CLASIFICACIÓN DE EMOCIONES EN AUDIOS DE CALL CENTER UTILIZANDO CIENCIA DE DATOS

JOHAN SEBASTIAN MARULANDA ALMANZA

Código: 8984592

jsmarulanda@javerianacali.edu.co

Director

Gloria Ines Alvarez Vargas.

galvarez@javerianacali.edu.co

Co-director

Diego Luis Linares Ospina.

dlinares@javerianacali.edu.co

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI

2025

Agradecimientos

Son muchas las personas a quienes debemos agradecerles por el apoyo a lo largo de la Carrera y especialmente en este Proyecto:

Principalmente a Dios por darme la salud, la voluntad y oportunidad de estudiar este posgrado.

A mi familia por la confianza depositada y por su esfuerzo, dedicación y apoyo incondicional a lo largo de la vida.

A mis directores Gloria Ines Alvarez Vargas y Diego Luis Linares Ospina, por su profesionalidad, paciencia, apoyo y colaboración en todo momento a lo largo de este proyecto.

A la Pontificia Universidad Javeriana y a todos nuestros profesores, por permitirnos adquirir conocimientos, ayudarnos a crecer profesional y personalmente.

A todas aquellas personas que de una u otra forma contribuyeron en este proyecto.

Tabla de Contenido

Lista de Figuras

Lista de tablas

Resumen

Este proyecto se desarrolló con el objetivo de clasificar emociones en llamadas de call center utilizando transcripciones de audio y técnicas de machine learning, tomando como caso de estudio el centro de contacto de una Universidad de Cali. La investigación se enmarca dentro de una iniciativa más amplia en la que se exploraron un enfoque de análisis de transcripciones textuales, el presente trabajo se centró exclusivamente en la información textual derivada de los audios, evaluando la efectividad de diferentes modelos de clasificación.

El principal desafío fue desarrollar un clasificador capaz de identificar emociones de manera automatizada y eficiente a partir de datos textuales. Para ello, se realizó una limpieza y normalización de datos, seguida de un entrenamiento supervisado con modelos como Logistic Regression, Random Forest y Multi-Layer Perceptron (MLP). Se aplicó un ajuste de hiperparámetros utilizando Grid Search, optimizando el rendimiento de los modelos.

Los resultados obtenidos muestran que Random Forest alcanzó el mejor desempeño con una precisión del 72 % y un F1-score ponderado de 0.73, superando a Logistic Regression (67 %) y MLP (67 %). Esto sugiere que, a pesar de la complejidad del lenguaje en las llamadas de call center, el procesamiento basado en transcripciones puede ser una herramienta efectiva para la clasificación de emociones.

Este estudio demuestra el potencial de aplicar técnicas de machine learning en el análisis de emociones a partir de texto, ofreciendo una solución que puede mejorar la evaluación de la satisfacción del cliente y la calidad del servicio en entornos de atención telefónica.

Introducción

La creciente digitalización de los servicios ha llevado a las organizaciones a buscar formas innovadoras de mejorar la experiencia del cliente. En este contexto, los call centers desempeñan un papel crucial al ser el principal punto de contacto entre las organizaciones y sus clientes. Sin embargo, comprender y analizar las emociones expresadas durante estas interacciones sigue siendo un desafío significativo, ya que la comunicación humana es altamente compleja y matizada.

Este trabajo de grado se centra en la clasificación automática de emociones en llamadas de call center mediante el análisis de transcripciones de audio. La identificación de emociones en este tipo de interacciones es fundamental para evaluar la satisfacción del cliente y mejorar la calidad del servicio. Sin embargo, el procesamiento de texto en este contexto presenta múltiples desafíos, como la variabilidad del lenguaje, el ruido semántico y la ambigüedad emocional.

Para abordar este problema, se implementaron técnicas de aprendizaje supervisado [1], utilizando modelos de machine learning como Logistic Regression, Random Forest y Multi-Layer Perceptron (MLP). Se desarrolló un proceso de preprocesamiento y normalización de datos, asegurando que las transcripciones fueran adecuadas para el entrenamiento de los modelos. Posteriormente, se aplicó un ajuste de hiperparámetros para optimizar su rendimiento.

Los resultados obtenidos muestran que Random Forest alcanzó la mayor precisión con un 72 % de exactitud y un F1-score ponderado de 0.73, demostrando que el análisis de texto puede ser una herramienta efectiva para la detección de emociones en interacciones de servicio al cliente. Adicionalmente, se midió el tiempo de respuesta de cada transacción, permitiendo evaluar el desempeño en la atención telefónica.

Este estudio no solo representa un avance en la aplicación del procesamiento del lenguaje natural (NLP) en la clasificación de emociones, sino que también ofrece una solución práctica para la evaluación y mejora de la experiencia del usuario en servicios educativos. Además, se respetaron todas las normativas de privacidad de datos, garantizando la confidencialidad y protección de la información de los usuarios.

[1] El aprendizaje supervisado es un enfoque de Machine Learning en el cual un modelo se entrena utilizando un conjunto de datos etiquetado [1]

Capítulo 1

Contexto y objetivos

1.1. Problema de Investigación

1.1.1. Planteamiento del problema

La satisfacción del cliente es un factor crítico para cualquier organización que brinde bienes o servicios, incluyendo instituciones educativas como la Universidad estudiada. En este contexto, el call center desempeña un papel fundamental en la interacción con estudiantes, docentes y personal administrativo, facilitando la gestión de consultas, solicitudes y problemáticas. Sin embargo, evaluar de manera efectiva la calidad de estas interacciones y comprender el impacto emocional de cada llamada sigue siendo un desafío significativo.

Uno de los principales retos radica en la identificación y clasificación automática de emociones en conversaciones telefónicas. En un entorno donde se procesan múltiples llamadas por hora, depender exclusivamente de métodos manuales para evaluar la satisfacción del usuario es inviable debido a la subjetividad, el tiempo y los recursos que ello implica. Es por esto que la automatización de este proceso mediante técnicas de ciencia de datos se convierte en una solución viable para el análisis eficiente de emociones en interacciones telefónicas.

El procesamiento del lenguaje natural (NLP) y los algoritmos de aprendizaje automático han demostrado ser herramientas eficaces en la clasificación de emociones en textos. Aplicados a las transcripciones de llamadas del call center, estos métodos pueden identificar patrones lingüísticos asociados a emociones como amabilidad, desmotivación, duda e interés, proporcionando una evaluación cuantificable y objetiva del estado emocional de los usuarios. No obstante, el análisis de texto en este contexto presenta desafíos adicionales, como la variabilidad del lenguaje, la ambigüedad en la expresión de emociones y la influencia del contexto conversacional.

Este trabajo de grado se centra en el desarrollo de un modelo de clasificación de emociones basado en transcripciones de llamadas, con el fin de evaluar de manera automatizada la experiencia del usuario en el call center de una Universidad de Cali. A través de técnicas avanzadas de aprendizaje supervisado, este estudio busca proporcionar una herramienta que optimice la gestión de interacciones telefónicas, mejorando la calidad del servicio y contribuyendo a una mejor comprensión de las necesidades de los usuarios.

1.1.2. Formulación del problema

En la actualidad, el call center de la Universidad se enfrenta al reto de gestionar un volumen considerable de llamadas telefónicas, abordando diversas consultas, inquietudes y solicitudes de los usuarios. La interacción en estas llamadas no solo implica la transmisión de información, sino también la expresión de emociones, que pueden influir en la satisfacción del cliente y en la calidad del servicio percibido. Sin embargo, la capacidad de identificar y analizar automáticamente estas emociones sigue siendo un desafío.

El procesamiento del lenguaje natural (NLP) ha demostrado ser una herramienta eficaz para analizar transcripciones de conversaciones, permitiendo extraer información relevante y clasificar emociones a partir del contenido textual. En este contexto,

¿Cómo puede desarrollarse un modelo de machine learning capaz de clasificar automáticamente emociones en transcripciones de llamadas de un centro de contacto universitario, con el fin de mejorar la evaluación de la satisfacción del cliente?

Para responder esta pregunta, se plantean las siguientes interrogantes específicas

- ¿Cómo preparar los datos de transcripciones de llamadas para el desarrollo del modelo de machine learning?
- ¿Cómo aplicar técnicas de machine learning para obtener un modelo eficiente en la clasificación automática de emociones en texto?
- ¿Cómo evaluar la efectividad del modelo de machine learning en la detección de emociones?

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un modelo de aprendizaje supervisado para clasificar las emociones presentes en las transcripciones de audios del centro de contacto de una universidad.

1.2.2. Objetivos específicos

- Preparar un conjunto de datos de transcripciones de llamadas telefónicas del centro de contacto de la Universidad.
- Desarrollar modelos de aprendizaje semi supervisado para la clasificación de emociones en texto.
- Evaluar el rendimiento de los modelos utilizando métricas estándar de clasificación (precisión, recall, F1-score).

1.3. Alcance

El alcance de este proyecto de grado es desarrollar un modelo de aprendizaje automático que pueda analizar y clasificar las emociones predominantes en las llamadas del call center de una Universidad

de Cali. Este análisis puede ser implementado como insumo para mejorar la satisfacción del cliente, esto queda a consideración de la Universidad.

Aunque el objetivo es desarrollar un prototipo funcional, hay aspectos que quedarán para trabajos futuros. Estos pueden incluir la mejora del modelo con más datos, la expansión del modelo para analizar otros tipos de interacciones con los clientes, y la integración del modelo en los sistemas existentes de la universidad.

Como restricciones se encuentran:

- Completar el proyecto en el término de un año.
- Limitaciones que impone la normativa legal vigente sobre el uso de datos personales con fines de investigación.
- Desafíos relacionados con el almacenamiento y procesamiento de audios (en términos de poder computacional) en los modelos a emplear

1.4. Justificación

La pertinencia y utilidad de este proyecto se puede argumentar en tres grandes grupos de posibles beneficios a diferentes grupos de interés.

Por una parte, la mejora de la experiencia del cliente y la optimización de las comunicaciones cliente-empresa, lo que se puede lograr al permitir una clasificación automática de las emociones de las llamadas en un call center, puesto que, al comprender las emociones expresadas por los clientes las empresas pueden ajustar sus estrategias para satisfacer mejor sus necesidades y expectativas. Además, al conocer la dinámica de la expresión emocional durante dichas conversaciones, se pueden implementar mejoras en el proceso de comunicación y atención que prevengan la aparición de emociones negativas en el flujo de las conversaciones

Adicionalmente, es posible impactar en el mejoramiento de indicadores relacionados con la optimización de la gestión del talento humano de la empresa. El analizar las emociones en las interacciones con los clientes proporciona una herramienta valiosa para la formación y desarrollo del personal del call center. Los supervisores pueden utilizar la retroalimentación basada en emociones para mejorar las habilidades de comunicación y la gestión de situaciones emocionalmente cargadas a los agentes del centro de contacto. Por otra parte, la automatización de la clasificación de emociones a través de modelos semi-supervisados de machine learning permite una mayor eficiencia operativa en el análisis de grandes volúmenes de conversaciones. Esto libera recursos humanos para tareas más estratégicas, mientras se mantiene un monitoreo continuo de la satisfacción del cliente.

Finalmente, el proyecto contribuye al ámbito académico al explorar y desarrollar técnicas avanzadas de procesamiento de lenguaje natural y machine learning aplicadas a contextos específicos, como el análisis de emociones en conversaciones de call center. Los resultados pueden servir como base para futuras investigaciones y desarrollos en el campo.

En cuanto a la viabilidad del proyecto, se puede afirmar que existen una serie de factores que indican que se pueden llegar a conseguir los objetivos del mismo y algunos que se deben controlar y aterrizar para no poner en juego la viabilidad de este.

En primer lugar, el acceso del corpus de datos y la calidad de los mismos es un punto clave de partida. La apuesta principal es establecer una colaboración efectiva con el call center de la Universidad Javeriana Cali, que será esencial para garantizar el acceso a los datos necesarios y comprender los contextos específicos de las interacciones. La cooperación también puede facilitar la implementación exitosa de las soluciones propuestas.

