

**UNIVERSIDAD EUROPEA DE MADRID**

**ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO**

**GRADO EN INGENIERÍA MATEMÁTICA**

**APLICADA AL ANÁLISIS DE DATOS**

PROYECTO FIN DE GRADO

**AFFECTIVE COMPUTING: HERRAMIENTA DE ACOMPAÑAMIENTO BASADA EN EL ANÁLISIS DE EMOCIONES**

**MARTA ALMENDRO ÁLVAREZ**

**Dirigido por**

**ANA DEL VALLE CORRALES PAREDES**

**CURSO 2020-2021**

**TÍTULO**: AFFECTIVE COMPUTING: HERRAMIENTA DE ACOMPAÑAMIENTO BASADA EN EL ANÁLISIS DE EMOCIONES

**AUTOR**: MARTA ALMENDRO ÁLVAREZ

**TITULACIÓN**: GRADO EN INGENIERÍA MATÉMATICA APLICACA AL ANÁLISIS DE DATOS

**DIRECTOR/ES DEL PROYECTO**: ANA DEL VALLE CORRALES PAREDES

**FECHA**: JUNIO DE 2021

# RESUMEN

Las emociones habitan en nuestro interior, como un fiel compañero que responde por nosotros a los estímulos externos. Se pueden definir de manera biológica como las reacciones psicofisiológicas que empleamos como modo de adaptación a lo que percibimos a nuestro alrededor. Desde una perspectiva un poco más filosófica, las emociones son la expresión de nuestros sentimientos, la afección del alma, o como decretó Descartes, las ‘pasiones del alma’. Estas pasiones estan ahí para ser sentidas y disfrutadas, pero no se debe vivir a merced de ellas.

Este proyecto plantea una fusión entre lo antiguo y lo nuevo; cómo se podría integrar el estudio clásico de las emociones con las tecnologías emergentes a través de la Inteligencia Artificial (IA). Para ello se ha empleado el Deep Learning, una de las herramientas novel de la IA que simula la mecánica neuronal de la mente humana. A partir de este concepto se ha desarrollado un modelo basado en redes neuronales profundas artificiales que pueda predecir de manera eficaz las emociones de un usuario a través de ciertas características residentes en su voz.

Para poder hacer accesible este sistema al público, se integra con un componente de desarrollo de aplicaciones. A tal efecto se juntan todas las piezas del puzle en una única infraestructura con capacidad de dar respuesta a múltiples peticiones en tiempo real y ofrecer la máxima seguridad y confidencialidad de los usuarios. El resultado es una aplicación web y móvil con una interfaz de usuario que haga de este proceso de autoconocimiento toda una experiencia estética para el consumidor.

**Palabras clave:** Computación afectiva, aprendizaje profundo, clasificación de emociones a través de la voz, extracción de características, desarrollo Front-end, computación en la nube

# ABSTRACT

Emotions lie within un, like a loyal companion that will look out on external stimuli for us. If defined in biological context, they are the psychophysiological reactions we use to adapt to what we perceive around us. From a philosophical point of view, emotions are the expression of our feelings, our soul’s affection, or how Descartes once defined them, the ‘passions of our soul’. These passions are there to be felt and enjoyed, but never to succumb to.

The following project sets out a fusion between the archaic and the new; how to meld classic studies about emotions with emerging technologies through Artificial Intelligence (AI). To gain this, Deep Learning, one of IA’s novel tools that pursues mimicking the neurological mechanics behind the human mind, has been employed. Parting from this concept, an artificial neural network has been developed to efficiently predict a user’s emotions from certain characteristics extracted from his/her voice.

In order to make this product accessible to public, it is integrated with an app development module. All the puzzle’s pieces are put together in one single infrastructure with real time capacity to serve multiple requests and offer maximum security and confidentiality to users. The result is a web and mobile app with a user interface that hopefully makes this self-awareness process a whole new aesthetic experience.

**Keywords:** Affective Computing, Deep Learning, Speech Emotion Recognition, Feature Extraction, Front-End, Cloud Computing

**AGRADECIMIENTOS**

En ocasiones se incluye este apartado para agradecer a aquellos que han ofrecido su ayuda en el desarrollo del trabajo, ya sea técnica o de otro tipo.

**Cita - frase célebre / Dedicatoria**

Esta página es del todo opcional, pero resulta una muy buena forma de presentar el trabajo académico más importante de todo el grado.

# TABLA RESUMEN

|  |  |
| --- | --- |
|  | DATOS |
| Nombre y apellidos: | Marta Almendro Álvarez |
| Título del proyecto: | Affective Computing: Herramienta de Acompañamiento Basada en el Análisis de Emociones |
| Directores del proyecto: | Ana del Valle Corrales Paredes |
| El proyecto ha implementado un producto:  (esta entrada se puede marcar junto a la siguiente) | SI |
| El proyecto ha consistido en el desarrollo de una investigación o innovación:  (esta entrada se puede marcar junto a la anterior) | SI |
| Objetivo general del proyecto: | Desarrollar una aplicación multi plataforma integrada con un sistema de reconocimiento de las emociones del usuario a través de su voz. |

**Índice**

[RESUMEN 4](#_Toc37845866)

[ABSTRACT 5](#_Toc37845867)

[TABLA RESUMEN 8](#_Toc37845868)

[Capítulo 1. RESUMEN DEL PROYECTO 14](#_Toc37845869)

[1.1 Contexto y justificación 14](#_Toc37845870)

[1.2 Planteamiento del problema 14](#_Toc37845871)

[1.3 Objetivos del proyecto 14](#_Toc37845872)

[1.4 Resultados obtenidos 14](#_Toc37845873)

[1.5 Estructura de la memoria 15](#_Toc37845874)

[Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE 16](#_Toc37845875)

[2.1 Estado del arte 16](#_Toc37845876)

[2.2 Contexto y justificación 24](#_Toc37845877)

[2.3 Planteamiento del problema 25](#_Toc37845878)

[Capítulo 3. OBJETIVOS 29](#_Toc37845879)

[3.1 Objetivos generales 29](#_Toc37845880)

[3.2 Objetivos específicos 29](#_Toc37845881)

[3.3 Beneficios del proyecto 30](#_Toc37845882)

[Capítulo 4. DESARROLLO DEL PROYECTO 31](#_Toc37845883)

[4.1 Planificación del proyecto 31](#_Toc37845884)

[4.2 Descripción de la solución, metodologías y herramientas empleadas 36](#_Toc37845885)

[4.3 Recursos requeridos 52](#_Toc37845886)

[4.4 Presupuesto 53](#_Toc37845887)

[4.5 Viabilidad 54](#_Toc37845888)

[4.6 Resultados del proyecto 54](#_Toc37845889)

[Capítulo 5. DISCUSIÓN 63](#_Toc37845890)

[Capítulo 6. CONCLUSIONES 64](#_Toc37845891)

[6.1 Conclusiones del trabajo 64](#_Toc37845892)

[6.2 Conclusiones personales 64](#_Toc37845893)

[Capítulo 7. FUTURAS LÍNEAS DE TRABAJO 65](#_Toc37845894)

[Capítulo 8. REFERENCIAS 66](#_Toc37845895)

[Capítulo 9. ANEXOS 71](#_Toc37845896)

**Índice de Figuras**

[Figura 1 Sistema HMI: usuario humano, interfaz y máquina (Papetti,2013) 17](#_Toc72522329)

[Figura 2 Concentración de mercado del sector del Affective Computing (Market, 2021) 18](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522330)

[Figura 3 Estrella de emociones de Plutchik (Donaldson, 2017) 19](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522331)

[Figura 4 Sistema SER tradicional (Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. , 2019) 20](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522332)

[Figura 5 Características Segmentales y Suprasegmentales (Anagnostopoulos, Iliou & Giannoukos, 2012) 22](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522333)

[Figura 6 Método tradicional de Machine Learning vs. Deep Learning (Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. , 2019) 23](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522334)

[Figura 7 Página de entrada al programa Empath 24](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522335)

[Figura 8 Plantilla de la UI de la app Vmote 24](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522336)

[Figura 9 Página principal de la aplicación web Interview Simulator 25](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522337)

[Figura 10 Evolución del sentimiento del miedo en España durante el periodo del 1 del marzo 2020 hasta el 30 de abril 2020 (de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI., 2020) 27](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522338)

[Figura 11 Noticias relacionadas con el miedo y el Covid-19 en los medios de comunicación (de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI., 2020) 28](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522339)

[Figura 12 Pipeline principal detrás del sistema de reconocimiento de voz 37](#_Toc72522340)

[Figura 13 Representación del funcionamiento detrás de la técnica de oversampling SMOTE (Indresh Bhattacharyya, 2018) 41](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522341)

[Figura 14 Proceso de extracción de los coeficientes MFCC durante el Feature Extraction (Gong, S., Dai, Y., Ji, J., Wang, J., & Sun, H., 2015) 43](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522342)

