uc3m | Universidad Carlos III de Madrid

Grado Universitario
Doble Grado Ingeniería Informática y
Administración de Empresas (2021-2022)

Trabajo Fin de Grado

"Análisis de Sentimiento en Audio mediante Inteligencia Artificial orientado al idioma Español"

Javier Moncada Gutiérrez

Tutores:
Miguel Angel Patricio Guisado
Antonio Berlanga de Jesús

RESUMEN

En este proyecto se presenta un trabajo como proyecto de investigación de la cátedra entre la Universidad Carlos III de Madrid y el Grupo Masmovil.

El objetivo es crear una Inteligencia Artificial capaz de clasificar la emoción de los clientes en las llamadas dentro del Call Center. Pero la naturaleza agnóstica del problema permite enfocarlo desde una perspectiva más amplia, a partir de la base de clasificar emociones en audio para el idioma Español, es posible adaptar este trabajo a diferentes propósitos.

A partir de los principios del reconocimiento de emociones en el habla, se estudian distintas aproximaciones para transmitir el conocimiento de lo que representa una emoción a un modelo matemático. Además, se combinan los resultados con el sentimiento que transmite el aspecto textual haciendo uso de algoritmos de aprendizaje automático.

Además, este trabajo viene integrado en una interfaz de una aplicación de mensajería para probarlo en tiempo real.

Palabras clave:
IA: Inteligencia Artificial.
ML: Machine Learning.
SR: Speech Recognition.
PLN: Procesamiento del Lenguaje Natural.
STT: Speech To Text.
IVR: Interactive Voice Response.
MFCC: Mel Frequency Cepstral Coefficients.

ABSTRACT

This present document portraits a project carried out as part of the chair between the Carlos III University of Madrid and the Masmovil Group.

The proposed objective is to create an Artificial Intelligence capable of classifying the emotion of customers in calls inside the Call Center.

The agnostic nature of the problem allows to approach it from a broader perspective, from the basis of classifying emotions in audio for the Spanish language, it is possible to adapt this work for different purposes.

Based on the principles of emotion recognition in speech, different approaches are studied to convey the knowledge of what an emotion represents to a mathematical model. In addition, the results are combined with the sentiment conveyed by the textual aspect using machine learning algorithms.

Besides, this work is integrated in a messaging application interface to test it in real time.

DEDICATORIA

Como último trabajo de mi etapa universitaria, quiero agradecer a las personas que me han acompañado durante estos seis años.

Por un lado, a mis tutores Miguel Angel y Antonio, y a Kevin, sin los cuales este trabajo no habría visto la luz.

Por otro lado, a mis compañeros de carrera, ahora Presentes Ingenieros, de los cuales tanto he aprendido y encontrado siempre una mano.

Por último y no menos importante, a mis padres, amigos y a mi novia, por ser apoyo incondicional y haber confiado en mí.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

Desde la famosa columna "¿Qué pasa con el reconocimiento de voz?" escrita por John Robinson Pierce en la revista de la *Sociedad Acústica de América* en Junio de 1969 [1], el campo del reconocimiento del habla se consideró un problema demasiado complejo, el cuál no merecía los caudalosos fondos para los resultados que estaban obteniendo, en otras palabras, la posibilidad de refutar el *Entscheidungsproblem* de Turing [2] se alejaba cada vez más.

No fue hasta finales de los ochenta que se despertó el estudio en el campo gracias a la utilidad práctica del modelo lingüísitico generativo de n-gramas *Katz back-off*, el cual estimaba la probabilidad condicional de una palabra dada su historia en el n-grama, y permitía distinguir entre idiomas.

Desde entonces hasta los modelos del lenguaje más recientes como *GPT-3* [3] del estudio Open AI o *No Lenguage Left Behind* [4] de Meta con billones de parámetros, han ido apareciendo contínuas innovaciones y descubrimientos, como las Cadenas de Markov [5], las Redes Neuronales [6] y los famosos Transformers [7].

La Inteligencia Artificial es multidisciplinar, y existe un subcampo de estudio que trata en cuestión el reconocimiento del habla, conocido como *Speech Recognition (SR)*. Este aúna a profesionales de todo el campo de la informática y la lingüística computacional, para desarrollar metodologías y tecnologías que permitan el reconocimiento y la traducción del lenguaje hablado a texto.

Una de las especialidades de estudio más interesantes dentro del SR trata sobre el Procesamiento Emocional en el habla, el cuál comparte motivaciones con el área de Análisis de Sentimiento en el *Procesamiento del Lenguaje Natural (PLN)* gracias a las herramientas de translación de audio a texto *Speech To Text (STT)*.
En ambos casos, el objetivo consiste en el uso de la tecnología para ayudar a las personas en la identificación de las emociones humanas.

Este problema no es trivial, dado que ya sea a través de texto, un vídeo o un audio, si la precisión de las personas para reconocer las emociones de los demás es relativa, la definición lógica por tanto de las mismas es compleja.
Es por ello que por lo general, las aplicaciones que están surgiendo funcionan mejor si trabajan en un contexto concreto y utilizan múltiples tecnologías de trabajo.

Hasta la fecha, la mayor parte de los trabajos se han centrado en la automatización del reconocimiento de las expresiones faciales a partir de un *input* audiovisual, combinando ambas fuentes de información.
Asimismo, han experimentado un gran avance las investigaciones que extraen sentimientos a partir de un texto, ya que al ser más objetivo para categorizar las emociones basta con contar con un equipo de lingüistas que genere un corpus de palabras que representen cierta emoción en el contexto adecuado, por lo que el objeto de investigación se simplifica para el reconocimiento de emociones.

En cambio, para el trabajo que se presenta, el único *input* de partida es el audio, un campo de investigación más inexplorado.

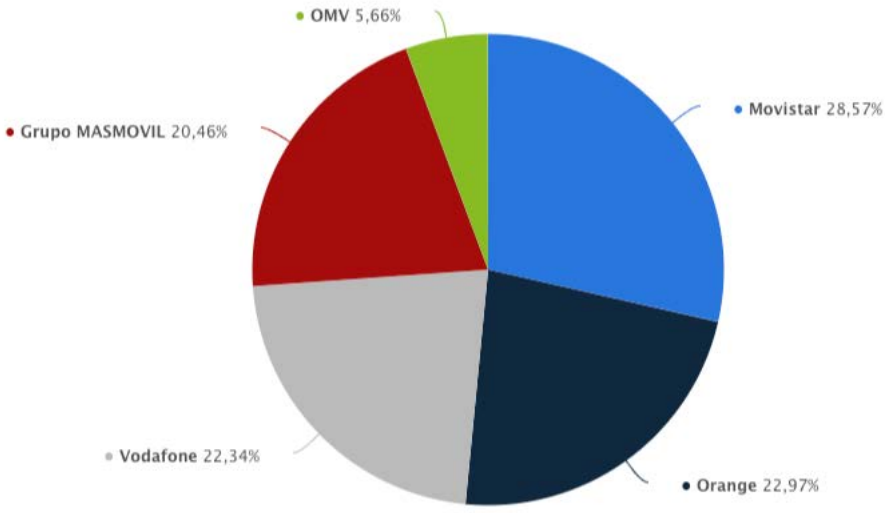
1.1. Motivación

En el contexto de la colaboración de Investigación entre la multinacional **Grupo Masmovil** y la **Universidad Carlos III de Madrid**, se plantea un problema.

Uno de los aspectos más importantes a tener en cuenta en una empresa, sobre todo en las que ofrecen un servicio como producto de pago recurrente, se trata de la atención al cliente. En este ámbito radica el éxito de la compañía en el medio largo plazo. En la industria de las telecomunicaciones y los servicios de telefonía móvil en España, el mercado es competitivo ya que la cuota de mercado está repartida entre unos pocos grandes competidores. Es por ello que el esfuerzo en inversión para retención del cliente dentro del grupo es esencial. En la siguiente imagen se muestra el reparto del mercado a diciembre de 2021, es importante destacar que a día de hoy en el Q3 de 2022 se está ultimando la fusión al 50 % de Orange y el Grupo Masmovil, lo que resultaría en un 43 % del total.

Dentro de la relación con el cliente, la mayoría de este tipo de compañías cuentan con un *Call Center*, para dar soporte a todos los usuarios actuales y potenciales. En este se encuentran agentes especializados en todos los aspectos de la compañía que puedan requerir asistencia de un profesional, desde la reparación técnica de un dispositivo hasta la modificación del contrato.

A pesar de las ventajas que supone para una empresa disponer de un centro de llamadas, hay ciertos problemas que no son fáciles de resolver, entre ellos la eficiencia en la gestión y las transferencias entre agentes. De cara a resolver este problema se implantó un sistema de voz interactiva (IVR) para dar asistencia primaria a los usuarios. Este sistema se encarga de recopilar una breve información de motivo de llamada, para la posterior selección de un agente. Si bien es cierto que el sistema funciona en su mayoría correctamente, hay margen de mejora si se decide utilizar un software de análisis avanzado, como el que se procura desarrollar en este trabajo.



Fuente: Statista.

Fig. 1.1. Cuota de mercado de las empresas de telefonía móvil en España.

La mejoría que se plantea aquí, consiste en asistir con antelación aquellos usuarios que se perciban con una emoción negativa, así como de informar al agente del estado del usuario previo al inicio de la llamada. Asimismo, de cara al estudio del comportamiento del cliente durante el proceso de la llamada, sería de utilidad un sistema de medición de las emociones en el transcurso de la misma.

Pese a ser una propuesta orientada a este contexto, el trabajo supone un proyecto agnóstico que puede reorientarse a ser de multi-propósito.

1.2. Objetivos

El objetivo de este proyecto consiste en la implementación de una Inteligencia Artificial (IA) que permita, a partir de una entrada de audio, discernir el tipo de emoción transmitida. Para ello, se debe investigar, experimentar, desarrollar, modelar y finalmente, implementar una infraestructura base sobre la cual presentar el proyecto.

Teniendo en cuenta este marco de trabajo, se especifican los siguientes objetivos alineados con la propuesta original del cliente.

- Dado que uno de los principales ejes de innovación radica en la construcción de una IA para el idioma Español, el primer objetivo será buscar o desarrollar una base de datos sobre la que trabajar. La particularidad del proyecto impide contar con unos datos iniciales.

- Establecer un procesamiento de los datos adecuado al problema a resolver, teniendo en cuenta la importancia de la calidad del dato para el posterior entrenamiento.
- Desarrollar diferentes modelos predictivos, basados en diferentes técnicas de aprendizaje automático, en busca del algoritmo que mejor ventaja competitiva suponga.
- Establecer una arquitectura de modelos para producir el mejor resultado combinando audio y texto.
- Proponer una solución de despliegue de la IA que sea útil y escalable para el usuario final.
- Realizar un análisis de los resultados obtenidos para definir la eficacia de los modelos y sus limitaciones, así como para determinar futuras líneas de investigación.

1.3. Entorno Socio Económico

Aunque todavía es difícil predecir el impacto que los avances en el área de la Inteligencia Artificial tendrán en los centros de atención al cliente, ya podemos empezar a ver cómo muchas empresas han tomado la iniciativa de aplicar técnicas analíticas a los datos que han ido recogiendo a lo largo de los años.

Lo que sí sabemos es que la aplicación de estos nuevos tipos de algoritmos tendrá el mismo objetivo que tenían originalmente los Call Centers: una experiencia mejor y más personalizada a un menor coste.

Según la firma de marketing IDC, en 2020, alrededor del 85 % de todas las operaciones realizadas en estos centros se gestionan mediante IA. Además, estudios de Accenture [8] demuestran que más del 50 % de los directores de tecnología reconocen que los asistentes virtuales conversacionales tendrán un impacto disruptivo en sus industrias.

La irrupción de los asistentes inteligentes, chat bots, modelos predictivos, etc., hace que todas las empresas que tienen u ofrecen servicios de atención al cliente y soporte telefónico se planteen una evolución natural de sus servicios en base a los datos recogidos hasta la fecha. Sin embargo, es importante tener en cuenta que debe haber un proceso en el que se prueben todos estos nuevos modelos y en el que coexistan las técnicas más tradicionales y las más avanzadas.

Desde un punto de vista más general y global, a medida que los modelos de Inteligencia Artificial permitan implementaciones más precisas y eficientes, su impacto en la economía será mayor.

1.4. Marco Regulador

Este apartado es de suma importancia de cara a la futura implementación de la IA. Cualquier proyecto donde se utilice, procese y alamacene información sensible de usuarios, debe estar protegida por un marco regulatorio de protección de datos.

El proyecto debe partir de la premisa de que ninguna información sensible será vista por nadie ajeno al consenso del usuario, así como vendida a un tercero para por ejemplo caracterización de perfiles.

De manera que esto se cumpla, se establece una regulación de procesamiento de datos a nivel nacional en la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales [9] y en el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016 [10].

Dados los recientes avances en el campo del análisis de datos y el de la Inteligencia Artificial, la aplicabilidad de los algoritmos de aprendizaje automático en la industria final se ha vuelto más plausibles ya que los resultados obtenidos son cada vez más acertados. Es por ello, que se han conformado una serie de principios éticos [11] para marcar el camino a esta nueva irrupción de forma responsable. Si bien es cierto, para el desarrollo del proyecto se han partido de datos ajenos a compañías, empleando bases de datos cuya licencia ha sido adquirida para su uso, la cuál se citará más adelante.

1.5. Estructura

En este apartado se hace un resumen sobre cada uno de los capítulos que tiene este documento:

- **Introducción:** En el capítulo que se trata en cuestión, se hace una introducción al problema y su dominio. Además, se presentan los objetivos, el entorno socioeconómico, el marco regulador y la estructura del documento.
- **Estado del Arte:** Este capítulo tiene como objetivo presentar un esquema actual sobre las diferentes técnicas, tecnologías y modelos que se encuentran en el mercado, así como de las vías de investigación abiertas. Partiendo de ese esquema, se presenta la propuesta llevada a cabo durante la realización del trabajo.
- **Base de Datos:** Se comienza realizando una definición de la base de datos a utilizar, y las diferentes propuestas que se han tenido en cuenta. Una vez escogido el conjunto de datos, se realiza una explicación sobre el preprocesado realizado a los datos, así como las transformaciones pertinentes.

- **Modelos de ML:** Después, se explican las métricas utilizadas para comparar a los diferentes modelos generados y la estructura que llevarán los experimentos realizados. Finalmente, se explican de forma esquematizada en cinco fases los pasos que han tenido lugar para la generación y validación de los modelos finales.
- **Despliegue:** En este último apartado se explica la construcción y funcionamiento de la herramienta de uso SentiAudio, un bot conversacional para probar la IA.
- **Conclusiones y Futuros Trabajos:** Por último se trazan las conclusiones globales del proyecto, incluyendo un apartado sobre los posibles trabajos futuros.

2. ESTADO DEL ARTE

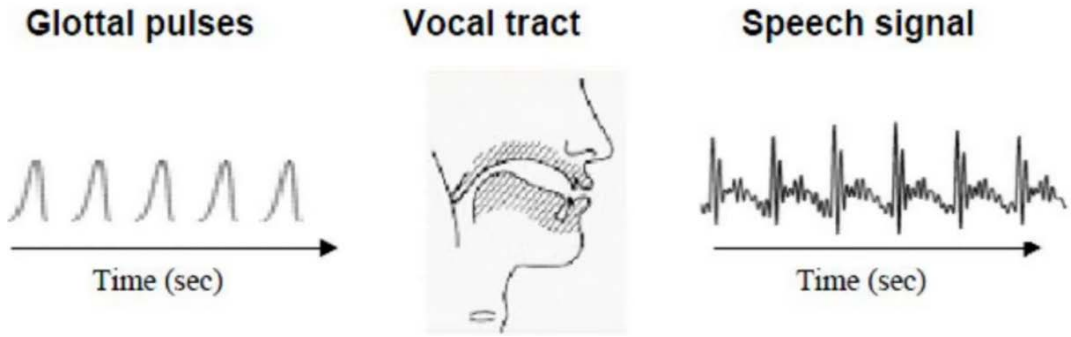
El estado del arte sirve para justificar el trabajo a realizar. Su objetivo es transmitir los conocimientos que se han establecido sobre un tema o problema, indicando los puntos fuertes y débiles, recopilar los trabajos académicos y de industria más relevantes y, en definitiva, organizar todo lo que se ha encontrado estableciendo relaciones entre las soluciones desarrolladas destacando los puntos importantes para el trabajo actual.