En segundo lugar, se conforma un equipo interdisciplinario de profesionales con experiencia en procesamiento de lenguaje natural (PLN) y machine learning. Este equipo puede diseñar, implementar y afinar modelos de clasificación de emociones, así como abordar desafíos específicos relacionados con el análisis de audio en el contexto de conversaciones de call center y con la comprensión, clasificación y etiquetado de las emociones [35]

Contribuye a una atención al cliente más efectiva y personalizada al identificar y abordar las emociones de los clientes durante las interacciones [4]. Esto puede generar una mayor satisfacción y fidelidad del cliente, mejorando así las relaciones empresa-cliente y esto puede redundar directamente en el mejoramiento de la percepción social de las personas acerca de su relación con las empresas que les prestan servicios en diferentes áreas.

Facilita la identificación de situaciones emocionalmente desafiantes para los agentes de call center, permitiendo a las empresas implementar medidas para gestionar el estrés y mejorar el bienestar laboral.

Optimiza los recursos al automatizar la clasificación de emociones, lo que puede llevar a una mayor eficiencia operativa y reducción de costos asociados al análisis manual de conversaciones.

La mejora en la experiencia del cliente puede contribuir a la retención de clientes, ya que un servicio más personalizado y sensible a las emociones puede influir positivamente en las decisiones de compra y lealtad.

La retroalimentación basada en emociones puede informar programas de formación y desarrollo del personal, aumentando la eficacia de los agentes de call center y, por ende, mejorando la eficiencia en la gestión de clientes.

Contribuye al desarrollo y aplicación de técnicas avanzadas de procesamiento de lenguaje natural y machine learning en un contexto específico, avanzando en la investigación y desarrollo de la IA aplicada a la comprensión de emociones en conversaciones.

Introduce innovaciones tecnológicas en la operación de call centers al incorporar herramientas de análisis de emociones, permitiendo a las empresas mantenerse a la vanguardia en la mejora continua de servicios.

Capítulo 2

Marco Referencial

2.1. Glosario

- **Inteligencia Artificial** [2]: Es una rama de las ciencias computacionales cuyo principal objetivo es construir sistemas artificiales con capacidad de emular una habilidad que denote una inteligencia similar a la de los seres vivos, para ello, es necesaria una serie de instrucciones que especifiquen las acciones que debe ejecutar (un algoritmo).
- **Machine Learning** [3]: Es un campo de la Inteligencia Artificial, tiene el propósito de imitar el aprendizaje humano al generar conocimiento con la ayuda de ejemplos y de la experiencia, su principal característica es la capacidad que tiene de aplicar automáticamente cálculos matemáticos complejos a grandes datos una y otra vez, más y más rápido. Hace uso de diferentes algoritmos para procesar datos. Una de las mayores ventajas de Machine Learning es que permite entrenar un algoritmo de aprendizaje para que se adapte y vaya cambiando de acuerdo a las características de un histórico de datos que es suministrados ya sea con datos previamente etiquetados (Aprendizaje supervisado) o con los desconocidos para el analista y la computadora (Aprendizaje no supervisado).
- **Aprendizaje supervisado** [4]: El término aprendizaje supervisado se refiere a darle al algoritmo un conjunto de datos donde se tienen las "respuestas correctas". Hace predicciones a partir de datos compuestos por vectores de entrada $x = (x_1, x_2, ..., x_n)$ con sus correspondientes vectores de salida $y = (y_1, y_2, ..., y_n)$, a esto se le conoce como conjunto de datos históricos $S = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

  El aprendizaje supervisado realiza tareas predictivas las cuales se dividen en dos tipos: regresión y clasificación.
- **Aprendizaje no supervisado** [4]: En el aprendizaje no supervisado, se trabaja con datos compuestos solamente por el vector de entrada $x = (x_1, x_2, ..., x_n)$, no existe un vector de salida y, es decir, no se cuenta con un conjunto de datos de entrenamiento, por lo que no existe un porcentaje de predicción. El aprendizaje no supervisado es útil para tareas de clustering (agrupamiento).
- **Procesamiento de lenguaje natural (PNL)** [5]: Son sistemas capaces de reconocer, pro-

cesar y emular el lenguaje humano. El surgimiento de esta área fue gracias al test de Turing.

2.2. Marco Teórico

Teniendo en cuenta que nuestro proyecto gira en torno al desarrollo de un modelo de machine learning que permita la clasificación de conversaciones en términos del contenido emocional de las mismas, el marco teórico del presente trabajo debe estar comprendido por: la revisión de teorías y conceptos asociados a las diferentes técnicas de machine learning disponibles y que sean potencialmente útiles para el desarrollo del modelo; también será importante identificar un modelo explicativo de las expresiones emocionales que nos permita fundamentar la idea de que es posible establecer una relación entre el contenido acústico de las conversaciones (señales de audio) y la expresión de una emoción en particular y; un marco teórico que permita comprender el concepto de satisfacción del cliente y su relación con el contenido emocional de las conversaciones en el contexto de los centros de contacto.

2.2.1. Características del Audio

La formalización cualitativa del habla implica la convolución de la respuesta en frecuencia del tracto vocal con el pulso glótico. En términos simples, esto significa que el pulso glótico, originado en las cuerdas vocales, se filtra a través del tracto vocal, dando lugar a la señal acústica. El pulso glótico, donde se origina el fundamento de la frecuencia sonora o tono, determina las frecuencias más altas. La forma específica del tracto vocal da lugar a la generación de fonemas o al timbre característico de la señal. [6]

A continuación una imagen que muestra cada uno de los elementos anteriormente mencionados [6]:

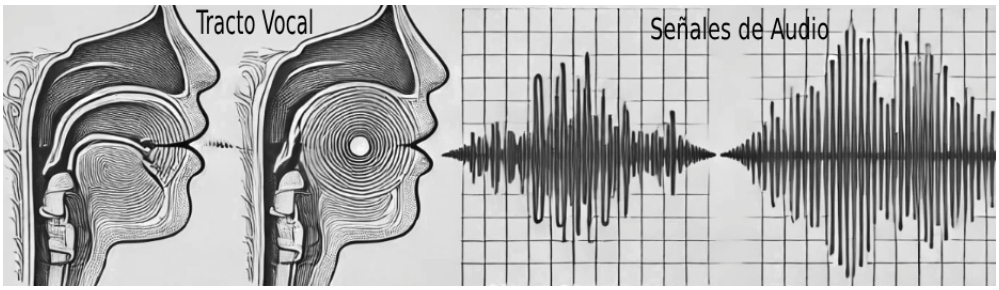

Figura 2.1: Esquema del Habla

Frente a esta señal del habla, el procesamiento del audio cuenta con sus propias características mencionadas por el Phd. J.R. Zapata Gonzalez que resume de la siguiente manera [7].

- **Energía:** La Energía de una señal se refiere a la suma total de las magnitudes de dicha señal, siendo indicativa de la intensidad o volumen de una señal de audio.
- **Zero Crossing Rate (Número de cruces por cero):** Este parámetro indica cuántas veces una señal cruza el eje horizontal por cero, proporcionando información sobre la frecuencia y la variación de la señal.

- **STFT - Transformada corta de Fourier (Short-Time Fourier Transform):** Las señales de audio no estacionarias, es decir, cuyas características cambian con el tiempo, pueden analizarse mediante la STFT. Esta técnica divide la señal en segmentos temporales para examinar sus características de frecuencia en diferentes intervalos de tiempo.
- **Espectrograma:** El Espectrograma representa la intensidad de las frecuencias a lo largo del tiempo y se obtiene al calcular la magnitud al cuadrado de la STFT. En el procesamiento de audio, a menudo se enfoca en la amplitud en escala logarítmica (dB), omitiendo la información de fase.
- **Mel-Espectrograma:** Utilizando la escala Mel, que relaciona la frecuencia percibida con la medida real, el Mel-Espectrograma se centra en la amplitud logarítmica debido a la percepción humana logarítmica de la intensidad del sonido.
- **Mel Frequency Cepstral Coefficients (MFCCs):** Los MFCCs son coeficientes utilizados en la representación del habla, basados en la percepción auditiva humana. Estos coeficientes se centran en información valiosa, excluyendo detalles irrelevantes como el ruido de fondo, emociones, volumen y tono, siendo comúnmente empleados para describir el timbre.
- **Características Espectrales:** Para la clasificación de sonidos, se utilizan momentos espectrales como el centroide, ancho de banda, asimetría, curtosis, entre otros, que proporcionan información sobre la distribución espectral. El centroide espectral, por ejemplo, indica la frecuencia central de la energía espectral, similar a una media ponderada.
- **Ancho de Banda espectral (Spectral Bandwidth):** Calcula el ancho de banda espectral, proporcionando información sobre la dispersión de las frecuencias en un espectro.
- **Contraste Espectral (Spectral Contrast):** Considera los picos y los valles espectrales, evaluando las diferencias entre ellos en cada sub banda de frecuencia.
- **Desplazamiento espectral (Spectral Rolloff):** Indica la frecuencia por debajo de la cual se encuentra un porcentaje específico de la energía espectral total.

2.2.1.1. Las emociones humanas

El estudio de las emociones humanas constituye un fascinante campo multidisciplinario que involucra a la psicología, la neurociencia, la filosofía y otras disciplinas afines. La complejidad de las emociones radica en su capacidad para influir en el pensamiento, el comportamiento y la experiencia subjetiva de los individuos. Los investigadores buscan comprender la naturaleza de las emociones, su origen evolutivo, y cómo interactúan con la cognición y el cuerpo humano. Desde la alegría hasta el miedo, las emociones desempeñan un papel crucial en la toma de decisiones, las relaciones interpersonales y la salud mental. El avance tecnológico, incluyendo técnicas de imagen cerebral y análisis de datos, ha permitido un mayor conocimiento sobre la base neural de las emociones. Este campo en constante evolución no sólo arroja luz sobre los misterios de la mente humana, sino que también tiene aplicaciones prácticas en la mejora de la salud mental, la inteligencia artificial y el diseño de interacciones humanas más efectivas. [8]

2.2.1.2. Emociones básicas

Paul Ekman es un autor que defiende la idea de la existencia de un grupo de emociones básicas universales, cuya clasificación es ampliamente utilizada con fines de clasificación discreta de las emociones en distintas áreas, incluyendo recientemente la mayoría de trabajos en ciencia de datos que buscan la identificación automática de las emociones. Estas emociones básicas son: alegría, tristeza, miedo, ira, desagrado y sorpresa. [9]

Las categorías empleadas en este proyecto son las siguientes:

- **Interés:** Conversaciones en las que el cliente demuestra activamente intención de continuar con el proceso, mediante expresiones de curiosidad, solicitud de más información o afirmaciones positivas sobre el mismo.
- **Desagrado:** Fragmentos en los que el cliente o sus acompañantes expresan de forma explícita críticas, molestias o desacuerdos con aspectos del proceso de admisión.
- **Desmotivado:** Intervenciones en las que el cliente manifiesta desinterés, desánimo o baja disposición para continuar con el proceso, usando expresiones verbales que reflejan apatía o falta de motivación.
- **Amabilidad:** Conversaciones donde se evidencian expresiones corteses o respetuosas por parte del cliente, sin que necesariamente se asocien a un interés específico en el proceso.
- **Decepción:** Casos en los que el cliente manifiesta insatisfacción o expresa que sus expectativas no fueron cumplidas, mediante frases que reflejan frustración o desencanto.
- **Incomunicado:** Situaciones en las que se evidencia una falta de entendimiento entre los participantes, reflejada en repeticiones constantes, malentendidos explícitos o falta de respuesta a lo solicitado.

Nota: en el corpus que se obtuvo practicando con estas categorías, Desagrado tiene un desbalanceo muy grande frente a las otras categorías, por lo que termina incluyéndose en la categoría Decepción.

