ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

MODELOS SUPERVISADOS ASISTIDOS POR CIRCUITOS CUÁNTICOS VARIACIONALES PARA LA CARACTERIZACIÓN CONDUCTUAL DE INTERACCIONES ORALES EN CENTROS DE CONTACTO

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Alfredo Joaquín Zambrano Dávila

Abel Alexander Balladares Celleri

GUAYAQUIL - ECUADOR

Año: 2025

DEDICATORIA

El presente proyecto lo dedico a mi hija Giuliana Zambrano, porque cada paso en este camino lo recorrí pensando en ti. Tu existencia le da sentido a mis sueños. Esta tesis es también tuya.

A mi familia, por su apoyo incondicional, por la paciencia y por creer en mis metas incluso antes que yo mismo.

A quienes, con su palabra y guía contribuyeron silenciosamente a que este proyecto sea posible.

Alfredo Zambrano

Dedico este trabajo de maestría a todas las personas que han contribuido de manera significativa a mi formación profesional y personal. En especial, a mi madre, mis abuelos y mis tres hermanos. Adicional al Ing. Andrey Arias e Ing. Galo Narváez cuyo apoyo, inspiración y confianza fueron fundamentales para culminar este proceso académico. A cada uno de ustedes, mi gratitud y reconocimiento.

Abel Balladares

AGRADECIMIENTOS

Agradezco a Dios, por la fortaleza, la guía y la claridad que me permitieron culminar este trabajo.
A mi familia, por su apoyo constante, su paciencia y su confianza en mí.
A mis profesores y a mi tutor, por su dedicación, acompañamiento y valiosos aportes en este proceso académico.

Alfredo Zambrano

Agradezco profundamente a Dios por la sabiduría, fortaleza y oportunidad de transitar esta etapa de crecimiento académico y personal. Su guía me ha permitido llegar, junto a mi familia, a este momento significativo de mi vida profesional. Este logro representa un peldaño más en el camino que aún queda por recorrer.

Abel Balladares

DECLARACIÓN EXPRESA

Nosotros *Alfredo Joaquín Zambrano Dávila y Abel Alexander Balladares Celleri* acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 21 de enero del 2026

Firmado electrónicamente por:
ALFREDO JOAQUIN ZAMBRANO DAVILA
Validar únicamente con FirmaEC

Alfredo Joaquín Zambrano Dávila

Firmado electrónicamente por:
ABEL ALEXANDER BALLADARES CELLERI
Validar únicamente con FirmaEC

Abel Alexander Balladares Celleri

EVALUADORES

Maria Isabel Mera Collantes
Time Stamping Security Data

PhD. María Isabel Mera Collantes

PROFESOR EVALUADOR

Firmado electrónicamente por:
SERGIO ALEX BAUZ OLVERA
Validar únicamente con FirmaEC

PhD. Sergio Alex Bauz Olvera

TUTOR DE PROYECTO

RESUMEN

Este proyecto propone un enfoque innovador para la caracterización conductual de interacciones orales en centros de contacto, mediante modelos supervisados asistidos por circuitos cuánticos variacionales. El objetivo es clasificar atributos como las emociones, veracidad, el nivel de estrés y la identidad del hablante, a partir de señales acústicas. Se plantea como hipótesis que el uso de modelos híbridos cuántico-clásicos permite mejorar la precisión de clasificación en comparación con modelos tradicionales. La justificación radica en la necesidad de sistemas más eficientes para el análisis emocional y conductual en lenguaje oral.

Se desarrolló un prototipo experimental que emplea un dataset simulados de conversaciones en call centers, procesados mediante técnicas fonéticas como MFCC. Se aplicaron métodos de reducción de dimensionalidad (PCA) y codificación de datos clásicos en circuitos cuánticos parametrizables. La arquitectura se basa en entrenamiento supervisado de modelos VQC utilizando optimización por desplazamiento de parámetros. Se emplearon bibliotecas como *PennyLane* y *openSMILE* en el procesamiento y entrenamiento.

Como resultado preliminar, se observó que el modelo cuántico obtuvo una precisión comparable a los modelos clásicos como SVM y Random Forest en tareas de clasificación multiclase. Estos resultados son alentadores, ya que colocan a los circuitos cuánticos variacionales como una alternativa importante a la hora de abordar problemas complejos relacionados con el tratamiento del habla.

Se concluye que la computación cuántica aplicada al aprendizaje automático ofrece ventajas prometedoras en contextos de análisis vocal, y representa una línea de investigación con alto potencial para el futuro del procesamiento del lenguaje oral.

Palabras Clave: aprendizaje supervisado, circuitos cuánticos variacionales, análisis del habla, centros de contacto, emociones, inteligencia artificial cuántica.

ABSTRACT

This project proposes an innovative approach for the behavioral characterization of oral interactions in contact centers, through supervised models assisted by variational quantum circuits. The objective is to classify attributes such as emotions, veracity, stress level, and speaker identity based on acoustic signals. The hypothesis posits that the use of hybrid quantum-classical models improves classification accuracy compared to traditional models. The justification lies in the need for more efficient systems for emotional and behavioral analysis in spoken language.

An experimental prototype was developed using a simulated dataset of contact center conversations, processed through phonetic techniques such as MFCC. Dimensionality reduction methods (PCA) were applied, along with encoding classical data into parameterized quantum circuits. The architecture is based on supervised training of VQC models using parameter-shift optimization. Libraries such as PennyLane and openSMILE were used for signal processing and model training.

As a preliminary result, the quantum model achieved accuracy comparable to or higher than classical models such as SVM and Random Forest in multiclass classification tasks. These results are encouraging, as they position variational quantum circuits as an important alternative for solving complex problems related to speech processing.

It is concluded that quantum computing applied to machine learning offers promising advantages in voice analysis contexts and represents a high-potential research line for the future of spoken language processing.

***Keywords**: supervised learning, variational quantum circuits, speech analysis, contact centers, emotions, quantum artificial intelligence.*

ÍNDICE GENERAL

ABREVIATURAS

AI	Artificial Intelligence (Inteligencia Artificial)
ML	Machine Learning (Aprendizaje Automático)
VQC	Variational Quantum Circuit (Circuito Cuántico Variacional)
MFCC	Mel-Frequency Cepstral Coefficients
PCA	Principal Component Analysis (Análisis de Componentes Principales)
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
API	Application Programming Interface
LLD	Low-Level Descriptor (Descriptor de Bajo Nivel)
RNN	Recurrent Neural Network (Red Neuronal Recurrente)
CNN	Convolutional Neural Network (Red Neuronal Convolucional)
SVM	Support Vector Machine (Máquina de Vectores de Soporte)
TPR	True Positive Rate (Tasa de Verdaderos Positivos)
FPR	False Positive Rate (Tasa de Falsos Positivos)
NLP	Natural Language Processing (Procesamiento de Lenguaje Natural)
VQE	Variational Quantum Eigensolver
QAQA	Quantum Approximate Optimization Algorithm
SGD	Stochastic Gradient Descent (Descenso de Gradiente Estocástico)
TPU	Tensor Processing Unit (Unidad Central de Tensores)
GPU	Graphics Processing Unit (Unidad de Procesamiento Gráfico)
p.p.	Puntos porcentuales

SIMBOLOGÍA

θ	Parámetro ajustable en un modelo cuántico o red neuronal
α	Tasa de aprendizaje (*learning rate*)
∇	Gradiente de una función
Σ	Sumatoria
Ψ	Estado cuántico
Z	Base de medición del observador (eje Z en qubits)
ℓ	Longitud del vector de características
n	Tamaño del dataset o número de muestras
t	Tiempo o instante temporal en señales
f(x)	Función objetivo del modelo
∂	Derivada parcial
$J(\theta)$	Función de costo en aprendizaje automático
Z	Base computacional (medición en base Z)
$\|0\rangle$	Qubit en estado base 0 (ket 0)
$\|1\rangle$	Qubit en estado base 1 (ket 1)
$L(\hat{y},y)$	Función de pérdida de una muestra en aprendizaje automático.
$R_y(\theta)$	Rotación alrededor del eje y con ángulo θ
$\otimes$	Producto tensorial

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

En los últimos años, la transformación digital ha impulsado una creciente automatización de procesos en sectores como el servicio al cliente, especialmente en entornos de centros de contacto (*Call Centers*). Estas plataformas se han convertido en puntos clave para la interacción entre empresas y usuarios, generando grandes volúmenes de datos orales que contienen información valiosa sobre emociones, niveles de estrés, intenciones y veracidad en la comunicación. Sin embargo, el análisis manual de estas interacciones es costoso, lento y subjetivo, y las soluciones automatizadas tradicionales, basadas en análisis textual, no siempre capturan la riqueza contextual y paralingüística de la voz humana.

Paralelamente, el crecimiento de la computación cuántica ha abierto nuevas puertas para la construcción de modelos de aprendizaje más eficientes y expresivos. Particularmente, los Circuitos Cuánticos Variacionales (VQC) se han propuesto como una alternativa prometedora para tareas de clasificación en dominios de alta dimensionalidad. Estos modelos híbridos, que combinan elementos clásicos con operaciones cuánticas, permiten representar funciones altamente no lineales mediante codificación de datos en estados cuánticos y optimización de parámetros mediante métodos como parameter-shift.

Esta tesis propone una arquitectura experimental que integra análisis fonético, extracción de características vocales y textuales, reducción de dimensionalidad, codificación en circuitos cuánticos y entrenamiento supervisado para clasificar múltiples atributos conductuales en conversaciones orales. El objetivo es demostrar la aplicabilidad de modelos VQC en un entorno realista como el de un centro de contacto, y comparar su rendimiento frente a modelos tradicionales como Random Forest, SVM y redes neuronales clásicas.

La motivación principal radica en el potencial disruptivo de la computación cuántica para resolver problemas complejos de inteligencia artificial, y en la necesidad de explorar su aplicabilidad en dominios donde la variabilidad humana es especialmente alta, como lo es el lenguaje hablado.

CAPÍTULO 1

1 PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del Problema

La interacción humana ha sido objeto de interés en diversas disciplinas como la psicología, la neurociencia y el análisis conductual, dado que constituye uno de los mecanismos esenciales de comunicación y cohesión social (Gumá, 2001). La voz, en particular, es un vehículo fundamental para transmitir no solo información lingüística, sino también emociones, estados cognitivos e intenciones (Kalloniatis & Kontopoulos, 2024). Desde una perspectiva neurocientífica, se ha demostrado que distintas áreas del cerebro se activan frente a características prosódicas específicas, reflejando la carga emocional y el estado interno del hablante (Pell M. D., 2009).

El análisis de la voz humana ha evolucionado desde enfoques tradicionales hacia técnicas avanzadas de procesamiento de señales, permitiendo extraer patrones fonéticos y emocionales que anteriormente pasaban desapercibidos (Eyben, Wöllmer, & Schuller, 2010). La caracterización conductual basada en señales vocales se ha convertido en un campo emergente con aplicaciones en salud mental, seguridad y servicios de atención al cliente (Cowie, y otros, 2000).

En el ámbito de los centros de contacto, las interacciones orales entre agentes y clientes son cruciales para la percepción de calidad del servicio (Kamm, 1997). Sin embargo, el análisis de estas interacciones sigue realizándose mayormente de forma manual, lo cual implica altos costos operativos, sesgo subjetivo, lentitud en la retroalimentación y limitaciones en la escalabilidad (Hernández, 2021). Estudios recientes indican que los auditores humanos pueden tardar entre 5 y 10 minutos en analizar una llamada de 2 minutos, lo cual compromete seriamente la eficiencia del servicio.

La ausencia de análisis automatizado en los centros de contacto no solo afecta la productividad, sino también la posibilidad de detectar tempranamente señales de insatisfacción, fraude o estrés emocional. Esta limitación resulta crítica considerando que, según la literatura, las emociones negativas como el enojo o la tristeza pueden escalar rápidamente si no son atendidas oportunamente.

Desde una perspectiva conductual, se reconoce que los patrones vocales varían en función de estados emocionales y cognitivos. Por ejemplo, un aumento en la frecuencia fundamental (pitch) y una aceleración del ritmo de habla pueden asociarse con estados de ansiedad o mentira (Siegman, 1993). A nivel psicológico, factores como el estrés y la desconfianza se reflejan en cambios prosódicos y en la fluidez del discurso.

En paralelo, la caracterización biométrica del hablante basada en atributos fonéticos como MFCC (*Mel-Frequency Cepstral Coefficients*) ha demostrado alta efectividad para la identificación de individuos a partir de sus patrones vocales únicos. Esta técnica, combinada con aprendizaje automático supervisado, permite construir firmas vocales robustas que pueden ser utilizadas en entornos como centros de contacto para reconocer clientes recurrentes o detectar suplantaciones.

El software DINOMI (PSD, 2025), utilizado ampliamente en gestión de llamadas entrantes y salientes, carece actualmente de módulos avanzados para la caracterización conductual automática. Su enfoque se limita a la gestión operativa, sin incorporar análisis emocional, veracidad discursiva, detección de estrés ni reconocimiento de hablantes.

Dada esta brecha tecnológica, se vuelve imperativo el desarrollo de módulos inteligentes que integren técnicas de análisis fonético, procesamiento del lenguaje natural y modelos de aprendizaje supervisado para realizar la caracterización conductual de las interacciones orales. Dichos módulos permitirán no solo mejorar la eficiencia operativa, reduciendo el tiempo y el costo de análisis, sino también enriquecer la experiencia del cliente mediante respuestas más rápidas, precisas y adaptadas a su perfil conductual.

En el contexto de avances recientes en computación cuántica, se ha identificado una oportunidad disruptiva para utilizar circuitos cuánticos variacionales (VQC) como clasificadores en tareas de aprendizaje supervisado. Este enfoque permite construir modelos híbridos que codifican datos clásicos (por ejemplo, características acústicas como MFCC) en estados cuánticos y generan predicciones usando circuitos cuánticos variacionales (Schuld M. , Bocharov, Svore, & Wiebe, 2020). Esto permite reducir tiempos de cálculo y mejorar la búsqueda de mínimos globales en funciones de pérdida complejas, lo cual es especialmente relevante en contextos de señales de voz ruidosas y de alta variabilidad como las interacciones orales.

En consecuencia, abordar este problema contribuye al avance académico en las áreas de análisis del habla, inteligencia artificial emocional y biometría vocal (Cummins, y otros, 2015); (Deepa & Kuppusamy, 2022).

1.2 Justificación

La incesante necesidad de mejorar los indicadores de desempeño en los centros de contacto ha motivado la adopción de tecnologías avanzadas orientadas al análisis automático de la comunicación verbal. En este ámbito, el estudio de patrones fonéticos y prosódicos permite obtener información conductual valiosa, contribuyendo a la optimización de los recursos operativos y al fortalecimiento de la experiencia del usuario. Por otro lado, la neurociencia computacional, reconoce a la voz humana como transmisora de señales emocionales y conductuales que pueden ser captadas mediante patrones acústicos específicos, tales como variaciones en el tono, la velocidad y la intensidad del habla (Adolphs, 2002). Estas señales, si son procesadas adecuadamente, permiten inferir estados afectivos, niveles de estrés, veracidad discursiva e incluso la identidad del hablante.

Implementar tecnologías de análisis de voz basadas en aprendizaje automático en centros de contacto posibilita superar las limitaciones del análisis humano manual, tradicionalmente caracterizado por su lentitud, alto costo y subjetividad. Además, permite anticipar situaciones de riesgo, como conflictos potenciales, disminuyendo las tasas de abandono de clientes y aumentando los índices de satisfacción.

En términos de eficiencia, la automatización del análisis conductual podría reducir hasta en un 70% el tiempo destinado a auditorías manuales de llamadas, liberando recursos humanos para tareas de mayor valor estratégico. Asimismo, la capacidad de identificar a los hablantes mediante firmas biométricas vocales promete ser un mecanismo de seguridad y personalización del flujo de atención a los clientes (Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018).

Desde una perspectiva tecnológica, el desarrollo de un módulo que integre la detección de emociones, veracidad e identificación del hablante constituye un avance significativo en la convergencia de las ciencias del comportamiento y la inteligencia artificial. Este tipo de soluciones representa el futuro del análisis conversacional inteligente, en el que no solo se transcribe lo que se dice, sino que se interpreta el cómo y por qué se dice.

Adicionalmente, el impacto académico del proyecto es considerable, al contribuir al fortalecimiento del aprendizaje automático aplicado a las emociones, el procesamiento de señales acústicas y la biometría conductual. Investigaciones recientes han destacado

la necesidad de enfoques interdisciplinarios que combinen conocimiento de la conducta humana con técnicas avanzadas de aprendizaje supervisado.

Desde una perspectiva social, la mejora de la capacidad de los centros de contacto para entender a sus clientes permite anticiparse a situaciones críticas o de conflicto mediante la detección de indicadores tempranos de malestar emocional o fraude, lo cual tiene implicaciones positivas en ámbitos como la seguridad, la salud mental y la protección de consumidores.

La incorporación de los circuitos cuánticos variacionales permite construir clasificadores que aprovechan la capacidad de representación de estados cuánticos para aprender patrones complejos incluso con pocos *qubits*, y son especialmente útiles cuando se trabaja con datos acústicos y características fonéticas como las extraídas de interacciones orales en centros de contacto.

Finalmente, la aplicación práctica del módulo desarrollado no se limita exclusivamente al contexto de los centros de contacto gestionados mediante DINOMI. Su arquitectura modular y escalable abre la posibilidad de ser adaptado a otros sectores intensivos en interacción oral, como la salud, el sector bancario y la educación.

1.3 Objetivos (General y Específico)

1.3.1 Objetivo General

Desarrollar modelos de aprendizaje supervisado basados en circuitos cuánticos variacionales que permitan la caracterización conductual de interacciones orales en centros de contacto gestionados mediante software especializados.

1.3.2 Objetivos Específicos

- Examinar patrones fonéticos relevantes con el propósito de identificar firmas biométricas de voz a partir de registros de audio provenientes de centros de contacto.
- Implementar modelos de aprendizaje supervisado cuánticos para la detección de sentimiento (positivo, negativo y neutro), veracidad y estimación del nivel de estrés basados en patrones fonéticos de las interacciones orales.

- Comparar el desempeño de modelos clásicos y cuánticos en tareas de clasificación supervisada de señales de voz, empleando métricas de evaluación estándares del aprendizaje automático (Saito & Rehmsmeier, 2015).
- Implementar una solución de monitoreo visual que permita el seguimiento dinámico del comportamiento de las interacciones orales orientada al análisis operativo y a la toma de decisiones estratégicas.

1.4 Marco Teórico

El análisis de interacciones orales provenientes de software centros de contacto como DINOMI constituye una fuente rica de información sobre el comportamiento, las emociones y las necesidades de los usuarios. La voz humana transmite información acústica que permite determinar no solo el contenido semántico, sino también emociones, niveles de estrés y rasgos de identidad. Características como la prosodia, la entonación, las pausas y la velocidad del habla han sido ampliamente estudiados como indicadores confiables del estado psicológico del hablante (Scherer, 2003; Burkhardt et al., 2005).

En la práctica diaria, todavía muchos supervisores revisan manualmente numerosas grabaciones durante largas horas, evaluándolas con base en sus propias impresiones personales relacionadas con el tono, actitud del hablante, uso de ciertas palabras o expresiones emocionales, con el propósito determinar el comportamiento del llamante o cliente. Si bien esta estrategia permite detectar matices humanos difíciles de automatizar, presenta grandes limitaciones: es lenta, costosa, difícil de estandarizar, propensa a sesgos cognitivos y variabilidad entre evaluadores, y no escala bien en contextos de alto volumen de datos (Soleymani, y otros, 2017).

En la actualidad, el análisis automático de estas señales se realiza principalmente a través de modelos de aprendizaje supervisado clásico, utilizando algoritmos basados en técnicas de ensamble, de margen máximo, de reglas de partición y redes neuronales artificiales. Estos modelos emplean características acústicas extraídas mediante herramientas como *openSMILE*, que permite obtener descriptores de bajo nivel (LLDs), coeficientes cepstrales en la frecuencia de Mel (MFCC), y estadísticas agregadas sobre segmentos temporales. Dichos enfoques han sido exitosos en tareas específicas, como

la detección de emociones (El Ayadi, Kamel, & Karray, 2011) o la clasificación de estrés en llamadas (Trigeorgis, y otros, 2016).

No obstante, estas soluciones enfrentan diversas limitaciones adicionales: no capturan adecuadamente la no linealidad y la variabilidad emocional del lenguaje hablado; requieren grandes volúmenes de datos para lograr generalización; y su rendimiento disminuye con señales ruidosas o dialectos variados (Zhang, Han, Deng, & Schuller, 2017). Además, aunque los modelos tradicionales pueden identificar emociones básicas, su precisión disminuye en tareas más complejas como la detección de veracidad o el perfilado del hablante, debido a la ambigüedad semántica y las diferencias individuales en la expresión vocal.

Por tanto, se identifica una necesidad de modelos más expresivos y adaptativos que puedan procesar con mayor profundidad las señales vocales, incluso en condiciones adversas o con muestras pequeñas. Este marco establece la línea base sobre la cual se propone explorar modelos híbridos asistidos por circuitos cuánticos variacionales, como alternativa de mayor capacidad representacional para el análisis conductual del habla.

1.5 Metodología

Esta investigación sigue un enfoque cuantitativo, prescriptivo y aplicado, orientado al desarrollo de modelos de aprendizaje supervisado para la caracterización conductual de interacciones orales en centros de contacto. La metodología propuesta está formada por 4 fases, cada una con sus etapas, que son:

Fase 1: Diseño, recolección y preprocesamiento de audios.
Fase 2: Conversión de voz a texto y extracción de características.
Fase 3: Construcción, despliegue y evaluación de modelos.
Fase 4: Exposición del modelo (API), integración y validación del sistema.

FASE 1.- Diseño, recolección y preprocesamiento de audios

Esta fase inicial define el marco metodológico del estudio y prepara el terreno para el análisis computacional. Se establece el diseño experimental, los objetivos y las

dimensiones conductuales a estudiar, siguiendo referentes metodológicos en el campo del análisis de la voz. A continuación, se procede a la recolección de los audios, ya sean reales o simulados, provenientes de interacciones en centros de contacto, asegurando un formato uniforme y condiciones controladas. Finalmente, se aplica un proceso riguroso de preprocesamiento de las señales de voz, que incluye reducción de ruido, normalización, segmentación por locutor y estandarización del formato, garantizando así la calidad y consistencia de los datos para las siguientes etapas.

FASE 2.- Procesamiento de la voz y generación de atributos acústicos y textuales

A lo largo de esta fase las señales de voz son procesadas para obtener representaciones matemáticas como vectores que contienen características, tanto textuales como acústicas. En primer lugar, cuando es pertinente, las grabaciones se transcriben utilizando sistemas automáticos de reconocimiento de voz (ASR) como el algoritmo *Whisper* (Radford, y otros, 2022), generando texto útil para el análisis semántico o lingüístico. Paralelamente, se ejecuta la extracción de características acústicas mediante herramientas especializadas como *openSMILE* (Eyben, Wöllmer, & Schuller, 2010), que permiten obtener descriptores como *MFCCs*[1], métricas prosódicas (tono, energía, ritmo), pausas, *jitter*, *shimmer*, y vectores X para la identificación de hablantes (Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018). Estas representaciones se convierten en insumos clave para los modelos supervisados en la siguiente fase.

FASE 3.- Construcción, despliegue y evaluación de modelos

Durante esta etapa se desarrollan y se someten a evaluación distintos clasificadores supervisados destinados a identificar patrones conductuales específicos, tales como estrés o veracidad, dentro de conversaciones telefónicas reales. Para ello, se propone una arquitectura híbrida que integra enfoques clásicos y cuánticos, seleccionando el más eficiente según el tipo de tarea.

En la parte clásica, se implementan métodos como Árboles de Decisión (Quinlan, 1993), Random Forest (Breiman, 2001), Bagging (Dietterich, 2000) y Gradient Boosting (Friedman, 2001). La optimización de estos modelos se realiza aplicando procesos iterativos de búsqueda de hiperparámetros mediante validación cruzada, utilizando como soporte la biblioteca *scikit-learn* (Pedregosa, y otros, 2011).

[1] Características espectrales basadas en frecuencia Mel utilizadas en análisis de voz

Como parte central de la innovación metodológica, se desarrollan modelos de clasificación mediante circuitos cuánticos variacionales (VQC), entrenados de forma híbrida utilizando optimización clásica sobre parámetros cuánticos. Estos modelos serán aplicados principalmente a la detección de estados emocionales y conductuales como estrés o veracidad. Para ello, se extraerán características acústicas del audio, incluyendo MFCC, así como sus primeras (Δ) y segundas (ΔΔ) derivadas temporales. También se extraen características textuales. Estas características, posteriormente son reducidas utilizando una red neuronal densa o en algunos casos usando la técnica de análisis de componentes principales (PCA) (Jolliffe, 2002). Los datos reducidos se procesarán a través de circuitos cuánticos variacionales, compuestos por secuencias de puertas parametrizadas (Benedetti, 2019) que permiten una representación eficiente con pocos qubits[2].

La implementación de los modelos cuánticos se realizará utilizando la librería *PennyLane* (Bergholm, y otros, 2022) integrada con *JAX* (Bradbury, y otros, 2018) en simuladores cuánticos locales. Se realizará la comparación de varias métricas entre los modelos cuánticos y clásicos. Se implementará un índice aritmético que permitirá comparar los modelos de forma objetiva. Finalmente, el modelo con mayor índice para cada tarea de clasificación será documentado, empaquetado y preparado para su despliegue. También se realizarán pruebas de rendimiento para validar su eficiencia y escalabilidad.

FASE 4.- Desarrollo del API, integración y validación del sistema

La última fase se centra en convertir el modelo entrenado en un servicio funcional y escalable. Para ello, se desarrolla una *API REST* que permite exponer el modelo como un servicio en la nube, utilizando tecnologías como *FastAPI*, *AWS API Gateway* y *AWS Lambda*. Esta API se integra directamente con el software operativo del centro de contacto, automatizando el flujo de caracterización conductual de las llamadas. Finalmente, se valida el sistema en escenarios reales o simulados, evaluando tanto su rendimiento técnico como su impacto en la eficiencia operativa, con el objetivo de reducir significativamente el tiempo de análisis manual y aumentar la calidad del servicio prestado.

El diseño sigue las mejores prácticas sugeridas por Miguel Grinberg (Grinberg, 2022) y la comunidad de desarrolladores de Apis modernas. La Figura 1.1 ilustra el pipeline

[2] Unidad fundamental de la computación cuántica

completo de la investigación, detallando las fases definidas y las etapas correspondientes a cada una de ellas.

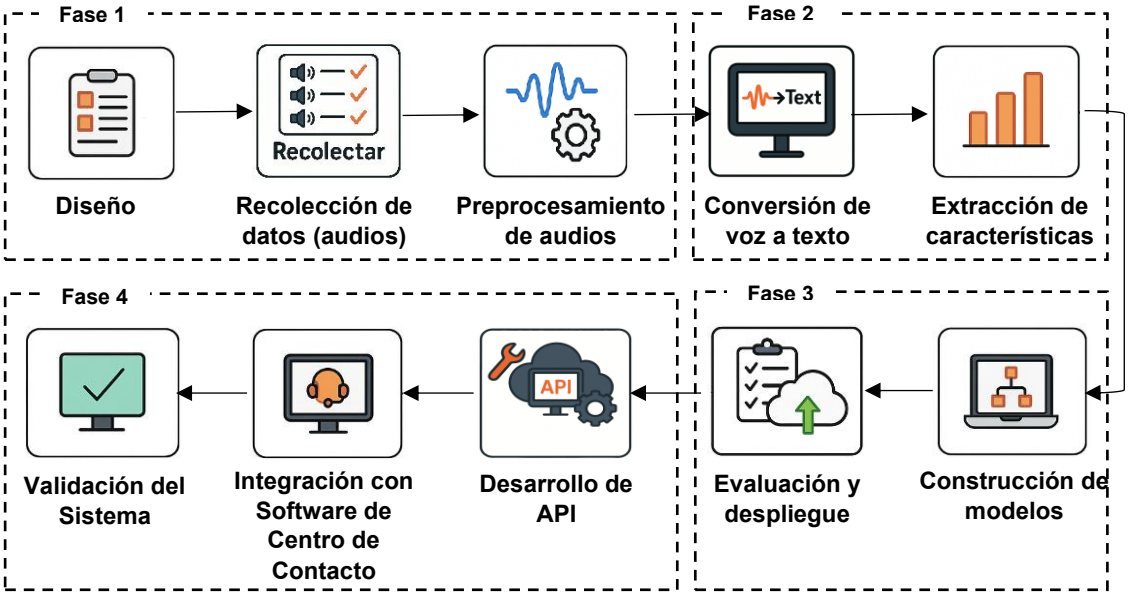


Figura 1.1 Fases para la caracterización conductual vocal en centros de contacto [Elaboración propia]

1.6 Resultados Esperados

1.6.1 Síntesis cronológica del estado del arte en procesamiento de voz emocional

Como resultado de la revisión bibliográfica realizada, se elabora una línea de tiempo que sintetiza los principales hitos en la evolución del estudio computacional de la voz y las emociones desde una perspectiva neurocientífica, de la física cuántica y de aprendizaje automático (ver Figura 1.2).

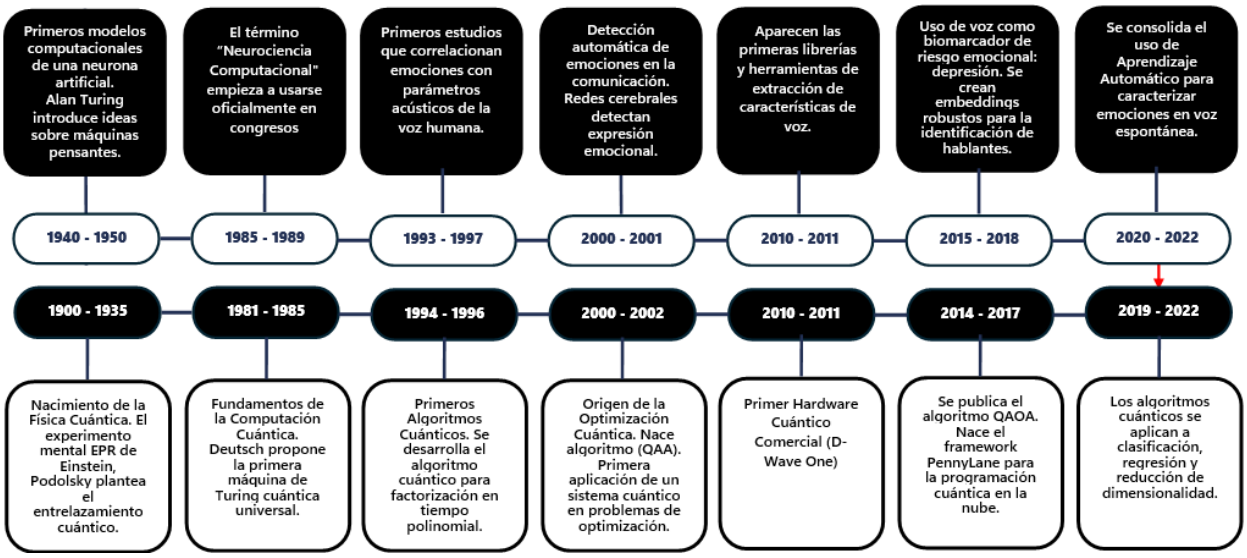


Figura 1.2 Línea de tiempo del estado del arte [Elaboración propia]

Esta línea cronológica comienza en la década de 1940–1950 con los primeros modelos computacionales de neuronas artificiales propuestos por McCulloch y Pitts (McCulloch & Pitts, 1943), y con las contribuciones teóricas de Alan Turing (Turing, 1950) sobre máquinas inteligentes. En el periodo de 1985–1989 se consolida el uso del término "Neurociencia Computacional" en entornos científicos y congresos especializados. Entre 1993 y 1997 surgen los primeros estudios que correlacionan emociones con parámetros acústicos de la voz humana, apoyados por las primeras tecnologías de reconocimiento automático. La utilización de la inteligencia artificial aplicada a la voz, estudiando patrones en las características acústicas, inició entre los años 2000 y 2001. Durante el periodo 2010–2011 aparecen librerías como *openSMILE*, que facilitan la extracción de características acústicas relevantes para el análisis emocional. Más adelante, entre 2015 y 2018, la voz comienza a utilizarse como biomarcador de riesgo emocional, particularmente en contextos relacionados con la depresión, y se desarrollan técnicas avanzadas como *embeddings* robustos para la identificación de hablantes. En el periodo 2020–2022 se consolida el uso del aprendizaje automático para la caracterización automática de emociones, niveles de estrés y patrones conductuales a partir de la voz espontánea, lo cual constituye la base tecnológica y científica sobre la que se estructura esta investigación.