[Figura 15 Mel filter banks basis functions using 20 Mel-filters in the filter bank (Yusnita, M. A., Paulraj, M. P., Yaacob, S., Yusuf, R., & Shahriman, A. B., 2013) 44](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522343)

[Figura 16 Secuencia del preprocesamiento de datos 45](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522344)

[Figura 17 Arquitectura de un modelo híbrido 1D CNN + LSTM (Hamad, R. A., Yang, L., Woo, W. L., & Wei, B., 2020) 47](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522345)

[Figura 18 Arquitectura de la estructura seleccionada y las dimensiones de los datos de salida de cada capa 50](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522346)

[Figura 19 Esquema del script desarrollado para la fase de entrenamiento 51](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522347)

[Figura 20 Estructura y flujo de datos de la aplicación propuesta 52](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522348)

[Figura 21 Logos de las tecnologías empleadas 53](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522349)

[Figura 22 Diagrama de barras mostrado el conteo de cada emoción antes de aplicar oversampling a los datasets femenino (izquierda) y masculino (derecha) 55](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522350)

[Figura 23 Diagrama de barras mostrado el conteo de cada emoción después de aplicar oversampling a los datasets femenino (izquierda) y masculino (derecha) 56](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522351)

[Figura 24 Espectograma de MFCCs de un audio de muestra: 03-01-05-02-01-01-08.wav, emoción de enfado de una mujer. Extracción de 20 coeficientes 56](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522352)

[Figura 25 Espectograma de MFCCs de un audio de muestra: 03-01-05-02-01-01-08.wav, emoción de enfado de una mujer. Extracción de 50 coeficientes 57](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522353)

[Figura 26 Gráficas de Accuracy y Loss durante el entrenamiento del segundo modelo de la tabla 9. 59](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522354)

[Figura 27 Comportamiento del Loss en función del Learning Rate (Rosebrock, 2021) 60](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522355)

[Figura 28 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género mixto 60](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522356)

[Figura 29 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género masculino 61](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522357)

[Figura 30 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género femenino 61](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522358)

[Figura 31 Matriz de confusión para las predicciones de emociones del dataset femenino 62](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522359)

[Figura 32 Matriz de confusión para las predicciones de emociones del dataset masculino 63](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522360)

[Figura 33 Sistema de reconocimiento de emociones 63](file:///C:\Users\almen\Documents\TFG\MEMORIA\memoriaTFG_v01.docx#_Toc72522361)

**Índice de Tablas**

[Tabla 1 Descripción de la identificación de archivos de audio de la base de datos RAVDESS 37](#_Toc72522362)

[Tabla 2 Descripción de la identificación de archivos de audio de la base de datos CREMA-D 39](#_Toc72522363)

[Tabla 3 Técnicas de Data Augmentation, descripción y librería de Python utilizada para su aplicación 41](#_Toc72522364)

[Tabla 4 Dimensiones de los diferentes conjuntos de datos utilizados 45](#_Toc72522365)

[Tabla 5 Hiperparámetros del modelo final 1D CNN + LSTM 47](#_Toc72522366)

[Tabla 6 Arquitectura seleccionada para la red neuronal artificial detrás del sistema final 48](#_Toc72522367)

[Tabla 7 Listado y descripción de las librerías empleadas en Python durante el proyecto 51](#_Toc72522368)

[Tabla 8 Estimación del presupuesto del trabajo realizado 54](#_Toc72522369)

[Tabla 9 Arquitectura de los modelos entrenados para el conjunto de datos de género mixto y los resultados de validation accuracy y los para cada etapa 57](#_Toc72522370)

[Tabla 10 Arquitectura de los dos últimos modelos entrenados para el conjunto de datos de género femenino y los resultados de validation accuray y los para cada etapa 57](#_Toc72522371)

[Tabla 11 Arquitectura de los modelos entrenados para el conjunto de datos de género masculino y los resultados de Validation Accuracy y Loss para cada etapa 57](#_Toc72522372)

# RESUMEN DEL PROYECTO

## Contexto y justificación

El estudio de la relación entre las emociones y la salud en los humanos se remonta a la Antigüedad. En la antigua Grecia esta idea se sustentaba en el hecho de que el cuerpo y la mente debían estar en armonía. La filosofía “mens sana in corpore sano”, que se traduce a “mente sana en un cuerpo sano”, era uno de los pilares de la medicina hipocrática.

La comunicación hombre-máquina se ha vuelto cada vez más ‘habladora’: Alexa, Cortana, Siri [3], entre muchos otros sistemas inteligente de diálogo, han triunfado en el mercado de consumo de manera destacada, pero ¿y si estos realmente pudiesen notar nuestras emociones y reaccionar a ellas como un humano haría? Esto haría de la HMI una comunicación mucho más natural, efectiva y agradable. La disciplina de reconocimiento automático de las emociones humanas y estados afectivos a través de la voz es conocida como Speech Emotion Recognition (SER).

## Planteamiento del problema

El contacto físico hoy en día supone un peligro grave para muchos, por lo que para muchos el hogar es su día a día. Este día a día incluye el cuidado de la salud, de modo que, este proyecto quiere ayudar a crear una alternativa al contacto físico para proporcionar seguridad en momentos de mayor malestar emocional.

Este proyecto trata de conseguir una solución que pueda aliviar a los sistemas sanitarios, y al mismo tiempo incrementar el estado afectivo de los usuarios, que podrán mantener un estado físico y mental saludables desde la comodidad del hogar. Sin embargo, su uso puede ir más allá y puede ser de utilidad, por ejemplo, para trabajadores en oficinas.

## Objetivos del proyecto

Este proyecto propone una aplicación de reconocimiento de emociones a través de la voz, siguiendo la dinámica de la interacción hombre-maquina, y sumergiéndose en el campo del Affective Computing. Para ello se plantea desarrollar un algoritmo de reconocimiento, como estructura de red neuronal artificial, empleando diferentes estructuras y experimentando con diferentes parámetros. A su vez, se persigue construir una arquitectura front-end con una interfaz de usuario (UI) adecuada y un diseño que haga de la experiencia de usuario (UX) lo más dinámica y satisfactoria posible. Por último, conectar ambos componentes back-end y front-end en una arquitectura alojada en servicios de computación en la nube que den servicio en formato de aplicación multiplataforma.

## Resultados obtenidos

Incluye el resumen de los resultados obtenidos al final de tu proyecto. En el capítulo correspondiente a resultados se detallan estos resultados.

## Estructura de la memoria

La estructura de la memoria corresponde a los siguientes apartados:

**Capítulo 1:** contiene un breve resumen del contexto del proyecto, el planteamiento del problema, los objetivos y los resultados,

**Capítulo 2:** pone en contexto el proyecto y se especifican aquellos aspectos que resguardan las técnicas estudiadas y aplicadas a través del estado del arte,

**Capítulo 3:** incluye una descripción detallada de los objetivos, general y específicos, y los beneficios del proyecto,

**Capítulo 4:** expone el desarrollo completo del proyecto, desde la planificación, las herramientas y recursos empleados, en este caso, mayoritariamente computacionales, hasta la viabilidad, el presupuesto estimado y los resultados obtenidos, además del componente de innovación,

**Capítulo 5:** plantea una discusión de los resultados obtenidos,

**Capítulo 6:** incluye las conclusiones obtenidas a raíz de los resultados,

**Capítulo 7:** propone las futuras líneas de trabajo consideradas tras la finalización del proyecto.

# ANTECEDENTES / ESTADO DEL ARTE

## Estado del arte

*Human-Machine Interaction*

Para poder emplear la tecnología en el campo de reconocimiento de emociones humanas, debe existir la comunicación entre un humano y una máquina. De aquí nace el concepto Interacción Hombre-Máquina (HMI), descrito como la interacción y comunicación entre un usuario humano y una máquina (sistema técnico dinámico), a través de una interfaz hombre-máquina [8]. Para esta última parte, hay que tener en cuenta tanto la User Interface (UI) y la User Experience (UX).

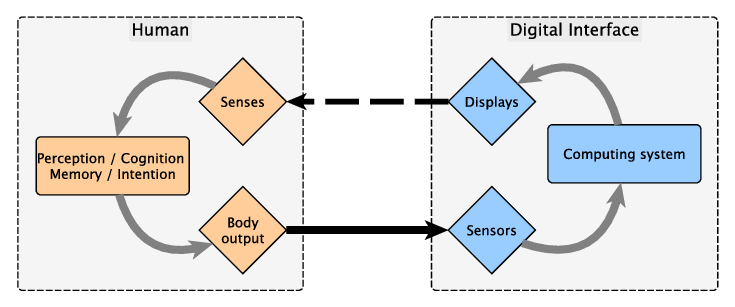


Figura 1 Sistema HMI: usuario humano, interfaz y máquina (Papetti,2013)

*Affective Computing*

Una de las grandes fuentes de inspiración de este proyecto viene de la idea del Affective Computing (AC). Este término fue propuesto por Rosalind Picard en 1997 como la computación que se relaciona, emerge o influye en las emociones [6]. Esto lo convierte en una modalidad de computación que nos permite comunicarnos e interactuar con ordenadores, robots u otras tecnologías a través de nuestras emociones [9]. El AC persigue formar una base sólida en la investigación sobre la relación entre los estados afectivos, cognitivos y físicos del ser humano [10]. Un valor añadido de esta modalidad de computación es que se pueden acumular grandes cantidades de datos para el desarrollo de las ciencias cognitivas que servirán para el entrenamiento de tecnologías avanzadas que puedan, entre otras cosas, reforzar la capacidad de tomar decisiones. A través de tecnologías que monitorean y analizan las emociones, se pueden detectar amenazas a la salud tanto mental como física, incluso logrando reducir el problema de la detección de enfermedades tardías.