2.2.1.3. Percepción e identificación de las emociones

En cuanto al enfoque principal del abordaje de la expresión emocional del presente trabajo, el principal referente teórico es Ekman, quien defiende una postura acerca de la existencia de formas de expresión emocional universales a toda la especie humana, lo que incluye, además de una fuerte evidencia acerca de la universalidad de las expresiones faciales, unas hipótesis fuertes alrededor de la universalidad en todos los canales de expresión emocional, incluyendo la señal acústica del habla, que además implicaría una universalidad de dichos patrones acústicos. [10]

2.2.1.4. Machine Learning

Algunos modelos muy empleados para el procesamiento de las características del Audio son los siguientes:

- **Regresión Logística:** Se trata de un modelo estadístico empleado para estimar la probabilidad de ocurrencia de un evento particular. Comparte similitudes con la regresión lineal,

pero la diferencia radica en que la variable objetivo es binaria, (1 o 0) en lugar de generar un resultado. [11]

En el modelo de regresión logística, la variable respuesta (Y) es dicotómica, posee valores 1 o 0, con probabilidad π_i para $Y_i = 1$ y probabilidad $1 - \pi_i$ para $Y_i = 0$, según Hosmer y Lemeshow (2000). [12]

El análisis de Regresión Logística comprende la estimación de la probabilidad de que ocurra un evento (variable de respuesta dicotómica; con valores 0 y 1) como función de los valores de p variables independientes (predictoras).

Consideremos Y una variable respuesta y una colección de p variables independientes expresado por el vector $X' = (x_1, x_2, \ldots, x_p)$.

La forma específica del modelo logístico con p variables predictoras está representado por:

$$\pi = \pi(x) = P(Y = 1|x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

Que representa que la probabilidad condicional de que el evento $Y = 1$ ocurra dada la ocurrencia de un conjunto de variables X (Probabilidad de éxito).

Una transformación de $\pi(x)$ que es fundamental en el estudio de la regresión logística es la transformación logit. Esta transformación se define en términos de $\pi(x)$, como:

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Donde β_0 es la constante y los β_i son los coeficientes de los predictores x_i del modelo.

La importancia de esta transformación es que $g(x)$ posee muchas de las propiedades deseables de un modelo de regresión lineal. La función logit es lineal en sus parámetros, puede ser continuo y variar de $-\infty$ a $+\infty$, dependiendo del rango de x.

- **Máquinas de soporte vectorial (SVM):** Modelo que busca identificar un hiperplano óptimo que logre la mejor separación entre dos clases diferentes de puntos de datos. lo que implica encontrar aquel que tenga el margen más amplio entre dso clases. [13]
- **Árboles de decisión:** Modelo que se asemeja a una estructura en forma de árbol, donde los nodos internos representan características o atributos, las ramas denotan reglas de decisión, y cada nodo hoja indica el resultado. La selección de nodos se realiza mediante conceptos como la entropía o la ganancia de información, propiedades estadísticas que evalúan cómo un atributo específico separa los ejemplos de entrenamiento según su clasificación objetivo. [14]
- **Random Forest (RF):** Modelo que involucra varios árboles de decisión combinados con bagging. Al usar bagging, cada árbol ve distintas porciones de los datos, ninguno usa todos los datos de entrenamiento. Esto hace que cada uno se entrene con distintas muestras para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor. [14]

- **Redes neuronales de perceptrón multicapa (FFNN):** Una red neuronal se presenta como un procesador masivamente distribuido y en paralelo con una inclinación inherente para capturar conocimiento experimental y hacerlo accesible para su aplicación. Este modelo guarda similitudes con el cerebro en dos aspectos fundamentales: 1. La red adquiere conocimiento a través de un proceso de aprendizaje. 2. Las fuerzas de conexión entre neuronas, denominadas ponderaciones sinápticas, se emplean para retener la información adquirida.[24] El perceptrón, la forma más básica de red neuronal utilizada para clasificar patrones linealmente separables, se expande en el perceptrón multicapa, incorporando múltiples capas. Este diseño incrementa su capacidad para abordar problemas que no presentan una separación lineal clara. [15]

  Representamos el error en el nodo j en el punto de datos n como $e_j(n) = d_j(n) - y_j(n)$, donde d es el valor objetivo y y es el valor producido por el perceptrón. Cuando hacemos las correcciones a los pesos de los nodos basados en estas correcciones, las cuales minimizan el error en toda la salida, dado por

  $$\mathcal{E}(n) = \frac{1}{2}\sum_j e_j^2(n).$$

  Usando un *descenso de gradiente*, encontramos que nuestro cambio en cada peso es

  $$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

  Donde y_i es la salida de la neurona anterior y η es el *ritmo de aprendizaje*, que se selecciona cuidadosamente para que los pesos converjan a una respuesta lo suficientemente rápida, sin producir oscilaciones. En la programación de aplicaciones, este parámetro generalmente va de 0.2 a 0.8.

  La derivada a ser calculada depende del campo local inducido v_j, el cual varía por sí mismo. Es fácil probar que para un nodo de salida, esta derivada puede ser simplificada a

  $$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n))$$

  Donde ϕ' es la derivada de la función de activación descrita más arriba, la cual no varía. El análisis es más complicado para el cambio en pesos en un nodo oculto, pero puede mostrarse que la derivada relevante es

  $$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'(v_j(n))\sum_k \left(-\frac{\partial \mathcal{E}(n)}{\partial v_k(n)} w_{kj}(n)\right).$$

  Esto depende en el cambio en los pesos de los nodos k, que representan la capa de salida. Entonces, para cambiar los pesos de la capa oculta, debemos conocer los errores en la capa de salida de forma acorde a la derivada de la función de activación, y este algoritmo representa una retrospectiva de la función de activación.

- **Redes neuronales recurrentes (RNN):** Es un tipo de red neuronal artificial que utiliza datos secuenciales o datos de series temporales. Muy empleada para procesamiento de audio. [16]
- **Redes neuronales convolucionales (CNN):** Es un tipo de red que es un variante del perceptrón multicapa, diseñada específicamente para trabajar con matrices bidimensionales. Esta característica las hace especialmente adecuadas para tareas relacionadas con la visión artificial. [17]

2.2.1.5. Métricas de evaluación de aprendizaje supervisado

A continuación, se presentan algunas de las técnicas utilizadas para evaluar modelos de aprendizaje semi-supervisado [18] [19]. Es de utilidad hacer uso de la matriz de confusión para representar de forma gráfica los valores sobre los que basamos varias métricas:

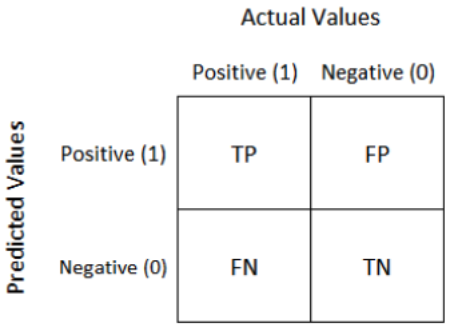

Figura 2.2: Matriz de Confusión
Fuente: Revista Towards Data Science

Donde:
True Positives (TP): Valores clasificados por el algoritmo correctamente.
True Negatives (TN): Valores no clasificados por el algoritmo correctamente.
False Positives (FP): Valores clasificados por el algoritmo pero que no pertenecen.
False Negatives (FN): Valores no clasificados por el algoritmo pero que pertenecen.

- **Exactitud:** Esta métrica representa la proporción de predicciones correctas en relación con el total de predicciones realizadas. Se calcula dividiendo la suma de verdaderos positivos y verdaderos negativos entre el número total de casos.

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precisión:** La precisión mide la proporción de predicciones positivas que son acertadas. Su cálculo se realiza dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos positivos.

$$\text{Precisión} = \frac{TP}{TP + FP}$$

- **Recall o Sensibilidad:** Esta métrica evalúa la proporción de casos positivos reales que son correctamente identificados por el modelo. Se calcula dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.

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

- **F1-score:** El F1-score combina precisión y recall en una única medida. Su cálculo se realiza mediante el doble del producto de precisión y recall dividido por la suma de precisión y recall.

$$\text{F1-score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$$

- **Curva ROC y AUC:** La Curva ROC (Receiver Operating Characteristic) es un gráfico que ilustra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para diversos umbrales de clasificación. El AUC (Área Bajo la Curva) representa la probabilidad de que el modelo clasifique correctamente casos positivos y negativos seleccionados al azar. Cuanto más cercano sea el AUC a 1, mejor será el rendimiento del modelo.

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

El AUC se calcula como el área bajo la curva ROC:

$$\text{AUC} = \int_0^1 \text{TPR}(x)\, dx$$

2.2.1.6. Técnicas de etiquetado semi automático

Para la generación de etiquetas para los datos no etiquetados a partir de los etiquetados, contamos con las siguientes técnicas [20]

- **Co-entrenamiento:** La estrategia de co-entrenamiento implica la capacitación simultánea de dos o más modelos, cada uno con conjuntos de características distintas para un mismo ejemplo. Posteriormente, se emplean las predicciones de cada modelo para asignar etiquetas a los datos no etiquetados del otro modelo. Este enfoque se basa en la suposición de que las características son independientes y se complementan entre sí.
- **Aprendizaje auto-didacta:** En el aprendizaje auto-didacta, se entrena un modelo utilizando datos etiquetados, y luego se emplean las predicciones de dicho modelo para etiquetar los datos que carecen de etiquetas. La premisa subyacente es que el modelo tiene la capacidad de generalizar de manera efectiva a partir de los datos etiquetados.

- **Aprendizaje basado en grafos:** La metodología de aprendizaje basado en grafos implica la representación de datos en forma de un grafo, donde los ejemplos son nodos y las similitudes entre ellos se expresan mediante aristas. Posteriormente, las etiquetas de los nodos etiquetados se propagan a los nodos no etiquetados a través de las aristas, utilizando algún criterio de optimización o inferencia.
- **Aprendizaje basado en modelos generativos:** Este enfoque implica modelar la distribución conjunta de las características y las etiquetas. Luego, se utiliza el algoritmo de Expectation-Maximization (EM) para estimar los parámetros del modelo y asignar etiquetas a los datos no etiquetados. Este método parte de la suposición de que los datos siguen una distribución probabilística específica.

2.2.1.7. Procesamiento del lenguaje natural (PLN)

Una de las tareas fundamentales de la Inteligencia Artificial (IA) es la manipulación de lenguajes naturales usando herramientas de computación que forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina.

El PLN consiste en la utilización de un lenguaje natural para comunicarnos con una computadora, debiendo esta entender las oraciones proporcionadas, el uso de lenguajes naturales, facilitan el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje. [21]

Aplicaciones del PLN
Las aplicaciones del PLN son muy variadas, y han dado lugar a muchas líneas de investigación relacionadas con las diferentes tareas que hay que realizar. Algunas de estas aplicaciones y tareas dentro del PLN son:

- Recuperación y extracción de información.
- Generación automática de resúmenes a partir de textos.
- Reconocimiento de entidades y conceptos.
- Análisis del discurso.
- Comprensión del lenguaje natural y formalización utilizando lenguajes formales.
- Generación del lenguaje natural en diferentes lenguas.
- Reconocimiento óptico de caracteres-del inglés Optical Character Recognition (OCR).
- Reconocimiento de la voz.
- Etiquetado morfológico, sintáctico y semántico.
- Análisis de sentimientos para determinar la polaridad del contenido.

Componentes básicos de cualquier técnica de PLN

- Análisis morfológico o "léxico": Consiste en el análisis de las palabras que forman las sentencias para extraer sus lemas o raíces, rasgos flexivos, unidades léxicas compuestas, etc.

- Análisis sintáctico: Consiste en el análisis de la estructura de las secuencias, de acuerdo con una gramática del lenguaje analizado.
- Análisis semántico: Orientado a la extracción del significado de las sentencias y eliminación de ambigüedades.
- Análisis pragmático: Orientado al análisis del contexto donde se encuentra inmerso el texto analizado y cómo influye éste en el significado del mismo.

