regulares para encontrar patrones en los comentarios. A continuación, se implementará el modelo mejorado en Python y se validará su desempeño.

Por consiguiente, se expondrán los resultados logrados durante la ejecución del proyecto. En primer lugar, los resultados de la automatización del proceso ETL en la experiencia de sucursales fueron beneficiosos, ya que se logró reducir significativamente el tiempo de ejecución: de aproximadamente 9 minutos a solo 1.5 minutos. Además, se obtuvo una mayor cantidad de información en comparación con la que se brindaba previamente. Respecto al modelo de categorización, se logró clasificar aproximadamente el 80% de los comentarios en alguna categoría, mientras que el resto quedó sin clasificar. Es posible que algunos de los comentarios no clasificados puedan ser asignados a su correspondiente categoría.

El proyecto se estructurará siguiendo una secuencia metodológica que contempla el establecimiento de objetivos claramente definidos, la revisión sistemática del marco teórico para sustentar adecuadamente todas las etapas del trabajo, la aplicación de una metodología rigurosa y

apropiada para alcanzar los objetivos propuestos, la obtención de resultados precisos, la formulación de conclusiones pertinentes y la elaboración de recomendaciones útiles para futuros trabajos.

1 Objetivos

1.1 Objetivo general

Automatizar los procesos de extracción, transformación, carga de sucursales y categorización de los comentarios de los clientes del área de Inteligencia Experiencia del Cliente de Bancolombia S. A

1.2 Objetivos específicos

1. Diagnosticar los procesos de extracción, transformación y carga para las sucursales, que se realizan actualmente en el área de Inteligencia Experiencia del cliente.

2. Diagnosticar el proceso de categorización de comentarios de los clientes del área de Inteligencia Experiencia del Cliente.

3. Proponer un proceso de automatización para los procesos de extracción, transformación y carga para las sucursales que se realiza en el área de Inteligencia Experiencia del cliente.

4. Proponer un modelo de categorización de los comentarios de los clientes del área de Inteligencia Experiencia del Cliente.

5. Implementar el proceso automatizado en el área de Inteligencia Experiencia del cliente.

6. Implementar el modelo de categorización de los comentarios de los clientes del área de Inteligencia Experiencia del Cliente.

7. Validar el proceso automatizado y el modelo de categorización propuesto en el área de Inteligencia Experiencia del cliente.

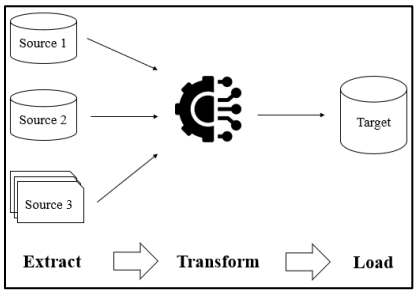
2 Marco teórico

El marco teórico juega un papel fundamental en la elaboración de un proyecto de investigación, ya que proporciona una base sólida para el desarrollo de las ideas y la identificación de las relaciones existentes entre los diferentes elementos que conforman el estudio. En este proyecto, abordaremos el marco teórico desde dos perspectivas diferentes: en primer lugar, desde la automatización de ETL, enfocada en señalar las fases y herramientas para la extracción, transformación y carga de datos. En segundo lugar, se considerará la categorización de comentarios, implementando técnicas y herramientas de minería de opiniones para el análisis profundo de las reacciones y opiniones de los usuarios. De esta manera, se podrán identificar las principales tendencias y patrones de comportamiento que serán analizados en la revisión de literatura.

2.1 ETL

El proceso de construcción de un Data Warehouse (DW) se lleva a cabo a través de la herramienta ETL, la cual consta de tres tareas fundamentales. Según El-Sappagh, Hendawi y El Bastawissy (2011), estas tareas son: (1) extracción de datos de diferentes fuentes, (2) transformación y limpieza de los datos en el área de ensayo, y (3) carga de los datos en el almacén de datos. Para comprender de manera más clara las etapas del proceso ETL desde un punto de vista visual, se presenta la Figura 1, que ilustra el flujo de cada fase del ETL.

Figura 1:Fases proceso ETL



Nota. Adaptado de Pastor, G. (2022).

El-Sappagh et al. (2011) afirman que la tarea de extracción de datos es responsable de la recolección de datos de los sistemas de origen, los cuales tienen características distintas que deben ser administradas de manera efectiva para el proceso ETL. El paso de transformación hace limpieza y ajuste a los datos entrantes para obtener datos precisos y sin ambigüedades. Este proceso incluye la limpieza, transformación e integración de datos. Finalmente, el paso de carga de datos en la estructura multidimensional de destino es el paso final de ETL. En este paso, los datos extraídos y transformados se escriben en las estructuras dimensionales a las que realmente acceden los usuarios finales y los sistemas de aplicación (El-Sappagh et al., 2011).

Martínez et al., (2013) nos presenta de manera visual el proceso ETL, con las entradas, operaciones y resultados de cada tarea, tal como se puede apreciar en la Tabla 1.

Tabla 1: Entrada, operaciones y salidas de cada una de las fases del ETL

Componente	**Elementos Objetivos (entrada)**	**Operaciones realizadas (proceso)**	**Resultado de la tarea (salida)**
Extracción	Fuentes de datos, sistemas transaccionales, hojas de cálculo, archivos de texto.	Selección	Datos crudos (cargados en memoria)
Transformación	Datos crudos (cargados en memoria)	Limpieza, transformación, personalización, realización de cálculos y aplicación de funciones de agregación	Datos formateados, estructurados y resumidos de acuerdo con las necesidades (aún en memoria)
Carga	Datos formateados, estructurados y resumidos de acuerdo con las necesidades (aún en memoria)	Inserción	Datos formateados, estructurados y resumidos con persistencia en el DW

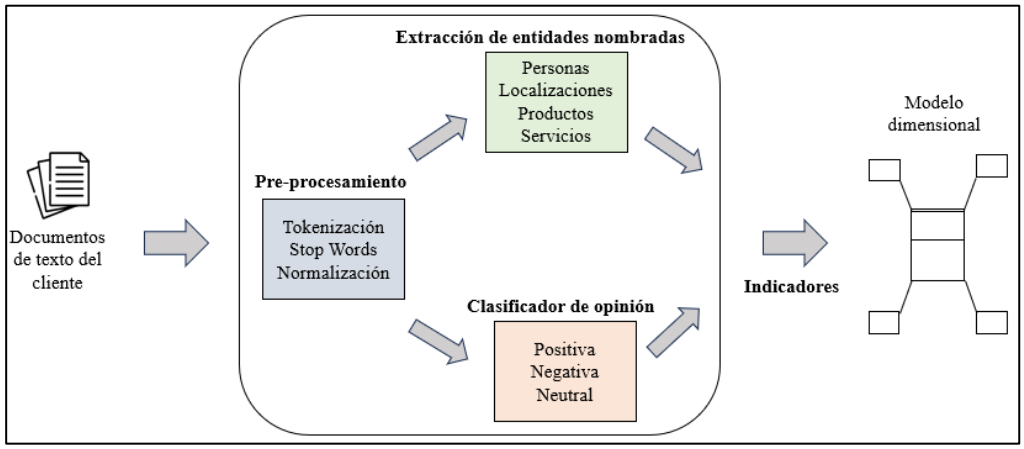
Nota. Adaptado de (Martínez et al., 2013).

2.2 Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (PLN) es una disciplina de la inteligencia artificial que tiene como objetivo dotar de la capacidad de comprensión de textos y palabras a los ordenadores, en la misma forma en la que lo hacemos los seres humanos (Barbosa, 2021). Según Damarta et al. (2021) para lograr esto, es necesario realizar diversas fases en el preprocesamiento de los textos. En primer lugar, se deben normalizar los textos, convirtiéndolos a minúsculas y eliminando caracteres repetidos, marcas y puntuaciones. Luego, se procede a la tokenización, que consiste en dividir el texto en elementos individuales o "tokens". Posteriormente, se eliminan los tokens que no agregan valor, como las llamadas "palabras vacías" o stopwords, y se realiza la lematización para convertir las palabras a su forma base. Todo esto con el objetivo de disminuir el volumen de datos y mejorar el análisis posterior.