De cara a contextualizar el proyecto, el capítulo se ha repartido en dos grupos. Por un lado, el área de Speech Recognition, donde se explican las principales técnicas en audio y cómo funcionan, así como trabajos actuales. Por otro, centrado en el área del Procesamiento del Lenguaje Natural, con un mismo esquema orientado al reconocimiento de sentimiento en texto.
Por último, se aúnan ambas direcciones en la arquitectura del presente trabajo y se explican las tecnologías implementadas.

2.1. Speech Recognition

Previamente a entrar en detalle sobre la ciencia detrás de este campo, es importante comentar los conceptos que rodean la voz humana para comprender los fundamentos.

El habla se puede formalizar cualitativamente como la convolución de la respuesta en frecuencia del tracto vocal con el pulso glótico. En otras palabras, implica que el pulso glótico, producido en las cuerdas vocales, es filtrado en el tracto vocal provocando la señal acústica. Es en el pulso glótico donde se genera el fundamento de la frecuencia sonora o *pitch*, que define las altas frecuencias. Dependiendo de la forma del tracto vocal se generan los fonemas o timbre de la señal.
En la siguiente imagen [12] se pueden discernir visualmente cada concepto.



Fuente: Obtenida del *Journal of Physics Conference Series*

Fig. 2.1. Esquema del proceso del habla.

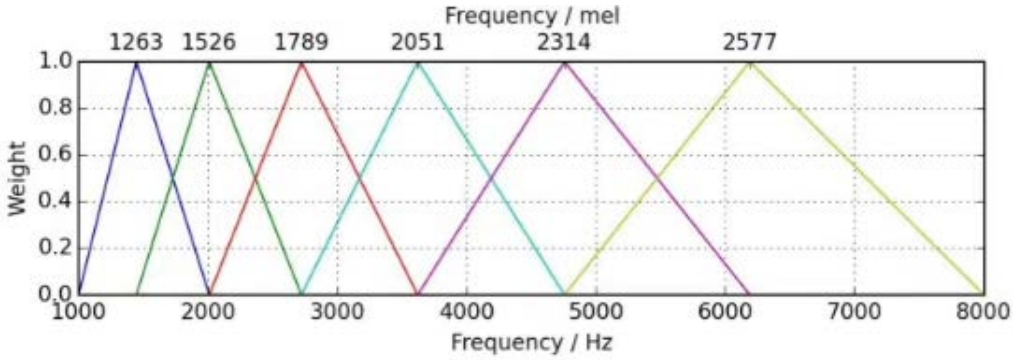
A partir de estos simples conceptos, se han fundamentado las investigaciones relacionadas con el procesamiento del habla humana. En este sentido, han ido surgiendo diferentes técnicas de procesamiento de señales específicas para cada tarea. Entre ellas, se encuentran entre las más simples la varianza o su derivada expresada como la energía del audio, o el cociente abierto (OQ) [13] o su parámetro complementario, el cociente cerrado (CQ), el cúal pertenece a los parámetros clásicos para cuantificar las propiedades de vibración de las cuerdas vocales.
Haciendo uso de la transformada de Fourier [14], que se especifica más adelante, se introdujeron técnicas más complejas, como la energía cromática, conocida como la transformada Q, la cual transforma una serie de datos al dominio de la frecuencia, y que está muy estrechamente relacionada con la transformada de ondícula compleja de Morlet[15]. Otro método de análisis estudia el espectro del sonido, el cual representa la distribución de energía sonora en función de la frecuencia, este concepto es el contraste espectral [16].

2.1.1. Procesamiento Emocional del Habla

En el presente trabajo, se busca el método más adecuado que pueda extraer información que permita reconocer emociones. Este es un procedimiento que convierte la voz de un ser humano en un símbolo emocional, como el enfado, la tristeza o la felicidad. Se han realizado muchos estudios sobre el reconocimiento de emociones. La mayoría de ellos utilizan información prosódica y lingüística para la extracción de atributos [17]. Sin embargo, la precisión del reconocimiento de emociones con estos métodos es baja. En particular, la precisión cae por debajo del 50 % en la mayoría de los sistemas de reconocimiento de emociones independientes del hablante para cuatro (o más) emociones. Esto puede deberse a que los rasgos prosódicos se componen tan solo de la frecuencia fundamental y la energía, lo que significa que sólo hay dos componentes independientes en cada señal. En realidad, hay más componentes independientes en las características fonéticas del habla que en las prosódicas.

Para contrarestar este problema, se desarrollaron técnicas alrededor de la escala de Mel. Es una escala de tonos que el oído humano percibe generalmente como equidistantes entre sí. A medida que aumenta la frecuencia, aumenta el intervalo, en herzios, entre los valores de la escala mel. El espectrograma mel [18] que se muestra a de la imagen 2.3 contiene los distintos bancos de filtros en la escala representados en los triángulos de distinto color.

A partir del estudio de las frecuencias en la escala de mel, es posible obtener los parámetros necesarios que definen una señal acústica para el reconocimiento de emociones. En el proceso convencional de extracción de atributos del habla, cada vector de características se genera utilizando todo el espectro de frecuencias de una trama o *frame* de habla



Fuente: Obtenida del *Polish-Japanese Academy of Information Technology*.

Fig. 2.3. Espectrograma de mel.

determinada. Por lo tanto, cuando la señal del habla está parcialmente degradada por una anomalía localizada en el tiempo en el tiempo y la frecuencia, los vectores de características que se generan en el de la anomalía están completamente contaminados. En tales casos, sin embargo, es probable que las partes no afectadas de las regiones espectrales contengan información útil para la discriminación de la emoción del hablante. Una forma lógica de abordar este problema es dividir las frecuencias en una serie de subregiones y utilizar la información espectral contenida en cada una de estas regiones para generar vectores de características independientes. Esta técnica se conoce comúnmente como análisis de sub-bandas (SBC) [19], la cual fue valorada al inicio de este proyecto.

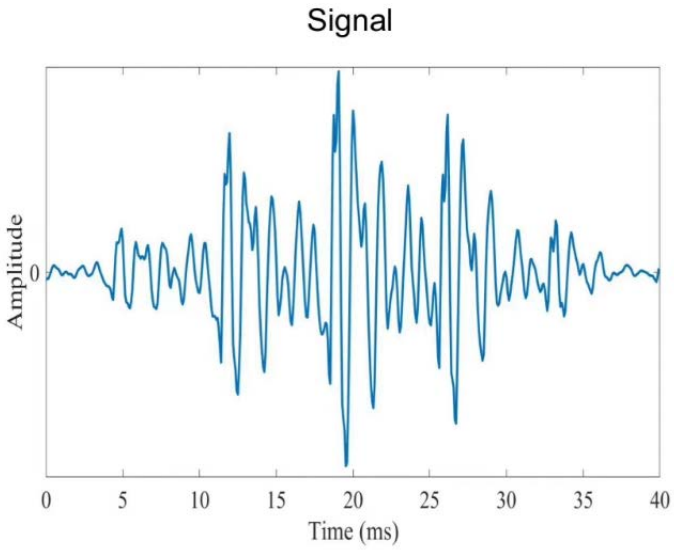
Sin embargo, repasando literatura [20] para el contexto concreto y de cara a la practicidad del proyecto, se observó que SBC supera otras técnicas para la identificación del hablante, pero para el estudio de las emociones, existe un método cuya precisión excede [21] y el cuál, ha sido empleado para este proyecto, que son los coeficientes cepstrales de mel (MFCC).

Coeficientes Cepstrales de Mel

En esta sección se detallan todos los pasos para obtener los atributos de los audios a entrenar.

Partiendo de la señal acústica del habla, se puede representar como una forma de onda donde el eje **X** (abscisas) es el tiempo y el eje **Y** (ordenadas) es la intensidad, medida en decibelios (dB). Esta intensidad o amplitud no es muy informativa, ya que sólo habla del volumen de la grabación de audio. Para entender mejor la señal, es necesario transformarla en el dominio de la frecuencia.

En la siguiente imagen se grafica una señal acústica del habla humana.



Fuente: Obtenida del medio *The Sound of AI [22]*.

Fig. 2.5. Gráfica de una señal de audio.

La representación en el dominio de la frecuencia de una señal indica qué frecuencias diferentes están presentes en la señal. La Transformada de Fourier (DFT) es un concepto matemático que permite convertir una señal continua del dominio del *tiempo* al dominio de la *frecuencia*. Una señal de audio es una señal compleja compuesta por múltiples "ondas sonoras de frecuencia única"que viajan juntas como una perturbación en el medio. Cuando se graba un sonido, sólo se captan las amplitudes resultantes de esas múltiples ondas. La DFT es un concepto matemático que permite descomponer una señal en las frecuencias que la componen, no sólo indica las frecuencias presentes en la señal, sino también la magnitud de cada una de ellas.

Por otro lado, la Transformada de Fourier Inversa (IDFT) es justo lo contrario, toma la representación en el dominio de la frecuencia de una señal dada como entrada y sintetiza matemáticamente la señal original.
La imagen a continuación muestra visualmente lo explicado anteriormente.

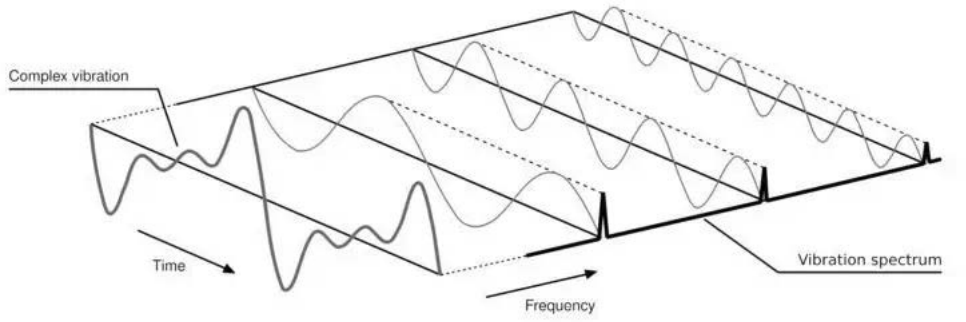
Además de la DFT, se aplica el logaritmo a la señal auditiva. Esto permite tratar de forma independiente ambos factores. Expresando esto matemáticamente queda:

Si se define el habla como:

$$x(t) = e(t) \cdot h(t) \tag{2.1}$$

De donde:

- **e**: pulso glótico.
- **h**: frecuencia del tracto vocal.



Fuente: Obtenida del medio *Towards Data Science [23]*.

Fig. 2.7. Diagrama de la aplicación de la DFT a una señal de audio.

Habiendo aplicado la DFT:

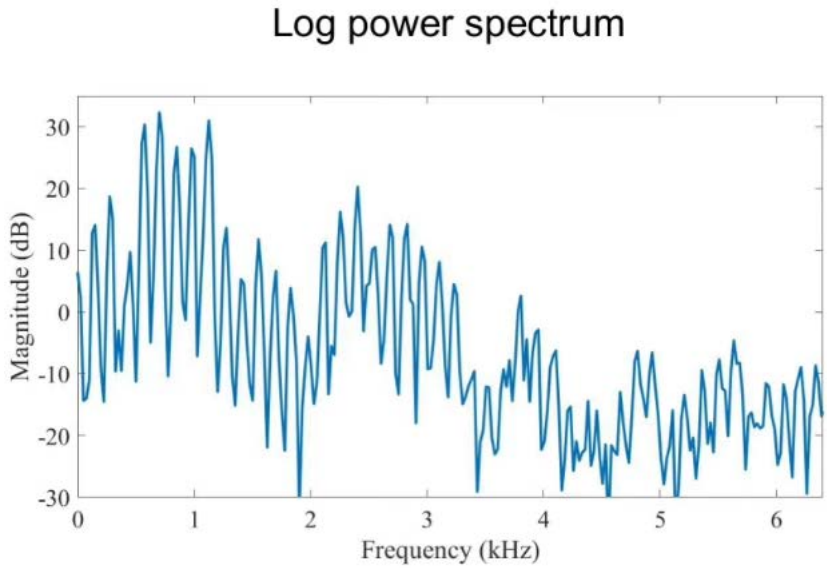
$$X(t) = E(t) \cdot H(t) \tag{2.2}$$

Se calcula el logaritmo:

$$\log(X(t)) = \log(E(t) \cdot H(t)) \tag{2.3}$$

$$\log(X(t)) = \log(E(t)) + \log(H(t)) \tag{2.4}$$

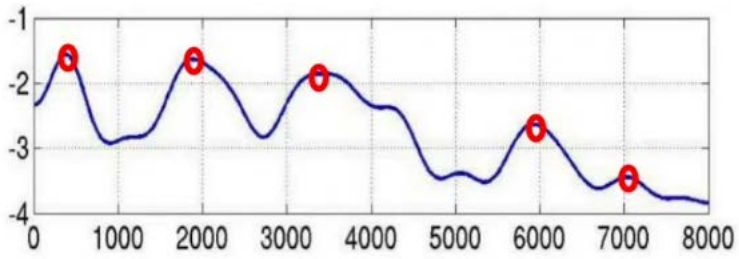
Aplicando esta transformación se obtiene la señal con el eje de abscisas en el dominio de las frecuencias (Hz), y puede tratarse cada factor de manera independiente. De ambos factores, interesa la frecuencia del tracto vocal, que puede definirse como el espectro de potencia logarítmica.



Fuente: Obtenida del medio *The Sound of AI*.

Fig. 2.9. Diagrama del resultado de aplicar la DFT a una señal de audio.

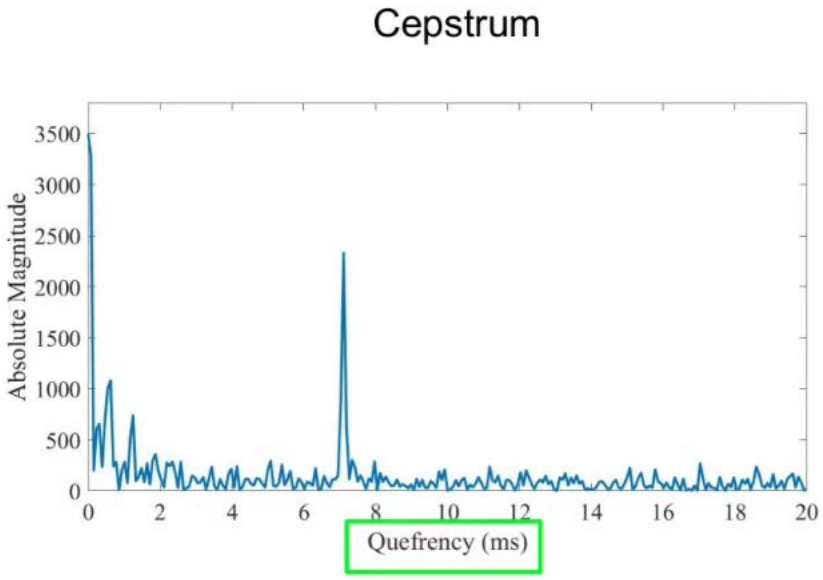
De esta gráfica, aplicando la escala de mel, se pueden obtener los "formantes", los cuales representan los fonemas dentro de la señal acústica y proporcionan identidad al sonido, por lo que son de vital importancia para el procesamiento del habla.



Fuente: Obtenida del medio *The Sound of AI*.

Fig. 2.11. Diagrama de los formantes del espectro de potencia logarítmica.

Para llegar a obtener los cepstrales, una vez obtenido el espectro logarítmico de potencia, se aplica una transformada inversa IDFT. El propósito consiste en transformar del dominio de las frecuencias a su inversa la cual, según la literatura, denominan "quefrency". Al descomponer la señal sobre la inversa del espectro se obtiene el valor de la importancia de las frecuencias.



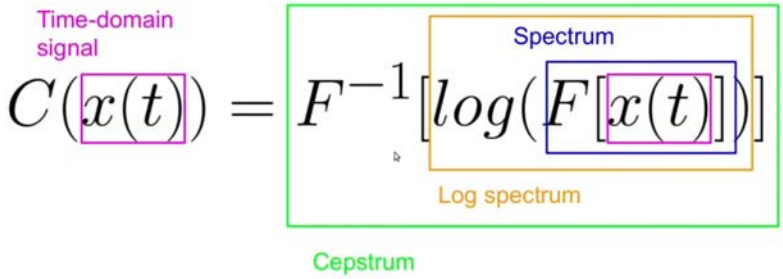
Fuente: Obtenida del medio *The Sound of AI*.

Fig. 2.13. Diagrama de los cepstrales en el eje de la inversa de la frecuencia.