El presente trabajo se centra en el primer objetivo principal mencionado anteriormente, que está relacionado con la clasificación de emociones. En los siguientes apartados se hará una revisión de las principales técnicas de procesado de textos.

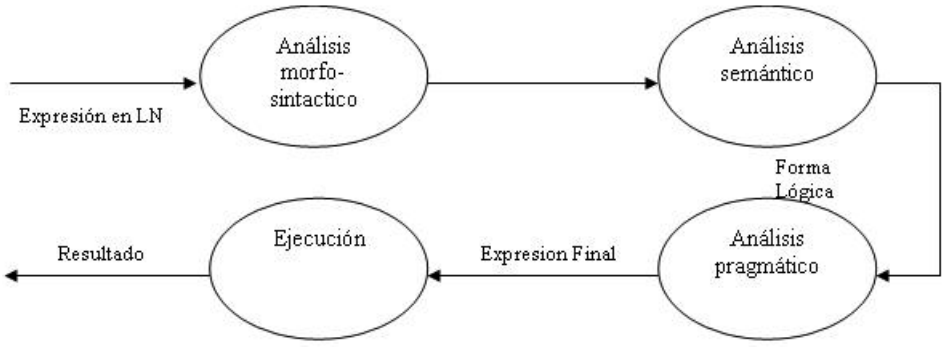

Figura 2.3: Arquitectura de un Sistema de Procesamiento de Lenguaje Natural

2.3. Antecedentes

A continuación, se enumeran los antecedentes más importantes encontrados en lo referente a la identificación y clasificación de emociones a partir de señales auditivas y luego se presenta la evidencia de lo revisado hasta el momento que tiene relación directa con el contexto de los centros de contacto.

- En el campo del reconocimiento emocional por medio de señales de voz, es importante mencionar el trabajo de tesis doctoral de Humberto Perez [22], quien utilizó un modelo de clasificación continuo de las emociones, haciendo uso de un auto entrenamiento semi-supervisado, el uso de agrupamiento difuso mediante la técnica Fuzzy C-means (FCM) y la técnica probabilística de campos aleatorios de Markov (CAM), con el fin de generar un método de predicción y reconocimiento de patrones emocionales espontáneos. De este trabajo puede ser útil el modelo de clasificación continuo de las emociones y su uso en el posible etiquetado automático del audio.
- También vale la pena mencionar el trabajo de grado de Sanchez [23], quien utiliza una técnica de aprendizaje automático para entrenar un modelo a partir de imágenes y voz e identificar emociones discretas. De este trabajo cabe resaltar que los resultados arrojaron una mejor predicción de las imágenes que de la voz y una mejor predicción cuando se mezclaban imágenes y voz que cuando se hacían por separado. De este trabajo nos valdremos para sustentar la

importancia de desarrollar modelos más confiables de predicción con base en voz, en el contexto de los centros de contacto.

- Por otra parte, el trabajo de Bello y colaboradores [24], ilustra un abordaje del reconocimiento de las emociones por medio del análisis de fragmentos de la voz que utiliza la transformada rápida de Fourier (FFT) y en coeficientes de correlación de Pearson. El trabajo resulta interesante, además, porque presenta una metodología completa basada en evidencia científica para lograr etiquetar las emociones, basados en parámetros observables de la señal del habla.
- En otro trabajo, se utilizó la base de datos RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) [25] para entrenar un modelo a través de técnicas de aprendizaje supervisado, redes neuronales artificiales y redes neuronales convolucionales. El modelo logró un nivel de precisión del 85 % de clasificación de las emociones de personas en videos con audio que mantienen una configuración parecida a las de la base de datos de entrenamiento [26]. Este trabajo nos permitirá sustentar e informar el uso de las técnicas de aprendizaje supervisado y redes neuronales con el fin de escoger la más adecuada para el contexto de nuestro trabajo con audios de conversaciones reales.
- Así mismo, el trabajo de Pervaiz y Ahmed [27] se centra en mejorar la comunicación entre humanos y sistemas tecnológicos al reconocer y comprender las emociones en el discurso hablado en español. Se identifican problemas como la falta de arquitecturas de reconocimiento de emociones y la ausencia de API para investigación en este ámbito. La solución propuesta incluye el desarrollo de un clasificador de emociones, comparando su rendimiento con redes neuronales recurrentes y convolucionales. Además, se integra el clasificador en un framework multimodal y se crea una API para facilitar su uso. La metodología se basa en dos conjuntos de datos etiquetados con emociones básicas, abordando la diversidad de expresiones emocionales en el habla en español.

  En resumen, el trabajo busca mejorar la interacción persona-máquina al reconocer emociones en el discurso hablado en español, proponiendo soluciones prácticas y relevantes. Este trabajo nos permite guiar el desarrollo de nuestro modelo de identificación de emociones y alimentar la construcción de código propio y mejorado a partir de las soluciones planteadas de la comparación entre redes neuronales recurrentes y convolucionales, además de servirnos de guía para el posible desarrollo de una API. Este trabajo es especialmente relevante por utilizar el idioma español.
- En cambio, el proyecto de reconocimiento de emociones en la voz de Lerache y Elkfury [28] aborda la dificultad de analizar emociones en el discurso hablado en español rioplatense, especialmente en entornos interactivos con robots. Se reconoce que la voz humana tiene limitaciones en la transmisión de emociones, y se propone utilizar espectrogramas y técnicas de aprendizaje profundo, como redes neuronales convolucionales (CNN) y recurrentes (RNN), para desarrollar un clasificador de emociones. Además, se busca mejorar la comunicación persona-máquina mediante la creación de una API para la explotación del modelo, la comparación de rendimiento entre los clasificadores y la integración en un framework multimodal. Se plantea un método para la conversión de enfoques categóricos a dimensionales y se establecen bases para la mejora continua mediante la adquisición de nuevos datos para el entrenamiento del modelo. En conjunto, la investigación busca avanzar en la comprensión y aplicación de las emociones en el discurso hablado, contribuyendo a la mejora de la interacción emocional

en entornos tecnológicos. Este trabajo es de alta relevancia en nuestro proyecto, puesto que trabaja con particularidades propias de la señal de audio en idioma español, lo cual es escaso en la literatura.

- Asimismo, los mismos autores abordan [29] el reconocimiento de emociones en señales de voz mediante el uso de redes neuronales profundas. La identificación precisa de emociones en el habla es crucial para diversas aplicaciones, y el estudio se centra en evaluar los efectos de funciones de pérdida y técnicas de aumento de datos en la clasificación de siete emociones. Motivado por la colaboración con la empresa SEAT para evaluar sistemas basados en deep learning en situaciones de conducción, el proyecto tiene como objetivo principal implementar un sistema eficaz de reconocimiento de emociones en señales de voz, aprovechando información espectral. Las conclusiones resaltan el éxito de la implementación, subrayando la necesidad de evaluar configuraciones y técnicas para mejorar la precisión del modelo y destacando su potencial aplicabilidad en situaciones prácticas, incluyendo el reconocimiento de emociones en contextos de conducción.
- El trabajo de grado realizado por M. Patricio G. y A. Berlanga en 2022 se enfocó en la implementación de un análisis de emociones a través de la conversión de audio a texto. Este proceso de análisis fue llevado a cabo mediante el entrenamiento de un modelo utilizando un corpus de EMOFILM y la aplicación de herramientas de Procesamiento del Lenguaje Natural (NPL). El resultado final consistió en la implementación de un chatbot que permite a las personas enviar archivos de audio y recibir la clasificación emocional correspondiente [30]. Este documento presenta nuevos enfoques de modelos utilizados para el análisis de audios, centrándose en las características específicas de grabaciones de corta duración. Creemos que estos enfoques pueden ser aplicados de manera efectiva en nuestro trabajo.
- Finalmente, en un trabajo muy cercano a los intereses del presente proyecto, Bolo y colaboradores establecieron correlación entre los niveles de satisfacción reportada por el cliente y la detección automática de patrones emocionales de la señal del habla en una base de datos de 160630 llamadas de call center, analizando únicamente la parte de la llamada correspondiente al cliente y haciendo uso de técnicas de reconocimiento automático de texto y de etiquetado manual y técnicas de regresión, con el fin de generar modelos predictivos que combinaron el texto y la señal auditiva para predecir la valencia emocional y la ira, en particular [31]
- La empresa alemana Audeering ofrece en su portafolio de servicios una serie de dispositivos para análisis de voz, detección de emociones, biomarcadores de voz que permiten la detección de enfermedades como parkinson y Covid-19. Sus soluciones para empresas son implementadas para la detección de emociones en videojuegos de realidad aumentada, call center e investigaciones de mercado. Han sido pioneros desde hace 20 años en este tipo de investigaciones, sin embargo, al ser un producto costoso por la alta tecnología de sus dispositivos, no se ha podido masificar su comercialización, al menos en latinoamérica. [32]
- La Unión Europea en 2020 financió el proyecto Mixed Emotions desarrollado por diferentes colaboradores entre los que se encuentra el Dr. Paul Buitelaar como Director del Proyecto. Han tenido grandes avances en aplicaciones multilingües y multimodales de análisis de grandes datos, logrando determinar un perfil muy completo de la emoción del usuario, aunque contiene código abierto y han realizado aplicaciones con fines comerciales, han sido como proyectos piloto y hacen la advertencia de que lograr la interpretación de diferentes usuarios y diferentes

fuentes de información, estilos e idiomas es un reto muy amplio y aún no se ha estandarizado para poder realizarlo en un contexto industrial. [33]

Capítulo 3

Preparación de datos

En este capítulo se describe el proceso de recolección, organización y Preparación de los Audiosón de las transcripciones de llamadas telefónicas del centro de contacto de la Universidad. Se detalla el procedimiento de transcripción de los audios a texto, la limpieza de los datos textuales eliminando el ruido y la normalización del texto. Este trabajo es fundamental para asegurar la calidad y precisión del análisis posterior.

3.1. Preparación de datos

3.1.1. Adquisición de datos

Las transcripciones utilizadas en este estudio provienen del centro de contacto de una Universidad. Estas transcripciones fueron generadas a partir de grabaciones de interacciones entre los agentes del centro de contacto y los usuarios que llamaron para recibir información o asistencia.

Se nos entregaron grabaciones del año 2023 entre los meses de abril y agosto, organizadas en tres paquetes de datos:

Envíos de grabaciones	**Número de audios**
Grabaciones-1	258
Grabaciones-2	960
Grabaciones-3	343
Total	1561

Tabla 3.1: Registros de grabaciones del Call Center según paquete compartido

3.2. Descripción del Dataset

El dataset utilizado en este estudio está compuesto principalmente por las transcripciones textuales de llamadas telefónicas realizadas al call center de una institución educativa. Estas transcripciones fueron generadas mediante algoritmos de reconocimiento automático del habla (ASR) aplicados sobre los audios originales.

A partir de estas transcripciones, y con el objetivo de enriquecer el análisis, se extrajeron también algunas características adicionales a partir de los audios originales, como la duración de la llamada y la proporción de silencio. No obstante, el análisis exploratorio y la clasificación de emociones se basan en los datos textuales.

Cada registro del dataset contiene la siguiente información:

- **Transcripción del audio:** Representa el contenido verbal de la llamada, obtenido a partir de algoritmos de reconocimiento de voz.
- **Duración del audio:** Longitud de la llamada en segundos. Se utiliza como información contextual complementaria.
- **Proporción de silencio:** Tiempo relativo en el que no se detectó voz en la grabación, útil como indicador de pausas o silencios prolongados.
- **Etiqueta de emoción:** Categorización del contenido textual en una de las emociones predefinidas mediante un proceso de etiquetado semiautomático.

3.3. Etiquetado semiautomático

El análisis de la duración de las llamadas es importante para comprender las características del corpus de audio utilizado en el entrenamiento del modelo. Este análisis permite establecer la longitud típica de los fragmentos de texto transcritos y, por lo tanto, ayuda a definir estrategias adecuadas de representación y clasificación textual.