Pabón et al. (2020) indica que el proceso de clasificación de texto se puede dividir en un conjunto de fases. Primero, se procesa el texto para extraer información relevante. Luego, se realiza una clasificación de sentimiento y se identifica a quién se refiere el texto. Finalmente, se valida el resultado a través de indicadores y se realiza el análisis más detallado. La Figura 2 presenta una representación de las fases del análisis de texto.

Figura 2: Fases del análisis de texto



Nota. Adaptado de (Pabón et al., 2020).

Un componente importante que se debe considerar en el PLN son las expresiones regulares. Según Plou (2017), las expresiones regulares son un lenguaje que sirve para especificar búsquedas de cadenas de texto. Una cadena de texto es "una secuencia de caracteres alfanuméricos (letras, números, espacios, signos de puntuación)" y según Sguerra (2006) el objetivo de éstas es buscar de manera inteligente información dentro de datos anárquicos para extraer de ellos lo que siempre se busca con la informática: información bajo contexto.

Las expresiones regulares permiten especificar búsquedas de cadenas de texto y detección de patrones en el texto para su posterior análisis y tratamiento. A través de ellas, se pueden codificar estructuras que marcan los patrones detectados, lo que permite realizar tareas como el etiquetado y la detección y clasificación de entidades (Moneda, 2018).

La librería RE es ampliamente utilizada en Python para la manipulación de expresiones regulares. Según la documentación proporcionada por Python Software Foundation (2021), una expresión regular, también conocida como RE, se define como un conjunto de cadenas que coincide con ella. El módulo de esta librería brinda funciones que permiten la comprobación de si una cadena determinada concuerda con una RE específica, así como también si una expresión regular determinada concuerda con una cuerda particular que se traduce a lo mismo.

Por otro lado, una librería muy usada para el PLN es la NLTK y es usada para el procesamiento de texto. Proporciona diferentes tipos de datos, como tokens, etiquetas, bloques, árboles y estructuras de características, y algoritmos animados, tutoriales y conjuntos de problemas. Además, incluye interfaces e implementaciones de referencia para herramientas como tokenizers, stemmers y taggers (regexp, ngram y brill). En resumen, NLTK es una herramienta invaluable para el análisis y procesamiento de lenguaje natural (Bird, 2006).

La consideración del sentimiento es un aspecto crucial en la tarea de etiquetado de texto. Con el fin de proporcionar una identificación precisa del sentimiento en un texto, es posible

emplear la librería Pysentimiento para Python. Dicha paquetería, según los hallazgos expuestos por Barbosa (2021), representa una herramienta especializada en el Análisis de Sentimientos y otras aplicaciones sociales relacionadas con el Procesamiento de Lenguaje Natural. Es importante destacar que Pysentimiento se sustenta en los modelos más avanzados disponibles tanto en inglés como en español.

3 Metodología

La metodología empleada para alcanzar los objetivos se dividió en dos partes. En primer lugar, se implementó una metodología de automatización del proceso ETL destinada a mejorar el análisis de los datos para experiencia de sucursales. En segundo lugar, se aplicó una metodología específica para el desarrollo del modelo de categorización de comentarios basada en expresiones regulares y en procesamiento de lenguaje natural.

3.1 Automatización del proceso ETL de la experiencia de sucursales

La metodología comenzara con la exposición detallada de la metodología aplicada para la automatización del proceso de ETL, la cual se compone de cuatro etapas que se describirán de manera precisa.

3.1.1 Diagnóstico

Durante la etapa inicial del proyecto, se llevó a cabo un análisis exhaustivo del proceso ETL utilizado actualmente. Este diagnóstico se realizó mediante la diagramación de las distintas etapas del proceso ETL, donde se identificaron las fuentes de datos utilizadas y su ubicación, las transformaciones de los datos a llevar a cabo, el resultado final esperado y un flujograma usando la herramienta de Bizagi que refleja paso a paso la metodología actualmente utilizada. De esta manera, se logró un mayor entendimiento del proceso ETL en curso.

3.1.2 Identificación de mejora

Esta etapa implicó la identificación de mejoras a partir de un diagnóstico previo, cuyo objetivo consistió en reducir la cantidad de pasos necesarios para lograr una automatización eficiente. Adicionalmente, se definieron las nuevas etapas del proceso de ETL, a fin de incorporar múltiples fuentes de información y aumentar el número de tablas resultantes que proporcionaran la información requerida.

3.1.3 Automatización e implementación

Durante esta etapa del proyecto, se llevó a cabo la automatización del ETL mediante el uso del lenguaje de programación Python. El objetivo principal de esta fase fue reducir los pasos encontrados en la etapa anterior que no generan valor y el número de conexiones con tablas que se requerían de la zona de aterrizaje (LZ), con el fin de alcanzar una codificación más eficiente. Cabe destacar que dichas conexiones incrementaban el tiempo de ejecución del código. Con el fin de asegurar una correcta comprensión del funcionamiento del nuevo proceso, se diseñó un flujograma detallado que documenta los pasos más relevantes del nuevo proceso ETL. De manera complementaria, se elaboró un archivo .bat que permite procesar los códigos del ETL con mayor velocidad.

3.1.4 Validación

Durante esta etapa se llevó a cabo una evaluación de los resultados obtenidos del proceso ETL, en aras de determinar su adecuación y veracidad con respecto a los resultados previos. Para tal fin, se procedió a comparar las tablas resultantes de ambos métodos, lo cual implicó un análisis exhaustivo de aquellas tablas no contenidas en los resultados previos. Con el fin de facilitar este proceso, se utilizó la herramienta HUE, la cual posibilita la utilización del motor de consulta Impala. Esto permitió establecer una conexión con las bases de datos alojadas en la LZ y llevar a cabo una revisión meticulosa de los resultados mediante técnicas manuales.

3.2 Modelo de categorización de los comentarios de los clientes

La metodología comenzara con la exposición detallada de la metodología aplicada para la categorización de los comentarios de los clientes, la cual se compone de cuatro etapas que se describirán de manera precisa.

3.2.1 Diagnóstico

En esta etapa, se llevó a cabo una evaluación minuciosa de la metodología utilizada para llevar a cabo la categorización de comentarios. En primer lugar, se describió el proceso actual que se sigue para obtener el resultado de la categorización y la identificación del sentimiento de cada

comentario. Se elaboró un flujograma que describa el funcionamiento del código de categorización, así como de la paquetería empleada para predecir el sentimiento del comentario en general. Es importante mencionar que tanto la categorización como la identificación del sentimiento se llevaban a cabo de manera separada. Además, se dispone de una estructura detallada de la salida que genera el método utilizado anteriormente.

3.2.2 Identificación de mejora

En esta etapa se llevó a cabo la identificación de mejoras, utilizando como punto de partida un diagnóstico previo. Para ello se contó con la colaboración de los expertos del negocio, quienes se encargaron de aportar su conocimiento acerca de aquellos aspectos que pudieran agregar valor al modelo para que fácilmente pudieran detectar las brechas de satisfacción, y que debían ser potencializados en cada experiencia brindada al usuario. Con el objetivo de contar con una fase clara y definida, se consideró vital tener en cuenta este análisis previo para definir el modelo y los elementos que debían ser incluidos en el mismo.

3.2.3 Creación del modelo de categorización e implementación

En esta etapa se procedió a la construcción del modelo mediante el uso del lenguaje de programación Python. En primer lugar, se trabajó conjuntamente con el equipo de negocio para definir las categorías y sus respectivas definiciones, con el fin de asegurar una comprensión clara por parte de todo el equipo. Posteriormente, se llevó a cabo un proceso de clustering basado en características similares de la experiencia, también con la colaboración del equipo de negocio.

Una vez creadas las categorías y agrupadas las experiencias, se estableció una estructura parametrizada en Excel, la cual contenía las expresiones regulares que fueron creadas con insumos manuales y también con unas funciones que brinda la paquetería NLTK. Este aspecto era fundamental para lograr la clasificación eficiente de los comentarios y la detección de patrones asociados a las distintas categorías.

Posteriormente, en otra hoja de Excel se incluyeron las experiencias previamente agrupadas y definidas. Finalmente, se construyó el modelo en Python, el cual tuvo la capacidad de tomar las expresiones regulares de la categoría y el sentimiento correspondiente para clasificar el comentario.