Al aplicar diferentes ordenes de frecuencias en orden ascendente se consigue que en el eje de las abscisas (a menor quefrency), se obtengan los valores más significativos del audio, que contengan los fonemas, timbre e identidad. Gracias a la IDFT, es posible correlacionar la energía dentro de los diferentes bancos o bandas de la escala de mel, lo que permite posteriormente diferenciar las características en atributos linealmente independientes, lo más adecuado para los algoritmos de *Machine Learning*.

Estos valores se agrupan de menor a mayor escala en los coeficientes cepstrales de mel (MFCC).

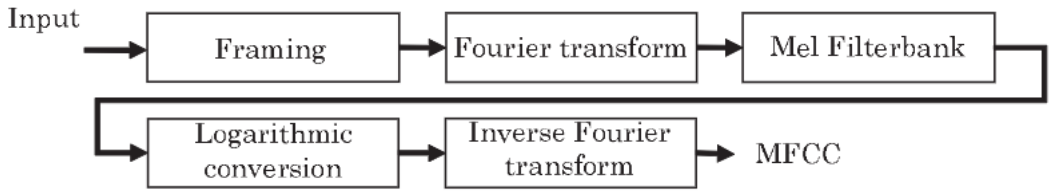
Como se ha mencionado, la importancia significativa de los MFCC va disminuyendo a medida que se obtienen un orden mayor, por lo que la información relativa que aportan al audio de entrada también se ve afectada. Es por ello que en la literatura, para el reconocimiento de emociones, suelen centrarse en un intervalo de entre los 12-25 primeros coeficientes.
De forma que aúne todos los pasos llevados a cabo para el proceso de obtención de los coeficientes :



Fuente:Elaboración propia.

Fig. 2.15. Fórmula de obtención de los cepstrales.

La siguiente imagen muestra el procedimiento a vista de águila de la obtención de los mfcc, extraída de la Revista del Procesamiento Natural [24].



Fuente: Association for Natural Language Processing.

Fig. 2.17. Gráfico de extracción de características de los MFCC.

Es importante destacar que este gráfico muestra el procedimiento por cada *frame*, y en el procesado final se recopilan las medias aritméticas de cada MFFC para el conjunto de *frames* que compone todo el audio.

2.1.2. Trabajos previos

Existen múltiples investigaciones y proyectos en el campo del reconocimiento de emociones haciendo uso de los MFCC. A continuación se detallan algunos relevantes para el contexto del proyecto.

El paper presentado por el departamento de la escuela de ingeniería Amrita en la India, por S. Lalitha et al. [25], destaca en el elevado número de posibles emociones detectadas en alemán empleando EMO-DB, que se explica más adelante. El proyecto mide la eficiencia de los MFCC con los 19 primeros para identificar 7 emociones distintas haciendo uso de redes neuronales con un porcentaje de acierto del 85,7 %.
Un trabajo a señalar realizado por la Universidad Pontificia Javierana de Bogotá por el estudiante Esteban Orozco Castaño [26], ha sido de ayuda para el trabajo ya que se basa en la misma problemática de reconocimiento de emociones en Español para un sistema de un Call Center. En este caso, tanto la base de datos empleada como la arquitectura de modelos no obtiene los resultados esperados, pero ha servido para prevenir las distintas problemáticas que se han encontrado durante el desarrollo del proyecto. La principal, la importancia de la calidad de los datos de entrenamiento.
En aplicaciones ya prácticas, destaca un proyecto dentro del mundo de la salud mental, aplicada a la monitorización de emociones a través de conversaciones interactivas de la mano del proyecto MENHIR [27].
Por último, resulta interesante que el uso de los MFCCs puede ser reorientado a propósitos ajenos al reconocimiento de emociones tal como la identificación del hablante, como el proyecto publicado en el *International Journal on Emerging Technologies* [28].

2.2. Procesamiento del Lenguaje Natural

Para esta segunda sección, pese a buscar el mismo objetivo, se emplean otras tecnologías para el campo del lenguaje natural en texto. El PLN es un subcampo interdisciplinar de la IA que se ocupa de las interacciones entre los ordenadores y los lenguajes humanos naturales a través del habla o el texto. Los programas informáticos que usan PLN nos ayudan en nuestra vida cotidiana de diversas maneras, como por ejemplo los asistentes personales de los dispositivos *smartphone*, el caso de Siri en la marca Apple o Cortana en Microsoft.

El PNL se divide principalmente en dos campos: La lingüística y la informática. Por un lado, la parte Lingüística se centra en la comprensión de la estructura del lenguaje, incluyendo subcampos como la Fonética y la Fonología, que estudian los sonidos y sistemas sonoros del lenguaje humano, o la Morfología y la Semántica, que tienen que ver con el estudio de la formación de la estructura interna tanto de las propias palabras como de las oraciones, así como del contexto semántico de las mismas para objetivos comunicativos concretos.
Por otro lado, la parte informática se ocupa de traducir los conocimientos lingüísticos y la experiencia del dominio en programas informáticos con la ayuda de la IA.

Las aplicaciones de estudio son muy extensas y están en continua evolución. Entre los problemas mayormente resueltos, se encuentran la clasificación de textos (detección

de spam en el correo), la clasificación de la morfología de una palabra (verbo, adjetivo, pronombre, etc.) así como la clasificación de entidades como lugares, personas, etc y también la corrección automática de palabras.

Existen también avances en problemáticas más complejas como los auto completadores de los motores de búsqueda, tráidos de la mano del equipo de Google con *word2vec* [29], que permitió resolver la resolución de correferencia. Dada una frase, determinar qué palabras se refieren a los mismos objetos. También la desambiguación del sentido de las palabras, ya que muchas palabras tienen más de un significado y se debe seleccionar el significado que tenga más sentido en función del contexto.

En 2017, apareció un modelo de ML, nuevamente del equipo de Google, que cambiaría por completo el estudio del lenguaje natural, los Transformers [30]. Sin entrar en detalle, para valorar la disrupción del nuevo modelo se debe mencionar que incorpora dos factores clave, que se denominan *Mecanismo de Atención* y *Multi-cabecera*.

- **Self-attention**: El objetivo de la atención es captar las relaciones contextuales entre las palabras de la frase creando un vector basado en la atención de cada palabra de entrada. Los vectores basados en la atención ayudan a comprender la relevancia de cada palabra de la frase de entrada con respecto a otras palabras de la frase (como consigo misma). Este mecanismo resolvía el problema que sufrían las redes neuronales recurrentes (RNN) con las dependencias de largo alcance, en otras palabras, con documentos de texto largos las RNN no eran capaces de almacenar la información de contexto de la oración.
- **Multi-head**: El procedimiento del cálculo de la atención en cada palabra no se realiza una sola vez, sino que se vuelve a computar tantas veces con respecto al resto de palabras haya. El objetivo es generar varios vectores de atención para la misma palabra. Esto ayuda al modelo a tener diferentes representaciones de las relaciones de las palabras en una frase. Esto implica también que son mucho más rápidos y eficientes, gracias a la capacidad de paralelización. Posteriormente, las diferentes matrices basadas en la atención generadas a partir de las distintas cabezas se concatenan y se pasan por una capa lineal para reducir el tamaño hasta el de una única matriz.

Gracias a esta innovación, un año después se presentó otro trabajo, el cual se considera hasta la fecha como el modelo del estado del arte en el área del PLN, el *Bidirectional Encoder Representations from Transformers*, conocido como BERT [31]. Este es un modelo preentrenado multipropósito ya que puede ser ajustado (*fine-tuning*) para crear modelos de última generación para una amplia gama de tareas, como la respuesta a preguntas y la inferencia lingüística, sin necesidad de realizar modificaciones sustanciales en la arquitectura específica de la tarea.

Gracias a estos últimos avances, ha surgido un área muy reciente del estudio de las emociones en texto, conocido como *Sentiment Analysis*.

2.2.1. Análisis de Sentimiento

El análisis de sentimientos se centra en el estudio de la polaridad de un texto (positivo, negativo, neutro), pero también va más allá de la polaridad para detectar sentimientos y emociones específicas (enfado, felicidad, tristeza, etc.), la urgencia (urgente, no urgente) e incluso las intenciones (interesado frente a no interesado).

Para poder tratar el audio de entrada como texto se han desarrollado una serie de herramientas en el campo del PNL que son capaces tanto de transcribir un audio a texto (STT) como del proceso inverso (TTS).

A continuación se muestran las herramientas escogidas para el trabajo, los modelos que analizan el sentimiento y las herramientas *Speech to Text*.

BETO

Para este proyecto se ha requerido de una adaptación de BERT desarrollada por la Universidad de Chile para el desarrollo de aplicaciones en Español, llamada BETO [32].

Dentro de este modelo, desde la Universidad de Buenos Aires se ha publicado en 2021 un proyecto de herramientas multilingües para el análisis de sentimiento y otras tareas de PLN en código Python, cuya librería Pysentimiento [33] se puede implementar a través de la plataforma *Hugging Face*.

En este paquete de herramientas, el modelo preetrenado para el análisis de sentimiento llamado RoBERTuito [34] ha sido utilizado en este proyecto.

STT

La aportación de las herramientas de translación automática ofrece todo un abanico de posibilidades. Pese a que cada una de las aplicaciones es independiente, la ciencia detrás del software es la misma. El procesamiento del audio comienza con la conversión del fichero en un secuencia de unidades acústicas. A continuación, dichas unidades se emparejan con los fonemas existentes, que son los sonidos que utilizamos en nuestro lenguaje para formar expresiones con sentido. El componente lingüístico se encarga de convertir esta secuencia de unidades acústicas en palabras, frases y párrafos. Esto se consigue al analizar todas las palabras precedentes y su relación para estimar la probabilidad de qué palabra debe usarse a continuación, de esto se encargan modelos como las cadenas de Markov, empleadas a menudo en los programas de reconocimiento del habla.

Para la realización del proyecto, se han valorado distintas herramientas STT para la implementación de la IA, en búsqueda de una solución escalable, eficiente y de calidad.

Se resumen brevemente algunas de las más relevantes.

La herramienta *open source* desarrollada por la startup Coqui.ai [35], fue considerada inicialmente debido a ser de los pocos proyectos de reconocimiento de voz para la comunidad de desarrolladores de forma gratuita, sin que eso suponga una pérdida de calidad en la transcripción. Sin embargo, durante la experimentación, se encontraron problemas recurrentes en la instalación, entre ellos la complejidad de paquetes añadidos para crear el módulo. Al ser una herramienta libre, requiere de la capacidad de cómputo del propio procesador, a diferencia del resto de herramientas las cuáles suelen funcionar a través una api en la cuál mandan el contenido a transcribir.

Por ello se decidió emplear un software de terceros, el cual aparte de los criterios iniciales, debía ser rápido en la respuesta. Haciendo uso de la librería en Python para reconocimiento de audio *Speech Recognition - Pypi*, se pudo conectar con las siguientes herramientas STT : Sphinx, Google, Wit, Azure, IBM y Houndify. Después de un proceso de experimentación con cada una, la escogida fue la herramienta Google STT [36].

2.2.2. Trabajos previos

El análisis del sentimiento en texto puede aportar una perspectiva esencial para entender a los clientes o funcionamiento de una empresa. Desde tendencias de consumidor y análisis de productos para empresas del mundo del retail, como las reseñas en Amazon, o la categorización de temas en redes sociales, como por ejemplo los hashtags en Twitter. También se puede analizar el sentimiento respecto a ciertos activos, como la percepción en el mercado de una determinada acción.

Un ejemplo actual trata sobre el análisis de sentimiento de tendencias sobre iconos populares o temas actuales para mejorar la reputación de la marca y producir un marketing estratégico a las nuevas generaciones, que está realizando la empresa americana KFC. O en el caso de la tecnológica Google, el equipo de desarrollo se encuentra constantemente monitorizando foros para recibir retroalimentación de sus productos y reorientar hacia nuevos desarrollos o retirar características mal percibidas.

2.3. Trabajo Actual

En el presente trabajo, de las dos vías de investigación se ha priorizado el estudio para la parte de audio. Esto implica, una vez escogido la base de datos adecuada, la extracción de los coeficientes cepstrales de mel, la posterior experimentación con los mismos mfccs, el número adecuado, la correlación entre emociones y posteriormente, la predicción de emociones.

2.3.1. Arquitectura final de la IA.

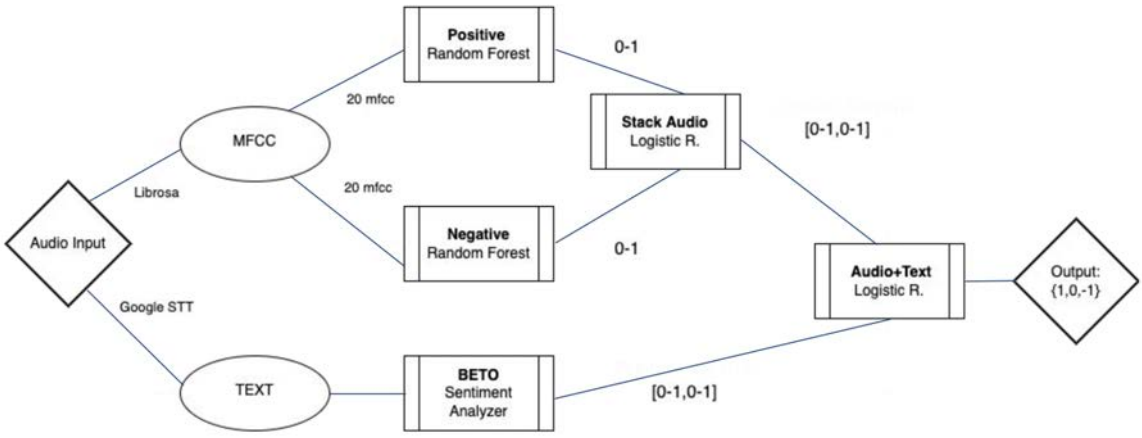
Para tratar los audios, se procuró inicialmente clasificar las emociones en cinco sentimientos diferentes: Alegría, Tristeza, Enfado, Desprecio y Miedo. Para ello, se construyeron varias alternativas para un modelo único que discerniera las emociones. De todas formas, los resultados no fueron los esperados.

Por otro lado, para el propósito final del proyecto, según las conversaciones con el cliente se concretó que bastaban tres resultados de emociones: Positivo, Neutro y Negativo. Esto se debe a que para el sistema del call center IVR, la granularidad de emociones carecía de sentido, y se objetó por simplificarlo en los tres puntos del abanico de emociones, extremos y medio.

Viendo los resultados de un modelo único, se trabajó en dos modelos de clasificación binarios independientes, uno positivo y otro negativo. Después de probar con distintas aproximaciones el algoritmo de ML que mejor resultados dió en ambos fue *Random Forest*, que se detallará en el capítulo de modelos. Para aunar ambos resultados y obtener un valor medio se combinan en un modelo (*stack*), el cual devuelve dos salidas, el mejor resultado se obtuvo con una regresión logística.

Para la parte de texto y apoyar el resultado obtenido, se transcribe el audio original a texto haciendo uso de la herramienta STT y posteriormente se procesa en un modelo de sentimiento (BETO), cuya salida sirve para determinar si la clasificación de audio es correcta. Combinando ambas partes, el mejor resultado se obtuvo también con una regresión logística, que devuelve tres salidas: 1 Positivo, 0 Neutro y -1 Negativo.

La siguiente imagen muestra la arquitectura final del proyecto.



Fuente: Elaboración propia.

Fig. 2.19. Arquitectura global del proyecto acorde a las necesidades del cliente

2.3.2. Tecnologías implementadas y Framework de Investigación

En el aspecto hardware, para que los modelos funcionasen correctamente se ha trabajado con las siguientes especificaciones:

- Procesador: MacBook Pro M1, SO: MacOS Monterey 12.3.1.
- Memoria: 8GB. CPU: 8 núcleos con 4 núcleos de rendimiento y 4 de eficiencia.
- GPU: 8 núcleos.

En el aspecto del aprendizaje automático, se ha trabajado con los siguientes modelos supervisados de Machine Learning, los cuales se explican brevemente y se analizan en el capítulo de modelos.

(I) **Regresión Logística (RLog):** Es un modelo estadístico que se utiliza para determinar la probabilidad de que ocurra un evento. Es similar a la regresión lineal, excepto que en lugar de un resultado gráfico, la variable objetivo es binaria, 1 o 0.

(II) **Árboles de decisión (DT):** Es una estructura de árbol similar a un diagrama de flujo donde los nodos internos representan una característica (o atributo), la rama una regla de decisión y cada nodo hoja el resultado. Para escoger el nodo existe la entropía o ganancia de información, propiedad estadística que mide cuánto un atributo dado separa los ejemplos de entrenamiento de acuerdo con su clasificación objetivo.