A continuación, se presenta la distribución de las llamadas según su duración total:

Duración en segundos	**Cantidad de llamadas**
0 – 30	65
30 – 60	372
60 – 90	381
90 – 120	184
120 – 150	136
150 – 180	87
180 – 210	77
210 – 240	51
240 – 270	50
270 – 300	30
Mayor a 300 (5 min)	128

Tabla 3.2: Distribución de llamadas del Call Center según su duración

Como se puede observar, la mayoría de las llamadas tienen una duración entre 30 y 90 segundos. Esto implica que los modelos de clasificación deben estar preparados para analizar textos breves, los cuales presentan retos específicos como la escasez de contexto y la variabilidad lingüística en mensajes cortos.

3.4. Preparación de los Audios

3.4.1. Preprocesamiento de Audios

El preprocesamiento de los audios se realiza utilizando la librería Whisper de Python. Los pasos incluyen:

- Eliminación de silencios: Se eliminan los segmentos de silencio de los audios para mejorar la precisión de la transcripción.
- Filtrado de audios basura: Se descartan los audios de baja calidad o que no contienen información relevante.

Al hacer la respectiva categorización de los audios, se estudió le interdependencia lineal entre los mismos con respecto a la variable a predecir. Se recomienda evaluar si el problema es agnóstico de los atributos que se van a introducir, haciendo uso de una aproximación de aprendizaje no supervizado mediante el método de Análisis de Componentes Principales (PCA). Para dicho análisis se plantean las siguientes preguntas:

Preparación de los Audios

Inicialmente, se contaba con un conjunto de datos compuesto por seis clases de emociones distintas, con la siguiente distribución:

- **Interés:** 413 muestras
- **Amabilidad:** 376 muestras
- **Desmotivado:** 355 muestras
- **Duda:** 232 muestras
- **Decepción:** 107 muestras
- **Incomunicado:** 78 muestras

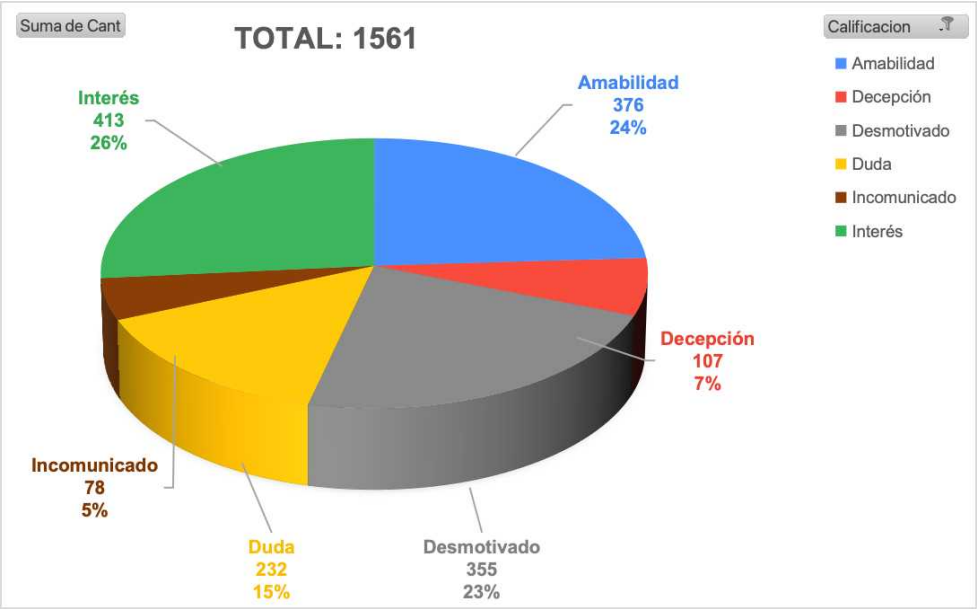

Figura 3.1: Gráfico circular: Primera clasificación
Fuente: Elaboración propia

Sin embargo, al analizar el contenido de estas clases, se evidenció un grado considerable de desbalanceo en algunas de ellas. Para abordar este problema y mejorar la representatividad de los datos en el entrenamiento del modelo, se realizaron los siguientes ajustes:

1. **Combinación de Decepción y Desmotivado:** Se identificó que ambas categorías representaban estados emocionales muy similares, por lo que se optó por fusionarlas bajo el nombre de **Desmotivado**.

2. **Eliminación de la clase Incomunicado:** Esta categoría no representaba una emoción en sí misma, sino que correspondía a llamadas en las que no se logró establecer comunicación. Debido a que estos registros no aportaban información útil para la clasificación de emociones, se decidió eliminarlos.

3. **Aumento de muestras para la clase Duda:** La clase **Duda** presentaba una cantidad de muestras significativamente menor en comparación con otras categorías, lo que podía afectar negativamente el rendimiento del modelo al clasificar este tipo de emociones. Para mitigar este desequilibrio, se aplicó una técnica de *oversampling*, incrementando las muestras de dicha clase. Es importante resaltar que este procedimiento se realizó **únicamente sobre el conjunto de entrenamiento**, luego de la partición de los datos, con el fin de evitar fuga de información (*data leakage*) y garantizar la validez de los resultados obtenidos en el conjunto de prueba.

Estos ajustes fueron fundamentales para garantizar un mejor balance en los datos y mejorar la capacidad del modelo para identificar patrones en todas las emociones representadas.

3.4.2. Transcripción de Audios a Texto

El proceso de transcripción de los audios a texto es fundamental para convertir las señales acústicas en datos textuales procesables para modelos de análisis emocional. Para esto, se exploraron distintas herramientas y configuraciones técnicas con el fin de seleccionar la más adecuada.

Primero, se evaluaron los formatos de audio. Los archivos fueron entregados inicialmente en formato .wav, con canales estéreo. Se realizaron pruebas de transcripción usando tanto el formato .wav como .mp3, y se observó que los resultados eran más precisos cuando los audios eran convertidos a formato .mp3. Adicionalmente, se convirtió la señal de estéreo a mono, lo que mejoró significativamente la claridad en la transcripción al evitar interferencias entre canales.

Posteriormente, se probaron dos herramientas de reconocimiento automático del habla (ASR):

- Google Speech-to-Text (STT), un servicio ampliamente utilizado por su accesibilidad y velocidad.
- Whisper, un modelo de código abierto desarrollado por OpenAI, el cual permite realizar transcripción localmente y está optimizado para múltiples lenguajes y ambientes ruidosos.

Para comparar el rendimiento de ambas herramientas, se seleccionó una muestra de 50 audios representativos del conjunto completo. Cada audio fue transcrito utilizando ambos sistemas. Las transcripciones fueron evaluadas manualmente con base en los siguientes criterios:

- Fidelidad semántica (preservación del mensaje original).
- Precisión fonética y gramatical.
- Cantidad de errores u omisiones en presencia de ruido o baja calidad.

Los resultados evidenciaron que Whisper superó consistentemente a Google STT en cada uno de los aspectos evaluados. Mientras que Google STT presentó omisiones y errores de interpretación, especialmente en audios ruidosos o con acentos regionales, Whisper mantuvo una mayor fidelidad y coherencia.

Adicionalmente, se probaron distintas versiones del modelo Whisper (tiny, base y medium). Se encontró que, si bien los modelos más grandes ofrecían mejores resultados en cuanto a precisión, también requerían mayor tiempo de procesamiento.

Como resultado de este análisis, se seleccionó Whisper como la herramienta principal de transcripción de audio para este proyecto, priorizando la calidad de los datos de entrada para las siguientes etapas de análisis.

3.5. Entendimiento de los datos textuales

3.5.1. Preprocesamiento de datos textuales

El preprocesamiento de las transcripciones es un paso esencial para garantizar que los datos sean adecuados para el análisis de emociones. Para obtener las transcripciones de los audios, se realizaron pruebas comparativas utilizando dos herramientas de reconocimiento de voz: Google Speech-to-Text y la librería Whisper, desarrollada por OpenAI.

Para tomar una decisión informada, se seleccionó una muestra representativa de audios con distintos niveles de calidad y complejidad. A continuación, se presenta un ejemplo de transcripción generada por cada herramienta para un mismo fragmento de audio:

y luego se viva Ya fue Ya fue Ya por eso es de la información Disculptame Sí pero yo no sé Es que lo de la vida Tengo más de dos semanas Y pues ahí se debe ver la búsqueda Si se le ha dado una página de la universidad Yo acabo de probar el programa y va Vale se parece bien si te comparte el información bien WhatsApp para que si si no es alguna duda no necesitas tu mis palabras notificas Vale sí Perfecto ya te compas entonces probas en que haré junto para poder te orientar en su momento en ningún inconveniente Te parece un señor Tapón Vale lo me siento Vale pero si Ya te siento les comparciendo espero tengas una excelente vida Y a mí gracias por acender mi llamada Claro que seguramente que se nos vendía Gracias igualmente hasta luego Hasta luego

Figura 3.2: Transcripción generada con Google Speech-to-Text
Fuente: Elaboración propia

Hola buenos días, es un amable comunicación, señora, mucho gusto, les habla Ingrid Anvila de la Universidad Javeriana de Cali, ¿cómo está? Espero que todo marche muy bien por su lado. Bien, muchas gracias. Programa de arquitectura, y en momento les vamos a presentar la información del programa. Ah, super, tengo que irme, lo que pasa es que yo en momento estoy como en realidad acá en un barco, ¿es posible que me llame más tarde o lo cuadramos el lunes? Disculpe. Ah, bueno, sí, señora, entonces agendamos una nueva llamada, no se preocupe. ¿O la llamo? Pues en momento la línea mantiene ocupada mi extensión es 94, yo le puedo agendar una nueva llamada, le puedo enviar por WhatsApp la información para que lo vayan mirando. Quedo atenta ¿vale? Bueno, señora. muchas gracias por su tiempo, que tenga buena tarde. Igualmente para usted, hasta luego.

Figura 3.3: Transcripción generada con Whisper
Fuente: Elaboración propia

Como puede observarse, Whisper produjo transcripciones más coherentes y completas, especialmente en aquellos audios con ruido de fondo, múltiples interlocutores o variaciones en la pronunciación. En cambio, Google Speech-to-Text omitió partes significativas del contenido o produjo frases con menor sentido semántico.

Con base en estos resultados, se decidió utilizar Whisper como herramienta principal de transcripción para todo el conjunto de datos.

Las transcripciones generadas fueron almacenadas en archivos CSV, cada uno de los cuales contiene la ubicación del audio original, su duración, la transcripción correspondiente y la clasificación de la emoción detectada.

Archivo	Transcripción	duratior	rms	zero_crossir	crest_facto	speech_rate	silence_ratio	Calificacion
/content/drive/My Drive/TESIS/Grabaciones-	Hola muy buenos días, ¿cómo va el día? Ah, hola, muy bien,	240.1	0.024339216	0.21858602	16.955467	2.09496043	0.3928368388	Desmotivado
/content/drive/My Drive/TESIS/Grabaciones-	Aló buenas, Si buenos días, tengo el gusto de hablar con la se	269.9	0.025398144	0.20574653	20.349537	2.25639125	0.41631206	Interés
/content/drive/My Drive/TESIS/Grabaciones-	¿Aló? Buen día, me comunico con Juan Manuel. Sí, señora. A	58.14	0.024456725	0.25080252	14.772935	2.32198142	0.32249527	Amabilidad
/content/drive/My Drive/TESIS/Grabaciones-	Hola, hola, hablo con Carlos Santiago. ¿Cómo estás? Estás co	114.28	0.026084483	0.21979135	14.997017	1.85509275	0.4486732149	Interés
/content/drive/My Drive/TESIS/Grabaciones-	Hola buenos días, Fabriana de Cali. Hola buenos días. Se hab	400.44	0.017255105	0.29331890	27.305267	1.48836280	0.4753577315	Interés
/content/drive/My Drive/TESIS/Grabaciones-	Comunica por favor con Mónica. Sí, con ella. Mucho gusto. C	174.9	0.01261008	0.24722360	37.995605	0.89193825	0.7126865352	Interés
/content/drive/My Drive/TESIS/Grabaciones-	¿Quieres hablar con Sofía López? Sí, señora. ¿Hablas con ella	112.22	0.023730222	0.20358460	17.178202	1.86241311	0.4341405275	Amabilidad

Figura 3.4: Fragmento de archivo de transcripciones
Fuente: Elaboración propia

Una vez obtenidas las transcripciones, se aplicaron las siguientes tareas de preprocesamiento textual para mejorar la calidad del análisis posterior:

- Eliminación de caracteres especiales y signos de puntuación innecesarios.
- Conversión de texto a minúsculas para estandarización.
- Eliminación de *palabras vacías* para reducir el ruido en el análisis.
- Lematización de palabras para normalizar las variaciones lingüísticas.