3.2.4 Validación

Para la validación del modelo, se seleccionó una muestra de **5000** comentarios, la cual fue sometida a revisión manual por parte del equipo de trabajo. Durante este proceso, se evaluaron tanto las categorías como los sentimientos asociados a estas categorías, los cuales fueron previamente identificados por el modelo. Para verificar la correcta categorización, se utilizó la herramienta Excel.

4 Resultados

A continuación, se presentan algunos resultados que se aproximan a lo obtenido en la ejecución del proyecto. Es importante destacar que los hallazgos reales son de naturaleza confidencial. Las herramientas utilizadas para dicho proceso fueron el lenguaje de programación Python, Excel y SQL. Además, se emplearon librerías propias de la compañía para establecer la conexión entre Hadoop y Python, así como otras librerías relevantes tales como re, numpy, pandas y NLTK.

4.1 Automatización del proceso ETL de la experiencia de sucursales

En esta sección se abordarán de forma detallada los pasos necesarios para lograr la automatización de ETL en la experiencia de sucursales. El primer paso consiste en llevar a cabo un diagnóstico, mediante el cual se evaluarán los procesos actuales y se identificarán posibles áreas de mejora. Una vez identificada la mejora potencial, se procederá a la automatización de dichos procesos, utilizando la herramienta de Python. Por último, se realizará una validación exhaustiva de los resultados obtenidos, con el fin de garantizar que los resultados sean los esperados y cumplan con los requisitos establecidos.

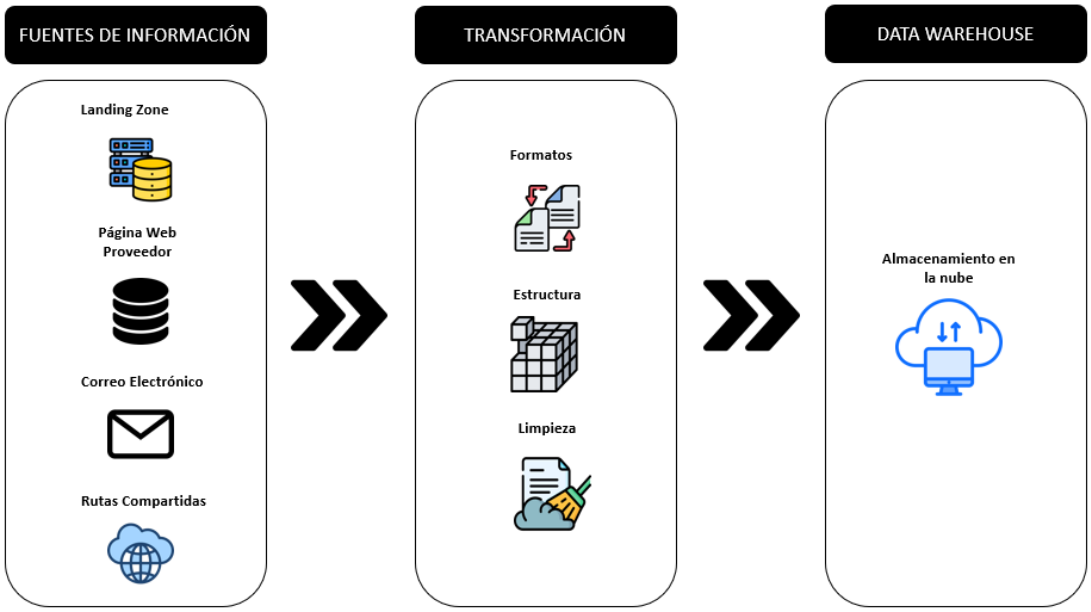
4.1.1 Diagnóstico

En el contexto de Bancolombia, más precisamente en la Gerencia de Experiencia Inteligente del Cliente, existen diversas fuentes de datos que se pueden obtener para fines de análisis. Entre estas fuentes se encuentra la llamada zona de aterrizaje, que constituye el servidor principal de almacenamiento de todas las bases de datos de la compañía. Es relevante destacar que no todos los colaboradores tienen acceso a todas las tablas, siendo necesario solicitar dicho acceso. También se puede obtener información relevante a través de la página web del proveedor, que funciona como fuente de datos para conocer, por ejemplo, la cantidad de respuestas de una sucursal. Asimismo, algunos de los proveedores envían datos por correo electrónico, los cuales deben ser tratados de forma manual. Finalmente, se pueden obtener datos específicos y de interés común de

todos los colaboradores que pertenecen al área a través de rutas compartidas de la nube, lo cual representa una fuente de información muy específica para esta área.

En la Figura 3, se realiza una representación visual de las fuentes de información utilizadas en el área de análisis, así como también se ilustran las transformaciones más habituales que se le aplican a los datos, como lo son la modificación de los formatos de las columnas, la reorganización de la estructura y la realización de tareas de limpieza de los datos. Además, se destaca que los resultados obtenidos frecuentemente son almacenados en la nube para su posterior uso.

Figura 3: Fuentes de datos, transformaciones y bodega de datos que son más utilizados en el área



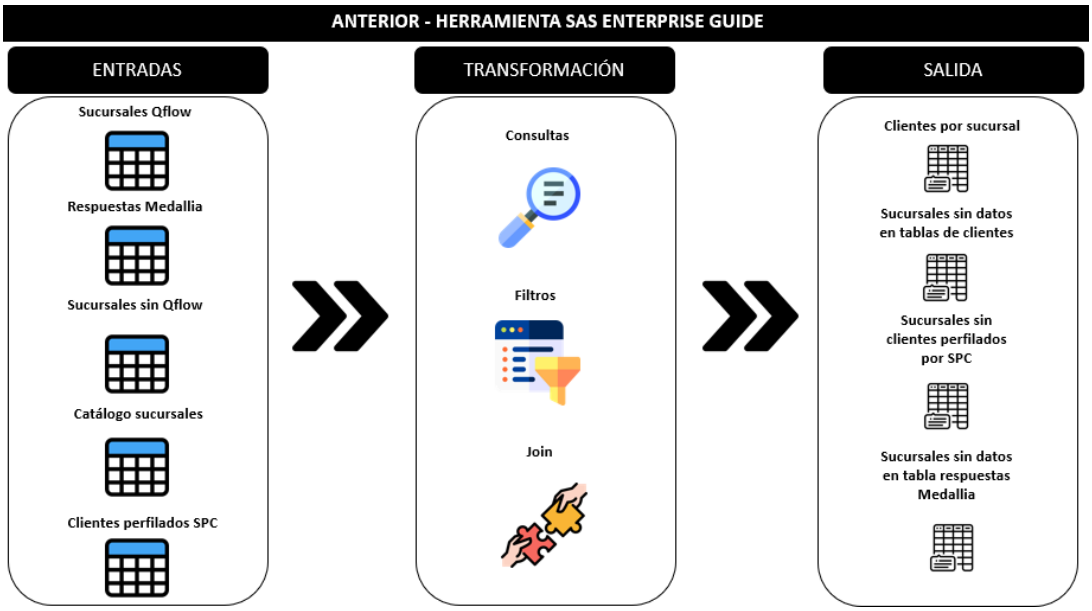
Nota. Elaboración propia

El proceso ETL actual ha sido implementado utilizando la herramienta SAS Enterprise Guide, la cual es de pago para la empresa. El resultado de este proceso consiste en la obtención del número de clientes que acudieron a cada sucursal, la identificación de sucursales que carecen de

datos de clientes en la tabla, la recopilación de las respuestas de las encuestas efectuadas en cada sucursal, la identificación de las sucursales que no llevaron a cabo el perfilamiento de clientes mediante el Área de Servicio para Clientes (SPC) para su posterior encuestado y la identificación de sucursales que no han recolectado respuestas de encuestas.

Para observar las entradas y las transformaciones empleadas para obtener las salidas previamente mencionadas, esto se puede evidenciar en la figura 4.

Figura 4: Las fases del proceso ETL del método actual



Nota. Elaboración propia

A fin de comprender de manera exhaustiva todas las entradas y transformaciones contenidas en la Figura 4, a continuación, se explica minuciosamente el significado de cada una de ellas en Tabla 2.