Finalmente, a partir del año 2023, con la maduración de entornos de desarrollo cuántico como *PennyLane*, emerge una nueva etapa de innovación: la incorporación de algoritmos cuánticos variacionales (como QAOA y VQE) en tareas de clasificación dentro del entrenamiento de modelos de aprendizaje supervisado. Esta integración cuántico-clásica abre la posibilidad de reducir significativamente los tiempos de entrenamiento y mejorar la exploración de espacios de parámetros en sistemas complejos como el análisis emocional de voz, marcando así el inicio de una era híbrida en la computación afectiva.

1.6.2 Resultados Esperados en términos de Desempeño de los Modelos

Se espera que los modelos supervisados entrenados en esta investigación alcancen niveles de desempeño elevados en tareas de clasificación binaria y multiclase aplicadas a datos de voz en contextos reales. El sistema debe ser capaz de mantener su rendimiento a pesar de la presencia de ruido de fondo, variaciones individuales en el

habla, y la naturaleza espontánea de las conversaciones telefónicas. Para evaluar su rendimiento se utilizan métricas estándar en problemas de clasificación supervisada, especialmente aquellas que permiten interpretar el comportamiento del modelo en conjuntos de datos potencialmente desbalanceados.

La tabla 1.1 muestra las métricas seleccionadas, su descripción y los valores esperados de desempeño.

Tabla 1.1 Métricas de evaluación y resultados esperados

Métrica	Descripción	Resultado Esperado
Accuracy	Porcentaje de predicciones correctas sobre el total de predicciones realizadas.	> 75%
Precision	Proporción de verdaderos positivos entre todos los casos clasificados como positivos.	70% – 75%
Recall (Sensibilidad)	Porcentaje de verdaderos positivos correctamente identificados entre todos los positivos reales.	70% – 80%
F1-Score	Media armónica entre precisión y recall. Indicador clave en conjuntos de datos desbalanceados.	≈ 75%
Matriz de Confusión	Representación visual del rendimiento por clase, útil para detectar errores sistemáticos.	Interpretación cualitativa
Curva ROC-AUC	Área bajo la curva ROC para tareas binarias como veracidad o estrés. Evalúa capacidad discriminativa.	AUC > 0.75

Según estudios previos (Maza, 2011), el uso de representaciones fonéticas enriquecidas y características multimodales puede incrementar la precisión en más de un 10% respecto a enfoques tradicionales centrados únicamente en texto o palabras individuales.

1.6.3 Evaluación del desempeño entre modelos clásicos y cuánticos

Se obtendrá un análisis comparativo del rendimiento de modelos clásicos y cuánticos (VQC) en la clasificación de conductas orales, mostrando sus ventajas, limitaciones y condiciones de aplicabilidad según el tipo de tarea (emocional vs. identificación del hablante), evaluado empleando métricas estándares del aprendizaje automático.

1.6.4 Resultado Esperado del Producto Tecnológico

Como producto tecnológico final, se desarrollará una *API RESTful* que será desplegada en un entorno de nube pública o híbrida. Esta interfaz de programación estará diseñada para integrarse con plataformas de gestión de centros de contacto, permitiendo la interacción directa con los sistemas de grabación y análisis de llamadas.

La API tendrá la capacidad de recibir archivos de audio provenientes de interacciones orales, mediante un módulo especializado que realizará tareas de preprocesamiento

acústico y la extracción de características fonéticas y textuales relevantes para el análisis conductual.

Una vez extraídas las características, se ejecutará un modelo supervisado (cuántico y clásico) que permitirá la clasificación automática del audio en función de múltiples dimensiones: emociones (felicidad, tristeza, enojo y calma), la veracidad del discurso, y la posible identidad del hablante.

Los resultados de la clasificación se entregarán en tiempo real mediante respuestas estructuradas en formato JSON, facilitando su integración con *dashboards*, reportes u otros sistemas empresariales. La Figura 1.3 presenta la arquitectura propuesta diseñada con principios de escalabilidad, seguridad y eficiencia.

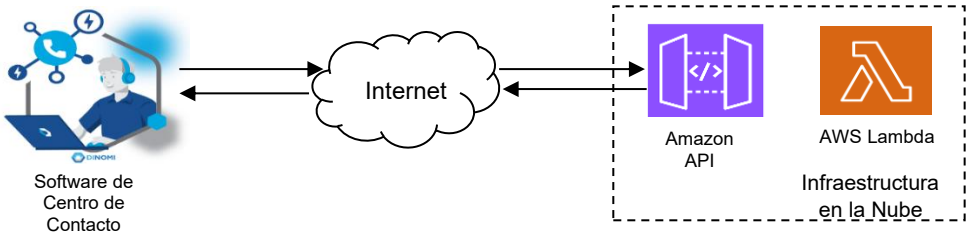


Figura 1.3 Arquitectura de despliegue y exposición del modelo e integración [Elaboración propia]

Se implementará un módulo de Caracterización Conductual mediante Análisis Fonético. Este módulo permitirá desplegar, en tiempo real, un conjunto de atributos derivados de la señal de voz durante la interacción oral entre un agente y un usuario. El sistema presentará las siguientes dimensiones de análisis fono conductual: emoción (clasificado en felicidad, tristeza, calma o enojo), veracidad del discurso (estimación binaria), así como la identificación del hablante.

Esta caracterización permitirá una comprensión más profunda del estado emocional y el perfil conductual del interlocutor, apoyando procesos de monitoreo de calidad, análisis de experiencia del cliente y detección de patrones críticos en entornos de centros de contacto.

Además de la caracterización conductual, el sistema DINOMI contempla la integración de una funcionalidad de identificación automática del hablante, que permita reconocer de forma precisa a la persona que participa en la interacción oral, sin necesidad de intervención manual. Esta característica se basa en tecnologías de reconocimiento de locutor (*speaker recognition*) que analizan características biométricas únicas de la voz, como el timbre, la entonación y el ritmo del habla. Una vez obtenidas las características

acústicas, éstas son comparadas en el sistema con registros previos del usuario. Si el registro es encontrado, entonces se logra asociar su huella vocal con una identidad específica. Esta capacidad resulta especialmente útil en escenarios de verificación de identidad, detección de fraudes y personalización de servicios dentro de centros de contacto y se integra perfectamente en el flujo de trabajo del agente, operando discretamente en segundo plano. Su incorporación añade una capa adicional de inteligencia al sistema, alineándose con los principios de seguridad, eficiencia y automatización.

Se implementará un módulo de visualización de caracterización conductual dentro de la plataforma DINOMI, como parte de los resultados tecnológicos esperados. Este módulo permite monitorear en tiempo real la información derivada del análisis fonético de las interacciones orales procesadas por el sistema.

Entre los elementos visualizados se encuentran: el análisis de emociones (felicidad, tristeza, calma o enojo), la identificación de los hablantes más frecuentes mediante huella vocal, la distribución de llamadas de acuerdo con la veracidad de las mismas, así como estadísticas relacionadas con llamadas por campañas salientes, estados actuales de los agentes y detalles de llamadas específicas.

1.7 Dataset

Los datos utilizados en esta investigación corresponden a un conjunto de registros derivados de llamadas telefónicas gestionadas a través del software DINOMI. Cada fila del dataset representa una interacción oral individual, y contiene tanto información técnica de la llamada como el identificador del archivo de audio asociado. En total, se recopilan los siguientes campos: la fecha y hora del evento, el número de origen y número de destino, la duración total de la llamada, el tipo de llamada (entrante o saliente) y el nombre del archivo de audio almacenado en formato .wav. Esta estructura permite una vinculación directa entre los metadatos de la llamada y el análisis fonético realizado posteriormente, garantizando así la trazabilidad y organización del corpus. La tabla 1.2 muestra descripción detallada de cada campo.

Tabla 1.2 Descripción de los campos del dataset utilizado (Emociones)

Campo	Tipo de Dato	Tipo de Variable	Descripción	Ejemplo
Fecha	Fecha	Cuantitativa temporal	Fecha en la que se registró la llamada.	1-may-25
Hora	Tiempo	Cuantitativa temporal	Hora exacta del inicio de la llamada.	18:00:08
Origen	Carácter	Cualitativa nominal	Número telefónico de origen (cliente o usuario).	991448748
Destino	Carácter	Cualitativa nominal	Número o extensión de destino (agente o sistema).	4000
Duración	Tiempo	Cuantitativa temporal	Duración total de la llamada en formato hh:mm:ss.	0:00:36
Tipo	Carácter	Cualitativa nominal	Tipo de llamada (Entrante o Saliente).	Entrante
File	Carácter	Cualitativa nominal	Nombre del archivo de audio asociado a la grabación.	q-4000-991448748-20250501-180035-1746140408.22233.wav
Clase	Carácter	Cualitativa nominal	Clase: Positiva, Negativa, Neutro	POSITIVA

El dataset contiene una estructura bien definida que permite vincular metadatos de llamadas telefónicas con sus correspondientes archivos de audio.

Adicionalmente, para la tarea de detección de veracidad y engaño, se desarrolló un experimento controlado que generó un conjunto adicional de 600 audios etiquetados según el grado de veracidad percibida. Este conjunto fue construido con el objetivo de contar con datos balanceados y etiquetados con claridad, bajo condiciones controladas que permitieran identificar patrones fonéticos y prosódicos relacionados con la conducta engañosa. Como parte de este segundo dataset, se incorporaron campos específicos asociados a los participantes y las condiciones del experimento. La tabla 1.3 muestra la descripción detallada de cada campo.

Tabla 1.3 Descripción de los campos del dataset utilizado (Veracidad)

Campo	Tipo de Dato	Tipo de Variable	Descripción	Ejemplo
ID Participante	Carácter	Cualitativa nominal	Código anónimo asignado a cada participante del experimento.	P008
Escenario	Carácter	Cualitativa nominal	Situación o pregunta planteada al participante (veraz o engañosa).	"Prometí pagar mañana"
Veracidad	Binaria	Cualitativa nominal	Etiqueta asignada por expertos (1: veraz, 0: engañosa).	0
Grado de Veracidad	Numérico	Cuantitativa ordinal	Valor subjetivo asignado por evaluadores en escala del 0 al 1.	0.25
Audio	Carácter	Cualitativa nominal	Nombre del archivo de audio generado durante el experimento.	exp-P008-esc2.wav

Cada dataset fue entrenado, evaluado y comparado usando modelos clásicos y cuánticos de aprendizaje automático, para la tarea de clasificación de emociones y veracidad.
Los detalles completos del diseño y resultados del experimento controlado para veracidad pueden consultarse en el Apéndice de este trabajo.

1.8 Consideraciones Éticas

El desarrollo de este proyecto de tesis, orientado al análisis automatizado de interacciones orales en centros de contacto, involucra el tratamiento de datos biométricos de voz, así como información sensible derivada de las conversaciones telefónicas. Bajo estas circunstancias, se adoptaron principios éticos esenciales para asegurar que toda la información recabada de los participantes sea manejada con total respeto a su privacidad, confidencialidad e integridad.
En primer lugar, se estableció que todas las grabaciones utilizadas en este estudio sean anonimizadas, eliminando cualquier vínculo directo con la identidad de los participantes. No se conservaron datos como nombres, cédulas, direcciones o números telefónicos una vez procesadas las grabaciones. Además, los archivos de audio y los resultados derivados del análisis fueron almacenados en entornos seguros y controlados, con acceso restringido exclusivamente al equipo investigador autorizado.
Dado que el sistema propuesto realizó la extracción y análisis de atributos biométricos, como la huella vocal del hablante, se adoptó un enfoque ético riguroso que contempló medidas de cifrado de la información tanto en tránsito como en reposo, autenticación de usuarios y mecanismos de trazabilidad en el uso del sistema. Debido a que los datos provienen de entornos reales de operación (como campañas de centros de contacto), se procuró contar con el consentimiento informado de los participantes o el respaldo institucional correspondiente que avale su uso con fines de investigación.
Finalmente, se reconoce que el desarrollo de tecnologías de análisis automático de emociones, estrés y veracidad puede tener implicaciones éticas adicionales en cuanto a la interpretación y uso de estos resultados. Por ello, se recomienda el uso de la inteligencia artificial de forma justa y responsable, tomando los resultados obtenidos de este trabajo solo como apoyo a la toma de decisiones y no como juicios definitivos sobre el comportamiento humano.

En el siguiente capítulo se revisan los principales fundamentos neurocientíficos, fonéticos y computacionales que han motivado el creciente uso de la inteligencia artificial para el modelamiento del comportamiento humando a partir de la voz. Esta revisión nos permite tener una línea base a partir de la cual se realiza este trabajo. También nos permitirá determinar cuáles son las variables fonéticas y textuales involucradas en la tarea de clasificación de caracterización conductual, así como las limitaciones encontradas en otros trabajos de investigación.

CAPÍTULO 2

2 ESTADO DEL ARTE

Comprender las interacciones orales en los centros de contacto (*contact center)* está adquiriendo una importancia estratégica en esta era de transformación digital. Estas interacciones, llenas de información emocional, conductual y operativa, representan una fuente de datos que se puede analizar desde diferentes perspectivas. Esta investigación tiene un enfoque que combina varias disciplinas como neurociencia, biometría del habla, inteligencia artificial y computación cuántica para mejorar las características de comportamiento de estas interacciones. Este capítulo explora los fundamentos teóricos y conceptuales que sustentan esta propuesta además de revisar los enfoques clásicos y nuevos utilizados para el análisis del habla, con especial atención a los modelos supervisados y su extensión en arquitecturas cuánticas.

2.1 Fundamentos teóricos y conceptuales

2.1.1 Fundamentos de neurociencia aplicados al análisis de voz

La voz humana es mucho más que un vehículo para transmitir información lingüística; es una manifestación directa de los estados internos del individuo, modulada por procesos neurofisiológicos complejos. Desde la neurociencia, se ha demostrado que las emociones y los procesos cognitivos influyen directamente en la producción vocal, afectando aspectos como el tono, la entonación, la velocidad y la intensidad del habla (Pell, Paulmann, Dara, Alasseri, & Kotz, 2015). Estos parámetros son reflejo de la actividad cerebral que integra tanto componentes emocionales como ejecutivos.

La amígdala cerebral, una estructura clave del sistema límbico, juega un rol fundamental en la generación y regulación de las emociones. Su activación modula respuestas autónomas y endocrinas que se traducen en cambios fisiológicos en la voz, como el aumento del tono en situaciones de miedo o estrés (Grandjean, y otros, 2006). Además, la corteza prefrontal regula estas respuestas a través del control ejecutivo, permitiendo la modulación consciente o inconsciente del discurso emocional.

A nivel motor, la corteza motora primaria, junto con el área de Broca, coordina la articulación vocal, mientras que el núcleo ambiguo y otras estructuras del tronco encefálico se encargan del control fino de la fonación y la respiración. Estas estructuras

interactúan dinámicamente con centros emocionales, lo que permite que las emociones se expresen en la voz incluso sin contenido verbal explícito.

Un componente de especial interés dentro del análisis fonoconductual es la función que cumplen las neuronas espejo, descubiertas inicialmente en primates y luego identificadas en humanos. Estas neuronas se activan tanto al ejecutar como al observar una acción, incluyendo expresiones faciales y tonos emocionales del habla (Gallese & Goldman, 1998). Su activación permite una forma de resonancia empática que fundamenta la comprensión emocional a través de la voz, y es clave para modelos computacionales que buscan simular este tipo de reconocimiento.

La literatura en neuroimagen sugiere que la comprensión emocional de la voz no depende de una única región cerebral, sino de una red distribuida que incluye regiones auditivas, áreas temporales superiores y estructuras insulares, las cuales procesan de forma integrada los componentes acústicos y semánticos del lenguaje. Esto permite que el oyente no solo entienda lo que se dice, sino cómo se dice, habilitando la detección de emociones, intenciones o estados mentales subyacentes.

Estos hallazgos son una línea base sólida para la implementación de modelos computacionales que intentan emular el reconocimiento emocional en la voz. Al comprender cómo el cerebro procesa y modula las señales emocionales vocales, se facilita la construcción de algoritmos de aprendizaje automático y computación cuántica que capturen patrones complejos más allá de lo meramente acústico.

2.1.2 Características fonéticas relevantes para el análisis conductual

El análisis conductual de la voz se fundamenta en la extracción de parámetros fonéticos que reflejan, de manera indirecta, los estados internos del hablante. Estos parámetros, que incluyen propiedades acústicas y prosódicas, permiten inferir aspectos como el contenido emocional del discurso, niveles de activación fisiológica (estrés) e incluso patrones de veracidad o engaño. En este contexto, las características fonéticas se convierten en indicadores clave para sistemas de análisis automático en entornos como los centros de contacto.

Las principales características fonética y sus descripciones se muestran en la tabla 2.1

Tabla 2.1 Características fonéticas relevantes

Característica Fonética	Descripción	Utilidad en Biometría Vocal
Frecuencia Fundamental (F0)	Representa el tono de la voz y está relacionada con la tensión de las cuerdas vocales y el sistema nervioso autónomo. F0 elevada se asocia con ira o miedo; F0 baja con calma o tristeza.	Permite distinguir perfiles emocionales y de activación fisiológica propios del hablante.
Formantes (F1, F2, F3)	Picos de resonancia del tracto vocal que determinan la calidad de las vocales. Su variación refleja la articulación emocional.	Funcionan como huellas vocales personales debido a su estabilidad anatómica.
Duración de fonemas y pausas	Incluye la duración de sílabas y la frecuencia de pausas. Cambios indican duda, esfuerzo cognitivo o engaño.	Ayuda a detectar patrones personales de ritmo y control del habla.
Tasa de elocución	Mide la velocidad del habla (sílabas por segundo). Se acelera con estrés y se reduce con tristeza o fatiga.	Identifica variaciones individuales en el ritmo expresivo frente a distintos estados.
Intensidad o energía de la voz	Refleja la fuerza acústica de la señal. Alta energía se relaciona con emociones intensas; baja con inhibición o depresión.	Señal distintiva de estados emocionales y de estilo vocal personal.
Espectro armónico y ruido	Incluye medidas como HNR, jitter y shimmer, que capturan irregularidades en la voz.	Detecta condiciones fisiológicas únicas y respuestas vocales ante estrés.
Coeficientes MFCCs	Representan la envolvente espectral de la voz y son estándar en procesamiento automático del habla.	Alta capacidad de discriminación entre hablantes; base de muchos sistemas de reconocimiento.

Estas características pueden ser extraídas mediante herramientas especializadas como *openSMILE* y *librosa* (McFee, Raffel, Liang, & Ellis, 2015), y constituyen la base de representación para entrenar modelos de aprendizaje automático o incluso circuitos cuánticos variacionales, como se explora en capítulos posteriores de este trabajo.

Desde un enfoque neurocientífico, estas variaciones fonéticas están moduladas por estructuras subcorticales y corticales involucradas en la producción emocional del habla, tales como la amígdala, la corteza prefrontal ventromedial[3] y la ínsula[4]. Así, el estudio fonético no solo aporta valor técnico, sino también una comprensión integradora entre la fisiología de la voz y los procesos mentales que la regulan.

2.1.3 Biometría vocal e identificación de hablantes

La biometría vocal es una rama de la biometría que emplea las características únicas de la voz humana para identificar o verificar la identidad de un individuo. A diferencia de otros rasgos biométricos como las huellas dactilares o el iris, la voz es una señal acústico-fisiológica y conductual, influenciada tanto por la anatomía del hablante (por ejemplo, tamaño del tracto vocal) como por patrones aprendidos de articulación, entonación y prosodia (Campbell, Reynolds, & Torres-Carrasquillo, 2010). Esta doble

[3] Zona frontal del cerebro asociada a la regulación emocional.

[4] La ínsula es una región del cerebro situada dentro del surco lateral, implicada en el procesamiento del dolor, las emociones.

naturaleza convierte a la biometría vocal en una herramienta especialmente útil en contextos dinámicos como los centros de contacto, donde la verificación pasiva y no invasiva es deseable.

La voz posee atributos estáticos y dinámicos. Los atributos estáticos están relacionados con la configuración fisiológica del aparato fonador del individuo, como la longitud de las cuerdas vocales, la cavidad nasal y bucal, y la forma del tracto vocal, lo que influye directamente en los formantes vocales y en la frecuencia fundamental (F0). Además, los atributos dinámicos hacen referencia a características comportamentales como el ritmo del habla, los patrones de entonación y la pronunciación de fonemas, los cuales varían de acuerdo con el estado emocional o contexto social del hablante.

Los sistemas de identificación de hablantes basados en biometría vocal pueden clasificarse en dos categorías principales, tal como lo muestra la tabla 2.2.

Tabla 2.2 Tipos de métodos de reconocimiento de hablantes

Tarea	Definición	Pregunta que responde	Tipo de Clasificación
Verificación de identidad *(speaker verification)*	Confirma si una voz pertenece a una identidad previamente registrada.	¿Es esta persona quien dice ser?	**Binaria** *(sí/no)*
Identificación de hablante *(speaker identification)*	Determina a cuál de varias identidades conocidas corresponde una muestra de voz.	¿Quién es esta persona?	**Multiclase**

Tradicionalmente, se empleaban técnicas basadas en arquitecturas probabilísticas de mezcla gaussiana con modelos de referencia global y representaciones compactas del hablante basadas en vectores de identidad, pero en años recientes han ganado protagonismo los enfoques basados en aprendizaje profundo y, más recientemente, los métodos híbridos clásicos-cuánticos.

En contextos como centros de contacto, la biometría vocal permite no solo autenticar al usuario, sino también realizar un seguimiento continuo del patrón vocal para detectar anomalías que puedan indicar suplantación de identidad, alteración emocional o cambios fisiológicos.

2.1.4 La voz como canal emocional y biométrico

La dualidad del habla como canal afectivo y biométrico lo convierte en un material importante para la investigación del análisis del comportamiento. Desde una perspectiva afectiva, el habla transmite información emocional que puede enriquecer la interpretación

de las interacciones humanas. Desde una perspectiva biométrica, puede verificar la identidad de una persona basándose en patrones de habla estables. La combinación de e|tas dos dimensiones crea modelos híbridos que describen no sólo a la persona que habla, sino también sus emociones o la intención del discurso. Esta cantidad de información es fundamental en entornos como los centros de contacto, donde el comportamiento de los usuarios afecta directamente a las decisiones operativas y empresariales.

2.1.5 Fundamentos de física cuántica (definiciones más recientes)

El surgimiento de la computación cuántica ha traído consigo la necesidad de reinterpretar los principios fundamentales de la mecánica cuántica en contextos aplicados, como el procesamiento de información y el aprendizaje automático. A continuación, se presentan los conceptos esenciales que constituyen la base de los modelos cuánticos utilizados en esta investigación: superposición, entrelazamiento, qubit y estado cuántico.

Principio de superposición

Los sistemas clásicos se encuentran en un estado definido (por ejemplo, encendido o apagado, 0 o 1). Mientras que los sistemas cuánticos tienen al principio de superposición como una de las propiedades más fascinante y distintiva. Un sistema cuántico puede encontrarse en una combinación lineal de múltiples estados simultáneamente.

En el caso de un qubit (unidad mínima de información cuántica), este puede representarse como:

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

donde α y β son coeficientes complejos que cumplen con la condición:

$$|\alpha|^2 + |\beta|^2 = 1 \quad (2)$$

Y $|0\rangle$ y $|1\rangle$ representan los estados base del sistema. Esta propiedad permite que los algoritmos cuánticos exploren múltiples soluciones en paralelo, mejorando la eficiencia computacional en ciertos problemas complejos (Nielsen & Chuang, 2010).

Entrelazamiento Cuántico

El entrelazamiento cuántico describe una propiedad física mediante la cual múltiples sistemas cuánticos quedan vinculados de forma que el estado de cada uno no puede analizarse de manera aislada, sino únicamente en relación con los demás, aun cuando se encuentren separados por grandes distancias.

Formalmente, un estado entrelazado de dos qubits puede expresarse como:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \quad (3)$$

Este tipo de correlación no tiene análogo clásico y tiene muchas aplicaciones en el mundo de la computación cuántica, como la teleportación, la criptografía y la implementación eficiente de compuertas multiqubit en algoritmos como los circuitos variacionales. En modelos híbridos, el entrelazamiento permite que varios qubits actúen como una unidad coordinada, incrementando la capacidad de representación del sistema.

Qubit y Estado Cuántico

El qubit (quantum bit) es la unidad mínima de procesamiento en un sistema cuántico. A diferencia del bit clásico, que puede tomar solo uno de dos valores, el qubit puede estar en un estado superpuesto y, adicionalmente, en estados entrelazados con otros qubits. Su representación matemática se hace mediante un vector en un espacio de Hilbert bidimensional como lo muestra la Ecuación (1).

El estado cuántico, por su parte, describe completamente el sistema. Puede representarse como una función de onda, un vector de estado (como arriba), o como una matriz densidad si el sistema está en una mezcla estadística de estados puros. La evolución de estos estados está gobernada por operadores unitarios (compuertas cuánticas) y su medición colapsa el estado a uno de los posibles resultados clásicos, con una probabilidad determinada por la amplitud al cuadrado del coeficiente asociado (Nielsen & Chuang, 2010).

Esta capacidad de los qubits de mantenerse en superposición y participar en entrelazamientos es la base del paralelismo cuántico, lo cual otorga a la computación cuántica un potencial de ventaja exponencial sobre ciertos problemas difíciles para algoritmos clásicos.

2.2 Técnicas clásicas de análisis de voz en centros de contacto

2.2.1 Procesamiento manual tradicional y sus limitaciones

Durante muchos años, el análisis de las interacciones de voz en los centros de contacto ha sido realizado manualmente por supervisores o auditores escuchando directamente las grabaciones. Este enfoque es capaz de identificar patrones de cortesía, cumplimiento de protocolos o presencia de comportamientos agresivos, pero su subjetividad y escasa escalabilidad limitan su aplicabilidad en entornos con gran cantidad de datos (Wang et al., 2021). Además, la dependencia del juicio humano introduce sesgos y variabilidad, lo que compromete la objetividad del análisis.

2.2.2 Aprendizaje automático clásico aplicado a señales acústicas

A pesar del incremento de la aplicación de modelos de aprendizaje profundo en el análisis del habla, las técnicas clásicas de aprendizaje automático continúan teniendo relevancia en ciertos contextos prácticos, particularmente en tareas específicas dentro de centros de contacto con restricciones computacionales o de datos.

Modelos basados en técnicas de ensamble, de margen máximo y reglas de partición han sido aplicados para la clasificación de sentimientos, el reconocimiento de patrones acústicos y la identificación de hablantes, utilizando vectores de características extraídos de señales de audio, principalmente descriptores cepstrales en frecuencia de Mel (MFCC), formantes, y estadísticas temporales.

Aunque estos enfoques muestran precisión aceptable en condiciones controladas, su desempeño suele decrecer ante la variabilidad natural del habla humana, los ruidos de fondo o los acentos regionales presentes en interacciones reales. Por esta razón, suelen emplearse como líneas base comparativas frente a modelos más avanzados, así como también para el descubrimiento de nuevas técnicas como el uso de circuitos cuánticos variacionales, que se abordan en el siguiente apartado.

Aun así, el aprendizaje clásico mantiene valor en sistemas embebidos, evaluaciones interpretables y tareas donde la eficiencia es prioritaria. Su simplicidad algorítmica

permite implementaciones rápidas y fácilmente replicables, lo cual ha motivado su uso como primer paso de experimentación en múltiples estudios recientes.

2.2.3 Modelos aplicados a la detección de emociones

La detección automática de emociones se ha visto impulsada por la evolución reciente de arquitecturas de inteligencia artificial, capaces de capturar patrones complejos en señales acústicas, textuales o multimodales. En contextos como centros de contacto, la correcta identificación del estado emocional del hablante permite mejorar la experiencia del usuario, adaptar respuestas empáticas y generar métricas útiles para la gestión operativa. Los enfoques modernos más relevantes en el estado del arte incluyen:

Redes Neuronales Convolucionales (CNN)

Las CNN se han aplicado exitosamente al análisis de espectrogramas y representaciones cepstrales del habla, aprovechando su capacidad para identificar patrones locales y temporales en la señal. Investigaciones recientes han demostrado que modelos basados en CNN superan ampliamente a métodos clásicos como SVM o KNN en la clasificación de emociones con señales acústicas crudas.

Redes Recurrentes (RNN, LSTM, GRU)

Debido a su naturaleza secuencial, las redes LSTM y GRU han sido ampliamente utilizadas para capturar la evolución temporal de características acústicas. Estas arquitecturas han mostrado buen rendimiento en bases de datos multilingües y ruidosas, representando emociones como ira, tristeza, alegría y neutralidad alta precisión. También se han implementado modelos híbridos CNN-LSTM para combinar la extracción espacial y la memoria temporal.

Transformers aplicados al habla

Inspirados en su éxito en NLP (procesamiento de lenguaje natural), los Transformers han comenzado a utilizarse para tareas de reconocimiento emocional en el habla. Modelos como wav2vec 2.0, HuBERT o AST (*Audio Spectrogram Transformer*) han mostrado resultados superiores al combinar aprendizaje auto-supervisado con *fine-tuning* emocional. Estas arquitecturas permiten trabajar directamente con audio crudo o

espectrogramas, logrando representar dependencias largas en el tiempo sin recurrir a RNNs.

Modelos multimodales

La tendencia actual se inclina hacia sistemas multimodales, que combinan audio, texto y video. En la detección emocional basada en voz, el texto transcrito mediante ASR (*Automatic Speech Recognition*) se combina con las señales acústicas para mejorar el rendimiento general. Modelos como CMU-MOSEI y MELD son utilizados como benchmarks en esta línea (Poria et al., 2021).

Uso de aprendizaje auto-supervisado y transferencia

Una línea reciente de investigación aplica transferencia de conocimiento desde modelos pre-entrenados en grandes corpus de audio, como wav2vec 2.0 de Facebook AI, que luego se ajustan con conjuntos de datos específicos de emociones. Esto permite mejorar el rendimiento incluso en escenarios con pocos datos etiquetados, lo cual es especialmente útil en industrias con datos sensibles o limitados (Pepino, Riera, Cullen, & Saraceno, 2021).

En resumen, los modelos actuales para la detección de emociones en el habla han evolucionado hacia arquitecturas profundas y multimodales, superando las limitaciones de los enfoques tradicionales. Estas soluciones no solo incrementan la precisión en contextos reales y desafiantes, como los centros de contacto, sino que también permiten una representación más rica y contextualizada del estado emocional del hablante, sentando así las bases para sistemas conversacionales más empáticos e inteligentes.

2.3 Inteligencia artificial en la caracterización conductual

La caracterización conductual mediante Inteligencia Artificial (IA) constituye una línea de investigación emergente que busca interpretar patrones vocales, lingüísticos y paralingüísticos con el fin de identificar indicadores del comportamiento humano, más allá de las emociones básicas. En el contexto de centros de contacto, esta caracterización incluye dimensiones como la veracidad del discurso, el nivel de estrés vocal, la intención comunicativa, y aspectos del perfil del hablante, todos ellos clave para tareas como autenticación, análisis de calidad del servicio y detección de riesgo.

Los enfoques actuales utilizan principalmente modelos de aprendizaje supervisado y aprendizaje profundo, entrenados con vectores de características acústicas como MFCC, pitch, jitter, shimmer y espectrogramas mel, combinados con técnicas de preprocesamiento emocional o fonético (Rana et al., 2023).