3.6. Limpieza y Normalización de Datos Textuales

3.6.1. Eliminación de Ruido

Después de la transcripción de los audios, el siguiente paso clave en el procesamiento de los datos fue la Limpieza y Normalización de Datos Textuales. Este proceso es esencial para garantizar que los textos transcritos sean consistentes y estén libres de ruido innecesario, lo cual facilita el análisis posterior. En primer lugar, se procedió a la eliminación de todos los caracteres no deseados, como puntuaciones innecesarias, números y otros símbolos que no aportan valor semántico al análisis. Para ello, se emplearon técnicas de procesamiento del lenguaje natural (NLP), que permitieron limpiar el texto de manera eficiente y precisa.

3.6.2. Normalización del Texto

Una vez realizada la limpieza, se llevó a cabo la normalización del texto. En primer lugar, se realizó la conversión de todas las palabras a minúsculas. Además, se utilizó la lematización. Finalmente, se procedió a la eliminación de palabras vacías, con el fin de reducir el volumen del texto y enfocarse en las palabras de mayor relevancia para el análisis de emociones.

3.7. Calidad de las Transcripciones

Dado que las transcripciones fueron generadas automáticamente por modelos de reconocimiento de voz, era necesario analizar su calidad antes de continuar con el proceso de etiquetado y modelado. Aunque Whisper demostró ser más preciso que otras herramientas evaluadas, como Google STT, aún existía la posibilidad de que ciertas transcripciones tuvieran errores, omisiones o ruido textual.

Para tener una primera aproximación a la calidad de las transcripciones, se analizó la longitud promedio del texto (en número de palabras) para cada emoción. Esto permitió identificar posibles diferencias estructurales entre los textos asociados a diferentes etiquetas, y detectar si había clases que sistemáticamente presentaban textos más extensos o más cortos.

Tabla 3.3: Longitud promedio de las transcripciones por emoción

Emoción	**Longitud Promedio (palabras)**
Amabilidad	18.2
Desmotivado	25.1
Duda	14.7
Interés	20.4

Como se observa en la tabla, las transcripciones etiquetadas como **Desmotivado** tienden a ser más extensas, lo cual sugiere que los usuarios que expresan esta emoción suelen argumentar o exponer con más detalle su malestar. En contraste, las transcripciones clasificadas como **Duda** presentan textos más breves, posiblemente por tratarse de consultas puntuales o interrupciones breves durante la llamada.

Este análisis también permitió detectar errores comunes en los textos, como palabras mal escritas, frases incompletas o repeticiones. Aunque estos errores no impidieron el entrenamiento del modelo,

se plantea como trabajo futuro el uso de técnicas de corrección ortográfica automática, detección de ruido textual, o incluso modelos de *text-cleaning* para mejorar la calidad del texto antes del modelado.

3.8. Etiquetado de Emociones

Cada transcripción del dataset fue asociada con una etiqueta emocional que representa la emoción predominante expresada por el usuario durante la llamada. Este proceso de etiquetado se llevó a cabo mediante un enfoque semiautomático, que combinó técnicas de procesamiento del lenguaje natural con revisión manual.

El proceso siguió los siguientes pasos:

1. Se aplicaron herramientas de análisis de sentimiento y detección de palabras clave emocionales para realizar una clasificación inicial automática.
2. Esta clasificación fue posteriormente revisada y corregida manualmente por el equipo, con base en un conjunto de criterios predefinidos para cada emoción.
3. En los casos en los que la emoción no era clara o no se podía determinar, se descartó la transcripción del análisis.

Las emociones consideradas inicialmente fueron: **Amabilidad**, **Decepción**, **Desmotivado**, **Duda**, **Incomunicado** e **Interés**.

3.9. Distribución y Balanceo de las Emociones

Tras el proceso de etiquetado, se observó un desequilibrio significativo en la cantidad de muestras por clase. Por ejemplo, la emoción **Interés** tenía más del doble de registros que categorías como **Decepción** o **Incomunicado**. Este desbalance podía afectar negativamente el rendimiento del modelo de clasificación, en especial para las clases minoritarias.

Para abordar este problema, se tomaron las siguientes medidas:

1. Se fusionaron las clases **Decepción** y **Desmotivado**, ya que mostraban patrones de lenguaje similares y se solapaban en la mayoría de los casos analizados.
2. Se eliminó la categoría **Incomunicado**, ya que correspondía a audios donde no se concretó una interacción verbal significativa.
3. Se aplicó *oversampling* a la clase **Duda** incrementando su representación hasta 300 muestras para mejorar el balance del dataset.

Tras aplicar estas estrategias, la nueva distribución de clases fue la siguiente:

- **Desmotivado:** 462 muestras.
- **Interés:** 413 muestras.
- **Amabilidad:** 376 muestras.
- **Duda:** 300 muestras.

Esta nueva estructura de datos permitió reducir el desbalance de clases y mejorar la capacidad del modelo para generalizar de manera equitativa entre todas las emociones.

- **Emoción predominante:** La categoría de Interés resultó ser la emoción más común en los audios, seguida por Amabilidad. Esto es consistente con la naturaleza de las interacciones en el call center, donde se resuelven problemas o se brindan respuestas positivas.
- **Menor representación:** Las emociones como decepción y desagrado están subrepresentadas en el dataset, lo que implica un desafío en términos de balance de clases para el modelo de clasificación.

Un gráfico de barras muestra la distribución de las emociones en el dataset.

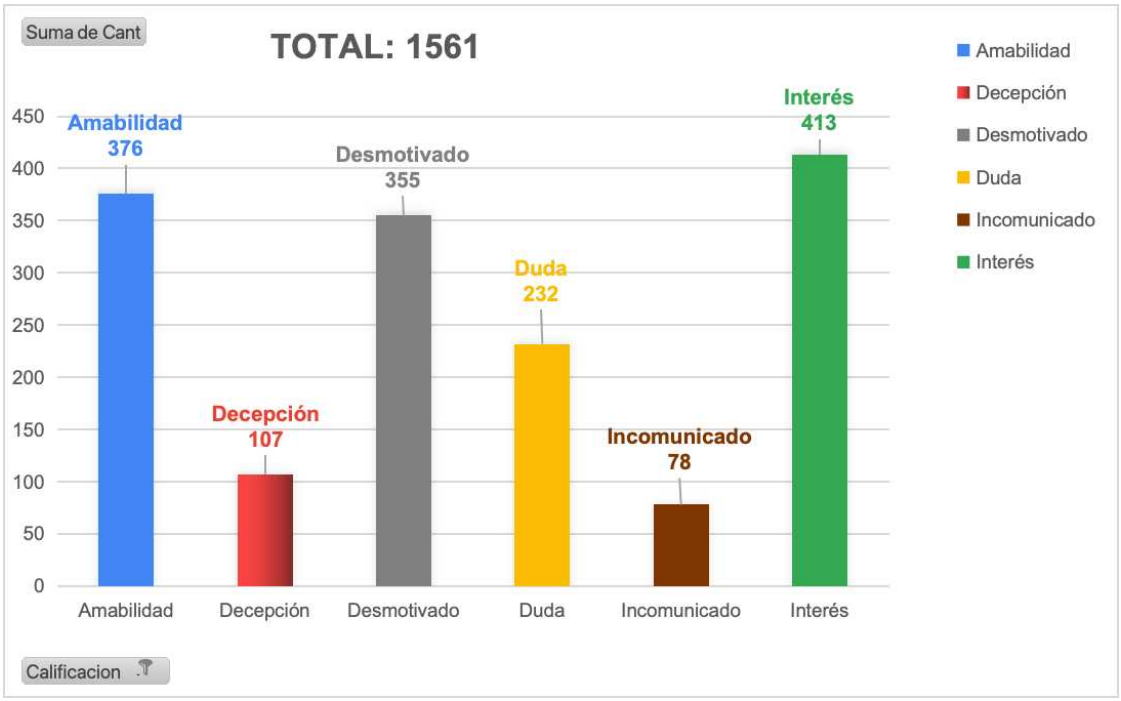

Figura 3.5: Gráfico de Barras Emociones dataset
Fuente: Elaboración propia

3.10. División del Dataset

El dataset utilizado para este estudio está compuesto por transcripciones de audios obtenidos del call center, los cuales han sido clasificados en distintas emociones. Para garantizar una evaluación justa de los modelos, se dividió el dataset en dos subconjuntos:

- **Conjunto de entrenamiento (80 %)**: Utilizado para entrenar los modelos de clasificación y ajustar sus hiperparámetros.
- **Conjunto de prueba (20 %)**: Destinado a la evaluación final de los modelos con datos no vistos.

Esta partición se realizó de manera estratificada para asegurar que la proporción de cada categoría de emoción se mantuviera constante en ambos conjuntos.

3.10.1. Vectorización de las Transcripciones

Antes de aplicar los modelos de clasificación, fue necesario transformar los textos en una representación numérica. Para ello se utilizó la técnica de vectorización **TF-IDF** (*Term Frequency - Inverse Document Frequency*), la cual asigna un peso a cada término según su frecuencia en una transcripción y su rareza en el conjunto completo de textos.

Previo a esta vectorización, los textos pasaron por un proceso de preprocesamiento que incluyó la normalización del texto, eliminación de signos de puntuación, conversión a minúsculas, eliminación de *stopwords* y lematización, como se detalló en secciones anteriores.

El resultado fue una matriz de características, donde cada fila representa una transcripción y cada columna un término del vocabulario. Esta matriz fue utilizada como entrada para los modelos de clasificación descritos en el siguiente capítulo.

El dataset original contenía un total de **1561 registros**, cada uno con las siguientes columnas:

- **Archivo:** Ruta o identificador del archivo de audio.
- **Transcripción:** Texto transcrito del audio.
- **duration:** Duración del audio (en segundos).
- **rms:** Nivel cuadrático medio de la señal.
- **zero crossing rate:** Tasa de cruce por cero del audio.
- **crest factor:** Relación entre el pico máximo y el valor RMS.
- **speech rate:** Tasa de palabras habladas por segundo.
- **silence ratio:** Proporción de silencio respecto al total del audio.
- **Calificación:** Etiqueta emocional asignada (variable objetivo).

Para el entrenamiento de los modelos, el conjunto de datos fue dividido en un **80 % para entrenamiento** y un **20 % para prueba**, preservando la distribución de clases mediante una partición estratificada. Específicamente:

- **Conjunto de entrenamiento:** 1248 muestras
- **Conjunto de prueba:** 313 muestras

Capítulo 4

Proceso de Modelado

En esta sección se presentan los resultados iniciales de los modelos de clasificación entrenados sobre el dataset, sin ajuste de hiperparámetros, con el objetivo de establecer una línea base de rendimiento. Posteriormente, se realiza un ajuste de hiperparámetros para mejorar la precisión y la generalización de los modelos. Finalmente, se presentan las comparaciones entre los modelos iniciales y los optimizados.