Según un informe de AllTheReasearch, los gigantes de la industria del AC son Google e IBM. Los productos se segregan entre hardware y software y entre sus más destacadas aplicaciones está la sanidad. El segmento de la atención médica tiene algunas de las aplicaciones más avanzadas y comercializadas de AC. Empresas como DeepMind y Babylon Health están realizando importantes esfuerzos para aumentar la contribución al dominio [3]. La empresa Babylon lanzó en 2016 una aplicación de consultas médicas online; a través de un sistema de reconocimiento de voz y de inspección de historiales clínicos, ofrece una ruta de acción adecuada a la circunstancia médica en cuestión [14]. La contribución de Google en el sector es Google Health, una plataforma que promueve el descubrimiento de nuevas oportunidades en el campo de la AI con el fin de mejorar la eficacia de las tecnologías sanitarias a nivel global [15].

En cuanto al impacto que ha tenido la crisis del Covid-19 en este mercado, el uso generalizado de los dispositivos móviles y la penetración de internet en tantos rincones del mundo ha fomentado la inclinación progresiva hacia el uso de tecnologías digitales como el reconocimiento facial y de voz para mantener las conexiones de manera virtual. Además, aparecen las soluciones informáticas que permiten emplear estas tecnologías de reconocimiento para la detección de temperatura en el control de la propagación del virus. Muchas industrias están invirtiendo fuertemente en I+D para desarrollar software que ayude a controlar la propagación del Covid-19 [mercado2].

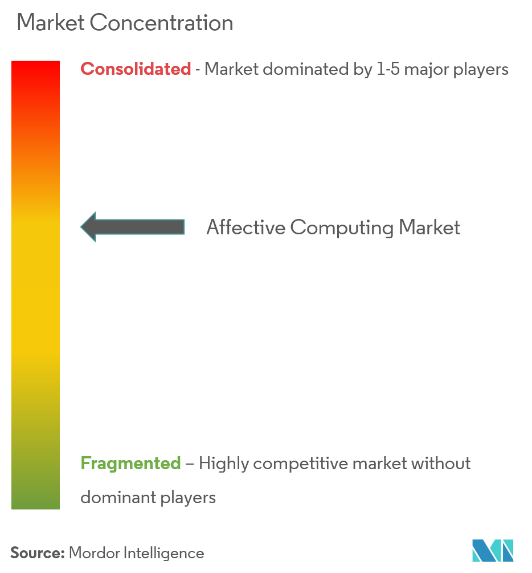


Figura 2 Concentración de mercado del sector del Affective Computing (Market, 2021)

*Speech Emotion Recognition (SER)*

Emplear las tecnologías en el reconocimiento automático de las emociones humanas y estados afectivos a través de la voz, normalmente conocido como Speech Emotion Recognition o SER, lleva siendo objeto de investigación varios años ya [SER]. Ejemplo de ello son los experimentos realizados (France et al. 2000) donde se descubrieron las propiedades acústicas del habla como identificadores de depresión y riesgo de suicidio [expt]. Otros campos donde se han empleado sistemas SER son la educación, el entretenimiento, la industria automotriz, y en sistemas de síntesis de voz natural [sectors].

Los humanos poseemos una habilidad natural de emplear todos nuestros sentidos para obtener la máxima percepción de los estímulos y mensajes que recibimos. La detección de emociones es un proceso natural e innato en los humanos, pero es una tarea compleja para las máquinas, ya que carecen de estas cualidades humanas. Por lo tanto, si el problema se desglosa, la tarea inicial es definir cómo se miden las emociones humanas. El estado emocional no puede ser reconocido de manera directa, por lo tanto, se queda en manos de las expresiones emocionales. Estas últimas se rigen por el sistema motor a través de varias formas; las principales son la voz y las expresiones faciales [6]. Sin embargo, las señales de voz son de las formas más rápidas y naturales de comunicación humana [SER]. Por ello, se convierten en el candidato ideal para servir de input en una interacción hombre-máquina más veloz y eficiente.

Lo primero es lo primero: abordar el reconocimiento automático de la emoción requiere un modelo apropiado de representación de la emoción. Comenzando por representar la emoción de una manera adecuada, un primer modelo es a través de clases discretas, en el que las emociones son categorías. Por lo general, se admite un grupo básico e innato de categorías emocionales, o patrones de respuesta emocional, esenciales en la supervivencia y evolución de la especie. Darwin (1872/1998) sugirió que las emociones se han desarrollado para servir una función de comunicación en la especie, y por tanto se expresan y reconocen de manera parecida de manera intercultural [BasicEmotions].

El grupo de emociones primordiales son conocidas como las "seis grandes" de Ekman (1992): sorpresa, alegría, ira, miedo, asco y tristeza [BasicEmotions]. A su vez, tenemos el concepto de ‘tonos emocionales’, desarrollado por Manfred Clynes, que se refiere a los siete tonos/energías emocionales que transmitimos los seres humanos a través de una sutil modulación del sistema motor [7]. Plutchik (1980) sugirió ocho emociones (enfado, anticipación, felicidad, confianza, miedo, sorpresa y tristeza) a través de un modelo tridimensional que modela los niveles de intensidad y las relaciones entre las emociones [colors].

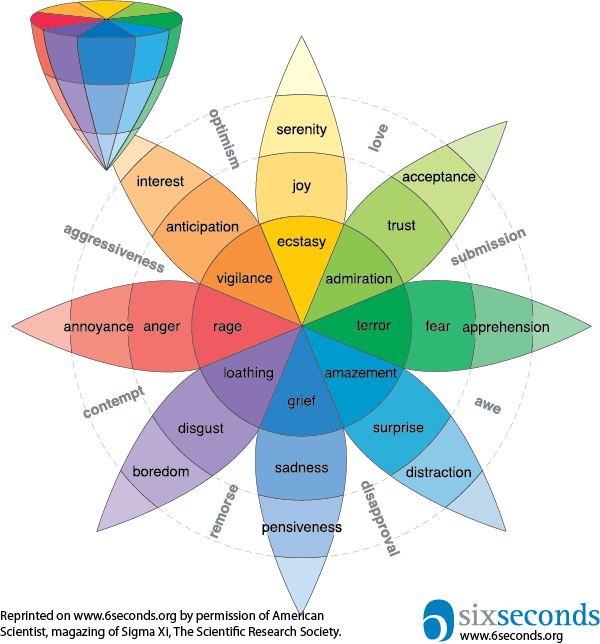


Figura 3 Estrella de emociones de Plutchik (Donaldson, 2017)

Uno de los mayores problemas a los que se enfrenta la disciplina del SER es que las emociones son conductas complejas, y cada persona siente emociones diferentes, y por lo tanto tiene formas distintas de mostrarlas [SERCNN)].  Las emociones se manifiestan y pueden medirse atendiendo a tres niveles o componentes diferentes [Ems]:

* **Cognitivo**: por ejemplo, a través de las expresiones verbales que manifiestan los sentimientos o experiencia subjetiva.
* **Conductual o motor**: por medio de cambios faciales o conductas de aproximación o retirada.
* **Fisiológico**: a través de los cambios viscerales u hormonales, y de los de tipo eléctrico y metabólico que se dan a nivel cerebral y periférico.

El enfoque del SER está compuesto principalmente por las fases conocidas como la extracción de características (Feature Extraction), la selección de estas características (Feature Selection) y la clasificación de las características (Emotion Classification) [Feature1].