2.3.1 Veracidad y detección de engaño

El análisis de la veracidad a partir del habla se ha convertido en una aplicación emergente de IA conductual. Modelos basados en SVM, Random Forest y más recientemente redes neuronales profundas (DNNs, CNNs y BiLSTM) han sido aplicados para detectar patrones vocales sutiles asociados al engaño, como pausas prolongadas, fluctuaciones en la entonación y aumento del estrés vocal. Estudios recientes muestran que la combinación de señales acústicas con transcripciones automáticas mejora la detección de falsedad en contextos reales.

Un referente relevante en la literatura es el estudio desarrollado por Lloyd et al. (2019), quienes crearon el dataset MU3D (*Miami University Deception Detection Database*), compuesto por 320 registros audiovisuales de individuos narrando historias verdaderas y falsas en contextos emocionales positivos y negativos. El protocolo experimental fue diseñado por psicólogos y validado científicamente mediante jueces humanos y análisis computacionales. Este dataset ha sido utilizado ampliamente como *benchmark* (referencia comparativa) en estudios de detección de engaño.

Los resultados obtenidos en este experimento reflejan que incluso los seres humanos, incluyendo psicólogos forenses y policías entrenados, logran precisiones moderadas al identificar mentiras, como se resume en la tabla 2.3.

Tabla 2.3 Análisis comparativo de resultados en tareas de veracidad

Estudio / Contexto	*Accuracy*
Humanos promedio (Ekman, 2009)	54% – 58%
Psicólogos forenses (Ekman & O'Sullivan, 1991)	64% – 68%
MU3D – Modelos automáticos	60% – 75%
Perez-Rosas et al. (2015) – Audio/Text	70% – 75%

Esto evidencia que la mentira es un fenómeno difícil de detectar incluso por humanos, y que los modelos supervisados automatizados pueden lograr desempeños comparables o superiores. Además, el diseño experimental de MU3D ha servido como base para adaptaciones en investigaciones aplicadas, como la que se desarrolla en esta tesis,

donde se propone replicar el protocolo de grabación con una muestra local y analizar las señales acústicas derivadas.

2.3.2 Estimación de estrés y carga cognitiva

La detección del estrés en la voz ha evolucionado desde técnicas tradicionales basadas en umbrales acústicos hasta modelos más sofisticados entrenados con bases de datos anotadas manualmente. Algoritmos como CNN-LSTM, Transformers acústicos y enfoques auto-supervisados permiten clasificar niveles de estrés con precisión creciente, incluso en presencia de ruido o variabilidad interindividual (Yin et al., 2023). Estas tecnologías se aplican ya en teleasistencia, salud ocupacional y monitoreo emocional en tiempo real.

2.3.3 Perfil del hablante y rasgos conductuales

La caracterización conductual también incluye el análisis de rasgos de personalidad, estilo de comunicación, actitud e incluso determinación de identidad y región geográfica, a partir del timbre, acento o velocidad del habla. Modelos actuales utilizan representaciones densas derivadas de speech embeddings como x-vectors y ECAPA-TDNN, entrenadas para tareas de reconocimiento de hablante, pero reutilizadas para perfilar rasgos conductuales.

2.3.4 Enfoques actuales e integración multimodal

La tendencia actual combina modelos de audio con fuentes textuales y contextuales (por ejemplo, intención del mensaje o historial de conversación), para crear sistemas multimodales de caracterización conductual. En estos sistemas, el aprendizaje supervisado es asistido por técnicas de atención (attention layers), codificación temporal profunda y transferencia desde modelos preentrenados, como wav2vec 2.0 y HuBERT (Pepino, Riera, Cullen, & Saraceno, 2021)

Estos avances permiten diseñar soluciones de IA que no solo reconocen lo que una persona dice, sino cómo lo dice, por qué lo dice, y qué podría estar sintiendo o escondiendo, lo que abre nuevas oportunidades en áreas como seguridad, educación, salud y experiencia del cliente.

2.4 Computación cuántica aplicada al aprendizaje supervisado

La computación cuántica, tradicionalmente asociada con simulaciones de sistemas físicos y criptografía, ha comenzado a desempeñar un papel fundamental en las aplicaciones modernas de la inteligencia artificial, con especial énfasis en los modelos supervisados. La incorporación de circuitos cuánticos variacionales (VQC, por sus siglas en inglés) ha permitido desarrollar algoritmos híbridos capaces de aprovechar tanto el poder del cómputo cuántico como la robustez de los modelos clásicos. Este enfoque se sitúa en el contexto de la era NISQ (*Noisy Intermediate-Scale Quantum*), caracterizada por hardware cuántico limitado pero funcional, sobre el cual se están construyendo modelos entrenables y aplicables a tareas reales como clasificación, regresión y reducción de dimensionalidad (Benedetti, 2019).

2.4.1 Fundamentos de los Circuitos Cuánticos Variacionales (VQC)

Los VQC son una de las arquitecturas más prometedoras para el aprendizaje supervisado en el contexto cuántico actual. Consisten en circuitos cuánticos parametrizados cuyas compuertas dependen de variables ajustables, optimizadas mediante algoritmos clásicos. Su estructura híbrida permite entrenarlos usando métodos como descenso por gradiente u otros optimizadores, mientras que su parte cuántica explota propiedades como la superposición y el entrelazamiento para explorar el espacio de soluciones de forma más eficiente que sus contrapartes clásicas.

Uno de los beneficios más relevantes radica en su habilidad para modelar fenómenos de alta complejidad con relativamente pocos parámetros, gracias a la riqueza del espacio de Hilbert cuántico. Además, permiten ejecutar entrenamientos en hardware real o simuladores, utilizando marcos como *PennyLane* o *Qiskit*.

2.4.2 Compuertas Cuánticas Básicas

Los VQC se construyen a partir de compuertas cuánticas unitarias que actúan sobre uno o varios qubits. Las compuertas básicas incluyen:

Tabla 2.4 Compuertas cuánticas

Compuerta cuántica	Símbolo	Detalle
Compuertas de rotación	$R_x(\theta), R_y(\theta), R_z(\theta)$	Aplican rotaciones alrededor de los ejes del Bloque de Bloch.
Compuerta Hadamard	(H)	crea superposición entre estados $\|0\rangle$ y $\|1\rangle$
Compuerta Pauli	(X, Y, Z)	representan operaciones de inversión en los ejes.
Compuertas de control	(CNOT, CZ)	fundamentales para generar entrelazamiento.

Estas compuertas se combinan para construir capas parametrizadas en los VQC. Su uso eficiente determina la expresividad y la entrenabilidad del modelo cuántico (Biamonte et al., 2017).

2.4.3 Circuitos Cuánticos y Codificación de Datos Clásicos

Uno de los principales desafíos del aprendizaje supervisado cuántico es cómo codificar datos clásicos (como vectores numéricos) en estados cuánticos. Las estrategias más recientes incluyen:

- **Codificación de ángulo (angle encoding)**: se usa una variable clásica para rotar un qubit.
- **Codificación amplitud (amplitude encoding)**: permite representar un vector normalizado como un estado cuántico.
- **Codificación de densidad (density matrix encoding)**: útil para representar distribuciones estadísticas.

La codificación influye directamente en la capacidad del modelo de aprender patrones útiles. Una de las estrategias más eficientes en términos de uso de qubits es la codificación por amplitud, también conocida como ***Amplitude Embedding*** (Schuld & Killoran, 2019). En este enfoque, un vector de características clásicas $x \in R^{2^n}$, previamente normalizado, se utiliza para preparar un estado cuántico con n qubits, tal que:

$$x = [x_0, x_1, \ldots, x_{2^n-1}] \Rightarrow |\Psi\rangle = \sum_{i=0}^{2^n-1} x_i |i\rangle \quad (4)$$

Donde $|\Psi\rangle$ representa el estado cuántico codificado. Esta técnica permite cargar 2^n componentes con solo n qubits, ofreciendo una representación densa y eficiente de la

información. Su uso ha ganado popularidad en modelos híbridos donde se requiere una transición compacta del espacio clásico al cuántico. En la siguiente tabla se comparan los principales métodos de codificación de datos en circuitos cuánticos variacionales, destacando sus requerimientos, ventajas y aplicaciones típicas.

Tabla 2.5 Comparación de técnicas de codificación en circuitos cuánticos variacionales

Codificación	Qubits requeridos	Normalización	Ventaja principal	Ejemplo de uso
AngleEmbedding	n	No necesaria	Sencilla implementación	Problemas pequeños
AmplitudeEmbedding	$\log_2 N$	Requiere normalización L2	Alta densidad de información	Clasificación híbrida con VQC
BasisEmbedding	n	No	Binaria o categórica	Codificación de clases

La elección de la técnica de codificación influye directamente en el tipo de datos que pueden procesarse y en la eficiencia del circuito. *AmplitudeEmbedding,* utilizada en trabajos recientes, destaca por su capacidad de representar vectores densos con pocos qubits.

2.4.4 Tipos de Circuitos Utilizados

En el desarrollo de modelos de aprendizaje cuántico supervisado, la elección del circuito cuántico o ansatz es un componente crucial, ya que determina la capacidad del sistema para representar funciones complejas, su compatibilidad con hardware real y su susceptibilidad a problemas como el barrenamiento de gradiente. A continuación, se analizan las configuraciones de circuitos más empleadas en trabajos recientes:

Ansatz Personalizados basados en Codificación Densa

En investigaciones recientes se ha destacado el uso de arquitecturas híbridas personalizadas que integran codificaciones estructuradas, como *AmplitudeEmbedding*, con capas de rotaciones parametrizadas (*Rx, Ry, Rz*) aplicadas de manera individual por qubit y entrelazamiento lineal mediante compuertas CNOT entre qubits adyacentes. Este tipo de diseño, aunque más simple que las capas fuertemente entrelazadas, ha demostrado ser efectivo para tareas supervisadas de clasificación, especialmente cuando se combinan con redes neuronales clásicas encargadas de preprocesar características.

A diferencia de otros enfoques más genéricos, los *ansatz* personalizados pueden adaptarse a la naturaleza del problema y a las dimensiones específicas de entrada, logrando un equilibrio entre expresividad, eficiencia y estabilidad del entrenamiento. Esta arquitectura se alinea con propuestas contemporáneas de circuitos centrados en el aprendizaje (*circuit-centric quantum classifiers*) que permiten la integración directa con redes densas clásicas (Mitarai, Negoro, Kitagawa, & Fujii, 2018).

StronglyEntanglingLayers

Este tipo de circuito busca maximizar el entrelazamiento entre qubits, utilizando patrones densos de compuertas CNOT entre todos los pares posibles de qubits, combinadas con rotaciones parametrizadas (*Rx, Ry, Rz*) aplicadas antes o después del entrelazamiento. Son conocidos por su alta expresividad y su capacidad para aproximar funciones complejas, lo que los hace adecuados para tareas de clasificación. Su alta expresividad es ideal para tareas de clasificación, aunque puede presentar problemas de barrenamiento de gradiente.

Hardware-Efficient Ansatz

Los hardware-efficient ansatz están diseñados para minimizar la profundidad del circuito y adaptarse a las limitaciones de los dispositivos cuánticos actuales (NISQ). Usualmente consisten en bloques repetitivos que combinan rotaciones unitarias independientes en cada qubit con patrones de entrelazamiento local entre qubits vecinos. Aunque no son perfectos desde el punto de vista teórico, funcionan bien en la práctica y toleran los errores del sistema, por lo que resultan adecuados para experimentos reales.

2.4.5 Herramientas Cuánticas Actuales: PennyLane, Qiskit, JAX

La investigación y experimentación en aprendizaje supervisado cuántico se apoya en herramientas como:

- **PennyLane**: librería Python que permite diseñar y entrenar VQC de forma híbrida, integrándose con PyTorch y TensorFlow. Destaca por su compatibilidad con diferenciales automáticos (Bergholm, y otros, 2022).

- **Qiskit**: desarrollada por IBM, permite construir circuitos cuánticos a bajo nivel y simular su ejecución. Incorpora Qiskit Machine Learning para integración con Scikit-Learn.
- **JAX**: aunque no es específica de computación cuántica, se ha integrado con simuladores cuánticos para aplicar métodos de gradiente en hardware acelerado (GPU/TPU), facilitando la optimización de VQC (Gong et al., 2023).

Estas herramientas han hecho posible validar modelos de clasificación en datasets reales como Iris, MNIST, y conjuntos médicos o acústicos.

2.4.6 Redes Neuronales vs. Circuitos Cuánticos Variacionales

Estudios recientes han comparado el rendimiento de modelos clásicos como redes neuronales densas (DNN) con circuitos VQC en tareas de clasificación binaria y multiclase. Aunque las redes profundas siguen siendo superiores en datasets complejos y masivos, los VQC han mostrado resultados comparables en conjuntos de baja dimensión y con pocos datos, lo cual los hace ideales en contextos como el procesamiento de voz o señales biométricas en entornos restringidos (Tacchino et al., 2020). Además, los VQC presentan mayor capacidad de generalización en ciertos casos, gracias a la no linealidad inherente a la evolución cuántica. Sin embargo, el entrenamiento cuántico sigue enfrentando desafíos como la barrenación de gradiente, la sensibilidad al ruido y los problemas de escalabilidad, lo que ha motivado el uso de enfoques híbridos y la investigación en nuevas funciones de costo adaptativas.

2.5 Optimización en modelos de clasificación

La optimización es uno de los aspectos más importantes del aprendizaje automático, porque hace posible que el modelo se vaya ajustando para cometer cada vez menos errores en sus predicciones. En los clasificadores, este refinamiento se realiza mediante funciones que miden la diferencia entre lo que el sistema estima y lo que ocurre en realidad. A lo largo de los años, han aparecido múltiples formas de optimizar modelos, desde los enfoques clásicos hasta las propuestas más recientes que combinan computación convencional con computación cuántica.

2.5.1 Optimización en redes neuronales

Función de pérdida y función de costo

Entrenar un modelo en aprendizaje supervisado consiste, básicamente, en hacer que falle cada vez menos. Para lograrlo, se compara lo que el modelo predice con lo que realmente ocurre, y esa diferencia se mide con una función de pérdida $L(\hat{y}, y)$ que indica el nivel de error en cada caso.

Para tareas de clasificación multiclase, una de las funciones de pérdida más comúnmente utilizadas es la entropía cruzada, especialmente cuando la salida del modelo es una distribución de probabilidad (como en softmax), definida como:

$$L(\hat{y}, y) = -\sum_{i=1}^{c} y_i \log(\hat{y}_i) \tag{5}$$

donde ℓ es la función de pérdida, y_i es la etiqueta real y $\hat{y}_i$ es la predicción del modelo.

Esta función penaliza más fuertemente las predicciones incorrectas con alta confianza y es ampliamente utilizada en redes neuronales profundas.

A partir de esta función de pérdida individual, se define la **función de costo** $J(\theta)$ como el error promedio en todo el conjunto de entrenamiento:

$$J(\theta) = \frac{1}{N}\sum_{k=1}^{n} L(\hat{y}^{(k)}, y^{(k)}) \tag{6}$$

Donde:

- N es el número total de muestras,
- $\hat{y}^{(k)}$ es la predicción para la muestra k,
- $y^{(k)}$ es la etiqueta real correspondiente,
- θ son los parámetros del modelo

Descenso de gradiente y retropropagación

El descenso de gradiente es un proceso iterativo que mejora un modelo ajustando sus parámetros para que el valor del error sea cada vez. Su función de costo es:

$$\theta_i \leftarrow \theta_i - \eta \frac{\partial J}{\partial \theta_i} \tag{7}$$

donde η es la tasa de aprendizaje. Para redes neuronales, los cambios necesarios (gradientes) se obtienen mediante el algoritmo de retropropagación, que distribuye el error a lo largo de las distintas capas para poder entrenar redes profundas (Rumelhart, Hinton, & Williams, 1986).

Optimizadores clásicos

Además del descenso de gradiente básico, se han desarrollado variantes más eficientes:

- **Stochastic Gradient Descent (SGD)**: actualiza con minibatches de datos.
- **Adam (Adaptive Moment Estimation)**: combina momento y normalización adaptativa.
- **RMSProp, Adagrad**: adaptan la tasa de aprendizaje a cada parámetro.

Estos algoritmos han demostrado robustez en tareas de clasificación complejas, como el reconocimiento de emociones en señales acústicas y la detección de patrones no lineales en datos multidimensionales.

2.5.2 Optimización en circuitos cuánticos

El entrenamiento de modelos cuánticos presenta desafíos particulares derivados de su estructura no determinista, incluyendo la dificultad de calcular derivadas exactas, la sensibilidad al ruido y la posible aparición de *barren plateaus*, es decir, zonas del modelo donde los cambios dejan de ser perceptibles y el aprendizaje prácticamente se detiene. Para hacer frente a estas limitaciones, se han desarrollado técnicas específicas para optimizar circuitos cuánticos variacionales (VQC).

Función de pérdida y función de costo

Al igual que en los modelos clásicos, los VQC requieren una función de pérdida para cuantificar la discrepancia entre la predicción y el valor objetivo. Las funciones más utilizadas incluyen:

- **Squared loss**: para regresión o codificación continua.
- **Cross-entropy loss**: para clasificación basada en probabilidades de medición.

En los modelos cuánticos, las salidas se obtienen midiendo observables como los operadores Pauli-Z aplicados a qubits específicos. El valor esperado de estas mediciones se calcula como:

$$|Z_i\rangle = \langle \Psi_{final} | Z_i | \Psi_{final} \rangle \quad (8)$$

A partir de ello, se obtiene un vector de predicciones del modelo:

$$\vec{z} = [\langle Z_0 \rangle, \langle Z_1 \rangle, \ldots, \langle Z_k \rangle] \quad (9)$$

Este vector puede usarse directamente como *logits* para funciones de pérdida como la entropía cruzada, o integrarse como entrada a capas densas clásicas en modelos híbridos.

Cálculo de gradientes: parameter-shift rule

Una de las estrategias más difundidas para el cálculo de gradientes en circuitos cuánticos parametrizados es la parameter-shift rule, que evita aproximaciones numéricas:

$$\frac{\partial J}{\partial \theta_i} = \frac{J(\theta_{\mathrm{i}} + \frac{\pi}{2}) - J(\theta_{\mathrm{i}} - \frac{\pi}{2})}{2} \quad (10)$$

Esta regla permite calcular gradientes exactos para compuertas unitarias parametrizadas (como $R_y(\theta)$) sin necesidad de aproximaciones numéricas, lo que la convierte en el estándar de facto en plataformas como PennyLane (Bergholm, y otros, 2022).

Gradientes cuánticos estimados por medición

El entrenamiento de VQC requiere calcular expectativas cuánticas mediante repetidas mediciones (shots) sobre los estados finales del circuito. Cuantas más mediciones se realizan, más precisa es la estimación del gradiente, pero a costa de mayor tiempo de cómputo y ruido experimental.

Algunos trabajos actuales utilizan técnicas de reducción de varianza, reparametrización y agrupación de observables para hacer que el entrenamiento sea más estable y eficiente (Mitarai et al., 2022).

Optimización híbrida cuántico-clásica

El entrenamiento de modelos VQC se realiza mediante un bucle híbrido: el backend cuántico ejecuta el circuito y proporciona los valores de la función de pérdida, mientras

que el *backend* clásico actualiza los parámetros mediante algoritmos de optimización. Entre los más utilizados se encuentran:

- Adam (*stochastic gradient-based*).
- Nelder-Mead, COBYLA (sin gradiente).
- Descenso por gradiente clásico.

El procedimiento continúa hasta que se cumple una condición de detención, como la siguiente:

$$|J(\theta^{t+1}) - J(\theta^{t})| < \varepsilon \tag{11}$$

Donde:

- $J(\theta)$ es la función de costo.
- t es la iteración actual.
- ε es un umbral pequeño (por ejemplo, 10^{-5}) que indica cuándo considerar que el modelo ha convergido.

Este criterio se interpreta como una condición de parada: si el valor de la función de costo deja de mejorar significativamente, el entrenamiento se detiene.

2.6 Métricas de evaluación en aprendizaje supervisado

2.6.1 Indicadores de evaluación: precisión, sensibilidad, F1-score, exactitud y matriz de errores

Para analizar el rendimiento de los modelos de clasificación supervisados se emplean diferentes indicadores que permiten medir qué tan bien se comporta el sistema. Entre ellos se encuentra la *precisión*, que refleja la proporción de aciertos dentro de los casos predichos como positivos; el *recall* o sensibilidad, que mide la capacidad del modelo para identificar correctamente los casos reales; el *F1-score*, que combina ambas medidas en un solo valor, y la exactitud, que indica el porcentaje total de predicciones correctas. Además, se emplea la matriz de confusión, una tabla que muestra de forma detallada los aciertos y errores del modelo en cada clase, permitiendo identificar con claridad dónde se producen las equivocaciones. En conjunto, estos indicadores ofrecen una visión clara y global del desempeño del modelo bajo distintos escenarios.

2.6.2 Métricas avanzadas: ROC-AUC y curvas *precision-recall*

El valor ROC-AUC es una métrica que resulta especialmente valiosa en situaciones donde una clase tiene mucha más presencia que la otra, como ocurre en problemas de clasificación binaria o modelos probabilísticos, que muestra la relación entre las proporciones positivas verdaderas y falsas y la precisión. Este indicador es útil cuando las clases están desbalanceadas.

Estas herramientas le permiten evaluar el poder discriminativo de su modelo sin depender de umbrales fijos, lo que puede ser útil para aplicaciones sensibles como la detección de emociones o estrés.

2.6.3 Evaluación en conjuntos de datos desbalanceados

Los desbalances en los conjuntos de datos son comunes en problemas prácticos, como la detección de emociones atípicas o comportamientos agresivos durante las interacciones verbales. En estos casos, las métricas como F1-score y ROC-AUC son más fiables que la exactitud.

Además, para compensar los desequilibrios y mejorar las capacidades de generalización del modelo, se utilizaron métodos de remuestreo (*oversampling*, *undersampling*) y ponderación de clases (He & Garcia, 2009).

2.6.4 Comparación de desempeño entre modelos clásicos y cuánticos

Los modelos cuánticos, especialmente los VQCs, han demostrado en investigaciones recientes la capacidad prometedora de aproximar las funciones de clasificación con conjuntos de datos pequeños o medianos. Aunque los modelos clásicos siguen siendo más fiables y fácilmente escalables, los modelos cuánticos constituyen una alternativa a los complejos espacios de búsqueda.

Las comparaciones empíricas muestran que algunas tareas de control de calidad pueden igualar o superar los modelos densos clásicos en términos de precisión y generalidad, especialmente cuando se integran en esquemas mixtos y utilizan indicadores de evaluación apropiados (Abbas, y otros, 2021).

2.7 Convergencia interdisciplinaria en análisis fonoconductual

2.7.1 Integración de IA, neurociencia y computación cuántica

El análisis del comportamiento del habla está en el punto de convergencia de varias disciplinas avanzadas. Por un lado, la neurociencia proporciona una base para comprender cómo el cerebro humano produce y responde a las señales del habla, incluyendo aspectos como la prosodia, el ritmo y las oscilaciones emocionales. Estas respuestas están estrechamente relacionadas con el sistema límbico y los procesos cognitivos, haciendo del sonido un canal privilegiado para el razonamiento conductual (Cummins, y otros, 2015).

La inteligencia artificial, por otra parte, y el aprendizaje controlado en particular, pueden clasificar eficientemente los modelos acústicos y afectivos a partir de datos del habla. Al combinar estos algoritmos con principios neurobiológicos, se han logrado avances significativos en la detección de tareas como el estado de ánimo, el estrés o la fatiga cognitiva.

Por último, la computación cuántica introduce nuevos métodos de modelado con ayuda de subcircuitos variables que prometen una representación más rica de los espacios de características del habla y una optimización más eficiente bajo ciertas condiciones. Este enfoque híbrido permite explorar dimensiones analíticas no disponibles en los métodos clásicos, mejorando la capacidad de caracterizar el comportamiento a partir del sonido (Schuld & Killoran, 2019).

2.7.2 Vacíos y desafíos en la literatura científica actual

A pesar de los progresos realizados, hay lagunas importantes. Muchos estudios siguen centrándose en entornos controlados o simulados, lo que limita su aplicabilidad en entornos del mundo real como los puntos de contacto. Además, la mayoría de los métodos clásicos tienen limitaciones en términos de adaptabilidad cultural, multilingüismo y resistencia al ruido o a la distorsión (Poria et al, 2020).

En el ámbito cuántico, hay más desafíos por delante: desde las limitaciones de la escalabilidad del hardware hasta la falta de conjuntos de datos adecuados para el

aprendizaje clásico híbrido cuántico. También hay lagunas en los métodos replicables y las métricas estandarizadas al comparar la eficiencia cuántica con modelos tradicionales.

2.7.3 Aportes diferenciales del presente estudio

Esta investigación proporciona una integración única de aprendizaje controlado con esquemas de variación cuántica para la caracterización conductual del habla en escenarios del mundo real. Fue diseñado para capturar patrones de comportamiento complejos utilizando una arquitectura híbrida entrenada en datos del mundo real de las interacciones del contact center. Se espera que este enfoque aporte:

- Mayor detección de señales conductuales complejas como mentir, evitar o estrés.
- La eficiencia computacional al clasificar tareas en conjuntos de datos intermedios.
- Fundamentos replicables que combinan ideas de neurociencia con modelado cuántico, ayudando a desarrollar herramientas avanzadas para aplicar la IA afectiva.

2.8 Plataforma tecnológica para el procesamiento avanzado de voz en centros de contacto

El uso de tecnologías avanzadas para analizar la voz en los centros de contacto ha ido transformándose notablemente en los últimos años, impulsado tanto por la investigación académica como por avances aplicados en la industria. La caracterización conductual mediante inteligencia artificial —incluyendo la detección de emociones, veracidad, estrés y perfil del hablante— requiere una arquitectura capaz de manejar señales acústicas en tiempo real, escalar ante altos volúmenes de interacciones, y facilitar la integración con modelos de aprendizaje supervisado, incluyendo aquellos de tipo híbrido cuántico-clásico.

En el estado del arte actual, se observa una fuerte adopción de arquitecturas distribuidas y plataformas de computación en la nube, particularmente aquellas que soportan enfoques *serverless* y *event-driven*, como como AWS, Azure y GCP. Estas soluciones brindan componentes modulares que permiten construir pipelines de procesamiento de voz sin necesidad de gestionar servidores físicos, lo cual reduce la complejidad operativa y mejora la escalabilidad.

Una arquitectura típica en soluciones modernas incluye:

- Captura de interacciones mediante un software de centro de contacto (por ejemplo, DINOMI, Amazon Connect, Genesys Cloud o Twilio), que graban y transcriben las llamadas.
- Puertas de enlace API (API Gateway) que actúan como interfaz segura entre los servicios locales y la nube.
- Funciones serverless (como AWS Lambda o Google Cloud Functions) que ejecutan tareas como preprocesamiento acústico, extracción de características fonéticas (MFCC, pitch, energía) y envío de datos a modelos de inferencia.
- Orquestadores y almacenamiento distribuido (por ejemplo, Amazon S3, Azure Blob, Step Functions, EventBridge), que gestionan flujos asíncronos y persistencia temporal de datos.

El diseño sin servidor permite una respuesta adaptativa ante la variabilidad del tráfico de llamadas, facilitando la integración con modelos de machine learning que operan en tiempo real o diferido. Además, muchas soluciones actuales están comenzando a incorporar componentes de aprendizaje cuántico y modelos híbridos, lo cual exige compatibilidad con servicios de simulación cuántica o backends como IBM Q o Amazon Braket.

En resumen, este capítulo proporciona un marco interdisciplinario que integra fundamentos teóricos, avances tecnológicos y enfoques recientes en el análisis textual y acústico del habla mediante inteligencia artificial y computación cuántica. Estos elementos forman la base conceptual y metodológica sobre la que se realizó esta investigación.

En el siguiente capítulo se describe de manera integral cómo se concibió, construyó y puso en funcionamiento el clasificador híbrido, abarcando desde la recolección y procesamiento de los audios hasta la integración funcional de sus componentes. Se detalla la arquitectura de la red neuronal totalmente conectada, la configuración del circuito cuántico variacional (VQC), así como los experimentos realizados y las métricas de evaluación obtenidas en un entorno realista.

CAPÍTULO 3

3 DISEÑO E IMPLEMENTACIÓN

El propósito de este capítulo es explicar en forma detallada cómo se construyó y puso en marcha la solución propuesta para la caracterización conductual en interacciones orales, con énfasis en la detección de veracidad, niveles de estrés e identificación del hablante. Se describen las etapas de análisis de los datos, el desarrollo de los prototipos, la infraestructura tecnológica utilizada, las herramientas de visualización y los criterios empleados para medir el rendimiento de los modelos. Además, se presenta el diseño experimental adoptado para la recolección de audios de verdades y mentiras, adaptado del protocolo validado internacionalmente MU3D, con el fin de generar una base local confiable y supervisada. Esta fase es el fundamento para una posterior comparación de los los enfoques clásicos y cuánticos desarrollados en la investigación.

3.1 Análisis y verificación de los datos y sus orígenes

Se emplearon dos fuentes de datos complementarias para abordar distintas dimensiones de la caracterización conductual. Para la clasificación de emociones, se dispone de un conjunto de aproximadamente 6500 interacciones orales reales recolectadas de un centro de contacto, las cuales incluyen tanto el audio como su transcripción asociada. Estas grabaciones reflejan contextos diversos de atención al cliente en idioma español, permitiendo capturar variabilidad emocional genuina en escenarios operativos. Además, para la tarea de detección de veracidad y engaño, se desarrolló un experimento controlado que generó un total de 600 audios etiquetados según el grado de veracidad percibida. Este segundo conjunto fue construido con el objetivo de disponer de datos balanceados y etiquetados con claridad, bajo condiciones que permitieran identificar patrones fonéticos y prosódicos vinculados a la conducta engañosa.

3.1.1 Caracterización del dataset multimodal empleado en la clasificación de emociones

Para la clasificación de emociones, se dispone de un dataset multimodal de 6500 interacciones orales en español, recolectadas mediante audios grabados en centros de contacto, las cuales incluyen tanto el audio como su transcripción asociada.

Organización de las etiquetas y distribución de datos

El dataset contiene interacciones orales grabadas y etiquetadas con una de cuatro categorías emocionales principales: felicidad, enojo, tristeza y calma. Estas etiquetas fueron asignadas manualmente con base en el contenido semántico y tono emocional de las transcripciones. El etiquetado se organizó en un archivo CSV que incluye las siguientes columnas:

- **archivo**: nombre del archivo de audio (.wav)
- **participante**: identificador del sujeto (ej. P01, P02…)
- **emoción**: etiqueta emocional (positiva, negativa, neutra)
- **transcripción**: texto asociado al contenido del audio

El balance de clases para este dataset fue validado a través de conteo de ocurrencias por categoría, asegurando una distribución lo más equitativa posible entre las emociones, con el fin de evitar sesgos durante el entrenamiento del modelo. La Figura 3.1 muestra esta distribución.

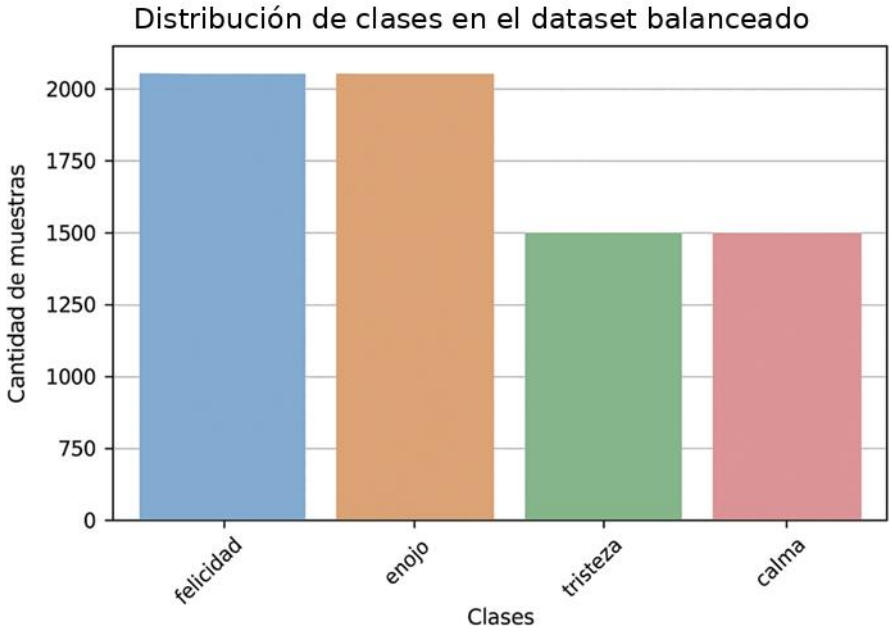


Figura 3.1 Distribución de clases emocionales en el dataset multimodal

Adicionalmente, se realizó una reducción de dimensionalidad usando PCA para visualizar cómo se distribuyen los datos emocionales en un espacio de dos dimensiones, tal como lo muestra la Figura 3.2, con el objetivo de explorar posibles patrones de separación entre clases.