4.1. Selección de Modelos

Para llevar a cabo la clasificación de emociones, se probaron los siguientes clasificadores, seleccionados por su capacidad de manejar datos de texto y su aplicabilidad en problemas de clasificación multiclase:

- **Random Forest Classifier:** Un modelo basado en árboles de decisión que combina múltiples árboles para mejorar la generalización y reducir el sobreajuste.
- **Logistic Regression:** Un modelo lineal utilizado para la clasificación binaria que puede extenderse a problemas multiclase.
- **MLP (Multi-layer Perceptron):** Una red neuronal artificial de múltiples capas utilizada para tareas de clasificación complejas.

Cada uno de estos modelos fue entrenado utilizando representaciones vectoriales de los textos mediante la técnica *TF-IDF*, descrita en el capítulo anterior. Esta representación permite transformar las transcripciones en una matriz numérica adecuada para el entrenamiento de algoritmos de aprendizaje automático.

4.2. Diagrama del proceso de modelado

El diagrama ilustra cada una de las etapas realizadas en el desarrollo del modelo, desde la adquisición y preprocesamiento de los datos hasta la clasificación final de emociones.

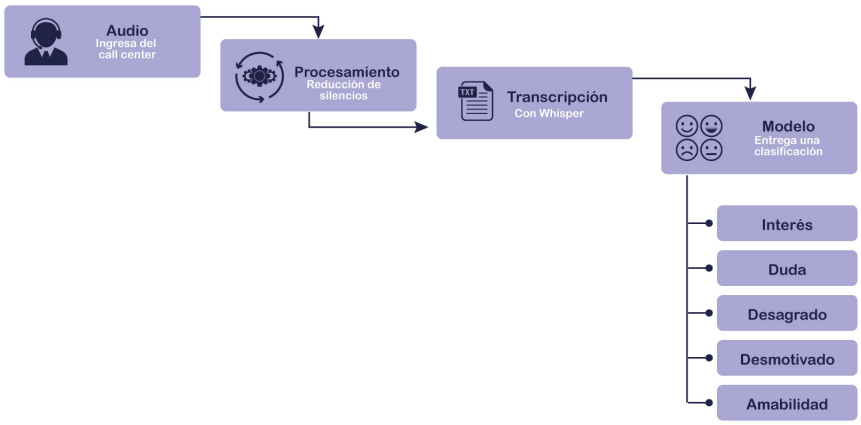

Figura 4.1: Diagrama flujo del proceso
Fuente: Elaboración propia

4.3. Resultados de los Modelos Iniciales

Los modelos fueron evaluados utilizando los siguientes métricas: exactitud, F1-score ponderado y recall. La Tabla 4.1 muestra los resultados obtenidos en la fase inicial del entrenamiento.

Los modelos fueron entrenados utilizando el 80 % del conjunto de datos y evaluados con el 20 % restante, reservado como conjunto de prueba. Las métricas de la Tabla 4.1 fueron calculadas exclusivamente sobre este conjunto de prueba.

Tabla 4.1: Resultados de los Modelos de Clasificación Iniciales

Modelo	**Exactitud**	**F1-score**	**Recall**
Random Forest	0.65	0.64	0.61
Logistic Regression	0.65	0.65	0.64
MLP	0.59	0.59	0.56

Estos resultados muestran que el modelo de Random Forest tuvo el mejor rendimiento inicial en comparación con los demás, indicando su capacidad para manejar la clasificación de emociones de manera efectiva.

4.4. Ajuste de Hiperparámetros

Para mejorar el rendimiento de los modelos, se realizó una búsqueda de hiperparámetros con validación cruzada. La Tabla 4.2 muestra los hiperparámetros explorados en cada modelo.

Tabla 4.2: Hiperparámetros Evaluados en Cada Modelo

Modelo	**Hiperparámetros Evaluados**
Random Forest	`n_estimators = [50, 100, 200], max_depth = [None, 10, 20, 30], min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4], criterion = [gini, entropy], bootstrap = [True, False]`
Logistic Regression	`penalty = ['l1', 'l2', 'elasticnet', 'none'], C = [0.01, 0.1, 1, 10], solver = ['liblinear', 'saga', 'lbfgs'], max_iter = [100, 500, 1000]`
MLP	`hidden_layer_sizes = [(100,), (50, 50), (100, 50), (100, 100)], activation = ['relu', 'tanh'], solver = ['adam'], alpha = [0.0001, 0.001, 0.01, 0.1], learning_rate = ['constant', 'adaptive'], max_iter = [200, 500, 1000]`

Tras la optimización, los mejores hiperparámetros encontrados fueron los siguientes:

Tabla 4.3: Mejores Hiperparámetros Encontrados

Modelo	**Mejores Hiperparámetros**
Random Forest	*n_estimators*=200, *max_depth*=20, *criterion*=gini
Logistic Regression	*C*=1, *max_iter*=100, *penalty*=l2
MLP	*hidden_layer_sizes*=(100,), *activation*=relu, *alpha*=0.1

4.5. Resultados de los Modelos Optimizados

Tras la optimización de los hiperparámetros, los modelos mostraron mejoras en sus desempeños:

Tabla 4.4: Resultados de los Modelos Optimizados

Modelo	**Exactitud**	**F1-score**	**Recall**
Random Forest	0.72	0.73	0.73
Logistic Regression	0.67	0.66	0.65
MLP	0.67	0.66	0.66

El modelo de Random Forest continuó mostrando el mejor desempeño tras la optimización de hiperparámetros, validando su efectividad para la clasificación de emociones en audios del call center. No obstante, el modelo de MLP también logró mejorar su rendimiento, acercándose a los resultados obtenidos por Random Forest.

4.6. Conclusión

Los resultados muestran que, tras el ajuste de hiperparámetros, el modelo Random Forest sigue siendo el más efectivo, con una exactitud del 65 % antes del ajuste y del 72 % después. Esto resalta la importancia de la optimización de hiperparámetros en el desempeño de los modelos. Logistic Regression y MLP también mostraron mejoras, aunque su desempeño no alcanzó el nivel de Random Forest. En futuras iteraciones, se podrían explorar técnicas adicionales como el uso de embeddings preentrenados para mejorar la clasificación de emociones en los audios del call center.

Capítulo 5

Análisis de Resultados

El análisis de los resultados obtenidos en la clasificación de emociones en audios de call center nos permite evaluar la eficacia del modelo, sus limitaciones y las implicaciones de su aplicación en entornos reales. A diferencia del análisis exploratorio de datos, donde se describe de manera objetiva la distribución y características del dataset, en este capítulo se presentan consideraciones críticas sobre los hallazgos obtenidos, discutiendo sus fortalezas, debilidades y el impacto que pueden tener en aplicaciones reales.

5.1. Limitaciones del Dataset y su Impacto en los Resultados

El dataset utilizado en este estudio fue construido a partir de transcripciones de audios de llamadas de un call center universitario, lo que introduce un primer obstáculo en la calidad del insumo de datos. Aunque se emplearon modelos de reconocimiento de voz avanzados, se identificaron errores semánticos y omisiones en las transcripciones. En particular, se evidenció que las palabras con pronunciación ambigua o poco clara tendieron a ser interpretadas de manera incorrecta, lo que inevitablemente afectó la precisión en la clasificación de emociones.

Otro aspecto crítico del dataset fue el desbalance de clases. En su versión original, la cantidad de registros para cada emoción variaba significativamente, con una sobre-representación de "Interés" y "Amabilidad" en comparación con "Decepción" e "Incomunicado". A pesar de aplicar técnicas de balanceo como *oversampling*, el problema de la representatividad de clases no fue completamente resuelto. La combinación de "Decepción" y "Desmotivado" ayudó a mitigar el problema, pero sigue existiendo un margen de mejora en la diversidad de ejemplos disponibles para entrenar el modelo.

5.2. Análisis del Desempeño de los Modelos

El modelo que obtuvo el mejor desempeño en la tarea de clasificación de emociones fue el **Random Forest**, alcanzando una exactitud del 72 % y un F1-score ponderado de 0.73. Aunque estos resultados son competitivos, revelan ciertos desafíos inherentes al análisis de emociones en textos provenientes de transcripciones automáticas.

Contrario a lo que se podría suponer, el modelo **Multi-Layer Perceptron (MLP)** no superó al

Random Forest, a pesar de su capacidad para capturar relaciones no lineales entre las variables. Esta diferencia de desempeño puede estar asociada a la naturaleza del dataset, compuesto por textos breves y ruidosos derivados del reconocimiento de voz, lo que limita la efectividad de modelos más complejos en contextos donde los datos no son del todo limpios o extensos.

Uno de los hallazgos más notables fue la dificultad de los modelos para distinguir entre las emociones **Desmotivado** y **Duda**. Esta confusión sugiere que las expresiones lingüísticas que caracterizan estas categorías presentan un alto grado de solapamiento, lo cual es un desafío común en tareas de análisis emocional del lenguaje natural, especialmente en interacciones telefónicas donde el tono, contexto o intención pueden no quedar explícitos solo con el texto.

Estos resultados evidencian la necesidad de contar con datos de mayor calidad, así como de explorar representaciones más robustas del lenguaje que permitan una mejor diferenciación entre emociones similares.

5.3. Factores que Influyeron en la Clasificación de Emociones

Uno de los factores clave en la clasificación de emociones fue la duración del audio. Como se observó en el análisis exploratorio, los audios más largos tendieron a reflejar emociones negativas como "Desmotivado" y "Duda", mientras que los audios más cortos se asociaron con emociones como "Amabilidad". Este hallazgo es fundamental, ya que indica que los clientes insatisfechos suelen expresar sus emociones de manera más extensa, lo que podría ser un indicador indirecto de la severidad de la queja o del nivel de frustración del usuario.

Otro factor determinante fue la calidad del lenguaje en las transcripciones. Se detectó que las frases con sarcasmo o ironía no fueron correctamente interpretadas por el modelo, lo que generó errores en la clasificación. Esto resalta la necesidad de incorporar modelos más avanzados de procesamiento del lenguaje natural (NLP) que sean capaces de detectar el contexto y la intencionalidad de las frases.

5.4. Repercusiones y Aplicabilidad de los Resultados

Los hallazgos de este estudio tienen implicaciones directas en la gestión de la experiencia del cliente en call centers. La posibilidad de detectar emociones en tiempo real permitiría tomar decisiones estratégicas, como redirigir llamadas problemáticas a agentes especializados o activar alertas cuando se detectan clientes insatisfechos. Sin embargo, los resultados también demuestran que una clasificación basada únicamente en texto tiene limitaciones. Para una detección más precisa, sería necesario complementar el análisis con características acústicas, como el tono de voz y la velocidad del habla.

En términos de aplicabilidad, la metodología utilizada en este trabajo podría extenderse a otros dominios, como el análisis de comentarios en redes sociales o la evaluación de encuestas de satisfacción. Esto sugiere que el impacto del modelo no está limitado a un contexto específico, sino que podría adaptarse a diversas aplicaciones en el ámbito empresarial y académico.

5.5. Aprendizajes Claves y Recomendaciones Futuras

A lo largo de este estudio, se han identificado varias lecciones clave:

- La importancia del balanceo de clases para evitar sesgos en la clasificación.
- La necesidad de mejorar la calidad de las transcripciones para reducir el impacto de errores en el reconocimiento de voz.
- La influencia de la duración del audio en la expresión de emociones.
- La dificultad inherente de clasificar emociones en texto sin considerar el contexto y la entonación.

Para futuras investigaciones, se recomienda:

1. Implementar modelos híbridos que combinen análisis de texto y audio.
2. Ampliar el dataset para incluir más ejemplos de emociones minoritarias.
3. Explorar el uso de modelos preentrenados en análisis de sentimientos, como BERT o GPT.
4. Incluir técnicas de postprocesamiento para mejorar la interpretación de las transcripciones automáticas.

Los resultados de este estudio representan un paso importante en la clasificación de emociones en interacciones de servicio al cliente, pero también destacan la necesidad de enfoques más sofisticados para abordar los desafíos inherentes al análisis del lenguaje natural en entornos reales.