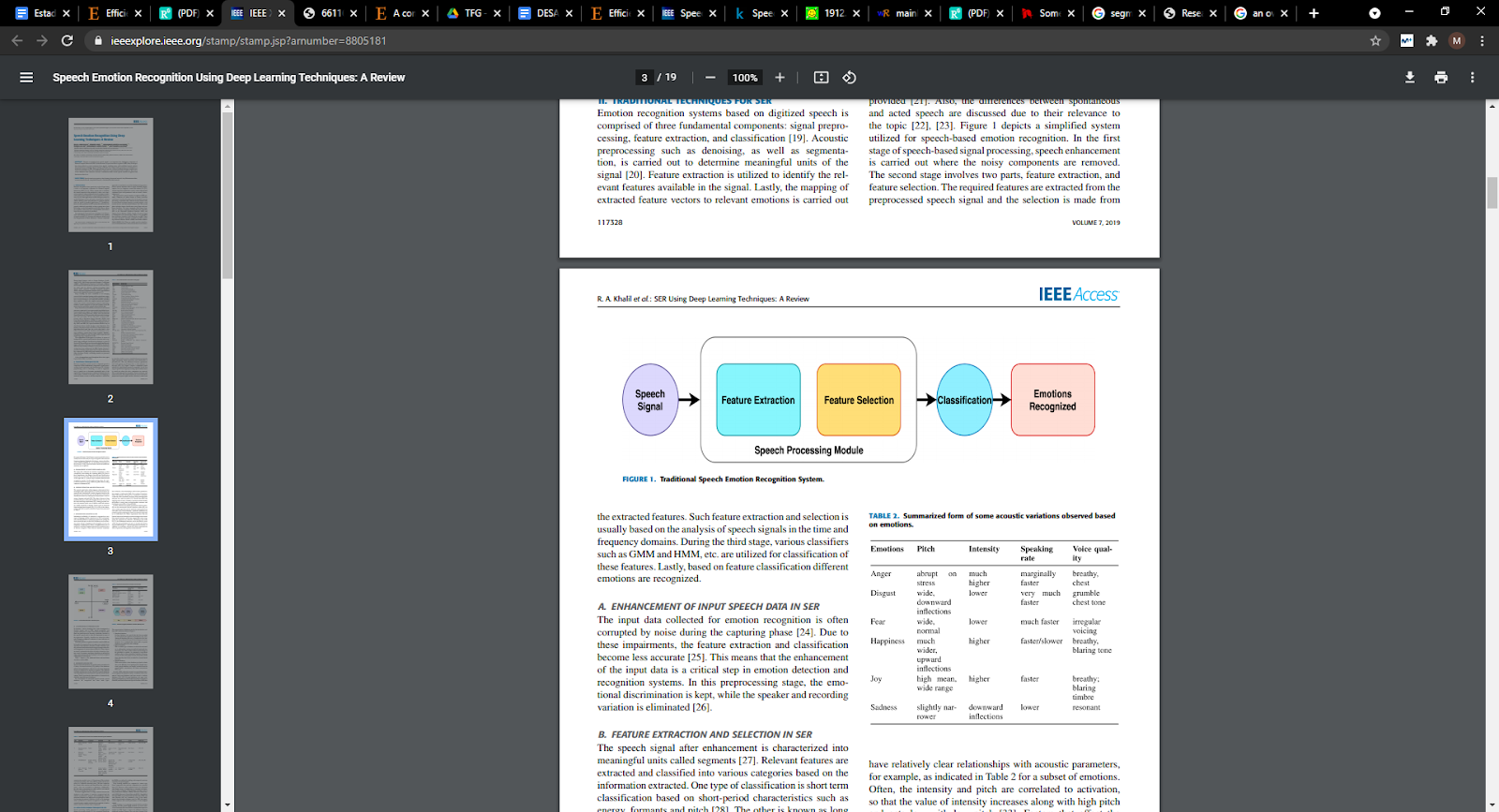


Figura 4 Sistema SER tradicional (Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. , 2019)

*SER: Feature Extraction*

El caso del tono de voz se cataloga como una respuesta fisiológica [6]. Realizar un reconocimiento de las emociones a través de la voz significa extraer estas emociones de las señales de voz de la persona en cuestión. Sin embargo, estas señales son complejas y almacenan información variada; el mensaje, quien habla, el género, el idioma y las emociones. SER es una tarea complicada ya que los tonos de voz son únicos para cada individuo [feature2]. Esta alta variabilidad del tono y la frecuencia puede resultar en que una misma frecuencia en cierto punto sugiera dos o más emociones. Por lo tanto, diferenciar entre las diferentes porciones es una tarea compleja [SER].

Un modelo no puede entender la información recibida en formato de audio, por lo tanto, es crucial convertir estos datos en un formato entendible. Para ello se utiliza la extracción de características, conocido como Feature Extraction. El Feature Extraction pretende extraer características particulares para cada una de estas porciones, denominadas speech utterances [feature3].

Las características del audio se categorizan generalmente en:

**Características temporales**

Estas características son fáciles de extraer, ya que se extraen directamente y por lo tanto requieren de menor poder de cómputo. Estas características proporcionan una manera más sencilla de analizar las señales de audio [SER2]. Estas incluyen:

* **Zero crossing rate**
* **short-term energy**
* **maximum amplitude**
* **minimum energy**
* **entropy energy**

**Características espectrales**

Relacionadas con el tracto vocal, suelen representar la distribución de la energía en la frecuencia del habla. Revelan patrones más profundos de las señales de audio, por lo que son las características que resultan más determinantes para determinar las emociones subyacentes [SER2].

* **MFCC**
* **LPCC**
* **Formants**
* **DFT**
* **Spectral centroid**
* **Linear prediction**
* **Chroma coeficcients**

**Características prosódicas**

En muchos idiomas, son las responsables de transmitir información semántica al oyente, y forman las bases del comportamiento lingüístico. Por lo tanto, son útiles para detectar las emociones del locutor [4]. En varias investigaciones, estas características se han empleado para tareas de reconocimiento de emociones [SER4] [SER5].

* **noise**
* **stretch**
* **shift**
* **pitch**
* **higher\_speed**
* **lower\_speed**

Los vectores de características se dividen entre segmentales y suprasegmentales, acorde a su estructura temporal. Las características prosódicas se califican del segundo tipo, y estas no segmentan la señal, sino que se aplican sobre la extensión entera. Al contrario que las características segmentales (temporales y espectrales) a las que se les aplica técnicas de windowing y se calculan una vez cada cierto segmento de tiempo (unos 20-50 mseg) [feature4].

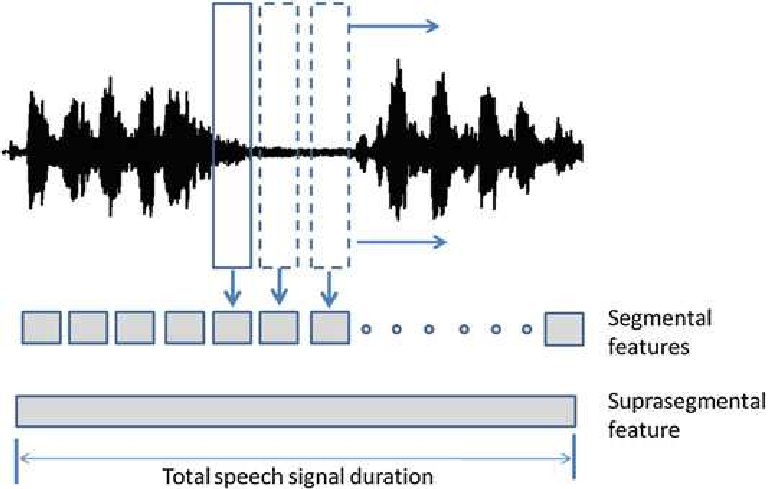


Figura 5 Características Segmentales y Suprasegmentales (Anagnostopoulos, Iliou & Giannoukos, 2012)

En estudios más tempranos, el foco se ponía principalmente en las características prosódicas, mientras que en investigación más reciente el uso de características espectrales ha crecido exponencialmente [feature4]. Escoger qué features son los más útiles para el entrenamiento puede ser una tarea muy confusa [SER]. Es importante que la selección de características no se vea afectada por la cultura, la región o el acento del locutor [feature2]. Dar con la combinación de características acústicas más robusta para el reconocimiento automático de la emoción de un locutor es uno de los pilares del SER [feature2].

MFCC

Los Coeficientes Cepstrales en la Escala de Mel (MFCC), alias ‘Most Frequently Considered Coeficcients’, representan la amplitud del espectro del habla de manera compacta, esto los convierte en la técnica de extracción de características más usada en reconocimiento del habla. Cualquier sonido generado por el ser humano está determinado por la forma de su tracto vocal, por lo tanto, representar este sonido va a depender de poder determinar la forma correctamente [MFCC]. Además, para lograr la mayor eficiencia del sistema de extracción de características, lo ideal sería imitar el comportamiento frecuencial del oído humano. De aquí surge el concepto de los coeficientes MFCC, basado en la escala de frecuencia MEL para imitar el comportamiento de tonos con distinta frecuencia dentro del oído humano.

*SER: Feature Classification*

Para la tarea de clasificación de características se emplean clasificadores lineales y no lineales. Los clasificadores lineales más usados incluyen las Bayesian Networks (BN) o el Maximum Likelihood Principle (MLP) y el Support Vector Machine (SVM). Sin embargo, la señal de voz se considera no estacionaria, por lo que un clasificador no lineal es óptimo para tareas SER [dibujos]. Existen muchos clasificadores no lineales, los más usados en la detección de emociones por medio de voz son el Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) [16].