(III) **Random Forest (RF):** Es un conjunto (*ensemble*) de árboles de decisión combinados con *bagging*. Al usar *bagging*, cada árbol ve distintas porciones de los datos, ninguno usa todos los datos de entrenamiento. Esto hace que cada uno se entrene con distintas muestras para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

(IV) **Redes neuronales de perceptrón multicapa (FFNN):** El perceptrón es la forma más simple de una red neuronal usada para la clasificación de un tipo especial de patrones, los linealmente separables. En el perceptrón multicapa existen múltiples capas, de tal manera que tiene capacidad para resolver problemas que no son linealmente separables.

(V) **Redes neuronales convolucionales (CNN):** Es una variación del perceptrón multicapa, ya que está diseñada para funcionar con matrices bidimensionales, lo que las hace efectivas para tareas de visión artificial.

(VI) **Máquinas de soporte vectorial (SVM):** El objetivo del algoritmo es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos, lo que implica aquel con el margen más amplio entre las dos clases. Solo se puede encontrar este hiperplano en problemas que permiten separación lineal.

Por último, para el desarrollo software, todo el código ha sido escrito en Python y se han usado las principales librerías:

Librería	**Descripción de Uso**	**Ref.**
Streamlit	Para el desarrollo de los cuadernos de trabajo, es una librería en local que permite tomar apuntes y facilita la investigación de los científicos de datos	[37]
Pandas	Proporciona estructuras de datos rápidas, flexibles y expresivas, diseñadas para que trabajar con datos relacionales o etiquetados sea fácil e intuitivo.	[38]
Numpy	Da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas	[39]
Matplotlib	Permite crear visualizaciones estáticas, animadas e interactivas en Python.	[40]
Seaborn	Sirve para hacer gráficos estadísticos en Python. Está construida sobre matplotlib y estrechamente integrada con las estructuras de datos de pandas.	[41]
Sklearn	Es una biblioteca de aprendizaje automático de código abierto que soporta el aprendizaje supervisado y no supervisado. También proporciona varias herramientas para el ajuste de modelos, el preprocesamiento de datos, la selección de modelos, la evaluación de modelos y demás.	[42]
AugLy	Es una biblioteca de aumentos de datos que actualmente admite cuatro modalidades (audio, imagen, texto y vídeo).	[43]
Pickle	Empleado para almacenar los modelos entrenados de ML, ya que tiene protocolos binarios para serializar y des-serializar una estructura de objetos de Python.	[44]
Librosa	Para la obtención de los MFCCs. Ofrece herramientas para el análisis de música y audio.	[45]
Speech_Recognition	Utilizado para la transcripción del audio a texto.	[46]
Pysentimiento	Una biblioteca basada en Transformer para tareas de PLN, como el análisis de sentimiento.	[47]
Pydub	Para la transcripción de audios y la conversión de codecs.	[48]
Requests	Para la fase final de implementación de la IA en un bot en Telegram, ha sido necesario hacer llamadas HTTP	[49]
Telegram bot	Además de la implementación pura de la API, esta biblioteca cuenta con una serie de clases de alto nivel para que el desarrollo de bots sea fácil y sencillo. Estas clases están contenidas en el submódulo telegram.ext.	[50]

TABLA 2.1. PAQUETES DE PYTHON REQUERIDOS PARA EL PROYECTO.

3. BASE DE DATOS DE ESTUDIO

El capítulo que sigue resume la búsqueda y selección de la base de datos escogida para el proyecto. La decisión de los datos es crucial para el desarrollo de un modelo y la plausibilidad de éxito.
A continuación se descibren los parámetros de búsqueda y el proceso que se ha llevado a cabo hasta conseguir la base de datos definitiva

3.1. Selección de Datos

Para una selección adecuada de los datos, es necesario que cumplan una serie de requisitos comunes, que se resumen en la siguiente tabla:

Atributo	Valor
Calidad de Audio	Media (ruido ambiente)
Variedad lingüística	Media.
Instancias en Español	>20 %
Emociones presentes	>3 (pos y neg)
Instancias totales	>1000

TABLA 3.1. CRITERIOS DE BÚSQUEDA DE LA BBDD DEL PROYECTO

Antes de explicar finalmente los datos escogidos, se presenta a continuación una serie de trabajos valorados y las razones para su descarte.

Artificial - RAVDESS

El trabajo realizado por la universidad de Wisconsin [51] ha sido de especial interés dado que ha sido muy beneficioso para el campo del *Speech Recognition*, ya que consiste en un extenso conjunto de audios en formato libre con voces de actores categorizadas en emociones.
Según lo describen en su web "El *Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)* contiene 7356 archivos (tamaño total: 24,8 GB). La base de datos contiene 24 actores profesionales (12 mujeres y 12 hombres), que vocalizan dos afirmaciones de léxico similar con acento neutro norteamericano. El discurso incluye expresiones de calma, alegría, tristeza, enfado, miedo, sorpresa y asco, y la canción contiene emociones de calma, alegría, tristeza, enfado y miedo. Cada expresión se produce en dos niveles de intensidad emocional (normal, fuerte), con una expresión neutral adicional. Todas las condiciones están disponibles en tres formatos de modalidad: Solo audio (16bit,

48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), y Solo Video (sin sonido)."

El motivo por el cual no se ha empleado este conjunto de datos radica en que incumple las condiciones de variedad lingüística y de número de instancias en Español.

Artificial - EMO-DB

Siguiendo la misma línea de trabajo, se encuentra el proyecto de Emo-DB *Berlin Database of Emotional Speech* [52], un trabajo de código abierto realizado en la cámara anecoica del departamento de acústica técnica de la Universidad Técnica de Berlín, como parte de un proyecto de investigación, del cual luego se extrajeron importantes trabajos.

Al igual que en RAVDESS, se descartó debido a que incumple las condiciones de variedad lingüística y de número de instancias en Español. De todas formas, es un trabajo muy interesante ya que trabajaron el sonido ambiente.

Artificial - SAVEE

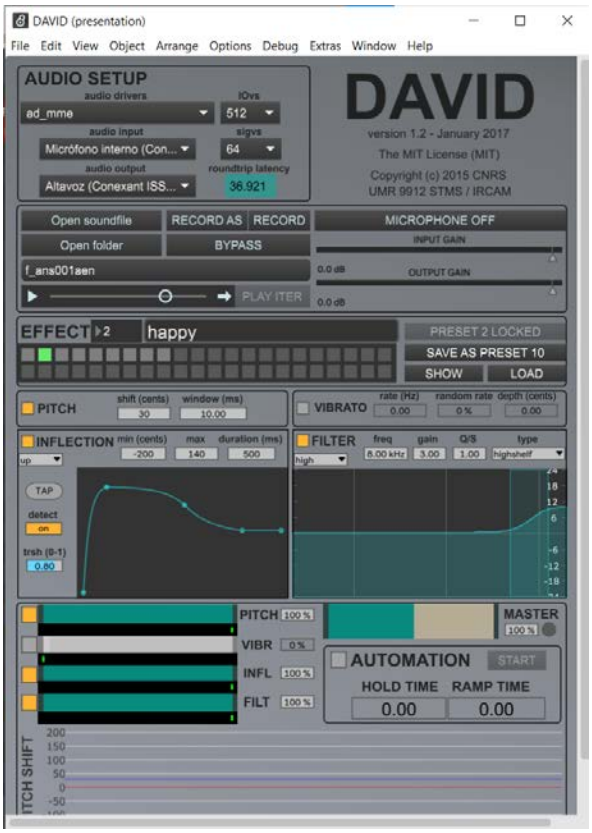
La base de datos *Surrey Audio-Visual Expressed Emotion (SAVEE)* [53] ha sido grabada como requisito previo para el desarrollo de un sistema de reconocimiento automático de emociones. La base de datos consta de grabaciones de 4 actores masculinos en 7 emociones diferentes, 480 enunciados en inglés británico en total.

No cumple los requisitos de variedad lingüística, de número de instancias en Español y de número total de instancias.

Sintética - DAVID

Dado que uno de los pilares del proyecto radica en el entrenamiento dirigido a la caracterización de emociones en Español, se valoró la propuesta realizada por el Instituto de Investigación y Coordinación en Acústica y Música (IRCAM) y la Universidad de la Sorbona, en París.
El trabajo consiste en un software de código abierto que transforma las emociones expresadas por las señales de voz utilizando efectos de audio como el cambio de tono, la inflexión, el vibrato y el filtrado, llamado DAVID [54].

Sin embargo, los resultados no eran los esperados, y pese a ser una apuesta prometedora, el trabajo que requiere montar una base de datos no era viable.



Fuente: Elaboración propia

Fig. 3.1. Captura de muestra programa DAVID

General - SER DATASETS

Es importante destacar el repositorio compuesto por Ayoub Malek (Machine Learning Engineer). El *Spoken Emotion Recognition Datasets (SER)* [55] es una colección de conjuntos de datos para el reconocimiento/detección de emociones en el habla. Incluye enlaces a proyectos ordenados cronológicamente con una descripción del contenido de cada conjunto de datos junto con las emociones incluidas, así como los idiomas hablados.

De aquí, después de todas las propuestas previas, se escogió para el trabajo la siguiente base de datos.

3.1.1. EMOFILM

La base de datos escogida finalmente es EMOFILM [56]. EmoFilm es un corpus de habla emocional multilingüe que comprende instancias de audio producidas en inglés, italiano y español. Los clips de audio se extrajeron de 43 películas (originales en inglés y sus versiones dobladas en italiano y español). Se consideraron géneros como la comedia, el drama, el terror y el thriller; se tuvieron en cuenta los estados emocionales de enfado, desprecio, felicidad, miedo y tristeza.

Cuenta con las siguientes características:

PROPIEDADES EMOFILM	
Instancias de Audio	1.115
Variedad de Idiomas	Español, Inglés e Italiano
Presencia de Español	34 %
Estados de Emoción	5
Voces	M y F
Media de Duración	3.5 sec
Desviación Estándar	1.2 sec
Frecuencia de muestreo	48kHz de 16-bit

TABLA 3.2. RESUMEN DE LAS PRINCIPALES PROPIEDADES DE EMOFILM.

3.2. Procesado de los Datos

En esta sección se tratan todos los procesos que conciernen a la exploración, limpiado y transformación de los datos para el posterior entrenamiento.
Sin este trabajo previo no es posible obtener resultados precisos y reales. El modelo aprende de los datos, un *input* defectuoso produce un modelo *output* erróneo.
Además, gracias a una buena arquitectura de los datos, es posible plantear las preguntas adecuadas a resolver, así como encontrar posibles fallas o futuros problemas / casuísticas no previstas.

Planteamiento del nuevo modelo.
Como se ha explicado previamente, se ha escogido la base de datos EMOFILM para el entrenamiento de los modelos. En la tabla 3.2 se pueden observar las principales características.
A continuación, se procede a la fase inicial exploratoria de los datos.

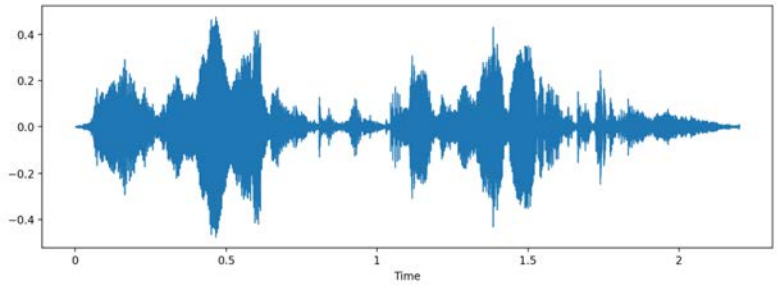
3.2.1. Primera fase: Análisis preliminar

Al enfrentarnos a una nueva base de datos, se debe primero diferenciar entre atributos y variable a predecir. En una primera instancia, se cuenta con unos audios categorizados en distintas emociones. Siendo esta la variable a predecir, es necesario extraer unos atributos que definan las características auditivas suficientes para cada audio. Una vez extraídos los atributos del audio, se estudia la interdependencia lineal entre ellos con respecto a la variable a predecir.

Pese a ser el enfoque más adecuado, siempre es recomendable valorar si el problema a resolver es agnóstico de los atributos que se van a introducir, haciendo una aproximación de aprendizaje no supervisado mediante el método de Análisis de Componentes Principales (PCA).
Partiendo de esta premisa, se realizan las siguientes preguntas:

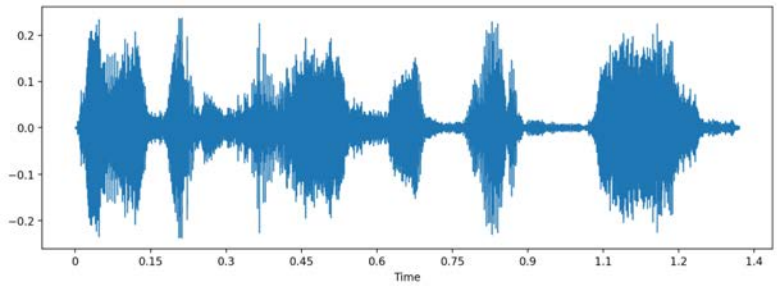
¿Existe una correlación visual entre los histogramas y las emociones?

Por ejemplo, entre el intervalo de las frecuencias. Algunos ejemplos como los siguientes demuestran que no es relevante.



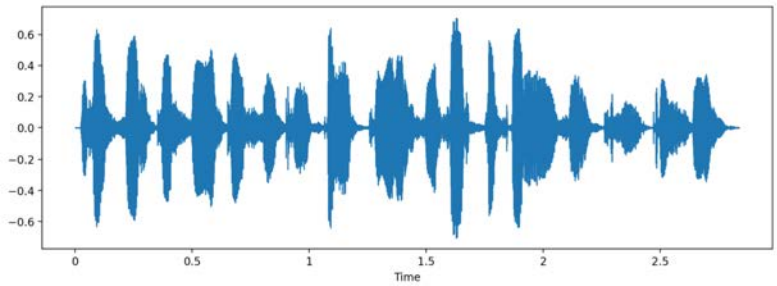
Fuente: Elaboración propia

Fig. 3.3. Histograma muestra f_ans001aes E:Miedo.



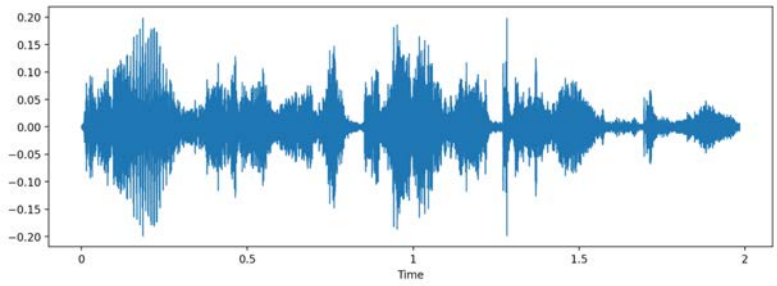
Fuente: Elaboración propia

Fig. 3.5. Histograma muestra f_rab006aes E:Enfado.



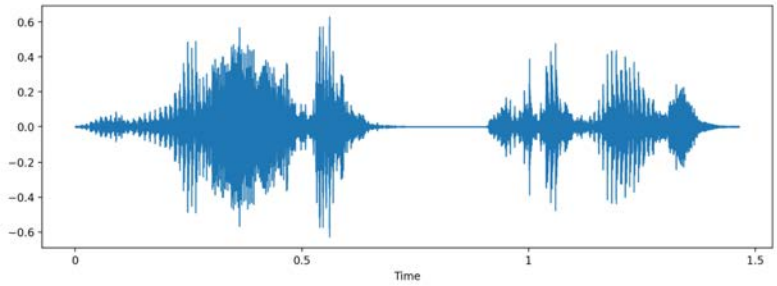
Fuente: Elaboración propia

Fig. 3.7. Histograma muestra f_dis008aes E:Desprecio.



Fuente: Elaboración propia

Fig. 3.9. Histograma muestra f_tri028aes E:Tristeza.