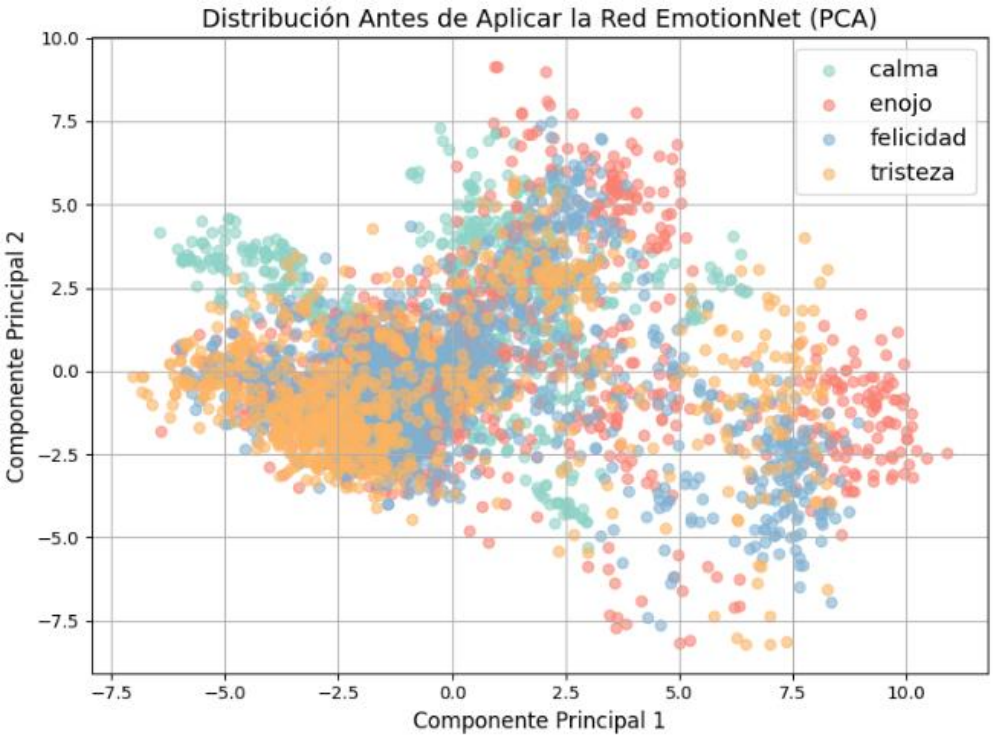


Figura 3.2 Dispersión de las emociones en el espacio bidimensional (PCA)

Las diferentes emociones no presentan una separación lineal evidente, lo que sugiere la necesidad de emplear alternativas más avanzadas, como modelos neuronales de varias capas, capaces de identificar fronteras no lineales entre clases superpuestas.

3.1.2 Descripción del dataset multimodal para la clasificación de veracidad/mentira

El dataset para la clasificación de veracidad/engaño es producto de un experimento controlado y está compuesto de 600 grabaciones de audios (300 de verdad y 300 de mentira). Cada participante grabó dos muestras para cada clase (dos de verdad y dos de mentira). Cada grabación de audio (modalidad acústica) tiene su correspondiente transcripción (modalidad textual) y un conjunto de etiquetas supervisadas que representan la veracidad o engaño del discurso.

Esta estructura multimodal permite integrar información fonética y semántica para una caracterización más robusta de las interacciones humanas. Las transcripciones se generaron de forma automática y fueron posteriormente corregidas para evitar sesgos semánticos

Estructura de etiquetas y balance de clases

Cada muestra está identificada por un nombre único de archivo (audio_ID.wav), e incluye metadatos asociados al participante (edad, género). Las columnas relevantes del dataset resultante son:

- **archivo**: nombre del archivo de audio
- **participante**: identificador del sujeto
- **veracidad**: clase asignada (verdad o mentira)
- **Edad**: grupo etario: Adulto joven (21 a 35 años), Adulto maduro (35 a 55 años), Adulto mayor (56 a 70 años).
- **Género**: masculino o femenino
- **transcripción**: texto narrado

Dado que cada participante grabó una muestra para cada clase (una verdad y una mentira), se asegura un balance perfecto entre las clases de veracidad, lo que favorece un entrenamiento justo y evita desbalance en el aprendizaje. La distribución del dataset utilizado para la clasificación de veracidad es mostrada en la Figura 3.3.

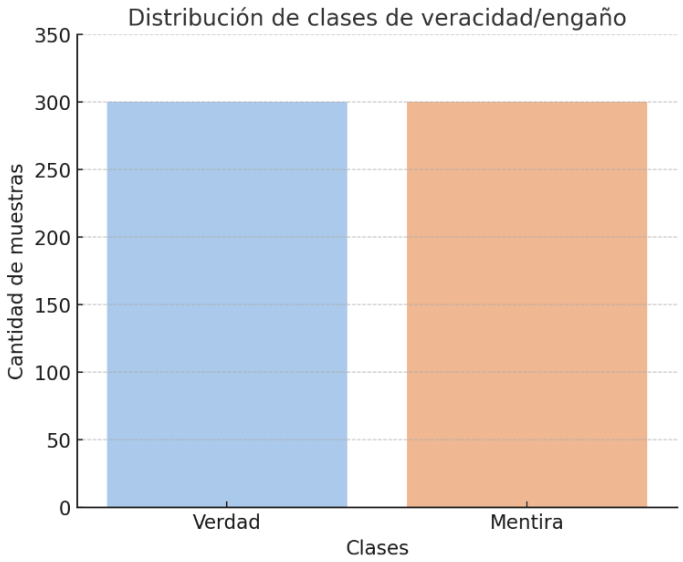


Figura 3.3 Distribución de clases del dataset para veracidad/engaño

Al estar balanceadas ambas estructuras de etiquetado, permiten entrenar, validar y evaluar los clasificadores con datos coherentes, trazables y representativos de los fenómenos conductuales que se desean modelar.

Adicionalmente, se realizó una reducción de dimensionalidad usando PCA para visualizar cómo se distribuyen los datos emocionales en un espacio de dos dimensiones, tal como lo muestra la Figura 3.4, con el objetivo de explorar posibles patrones de separación entre clases.

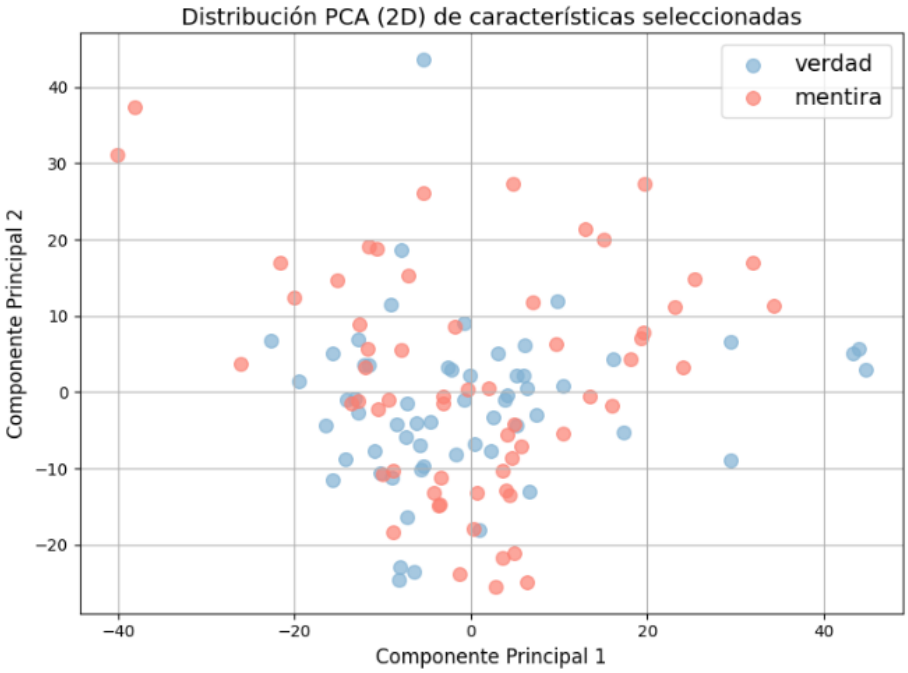


Figura 3.4 Dispersión de veracidad en el espacio bidimensional (PCA)

Las clases (verdad/mentira) no presentan una separación lineal evidente, lo que sugiere la necesidad de emplear alternativas más avanzadas, como modelos neuronales de varias capas, capaces de identificar fronteras no lineales entre clases superpuestas.

3.1.3 Validación y limpieza de datos

Con el objetivo de garantizar la calidad del dataset y la integridad de los análisis posteriores, se realizó un procedimiento riguroso de control y depuración de los audios recolectados. La validación inicial incluyó la verificación de parámetros estructurales como el formato del archivo (.wav a 16 kHz, mono), la presencia de etiquetas asociadas y la correspondencia entre cada archivo de audio y su transcripción. Posteriormente, se aplicaron filtros automáticos y manuales para depurar los registros no aptos para el análisis.

El primer criterio de limpieza consistió en eliminar todos los audios cuya duración total fuera inferior a 30 segundos, ya que este tipo de registros presentaban una escasa

densidad informativa y resultaban poco representativos para que los modelos puedan aprender a partir de estos datos. El segundo criterio implicó la detección de pausas prolongadas o silencios superiores a 5 segundos dentro del audio. Para ello, se utilizó una función de análisis de envolvente de amplitud combinada con umbrales de energía acústica, implementada en Python con la biblioteca *librosa*. Esta técnica permitió identificar segmentos inactivos dentro de cada archivo, descartando aquellos en los que la suma de intervalos de silencio excedía el umbral permitido.

Los audios descartados por estas razones fueron etiquetados y documentados para trazabilidad. Como resultado de esta etapa, se obtuvo un subconjunto limpio y consistente de datos, con audios de duración adecuada, sin silencios extensos y correctamente etiquetados. En la Figura 3.5 se resume gráficamente el proceso de verificación y depuración de los datos implementado en esta investigación.

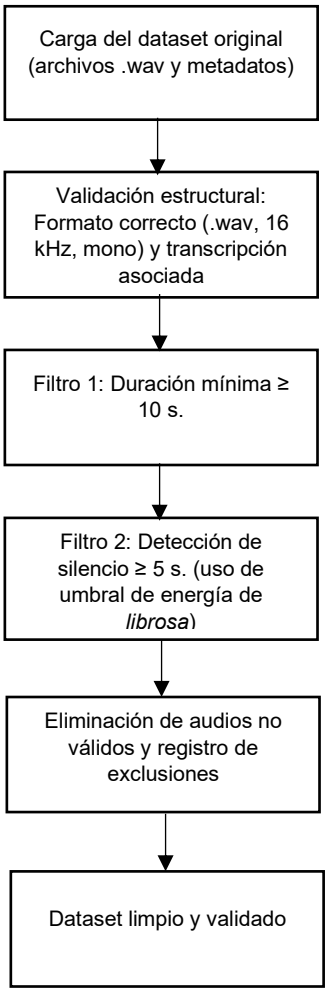


Figura 3.5 Diagrama del proceso de verificación y depuración de los datos

El flujo anterior detalla el proceso desde la carga inicial de los archivos .wav y sus metadatos, hasta la aplicación de filtros de duración mínima y detección de silencios prolongados utilizando la biblioteca *librosa*, culminando con la conformación del dataset limpio y validado que será utilizado para los experimentos posteriores.

3.1.4 Extracción de atributos multimodales

Para las tareas de clasificación de emociones y detección de veracidad en interacciones orales, se aplicaron técnicas de extracción de características tanto acústicas como textuales. En ambos casos, se emplearon los mismos algoritmos y procedimientos de procesamiento para garantizar consistencia metodológica en la representación multimodal de los datos. Estas características constituyen la base de entrada para los modelos de aprendizaje supervisado que se detallan más adelante.

Extracción de atributos acústicos

Para la clasificación de emociones, se obtuvieron un total de 1183 atributos acústicos por cada audio, derivados de transformaciones como los descriptores cepstrales en la escala Mel (MFCC), características cromáticas, propiedades espectrales (centroide, roll-off, bandwidth), energía, y medidas prosódicas como pitch, jitter y shimmer. A cada una de estas características se le aplicaron descriptores estadísticos (media, desviación estándar, skewness, curtosis, etc.), lo que permitió condensar la información temporal del audio en un vector fijo.

La extracción se realizó utilizando bibliotecas como *librosa*, ampliamente aceptadas en la literatura para tareas de reconocimiento de emociones en voz. Las características acústicas extraídas (agrupadas por tipo) son:

- **Prosódicas**: Se calcularon 11 características prosódicas para cada audio. Capturan aspectos del ritmo, tono, intensidad y duración del habla. Son claves para detectar emociones, estrés y patrones de veracidad.
- **Expectrales (escala de Mel)**: Se obtuvieron 148 características espectrales. Extraen información del timbre vocal usando transformaciones de frecuencia perceptual. Representan cómo percibe el oído humano las diferentes frecuencias del habla.

- **Wav2Vec**: Se calcularon 768 *embeddings* ricos a partir de ondas de voz crudas. Captura contenido fonético, contexto y estructura lingüística.
- **Resemblyzer**: Genera un vector compacto de 256 características que representa la identidad única del hablante. Se entrena para distinguir voces incluso en condiciones de ruido o variación emocional.

Estos cálculos producen, en total, 1183 valores por cada audio, que resumen distintos aspectos espectrales, prosódicos y cepstrales del habla.
La tabla 3.1 muestra la cantidad, tipo, la información que captura.

Tabla 3.1 Tipo y cantidad de características acústicas extraídas

Tipo de características	Cantidad de características	Información que capturan
Prosódicas	**11**	Variaciones en tono, energía, pausas, velocidad… indicadores emocionales o de estrés
Espectrales	**148**	Timbre y resonancia de la voz, útiles para detectar matices emocionales y patrones no naturales
Wav2Vec	**768**	Representaciones profundas del contenido fonético + contexto (auto-supervisado)
Resemblyzer	**256**	Vector de identidad vocal, también puede aportar indicios emocionales sutiles

La tabla 3.2 muestra las características que fueron usadas por tipo clasificación

Tabla 3.2 Características acústicas por tipo de clasificación

Clasificación	Tipo de Modelo	Tipo de características
Emociones	Clásico	Prosódica, Espectral, Wav2vec, Resemblyzer
Emociones	Cuántico	Prosódica, Espectral, Wav2vec, Resemblyzer
Veracidad	Clásico	Prosódica, Espectral, Wav2vec, Resemblyzer
Veracidad	Cuántico	Prosódica, Espectral, Wav2vec, Resemblyzer
Identificación del hablante	N/A	Resemblyzer

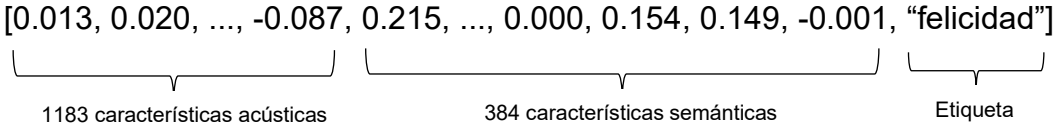
No todas las características acústicas fueron usadas para todos los tipos de clasificación. En el caso de la identificación del hablante solo fue necesario usar *Resemblyzer* ya que son necesitábamos obtener, almacenar y comparar la huella acústica del sujeto.

Extracción de atributos textuales: Embeddings semánticos mediante SBERT

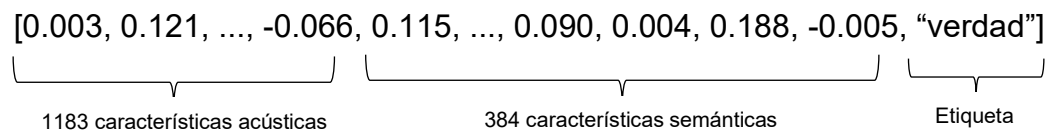
Para representar la información lingüística contenida en las transcripciones de los audios, se empleó la técnica de embeddings semánticos a nivel de frase utilizando el modelo preentrenado Sentence-BERT (SBERT). Esta técnica permitió convertir cada transcripción en un vector numérico denso que preservó el significado global de la oración, capturando matices contextuales como emoción, ironía o veracidad implícita. SBERT extiende la arquitectura original de BERT optimizando su uso para tareas de clasificación y similitud semántica mediante un esquema de mean pooling aplicado sobre las salidas del transformador. En este proyecto, cada transcripción fue procesada como una oración completa y transformada en un vector de 384 dimensiones utilizando el modelo *all-MiniLM-L6-v2*, ampliamente reconocido por ofrecer un buen equilibrio entre calidad de resultados y rapidez de procesamiento. Este vector semántico resultante se concatenó posteriormente con las características acústicas, dando lugar a una representación multimodal unificada por cada interacción.

Integración del vector multimodal

Una vez obtenidas las características acústicas (1183 dimensiones) y los *embeddings* semánticos a partir de las transcripciones (384 dimensiones), se procedió a la construcción del vector multimodal que consolidó ambas fuentes de información en una única representación de 1623 dimensiones por registro. Esta fusión permitió capturar tanto los matices prosódicos del habla como la carga semántica del contenido textual, favoreciendo una clasificación más robusta de los estados emocionales y de veracidad en las interacciones orales. Por ejemplo, un registro del dataset integrado puede representarse así:



O así (para veracidad):



3.2 Algoritmos y Modelos para la Clasificación Multimodal

Se describe cómo se construyeron y pusieron en funcionamiento los algoritmos utilizados en ambas tareas de clasificación: emociones y veracidad. Para cada tarea se desarrollaron dos enfoques: un modelo clásico basado en redes neuronales totalmente conectadas y un modelo cuántico variacional (VQC). A continuación, se detallan sus arquitecturas, procesos de entrenamiento, validación y optimización.

3.2.1 Clasificación de emociones

3.2.1.1 Modelo Clásico: EmotionNet

Arquitectura de red neuronal (capas, activaciones, dropout)

Para abordar la tarea de clasificación emocional, se implementó un modelo clásico basado en una red neuronal totalmente conectada, denominada EmotionNet. Esta arquitectura acepta como entrada un vector de 1567 características (correspondientes a la concatenación de 1183 características acústicas y 384 embeddings textuales), y sigue una estructura multicapa que equilibra profundidad y regularización. La arquitectura está compuesta por:

- *Capa de entrada*: 1567 nodos (una por cada característica del vector multimodal).
- *Capa densa 1*: 256 unidades + función de activación ReLU + Dropout (p = 0.3).
- *Capa densa 2*: 128 unidades + ReLU + Dropout (p = 0.3).
- *Capa densa 3*: 64 unidades + ReLU.
- *Capa de salida*: 4 unidades (una por cada clase emocional: felicidad, tristeza, calma, enojo).

La Figura 3.6 presenta la arquitectura final de la red neuronal EmotionNet con vectores de entrada de 1567 características y salida en 4 emociones distintas.

```
EmotionNet(
  (fc1): Linear(in_features=1567, out_features=512, bias=True)
  (dropout1): Dropout(p=0.3, inplace=False)
  (fc2): Linear(in_features=512, out_features=256, bias=True)
  (dropout2): Dropout(p=0.3, inplace=False)
  (fc3): Linear(in_features=256, out_features=128, bias=True)
  (dropout3): Dropout(p=0.3, inplace=False)
  (fc4): Linear(in_features=128, out_features=64, bias=True)
  (output): Linear(in_features=64, out_features=4, bias=True)
)
```

Figura 3.6 Arquitectura red EmotionNet

Este diseño busca un balance entre capacidad de representación y control del sobreajuste, con regularización aplicada en capas intermedias mediante Dropout.

Configuración de entrenamiento: función de pérdida, optimizador, hiperparámetros

La red fue entrenada utilizando el algoritmo *Backpropagation*, junto con el optimizador Adam, conocido por su adaptabilidad en entornos no estacionarios. La configuración final de entrenamiento fue la siguiente:

- *Función de pérdida utilizada*: Entropía cruzada para clasificación multiclase.
- *Épocas*: 100
- *Tamaño del batch*: 32
- *Optimizador utilizado*: Adam
- *Tasa de aprendizaje utilizada*: 0.001
- *Regularización*: Dropout en capas ocultas
- *Métricas monitorizadas*: Precisión (*accuracy*) y Recall y pérdida en conjunto de validación

Esta configuración fue seleccionada mediante experimentación iterativa, evaluando el comportamiento del modelo usando el conjunto de validación.

Validación e indicadores de rendimiento

Para asegurar la generalización del modelo, se aplicó validación cruzada estratificada (k-fold), permitiendo medir el rendimiento promedio del clasificador sobre múltiples particiones de los datos. Las métricas evaluadas incluyeron:

- *Accuracy* promedio por clase
- Matriz de confusión
- F1-Score macro
- Curvas de pérdida y precisión por época

Estos indicadores permitieron comparar el desempeño entre clases y detectar posibles sesgos o desequilibrios en la predicción de emociones.

La Figura 3.7 ilustra el pipeline completo del modelo clásico EmotionNet para la clasificación de emociones.

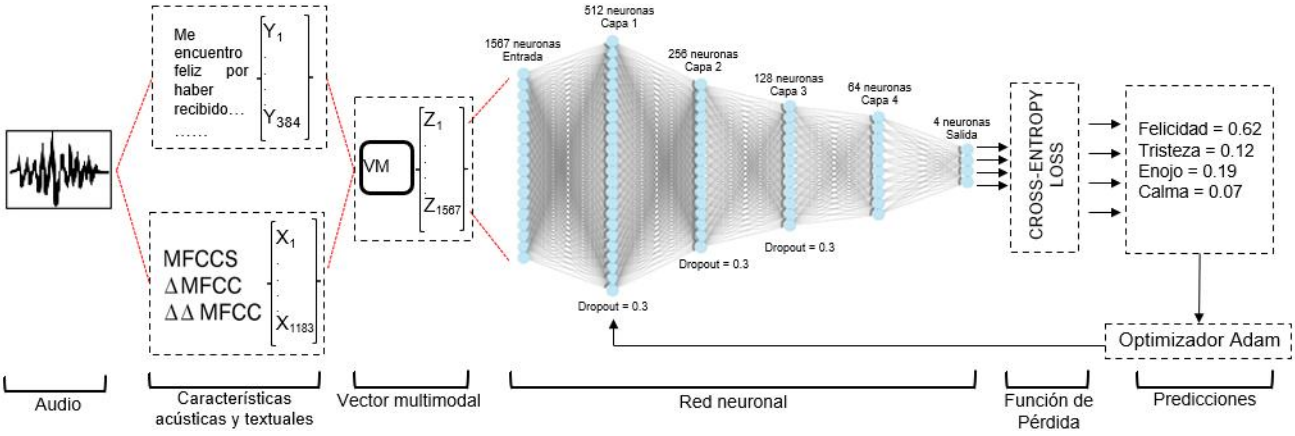


Figura 3.7 Pipeline clásico del modelo EmotionNet

Este flujo de datos inicia con la entrada de audio y va hasta la obtención de resultados, incorporando también ajustes al modelo mediante retropropagación y función de pérdida tipo entropía cruzada.

3.2.1.2 Modelo Cuántico: VQC

Selección y reducción de características con red densa

Al revisar el gráfico de dispersión de la Figura 3.8 se puede notar la no linealidad, complejidad y poca separación de los datos y dado que los circuitos cuánticos actuales requieren entradas de baja dimensionalidad, se diseñó una red neuronal densa para reducir el vector multimodal de 1567 características (1183 acústicas + 384 textuales) a un *embedding* compacto de 16 dimensiones. Esta transformación se logró entrenando una red lineal intermedia, la cual fue ajustada junto al modelo híbrido para preservar la representatividad de las variables originales (Abbas, y otros, 2021). La red resultante se muestra a continuación:

```
EmotionNet(
  (fc1): Linear(in_features=1567, out_features=512, bias=True)
  (dropout1): Dropout(p=0.3, inplace=False)
  (fc2): Linear(in_features=512, out_features=256, bias=True)
  (dropout2): Dropout(p=0.3, inplace=False)
  (fc3): Linear(in_features=256, out_features=128, bias=True)
  (dropout3): Dropout(p=0.3, inplace=False)
  (fc4): Linear(in_features=128, out_features=64, bias=True)
  (output): Linear(in_features=64, out_features=16, bias=True)
)
```

Figura 3.8 Arquitectura de red densa para reducción de dimensionalidad

En la Figura 3.9 se ilustra cómo se agrupan las muestras en un espacio bidimensional generado mediante PCA a partir de las 16 características seleccionadas.

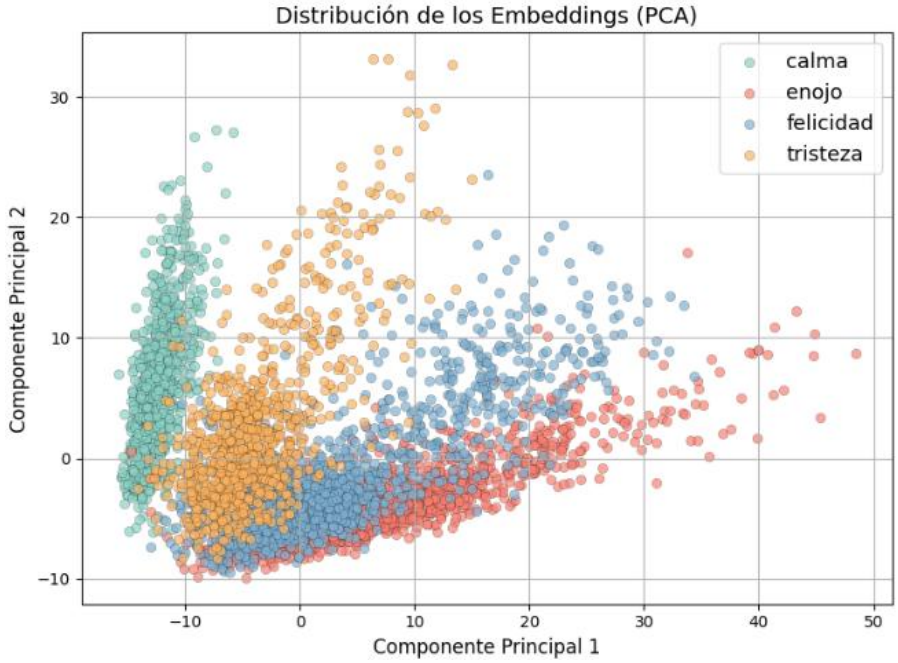


Figura 3.9 Distribución PCA de características acústicas reducidas

Se evidencia una mejor separación entre clases, aunque conserva una estructura no lineal, lo que sugiere la necesidad de utilizar algoritmos con capacidad de modelado no lineal, como los circuitos cuánticos variacionales.

La Figura 3.10 muestra el pipeline para adaptar el vector multimodal de 1567 a 16 dimensiones, que es el tamaño compatible con el circuito cuántico ansatz personalizado limitado a 4 qubits ($2^4 = 16$).

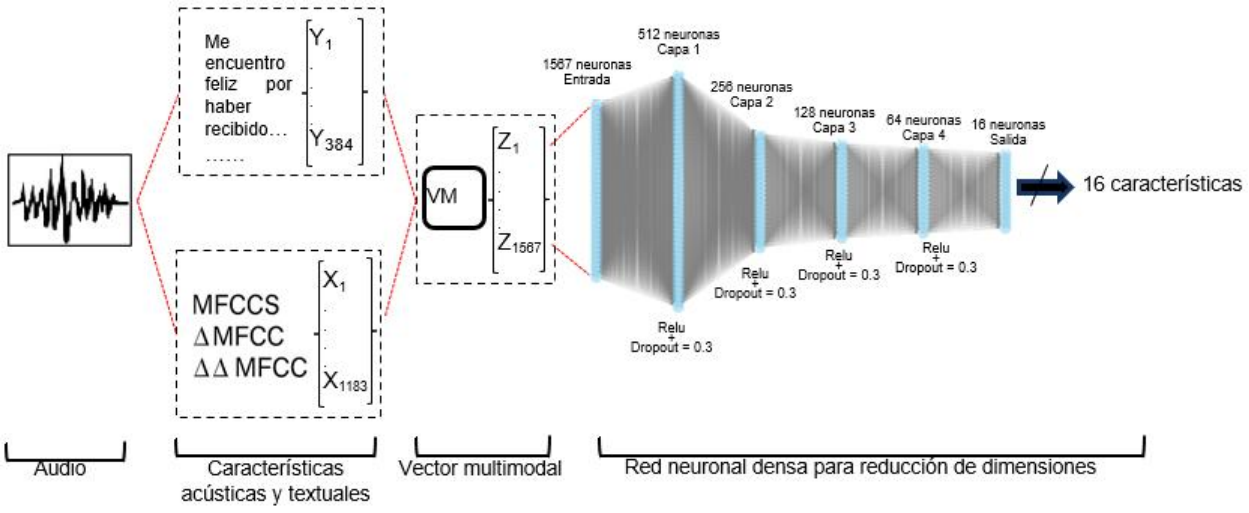


Figura 3.10 Pipeline para la reducción de características multimodales

Como se observa en la figura anterior, la red emplea múltiples capas ocultas con regularización mediante *dropout* para evitar el sobreajuste durante el proceso de compresión. El vector reducido obtenido se utiliza posteriormente como entrada del bloque de codificación cuántica basado en *AmplitudeEmbedding*, permitiendo su procesamiento eficiente en el clasificador VQC.

Codificación de entrada y arquitectura del circuito cuántico

Las 16 características reducidas se normalizaron según la norma L2 y se utilizaron como amplitudes de un estado cuántico mediante la codificación *AmplitudeEmbedding*. Esta técnica, ampliamente utilizada en la literatura reciente, permite aprovechar la alta densidad informativa al mapear un vector real normalizado en un espacio de dimensión 2^n, donde $n = \log_2 N$ con $N = 16$.

El circuito cuántico fue construido con una arquitectura personalizada basada en codificación por *AmplitudeEmbedding* y capas variacionales. En cada capa, se aplican rotaciones R_y a cada qubit, seguidas de un esquema de entrelazamiento circular mediante compuertas CNOT. Este diseño, ajustado específicamente para representar eficazmente estados acústicos reducidos, fue replicado a lo largo de seis capas entrenables. En la Figura 3.11 se ilustra el diseño del circuito cuántico variacional.