Las investigaciones más recientes se han desarrollado en el terreno de los sistemas de aprendizaje profundo. Estos sistemas infieren una representación jerárquica de los datos de entrada, lo que facilita su categorización [SERCNN2]. Estos sistemas forman parte de la rama de estudio conocida como el Deep Learning, cuya investigación y aplicación ha experimentado un fuerte auge durante los últimos años [DL]. Los modelos más utilizados en el campo del SER con los Convolutional Neural Networks (CNN) y las arquitecturas recurrentes como las Recurrent Neural Networks (RNN) y el Long Short-Term Memory (LSTM), que han resultado ser mucho más efectivos en las tareas de reconocimiento de voz como el Natural Language Processing (NLP) y el SER [SERCNN3].

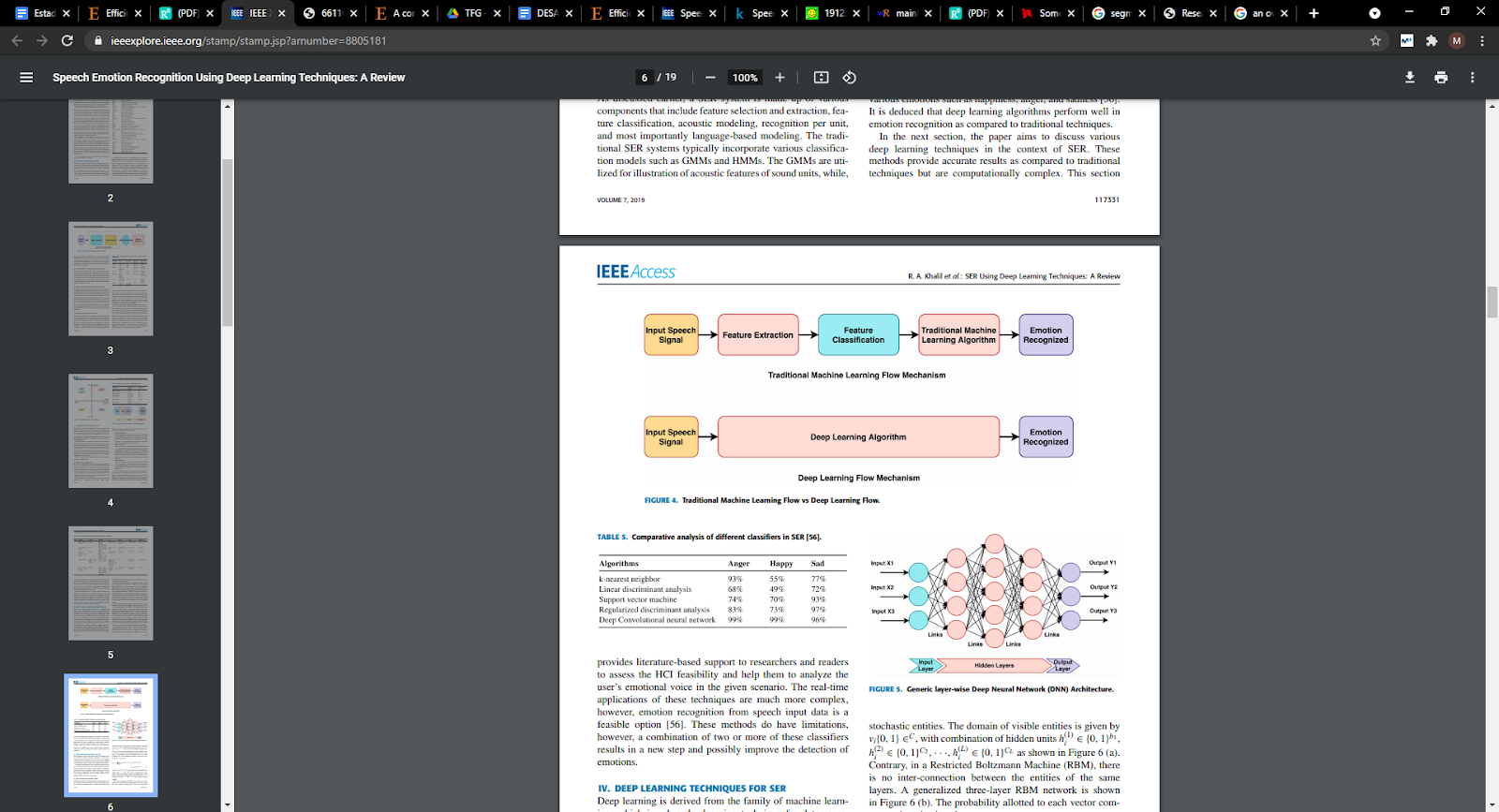
Las técnicas de Deep Learning tienen ciertas ventajas sobre los métodos tradicionales. Una de las más destacadas es su habilidad de detectar estructuras complejas de las características (features) sin la necesidad de hacer una extracción manual [dibujos].

Figura 6 Método tradicional de Machine Learning vs. Deep Learning (Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. , 2019)

Los estudios más recientes que han implementado estos modelos son aquellos dirigidos a la implementación de sistemas de detección automatizada del Covid-19. Md. Zabirul Islam et al. (2020) aplicaron una técnica de aprendizaje profundo basada en la combinación de una red neuronal convolucional (CNN) y una memoria a largo-corto plazo (LSTM) para diagnosticar COVID-19 automáticamente a partir de imágenes de rayos X. En este sistema, CNN se usa para la extracción profunda de características y LSTM se usa para la detección usando la característica extraída [covid].

*Aplicaciones representativas*

**Empath** [empath]

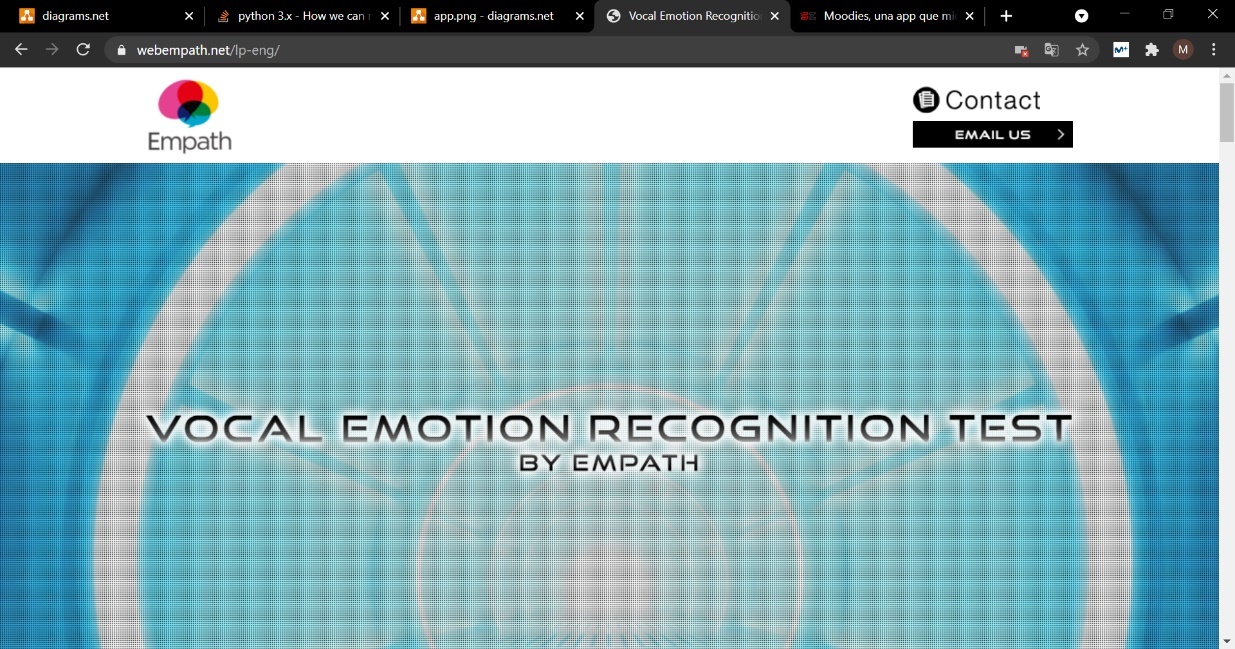
Programa desarrollado por Smartmedical Corp. Su algoritmo original identifica la emoción del usuario mediante el análisis de las propiedades físicas de su voz. Basado en decenas de miles de muestras de voz, detecta su ira, alegría, tristeza, calma y vigor. Sus servicios son ofrecidos a través de una página web propia (figura 7).