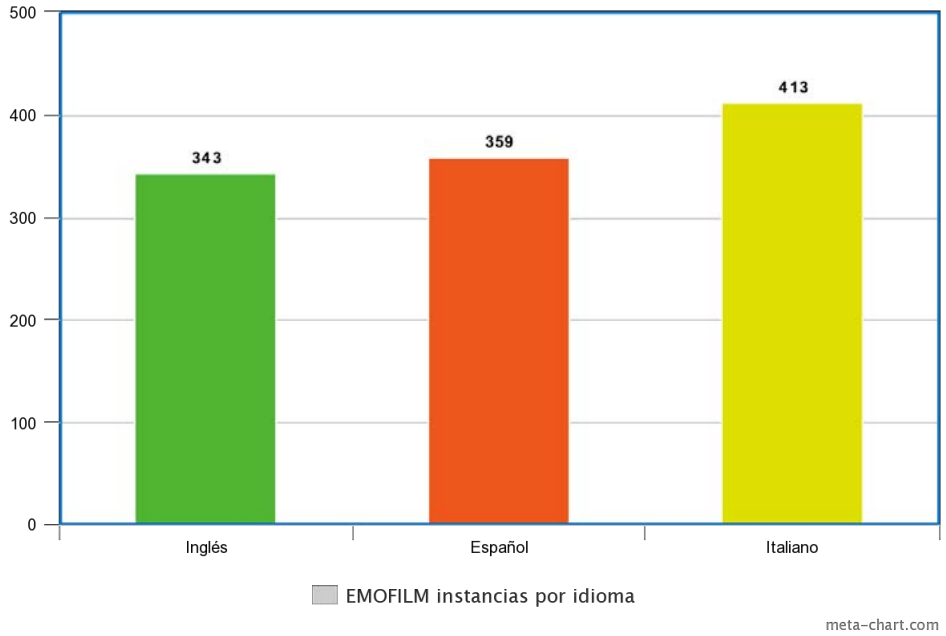


Fuente: Elaboración propia

Fig. 3.11. Histograma muestra f_gio002aes E:Felicidad.

¿Están desbalanceadas las clases de idiomas?

No se aprecia un desbalanceo significativo que pueda afectar el comportamiento del modelo.

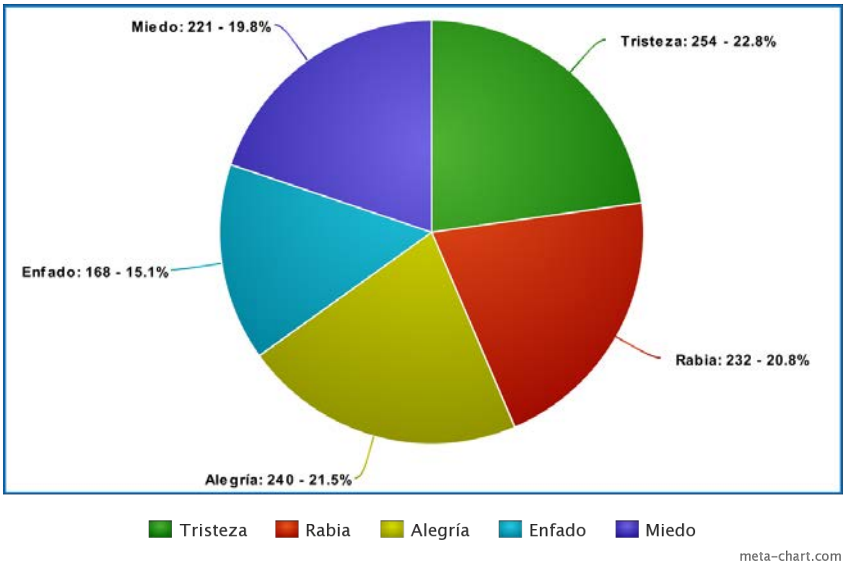


Fuente: Elaboración propia

Fig. 3.13. Gráfico circular de balanceo de todas las emociones.

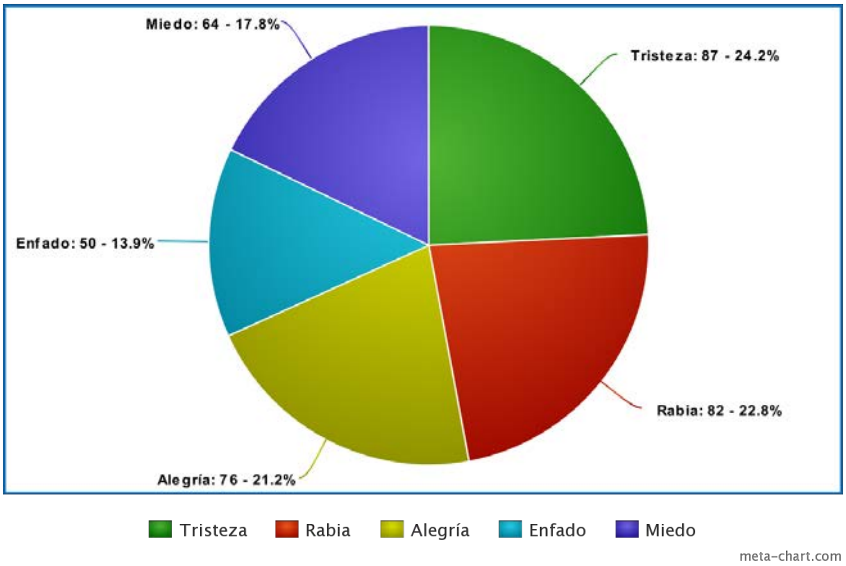
¿Están desbalanceadas las clases de emociones?

No se aprecia un desbalanceo significativo ni genéricamente ni por idioma, que pueda afectar el comportamiento del modelo.



Fuente: Elaboración propia

Fig. 3.15. Gráfico circular de balanceo de todas las emociones.

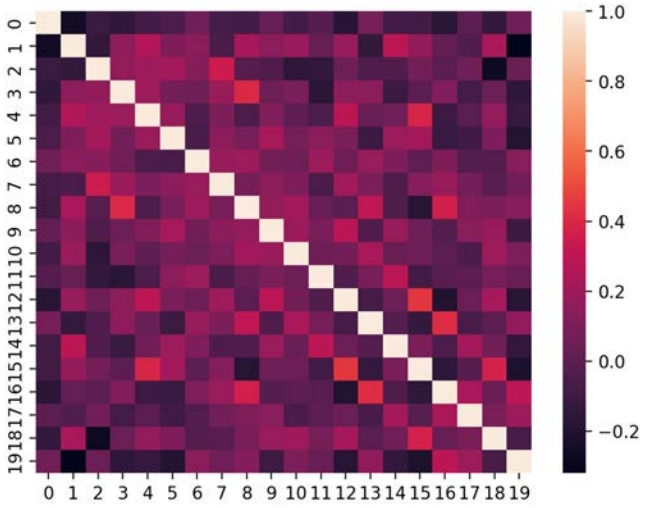


Fuente: Elaboración propia

Fig. 3.17. Gráfico circular de balanceo de las emociones en idioma Español.

En el conjunto general de datos, ¿existe una correlación visual entre los coeficientes cepstrales y las emociones?

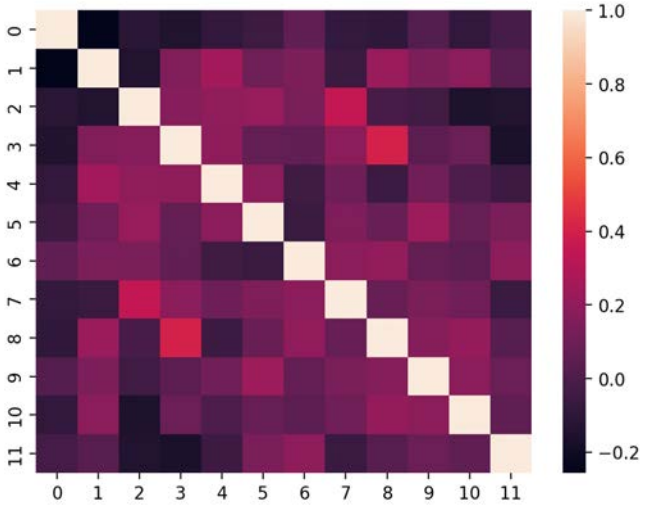
No se aprecia una correlación significativa que pueda afectar el comportamiento del modelo.



Fuente: Elaboración propia

Fig. 3.19. Mapa de calor de los mfcc en todas las instancias.

En el contexto de una emoción en particular, véase alegría por encontrarse en el extremo del espectro emocional, ¿existe una correlación visual entre los coeficientes cepstrales y la emoción?
No se aprecia una correlación significativa que pueda afectar el comportamiento del modelo.



Fuente: Elaboración propia

Fig. 3.21. Mapa de calor de los mfcc en todas las instancias de Alegría.

3.2.2. Segunda fase: Preprocesado

Habiendo realizado un primer análisis de los datos generales, se procede a la limpieza y transformación de los datos “per se".
Para la obtención de los coeficientes cepstrales, se hace uso de la herramienta que viene

implementada en librosa [57]. Como se explica en el capítulo del estado del arte, se realizó un estudio para el número de coeficientes. Según la literatura, oscila entre los primeros 12-25 MFCCs.

A continuación una muestra de la conversión a mffcs de los audios:

	mffc1	mffc2	mffc3	mffc4	mffc5	mffc6	mffc7	mffc8	mffc9	mffc10	mffc11	mffc12	Speaker_	Language	Emotion
0	-294.3086	90.9930	4.0629	2.6460	-17.1554	-6.9619	-17.7467	9.7623	-18.0579	1.7301	-6.8276	-8.9685	mujer	ingles	miedo
1	-239.0695	68.6995	23.1004	4.6364	-10.3547	8.2535	-1.3599	-22.7011	-30.2320	11.4800	-11.2950	-5.5395	mujer	espanol	miedo
2	-278.9554	79.3875	1.0997	12.7229	-17.5629	-0.2494	-11.5228	-24.7372	-36.7952	-11.1705	-13.3867	-17.9567	mujer	italiano	miedo
3	-223.6543	41.4487	14.4913	14.5926	-10.5763	-8.1247	-22.2398	-18.6500	-29.7213	-0.7236	-14.4400	1.4475	mujer	espanol	miedo
4	-271.7888	63.9482	4.6553	18.7108	-2.2028	-5.0951	-25.7172	-25.0449	-35.9950	-16.8657	-12.0768	-13.4977	mujer	italiano	miedo

Fuente: Elaboración propia

Fig. 3.23. Muestra de conjunto de datos con los MFCCs sin normalizar.

Como se puede observar en la imagen, se han sacado los primeros doce. Además, gracias al nombre del fichero de audio, se pueden obtener las siguientes variables extra:

1. IDIOMAS (dos últimas letras del nombre del archivo): Italiano - it Español - es Inglés - es
2. EMOCIONES (tres letras después de _): Miedo - ans Desprecio - dis Alegría - gio Enfado - rab Tristeza - tri
3. GÉNERO del hablante (primera letra): Mujer - f Hombre - m

Sin embargo, estas variables adicionales se descartaron después de conversaciones con el cliente por motivos de practicidad final. De los MFCCC, se confirma, como se explica en la literatura, que van decreciendo en intervalo de variación de mayor a menor a medida que avanzan los coeficientes.

Esto se puede observar en la descripción del contjunto de datos que se muestra:

	mffc1	mffc2	mffc3	mffc4	mffc5	mffc6	mffc7	mffc8	mffc9	mffc10	mffc11	mffc12
count	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000	1,115.0000
mean	-262.5660	83.1261	-23.7573	13.6108	-8.8969	-11.4435	-16.8508	-10.4943	-19.4053	-8.5012	-14.2463	-3.5276
std	59.2558	35.5894	28.9019	19.5456	17.3498	14.8846	12.7203	11.6917	12.3931	9.7379	8.6963	7.6018
min	-465.3603	-34.4424	-162.7382	-51.6190	-62.7231	-60.4714	-59.2015	-48.7161	-71.2896	-42.5520	-48.0153	-28.4767
25%	-299.4079	61.3981	-41.1382	1.1974	-20.2975	-21.0315	-24.8928	-18.5471	-27.3498	-15.2621	-19.9971	-8.7385
50%	-257.5787	82.8452	-20.6475	13.4515	-8.5788	-10.7136	-16.7176	-10.3773	-18.8483	-8.0659	-14.3733	-3.5578
75%	-220.5493	103.9828	-4.5096	25.8307	2.9314	-1.8532	-8.7495	-2.8356	-11.4234	-1.9752	-8.6934	1.6499
max	-120.4143	211.9293	71.9298	85.0573	40.9359	43.2757	32.6220	34.9032	16.6629	23.2257	17.0484	29.0258

Fuente: Elaboración propia

Fig. 3.25. Descripción de la base de datos sin normalizar.

Esto puede llegar a ser problemático para el entrenamiento de los modelos de ML. Por tanto, para los atributos númericos se aplica una técnica de normalización de forma que las variables tengan un peso similar a la hora de contabilizarlas en el entrenamiento.

Pese a que los MFCCs van perdiendo importancia relativa, la información de las emociones tiene matices que no llegarían a obtenerse de no escalar los datos. Como se puede observar por ejemplo en los valores de la desviación estándar (std), que los saltos en escala de importancia a medida que avanzan los coeficientes tienen una importancia significativa.
Principalmente para los algoritmos de redes neuronales, la importancia de unos datos escalados es clave para el funcionamiento del cálculo de la activación de las neuronas en todas las capas, si ocurriera que unos datos estuvieran con valores muy superiores, por ejemplo un atributo que oscilara entre 100 y -100 frente a otro de igual importancia entre 1 y -1, llegaría a ocurrir que en los cálculos internos del traspaso de pesos la segunda variable pasaría a ser irrelevante.

Es por ello, que se ha procedido a aplicar una normalización pertinente a todos los atributos, siguiendo la función MinMax:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

De donde:

- **X**: Atributo a calcular.

El resultado es el siguiente:

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.7297	0.4550	0.4716	0.5351	0.5490	0.3752	0.8821	0.8676	0.6768	0.3956	0.3352	0.4549
1	0.8911	0.2281	0.5023	0.2853	0.4492	0.4487	0.5470	0.3716	0.5388	0.5509	0.2459	0.1720
2	0.6153	0.5069	0.4389	0.5084	0.3701	0.4727	0.5844	0.7029	0.6373	0.5707	0.7971	0.3999
3	0.5359	0.5959	0.4785	0.5375	0.4473	0.3834	0.3944	0.3479	0.8924	0.6824	0.7303	0.1495
4	0.5023	0.4785	0.5977	0.3217	0.6507	0.5575	0.2597	0.7056	0.3003	0.0625	0.4665	0.5396

Fuente: Elaboración propia

Fig. 3.27. Muestra de conjunto de datos con los MFCCs normalizados .

3.2.3. Data Augmentation

Esta sección ha sido añadida en el capítulo de datos ya que corresponde a una transformación de la base de datos de entrenamiento, pese a ser un paso posterior a la primera aproximación (Clasificador General) del próximo capítulo de modelos de ML.

Como se explica en el estado del arte, después del intento de aunar la predicción de emociones en un mismo modelo, se viró hacia una estrategia diferenciativa hacia ambos extremos de emociones, dado que los resultados no estaban siendo los esperados.

Para ello, se optó por modelar clasificadores binarios para dos emociones, positiva y negativa. El problema que surgió era lo que se conoce como sobremuestreo, es decir, el número de ejemplos de entrenamiento para cada emoción se encontraba desbalanceado. Sólo en problemas de clasificación en dónde la variable a predecir es una anomalía, como por ejemplo en clasificaciones de impago hipotecario, es interesante trabajar con datos desbalanceados, ya que se quiere enseñar al modelo a discernir los casos específicos de la mayoría. En este caso, es necesario nivelar el número de ejemplos para ambos modelos.

Con este objetivo, se hace uso de la herramienta desarrollada por el equipo de investigación en Meta AI, llamada AugLy [58].
Según la literatura, los atributos acústicos que diferencian las emociones en el habla [59] son los siguientes: Partiendo de esta premisa, se aplican una serie de transformaciones a

Table V. Acoustic Profiles of Fear, Anger, Sadness and Joy

	Fear	Anger	Sadness	Joy
Volume (PKAMP)	High	High	Low	High
Volume variance (VPKAMP)	Low	High	High	Moderate
Pitch (PKFREQ)	High	Low	Low	Low
Pitch variance (VPKFREQ)	High	High	Low	High
Rate of utterance (TLKTOT)	Fast	Fast	Slow	Moderate
Duration of speech (SPTOT)	Short	Short	Long	Long
Duration of pauses (GPTOT)	Short	Short	Long	Long
Number of pauses (GPN)	Few	Few	Many	Some

Fuente: Revista de Investigación Psicolingüística.