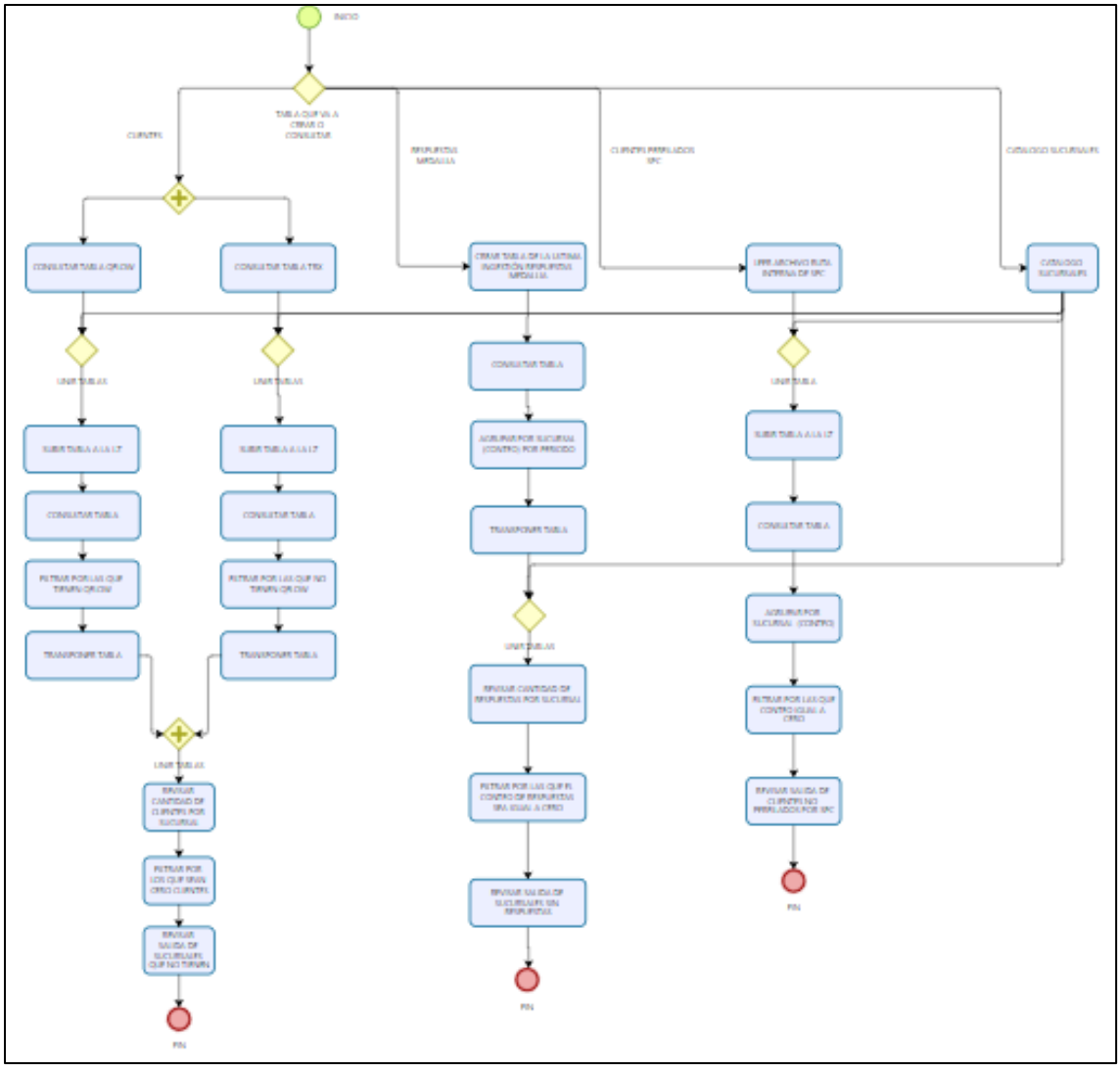
Tabla 2: Entradas y transformaciones del ETL actual

Fase	Insumo y cambios	Contenido
Entradas	Sucursales Qflow	Son las sucursales que cuentan con un sistema automatizado para registrar las transacciones de los clientes.
	Respuestas Medallia	Medallia es un proveedor que mide las experiencias de los clientes y aloja una tabla con los resultados en la LZ.
	Sucursales sin Qflow	Son las sucursales que requieren una transcripción manual de los datos del cliente y su transacción.
	Catálogo de sucursales	Es la lista de todas las sucursales activas en la compañía.
	Clientes perfilados por SPC	Son los clientes que han pasado por una serie de filtros establecidos por el área de servicio al cliente.
Transformaciones	Consultas	Son consultas que se le realizan a una tabla en específico, ya que no se necesita la tabla completa, cambia la estructura de la tabla
	Filtros	Son segmentaciones que se realizan en los datos para sacar información de interés
	Join	Son uniones que se hacen con otras tablas.

Nota. Elaboración propia

Posteriormente, en la Figura *5* se presenta un flujograma que detalla cada paso del método actual, permitiendo identificar posibles mejoras para incrementar la eficiencia del tiempo de ejecución. En este caso, se llevan a cabo alrededor de 27 actividades para la consecución de los resultados.

Figura 5: Flujograma del proceso de ETL actual



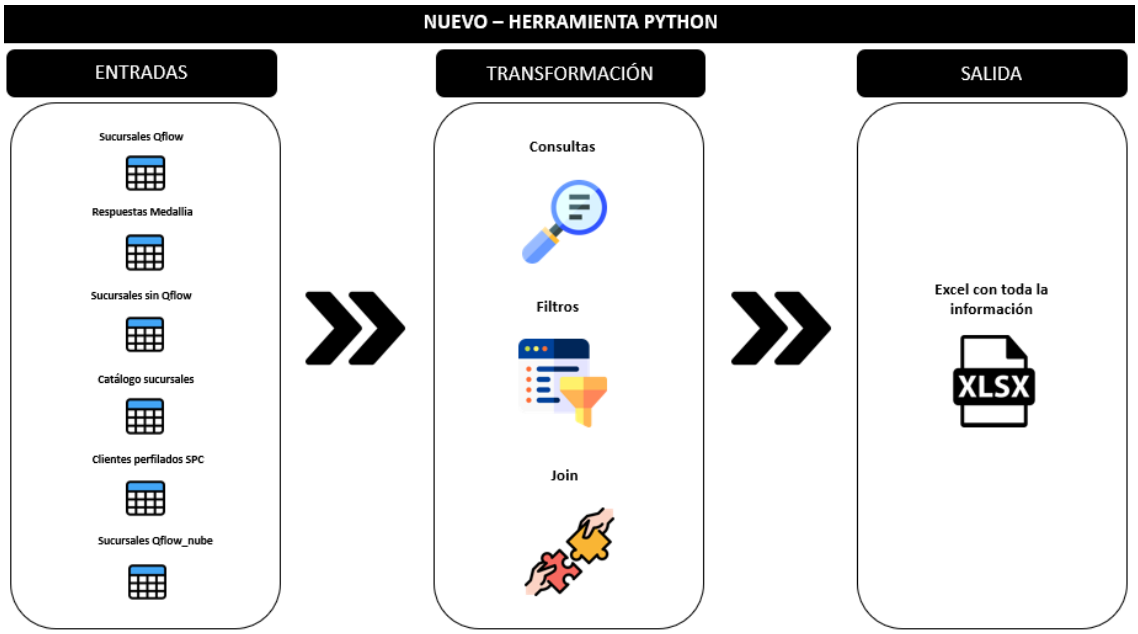
Nota. Elaboración propia

4.1.2 Identificación de mejora

Se debe tener en cuenta la Figura 5 como un insumo para detectar las tareas que no agregan valor. De acuerdo con el método actual, se realizan conexiones a la LZ que podrían ser realizadas sólo una vez mediante el lenguaje Python. Esto resultaría en una reducción del tiempo ya que la conexión es un proceso dispendioso. Asimismo, algunos cruces de tablas innecesarios pueden ser eliminados.

Se presentan las nuevas entradas, transformaciones y salidas para el método mejorado, esto lo podemos verificar en la Figura 6. Cabe mencionar que en algunas de estas fases se han incluido algunas tablas adicionales, las cuales deben ser consideradas. También se han presentado algunas transformaciones diferentes. Como resultado, sólo se obtendrá una salida en Excel, pero ésta contendrá varias hojas.