Figura 7 Página de entrada al programa Empath

**Vmote** [vmote]

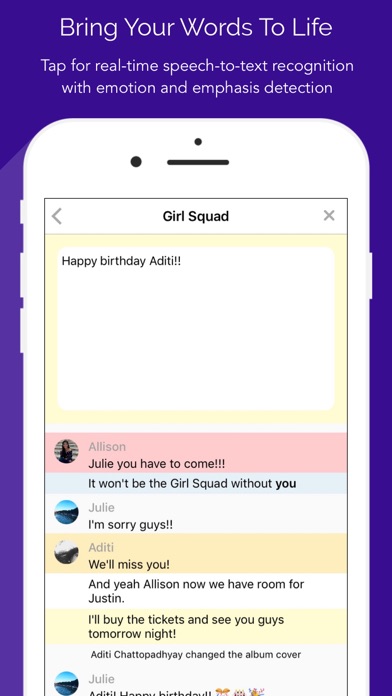
Vmote es una aplicación de mensajería cuya interfaz de usuario esta diseñada en torno a la voz.  A través del Speech to Text, te permite enviar mensajes más rápido. Además, como los mensajes a menudo pueden malinterpretarse, Vmote también incorpora reconocimiento basado en emociones y detección de énfasis para proporcionar más contexto a los mensajes de voz.

Figura 8 Plantilla de la UI de la app Vmote

**Interview Simulator** [intsim]

Plataforma de reconocimiento de emociones multimodal para analizar las emociones de los candidatos al empleo, en colaboración con la Agencia Francesa de Empleo. Se analizan las emociones faciales, vocales y textuales, utilizando principalmente enfoques basados ​​en el aprendizaje profundo. Consiste en una aplicación web desplegada en Flask.

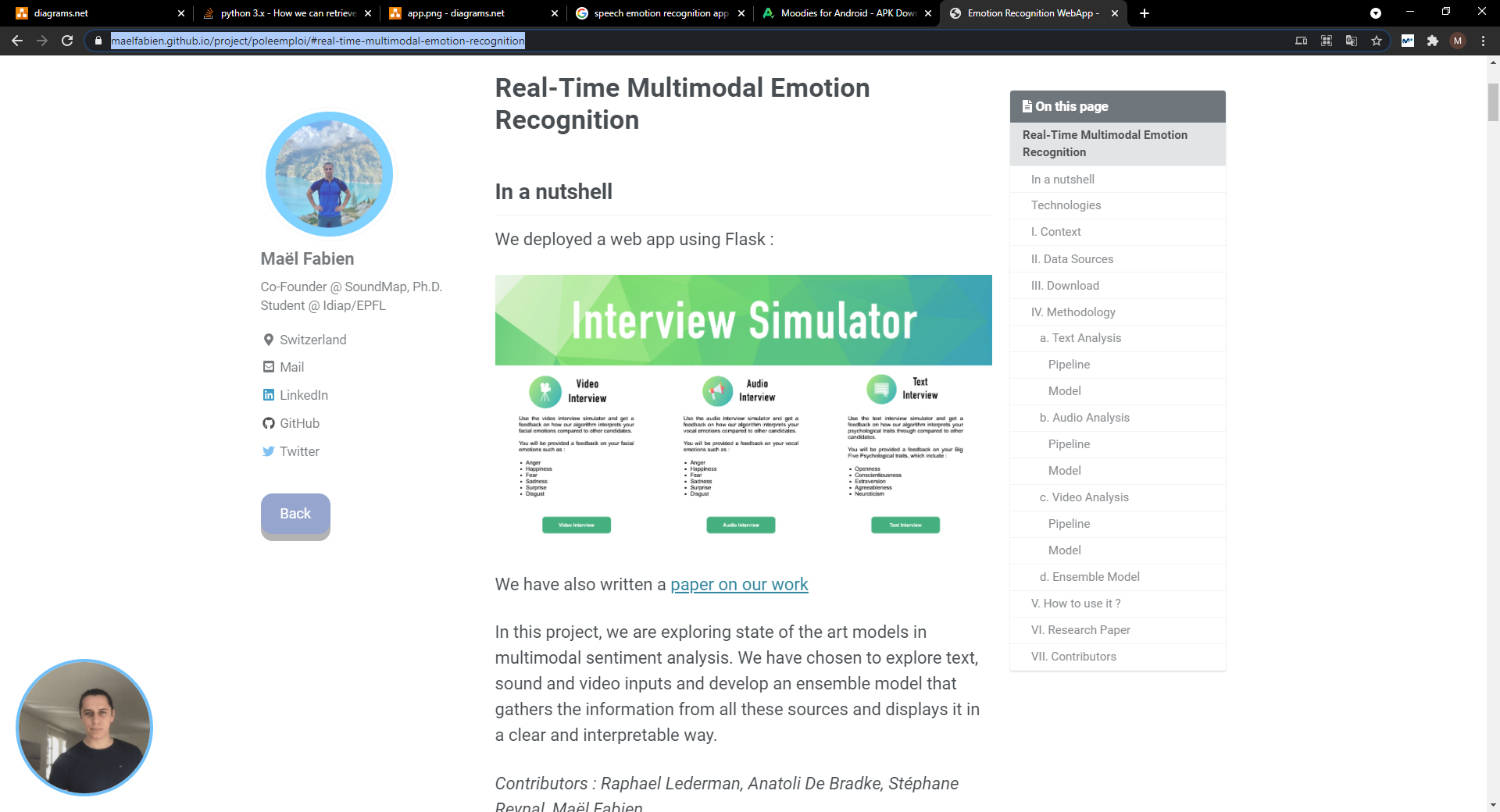


Figura 9 Página principal de la aplicación web Interview Simulator

## Contexto y justificación

La medicina hipocrática defiende la necesidad de un estado de equilibrio y armonía entre el cuerpo y la mente. El padre y pionero de esta perspectiva de la medicina es Hipócrates (460-357 a.C.), una eminencia de la medicina de la Antigua Grecia [1]. Este mismo fue quien afirmó que para prevenir un ataque de asma, “el asmático debe protegerse de su propia ira” [1]. Hoy en día, más de 2 mil años después, se persigue una filosofía similar: “La salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades” [2], la definición que proporcionó la Organización Mundial de la Salud en 1948. De ahí la aparición de la medicina holística, que defiende que el cuerpo, la mente y el ambiente contribuyen de manera equitativa a la salud [3].

¿Puede la mente curar el cuerpo? La corriente de pensamiento budista ha fomentado sus creencias sobre las capacidades curativas de la mente desde hace más de dos mil años [buddist], a lo que se han ido sumando cada vez más los científicos occidentales. Prueba de ello son numerosos encuentros entre el Dalai Lama y prominentes psicólogos, médicos y profesores de meditación, que han arrojado nueva luz sobre esta conexión cuerpo-mente [Dalai lama].

Ya existen numerosos estudios que analizan el estado emocional y afectivo como factor determinante para el estado de salud de las personas. Cohen demostró que el estrés incrementa el riesgo de infecciones respiratorias agudas [4]. Con la aparición de la pandemia del Covid-19, es posible que un cuidado de la salud mental pueda fortalecer las defensas y ser un aliado a la hora de combatir esta enfermedad. Ejemplo de ello es un estudio que comprobó que mantener una actitud positiva durante momentos de elevado estrés promovía los niveles de S-Iga (inmunoglobulina-A salival), reconocida como una sustancia que protege contra las enfermedades respiratorias [5].

Si las emociones juegan un papel tan importante en la salud, fomentar un correcto estado de salud mental podría mejorar la salud de muchos ciudadanos. Especialmente ahora que vivimos en tiempos sin precedentes; la pandemia del Covid-19 ha impactado cada aspecto de nuestras vidas y nos ha llevado a una nueva realidad. Una realidad que está afectando a la salud tanto física como mental; es un buen momento para poder trabajar con ambas entrelazadas para asegurar el bienestar de aquellos más afectados. La monitorización masiva de la salud de la población es crucial, para así poder garantizar el bienestar de los mismos ciudadanos, como a su vez, ayudar a descongestionar los servicios sanitarios.

A su vez, vivimos en un mundo en el que se puede explotar mucha riqueza de las nuevas tecnologías. El auge de estas últimas amplía el horizonte de oportunidades y nos permite abordar problemas a través de soluciones menos tediosas, más rápidas y eficaces. Entonces, dado lo mucho que las máquinas se han integrado en nuestras vidas, ¿por qué no darles la oportunidad de que nos conozcan mejor? A muchos se les pueden venir a la cabeza historias terroríficas de ciencia ficción, resultado de un uso descontrolado. Sin embargo, sembrando las bases éticas, y con el control adecuado, el uso de las avanzadas tecnologías puede traer múltiples beneficios y mejoras a nuestro día a día. De la cuestión anterior, nace otra, ¿pueden

las máquinas detectar nuestras emociones? La respuesta a estas dos cuestiones se plantea en el siguiente trabajo.

Watson de IBM es uno de los pioneros en computación cognitiva orientada a la salud, con una de sus mayores aplicaciones en el campo de la oncología. La empresa Babylon lanzó en 2016 una aplicación de consultas médicas online. Es a través de un sistema de reconocimiento de voz y de inspección de historiales clínicos que ofrece una ruta de acción adecuada a la circunstancia médica en cuestión [14]. En el mercado también está Google Health, una plataforma de Google que promueve el descubrimiento de nuevas oportunidades en el campo de la AI para mejorar la eficacia de las tecnologías sanitarias a nivel global [15].