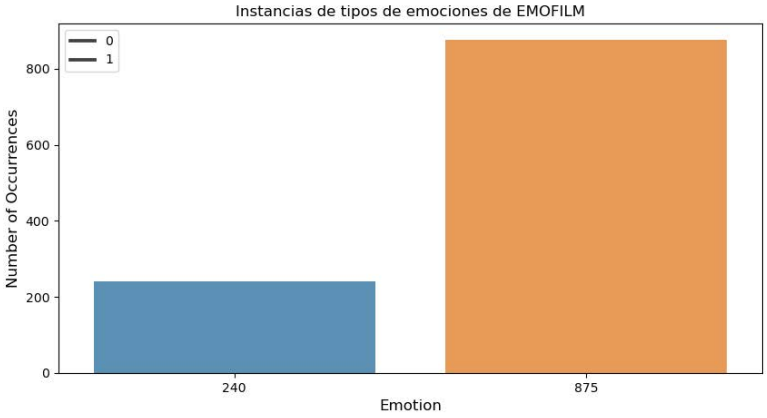
Fig. 3.29. Perfiles acústicos de las emociones.

los audios, de manera que el aumento de nuevos datos partiendo de los originales no contamine las muestras. Tanto para el aumento de los ejemplos positivos, que se extraen de la emoción Alegría, como de los negativos, extraídas de las emociones Enfado y Desprecio, se aplican distintas transformaciones en base a los parámetros que corresponden.

- **Pitch Shift:** Desplaza el tono del audio en n_pasos, donde cada paso es igual a un semitono.
- **Time Stretch:** Estira el audio en tiempo en una proporción fija. Si *rate* >1 el audio se acelerará en ese factor; si *rate* <1 el audio se ralentizará en ese factor.
- **Low Filter:** Permite el paso de señales de audio con una frecuencia inferior a la de un corte dado y atenúa las señales con frecuencias superiores a la del corte.

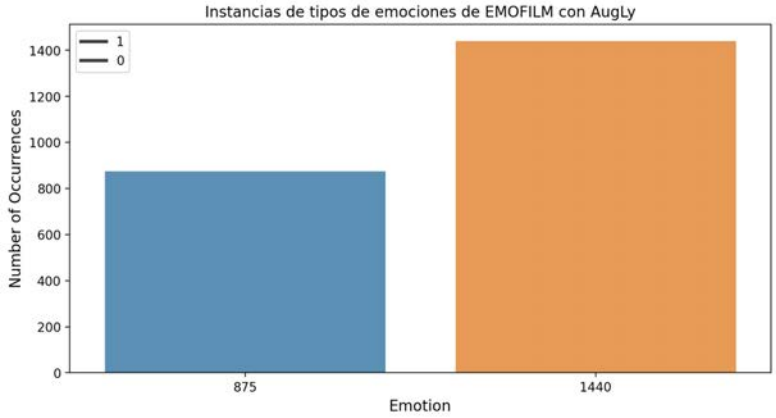
- **High Filter:** Al contrario del anterior, permite el paso de señales de audio con una frecuencia superior a la de un corte dado y atenúa las señales con frecuencias inferiores a la del corte.
- **Background Noise:** Añade sonido de fondo, desde ruido blanco, clicks intermitentes o sonido de calle.

Aplicando las transformaciones a los audios, para el caso de los ejemplos positivos (1) frente al resto (0), vemos lo siguiente.



Fuente: Elaboración propia.

Fig. 3.31. Instancias positivas (azul) infra-muestreadas antes de aplicar *Data Aug*.



Fuente: Elaboración propia.

Fig. 3.33. Instancias positivas (naranja) sobre-muestreadas después de aplicar *Data Aug*.

Posteriormente, para aleatorizar más la muestra, se aplica la función *RandomUnderSampler* de ImbLearn para igualar el número de instancias 0-1 al 50 % [60].

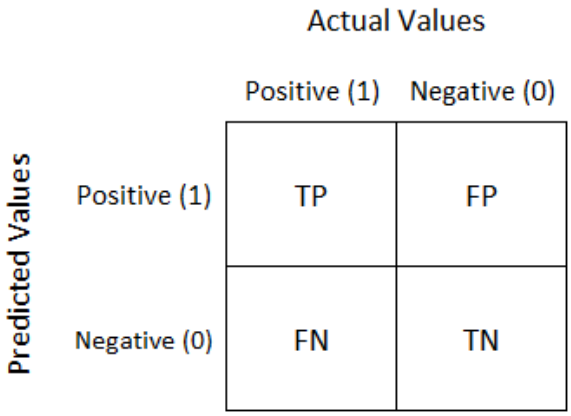
4. DESARROLLO DE MODELOS DE MACHINE LEARNING

En este capítulo, se redactan los distintos pasos necesarios que han permitido conformar la arquitectura final del proyecto.

No todas las fases que se resumen a continuación son prácticamente necesarias, pero al tratarse de un proyecto de investigación se deben mencionar todas las vías y sus errores.

4.1. Métricas

Antes de entrar en detalle, es importante declarar cuáles son las métricas sobre las que se han evaluado los distintos modelos.
En este sentido, es de mucha utilidad la matriz de confusión, ya que permite representar en una gráfica los valores sobre los que se basan varias métricas.



Fuente: Revista *Towards Data Science*

Fig. 4.1. Representación de la matriz de confusión.

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

De aquí se extraen las siguientes métricas:

- **Precision (Precisión)**: Permite medir la calidad del modelo.

$$\frac{TP}{TP + FP} \tag{4.1}$$

- **Recall (Exhaustividad)**: Permite medir la cantidad que el modelo es capaz de identificar.

$$\frac{TP}{TP + FN} \tag{4.2}$$

- **F1-score (F1-valor)**: Permite comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

$$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.3}$$

- **Accuracy (Exactitud)**: Permite medir el porcentaje de casos que el modelo ha acertado.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

4.2. Experimentación

Durante la experimentación se han tratado varios algoritmos de clasificación. Por un lado, los modelos más tradicionales de regresión logística (RLog), máquinas de soporte vectorial (SVM) y *random forest* (RF). Para el tratamiento de estos modelos, en concreto SVM y RF, se ha trabajado con una matriz de experimentación que se compone de los siguientes hiperparámetros:

Hiperparámetro	Valores
Número de estimadores	Fibonacci: [21, 23, 34, 55, 89, 144]
Mínimo de hojas para cálculo de nodo	Fibonacci: [1, 2, 3, 5, 8, 13]
Criterios de partición	Gini / Entropia
Máxima profundidad	12
Máximo valor de atributos a considerar en partición	log2(N)

TABLA 4.1. VALORES DE LOS HIPERPARÁMETROS DEL RF.

Hiperparámetro	Valores
Tipo de núcleo (kernel)	lineal / rbf / sigmoide
Parámetro de regularización (C)	[1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100]
Coeficiente del núcleo (gamma)	['scale','auto']

TABLA 4.2. VALORES DE LOS HIPERPARÁMETROS DEL SVM.

Por otro lado, se emplean los algoritmos de redes neuronales. Para este tipo de problemas, se usan *feed-forward* (FFNN) y convolucionales (CNN), cuya arquitectura ha ido variando para cada caso. Es importante mencionar que la variable a predecir de emociones, al ser categórica, debía ser transformada para estos modelos.

Para la CNN, basta con una codificación categórica. Sin embargo, para el perceptrón multicapa (FFNN), se aplica un codificador categórico siguiendo la técnica *one-hot*. Esto es porque algunos datos de entrada no tienen ninguna clasificación para los valores de las categorías, lo que puede dar lugar a problemas con las predicciones y a un rendimiento deficiente. Es entonces cuando una codificación *one-hot* ayuda. Hace que los datos de entrenamiento sean más útiles y expresivos, y se pueden reescalar fácilmente. Al utilizar valores numéricos, es más fácil determinar una probabilidad para cada uno de los valores. En particular, se utiliza una codificación *one-hot* para los valores de salida, ya que proporciona predicciones más matizadas que las etiquetas simples.

La salida antes

	Emotion
0	2
1	4
2	3
3	0
4	0
5	3
6	1
7	3
8	4
9	3

La salida transformada

	0	1	2	3	4
0	0	0	1	0	0
1	0	0	0	0	1
2	0	0	0	1	0
3	1	0	0	0	0
4	1	0	0	0	0
5	0	0	0	1	0
6	0	1	0	0	0
7	0	0	0	1	0
8	0	0	0	0	1
9	0	0	0	1	0

Fuente: Elaboración propia.

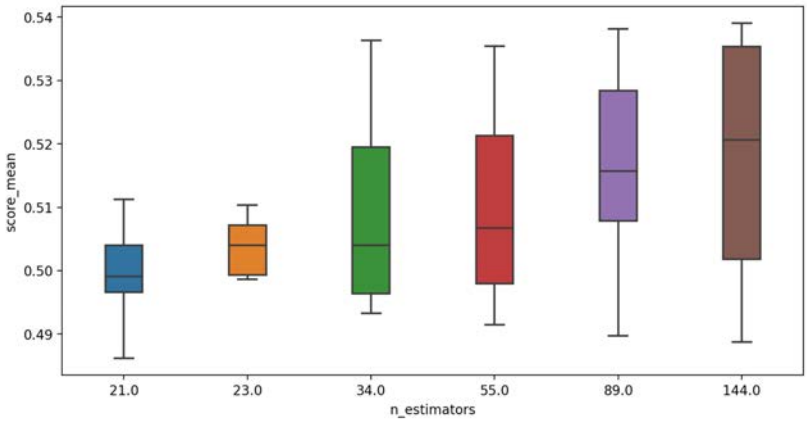
Fig. 4.3. Categorización de las emociones *one-hot*.

Por último, se ha llevado a cabo por regla general una partición 80-20, es decir, el 80 % de los datos para el conjunto de entrenamiento y validación, y el 20 % restante para el conjunto de test.
Además, para el mejor entrenamiento de los modelos se han empleado pliegues (*folds*). Un pliegue es un conjunto de registros (normalmente consecutivos) del conjunto de datos. Los *folds* se usan para generalizar el entrenamiento de los modelos. Durante la experimentación, se trabajan con 10 pliegues diferentes, de forma que cada modelo entrena en diez

ocasiones distintas, esto se conoce como validación cruzada. En concreto, para la generación de los pliegues se ha usado el método de pliegues estratificados. *Stratified K-fold* es una variante mejorada de la validación cruzada, que cuando hace los pliegues del conjunto de entrenamiento tiene en cuenta mantener equilibradas las clases. Por tanto, al general los pliegues de los clasificadores, se tiene en cuenta los porcentajes de representación de cada una de las emociones (clases) dentro de cada subconjunto de datos.

4.2.1. Apr. 1.0: Clasificador General

La primera aproximación consistió, como se ha mencionado ya previamente, en un clasificador genérico de las emociones el cuál, para darle la mayor brevedad posible, no dió los resultados esperados. Para los clasificadores tradicionales el mejor caso fue RF con una precisión del 62 %.



Fuente: Elaboración propia.

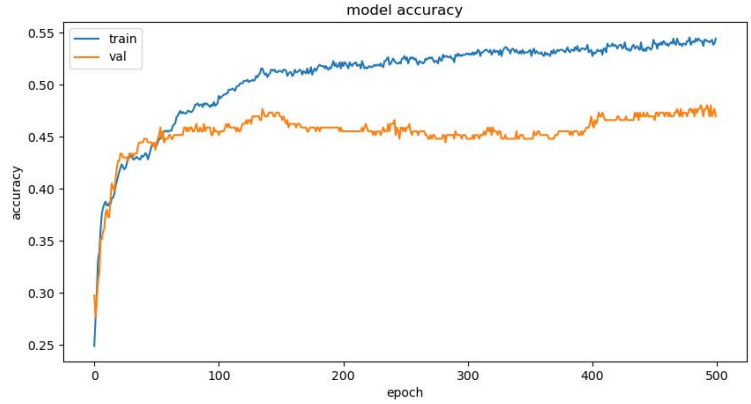
Fig. 4.5. Variación de precisión en número de estimadores para RF.

Mientras que para las redes neuronales, el perceptrón multicapa no logró superar la barrera del 50 % en precisión, por lo que se optó por cambiar la estrategia.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 12)                156

 dense_1 (Dense)             (None, 5)                 65

 dense_2 (Dense)             (None, 5)                 30

 dense_3 (Dense)             (None, 1)                 6

=================================================================
Total params: 257
Trainable params: 257
Non-trainable params: 0
```

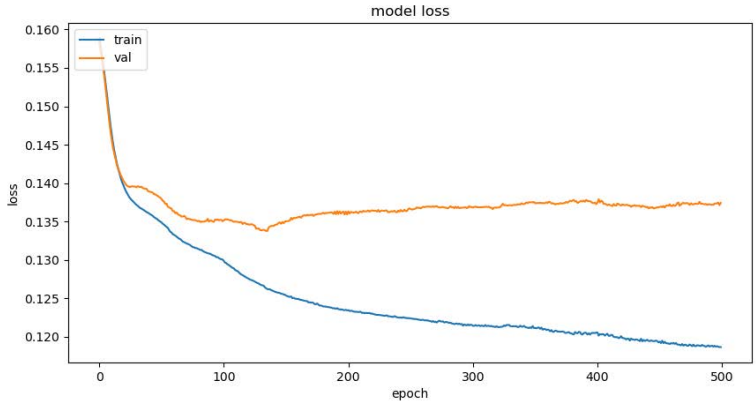
Fuente: Elaboración propia.

Fig. 4.7. Arquitectura simple perceptrón multicapa para clasificador general.



Fuente: Elaboración propia.

Fig. 4.9. Precisión FFNN para clasificador general.



Fuente: Elaboración propia.

Fig. 4.11. Error FFNN para clasificador general.

4.2.2. Apr. 2.0: Clasificador binario (Positivo, Negativo)

Dados los resultados previos, en lugar de realizar un clasificador general, se trabajó en unos independientes para las emociones contrarias, en particular positiva y negativa.

Clasificador positivo

En este caso, se convierte la variable a clasificar de forma binaria, donde 1 [emotion=alegría] y 0 [emotion=tristeza, miedo, enfado, desprecio], en total 875 audios por cada clase. Al entrenar los modelos con esta modificación, los resultados mejoran significativamente.

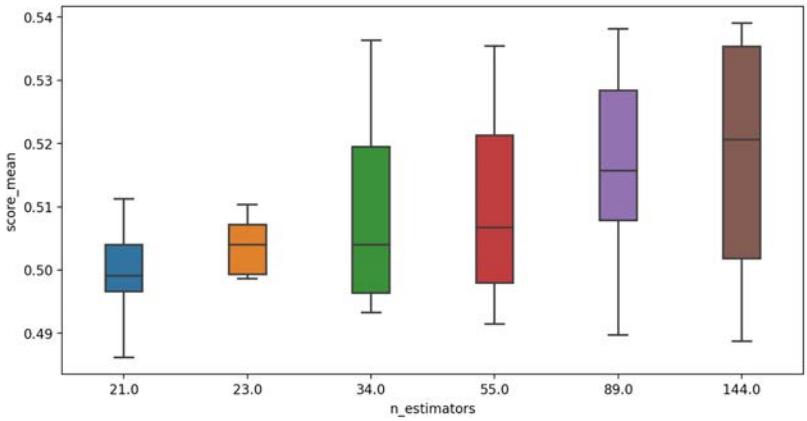
Después de entrenar los modelos con las especficaciones mencionadas previamente (*stratified k-fold*, 80-20 y matriz de experimentación), el mejor modelo obtuvo una precisión del 84,57 % , lo cual representó una mejora significativa respecto al anterior enfoque.

	f1_score_mean	f1_score_max	accuracy_mean	accuracy_max	score_mean	score_max
0	0.8026	0.8454	0.8034	0.8457	0.8034	0.8457

```
RandomForestClassifier(max_depth=12, max_features='log2', n_estimators=144, random_state=9103)
```

Fuente: Elaboración propia.

Fig. 4.13. Mejor modelo RF para clasificador positivo.



Fuente: Elaboración propia.

Fig. 4.15. Variación de precisión en número de estimadores para RF.

Posteriormente, para verificar los resultados se realizó una prueba de validación con sólo los audios originales, sin haber sido transformados. Los resultados de la matriz de clasificación fueron los esperados.

Valor	Precisión	Recall	F1-score	Support
0	1.00	1.00	1.00	175
1	1.00	1.00	1.00	175

TABLA 4.3. INFORME CLASIFICACIÓN RF POSITIVO.

Para el caso negativo, se convierte la variable a clasificar de forma binaria, donde 1 [emotion=enfado, desprecio] y 0 [emotion=tristeza, miedo, alegría], en total 715 audios por cada clase. Similar al clasificador positivo, el mejor modelo obtuvo una precisión del 86,71 %.

	f1_score_mean	f1_score_max	accuracy_mean	accuracy_max	score_mean	score_max
0	0.8097	0.8671	0.8098	0.8671	0.8098	0.8671