Figura 6: Fases del proceso ETL del método mejorado



Nota. Elaboración propia

En las fases más recientes del proceso de ETL se ha identificado la inclusión de una tabla adicional denominada "Qflow nube". Esta tabla es destinada para la migración de datos provenientes de aquellas sucursales que utilizan el sistema Qflow. Es importante destacar que su incorporación implica una variación con respecto al esquema de procesamiento actualmente implementado.

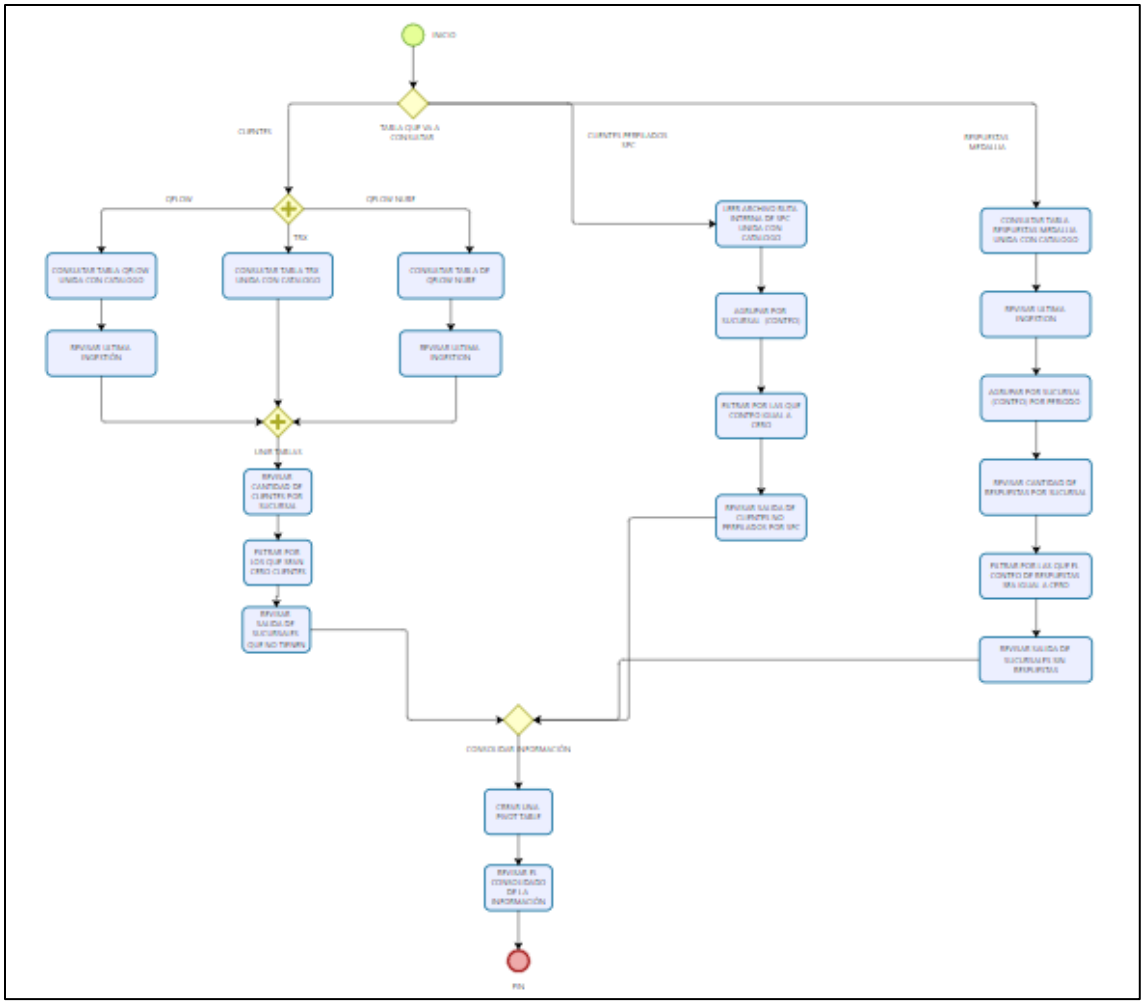
Considerando todas las etapas del proceso ETL, es viable proceder con la automatización de dichas fases.

4.1.3 Automatización e implementación

Para la automatización del proceso, se eliminaron las tareas que no generaban valor, como, por ejemplo, las conexiones a la LZ. Además, se agruparon tareas que podían ser realizadas conjuntamente y se redujo la cantidad de tareas, lo cual permitió obtener un código optimizado. Para llevar a cabo la automatización, se hizo uso de la paquetería de pandas, numpy, sparky_bc y helper. Los dos últimos paquetes se utilizaron para lograr la conexión entre Python y Hadoop, plataforma en la que se almacenan las bases de datos que se requerían. A través de código SQL, se pudo traer las tablas deseadas a Python.

En la Figura 7, se puede observar el flujograma que muestra de manera detallada cada una de las tareas realizadas y los pasos que se siguieron para obtener la salida esperada. En resumen, este es el funcionamiento del código.

Figura 7: Funcionamiento del código en Python



Nota. Elaboración propia

Con el fin de obtener un mayor nivel de comprensión acerca de la generación de la salida, se presenta la Tabla 3 que muestra las diversas hojas resultantes y la información observada en cada una de ellas. Es importante destacar que la hoja de consolidado representa la más relevante, ya que sintetiza toda la información necesaria y, además, permite obtener un mayor control sobre los datos.

Tabla 3: Información observada en cada una de las hojas del archivo de salida

Salidas	Contenido
Ingestión de datos tabla qflow	Son las ingestiones de datos en la tabla qflow
Ingestión de datos tabla qflow_nube	Son las ingestiones de datos en la tabla qflow_nube
Ingestión de datos tabla Medallia	Son las ingestiones de datos en la tabla de Medallia
Clientes diarios por sucursal	Cantidad de clientes que fueron a cada sucursal diariamente en el mes actual
Respuestas diarias por sucursal	Cantidad de clientes que respondieron de cada sucursal diariamente en el mes actual
Respuestas mes actual por sucursal	Cantidad de clientes que respondieron de cada sucursal en el mes actual
Sucursales sin datos en tablas del mes actual	Sucursales que no se están subiendo datos a la tabla de la LZ
Sucursales sin clientes perfilados SPC	Sucursales que SPC no perfilo clientes para ser encuestados
Consolidado	La cantidad de clientes diarios que fueron, la cantidad de respuestas diarias y los clientes perfilados por SPC, esto se muestra diario para el mes actual

Nota. Elaboración propia

Durante el estudio del ETL, se ha podido constatar que hubo una disminución significativa en el número de actividades del proceso, pasando de **27** a **20** actividades. Este hecho resulta aún más destacable si se tiene en cuenta que se trabajó con un mayor número de tablas de datos y se generó una mayor cantidad de información en la salida. Asimismo, es importante destacar que se ha logrado una notable optimización en el tiempo de ejecución. En concreto, se ha constatado que el modelo mejorado desarrollado en Python ha reducido significativamente el tiempo de ejecución del software utilizado inicialmente, disminuyendo el periodo de **9** minutos a tan solo **1.5** minutos.

4.1.4 Validación

Con el propósito de validar el ETL, a fin de determinar si los resultados obtenidos eran apropiados, se realizó una inspección en colaboración con el responsable de la experiencia en sucursales. Esta inspección se efectuó con la finalidad de verificar si se podían observar resultados acordes a lo esperado. Adicionalmente, se llevó a cabo una comparación entre los resultados obtenidos con el ETL anterior y los actuales, a través de una tabla donde se registraron cada uno de los resultados obtenidos de cada sucursal con ambos ETL. Para determinar si los resultados eran similares o diferentes, se utilizó una fórmula y se registró en una columna denominada Validador. Cuando la columna Validador era igual a "Verdadero", se consideraba que los resultados eran iguales. La Tabla 4 muestra este proceso de validación.