5.6. Desafíos y Limitaciones

A lo largo del desarrollo del proyecto, se identificaron varios desafíos:

1. **Calidad de las transcripciones:** Muchas de las transcripciones generadas a partir de los audios presentaban ruido, errores de segmentación o frases incompletas, lo que dificultó la clasificación precisa de las emociones.
2. **Desbalance de clases:** Algunas emociones tenían muchas más muestras que otras, afectando la capacidad del modelo para aprender patrones en las clases minoritarias.
3. **Dificultad en la diferenciación de emociones similares:** Emociones como *Desmotivado* y *Decepción* mostraron resultados de clasificación ambiguos, lo que sugiere que en algunos casos la separación entre categorías no es clara.
4. **Limitaciones del dataset:** Con solo 1,561 muestras iniciales, la cantidad de datos pudo haber sido insuficiente para que modelos más complejos, como redes neuronales, alcanzaran su máximo rendimiento.

Capítulo 6

Conclusiones

Este trabajo se enfocó en la clasificación de emociones en audios de un call center mediante técnicas de procesamiento de lenguaje natural y modelos de aprendizaje supervisado. A partir del análisis y modelado de los datos, se identificaron los principales desafíos asociados al uso de transcripciones automáticas en entornos reales, así como las oportunidades de mejora en la calidad de los datos y en la selección de características relevantes para la clasificación.

6.1. Conclusiones sobre el Desempeño de los Modelos

Los resultados obtenidos con los distintos modelos de clasificación permiten concluir que el enfoque basado en modelos tradicionales de aprendizaje supervisado, especialmente Random Forest, es adecuado para la clasificación de emociones a partir de transcripciones de llamadas de un call center. La capacidad del modelo para manejar datos ruidosos y desbalanceados lo convierte en una opción robusta dentro de contextos reales donde la calidad del texto no es óptima.

Sin embargo, el desempeño limitado de modelos como MLP sugiere que la complejidad de los datos actuales aún no permite aprovechar al máximo modelos más sofisticados. Esto evidencia que la calidad y el volumen de los datos de entrenamiento juegan un papel crucial, y que la mejora en la calidad de las transcripciones, así como el enriquecimiento del dataset, son aspectos clave para futuras investigaciones.

Además, la diferencia en el rendimiento entre modelos también refuerza la importancia de realizar un análisis cuidadoso de preprocesamiento, balanceo y selección de características, lo cual fue determinante para alcanzar resultados aceptables en este estudio.

6.2. Conclusiones Finales

- **Se alcanzó el objetivo general del proyecto**, al desarrollar y evaluar modelos de aprendizaje supervisado capaces de clasificar emociones en transcripciones de llamadas del call center. El sistema propuesto demostró un rendimiento satisfactorio, validando la viabilidad técnica del enfoque planteado.

- **El modelo de Random Forest se destacó como el más adecuado** para esta tarea, al ofrecer un buen equilibrio entre exactitud, capacidad de generalización y facilidad de interpretación. Su desempeño superó al de otros clasificadores como Logistic Regression y MLP.
- **La calidad de las transcripciones es un factor determinante** en el rendimiento de los modelos. La presencia de errores y ruido en los textos generados automáticamente afecta la capacidad de los clasificadores para identificar patrones emocionales. Mejorar esta etapa, por ejemplo mediante modelos más precisos o técnicas de postprocesamiento, es fundamental.
- **El desbalance entre clases emocionales sigue representando un desafío importante**. A pesar de aplicar estrategias como la fusión de clases y el oversampling, la clasificación de emociones minoritarias continúa siendo menos precisa. Futuros trabajos podrían explorar técnicas de generación de datos sintéticos o algoritmos robustos al desbalance.
- **La clasificación de emociones en contextos reales, como los call centers, es viable**, pero exige una combinación cuidadosa de limpieza de datos, selección de características relevantes y ajuste de modelos. Este trabajo sienta una base sólida para estudios posteriores que busquen mejorar la interacción entre usuarios y agentes mediante análisis emocional automatizado.

Capítulo 7

Trabajos Futuros

A partir de los hallazgos y limitaciones encontradas en el presente estudio, se identifican diversas líneas de trabajo futuro que podrían mejorar la clasificación de emociones en audios de call center. Entre las principales oportunidades de mejora se destacan:

- **Ampliación del dataset:** Uno de los principales factores que afectan la capacidad de generalización del modelo es la cantidad limitada de datos disponibles. Un dataset más grande, con una mayor representación de todas las emociones, permitiría mejorar el aprendizaje del modelo y reducir el sesgo hacia las clases mayoritarias. Además, obtener datos de distintas fuentes de call center ayudaría a evaluar la robustez del modelo en distintos escenarios de interacción con clientes.
- **Incorporación de características acústicas:** Hasta ahora, la clasificación se ha basado exclusivamente en el texto transcrito de los audios, lo que implica que la entonación, velocidad del habla y otras características del audio no han sido aprovechadas. Incorporar atributos acústicos como tono, pausas, intensidad y ritmo de habla podría mejorar significativamente la identificación de emociones, ya que muchas emociones se reflejan tanto en la voz como en el contenido verbal.
- **Exploración de representaciones avanzadas del texto:** Aunque en este estudio se utilizó TF-IDF para la vectorización del texto, sería recomendable probar técnicas más avanzadas como word embeddings (*Word2Vec, GloVe*) o modelos de lenguaje preentrenados como *BERT, RoBERTa o DistilBERT*. Estas representaciones permiten capturar mejor el significado contextual del texto, lo cual podría mejorar la precisión de los modelos en la clasificación de emociones.
- **Uso de arquitecturas de Deep Learning:** Los modelos de aprendizaje profundo, como las redes neuronales recurrentes (*RNN, LSTM, GRU*) y modelos basados en *Transformers*, han demostrado un alto rendimiento en tareas de análisis de lenguaje natural. Implementar estos enfoques podría permitir al modelo aprender mejor las relaciones semánticas en las transcripciones y mejorar su capacidad para detectar emociones en interacciones de call center.
- **Exploración de enfoques semi-supervisados y no supervisados:** La anotación manual de emociones es un proceso costoso y subjetivo. Métodos como el aprendizaje semi-supervisado

podrían permitir entrenar modelos con una menor cantidad de etiquetas, mientras que los enfoques no supervisados (como clustering y modelos de aprendizaje auto-supervisado) podrían ser útiles para descubrir patrones ocultos en los datos sin requerir etiquetas explícitas.

- **Evaluación del impacto de la calidad de las transcripciones:** Dado que el modelo depende completamente de la calidad del texto generado a partir del audio, sería relevante analizar hasta qué punto los errores en la transcripción afectan el rendimiento del modelo. Se podrían evaluar diferentes herramientas de transcripción automática y comparar su impacto en la clasificación de emociones.
- **Clasificación de emociones en tiempo real:** Una posible extensión del trabajo es la implementación de un sistema que pueda identificar emociones en tiempo real, lo cual sería útil para el monitoreo automático de interacciones en call centers. Esto requeriría optimizar el tiempo de inferencia del modelo y evaluar estrategias para procesar el audio de manera continua.
- **Implementación de un sistema de apoyo a la toma de decisiones:** En un contexto real de atención al cliente, un modelo de clasificación de emociones podría integrarse en sistemas de soporte que alerten a los agentes de call center cuando se detecten emociones negativas en las llamadas, permitiendo mejorar la experiencia del usuario en tiempo real.

Estas líneas de investigación futura permitirían mejorar la precisión y aplicabilidad de los modelos en entornos de atención al cliente, incrementando su utilidad en la automatización del análisis de emociones en interacciones telefónicas.

Bibliografía

[1] H. Chaviano Arteaga, "Técnicas de aprendizaje supervisado y no supervisado para el aprendizaje automatizado de computadoras," *Memorias del Primer Congreso Internacional de Ciencias Pedagógicas*, 2010. ISBN 978-9942-17-011-8, págs. 549–564.

[2] M. Acevedo and K., "Machine learning: Algoritmos de clasificación," *Estado de México, Naucalpan*, p. 133, 2017.

[3] SAS, "Machine learning: What it is and why it matters," 2018. Recurso en línea.

[4] E. A. Cingolani, "Evaluación de sistemas recomendadores de contenidos educativos a través de estudios de usuarios," *s. p.*, pp. 1–75, 2014.

[5] A. Gelbukh, "Procesamiento de lenguaje natural y sus aplicaciones," *Vol. I, p. 6*, 2010.

[6] R. B. Pittala, B. R. Tejopriya, and E. Pala, "Study of speech recognition using cnn," *ICAIS 2022*, 2022.

[7] J. R. Zapata, "Extracción de características - stft (transformada corta de fourier)," *Curso de Minería de Audio*. Recurso en línea.

[8] C. E. Izard, "Emotion theory and research: Highlights, unanswered questions, and emerging issues," *Annual Review of Psychology*, vol. 60, pp. 1–25, 2009.

[9] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3/4, 1992.

[10] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *University of Nebraska Press, Lincoln*, 1972.

[11] F. Santiago, "Regresión logística," 2011.

[12] M. Lizares, "Comparación de modelos de clasificación: Regresión logística y árboles de clasificación para evaluar el rendimiento académico," *Universidad Nacional Mayor de San Marcos*, 2017.

[13] L. Auria and R. A. Moro, "Support vector machines (svm) as a technique for solvency analysis," 2008.

[14] J. Orellana Alvear, "Árboles de decisión y random forest," 2018.

[15] IBM, "Redes neuronales: Perceptrón multicapa (ibm spss statistics)," 2024.

[16] IBM, "Redes neuronales recurrentes," 2024.

[17] IBM, "Redes neuronales convolucionales," 2024.

[18] X. Font, "Técnicas de clasificación (supervised learning)," 2019.

[19] I. Herrera and A. Figueroa, "Aprendizaje semi-supervisado de múltiples vistas para detectar temporalidad de preguntas," 2024.

[20] IBM, "¿qué es el etiquetado de datos?," 2024.

[21] A. Cortez M, "Procesamiento de lenguaje natural," *Universidad Nacional Mayor de San Marcos*, 2014.

[22] H. Pérez Espinosa, "Reconocimiento de emociones a partir de voz basado en un modelo emocional continuo," *Tesis de Doctorado, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla*, 2013.

[23] D. Sánchez Angón, "Reconocimiento de emociones a partir de imagen y voz," *Tesis de Licenciatura*, 2017.

[24] V. B. Ambario, M. M. Arroyo, J. A. M. Valverde, and J. M. H. Bravo, "Reconocimiento de emociones a través del análisis de la voz," *Memorias del Congreso Internacional de Investigación Académica Journals Celaya 2017*, 2017.

[25] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[26] L. F. Correa Pinto, "Reconocimiento automático de emociones en audio y video usando machine learning," *Proyecto Fin de Carrera, Universidad de los Andes, Facultad de Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica*, 2019.

[27] M. Pervaiz and T. Ahmed, "Emotion recognition from speech using prosodic and linguistic features," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, 2016.

[28] J. Ierache and F. Elkfury, "Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educación emocional," *XXIII Workshop de Investigadores en Ciencias de la Computación*, p. 623, 2021.

[29] V. E. Hernandez L., "Emociones en señales de voz: Reconocimiento con redes neuronales profundas," *Universitat Politècnica de Catalunya, Barcelona*, 2021.

[30] M. P. G. and A. Berlanga, "Análisis de sentimiento en audio mediante inteligencia artificial orientado al idioma español," *Tesis de grado, Administración de Empresas e Ingeniería de Sistemas, Universidad Carlos III de Madrid, España*, 2022.

[31] E. Bolo, M. Samoul, N. Seichepine, and M. Chetouani, "Quietly angry, loudly happy: Self-reported customer satisfaction vs. automatically detected emotion in contact center calls," *Interaction Studies*, vol. 24, no. 1, pp. 168–192, 2023.

[32] Audeering, "Devaice," 2022.

[33] MixedEmotions, "Mixedemotions project," 2023.