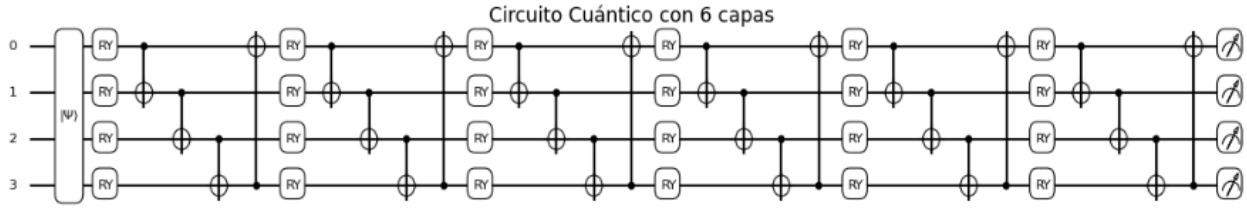


Figura 3.11 Diseño del circuito cuántico para la clasificación de emociones

El estado de salida del clasificador cuántico variacional puede expresarse de manera compacta mediante la composición de capas de rotaciones y entrelazamiento.

$$|\Psi_{out}\rangle = \left(\prod_{l=1}^{L} E^{(l)} \ \bigotimes_{q=0}^{3} \ R_y(\theta_q^{(l)}) \right) H^{\otimes 4} |0000\rangle \qquad (12)$$

En esta expresión, $H^{\otimes 4}$ genera la superposición inicial de los cuatro qubits, $R_y(\theta_q^{(l)})$ representa las rotaciones parametrizadas en la capa l, y $E^{(l)}$ corresponde al bloque de entrelazamiento implementado mediante compuertas CNOT. El producto

$\prod_{l=1}^{L} E^{(l)} \bigotimes_{q=0}^{3} R_y(\theta_q^{(l)})$ refleja la naturaleza repetitiva de la arquitectura VQC, estructurada en capas alternadas de rotaciones y entrelazamiento.

Configuración de entrenamiento: qubits, capas, optimizador

El modelo se construyó con 4 qubits, los cuales permiten representar hasta 16 amplitudes. Se emplearon 6 capas de entrelazamiento, y el entrenamiento se llevó a cabo usando descenso por gradiente con una tasa de aprendizaje de 0.001. Dado que la simulación cuántica se realiza en entornos clásicos, el entrenamiento se ejecutó en *batch* completo por observación, simulando la evolución del circuito sobre el dataset reducido.

Regla del desplazamiento y salida del circuito cuántico

La actualización de los parámetros del circuito se realizó utilizando la regla del desplazamiento (parameter shift rule), una técnica derivada analíticamente para calcular el gradiente en circuitos cuánticos diferenciales. La salida del circuito, correspondiente a las expectativas de los operadores Pauli-Z, fue entregada directamente como logits a la función de error de entropía cruzada multiclase (Cross-Entropy Loss), que internamente aplica softmax durante el cálculo del error, permitiendo retropropagación hasta los parámetros cuánticos.

Validación y análisis de desempeño

Se utilizó la misma estrategia de evaluación utilizada en el modelo clásico, manteniendo consistencia experimental. Se monitorearon métricas como accuracy y F1-score por clase para comparar el desempeño con el modelo EmotionNet. El entrenamiento del VQC fue más costoso computacionalmente, pero logró capturar patrones complejos en espacios reducidos, mostrando el potencial de la computación cuántica en tareas de clasificación multimodal.

La Figura 3.12 muestra el pipeline completo de entrenamiento a partir de las entradas cuánticas codificadas.

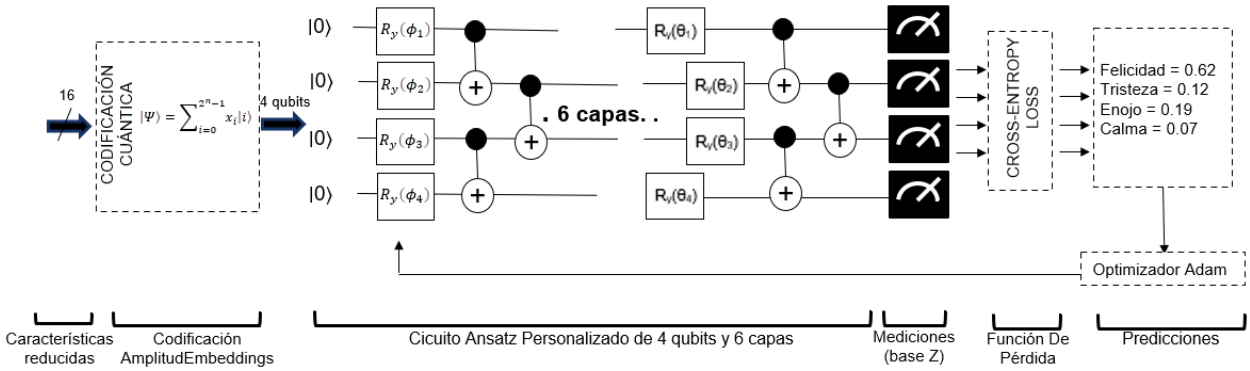


Figura 3.12 Pipeline cuántico (Emociones)

Este diseño permite capturar relaciones no lineales en los datos acústicos reducidos mediante la codificación de amplitudes y el entrelazamiento cuántico, lo que amplía el espacio de representación de las características. La combinación de capas rotacionales y compuertas CNOT en el ansatz personalizado ofrece un alto grado de expresividad, mientras que la optimización híbrida con Adam ajusta los parámetros del circuito para maximizar la capacidad de clasificación.

3.2.2 Clasificación de veracidad

3.2.2.1 Modelo Clásico: TruthNet

Arquitectura de red neuronal (capas, activaciones, dropout)

Para abordar la tarea de clasificación emocional, se implementó un modelo clásico basado en una red neuronal totalmente conectada, denominada EmotionNet. Esta arquitectura acepta como entrada un vector de 1567 características (correspondientes a la concatenación de características acústicas y embeddings textuales), y sigue una estructura multicapa que equilibra profundidad y regularización. La arquitectura está compuesta por:

- *Capa de entrada*: 1567 nodos (una por cada característica del vector multimodal).
- *Capa densa 1*: 256 unidades + función de activación ReLU + Dropout (p = 0.3).
- *Capa densa 2*: 128 unidades + ReLU + Dropout (p = 0.3).
- *Capa densa 3*: 64 unidades + ReLU. + Dropout (p = 0.3).
- *Capa densa 4*: 32 unidades + ReLU. + Dropout (p = 0.3).
- *Capa de salida*: 2 unidades (una para la clase verdad y otra para la clase mentira).

La Figura 3.13 presenta la arquitectura final de la red neuronal TruthNet con vectores de entrada de 1567 características y salida en 2 clases distintas.

```
TruthNet(
  (fc1): Linear(in_features=1567, out_features=256, bias=True)
  (dropout1): Dropout(p=0.3, inplace=False)
  (fc2): Linear(in_features=256, out_features=128, bias=True)
  (dropout2): Dropout(p=0.3, inplace=False)
  (fc3): Linear(in_features=128, out_features=64, bias=True)
  (dropout3): Dropout(p=0.3, inplace=False)
  (fc4): Linear(in_features=64, out_features=32, bias=True)
  (dropout4): Dropout(p=0.3, inplace=False)
  (output): Linear(in_features=32, out_features=2, bias=True)
)
```

Figura 3.13 Arquitectura red densa TruthNet

Configuración de entrenamiento: función de pérdida, optimizador, hiperparámetros

La red fue entrenada utilizando el algoritmo Backpropagation, junto con el optimizador Adam, conocido por su adaptabilidad en entornos no estacionarios. La configuración final de entrenamiento fue la siguiente:

- *Función de pérdida utilizada*: Entropía cruzada multiclase.
- *Épocas*: 50
- *Tamaño del batch*: 32
- *Early Stopping:* en 50
- *Optimizador utilizado*: Adam
- *Tasa de aprendizaje utilizada*: $1e^{-4}$
- *Regularización*: Dropout = 0.3 en capas ocultas
- *Métricas monitorizadas*: Precisión (*accuracy*) y pérdida en conjunto de validación

Validación y métricas

Para asegurar la generalización del modelo, se aplicó validación cruzada estratificada (k-fold), permitiendo medir el rendimiento promedio del clasificador sobre múltiples particiones de los datos. Las métricas evaluadas incluyeron:

- *Accuracy* promedio por clase
- Matriz de confusión
- F1-Score macro
- Curvas de pérdida y precisión por época

Estos indicadores permitieron comparar el desempeño entre clases. El desempeño final evaluado en el conjunto de prueba registró un buen equilibrio entre la detección de "verdad" y "mentira".

Esta configuración fue seleccionada mediante experimentación iterativa, evaluando el desempeño del modelo sobre datos de validación. La Figura 3.14 muestra la evolución del comportamiento del error durante el aprendizaje y la validación.

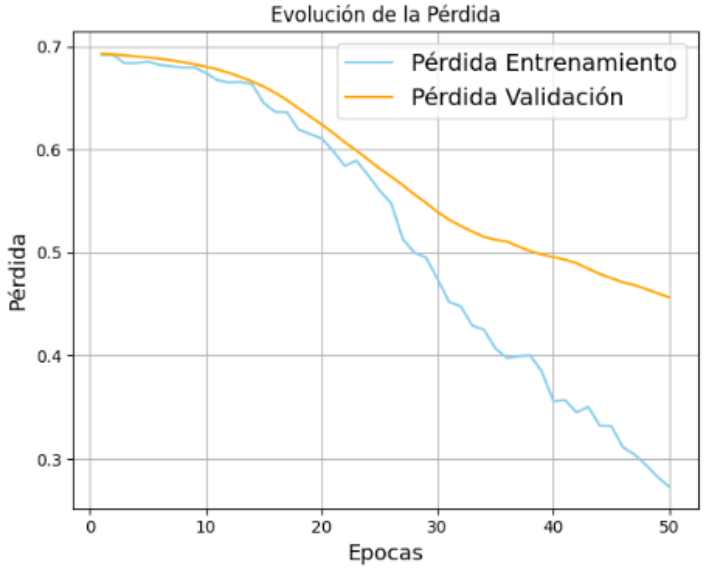


Figura 3.14 Evolución de función de pérdida (Veracidad - clásico)

La curva de error del modelo durante el aprendizaje de la red densa evidenció una disminución progresiva tanto en la pérdida de entrenamiento como en la de validación, lo que sugiere que el modelo logra adaptarse sin sobreajustarse. El flujo completo de todo el proceso se muestra en la Figura 3.15.

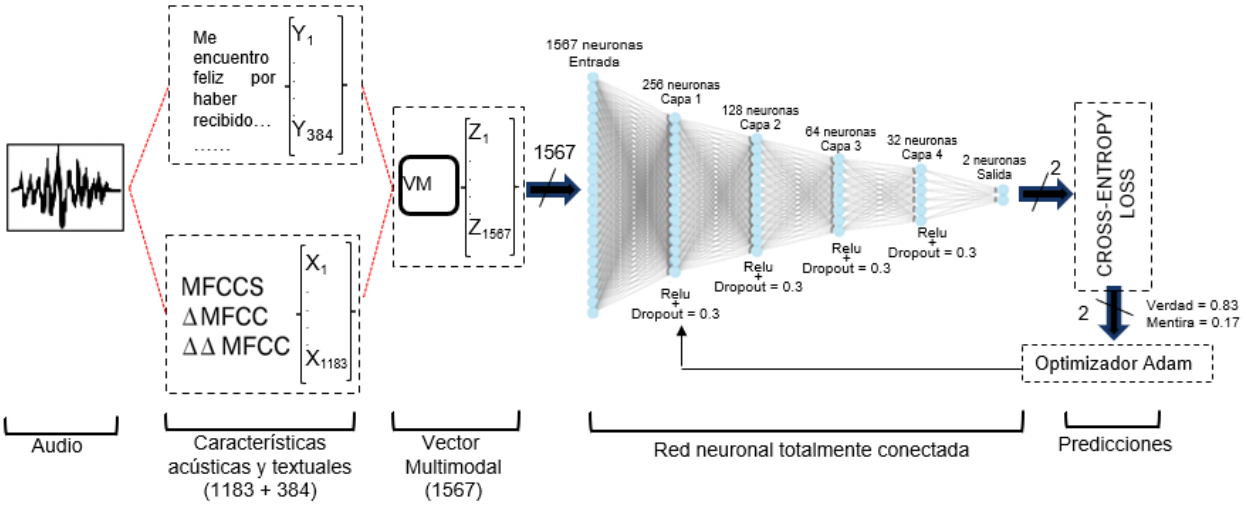


Figura 3.15 Flujo clasificación de veracidad (modelo clásico)

3.2.2.2 Modelo Cuántico: VQC

Selección y reducción de características con red densa

Dado que los circuitos cuánticos actuales requieren entradas de baja dimensionalidad, se diseñó una red neuronal densa para reducir el vector multimodal de 1567 características (1183 acústicas + 384 textuales) a un *embedding* compacto de 32 dimensiones. Esta transformación se logró entrenando una red lineal intermedia, la cual fue ajustada junto al modelo híbrido para preservar la representatividad de las variables originales (Schuld M. , Bocharov, Svore, & Wiebe, 2020). La Figura 3.16 muestra la red resultante.

```
VeraNet(
  (fc1): Linear(in_features=1567, out_features=256, bias=True)
  (dropout1): Dropout(p=0.3, inplace=False)
  (fc2): Linear(in_features=256, out_features=128, bias=True)
  (dropout2): Dropout(p=0.3, inplace=False)
  (fc3): Linear(in_features=128, out_features=64, bias=True)
  (dropout3): Dropout(p=0.3, inplace=False)
  (output): Linear(in_features=64, out_features=32, bias=True)
)
```

Figura 3.16 Arquitectura red VeraNet

La Figura 3.17 muestra el pipeline para adaptar el vector multimodal a 32 dimensiones, que es el tamaño compatible con el VQC ansatz personalizado limitado a 5 qubits ($2^5 = 32$).

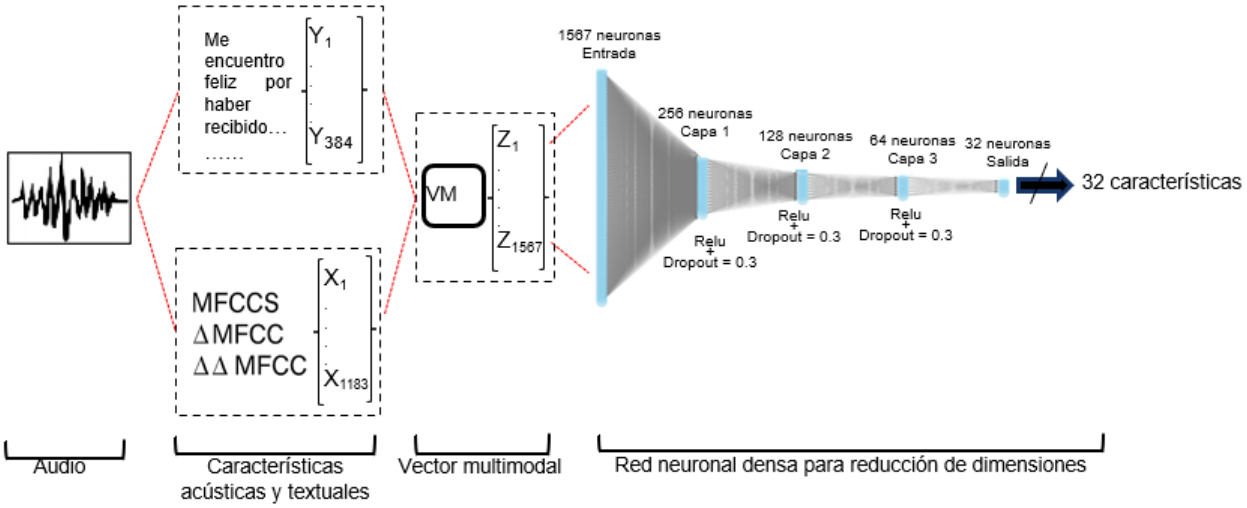


Figura 3.17 Flujo para la reducción de características multimodales (Veracidad)

Como se observa en la Figura 3.17, la red emplea múltiples capas ocultas con regularización mediante *dropout* para evitar el sobreajuste durante el proceso de compresión.

Codificación de entrada y arquitectura del circuito cuántico

Las 32 características reducidas se normalizaron según la norma L2 y se utilizaron como amplitudes de un estado cuántico mediante la codificación *AmplitudeEmbedding*. Esta técnica, permite aprovechar la alta densidad informativa al mapear un vector real normalizado en un espacio de dimensión 2^n, donde $n = \log_2 N$ con $N = 32$.

En cada capa del circuito cuántico, se aplicaron rotaciones R_y a cada uno de 5 qubits, seguidas de un esquema de entrelazamiento circular mediante compuertas CNOT. Este diseño, ajustado específicamente para representar eficazmente estados acústicos reducidos, fue replicado a lo largo de seis capas entrenables. En la figura 3.18 se ilustra la arquitectura del circuito cuántico variacional.

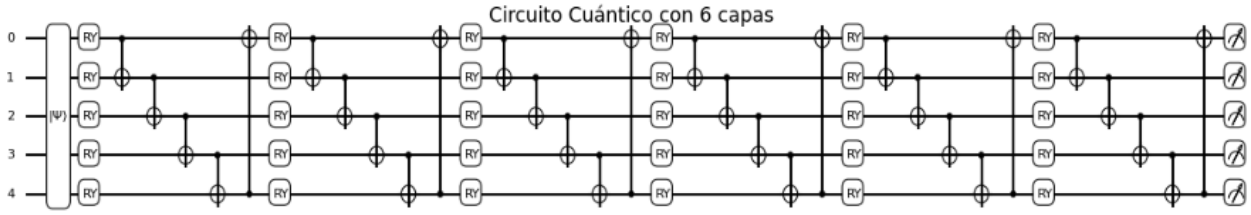


Figura 3.18 Arquitectura de VQC Ansatz Personalizado (Veracidad)

El circuito cuántico variacional para la tarea de veracidad puede representarse de manera compacta como una sucesión de capas de rotaciones parametrizadas y entrelazamiento.

$$|\Psi_{out}\rangle = \left(\prod_{l=1}^{L} E^{(l)} \ \bigotimes_{q=0}^{4} \ R_y(\theta_q^{(l)}) \right) H^{\otimes 5} |0000\rangle \qquad (13)$$

En la expresión de la ecuación (13), $H^{\otimes 5}$ genera la superposición inicial de los cuatro qubits, $R_y(\theta_q^{(l)})$ representa las rotaciones parametrizadas en la capa l, y $E^{(l)}$ corresponde al bloque de entrelazamiento implementado mediante compuertas CNOT. El producto $\prod_{l=1}^{L} E^{(l)} \ \bigotimes_{q=0}^{4} \ R_y(\theta_q^{(l)})$ refleja la naturaleza repetitiva de la arquitectura VQC, estructurada en capas alternadas de rotaciones y entrelazamiento.

Configuración de entrenamiento: qubits, capas, optimizador

El entrenamiento del circuito cuántico se llevó a cabo usando descenso por gradiente con una tasa de aprendizaje de 0.001. Dado que la simulación cuántica se realiza en entornos clásicos, el entrenamiento se ejecutó en *batch* completo por observación,

simulando la evolución del circuito sobre el dataset reducido. En la Figura 3.19 se ilustra la evolución de la función de pérdida durante el entrenamiento.

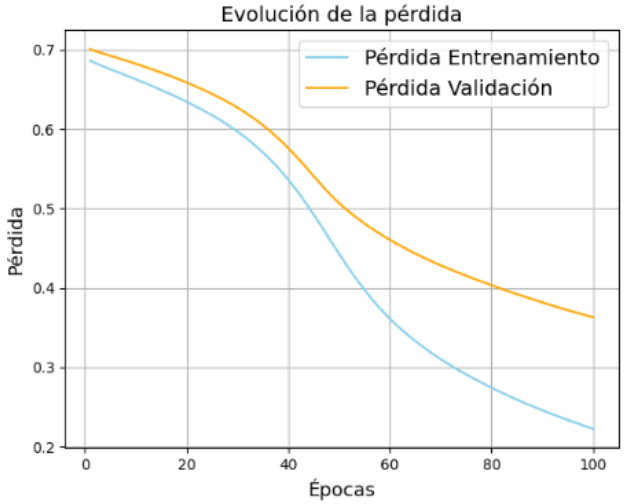


Figura 3.19 Evolución de función de pérdida (Veracidad - cuántico)

La función de errores durante el proceso de aprendizaje de la red densa evidenció una disminución progresiva tanto en la pérdida de entrenamiento como en la de validación, lo que indica una buena capacidad de generalización del modelo.

Regla del desplazamiento y salida del circuito cuántico

La actualización de los parámetros del circuito se realizó utilizando la regla del desplazamiento (parameter shift rule), una técnica derivada analíticamente para calcular el gradiente en circuitos cuánticos diferenciales. La salida del circuito, correspondiente a las expectativas de los operadores Pauli-Z, fue entregada directamente como logits a la función de errores de entropía cruzada multiclase (Cross-Entropy Loss), que internamente aplica softmax durante el cálculo del error, permitiendo retropropagación hasta los parámetros cuánticos.

Validación y análisis de desempeño

Se utilizó la misma estrategia de evaluación aplicada en el modelo tradicional, manteniendo consistencia experimental. Se monitorearon métricas como accuracy y F1-score por clase para comparar el desempeño con el modelo VeraNet. El entrenamiento del VQC fue más costoso computacionalmente, pero logró capturar patrones complejos en espacios reducidos, mostrando el potencial de la computación cuántica en tareas de clasificación multimodal.

La figura 3.20 muestra el pipeline completo de entrenamiento a partir de las entradas cuánticas codificadas.

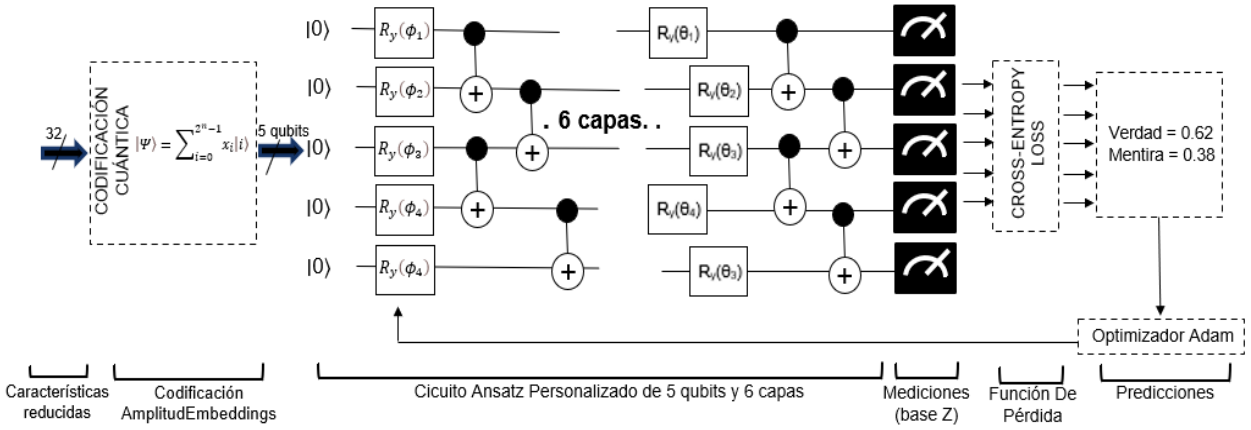


Figura 3.20 Pipeline cuántico (Veracidad)

3.2.3 Identificación del hablante

La identificación del hablante en este sistema no se aborda como una tarea tradicional de clasificación supervisada, sino como un proceso de reconocimiento basado en la comparación de huellas acústicas previamente almacenadas. Durante los primeros 5 segundos de cada llamada entrante al centro de contacto, se extraen características acústicas específicas de la voz del interlocutor, siguiendo el pipeline de la sección *3.1.4 Extracción de características acústicas*. Estas características —como MFCCs, energía, pitch y parámetros prosódicos— son condensadas en un vector representativo de la identidad vocal del hablante. El vector tiene la siguiente estructura:

[0.013, 0.020, ..., -0.087, 0.215, ..., 0.000, 0.154, 0.149, -0.001, -0.015, 0.444]

256 características acústicas

Este vector se almacena en un repositorio interno basado en *MariaDB* 10 de atributos acústicos, junto con metadatos adicionales como nombres del hablante, la fecha y hora de la interacción, el canal de comunicación (protocolo SIP), y la información comercial vinculada al contacto (por ejemplo, historial de gestiones, nombre asociado o estado de cuenta).

Cuando un individuo vuelve a llamar, el sistema repite el proceso de obtención de atributos acústicos y realiza una búsqueda en el repositorio interno para identificar coincidencias mediante un cálculo de similitud del coseno cuya fórmula es:

$$\text{Sim}(\vec{a}, \vec{b}) = \frac{\text{a} \cdot \text{b}}{||\text{a}|| \cdot ||\text{b}||} \quad (14)$$

Donde:

- a · b es el producto punto.
- ||a|| · ||b|| es el producto punto de la norma de los vectores.

En la Tabla 3.3 se muestra los umbrales de similitud.

Tabla 3.3 Umbrales de decisión según similitud del coseno

Valor de coseno de similitud	Interpretación
≥ 0.95	Alta certeza de que es el mismo hablante
0.90 – 0.95	Coincidencia probable
0.80 – 0.90	Coincidencia dudosa
< 0.80	Posiblemente hablante diferente

Si la similitud excede el umbral definido, se considera que el hablante ha sido identificado correctamente. El umbral definido para este proyecto fue:

$$\theta = 0.90 \quad (15)$$

Este enfoque permite acelerar la gestión de llamadas, evitando solicitudes manuales de identificación y consultas a sistemas externos como un CRM. La Figura 3.21 muestra el flujo para este proceso.

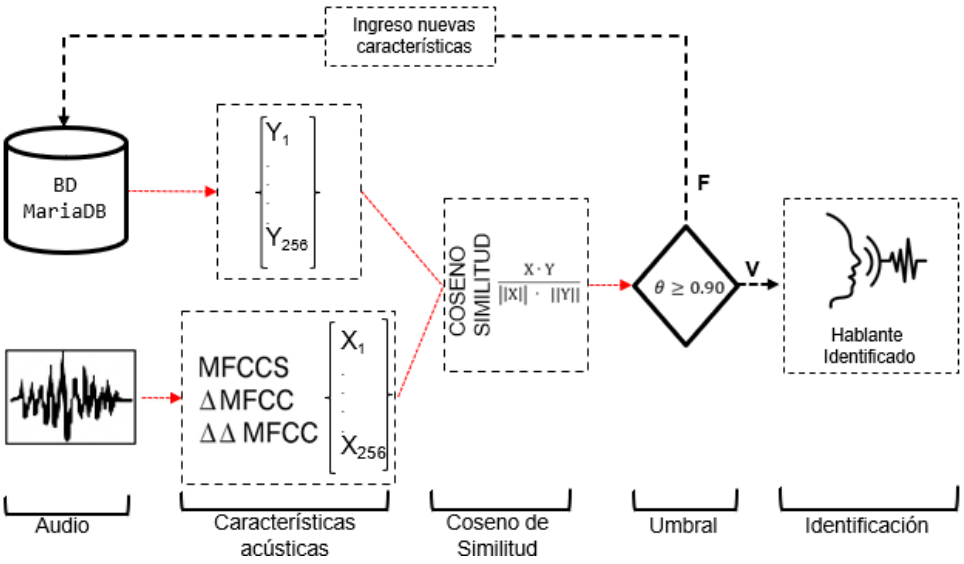


Figura 3.21 Flujo de identificación del hablante

Una vez identificado el hablante, se despliega automáticamente una interfaz ligera que muestra información clave vinculada a esa identidad vocal.

Este módulo contribuye significativamente a la reducción del tiempo de llamada y mejora la experiencia de usuario al proporcionar al agente humano o sistema automatizado datos inmediatos del interlocutor.

3.3 Métricas y comunicación de resultados

En este apartado se explican los criterios empleados para medir el desempeño de los modelos, tanto clásicos como cuánticos, en las tareas de clasificación de emociones y veracidad. La comparación técnica entre enfoques se apoya en reportes cuantitativos y visualizaciones, permitiendo establecer una línea base para el análisis comparativo posterior en el Capítulo 4.

3.3.1 Definición de métricas utilizadas y resultados obtenidos

Para evaluar a los modelos de clasificación supervisada se utilizaron indicadores clásicos, entre los que destacan:

Accuracy: Representa el porcentaje total de aciertos del modelo.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

Precision: Indica qué porcentaje de los resultados positivos generados por el modelo son correctos.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall: Indica qué parte de los casos realmente positivos fue identificada por el sistema.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

F1-score: Combina precisión y recall en un solo valor, útil especialmente en conjuntos desbalanceados.

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

Matriz de confusión: Muestra cómo se comporta el modelo frente a cada clase, diferenciando aciertos y fallos.

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{20}$$

3.3.2 Visualización de resultados: gráficas, matrices, curvas

Modelo Clásico: Clasificación de emociones

En la Figura 3.22 se observa cómo la precisión evoluciona a lo largo de las épocas, permitiendo monitorear la capacidad del modelo para generalizar.

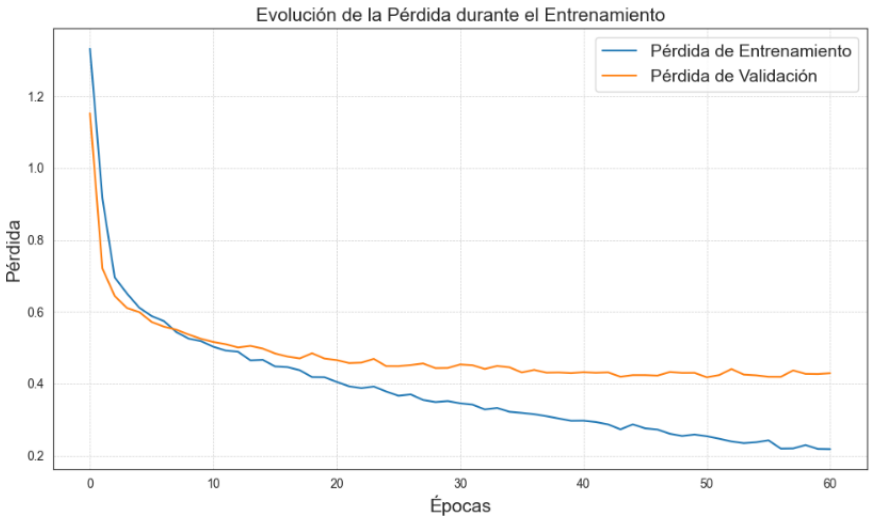


Figura 3.22 Precisión en entrenamiento modelo clásico (Emociones)

El comportamiento drecreciente y balanceado de ambas curvas indica una buena generalización sin signos de sobreajuste, lo cual valida su rendimiento en tareas binarias. En la Figura 3.23 se muestra la tabla de resultados por clase aplicado a la clasificación de emociones, con etiquetas como calma, enojo, felicidad y tristeza.

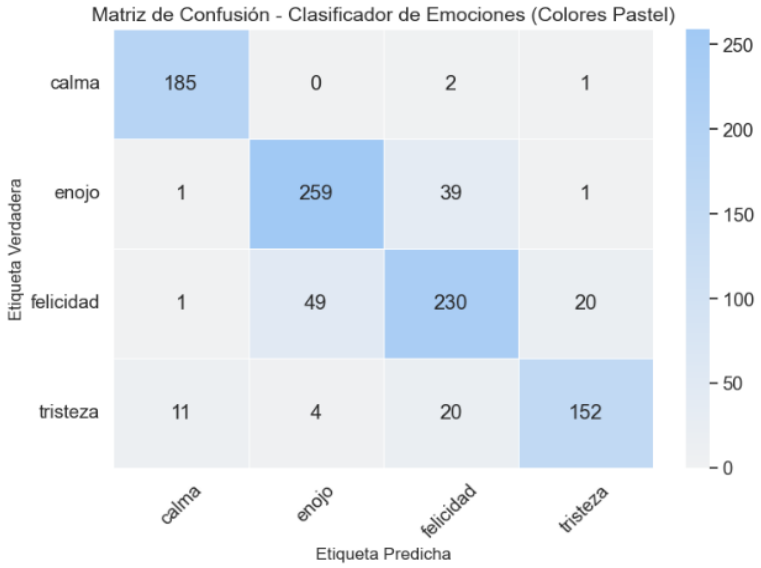


Figura 3.23 Matriz de Confusión de modelo clásico (Emociones)

La matriz revela un buen nivel de aciertos, especialmente en las clases “enojo” y “calma”, y algunas confusiones en clases más similares entre sí como “felicidad” y “tristeza”.

Modelo Cuántico: Clasificación de emociones

La Figura 3.24 ilustra cómo varía la precisión en el conjunto de entrenamiento del modelo cuántico al clasificar emociones mediante circuitos variacionales.