Tabla 4: Estructura de validación del ETL mejorado

Salidas	Resultado 1 SAS	Resultado 1 Python	….	Resultado n	Validador
Sucursal 1	100	100	….	13	Verdadero
Sucursal 2	80	80	….	17	Verdadero
….	….	….	….	5	Verdadero
Sucursal n	156	156	….	12	Verdadero

Nota. Elaboración propia

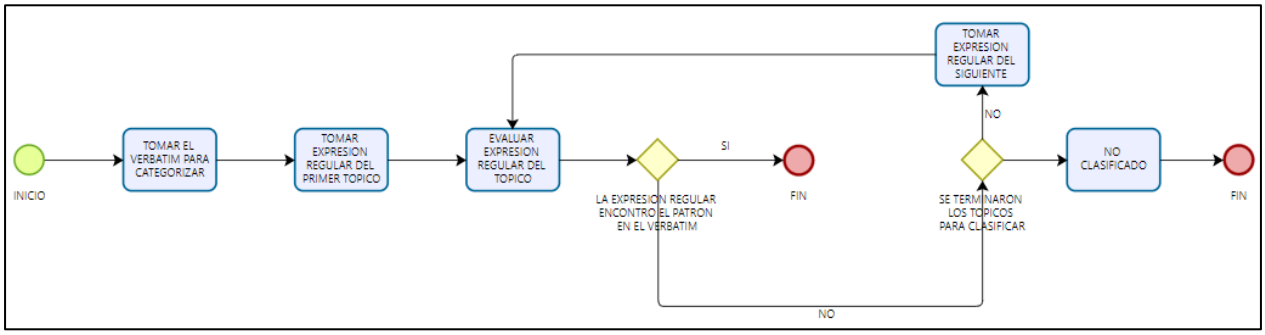
Tras llevar a cabo el proceso de validación pertinente, se detectaron algunas diferencias entre los datos obtenidos mediante el ETL utilizado previamente y los correspondientes a la fuente original. Estas divergencias fueron debidamente analizadas y se determinó que se debían a la existencia de filtros que no estaban debidamente contemplados en las tablas. Tras proceder a la organización de dichos errores, se llevó a cabo una nueva validación, la cual arrojó como resultado que todos eran idénticos entre sí. En virtud de lo anterior, se aceptó el nuevo ETL como la herramienta más adecuada para llevar a cabo la obtención de los datos relativos a las sucursales.

4.2 Modelo de categorización de los comentarios de los clientes

4.2.1 Diagnóstico

En el ámbito de la inteligencia de experiencia del cliente, se ha implementado un modelo de categorización basado en una lista de categorías establecida por un analista. El proceso consiste en identificar una categoría general basándose en el comentario y utilizar una herramienta de análisis de sentimientos llamada Pysentimiento para determinar el sentimiento general del comentario. Sin embargo, se ha observado que este modelo de categorización presenta una eficacia limitada debido a que utiliza expresiones regulares poco rigurosas que contienen la frase completa, lo que hace difícil encontrar la misma frase en otro comentario. Además, algunas categorías reciben un peso mayor, lo que infla la calificación de estas categorías y lleva a una posible categorización en una sola categoría, aunque el comentario podría haber sido categorizado en otras más. Por otro lado, la herramienta de análisis de sentimientos proporciona información poco útil ya que clasifica muchos comentarios como neutrales, lo que no aporta valor al objetivo final del proceso. Para comprender el proceso actual de categorización, se presenta la Figura 8 que ilustra de manera detallada el paso a paso de la función actual de categorización.

Figura 8: Flujograma de la función de categorizar actual



Nota. Elaboración propia

Por otra parte, se muestra en la Tabla 5 la estructura de salida que se obtenía a través de la metodología actual utilizada para la categorización y determinación del sentimiento. De dicha tabla se puede inferir la escasa eficiencia del modelo para identificar la categoría correspondiente.

Tabla 5: Estructura de la salida actual del modelo de categorización

Comentario	Categoría	Sentimiento
El precio de ustedes es muy alto, su servicio es malo, lo bueno es que están en todos lados	Servicio	Neutro
….	….	….
Comentario n	Categoría n	Sentimiento n

Nota. Elaboración propia

4.2.2 Identificación de mejora

En primer lugar, se llevó a cabo una reunión con el equipo de negocios para comprender las necesidades del modelo de clasificación, a fin de que pudiera brindar un valor significativo en la mejora de la experiencia del usuario. Por tanto, se requería un modelo capaz de identificar varias categorías y su sentimiento en un comentario. Es evidente, según la Tabla 5, que en los comentarios se abordan diferentes temas, algunos de naturaleza positiva y otros de naturaleza negativa, lo que demuestra la imposibilidad de categorización en una única categoría. Por lo tanto, se necesitaba una solución que abarcara todos los temas abordados en el comentario. Asimismo, se descartó la posibilidad de obtener un sentimiento general del comentario, dado que este podría ser neutral debido a la presencia de aspectos tanto positivos como negativos. Se decidió entonces que el sentimiento tendría que estar asociado a la categoría correspondiente. Por ejemplo, si se trataba de la categoría "precio", se debería especificar el sentimiento asociado a esa categoría, por ejemplo, "precio-negativo".

4.2.3 Creación del modelo de categorización e implementación

Para esta fase, se ha procedido a definir una estructura jerárquica compuesta por un tópico principal y subtemas que estén directamente relacionados con el tópico principal. El resultado de la concatenación de los dos elementos se identificó como una categoría. Posteriormente, se ha

definido una tabla en la cual se ha indicado la definición correspondiente para cada una de las categorías identificadas, con el fin de homologar el lenguaje utilizado por todos los miembros del equipo y evitar confusiones en la interpretación de los temas abordados. El propósito es lograr una comprensión unívoca por parte de todos los integrantes del equipo de negocio.

Con el fin de obtener una comprensión adecuada de la estructura de la tabla de definiciones, se presenta su disposición mediante la Tabla 6, con el propósito de proporcionar una mayor claridad y comprensión al respecto.

Tabla 6: Estructura de la tabla de definiciones de las categorías

Tópico	Subtopico	Categoría	Definición
Tópico 1	Subtopico 1	Tópico 1 – Subtopico 1	Esta categoría hace referencia a…
Tópico 1	Subtopico 2	Tópico 1 – Subtopico 2	Esta categoría hace referencia a…
….	….	….	….
Tópico n	Subtopico n	Tópico n – Subtopico n	Esta categoría hace referencia a…

Nota. Elaboración propia

Posteriormente se llevará a cabo el proceso de agrupación de experiencias, el cual se justifica por la similitud existente en las características de estas y su correspondiente terminología. Resulta relevante destacar que no todas las categorías serán aplicables en todas las experiencias, ya que algunas categorías no serán pertinentes de acuerdo con la naturaleza de la experiencia en particular. Por ejemplo, si la experiencia se refiere al uso de la herramienta digital de Bancolombia, no debería haber una categoría que esté relacionada con la atención personalizada, ya que la experiencia no se llevó a cabo de manera presencial. Esta práctica contribuirá a disminuir los posibles errores que se puedan presentar en el modelo. En consecuencia, se establece que es necesario definir adicionalmente las categorías que pueden ser relevantes para el grupo identificado, en función de la naturaleza de la experiencia. La estructura de lo anteriormente expuesto puede ser visualizada en la Tabla 7.

Tabla 7: Estructura de la agrupación de las experiencias y las categorías pertinentes para el grupo

Grupo	Experiencias	Categorías
Grupo 1	1,23,57	Categoría 1, categoría 4, …
Grupo 2	3,7,9	Categoría 1, categoría 2, …
….	….	….
Grupo n	67,100,123	Categoría 3, categoría 5, …

Nota. Elaboración propia

Luego, se procedió a la elaboración de un archivo parametrizado en Microsoft Excel, este archivo tenía una estructura compuesta por una hoja que contemplaba la tabla 6 y otra hoja que mostraba una estructura presente en la Tabla 8. Esta última, exhibía la categorización tanto para sentimientos positivos como negativos con el objetivo de poder realizar expresiones regulares, dependiendo del sentimiento que se quisiera asociar con la categoría específica. Además, incluía una columna que contenía las expresiones regulares.

Tabla 8: Estructura de hoja parametrizada con las expresiones regulares para la categoría y sentimiento

Categorías	Sentimiento	Expresiones regulares
Categoría 1	Positivo	[a-b].*
Categoría 1	Negativo	[a-b].*
Categoría 2	Positivo	[a-b].*
Categoría 2	Negativo	[a-b].*
….	….	….
Categoría n	Sentimiento n	[a-b].*

Nota. Elaboración propia

Cabe resaltar que no se disponía de un registro histórico de etiquetas en los comentarios, por lo tanto, se decidió emplear un modelo con expresiones regulares para identificar patrones en las opiniones de los clientes. Para realizar las expresiones regulares se contó con información manual, en la cual el equipo de negocio asignó etiquetas a los comentarios que facilitaron el proceso. De igual forma, se dispuso de un recurso de palabras repetidas, bigramas y trigramas proporcionados por la paquetería de NLTK en Python. Este recurso se puede visualizar en la Tabla 9 y permitió empezar a desarrollar expresiones regulares con mayor impacto, dado su frecuencia en los comentarios. Es importante mencionar que este recurso se particionó para cada una de las experiencias.