## Planteamiento del problema

La crisis sanitaria producida por el Covid-19 ha sembrado sentimientos de miedo, preocupación y estrés en la población. Ante esta nueva y desafiante realidad es importante que cuidemos tanto nuestra salud física como mental. La vida cotidiana ha cambiado, el distanciamiento físico requerido para mantener un estilo de vida seguro ha desembocado en trabajar o estudiar desde casa, el desempleo temporal y la falta de contacto con seres queridos y amigos [OPS]. La salud tanto física como mental se está viendo afectada, por lo que es un buen momento para poder trabajar con ambas entrelazadas para asegurar el bienestar de aquellos más afectados.

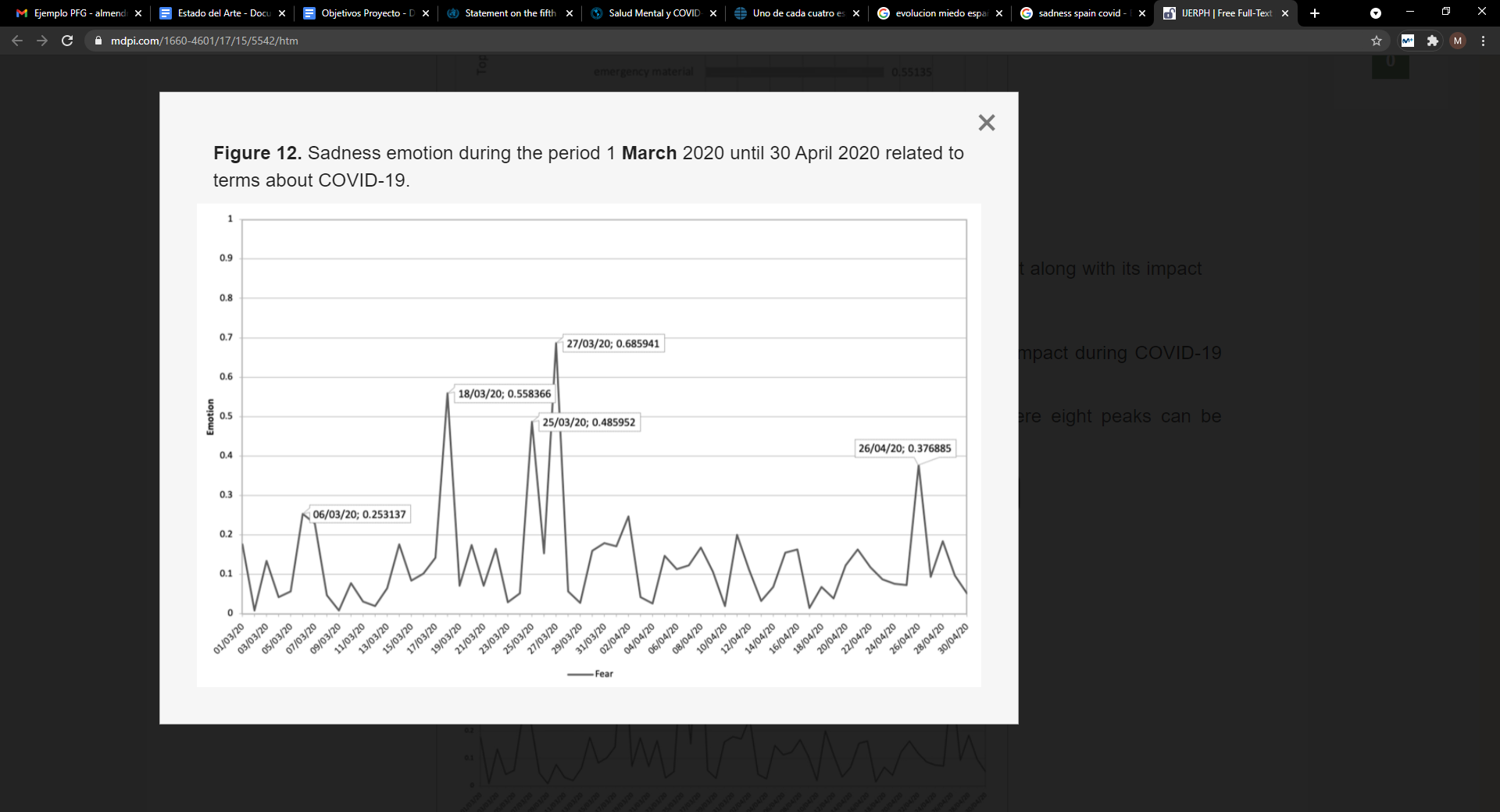
Un estudio sobre el análisis de sentimientos en España durante la pandemia del Covid-19 refleja el gran impacto que esta ha causado en el estado emocional de muchos españoles. Esta crisis es un reto para los gobiernos e instituciones sanitarias, a medida que los sentimientos de miedo, estrés y disgusto se propagan [16].

Figura 10 Evolución del sentimiento del miedo en España durante el periodo del 1 del marzo 2020 hasta el 30 de abril 2020 (de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI., 2020)

Se distinguen varios picos en los datos que coinciden con fechas específicas de ciertos titulares en los medios de comunicación. La figura 11 detalla la difusión de información en ciertos días del confinamiento que coinciden con los picos observados en la Figura 10.

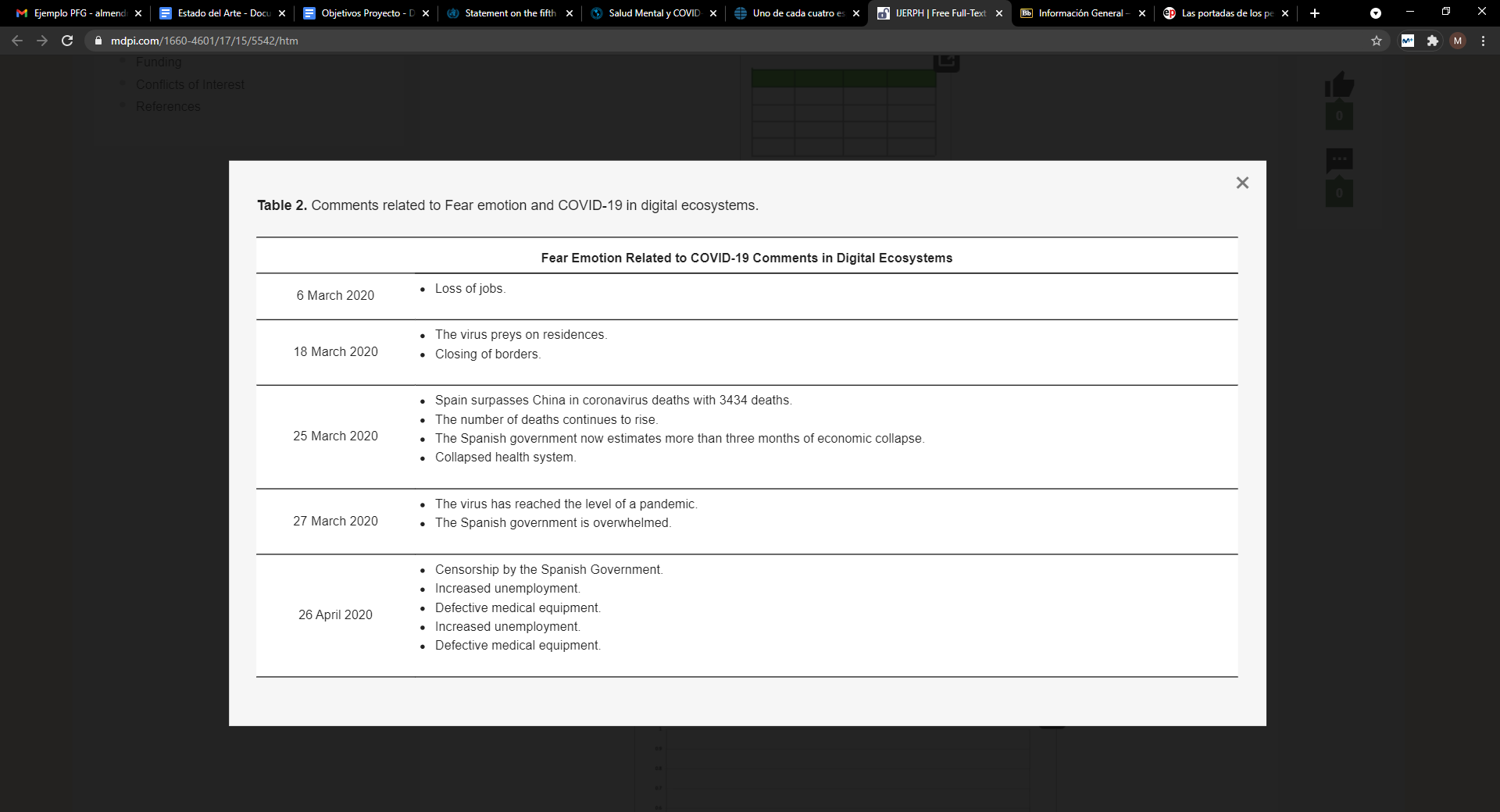
La OMS hizo un llamamiento en la Declaración sobre la reunión realizada el 29 de octubre de 2020 a la necesidad de fortalecer los servicios de salud mental ante la crisis sanitaria del Covid-19 [OMSComite].