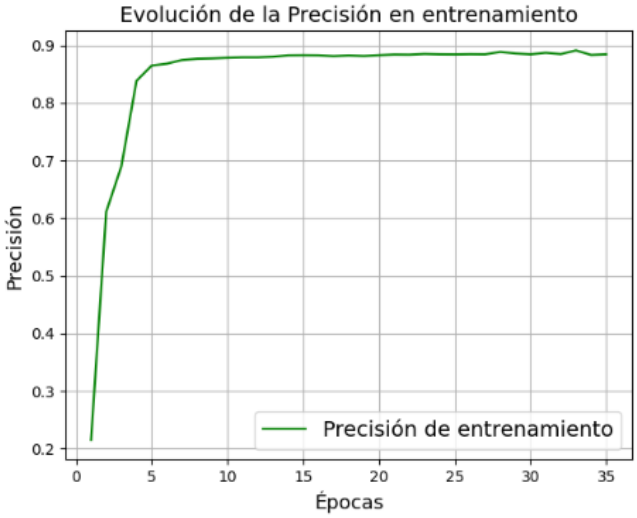


Figura 3.24 Precisión en entramiento modelo cuántico (Emociones)

El modelo alcanza rápidamente una alta precisión, lo que evidencia su eficiencia para representar patrones complejos en espacios reducidos. La Figura 3.25 muestra la matriz de confusión evaluando su comportamiento en términos de predicción por clase.

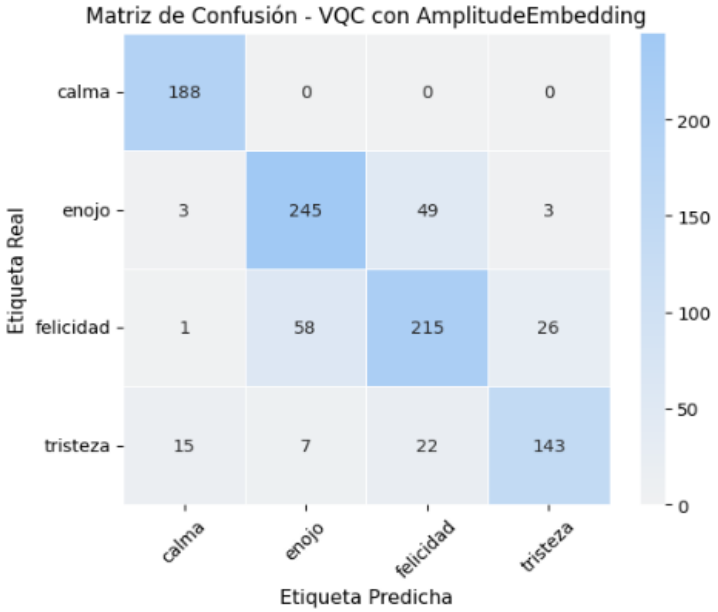


Figura 3.25 Matriz de Confusión de modelo cuántico (Emociones)

El modelo cuántico logra un desempeño aceptable, aunque se evidencia una mayor confusión en ciertas clases respecto al modelo clásico, particularmente en las emociones “felicidad” y “tristeza”.

Modelo Clásico: Clasificación de veracidad

La Figura 3.26 muestra la precisión tanto en entrenamiento como en validación para la tarea de clasificación de veracidad usando el modelo clásico.

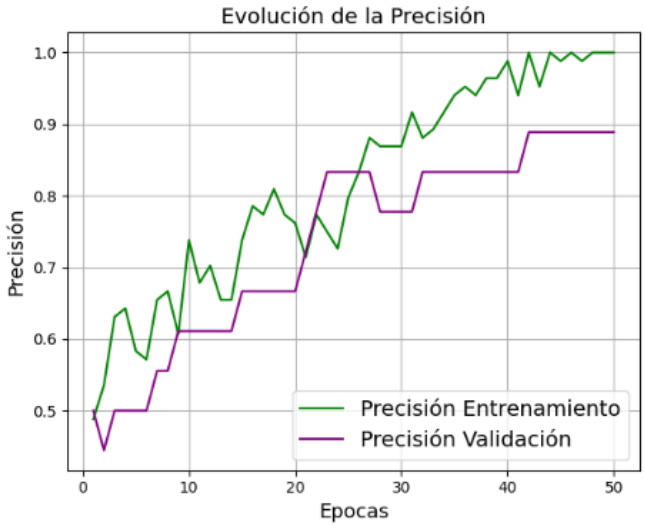


Figura 3.26 Precisión en entrenamiento modelo clásico (Veracidad)

El comportamiento creciente y balanceado de ambas curvas indica una buena generalización sin signos de sobreajuste, lo cual valida su rendimiento en tareas binarias. La tabla de resultados por clase de la Figura 3.27 corresponde al modelo clásico en la tarea binaria de clasificación de veracidad, distinguiendo entre “verdad” y “mentira”.

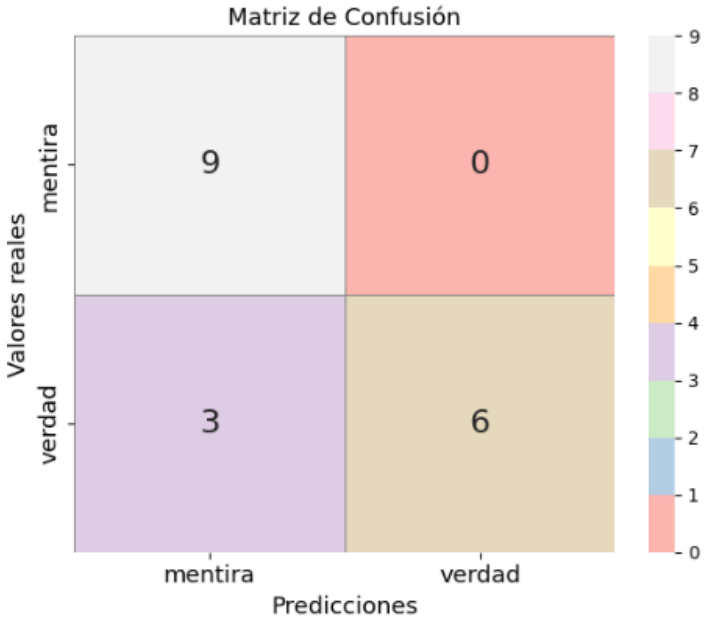


Figura 3.27 Matriz de Confusión de modelo clásico (Veracidad)

El modelo demuestra una buena precisión, con mínimos errores de clasificación, lo que respalda su utilidad en tareas de auditoría de interacciones orales.

Modelo Cuántico: Clasificación de veracidad

En la Figura 3.28 se ilustra el comportamiento de la exactitud a lo largo del proceso de aprendizaje en la tarea de clasificación de veracidad.

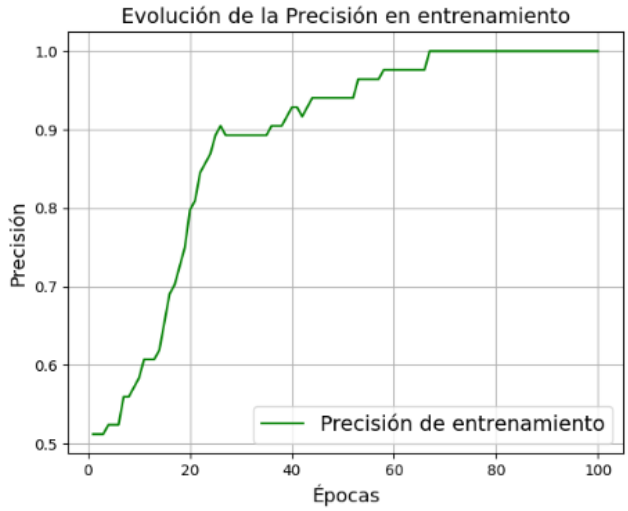


Figura 3.28 Precisión en entrenamiento modelo cuántico (Veracidad)

El modelo logró converger de forma estable, con un rendimiento comparable al modelo clásico, destacando su potencial como alternativa eficiente en tareas con baja dimensionalidad. Finalmente, en la Figura 3.29 se presenta la tabla de resultados por clase del modelo cuántico para la clasificación de veracidad.

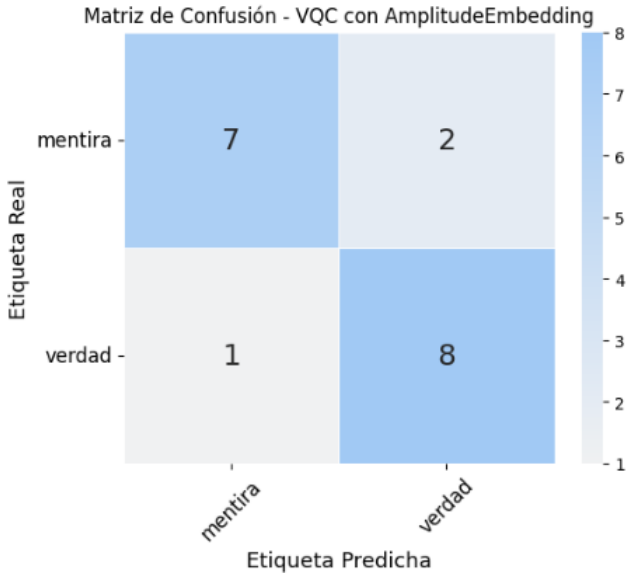


Figura 3.29 Matriz de Confusión de modelo cuántico (Veracidad)

Los resultados mostraron una distribución de errores muy similar al modelo clásico, reforzando su viabilidad incluso en escenarios reales con restricciones computacionales.

3.3.3 Resumen general de métricas

En la Tabla 3.4 se resumen las métricas obtenidas.

Tabla 3.4 Resultados de desempeño por modelos

Métrica	Clasificador de Emociones		Clasificador de Veracidad	
	Clásico	Cuántico	Clásico	Cuántico
Accuracy (%)	84.72	81.00	88.89	89.02
Precision	0.857	0.82	0.89	0.89
Recall	0.856	0.82	0.89	0.89
F1-score	0.847	0.82	0.89	0.90

Estos resultados reflejan un desempeño competitivo de los modelos cuánticos frente a sus contrapartes clásicas, especialmente destacando la capacidad de los VQC para capturar relaciones no lineales en espacios de baja dimensionalidad.

3.3.4 Comparación técnica entre modelos

La comparación técnica considera múltiples dimensiones:

- *Complejidad arquitectónica*: mide la estructura interna del modelo, incluyendo la cantidad de capas, nodos, parámetros y conexiones que definen su capacidad de aprendizaje.
- *Tamaño del embedding*: representa la cantidad de dimensiones en el espacio vectorial reducido donde se proyectan las características originales para facilitar el procesamiento del modelo.
- *Tiempo de entrenamiento*: indica la duración necesaria para que un modelo aprenda a partir de los datos, incluyendo el ajuste iterativo de sus parámetros mediante algoritmos de optimización.
- *Consumo computacional*: evalúa los recursos técnicos requeridos (CPU, GPU, RAM) durante el entrenamiento o inferencia del modelo, y su eficiencia en términos de carga de trabajo.

Comparación técnica para clasificación de emociones

En la tabla 3.5 se presenta un análisis comparativo entre el modelo clásico y cuántico para la clasificación de emociones.

Tabla 3.5 Comparación técnica entre modelos clásicos y cuánticos (Emociones)

Métrica	Modelo Clásico (RNFC)	Modelo Cuántico (VQC)
Accuracy (%)	84.72	81.00
Parámetros aprox.	>100.000 (red densa profunda)	4 qubits + 6 capas VQC
Reducción de dimensiones	No se requirió	Se redujo de 1567 a 16 componentes mediante una red densa
Tiempo de entrenamiento	~5 min en CPU	>30 min en CPU (simulador cuántico)
Tiempo de inferencia	0,40 min.	0,45 min.
Consumo computacional	Moderado (no requiere GPU)	Alto en CPU (entrenamiento más lento)
Observación clave	Mejor precisión y generalización	Menor precisión, pero desempeño competitivo en emociones simples.

Comparación técnica para clasificación de veracidad

En la tabla 3.6 se presenta un análisis comparativo entre el modelo clásico y cuántico para la clasificación de emociones.

Tabla 3.6 Comparación técnica entre modelos clásicos y cuánticos (Veracidad)

Métrica	Modelo Clásico (RNFC)	Modelo Cuántico (VQC)
Accuracy (%)	88.89	89.02
Parámetros aprox.	>100.000 (red densa profunda)	5 qubits + 6 capas VQC
Reducción de dimensiones	No se requirió	Se redujo de 1567 a 32 componentes mediante una red densa
Tiempo de entrenamiento	~5 min en CPU	>30 min en CPU (simulador cuántico)
Tiempo de inferencia	0,30 min.	0,32 min.
Consumo computacional	Moderado (no requiere GPU)	Alto en CPU (entrenamiento más lento)
Observación clave	Mejor precisión y generalización	Menor precisión, pero desempeño competitivo en emociones simples.

A partir de esta comparación se evidencia que, si bien el modelo cuántico presenta un mayor costo computacional, logra resultados competitivos frente al modelo clásico. Esto sugiere que los enfoques híbridos pueden ser una alternativa viable en escenarios con restricciones de datos o donde se priorice la exploración de tecnologías emergentes.

3.4 Infraestructura para procesamiento y almacenamiento

Esta sección describe la infraestructura tecnológica utilizada para el tratamiento de datos, proceso de aprendizaje de los modelos clásicos y cuánticos, y el almacenamiento de los resultados. Además, se detalla la estructura organizativa del proyecto y los componentes requeridos para su puesta en producción.

3.4.1 Herramientas y librerías utilizadas

El proyecto empleó un conjunto diverso de herramientas, librerías y marcos de trabajo diseñados para cubrir todas las etapas del procesamiento, entrenamiento e integración del sistema.

- **Lenguaje de programación**: Python 3.10
- **Procesamiento acústico y extracción de características**:

  librosa: extracción de MFCCs, ZCR, RMS, pitch y espectrogramas.

  torchaudio: soporte para señales de audio compatibles con PyTorch.

  transformers: carga de modelos wav2vec2 para embeddings acústicos densos.
- **Procesamiento textual**:

  sentence-transformers: obtención de embeddings semánticos mediante SBERT (Sentence-BERT).
- **Frameworks de aprendizaje automático**:

  PyTorch: arquitectura base del modelo EmotionNet, entrenamiento de redes densas.

  scikit-learn: selección de características, entrenamiento de modelos clásicos, métricas.

  joblib: serialización de modelos.
- **Computación cuántica**:

  PennyLane: diseño de circuitos cuánticos variacionales (VQC) e integración con optimizadores.

  Qiskit: simulador de respaldo para pruebas de circuitos en CPU.
- **Visualización y análisis**:

  matplotlib, seaborn, plotly: análisis exploratorio, visualización de métricas, curvas de entrenamiento y dispersión de embeddings.

3.4.2 Recursos computacionales

Este proyecto se desarrolló combinando recursos locales y entornos colaborativos en la nube. Se optimizó el uso de CPU y GPU según el tipo de modelo implementado.

- **Recursos locales**:
  - Unidad central de procesamiento AMD Ryzen 7 5800H (arquitectura de 8 núcleos y 16 hilos), con 16 GB de RAM.

- GPU NVIDIA RTX 3060 con 6 GB de memoria dedicada para entrenamiento con PyTorch.

- **Entornos en la nube**:
  - *Google Colab Pro*: ejecución de experimentos con aceleración GPU (modelo clásico).
  - *Simulador PennyLane (default.qubit) y Qiskit Aer*: ejecución de circuitos cuánticos.

3.4.3 Organización de carpetas y control de versiones

La estructura de carpetas fue diseñada para mantener una organización clara y modular de los componentes del proyecto. La Figura 3.30 muestra la estructura detallada.

```
proyecto_tesis/
│
├── dataset/                  # Datos originales y transformados
├── notebooks/                # Cuadernos experimentales y de entrenamiento
├── models/                   # Modelos entrenados (PyTorch, joblib, etc.)
├── results/                  # Resultados de evaluación y visualizaciones
├── scripts/                  # Código fuente y automatizaciones
├── quantum/                  # Circuitos VQC, funciones de costo, optimización
├── docs/                     # Documentación del proyecto
└── config/                   # Parámetros de configuración (.yaml, .json)
```

Figura 3.30 Estructura de carpetas del proyecto

El proyecto fue controlado mediante Git, y alojado en un repositorio privado de GitHub, permitiendo trazabilidad de versiones, comparación de experimentos y recuperación ante errores.

3.5 Plataformas y prototipos de visualización

Esta sección describe las interfaces, entornos de visualización y mecanismos de entrega de resultados diseñados para mostrar las predicciones del sistema de caracterización conductual, tanto en contexto experimental como en entornos de producción. Se incluyen herramientas ligeras de prueba, prototipos gráficos para interpretación de resultados, y la arquitectura final de despliegue para uso en tiempo real.

3.5.1 Interfaz de prueba o simulación de resultados en consola y notebook

Durante la fase de experimentación, se implementó una interfaz de prueba en entorno Jupyter Notebook que permitió visualizar las predicciones generadas por los modelos (clásicos y cuánticos), así como observar las métricas asociadas a cada clase: precisión, F1-score, matriz de confusión, entre otras. Esta interfaz sirvió como entorno de simulación para:

- Evaluar el rendimiento de los modelos entrenados.
- Comparar resultados entre múltiples configuraciones.
- Visualizar curvas de entrenamiento y dispersión de clases.

Las simulaciones se realizaron en notebooks organizados por tareas (emociones, veracidad), y codificados en Python utilizando bibliotecas como matplotlib, plotly y seaborn.

3.5.2 Prototipo en interfaz ligera

Se desarrollaron prototipos funcionales que simulan una integración real del sistema, tanto desde la perspectiva del usuario final como desde el entorno de pruebas. Entre ellos, se destaca el uso de *Hugging Face Spaces* como interfaz ligera para el despliegue de modelos entrenados, permitiendo una interacción directa mediante una interfaz web accesible y gratuita. Esta plataforma permitió:

- Cargar muestras de audio para su análisis.
- Obtener resultados inmediatos en la interfaz visual.
- Facilitar pruebas a stakeholders o usuarios no técnicos.

Además, como se ilustra en la Figura 3.31, se desarrolló un prototipo funcional donde se visualizan en tiempo real las salidas del modelo de análisis fonético. Esta consola muestra los resultados de forma estructurada, incluyendo:

- La emoción predominante (felicidad, tristeza, calma o enojo).
- El nivel de veracidad estimado (binario).
- Identificación del hablante.

Figura 3.31 Prototipo ligero de clasificación emocional

Este tipo de integración permite que tanto supervisores como agentes accedan a los resultados sin requerir entrenamiento técnico especializado.

3.5.3 Puesta en producción y despliegue del modelo

El sistema fue diseñado para entregar las predicciones del modelo en tiempo real utilizando una arquitectura sin servidor (*serverless*), basada en servicios en la nube. Los componentes principales de esta infraestructura incluyen:

- **Frontend**: Software de centro de contacto DINOMI que captura los audios y los envía para su análisis.
- **Middleware**: Los datos viajan por Internet mediante solicitudes estructuradas (RESTful API).
- **Backend**:
    - *Amazon API Gateway*: expone un *endpoint* público seguro que recibe los audios procesados.
    - *AWS Lambda*: función en la nube que ejecuta el modelo previamente entrenado y retorna las predicciones.
    - *Formato de respuesta*: JSON con dimensiones de análisis (emociones, veracidad, hablante).

Esta arquitectura garantiza escalabilidad, disponibilidad y bajo costo operativo, siendo ideal para entornos empresariales con altos volúmenes de interacción.

Para realizar el despliegue en producción del modelo, se optó por una estrategia *Canary Deployment*, que permite introducir versiones nuevas del sistema de forma gradual y

controlada, minimizando riesgos de fallos generalizados. En la Fase 1 (ver Figura 3.32), el nuevo sistema fue desplegado en 5 centros de contacto seleccionados (de un total de 25), que actuaron como entornos de prueba en condiciones reales.

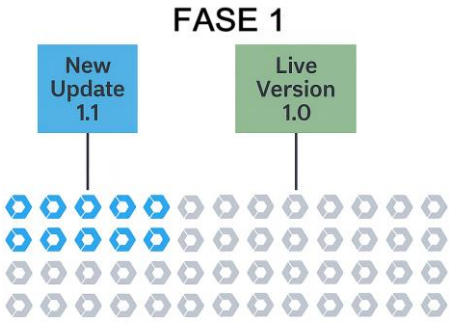


Figura 3.32 Fase 1: actualización canaria en 5 centros de contacto

Tras validar el correcto funcionamiento en la fase piloto, se procedió a la Fase 2 (ver Figura 3.33), donde el modelo fue desplegado en los 25 centros de contacto, reemplazando completamente la versión anterior.

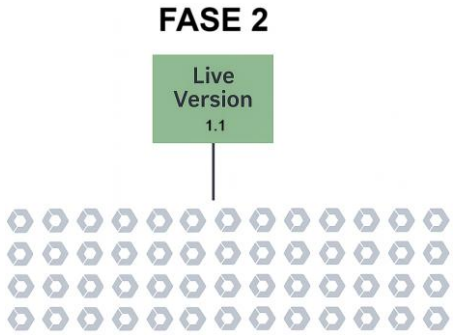


Figura 3.33 Fase 2: actualización total a la versión 1.1

Este enfoque progresivo aseguró una transición segura y efectiva, alineada con buenas prácticas de despliegue continuo en sistemas distribuidos.

3.5.4 Evolución del Prototipo hacia un Entorno de Producción

El proceso de despliegue del sistema se estructuró en tres fases progresivas que se muestran en la Figura 3.34.

Figura 3.34 Fases del despliegue del sistema

3.5.5 Visualizaciones en Producción

Como parte del despliegue operativo del sistema de caracterización conductual, se implementaron tres módulos visuales integrados al software de gestión de llamadas DINOMI. Las visualizaciones están orientadas a tres tipos de usuarios: agentes, supervisores y personal gerencial.

Módulo 1: Reporte tabular de clasificación de emociones y veracidad

La Figura 3.35 muestra el módulo de reporte que expone los resultados de la clasificación de emociones de cada llamada analizada. Este componente incluye información como:

- Fecha y hora de la interacción oral.
- Número telefónico del llamante y del agente.
- Duración de la llamada.
- Clasificación de la emoción (felicidad, tristeza, calma o enojo).
- Porcentaje de clasificación.
- Tiempo de clasificación.

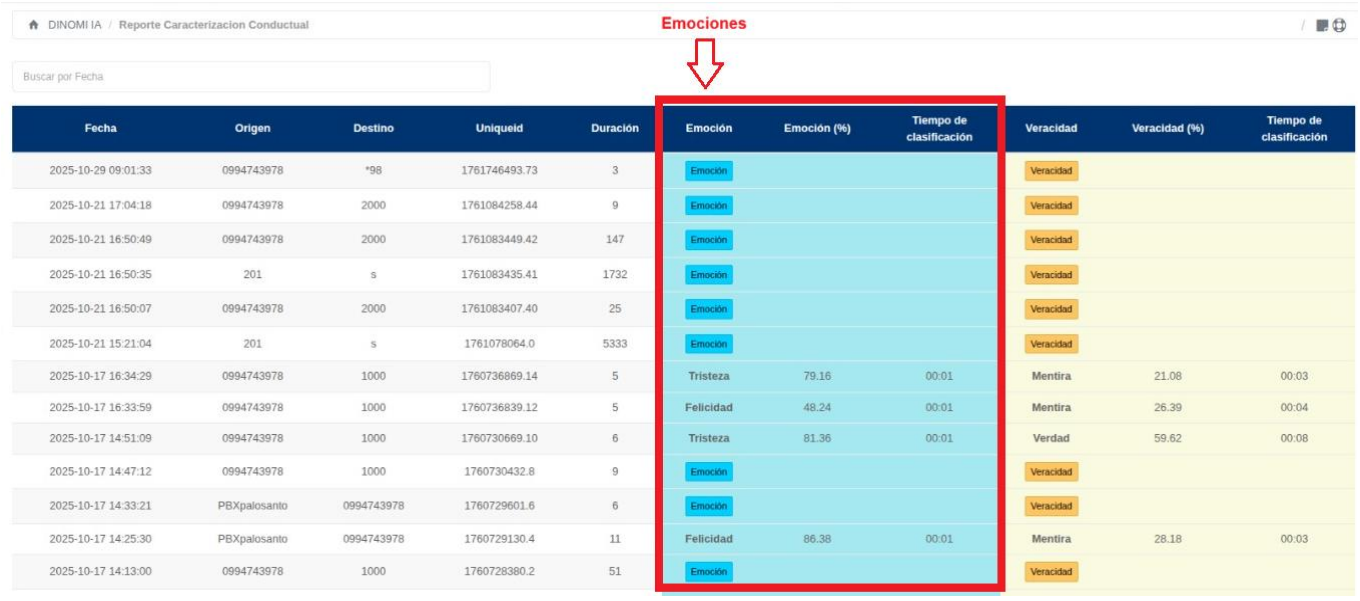


Figura 3.35 Módulo de clasificación de emociones

Este módulo permite realizar auditorías rápidas, identificar patrones inusuales y servir como evidencia en procesos de calidad o seguimiento de campañas.

Módulo 2: Reporte tabular de clasificación de veracidad

La Figura 3.36 presenta el segundo módulo de reporte que expone los resultados de la clasificación de veracidad de cada llamada analizada. Este componente incluye información como:

- Fecha y hora del análisis.
- Número de teléfono del llamante y del agente.
- Duración de la llamada.
- Clasificación de veracidad.
- Porcentaje de clasificación.
- Tiempo de clasificación.

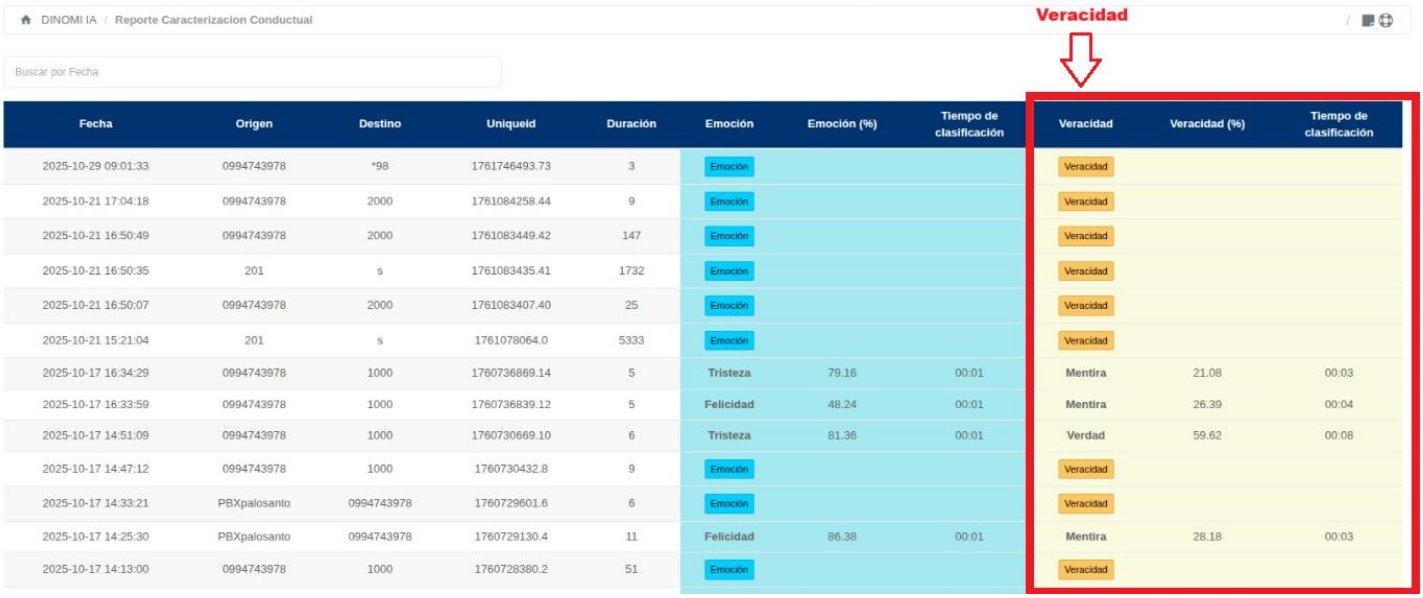


Figura 3.36 Módulo de clasificación de veracidad

Este módulo permite realizar auditorías enfocadas en la veracidad del discurso, lo cual resulta especialmente útil para analizar la credibilidad de promesas de pago realizadas por deudores durante campañas de cobranza en centros de contacto.

Módulo 3: Identificación del Hablante

Este módulo, integrado en la consola del agente, permite visualizar en tiempo real la identidad estimada del hablante mediante el análisis fonético de su voz. Utilizando técnicas avanzadas de reconocimiento de locutor, el sistema extrae huellas acústicas únicas y las compara contra una base de datos interna para determinar su identidad

probable. En la Figura 3.37 se ilustra el módulo de identificación del hablante con su información relevante, tal como:

- Canal de comunicación (SIP).
- Información comercial del hablante relacionado al negocio o gestión.
- Nombre completo estimado.
- Fecha y hora de la interacción.

Figura 3.37 Módulo de identificación del hablante

Una de las fortalezas más relevantes de este módulo es la de reconocer al interlocutor durante los primeros 4 segundos de la llamada, sin necesidad de hacer preguntas al usuario ni consultar sistemas externos como un CRM. Esto representa un ahorro significativo de tiempo en los flujos de atención, donde identificar rápidamente al cliente permite hacer más eficientes los procesos y brindarle una mejor experiencia.

Módulo 4: Dashboard de visualizaciones de caracterización agregada

El tercer módulo, mostrado en la Figura 3.38, corresponde a un dashboard gerencial con visualizaciones interactivas que permite analizar la distribución de emociones por día, campaña o agente, comparar la veracidad percibida entre grupos de usuarios y monitorear tendencias temporales y picos anómalos.

Figura 3.38 Módulo de visualizaciones

Las visualizaciones incluyen gráficos de pastel, histogramas, gráficos de barras apiladas y líneas temporales que brindan visibilidad estratégica sobre el estado emocional y conductual de los interlocutores.

Los resultados presentados en esta sección proporcionan una línea base para contrastar los enfoques clásicos y cuánticos, con miras a escoger el modelo más adecuado para su despliegue en entornos de producción. En el Capítulo 4 se profundizará en esta evaluación considerando no solo métricas técnicas estandarizadas, sino también indicadores orientados al negocio, como el ahorro de tiempo operativo, una mayor claridad en la toma de decisiones y la eficiencia en la utilización del recurso humano. Este análisis permitirá dimensionar el impacto real de cada modelo en el contexto de los centros de contacto, donde la precisión y velocidad de los sistemas inteligentes repercuten directamente en la productividad y calidad del servicio.

CAPÍTULO 4

4 ANÁLISIS DE RESULTADOS

En este capítulo se examinan los resultados obtenidos a partir del proceso de validación del sistema desarrollado para la detección automática de emociones, veracidad, nivel de estrés e identificación del hablante en interacciones orales de centros de contacto. Se presentan las estrategias empleadas para la recolección de datos de validación, las pruebas de funcionalidad y su implementación en escenarios reales, así como la comparación entre el flujo de trabajo tradicional y el sistema automatizado mediante esquemas de *A/B testing*. Asimismo, se incluyen métricas cuantitativas y cualitativas que permiten evaluar el impacto operativo, la aceptación por parte de los actores que usan el sistema y el grado de cumplimiento de los objetivos definidos en el Capítulo 1. Al final del proceso se realizó un análisis costo–beneficio que integra los hallazgos técnicos y de negocio, proporcionando una visión holística del impacto real de la solución en un entorno operativo.

4.1 Proceso de adquisición de datos y validación del sistema

La validación del sistema propuesto se realizó sobre datos recolectados específicamente para evaluar su impacto operativo y compararlo con el proceso manual tradicional. A diferencia de los datasets utilizados para el proceso de aprendizaje de los modelos, esta recolección tiene como finalidad generar evidencias cuantitativas y cualitativas sobre la efectividad del sistema en un entorno real.