Tabla 9: Estructura del insumo utilizando la paqueteria NLTK en Python

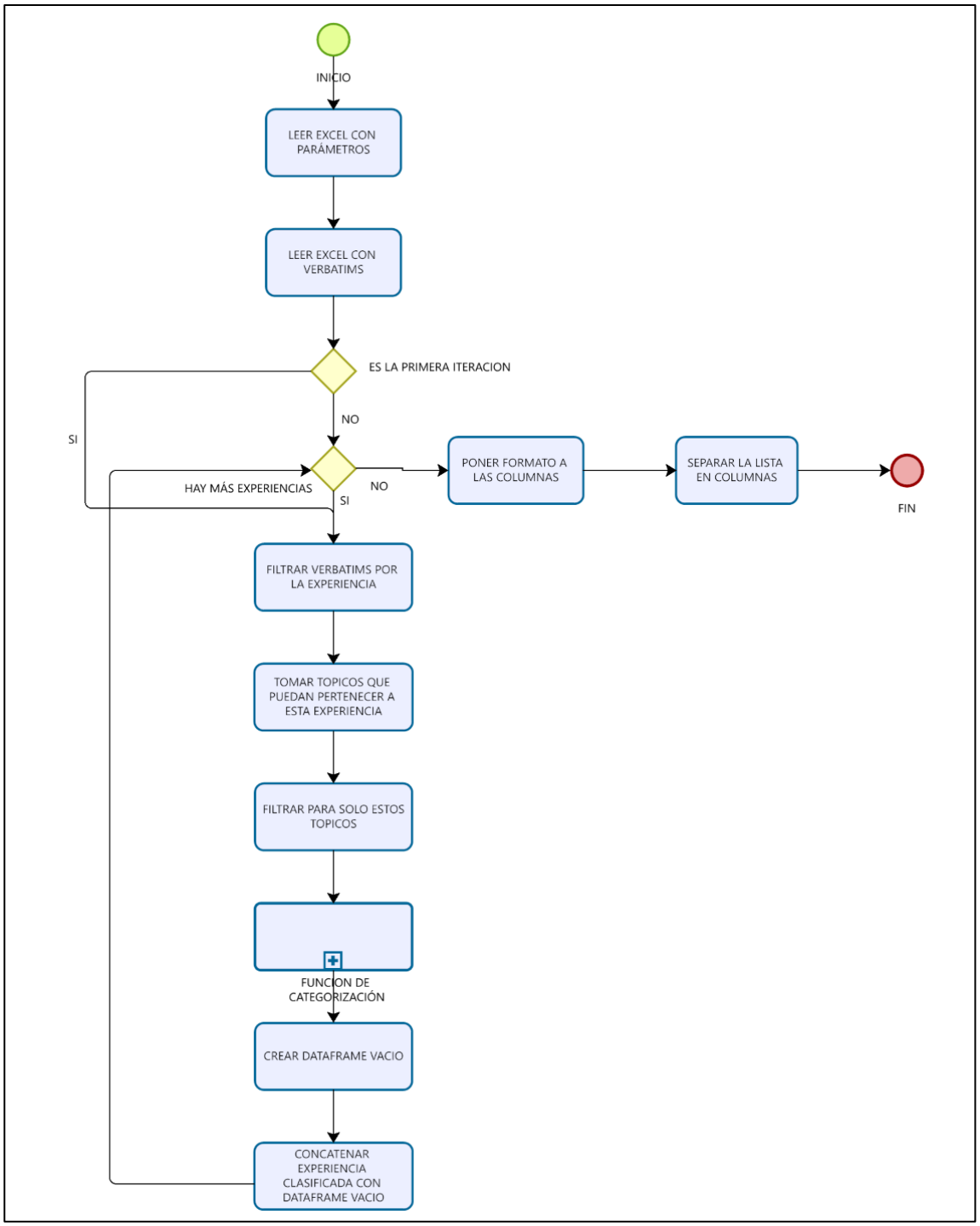
Repetidas	Bigramas	Trigramas	Experiencia
(plataforma, 30)	(plataforma mala, 18)	(la plataforma mala, 12)	100
….	….	….	100
(palabra n, frecuencia)	(palabras n, frecuencia)	(palabras n, frecuencia)	100

Nota. Elaboración propia

Una vez se tienen los parámetros en una hoja de Excel, se procede a construir el modelo en Python. Este modelo realiza diversas operaciones, comenzando por la lectura del archivo de Excel que contiene los parámetros. Posteriormente, se lee el archivo de comentarios de los clientes que se desean categorizar. Después, se lleva a cabo el proceso de limpieza en el texto, el cual consiste en remover las tildes, transformar todo a minúsculas, eliminar los espacios sobrantes y también suprimir los signos de puntuación. A continuación, se selecciona el primer grupo de experiencias y se procede a filtrar los comentarios que están relacionados con dicha experiencia específica. Acto seguido, se seleccionan las categorías correspondientes a ese grupo de experiencias y se realiza un filtrado en la hoja que contiene las expresiones regulares, únicamente para aquellas categorías seleccionadas.

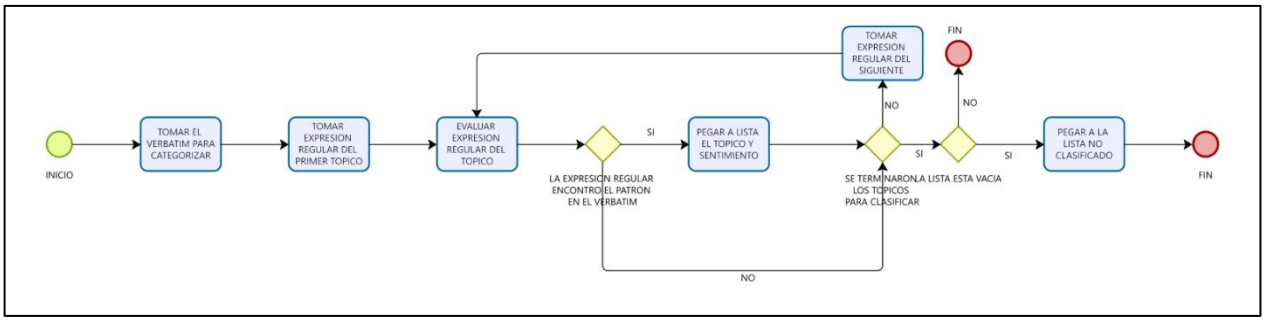
Posteriormente, se aplica una función de la paquetería RE de Python, la cual puede ser visualizada en la Figura 9. Si se encuentra el patrón de la expresión regular en el comentario, la categoría correspondiente se agrega a una lista. De lo contrario, se agrega una categoría de "No Clasificado" al final. Este proceso se repite para los demás grupos de experiencias. Se puede observar el proceso detallado del funcionamiento del código en la Figura 10, con el fin de comprender por completo cómo funciona el modelo.

Figura 9: Funcionamiento del código completo



Nota. Elaboración propia

Figura 10: Funcionamiento de la función que categoriza los comentarios



Nota. Elaboración propia

Finalmente, se procede a obtener una salida en formato de lista de listas. En cada una de estas listas internas se encuentra la categoría correspondiente junto al sentimiento asociado a dicha categoría. La estructura de esta salida se muestra en detalle en la Tabla 10.

Tabla 10: Estructura de salida del modelo de categorización

Experiencia	Comentario	Categoría
2	El precio de ustedes es muy alto, su servicio es malo, lo bueno es que están en todos lados	[[‘precio’, ‘negativo’], [‘servicio’, ‘negativo’], [‘cobertura’, ‘positivo’]]
….	….	….
Experiencia n	Comentario n	Categoría n

Nota. Elaboración propia

Tras la operación de clasificación de unos **515.000** comentarios recolectados de diversas experiencias, se constató que alrededor del **80**% de los comentarios fueron etiquetados con al menos una categoría específica, mientras que el restante fue asignado a la categoría “No Clasificado”. Algunos comentarios de este último grupo podrían haberse incluido en una categoría

determinada, mientras que otros ofrecían información insuficiente para tal fin. Este procedimiento inició la validación de la clasificación de comentarios con respecto a su adecuación a la categoría asignada por el modelo.