```
RandomForestClassifier(criterion='entropy', max_depth=12, max_features='log2', n_estimators=89,
random_state=9103)
```

Fuente: Elaboración propia.

Fig. 4.17. Mejor modelo RF para clasificador negativo.

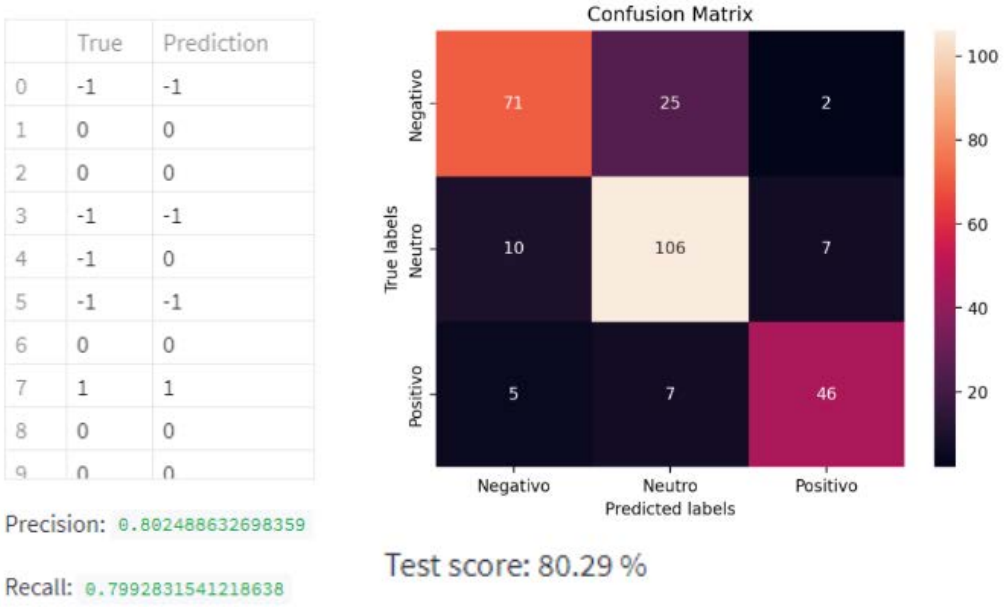
Valor	Precisión	Recall	F1-score	Support
0	1.00	1.00	1.00	143
1	1.00	1.00	1.00	143

TABLA 4.4. INFORME CLASIFICACIÓN RF POSITIVO.

4.2.3. Apr. 3.0: Stacking Logístico

Para combinar ambos resultados, se almacenan los modelos entrenados en un objeto serializable empleando la librería *Pickle*. Cargando ambos modelos, se entrenan los audios realizando una combinación de las salidas de los clasificadores binarios como entradas para un nuevo clasificador, esto se conoce como un *stacking*.

Este nuevo clasificador tendrá tres salidas, correspondientes a las tres emociones POS, NEU y NEG. Sin embargo, para la continuación del proyecto, bastan sólo POS y NEG, ya que la tercera es la parte restante sobre el porcentaje y no suma información para nuevos modelos.
Para este tipo de problema, el modelo que mejores resultados ha dado, tanto en precisión, exhaustividad y exactitud como en tiempo de respuesta, es la regresión logística.



Fuente: Elaboración propia.

Fig. 4.19. Resultados RLog para el stacking de audio.

Sin embargo, las redes convolucionales también se han comportado favorablemente al problema, pese a no estar diseñadas especificamente para este tipo de problemas.
La arquitectura de la red neuronal convolucional ha seguido el enfoque tomado por un trabajo similar de acceso público en *Github* que trata las redes en formato bidimensional [61].

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 12, 128)	768
activation (Activation)	(None, 12, 128)	0
dropout (Dropout)	(None, 12, 128)	0
max_pooling1d (MaxPooling1D)	(None, 2, 128)	0
conv1d_1 (Conv1D)	(None, 2, 128)	82048
activation_1 (Activation)	(None, 2, 128)	0
dropout_1 (Dropout)	(None, 2, 128)	0
flatten (Flatten)	(None, 256)	0
dense_4 (Dense)	(None, 5)	1285
activation_2 (Activation)	(None, 5)	0

Total params: 84,101
Trainable params: 84,101
Non-trainable params: 0

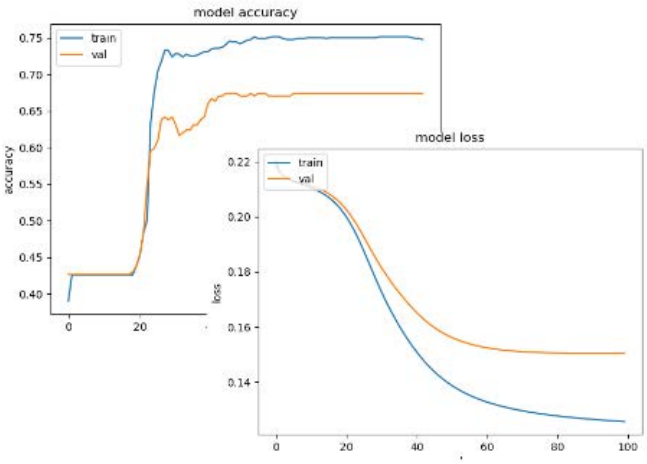
Fuente: Elaboración propia.

Fig. 4.21. Arquitectura bidimensional de red convolucional para stacking de audio.

Atributos y salidas quedan...

	RF_Pos	RF_Neg
850	0.7422	0.4110
672	0.8356	0.8226
507	0.7230	0.2451
882	0.3250	0.2485
219	0.4260	0.6977

	0	1	2
0	0	0	1
1	0	0	1
2	0	0	1
3	0	1	0
4	1	0	0



Fuente: Elaboración propia.

Fig. 4.23. Resultados CNN para el stacking de audio.

4.2.4. Apr. 4.0: Speech to Text. Bert

Una vez que se obtuvieron los resultados de audio favorables, se decidió extender la arquitectura incorporando el contexto de la parte escrita.
Para ello, era necesario encontrar una herramienta de transcripción del audio a texto, y posteriormente probar el modelo de extracción de sentimiento.

En el capítulo del estado del arte se detallan las herramientas que fueron valoradas. Para la transcripción fue escogida *Google STT*. Para poder transcribir el mayor número de audios, se realizó un experimento con los distintos códecs posibles permitidos según la documentación de la API:

La API de Speech-to-Text admite varias codificaciones diferentes. En la siguiente tabla, se enumeran los códecs de audio compatibles:

Códec	Nombre	Sin pérdida	Notas de uso
MP3	Capa de audio MPEG III	No	La codificación MP3 es una función beta y solo está disponible en v1p1beta1. Consulta la documentación de referencia de RecognitionConfig para obtener más información.
FLAC	Códec de audio sin pérdida gratuito	Sí	Se requieren 16 bits o 24 bits para transmisión continua.
LINEAR16	PCM lineal	Sí	Codificación de modulación lineal por impulsos codificados (PCM) de 16 bits. El encabezado debe contener la tasa de muestreo.
MULAW	Ley μ	No	Codificación PCM de 8 bits
AMR	Banda estrecha con tasas de transferencia múltiples adaptables	No	La tasa de muestreo debe ser de 8,000 Hz.
AMR_WB	Banda ancha con tasas de transferencia múltiples adaptables	No	La tasa de muestreo debe ser de 16,000 Hz.
OGG_OPUS	Tramas de audio con codificación Opus en un contenedor Ogg	No	La tasa de muestreo debe ser una de las siguientes: 8,000 Hz, 12,000 Hz, 16,000 Hz, 24,000 Hz o 48,000 Hz.
SPEEX_WITH_HEADER_BYTE	Banda ancha de Speex	No	La tasa de muestreo debe ser de 16,000 Hz.
WEBM_OPUS	WebM Opus	No	La tasa de muestreo debe ser una de las siguientes: 8,000 Hz, 12,000 Hz, 16,000 Hz, 24,000 Hz o 48,000 Hz.

Fuente: Google.

Fig. 4.25. Especificación códecs de audio para la API de Google.

Mediante la librería de python *Pydub*, se fueron transformando el conjunto de audios y posteriormente, probando la herramienta STT. A continuación se detallan los distintos códecs empleados.

Codec	**Ratio**
WAV	89,7 %
MP3	64,6 %
FLAC	52,4 %

TABLA 4.5. RESULTADOS TRANSCRIPCIÓN INSTANCIAS ESPAÑOL POR CODECS.

Como se puede observar, el códec que mejor resultados de transcripción fue *WAV*. Además, se verificó que los codecs no afectaban al comportamiento de los MFCCs.

Posteriormente, para la extracción de emociones del audio, se empleó el transformer basado en BETO denominado BERTuito. Un ejemplo de la entrada y salida en texto de este modelo, con las probabilidades de sentimiento para cada clase.

```
Esto es pésimo

AnalyzerOutput(output=NEG, probas={NEG: 0.948, NEU: 0.048, POS: 0.004})
```

Fuente: Elaboración propia.

Fig. 4.27. Ejemplo de entrada y salida del Transformer desarrollado en BETO.

La herramienta STT de Google se empleó para transcribir el inglés y español de la base de datos original, dado que el transformer no funciona con el italiano.
Empleando el modelo de extracción de sentimiento del texto, para la transcripción de los audios de idioma Español e Inglés, dió los siguientes resultados:

```
Numero de casos clasificados como positivo correctamente del total en % 240
71.25
Numero de casos clasificados como negativo correctamente del total en % 400
80.5
Numero de casos clasificados como neutro correctamente del total en % 475
82.52631578947368
```

Fuente: Elaboración propia.

Fig. 4.29. Resultados del Transformer desarrollado en BETO.

A continuación se muestra toda la información recopilada hasta el momento para el modelo.

RF_Pos	RF_Neg	Log_Neg	Log_Pos	Google	BERT	Emotion	Idioma
0.7565	0.5760	0.2434	0.6264	18 años voy a poder estar al mando de mi propia vida Ay Dios mío	NEG	alegria	es
0.2982	0.4881	0.4101	0.0371	50	NEU	miedo	en
0.2494	0.7621	0.8801	0.0038	6 birthday	NEU	enfado	en
0.7245	0.6220	0.3833	0.4767	a las chicas les gusta	NEU	alegria	es
0.4428	0.3028	0.1089	0.2129	A quién acudir	NEU	miedo	es
0.4710	0.6353	0.6815	0.0740	a un amigo tuyo	NEU	rabia	es
0.4763	0.4865	0.3644	0.1699	absolute classics	POS	alegria	en
0.5556	0.3054	0.0822	0.4385	además aseguro que ya me lo sé	NEU	alegria	es
0.2659	0.3405	0.1641	0.0447	alguien puede decirme dónde está el fallo de esa fotografía	NEU	enfado	es
0.4289	0.6220	0.6702	0.0562	all over you going to follow you around forever	POS	rabia	en
0.3661	0.7763	0.8907	0.0098	alright this apartment situation	NEG	rabia	en
0.5127	0.5732	0.5302	0.1514	always talk so much trouble	NEG	miedo	en
0.4294	0.3871	0.2074	0.1614	and I didn't feel anything	NEG	tristeza	en
0.4470	0.6389	0.6968	0.0587	and I'm really hoping for the money	POS	enfado	en
0.1347	0.2347	0.0728	0.0163	and if you leave I can't learn this way	NEG	tristeza	en
0.2174	0.4650	0.3639	0.0196	and then one day nakamas this wonderful little boy	POS	tristeza	en
0.5486	0.4583	0.2660	0.3131	and things have been different	NEU	tristeza	en
0.2386	0.8810	0.9531	0.0012	and this time really f****** you acted like you actually gave us	NEG	rabia	en
0.4386	0.4650	0.3376	0.1366	apenas se hablan	NEU	tristeza	es

Fuente: Elaboración propia.

Fig. 4.31. Entradas y resultados del stacking de audio.

4.2.5. Apr. 5.0: Stacking Dual (Audio+Texto)

En esta última parte, se entrena un modelo que haga *stacking* de ambas entradas, audio y texto. Para ello, se extraen de las columnas recopiladas aquellas que nos interesan.

En este sentido, por un lado las referentes a la regresión logística para emoción positiva y negativa, y por otro las probabilidades positiva y negativa de BERT. En la siguiente imagen se muestran las entradas y variable a predecir del conjunto de datos que conforman el *stacking* final.

Alegría = 1, Enfado y Rabia = -1, Miedo y Tristeza = 0

	Log_Neg	Log_Pos	BERT_NEG	BERT_POS	Emotion
0	0.1402	0.5698	0.0870	0.0120	tristeza
3	0.8520	0.0752	0.0040	0.0010	rabia
4	0.9468	0.0079	0.0550	0.0010	rabia
7	0.0536	0.4675	0.4240	0.0860	tristeza
10	0.0281	0.0697	0.0660	0.0040	tristeza
11	0.0399	0.7257	0.0640	0.0040	alegria
14	0.0227	0.6048	0.0370	0.0300	alegria
15	0.8722	0.0335	0.0210	0.0030	rabia
20	0.9516	0.0167	0.0550	0.0050	rabia
21	0.7588	0.0972	0.0230	0.0010	enfado

Fuente: Elaboración propia.

Fig. 4.33. Muestra de las entradas de datos para el stacking completo (Audio+Texto).

Al entrenar los posibles modelos, de nuevo los mejores resultados para la función de *stacking* fue la regresión logística.

	True	Prediction
0	0	-1
1	-1	-1
2	1	1
3	0	0
4	-1	-1
5	-1	0
6	0	-1
7	0	-1
8	0	-1
9	0	0

Precision: `0.7337972259842153`

Recall: `0.7304347826086957`

Fuente: Elaboración propia.

Fig. 4.35. Mejor modelo RLog para stacking completo.

Como se muestra en la imagen, la precisión y exhaustividad del modelo final ronda el 73 %. Esto no debe malinterpretarse con los resultados del *stacking* de solo audio. La bajada en las métricas radica en posibles problemas de transcripción de la base de datos empleada, dado que al tratarse de audios de películas de duración recortada suelen ser frases sacadas de contexto.

Es por ello, que posteriormente al practicar con audios reales, expresiones como el sarcasmo o la ironía transmiten emociones con facetas que no pueden verse reducidas a la parte acústica, y gracias al texto se pueden clasificar mejor.

5. DESPLIEGUE DE LA IA

De forma que se pueda practicar la eficacia de los modelos de una forma sencilla y pública, se propuso integrar la arquitectura de modelos en un *bot* dentro de la aplicación de mensajería Telegram.

5.1. Desarrollo del ChatBot

Haciendo uso de la librería de Python que conecta con la API de Telegram, se creó un bot inicial en la red social gracias a su servicio gratuito de generación de chatbots *BotFather*. El nombre del bot para el proyecto es SentiAudio [62].

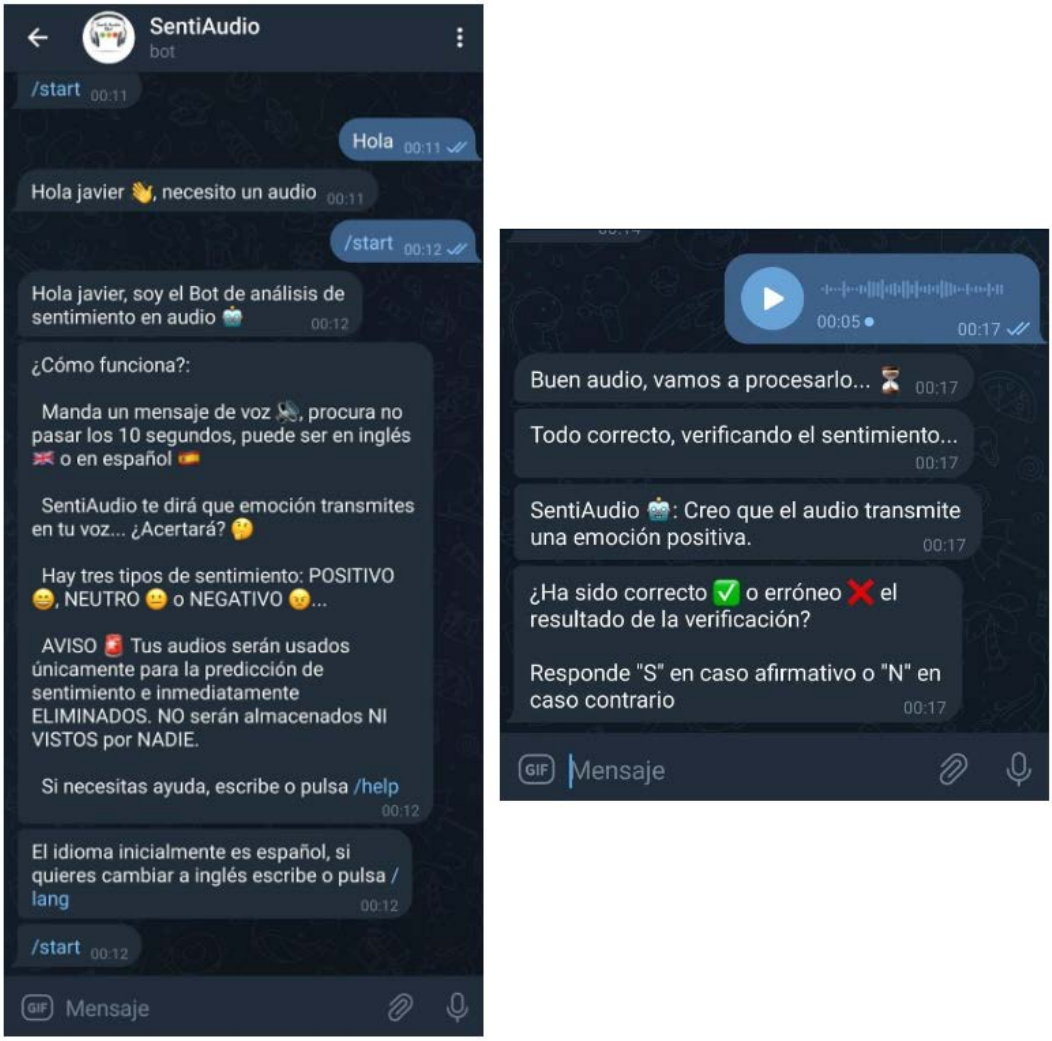
Fuente: Elaboración propia.

Fig. 5.1. Logo bot SentiAudio de Telegram.

Detrás de la interfaz, se creó un flujo de código que al recibir, después de ejecutar un comando de inicio, lanze un mensaje introductorio. Posteriormente, si recibe un audio, lo descarga, transforma a formato *WAV*, lo procesa la arquitectura de modelos y manda el resultado a SentiAudio.

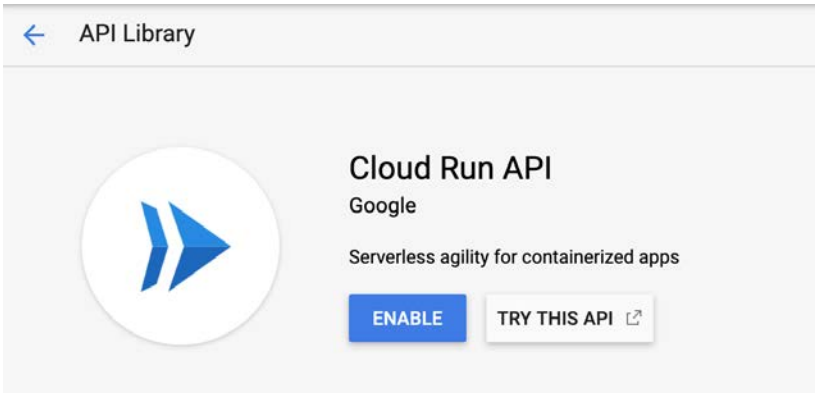
5.2. Implementación y facturación

Para mantener funcionando el bot en todo momento, es necesario buscar una solución en la nube, donde el código se ejecute con el comando de la petición inicial. En este propósito, existe un servicio dentro de la suite de Google, llamado *Cloud Run*, que tiene muchas ventajas, tanto de escalabilidad y latencia, como de costes por correr la máquina virtual. Actualmente el servicio está incorporado pero no activo.



Fuente: Elaboración propia.

Fig. 5.3. Muestra de conversación bot SentiAudio.



Fuente: Elaboración propia.

Fig. 5.5. Servicio *Cloud Run* de la nube de *GCP* para hostear el bot.

6. CONCLUSIONES, LIMITACIONES Y FUTUROS TRABAJOS

El proyecto que se detalla en el presente documento, tiene una perspectiva global, de principio a fin. Desde la parte de investigación de la literatura actual, de selección de datos, de desarrollo de modelos y de implementación final. Además, la problemática a resolver no es sencilla.

Dicho esto, el proyecto tiene mucha capacidad de mejora y crecimiento. Uno de los aspectos más importantes trata sobre la calidad del dato. Los actuales, pese a los buenos resultados de los modelos, no han conseguido plasmar la realidad completa de las emociones en una conversación telefónica.

De aquí se extraen propuestas de desarrollo como la creación de una base de datos categorizada de llamadas de servicio de atención al cliente, o la mejora del *corpus* de palabras del modelo de sentimiento introduciendo términos o expresiones concretas del ámbito de la telefonía móvil.

Otro aspecto es la exactitud de los MFCCs, ya que si se trata de audios de larga duración pierde sentido de manera exponencial.

Otra vía exploratoria, en la herramienta SentiAudio, sería la recopilación de *logs* y resultados para, cuidando la privacidad del usuario, extender la base de datos de trabajo para reentrenar los modelos y mejorar los resultados.

BIBLIOGRAFÍA

[1] J. R. Pierce, "Whither Speech Recognition?" *The Journal of the Acoustical Society of America*, vol. 46, p. 1049, 4B ago. de 2005. DOI: 10.1121/1.1911801. [En línea]. Disponible en: https://asa.scitation.org/doi/abs/10.1121/1.1911801.

[2] A. M. Turing, "ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM," 12.

[3] T. B. Brown et al., "Language Models are Few-Shot Learners," 2020.

[4] N. Team et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," [En línea]. Disponible en: https://github.com/facebookresearch/fairseq/tree/nllb..

[5] J. M. G. Segismundo S. Izquierdo y J. I. S. L. R. Izquierdo, "Techniques to Understand Computer Simulations: Markov Chain Analysis," ene. de 2009.

[6] W. S. McCulloch y W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics 1943 5:4*, vol. 5, pp. 115-133, 4 dic. de 1943. DOI: 10.1007/BF02478259. [En línea]. Disponible en: https://link.springer.com/article/10.1007/BF02478259.

[7] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999-6009, jun. de 2017. DOI: 10.48550/arxiv.1706.03762. [En línea]. Disponible en: https://arxiv.org/abs/1706.03762v5.

[8] *Del 'call center' a la conversación inteligente | Transformación Digital | Tecnología | EL PAÍS*. [En línea]. Disponible en: https://elpais.com/retina/2019/03/18/tendencias/1552915693_698195.html.

[9] J. D. Estado, "Disposición 16673 del BOE núm. 294 de 2018," 2018. [En línea]. Disponible en: http://www.boe.es.

[10] "REGLAMENTO (UE) 2016/ 679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO - de 27 de abril de 2016 - relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/ 46/ CE (Reglamento general de protección de datos),"

[11] *The Institute for Ethical AI Machine Learning*. [En línea]. Disponible en: https://ethical.institute/principles.html.

[12] C. P. Loizou y P. Christodoulides, "Voice signal analysis techniques for cognitive decline (stress) assessment," *Journal of Physics: Conference Series*, vol. 1687, 1 nov. de 2020. DOI: 10.1088/1742-6596/1687/1/012007. [En línea]. Disponible en: https://www.researchgate.net/publication/346445704_Voice_signal_analysis_techniques_for_cognitive_decline_stress_assessment.

[13] D. Roy y A. Pentland, "Automatic spoken affect classification and analysis," *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pp. 363-367, 1996. DOI: 10.1109/AFGR.1996.557292.

[14] D. H. Bailey y P. N. Swarztrauber, "A Fast Method for the Numerical Evaluation of Continuous Fourier and Laplace Transforms," *https://doi.org/10.1137/0915067*, vol. 15, pp. 1105-1110, 5 jul. de 2006. DOI: 10.1137/0915067. [En línea]. Disponible en: https://epubs.siam.org/doi/10.1137/0915067.

[15] L. A. Montejo y L. E. Suárez, "APLICACIONES DE LA TRANSFORMADA ONDÍCULA ("WAVELET") EN INGENIERÍA ESTRUCTURAL," [En línea]. Disponible en: http://civil.uprm.edu.

[16] J. Yang, F.-L. Luo y A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons q," [En línea]. Disponible en: www.elsevier.com/locate/specom.

[17] M. Pervaiz y T. Ahmed, "Emotion Recognition from Speech using Prosodic and Linguistic Features," *International Journal of Advanced Computer Science and Applications*, vol. 7, 8 2016. DOI: 10.14569/IJACSA.2016.070813. [En línea]. Disponible en: https://www.researchgate.net/publication/309624815_Emotion_Recognition_from_Speech_using_Prosodic_and_Linguistic_Features.

[18] *1: The relationship between the frequency scale and mel scale is shown... | Download Scientific Diagram*. [En línea]. Disponible en: https://www.researchgate.net/figure/The-relationship-between-the-frequency-scale-and-mel-scale-is-shown-on-the-left-The_fig1_336444658.

[19] P. Sivakumaran, A. M. Ariyaeeinia y M. J. Loomes, "SUB-BAND BASED TEXT-DEPENDENT SPEAKER VERIFICATION List of Unusual Symbols and Abbreviations Used,"

[20] K. V. K. Kishore y P. K. Satish, "Emotion recognition in speech using MFCC and wavelet features," *Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013*, pp. 842-847, 2013. DOI: 10.1109/IADCC.2013.6514336.

[21] R. Móstoles, D. Griol, Z. Callejas y F. Fernández-Martínez, "A proposal for emotion recognition using speech features, transfer learning and convolutional neural networks," pp. 56-60, mar. de 2021. DOI: 10.21437/IBERSPEECH.2021-12. [En línea]. Disponible en: https://www.researchgate.net/publication/

348929983_A_proposal_for_emotion_recognition_using_speech_features_transfer_learning_and_convolutional_neural_networks.

[22] *The Sound of AI – Valerio Velardo*. [En línea]. Disponible en: https://valeriovelardo.com/the-sound-of-ai/.

[23] *Understanding Audio data, Fourier Transform, FFT and Spectrogram features for a Speech Recognition System | by Kartik Chaudhary | Towards Data Science*. [En línea]. Disponible en: https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520.

[24] N. Sato e Y. Obuchi, "Emotion Recognition using Mel-Frequency Cepstral Coefficients," *Journal of Natural Language Processing*, vol. 14, pp. 83-96, 4 2007. DOI: 10.5715/JNLP.14.4_83.

[25] S. Lalitha, D. Geyasruti, R. Narayanan y M. Shravani, "Emotion Detection Using MFCC and Cepstrum Features," *Procedia Computer Science*, vol. 70, pp. 29-35, 2015. DOI: 10.1016/J.PROCS.2015.10.020. [En línea]. Disponible en: https://cyberleninka.org/article/n/585189.

[26] A. E. O. C. D. I. J. A. H. Londoño, "DETECCIÓN DE ESTADOS DE ÁNIMO MEDIANTE EL PROCESAMIENTO DE SEÑALES ACÚSTICAS," 2017.

[27] *MENHIR – Mental health monitoring through interactive conversations*. [En línea]. Disponible en: https://menhir-project.eu/.

[28] V. Tiwari, "MFCC and its applications in speaker recognition,"

[29] T. Mikolov, K. Chen, G. Corrado y J. Dean, "Distributed Representations of Words and Phrases and their Compositionality,"

[30] A. Vaswani et al., "Attention Is All You Need,"

[31] J. Devlin, M. W. Chang, K. Lee y K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171-4186, oct. de 2018. DOI: 10.48550/arxiv.1810.04805. [En línea]. Disponible en: https://arxiv.org/abs/1810.04805v2.

[32] J. Cã et al., "SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA," [En línea]. Disponible en: https://github.com/josecannete/spanish-corpora.

[33] J. M. Pérez, J. C. Giudici y F. Luque, "pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks," jun. de 2021. [En línea]. Disponible en: https://arxiv.org/abs/2106.09462v1.

[34] J. M. Pérez, D. A. Furman, L. A. Alemany y F. Luque, "RoBERTuito: a pre-trained language model for social media text in Spanish," nov. de 2021. [En línea]. Disponible en: http://arxiv.org/abs/2111.09453.

[35] *Coqui*. [En línea]. Disponible en: https://coqui.ai/.

[36] *Speech-to-Text: Automatic Speech Recognition | Google Cloud*. [En línea]. Disponible en: https://cloud.google.com/speech-to-text.

[37] *streamlit · PyPI*. [En línea]. Disponible en: https://pypi.org/project/streamlit/.

[38] *pandas · PyPI*. [En línea]. Disponible en: https://pypi.org/project/pandas/.

[39] *numpy · PyPI*. [En línea]. Disponible en: https://pypi.org/project/numpy/.

[40] *matplotlib · PyPI*. [En línea]. Disponible en: https://pypi.org/project/matplotlib/.

[41] *seaborn · PyPI*. [En línea]. Disponible en: https://pypi.org/project/seaborn/.

[42] *scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation*. [En línea]. Disponible en: https://scikit-learn.org/stable/index.html.

[43] *augly · PyPI*. [En línea]. Disponible en: https://pypi.org/project/augly/.

[44] *pickle — Python object serialization — Python 3.10.6 documentation*. [En línea]. Disponible en: https://docs.python.org/3/library/pickle.html.

[45] *librosa · PyPI*. [En línea]. Disponible en: https://pypi.org/project/librosa/.

[46] *SpeechRecognition · PyPI*. [En línea]. Disponible en: https://pypi.org/project/SpeechRecognition/#description.

[47] *pysentimiento · PyPI*. [En línea]. Disponible en: https://pypi.org/project/pysentimiento/.

[48] *pydub · PyPI*. [En línea]. Disponible en: https://pypi.org/project/pydub/.

[49] *requests · PyPI*. [En línea]. Disponible en: https://pypi.org/project/requests/.

[50] *python-telegram-bot · PyPI*. [En línea]. Disponible en: https://pypi.org/project/python-telegram-bot/.

[51] S. R. Livingstone y F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," abr. de 2018. DOI: 10.5281/ZENODO.1188976. [En línea]. Disponible en: https://zenodo.org/record/1188976.

[52] *Emo-DB*. [En línea]. Disponible en: http://emodb.bilderbar.info/start.html.

[53] *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. [En línea]. Disponible en: http://kahlan.eps.surrey.ac.uk/savee/.

[54] L. Rachman et al., "DAVID: An open-source platform for real-time emotional speech transformation," *bioRxiv*, p. 038 133, ene. de 2016. DOI: 10.1101/038133. [En línea]. Disponible en: https://www.biorxiv.org/content/10.1101/038133v1%20https://www.biorxiv.org/content/10.1101/038133v1.abstract.

[55] *Datasets — SuperKogito/SER-datasets documentation*. [En línea]. Disponible en: https://superkogito.github.io/SER-datasets/.

[56] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird y B. W. Schuller, "Categorical vs dimensional perception of Italian emotional speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, pp. 3638-3642, 2018. DOI: 10.21437/INTERSPEECH.2018-47.

[57] *librosa.feature.mfcc — librosa 0.10.0.dev0 documentation*. [En línea]. Disponible en: https://librosa.org/doc/main/generated/librosa.feature.mfcc.html.

[58] Z. Papakipos, J. Bitton y M. Ai, "AugLy: Data Augmentations for Robustness," [En línea]. Disponible en: https://github..

[59] C. Sobin y M. Alpert, "Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy," *Journal of Psycholinguistic Research 1999 28:4*, vol. 28, pp. 347-365, 4 1999. DOI: 10.1023/A:1023237014909. [En línea]. Disponible en: https://link.springer.com/article/10.1023/A:1023237014909.

[60] *RandomUnderSampler — Version 0.9.1*. [En línea]. Disponible en: https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

[61] *GitHub - yfliao/Emotion-Classification-Ravdess: Understanding emotions with Neural Networks (Python, Scikit-Learn, Keras) and the Ravdess dataset.* [En línea]. Disponible en: https://github.com/yfliao/Emotion-Classification-Ravdess/blob/master/EmotionsRecognition.ipynb.

[62] *Telegram: Contact @sentiespanol$_b$ot*. [En línea]. Disponible en: https://t.me/sentiespanol_bot.

DECLARACION DE ORIGINALIDAD

Yo, Javier Moncada Gutiérrez declaro que el TFG *"Análisis de Sentimiento en Audio mediante Inteligencia Artificial orientado al idioma Español"* es totalmente original mío, que no ha sido presentado en ninguna otra universidad como TFG y que todas las fuentes que han sido utilizadas han sido adecuadamente citadas y aparecen en las referencias bibliográficas.

Colmenarejo, a 07/09/2022:

Javier Moncada Gutiérrez.