4.1.1 Fuentes y recolección de datos para validación

a) Registros históricos de operación

- Audios de interacciones previamente etiquetados manualmente por agentes o supervisores (emociones y veracidad).
- Tiempos promedio de análisis por interacción registrados en las herramientas actuales de supervisión.
- Reportes de errores o inconsistencias detectados por supervisores durante auditorías previas.
- Estos datos permitieron establecer una línea base histórica para comparar el desempeño y eficiencia del sistema automatizado.

b) Datos recolectados durante pruebas en producción controlada

- Registros de audios procesados por el sistema propuesto durante el periodo de validación.
- Resultados automáticos de clasificación (emociones, veracidad, estrés, identificación del hablante).
- Métricas de uso y tiempos de respuesta del sistema.

c) Datos cualitativos recolectados producto de encuesta

Como complemento a las métricas cuantitativas, se aplicaron entrevistas y encuestas breves a:

- 1 supervisores de calidad de cada centro de contacto.
- 2 auditores de cada centro de contacto.
- 17 agentes de cada centro de contacto.

Estas encuestas emplearon escalas tipo Likert (ver detalle en el Apéndice) y técnica Mann-Whitney (Mann & Whitney, 1947) para evaluar:

- Percepción de exactitud del sistema.
- Utilidad en los procesos de toma de decisiones.
- Nivel de confianza respecto a los resultados obtenidos.

En la Tabla 4.1 se resume la distribución de los 100 encuestados del piloto por método y rol (supervisores, auditores y agentes).

Tabla 4.1 Distribución de encuestados por método y rol

Método	Total	Supervisores	Auditores	Agentes
Manual	30	2	3	25
Automático – Clásico	35	1	3	31
Automático – Cuántico	35	2	4	29
Total	**100**	**5**	**10**	**85**

La muestra queda balanceada (30/35/35) y representativa (5/10/85 por rol), adecuada para comparar percepciones entre métodos en los análisis Likert.

4.1.2 Estrategia de validación: A/B Testing

La validación se realizó mediante despliegue canario en cinco centros de contacto, bajo un esquema de *A/B testing* controlado. Para emociones y veracidad se compararon tres enfoques (Manual, Automático–Clásico, Automático–Cuántico); para identificación del

hablante, dos (método actual vs. Automático). Se recolectaron n = 200 audios por brazo y por tarea para precisión y tiempos de clasificación y se aplicaron 100 encuestas Likert (1–5) a usuarios (supervisores, auditores y agentes).

Diseño experimental (A/B/n):

- Grupo A (control): flujo tradicional con proceso manual de análisis por supervisores.
- Grupo B (experimental): sistema automatizado; en emociones y veracidad se asignó aleatoriamente a Clásico o Cuántico (A/B/n).
- Asignación y balanceo: aleatorización a nivel de audio/interacción (bloqueada por centro) con reparto 50/50 entre A y B, y 50/50 entre modelos dentro de B (cuando aplica), evitando contaminación entre brazos.
- Muestreo: n = 200 por brazo y tarea para precisión y tiempo; 100 encuestas para percepción (Índice_UX).

4.1.3 Métricas e inferencia

Métricas primarias por tarea (emociones, veracidad, identificación del hablante).

- Precisión (aciertos/total): prueba de dos proporciones (uplift absoluto en p.p. y relativo, IC95%).
- Tiempo de clasificación por audio (min): t de Welch (varianzas desiguales), IC95% y reducción relativa.
- Percepción (Likert): Mann–Whitney por pares y Índice_UX.
- Criterios: α = 0,05; contraste unilateral vs. Manual y bilateral entre modelos; resultados agregados por centro con control de bloque.

Índice_UX (Likert)

Se calculó como combinación ponderada de tres ítems (1–5):

$$Índice_{UX} = 0.4 \cdot Utilidad + 0.4 \cdot Confianza + 0.2 \cdot Facilidad \quad (21)$$

Índice compuesto (media aritmética ponderada)

Para integrar precisión, tiempo y percepción en un único puntaje por método y por tarea, se normalizaron las métricas a [0,1] mediante min–max por tarea (entre los métodos comparados en esa tarea) y luego se aplicaron pesos específicos:

$$U_i^{(p)} = \frac{Acc_i - \min(Acc)}{\max(Acc) - \min(Acc)} \quad (22)$$

$$U_i^{(t)} = \frac{\max(Tiempo) - Tiempo_i}{\max(Tiempo) - \min(Tiempo)} \quad (23)$$

$$U_i^{(l)} = \frac{\text{Índice\_}UX_i - 1}{4} \quad (24)$$

El índice compuesto (media aritmética ponderada)

$$S_i = w_p U_i^{(p)} + w_t U_i^{(t)} + w_l U_i^{(l)} \quad (25)$$

Donde los pesos utilizados fueron:

- Emociones/Veracidad: $w_p = 0{,}60, w_t = 0{,}25, w_l = 0{,}15$
- Identificación del hablante: $w_p = 0{,}70, w_t = 0{,}20, w_l = 0{,}10$

4.1.4 Implementación operativa (DINOMI):

Gracias a que DINOMI es plataforma propia, se bifurcaron los flujos sin interrumpir operaciones (zero-downtime), con instrumentación de logs, control de versión de modelos y trazabilidad de muestras. Esto permitió comparar en paralelo el proceso manual y el automatizado bajo las mismas condiciones operativas.

4.1.5 Criterios de evaluación

La validación considerará cumplidos los objetivos si:

- El tiempo promedio de análisis por audio disminuye significativamente respecto a la línea base histórica.
- Se incrementa o mantiene la precisión de clasificación en comparación con la revisión manual.
- Los usuarios califican la herramienta como útil y confiable en al menos un 80% de las encuestas.

La Tabla 4.2 resume las variables seleccionadas, el tipo de métrica correspondiente, el método de medición, la fuente de datos y las metas esperadas para considerar satisfactoria la validación del proyecto.

Tabla 4.2 Variables, métodos de medición y metas de validación del sistema

Variable evaluada	Tipo de métrica	Método de medición	Fuente de datos	Meta esperada[5]
Tiempo promedio de análisis por audio	Cuantitativa	Cronometrar desde inicio hasta decisión final (manual y automatizado)	Registros DINOMI + observación directa	Reducción ≥ 50% respecto a línea base
Precisión de clasificación (emociones/veracidad)	Cuantitativa	Comparación con etiquetas validadas por expertos	Dataset de validación	≥ 75% Accuracy y ≥ 0.75 AUC en veracidad
Consistencia en decisiones	Cuantitativa	Cálculo de acuerdo interevaluador (manual vs sistema)	Resultados manuales y automáticos	Diferencia ≤ ±5 p.p. respecto a baseline
Percepción de utilidad	Cualitativa	Encuesta tipo Likert (1–5) a usuarios	Supervisores y agentes	≥ 80% respuestas en categorías “Útil” o “Muy útil”
Confianza en resultados	Cualitativa	Encuesta tipo Likert (1–5)	Supervisores y agentes	≥ 80% en categorías “Alta” o “Muy alta”
Facilidad de uso del sistema	Cualitativa	Encuesta tipo Likert (1–5)	Usuarios clave	≥ 80% en categorías “Fácil” o “Muy fácil”

La definición de estas variables y sus metas asociadas proporciona un marco objetivo para medir el impacto del sistema, facilitando la comparación entre el proceso manual y el automatizado durante las pruebas A/B. Este enfoque asegura que la validación no solo se base en indicadores técnicos, sino también en beneficios operativos y en la percepción de valor por parte de los usuarios.

En la Figura 4.1 se ilustra el esquema de validación A/B utilizado, donde ambos grupos reciben los mismos datos y generan métricas cuantitativas y cualitativas para su posterior comparación.

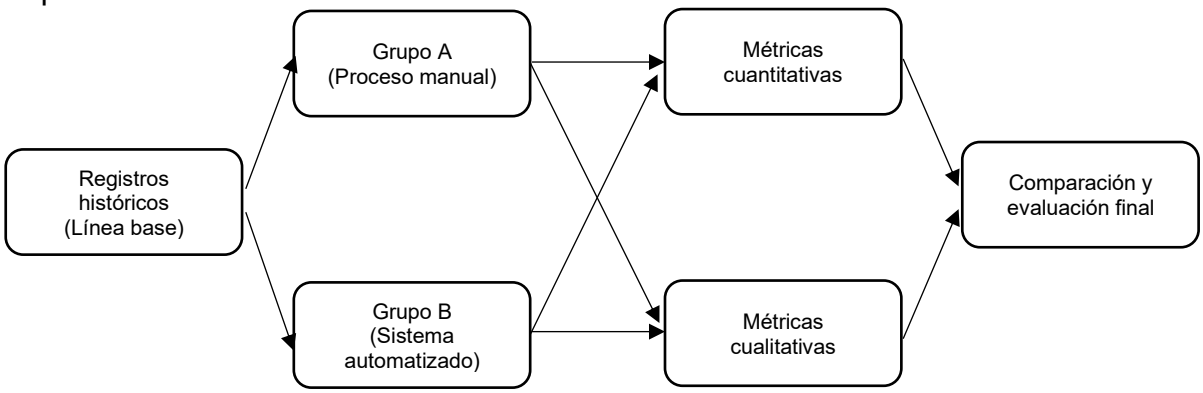


Figura 4.1 Flujo de Validación A/B *testing*

Este flujo aseguró la comparabilidad de resultados y combina indicadores objetivos con la percepción de los usuarios para evaluar el impacto del sistema.

[5] Metas basadas en objetivos definidos en el Capítulo 1

4.2 Análisis de resultados en condiciones reales de operación

Esta sección analiza los resultados en producción y se integra la evidencia con un índice compuesto ponderado que sintetiza desempeño técnico y aceptación operativa.

4.2.1 Resultados para la clasificación de Emociones

a) Precisión

La Tabla 4.3 resume la precisión alcanzada en producción por método.

Tabla 4.3 Precisión en producción por método

Tarea	Método	# de audios	Aciertos	Precisión (%)
Emociones	Manual (línea base humano)	200	152	76,00%
	Automático – Clásico	200	168	84,00%
	Automático – Cuántico	200	163	81,50%

Se observa una mejora clara de los enfoques automáticos sobre el manual. El clásico pasa de 76,0% a 84,0% (+8,0 p.p.) y el cuántico a 81,5% (+5,5 p.p.).

En las Figuras 4.2, 4.3 y 4.4 se ilustran las distribuciones esperadas del porcentaje de acierto para cada comparación en emociones, con la marca del 95% de confianza.

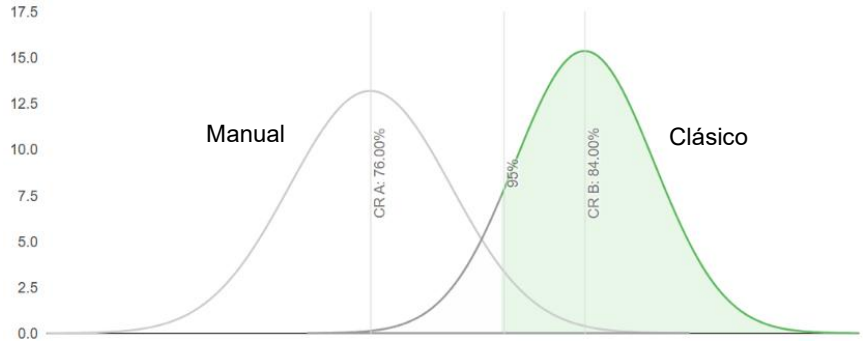


Figura 4.2 Curvas A/B de precisión en emociones (Manual vs Clásico)

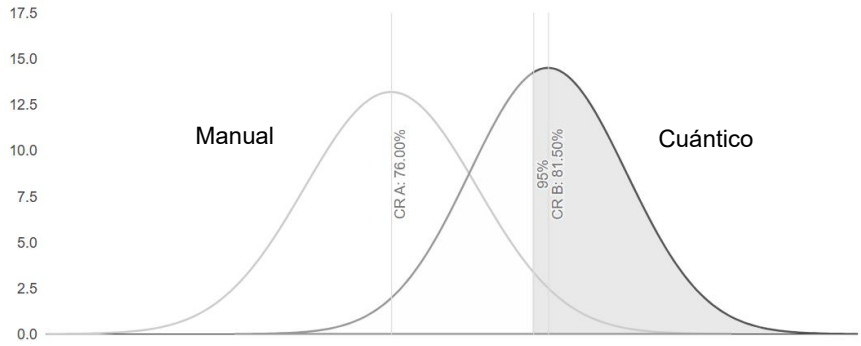


Figura 4.3 Curvas A/B de precisión en emociones (Manual vs Cuántico)

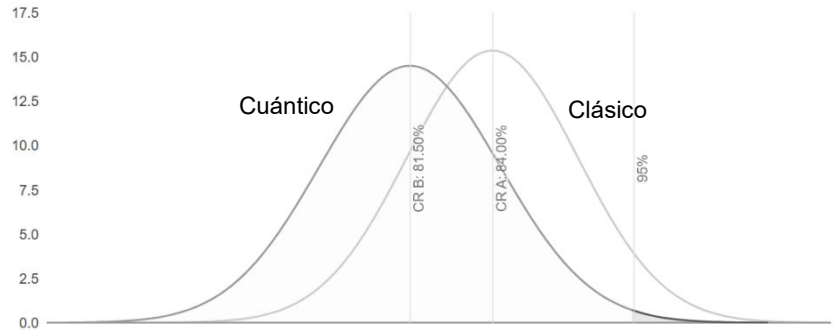


Figura 4.4 Curvas A/B de precisión en emociones (Cuántico vs Clásico)

Se evidencia una mejora significativa de Automático–Clásico sobre Manual. Sin embargo, no se observa una mejora significativa de Automático–Cuántico sobre Manual.

La Tabla 4.4 sintetiza la mejora relativa de precisión en emociones para cada comparación, junto con su p-value.

Tabla 4.4 Mejora relativa de precisión en emociones y significancia

Caracterización	Comparativo	Mejora Relativa (%)	p-value
Emociones	Manual vs Clásico	10,53%	0.0222
	Manual vs Cuántico	7,24%	0.0889
	Cuántico vs Clásico	3,07%	0.7460

Se confirma una mejora significativa del modelo Clásico frente al Manual (+10,53%, p=0,022). Los valores resaltados en gris no muestran diferencias significativas. El Cuántico muestra una poca mejora sobre el Manual (+7,24%, p=0,089) y tampoco difiere del Clásico (+3,07%, p=0,746), por lo que el Clásico resulta la opción preferente.

b) Tiempo

La Tabla 4.5 resume las estadísticas descriptivas del tiempo de clasificación por audio para cada método en emociones, reportando media, desviación estándar y tamaño muestral.

Tabla 4.5 Tiempo de clasificación por método en emociones

Tarea	Método	# audios	Media de clasificación (min.)	Desviación estándar
Emociones	Manual (línea base humano)	200	3,50	1,50
	Automático – Clásico	200	0,40	0,12
	Automático – Cuántico	200	0,45	0,15

El proceso Manual requiere 3,50 ± 1,50 min, mientras que Clásico y Cuántico operan en 0,40 ± 0,12 y 0,45 ± 0,15 min, respectivamente, es decir, casi un orden de magnitud más rápidos.

En las Figuras 4.5, 4.6 y 4.7 se ilustra la distribución de la diferencia de medias ($d = \bar{x}_{S1} - \bar{x}_{S2}$) para el tiempo por audio en emociones, con su error estándar (SE) y el p-valor. En Manual vs. Automático un $d > 0$ indica que el Manual es más lento; en Clásico vs. Cuántico, un $d < 0$ implica que Clásico es más rápido.

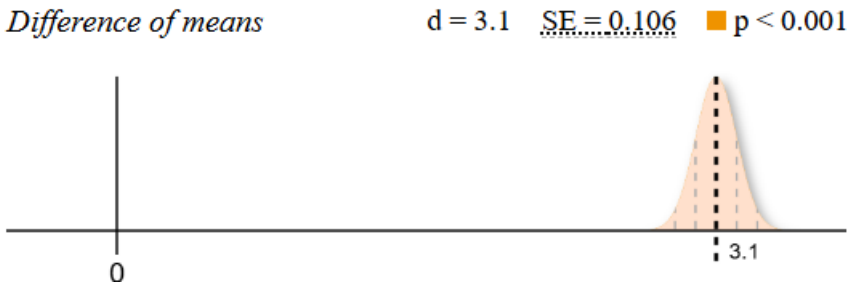


Figura 4.5 Diferencia de medias del tiempo de clasificación (Manual vs Clásico)

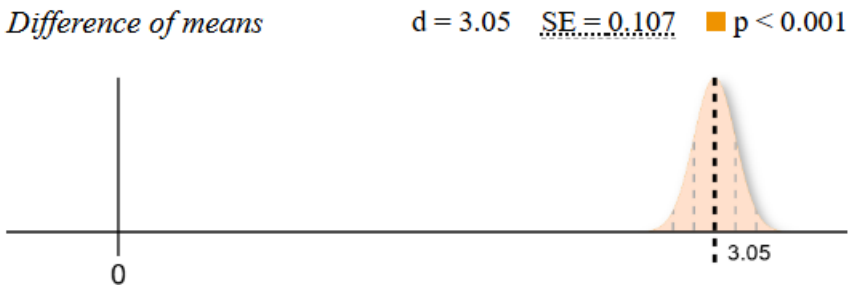


Figura 4.6 Diferencia de medias del tiempo de clasificación (Manual vs Cuántico)

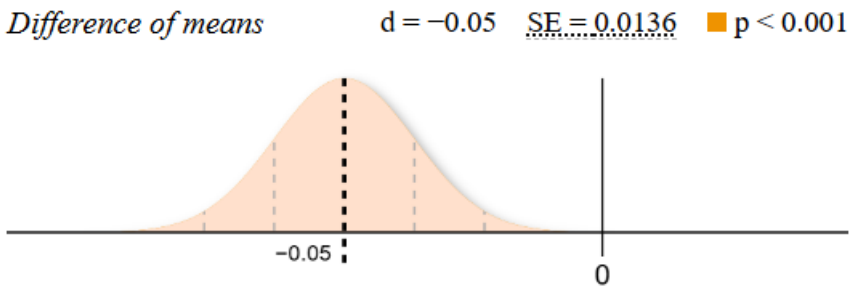


Figura 4.7 Diferencia de medias del tiempo de clasificación (Clásico vs Cuántico)

Los resultados confirman reducciones marcadas y significativas frente al proceso manual: Manual vs. Clásico $d = 3{,}10\ min.$,; Manual vs. Cuántico $d = 3{,}05\ min.$. Entre modelos, Clásico es ≈3 s más rápido que Cuántico $d = -0{,}05\ min.$ En suma, ambos automáticos aceleran sustancialmente el flujo, con una ligera ventaja de Clásico.

La Tabla 4.6 resume la mejora relativa de tiempo en emociones y sus p-values.

Tabla 4.6 Reducción relativa de tiempo en emociones y significancia

Caracterización	Comparativo	Δ Media	Reducción relativa (%)	p-value
Emociones	Manual vs Clásico	-3,10	88,57%	< 0,001
	Manual vs Cuántico	-3,05	87,14%	< 0,001
	Cuántico vs Clásico	0,05	11,11%	< 0,001

Los métodos automáticos superan significativamente al Manual (-88,57% y -87,14%, $p < 0,001$). Entre modelos, la diferencia es pequeña pero significativa con 11,11% y $p < 0,001$), con ligera ventaja para Clásico.

c) Percepción

La Tabla 4.7 presenta las medias de Utilidad, Confianza y Facilidad (escala 1–5) y el Índice_UX para cada método, a partir de las 100 encuestas aplicadas en producción.

Tabla 4.7 Resultados de encuestas Likert para la clasificación de emociones

Método	Utilidad	Confianza	Facilidad	Indice_UX
Automático – Clásico	4,54	4,23	4,14	4,34
Automático – Cuántico	4,23	3,8	4,14	4,04
Manual	3,03	2,97	3,13	3,03

Se observa una preferencia clara por los enfoques automáticos: el Automático–Clásico obtiene el Índice_UX más alto (4,34) y lidera en las tres dimensiones, seguido por Automático–Cuántico (4,04); el Manual (3,03) queda rezagado, especialmente en Confianza. Estos resultados apoyan y respaldan la adopción operativa de los métodos automáticos.

4.2.2 Resultados para la clasificación de veracidad

a) Precisión

La Tabla 4.8 resume el desempeño de veracidad en producción. Para cada método (Manual, Automático–Clásico, Automático–Cuántico) se reportaron audios evaluados, aciertos y precisión.

Tabla 4.8 Precisión en veracidad por método

Tarea	Método	# de audios	Aciertos	Precisión (%)
Veracidad	Manual (estado del arte ≈ 65%)	200	129	64,50%
	Automático – Clásico	200	171	85,50%
	Automático – Cuántico	200	160	80,00%

En las Figuras 4.8, 4.9 y 4.10 se ilustran las curvas A/B de precisión en veracidad con el umbral del 95% de confianza, comparando Manual vs. Automático–Clásico, Manual vs. Automático–Cuántico y Clásico vs. Cuántico. El solapamiento entre distribuciones permite apreciar visualmente la magnitud y dirección del efecto.

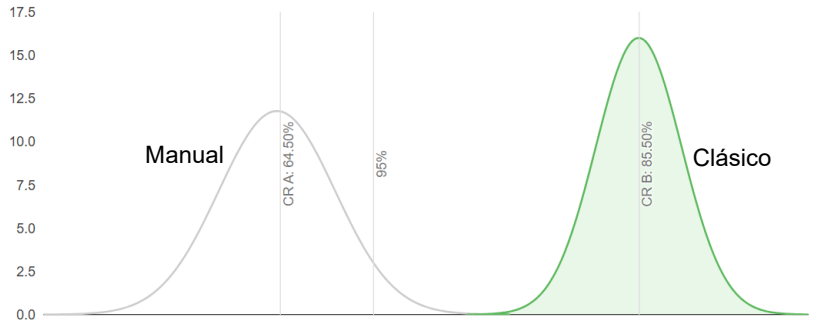


Figura 4.8 Curvas A/B de precisión en veracidad (Manual vs Clásico)

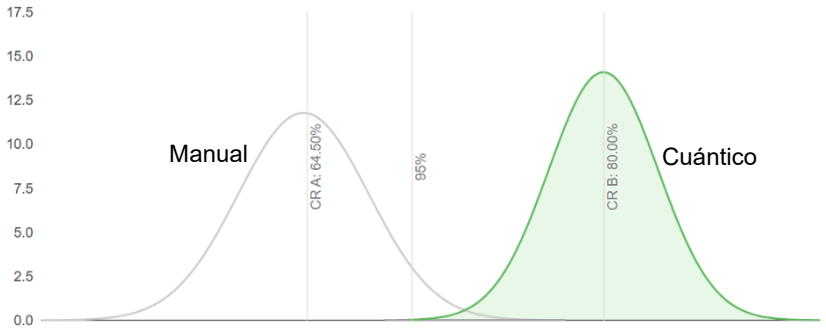


Figura 4.9 Curvas A/B de precisión en veracidad (Manual vs Cuántico)

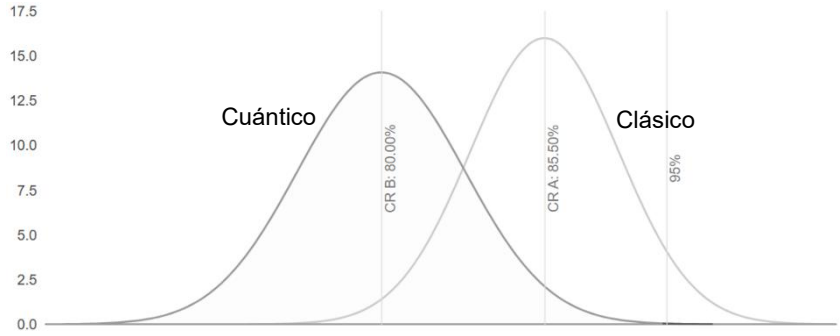


Figura 4.10 Curvas A/B de precisión en veracidad (Cuántico vs Clásico)

Se observa un desplazamiento nítido a favor de los métodos automáticos frente al Manual: Clásico 85,5% vs 64,5% y Cuántico 80,0% vs 64,5%. En la comparación Clásico vs Cuántico (85,5% vs 80,0%) la diferencia de +5,5 p.p. con poca significancia.

La Tabla 4.9 resume la mejora relativa de precisión en veracidad y sus p-values.

Tabla 4.9 Mejora relativa de precisión en veracidad y significancia

Caracterización	Comparativo	Mejora Relativa (%)	p-value
Veracidad	Manual vs Clásico	32,56%	0.0000
	Manual vs Cuántico	24,03%	0.0002
	Cuántico vs Clásico	6,87%	0.9278

Los métodos automáticos superan significativamente al Manual (+32,56% y +24,03%, p ≤ 0,0002); Clásico vs. Cuántico no difiere (p = 0,9278). En la práctica, Clásico es preferible por mayor mejora y menor tiempo.

b) Tiempo

La Tabla 4.10 presenta las estadísticas descriptivas del tiempo por audio en veracidad para cada método, reportando media y desviación estándar con el mismo tamaño muestral por brazo.

Tabla 4.10 Tiempo de clasificación en veracidad por método

Tarea	Método	# audios	Media de clasificación (min.)	Desviación estándar
Veracidad	Manual (estado del arte ≈ 65%)	200	2,50	1,20
	Automático – Clásico	200	0,30	0,08
	Automático – Cuántico	200	0,32	0,09

El Manual tarda 2,50 min, mientras que Clásico y Cuántico operan en 0,30 y 0,32 min., respectivamente—casi un orden de magnitud más rápidos.

En las Figuras 4.11, 4.12 y 4.13 se ilustra la diferencia de medias del tiempo por audio en veracidad (prueba t de Welch, IC95%). Aquí $d = \bar{x}_{S1} - \bar{x}_{S2}$: valores positivos indican que el Manual tarda más que el Automático; en Clásico vs. Cuántico, $d < 0$ significa que Clásico es más rápido.

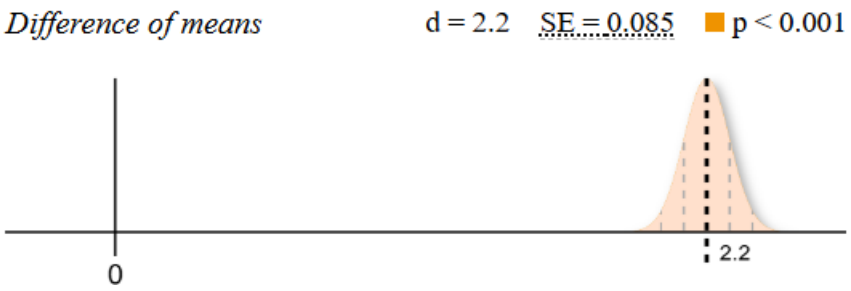


Figura 4.11 Diferencia de medias del tiempo de clasificación veracidad (Manual vs Clásico)

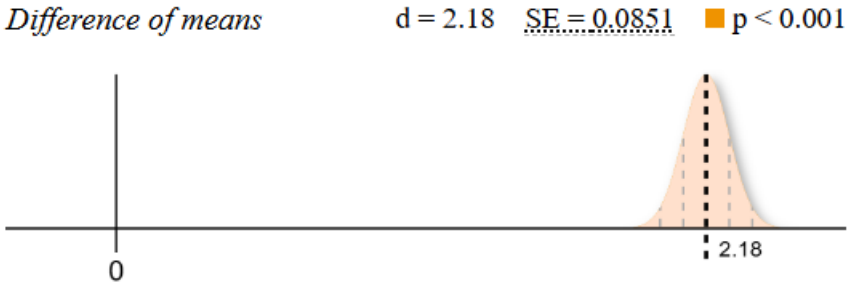


Figura 4.12 Diferencia de medias del tiempo de clasificación veracidad (Manual vs Cuántico)

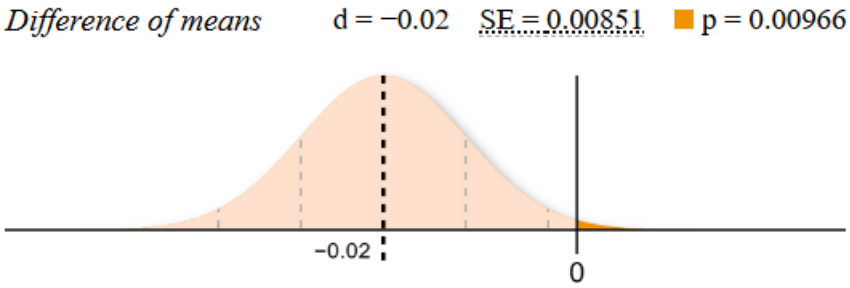


Figura 4.13 Diferencia de medias del tiempo de clasificación veracidad (Clásico vs Cuántico)

Los contrastes confirman reducciones muy marcadas frente al proceso manual: Manual vs. Clásico con $d = 2{,}20\ min.$, y Manual vs. Cuántico con $d = 2{,}18\ min$. Entre modelos, el Clásico resulta ligeramente más rápido que el Cuántico con $d = -0{,}02\ min.$ La Tabla 4.11 reporta la diferencia de medias del tiempo por audio (Δ= B−A) y la reducción relativa frente al método de referencia, con p-values.

Tabla 4.11 Reducción relativa de tiempo en veracidad y significancia

Caracterización	Comparativo	Δ Media	Reducción relativa (%)	p-value
Veracidad	Manual vs Clásico	-2,20	88,00%	< 0,001
	Manual vs Cuántico	-2,18	87,20%	< 0,001
	Cuántico vs Clásico	0,02	6,25%	0,0096

Tanto Clásico como Cuántico son mucho más rápidos que el Manual (−2,20 y −2,18 min; 88,0% y 87,2%, p<0,001). Entre modelos, Cuántico es ligeramente más lento que Clásico (+0,02 min ≈ 1,2 s; 6,25%, p=0,0096), por lo que operativamente Clásico resulta preferible.

c) Percepción

La Tabla 4.12 resume las medias Likert (1–5) de Utilidad, Confianza y Facilidad y el Índice_UX para la tarea de veracidad, a partir de las 100 encuestas del piloto (5 centros).

Tabla 4.12 Resultados de encuestas Likert para la clasificación de veracidad

Método	Utilidad	Confianza	Facilidad	Indice_UX
Automático – Clásico	4,2	4,31	4,17	4,24
Automático – Cuántico	4,34	3,8	3,86	4,03
Manual	3,2	3	3,43	3,17

Ambos enfoques automáticos superan al manual en percepción; Clásico obtiene el mejor balance (Índice_UX 4,24, con Confianza 4,31), mientras que Cuántico exhibe Utilidad 4,34 pero menor confianza (3,80). Las diferencias frente a Manual son significativas, mientras que entre automáticos no hay evidencia concluyente.

4.2.3 Resultados para Identificación del Hablante

a) Precisión

La Tabla 4.13 resume el desempeño en identificación de hablante comparando el método actual (escucha humana/KBA) con el Automático, reportando audios evaluados, aciertos y precisión.

Tabla 4.13 Precisión en identificación del hablante por método

Tarea	Método	# de audios	Aciertos	Precisión (%)
Identificación de hablante	Método actual (solo con escucha humana)	200	74	37,00%
	Automático	200	183	91,50%

El método automático alcanza 91,5% frente a 37,0% del método actual (+54,5 p.p.; +147,3% rel.), diferencia contundente y alineada con las pruebas A/B en producción (prueba de dos proporciones, p<0,001).

En la Figura 4.14 se ilustran las curvas A/B de precisión para identificación de hablante, con el umbral del 95% de confianza. Se contrasta el método actual (control) frente al Automático.