4.2.4 Validación

En el proceso de validación de nuestro modelo, se seleccionó una muestra de 5000 comentarios para su revisión manual. Cabe destacar que no se realizó ningún ajuste a las expresiones regulares para esta muestra de comentarios, lo que asegura la autenticidad y objetividad del proceso de validación. Durante la revisión manual, se evaluó la precisión de nuestro modelo al clasificar los comentarios y se registró la clasificación que había encontrado el modelo. Para asegurar la rigurosidad del proceso de validación, se consideró que un comentario estaba correctamente clasificado únicamente si se ajustaba perfectamente a todas las categorías y sentimientos predichos por el modelo, mientras que cualquier fallo en alguna categoría o sentimiento disminuiría la calificación de ese comentario a "malo". Se puede observar en la Tabla 11 cómo se estructuraba la revisión del modelo.

Tabla 11: Estructura de la revisión de las categorías brindadas por el modelo

Comentario	Categoría	Revisión
El precio de ustedes es muy alto, su servicio es malo, lo bueno es que están en todos lados	[['precio', 'negativo'], ['servicio', 'negativo'], ['cobertura', 'positivo']]	Bueno
….	….	….
Comentario n	Categoría n	Revisión n

Nota. Elaboración propia

Después de finalizar el proceso de revisión, se observó que aproximadamente el 75% de los comentarios estaban correctamente clasificados, mientras que los demás podían haber fallado en la categoría o sentimiento o pudieron haber tenido una categoría faltante o sobrante. Luego, se

procedió a crear un archivo .bat para su ejecución automática, asegurándose de que la salida se almacenara en una ruta compartida para su posterior análisis.

5 Análisis

5.1 Automatización del ETL de sucursales

Se ha evidenciado la importancia de la experiencia de sucursales para la Gerencia de Inteligencia del Cliente debido a la alta demanda de personas que visitan las sucursales y la necesidad de los gerentes de obtener una muestra significativa de respuestas para controlar la calidad del servicio ofrecido al usuario. Para ello, se requiere de datos precisos y actualizados que permitan generar bonificaciones a partir de las calificaciones obtenidas.

Teniendo en cuenta lo anterior, se observó la existencia de posibles mejoras en el ETL utilizado para obtener y procesar la información requerida, dado que su ejecución era demasiado lenta y no ofrecía suficientes resultados para analizar. Por tal motivo, se decidió migrar al lenguaje de programación Python, con la finalidad de acelerar el proceso y generar una mayor cantidad de datos en menor tiempo, logrando disminuir el tiempo de ejecución de 9 a 1.5 minutos. Asimismo, se logró reducir la cantidad de tareas de 27 a 20, lo que evidencia la optimización del código empleado.

No obstante, la migración al nuevo ETL presenta algunas desventajas, como la falta de un entorno visual que facilite la comprensión de las actividades realizadas por el código. Además, es importante mencionar que existen posibles riesgos en caso de intermitencias en la LZ, lo que podría dificultar la extracción de la información alojada. Cabe destacar, sin embargo, que el ETL anterior también se veía afectado por este problema.

5.2 Modelo de categorización de comentarios

El proyecto inició a partir de un modelo de comentarios previamente utilizado, que solo predecía una categoría para todo el comentario y les otorgaba un peso mayor a algunas categorías, lo que inflaba sus resultados y daba como resultado que los clientes hablaban predominantemente

de una categoría específica. Se utilizó la paquetería de pysentimiento para predecir el sentimiento del comentario, pero esto no fue de gran utilidad ya que muchos comentarios resultaron neutros. Por lo tanto, se decidió utilizar expresiones regulares, dado que se observaron patrones repetitivos en los comentarios por cada una de las experiencias, y no se consideró necesario un modelo de aprendizaje automático debido a la falta de datos categorizados. La decisión de vincular el sentimiento con la categoría resultó muy útil, ya que se puede determinar cuán negativo están hablando los clientes de una categoría específica.

Los resultados fueron la categorización del **80**% de los comentarios, mientras que el **20**% restante no se clasificó. Se realizó una validación y se encontró que el **75**% de la clasificación fue correcta. Sin embargo, una desventaja del modelo es el mantenimiento que requiere, ya que siempre se deben aplicar expresiones regulares para poder categorizar los comentarios, lo que puede resultar tedioso con el tiempo. Por lo tanto, se puede considerar utilizar modelos de aprendizaje automático o una red neuronal dado que ya se tienen datos etiquetados.

6 Conclusiones

En conclusión, se ha demostrado la importancia de la automatización del ETL de sucursales para la Gerencia de Inteligencia del Cliente, debido a la necesidad de contar con datos precisos y actualizados para controlar la calidad del servicio ofrecido al usuario. La migración al lenguaje de programación Python ha permitido acelerar el proceso y generar una mayor cantidad de datos de manera más eficiente, sin embargo, es necesario considerar posibles riesgos en caso de intermitencias en la LZ.

En relación con el modelo de categorización de comentarios, se ha observado que la vinculación del sentimiento con la categoría resultó muy útil para determinar el nivel de satisfacción del cliente en cada categoría. Sin embargo, se debe considerar que la aplicación de expresiones regulares para categorizar los comentarios puede resultar tediosa con el tiempo, por lo que podría ser beneficiosa la implementación de modelos de aprendizaje automático o una red neuronal. En general, ambos proyectos han demostrado el potencial de la tecnología en la optimización de procesos y la obtención de datos precisos y útiles para la gestión empresarial.

7 Recomendaciones

Para las recomendaciones de trabajos futuros del proyecto, se sugiere la posibilidad de abordar la data etiquetada mediante la incorporación de un modelo de machine learning multi-categórico o de redes neuronales, a finde generar un sistema de aprendizaje automático más eficiente y accesible para el área. Debido al carácter tedioso del mantenimiento de las expresiones regulares, resulta conveniente explorar esta alternativa y evaluar su conveniencia operativa.

Por otra parte, se recomienda analizar la inclusión de nuevas categorías en el modelo, a fin de abarcar temas que no hayan sido previamente considerados. En efecto, se ha observado que algunos comentarios presentan dificultades para ser asignados a las categorías definidas, lo que podría generar discrepancias en cuanto a su clasificación. Incorporar nuevas categorías que aborden estos temas no abarcados en la versión actual del modelo puede resultar en una solución ideal para fortalecer su eficacia y relevancia.

Referencias

Barbosa, G. (2022). Minería de opinión basada en aspectos: Un análisis comparativo de opiniones de estudiantes de educación superior en Colombia antes y después del COVID-19. Pontificia Universidad Javeriana, Colombia.

Bird, S. (2006, July). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 69-72).

Damarta, R., Hidayat, A., & Abdullah, A. S. (2021, January). The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality. In *Journal of Physics: Conference Series* (Vol. 1722, p. 012002). IOP Publishing.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University-Computer and Information Sciences*, *23*(2), 91-104.

Martínez, A. B., Lista, E. A. G., & Flórez, L. C. G. (2013). Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI. *Scientia et technica*, *18*(1), 185-191.

Moneda, N. (2012). Generación de expresiones regulares para la creación de reglas en aplicaciones de PLN. Universidad de Matanzas, Cuba.

Pabón, O. S., Torres, J. H., & Bucheli, V. A. (2020). Un enfoque de Análisis Inteligente de Datos para Apoyar la Relación con los Clientes. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (39), 52-66.

Pastor, P. (2022). Automatización de procesos ETL para la gestión de acuerdos de servicio (SLA). Universidad de Valladolid, España.

Plou, D. B. (s.f.). Expresiones Regulares: Una Herramienta Lógica para la Filosofía Experimental del Lenguaje. Universidad de Valparaíso, Chile.

Python Software Foundation. (2021). re - Regular expression operations - Python 3.10.0 documentation. Python 3.10.0 documentation. https://docs.python.org/3/library/re.html

Sguerra, Manuel Dávila. "Las expresiones regulares." *INVENTUM* 1.1 (2006): 31-37.