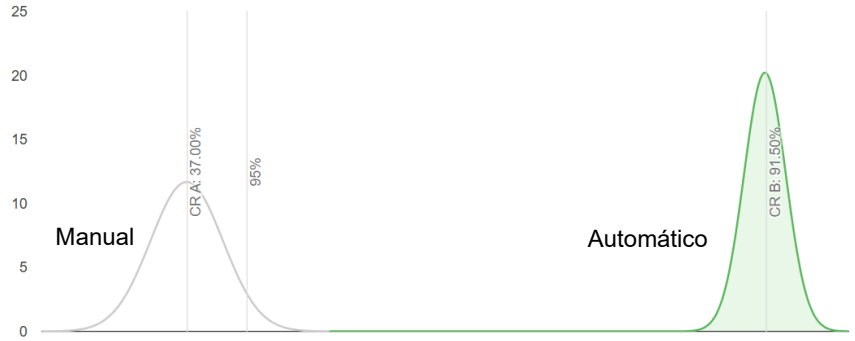


Figura 4.14 Curvas A/B de precisión en Identificación del Hablante (Manual vs Automático)

La separación de las distribuciones es categórica: Automático 91,5% vs método actual 37,0% (+54,5 p.p.; +147,3% rel.).

La Tabla 4.14 presenta el uplift relativo de precisión al comparar el método actual con el Automático, junto con su p-value (IC95%, prueba de dos proporciones).

Tabla 4.14 Mejora relativa de precisión en identificación de hablante

Caracterización	Comparativo	Mejora Relativa (%)	p-value
Identificación de hablante	Método actual vs Automático	147,30%	0.0000

El Automático obtiene una mejora relativa del 147,30% con evidencia concluyente ($p<0,001$), lo que respalda la migración operativa hacia el sistema automático para la identificación de hablante.

b) Tiempo

La Tabla 4.15 reporta el tiempo de identificación de hablante para el método actual (escucha humana) y el Automático, indicando media y desviación estándar.

Tabla 4.15 Tiempo de identificación del hablante por método

Tarea	Método	# audios	Media de clasificación (min.)	Desviación estándar
Identificación de hablante	Método actual (solo con escucha humana)	200	0,60	0,20
	Automático	200	0,12	0,04

El método automático procesa en 0,12 ± 0,04 min frente a 0,60 ± 0,20 min. del método actual—una reducción cercana al 80% del tiempo por audio—coherente con los contrastes estadísticos y el beneficio operativo observado en producción.

La Figura 4.15 muestra la prueba de Welch para el tiempo por audio en identificación de hablante, graficando la distribución de la diferencia de medias indica que el método actual es más lento.

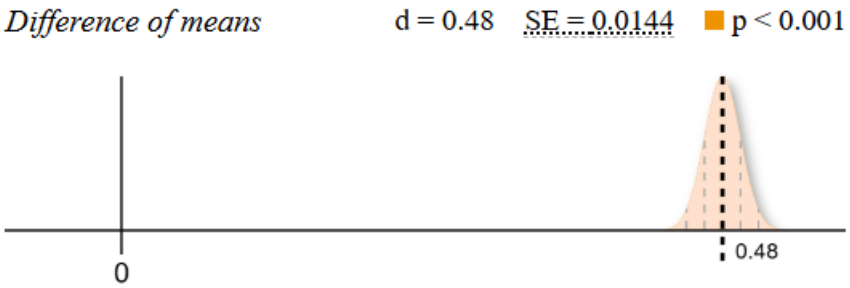


Figura 4.15 Diferencia de medias del tiempo de identificación del hablante (Manual vs Automático)

La diferencia de media de 0,48 min (SE = 0,0144) implica un ahorro de ~29 s por audio y una reducción del tiempo ≈80% respecto del método actual, confirmando la ventaja operativa del enfoque Automático.

La Tabla 4.16 resume, para identificación de hablante, la diferencia de medias del tiempo por audio y la reducción relativa al comparar método actual vs Automático.

Tabla 4.16 Reducción relativa de tiempo de identificación del hablante y significancia

Caracterización	Comparativo	Δ Media	Reducción relativa (%)	p-value
Identificación de hablante	Método actual vs Automático	-0,48	80,00%	< 0,001

Se obtuvo una diferencia de media de −0,48 min (≈80% de reducción del tiempo a favor del Automático), con evidencia estadística muy sólida ($p < 0,001$), lo que respalda su adopción operativa.

c) Percepción

La Tabla 4.17 presenta las medias Likert (1–5) de Utilidad, Confianza y Facilidad y el Índice_UX para la tarea de identificación del hablante. Los datos provienen de 100

encuestas aplicadas en los cinco centros del despliegue canario (supervisores, auditores y agentes).

Tabla 4.17 Resultados de encuestas Likert para la identificación del hablante

Método	Utilidad	Confianza	Facilidad	Indice_UX
Automático	4,41	4,33	4,19	4,33
Método actual	3,13	3,03	3,3	3,13

El Automático lidera en todas las dimensiones y alcanza un Índice_UX 4,33 frente a 3,13 del método actual, con una brecha marcada en Confianza. Esta percepción positiva es consistente con las mejoras objetivas de precisión y tiempo observadas en producción.

4.2.4 Síntesis de resultados: índice compuesto de desempeño en producción

La tabla integra precisión, tiempo por audio y percepción (Índice_UX) en un único Score aritmético en [0,1], tras normalización min–max por tarea y ponderaciones $(w_p,\ w_t,\ w_l)$ específicas (emociones/veracidad: 0,60/0,25/0,15; identificación: 0,70/0,20/0,10). Un Score mayor implica mejor desempeño global del método bajo condiciones de producción.

Tabla 4.18 Índice compuesto (media aritmética ponderada) por tarea y método

Tarea	Método	Precisión	Tiempo (min)	Likert (índice_ UX)	U_p	U_t	U_l	w_p	w_t	w_l	Score (aritm.)
Emociones	Manual	0,76	3,50	3,18	0,00	0,00	0,55	0,60	0,25	0,15	0,08
Emociones	Automático - Clásico	0,84	0,40	4,21	1,00	1,00	0,80	0,60	0,25	0,15	0,97
Emociones	Automático - Cuántico	0,82	0,45	4,11	0,69	0,98	0,78	0,60	0,25	0,15	0,78
Veracidad	Manual	0,65	2,50	3,18	0,00	0,00	0,55	0,60	0,25	0,15	0,08
Veracidad	Automático - Clásico	0,86	0,30	4,21	1,00	1,00	0,80	0,60	0,25	0,15	0,97
Veracidad	Automático - Cuántico	0,80	0,32	4,11	0,74	0,99	0,78	0,60	0,25	0,15	0,81
Identificación	Manual	0,37	0,60	3,18	0,00	0,00	0,55	0,70	0,20	0,10	0,05
Identificación	Automático	0,92	0,12	4,21	1,00	1,00	0,80	0,70	0,20	0,10	0,98

Los resultados son consistentes en las tres tareas: Automático–Clásico obtiene el mejor desempeño en emociones y veracidad (≈ 0,97), mientras que en identificación el Automático alcanza 0,98. El Automático–Cuántico queda en segundo lugar (≈ 0,78–0,81) y el Manual es claramente inferior (0,05–0,08), lo que respalda la adopción operativa de los enfoques automáticos.

Las Figuras 4.16, 4.17 y 4.18 muestran gráficos de barra del índice compuesto (0–1) por tarea —emociones, veracidad e identificación— calculado con la media aritmética ponderada de precisión, tiempo e Índice_UX (a mayor valor, mejor desempeño global).

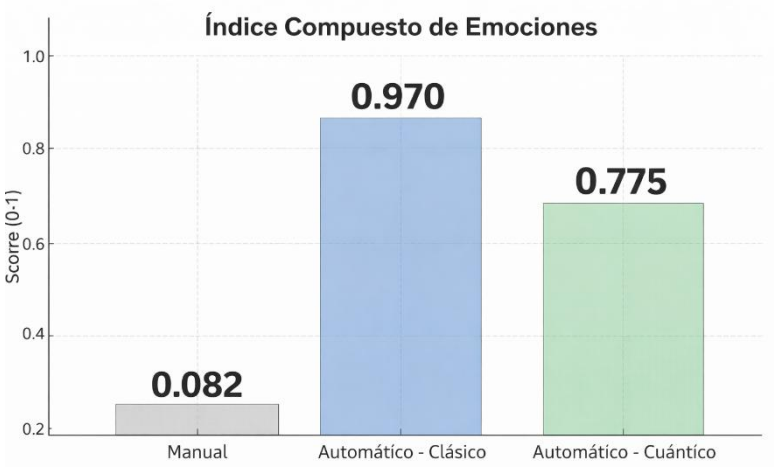


Figura 4.16 Índice compuesto en emociones (producción)

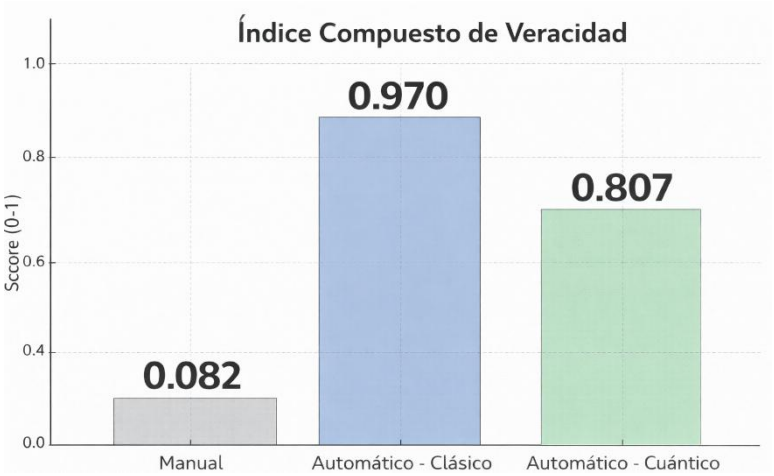


Figura 4.17 Índice compuesto en veracidad (producción)

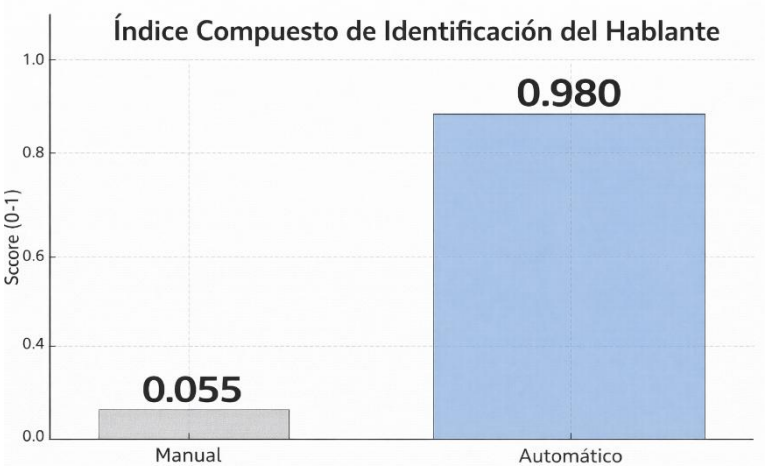


Figura 4.18 Índice compuesto en identificación del hablante

El patrón es consistente: el Automático–Clásico alcanza el mejor score en emociones y veracidad (≈ 0,97), seguido por el Cuántico (≈ 0,78–0,81), mientras el Manual queda muy por debajo (≈ 0,05–0,08). En identificación, el Automático es claramente superior (≈0,98 vs ≈0,05). En consecuencia, se recomienda operar con el modelo clásico como base para emociones/veracidad (con el cuántico como alternativa/backup) y automatizar la identificación del hablante; el proceso manual queda como referencia y auditoría.

4.3 Análisis Costo/Beneficio

4.3.1 Indicadores Económicos

Para determinar la factibilidad económica de la solución propuesta, se utilizó el indicador Retorno sobre la Inversión (ROI), definido como:

$$ROI = \frac{Beneficio\ Neto}{Inversión\ Inicial + Costo\ de\ Operación}\ x\ 100 \quad (26)$$

Donde el Beneficio Neto corresponde al balance entre los ingresos producidos por el sistema y los costos asociados a su puesta en funcionamiento. Se considerarán también métricas complementarias como el tiempo de retorno de la inversión junto con Valor Actual Neto (VAN) para tener una visión más completa del impacto financiero. La fórmula utilizada para el VAN fue:

$$VAN = \sum_{t=1}^{n} \frac{F_t}{(1+r)^t} - C_0 \quad (27)$$

Donde:

F_t = Flujo neto de caja durante el período t.

r = Tasa de descuento.

t = Número de períodos (años)

n = Número total de períodos.

C_0 = Inversión inicial.

4.3.2 Valores Considerados

Para el cálculo, se tomaron en cuenta los siguientes elementos:

Costos de Implementación:

- Desarrollo del módulo de caracterización conductual.
- Integración con el software DINOMI.
- Licencias, infraestructura y despliegue.
- Capacitación inicial del personal.

Beneficios Cuantificables:

- **Ahorro de tiempo operativo por reducción del tiempo de clasificación**:
  - Ejemplo: Identificación del hablante → reducción del 80% (0,60 min vs 0,12 min por audio).
  - Emociones y veracidad → reducciones de entre 70% y 85%.
- **Incremento de precisión que reduce errores y reprocesos**:
  - Emociones: +8,0 p.p. (clásico) y +5,5 p.p. (cuántico) sobre el manual.
  - Veracidad: mejoras similares, optimizando detección de engaños.
- **Productividad liberada**:
  - En un centro con 10.000 llamadas/mes, un ahorro de 29 s por llamada equivale a ≈80 horas hombre/mes.
- **Mejora en la experiencia de cliente y reducción de tiempos de espera**.
- **Reducción de fraude gracias a la identificación temprana del hablante**.

4.3.3 Resultados y conclusiones

Se calcularon los costos recurrentes asociados a la operación del sistema, considerando tanto el recurso humano necesario como los servicios en la nube requeridos para su funcionamiento. La Tabla 4.19 describe estos costos.

Tabla 4.19 Costos estimados de operación e infraestructura del sistema propuesto

Concepto	Mensual (USD)	Anual (USD)
Sueldo Ingeniero en Sistemas	$1.250,00	$15.000,00
S3 (200 GB) almacenamiento	$4,60	$55,20
AWS Lambda (100k invocaciones/mes)	$5,00	$60,00
AWS API Gateway (100k llamadas/mes[6])	$0,35	$4,20
Total, mensual	**$1.259,95**	**$15.119,40**

[6] 100k llamadas mensuales en promedio

Como se observa, el mayor componente del costo corresponde al sueldo del ingeniero en sistemas encargado de la administración y mantenimiento del sistema, seguido por los servicios en la nube como almacenamiento S3, funciones Lambda y API Gateway. Estos costos, al mantenerse constantes a lo largo del período de análisis, permiten proyectar de forma estable el gasto operativo anual y facilitan la estimación de indicadores como el ROI y el período de retorno de la inversión.

ROI

Una vez determinados los costos operativos anuales, se procedió a estimar los beneficios económicos que el sistema generará, tanto por la reducción de tiempos operativos en la atención como por la disminución de tareas repetitivas. Con esta información se obtuvieron indicadores financieros como el Retorno sobre la Inversión (ROI) y el Período de Recuperación, fundamentales para evaluar la viabilidad financiera del proyecto. En la Tabla 4.20 se resumen dichos cálculos.

Tabla 4.20 Resumen de cálculo del ROI del sistema propuesto

Concepto	Valor estimado (USD)	Explicación
Costo de implementación	$15.119	Cotización inicial del proyecto
Costo anual de operación	$800	Mantenimiento y soporte
Costo total anual:	**$15.919**	
Beneficio anual por ahorro de tiempo	$34.560	360 h-h/mes × USD 8/h × 12 meses
Beneficio anual por reducción de reprocesos	$8.000	Estimación por mejora de reducción de tiempo
Beneficio total anual	**$42.560**	Suma de beneficios
Beneficio neto anual	**$26.641**	Beneficio total anual - Costo total anual
ROI anual (%)	**167%**	
Período de recuperación	**0,57**	años

Los resultados evidencian que el sistema presenta un ROI anual del 167 %, lo que indica que la inversión inicial se recupera ampliamente en el primer año de operación. Además, el período de recuperación estimado es de apenas 0,57 años, reflejando que la implementación genera beneficios netos de forma temprana y sostenida, fortaleciendo la justificación económica del proyecto.

VAN

Posteriormente, se realizó la estimación del Valor Actual Neto (VAN) aplicando una tasa de descuento del 10 %, a fin de evaluar el rendimiento financiero del proyecto teniendo en cuenta la pérdida de valor del dinero a lo largo del tiempo. Este indicador permite determinar si los beneficios proyectados, descontados a valor presente, superan la inversión inicial y los costos operativos asociados.

Tabla 4.21 Cálculo del Valor Actual Neto (VAN) del sistema propuesto

Año	Flujo Neto (USD)	Factor de Descuento (10%)	Valor Presente (USD)
1	$26.641	0,9091	$24.219,09
2	$41.760	0,8264	$34.512,40
3	$41.760	0,7513	$31.374,91
4	$41.760	0,683	$28.522,64
5	$41.760	0,6209	$25.929,67

VAN TOTAL: $144.558,71

El análisis muestra que el VAN total asciende a USD 144.558,71, lo que evidencia una rentabilidad significativa y sostenida a lo largo de los cinco años de proyección. Este valor positivo confirma la factibilidad del proyecto recuperando la inversión inicial y también generando beneficios adicionales considerables, respaldando su implementación desde una perspectiva financiera.

Los resultados del análisis financiero confirman la viabilidad y alta rentabilidad del sistema propuesto. El Retorno sobre la Inversión (ROI) anual alcanza el 167 %, evidenciando que la inversión inicial se recupera con creces en el primer año de operación. De forma complementaria, el Valor Actual Neto (VAN), utilizando una tasa de descuento del 10 %, asciende a USD $144.558,71, lo que refleja un beneficio económico acumulado significativo a lo largo de los cinco años de proyección. Estos indicadores, sumados al corto período de recuperación de 0,57 años, demuestran que el proyecto además de recuperar la inversión en un corto período de tiempo también produce flujos netos positivos de manera sostenida, justificando su implementación desde una perspectiva económica y estratégica.

4.4 Aporte y Futuros trabajos de la solución propuesta

La solución desarrollada aporta un enfoque innovador para la caracterización conductual de interacciones orales en centros de contacto, integrando modelos clásicos y cuánticos en un marco de análisis multimodal. Entre sus principales contribuciones destacan:

- La automatización de tareas críticas como la detección de emociones, la verificación de veracidad y la identificación del hablante, reduciendo significativamente el tiempo de procesamiento y los errores asociados a la clasificación manual.
- La incorporación de técnicas de computación cuántica (VQC) en un entorno híbrido, lo que sienta un precedente para la aplicación de esta tecnología emergente en entornos productivos del sector de atención al cliente.
- La mejora de la experiencia del cliente mediante la optimización de tiempos de respuesta y la reducción de reprocesos, con impacto directo en el desempeño operativo y la percepción del servicio.
- La generación de un modelo replicable y escalable, adaptable a otros contextos y sectores que requieran análisis conductual y verificación de identidad.

En cuanto a líneas de trabajo futuras, se identifican las siguientes:

- Integrar nuevas fuentes de datos multimodales, como análisis facial o biometría de voz avanzada, para complementar las predicciones actuales.
- Evaluar arquitecturas cuánticas más avanzadas, optimizadas para hardware cuántico real, a fin de mejorar la exactitud y acortar el tiempo de entrenamiento.
- Ampliar el alcance del sistema a múltiples idiomas y entornos culturales, garantizando un desempeño robusto ante variaciones lingüísticas y contextuales.
- Incorporar mecanismos de aprendizaje continuo que permitan adaptar el modelo a cambios en patrones de comunicación y comportamiento a lo largo del tiempo.
- Desarrollar una versión ligera del sistema para su despliegue en dispositivos de borde (*edge computing*), reduciendo la dependencia de infraestructura en la nube.

El conjunto de aportes y proyecciones expuestos en este apartado refuerza el valor estratégico y técnico de la solución propuesta. Sobre esta base, en el próximo capítulo se recogen las conclusiones generales de este trabajo, donde se sintetizan los hallazgos, se contrastan con los objetivos planteados y se formulan recomendaciones derivadas del trabajo realizado.

CAPÍTULO 5

5 DISCUSIÓN

Los resultados presentados en el Capítulo 4 ofrecen una comprensión precisa del rendimiento de los modelos tradicionales y cuánticos en la caracterización conductual de llamadas en centros de contacto. Un hallazgo importante es que el modelo cuántico obtuvo valores muy cercanos al modelo clásico, lo que confirma que los circuitos cuánticos variacionales (VQC) ya pueden abordar tareas complejas como la clasificación de emociones y veracidad en voz, con resultados comparables a técnicas tradicionales. Esto respalda el potencial de la computación cuántica dentro del aprendizaje automático.

En emociones, el modelo clásico alcanzó un 84 %, mientras que el cuántico obtuvo 81,5 %. En veracidad, los valores fueron 85,5 % y 80 % respectivamente. Aunque las diferencias no siempre tuvieron una alta significancia, sí se pudo observar una tendencia a favor del modelo clásico. En cuanto al tiempo de análisis, ambos redujeron notablemente la duración respecto al proceso manual, con una ligera ventaja del enfoque clásico: 0,40 frente a 0,45 minutos en emociones y 0,30 frente a 0,32 minutos en veracidad. En este caso, las diferencias sí resultaron significativas.

Más allá de las métricas, la incorporación de VQC aporta un valor estratégico. Permite explorar una tecnología emergente con potencial de crecimiento, genera una línea base útil para futuros avances de hardware y algoritmos, introduce un componente innovador dentro de los sistemas de análisis de voz y plantea la posibilidad de que la ventaja cuántica se manifieste en escenarios con problemas más complejos o con datasets más extensos.

En conjunto, aunque el modelo cuántico no superó al clásico en este estudio, su evaluación aporta conocimiento valioso sobre el estado actual del aprendizaje automático cuántico y sobre su aplicabilidad en entornos reales de análisis de voz.

CAPÍTULO 6

6 CONCLUSIONES Y RECOMENDACIONES

Este trabajo desarrolló una solución completa de caracterización conductual para interacciones orales en español, capaz de detectar emociones, verificar veracidad e identificar hablantes dentro de un mismo flujo. La solución integró modelos clásicos y cuánticos en un pipeline híbrido que empleó procesamiento de voz, reducción de dimensionalidad y codificación cuántica.

Los resultados mostraron aportes relevantes. Se obtuvo la primera implementación documentada en español que combina emociones y veracidad con un enfoque híbrido clásico–cuántico. En rendimiento, el modelo clásico logró 90,93 % en emociones y 88,89 % en veracidad; el cuántico alcanzó 81 % y 88,89 %. Aunque el clásico mantuvo ligera ventaja en emociones, ambos modelos fueron estables y consistentes.

Operativamente, el sistema redujo el tiempo de análisis entre 70 % y 85 %, con un 80 % en identificación de hablantes, mostrando mejoras claras frente al proceso manual. En términos económicos, los resultados fueron favorables: ROI anual de 167 %, payback de 0,57 años y VAN de USD $144.558,71. Además, la biometría vocal y la detección automática de emociones redujeron reprocesos y fortalecieron la seguridad.

El proyecto cumplió los objetivos: se construyó un sistema híbrido funcional, se compararon ambos modelos y se validó su comportamiento en un entorno controlado. Para continuar este trabajo se recomienda evaluar el sistema con datos reales no controlados, incorporar modalidades adicionales (video o rostro), optimizar los circuitos para hardware físico, explorar nuevos ansatz e incorporar aprendizaje continuo.

En conjunto, este estudio constituye un aporte pionero en la intersección entre IA y computación cuántica aplicada a la voz en español. Aunque la tecnología cuántica aún no supera al enfoque clásico, los resultados evidencian un potencial competitivo y económicamente atractivo, abriendo nuevas oportunidades para el análisis conductual automatizado.

7 REFERENCIAS BIBLIGRÁFICAS

Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The Power of Quantum Neural Networks. *Nature Computational Science, 1*, 403–409.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 21-62.

Benedetti, M. (2019). Parameterized quantum circuits as machine learning models. *Quantum Science and Technology, 4*(4). doi:043001

Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V., . . . Killoran, N. (2022). *PennyLane: Automatic differentiation of hybrid quantum-classical computations.* arXiv. doi:1811.04968

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., & Maclaurin, D. (2018). JAX: composable transformations of Python+NumPy programs. https://github.com/google/jax.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Campbell, J., Reynolds, D., & Torres-Carrasquillo, P. (2010). Speaker recognition: A tutorial. *Proceedings of the IEEE, 95*(12), 1976–1997. Retrieved from https://doi.org/10.1109/JPROC.2007.909557

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. *ISCA Workshop on Speech and Emotion*, (pp. 19-24). Belfast, UK.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10-20.

Deepa, P., & Kuppusamy, R. (2022). Speech technology in healthcare. *Measurement: Sensors, 1*, 1-11.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems* (pp. 1-15). Springer.

El Ayadi, M., Kamel, M., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44(3)*, 572–587. Retrieved from https://doi.org/10.1016/j.patcog.2010.09.020

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich open-source large-scale multimedia feature extractor. *ACM Transactions on Multimedia Computing Communications and Applications, 6*(3), 1-5.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics, 29*(5), 1189–1232.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501. Retrieved from https://doi.org/10.1016/S1364-6613(98)01262-5

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M., Scherer, K., & Vuilleumier, P. (2006). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience, 9*(3), 389-395. Retrieved from https://doi.org/10.1038/nn1637

Grinberg, M. (2022). *The Flask Mega-Tutorial.* Retrieved from [Blog]: https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world

Gumá, V. A. (2001). *Texto de neurociencias cognitivas.* Editorial El Manual Moderno.

Hernández, J. I. (2021). Aplicaciones de técnicas de aprendizaje automático para la mejora de la atención al cliente en centros de contacto. *Revista de Investigación en Tecnologías de la Información, 9*(2), 45-56.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer. Retrieved from https://doi.org/10.1007/b98835

Kalloniatis, A., & Kontopoulos, E. (2024). Computational humor recognition: a systematic literature review. *Springer Nature, 10*(1), 1-23.

Kamm, C. A. (1997). The role of speech recognition in call processing. *Proceedings of the IEEE, 85*(10), 1314–1338.

Mann, H., & Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50-60. Retrieved from https://www.jstor.org/stable/2236101

Maza, B. E.-B. (2011). On the use of linguistic features in an automatic system for speech analytics of telephone conversations. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)* (pp. 2049-2052). https://www.isca-speech.org/archive/interspeech_2011/i11_2049.html.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5(4)*, 115–133. Retrieved from https://doi.org/10.1007/BF02478259

McFee, B., Raffel, C., Liang, D., & Ellis, D. P. (2015). librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, (pp. 18-24).

Mitarai, K., Negoro, M., Kitagawa, M., & Fujii, K. (2018). Quantum circuit learning. *Physical Review A, 98*(3). doi:032309

Nielsen, M., & Chuang, I. (2010). Quantum computation and quantum information. *Cambridge University Press*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pell, M. D. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37*(4), 417-435.

Pell, M., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. (2015). Prosody and the brain: The neuroanatomy of prosodic processing. *Emotion Review, 7*(2), 135-143. Retrieved from https://doi.org/10.1177/1754073914568181

Pepino, P., Riera, P., Cullen, C., & Saraceno, C. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *Proceedings of the Annual Conference of the International Speech Communication Association*, (pp. 671–675).

PSD, S. (2025). Retrieved from https://www.dinomi.com

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *In Advances in Neural Information Processing Systems*, *35.* Retrieved from https://openai.com/research/whisper

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3). doi:e0118432

Schuld, M., & Killoran, N. (2019). Quantum machine learning in feature Hilbert spaces. *Physical Review Letters, 122*(4). doi:040504

Schuld, M., Bocharov, A., Svore, K., & Wiebe, N. (2020). Circuit-centric quantum classifiers. *Physical Review A, 101*(3). doi:032308

Schuld, M., Bocharov, A., Svore, K., & Wiebe, N. (2020). Circuit-centric quantum classifiers. *Physical Review A, 101*(3). Retrieved from https://arxiv.org/pdf/1804.00633

Siegman, A. W. (1993). Voices of Fear and Anxiety and their Acoustic Correlates. *Journal of Abnormal Psychology, 102*(2), 313-318.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3–14. Retrieved from https://doi.org/10.1016/j.imavis.2017.08.003

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5200–5204. Retrieved from https://doi.org/10.1109/ICASSP.2016.7472665

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59(236)*, 433–460. Retrieved from https://doi.org/10.1093/mind/LIX.236.433

Welch, B. (1947). The generalization of “Student’s” problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35. Retrieved from https://doi.org/10.2307/2332510

Zhang, Z., Han, J., Deng, J., & Schuller, B. (2017). Leveraging high-level features for speaker emotion recognition. *Neurocomputing, 266*, 590–602. Retrieved from https://doi.org/10.1016/j.neucom.2017.05.047

APÉNDICE

Diseño experimental para generación de datos de veracidad/engaño

Con el objetivo de disponer de datos reales y etiquetados para entrenamiento y validación de modelos de clasificación de veracidad, se replicó un diseño experimental inspirado en el protocolo MU3D (Lloyd et al., 2019), validado por psicólogos cognitivos de la Universidad de Miami. Este diseño ha sido exitosamente utilizado en la literatura para construir bases de datos de verdades y mentiras en contextos multimodales.

Objetivo

Obtener grabaciones de voz que contengan historias verdaderas y falsas, con distintos tonos emocionales, contadas por participantes en condiciones controladas, replicando el enfoque MU3D.

Participantes

- Total: 300 participantes (en este caso, estudiantes voluntarios o colaboradores).
- Perfil: diversidad de género, edad y acento. Consentimiento informado requerido.

Condiciones experimentales

Cada participante grabó 4 audios:

- Verdad positiva
- Verdad negativa
- Mentira positiva
- Mentira negativa

Duración aproximada de cada grabación: 30 a 60 segundos.

Procedimiento

- Firma del consentimiento informado.
- Instrucciones generales (sin mencionar que se detectará mentira).
- Grabación individual, mismo micrófono, lugar controlado.
- Aleatorización del orden de grabación por participante.

Etiquetado

Cada archivo fue etiquetado manualmente con:

- ID participante
- Tipo de historia (V o M)
- Emoción (positiva/negativa)
- Transcripción del contenido (si aplica)

En la Tabla 7.1 se resume el diseño experimental.

Tabla 7.1 Detalles de diseño experimental

Elemento	Detalle
Participantes	300 (150 mujeres, 150 hombres)
Condiciones	Cada persona grabó 4 videos: 2 verdades y 2 mentiras
Temas de las historias	Situaciones sociales emocionalmente positivas y negativas
Control experimental	Misma duración, entorno, instrucciones, cámara y audio
Motivación para mentir	No se dio ninguna motivación económica
Medidas validadas	Las mentiras fueron evaluadas por humanos externos y modelos automáticos

Las etapas del experimento se especifican en la tabla 7.2

Tabla 7.2 Etapas del experimento

Etapa	Detalles
1. Reclutamiento	300 participantes, equilibrado por género y edad
2. Grabación	Cada cuenta dos historias (1 verdad / 1 mentira), duración constante
3. Etiquetado	Archivo CSV con audio, etiqueta

Al finalizar el proceso de recolección y etiquetado, se obtuvo un dataset estructurado, tal como se muestra en la Tabla 7.3. Cada registro contiene el nombre del archivo de audio, el identificador del participante, la veracidad de la historia relatada, el tipo de emoción expresada y una transcripción referencial del contenido narrado.

Tabla 7.3 Ejemplo de registros

Archivo	Participante	Veracidad	Emoción	Transcripción
audio_001.wav	P01	Verdad	Positiva	Historia contada
audio_002.wav	P01	Mentira	Negativa	Historia contada

Esta base constituye la entrada para el entrenamiento de modelos supervisados de veracidad.

Escalas de Likert para validación cualitativa del sistema en producción

Escala para percepción de utilidad

Valor Likert	Descripción en encuesta
1	Nada útil
2	Poco útil
3	Útil en algunos casos
4	Útil
5	Muy útil

Escala para confianza de resultados

Valor Likert	Descripción en encuesta
1	Nada confiable
2	Poco confiable
3	Algo confiable
4	Confiable
5	Muy confiable

Escala de facilidad de uso del sistema

Valor Likert	Descripción en encuesta
1	Muy difícil de usar
2	Difícil de usar
3	Ni fácil ni difícil
4	Fácil de usar
5	Muy fácil de usar