Figura 11 Noticias relacionadas con el miedo y el Covid-19 en los medios de comunicación (de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI., 2020)

Mucha gente tiene que luchar sola contra estas situaciones emocionalmente complejas, lo cual puede suponer un empeoramiento de la salud mental y física. La población de avanzada edad es especialmente vulnerable; para muchos tener que asistir a la consulta de un médico de manera física supone un alto riesgo, muchas veces vital. Por lo tanto, se considera importante poder ofrecer alternativas al contacto físico que puedan proporcionar seguridad en momentos de mayor malestar emocional. Este proyecto trata de conseguir una solución que pueda aliviar a los sistemas sanitarios, y al mismo tiempo incrementar el estado afectivo de los usuarios, que podrán mantener un estado físico y mental saludables desde la comodidad del hogar.

Sin embargo, el uso de la aplicación aquí propuesta puede ser mas casual. A veces surge la curiosidad de que emociones desprendemos cuando hablamos con otras personas. Poder detectar y monitorizar patrones en la voz puede ayudar a entender y gestionar la forma en la que nos comunicamos en el día a día. El seguimiento de estas emociones puede, a su vez, fortalecer la conciencia propia.

El principal problema que afronta el AC según Picard es que los perfiles de cada persona son muy particulares, lo que convierte la tarea predecir los estados afectivos en una con una alta variabilidad [6]. lo tanto, este estudio plantea el uso de características extraídas de la voz para mejorar la comprensión del estado afectivo de los usuarios. La voz se considera un input adecuado ya que toma diferentes formas para distintas emociones [6].

Una tarea indispensable es que la aplicación aprenda de manera autónoma. En este caso, se busca la estrategia de aprendizaje óptima a través de un algoritmo de clasificación que de los mejores resultados en las fases de entrenamiento y predicción y empleando previamente las mejores prácticas de preparado de datos.

# OBJETIVOS

## Objetivos generales

El objetivo del presente trabajo es desarrollar una aplicación multi plataforma que sea capaz de detectar el estado emocional del usuario a través de un sistema de reconocimiento de voz y pueda devolver esta predicción al usuario en tiempo real.

## Objetivos específicos

La arquitectura propuesta estará formada por un componente back-end y un front-end.

La analítica de datos tomará un papel importante en este proyecto. Para ello, se diseña un proceso ETL sobre el dataset previo a la implementación del algoritmo. Este proceso se desglosa en las siguientes transformaciones:

1. Extract:

Encontrar y extraer los conjuntos de datos adecuados para el entrenamiento del algoritmo. Para la búsqueda de fuentes de datos adecuadas se va a utilizar la plataforma Kaggle [25], donde se encuentra una de las comunidades de científicos de datos más destacadas a nivel mundial.

1. Transform:
   * Solucionar los problemas de calidad de los datos en el caso de que los hubiese
   * Tratado de datos para su correcta ingesta por el algoritmo.
   * En el caso de tareas de clasificación del habla, se requiere una extracción de características (*feature extraction*) adecuada. Por lo tanto, requiere estudiar las opciones de extracción de características.
   * Seleccionar las características que optimizan el aprendizaje del algoritmo.
2. Load:

Cargar los datos finales transformados en un lugar centralizado. Una vez estén los datos preparados para su uso, podrán ser empleados para entrenar el algoritmo.

A continuación, se especifica la parte técnica de inteligencia artificial:

1. Estudiar las opciones de algoritmos de sistemas de reconocimiento de voz.
2. Diseñar el sistema de reconocimiento de voz: la idea principal consistirá en un algoritmo de clasificación, en concreto una red neuronal artificial.
3. Decidir cuáles serán las métricas más significativas para escoger el modelo óptimo.
4. Análisis de los resultados obtenidos para cada iteración; realizar los ajustes necesarios en los parámetros, con el fin de escoger el mejor modelo.
5. Escoger el algoritmo con mejores resultados, capaz de predecir la emoción en datos nuevos.

La importancia de que el modelo muestre destreza a la hora de completar e interpretar estas tareas correctamente va a ser determinante a la hora de proporcionar las conclusiones adecuadas. Una vez el algoritmo haya sido entrenado, se pasa a la fase de desarrollo de la aplicación móvil.

Una vez el dispositivo haya realizado la tarea de clasificación, debe ser capaz de responder al usuario.

El objetivo es conseguir que el usuario se sienta mejor, para ello, los principales sistemas integrados serán una serie de habilidades:

1. Captar los datos necesarios a través del micrófono
2. Preprocesamiento de los datos; repetir transformaciones especificadas anteriormente
3. Analizar y dar respuesta a los datos

Para poder dar respuesta a los usuarios, se debe implementar una interfaz de usuario adecuada. Este proyecto propone una aplicación multi plataforma que pueda ser accesible desde cualquier smartphone. Los objetivos específicos que se persiguen para este bloque son:

1. Estudiar alternativas para el desarrollo de una aplicación web y móvil (multi plataforma). Con un sistema de registro e inicio de sesión personal incorporado y que sea capaz de reconocer las emociones en la voz de un usuario y devolver un resultado en tiempo real
2. Diseño de la interfaz de usuario
3. Implementar una arquitectura de cloud computing junto con un framework de programación de aplicaciones móviles para montar una aplicación con alojamiento web y móvil en la nube
4. Conectar todos los componentes implementando una BBDD adecuada para centralizar los datos que se intercambiaran entre los componentes back-end y front-end

## Beneficios del proyecto

Esta aplicación puede ser un beneficio muy grande para aquellos que se encuentren en una situación emocionalmente compleja, y deseen llevar una vida más autónoma, sin necesidad de desplazamientos. Con este sistema el usuario podrá mantener una monitorización del estado de sus emociones e impulsar una rutina hacia una vida más saludable.

La herramienta fundamental para poder usar el producto será el smartphone, de uso generalizado, por lo que es un producto altamente accesible. Permite poder usar la aplicación donde sea y sin la necesidad de tener acceso a un ordenador. No existe ninguna aplicación en el mercado con la funcionalidad de detectar las emociones a través de la voz del usuario, por lo que se podría satisfacer una necesidad a que aún no lo está.

Cabe destacar las grandes ventajas de emplear el cloud computing para llevar a cabo el hosting de este proyecto. De tal manera, la aplicación puede correr en remoto, sin la necesidad de que haya un dispositivo hardware pendiente de dar respuesta a las peticiones. Estas ultimas pueden llegar en grandes cantidades, sin perder la rapidez y eficacia del sistema de respuesta. A su vez, resulta en un ahorro en costes hardware, tanto a nivel de componentes como de personal necesario para mantenerlo.

# DESARROLLO DEL PROYECTO

## Planificación del proyecto

**Fase 1. Investigación previa**

1. Estudio del del contexto y el estado del arte de los conceptos y tecnologías del proyecto. Esto incluye la búsqueda de las alternativas y casos de uso actuales.
2. Análisis técnico de las tecnologías escogidas, tanto para el componente back-end y el front-end.

**Fase 2. Planificación y diseño**

1. Definir los objetivos del proyecto en base al planteamiento del problema detectado. Determinar el alcance de los objetivos y la distribución de tiempo correspondiente.
2. Definición de requisitos y recursos necesarios. Estos pueden ser requisitos tanto software como hardware.
3. Diseño de la arquitectura back- end del sistema. Esto consiste en canal de preparado de datos, extracción de características y en el algoritmo de reconocimiento de emociones y sus predicciones.
4. Diseño de la arquitectura front-end del sistema. Esta será una aplicación móvil capaz de captar la voz del usuario, enviarla al servidor donde se aloje el algoritmo, y recibir una predicción en tiempo real.
5. Diseño del proceso de validación y pruebas del sistema

**Fase 3. Preparación del entorno**

1. Reunir los requisitos necesarios para preparar el entorno de trabajo. Aplicaciones, programas y elementos como SDKs necesarios para el correcto funcionamiento en la maquina local.
2. Empezar a manejar aquellas tecnologías que son nuevas y no se hayan usado previamente al desarrollo del proyecto. Para ello se emplean videos y cursos online relacionados con estas competencias.

**Fase 4. Desarrollo e implantación de la arquitectura**

ETL:

1. Buscar los datasets adecuados para el entrenamiento. Los datos deben tener un formato parecido o igual y con un labelling adecuado.
2. Desarrollar un pipeline adecuado de ingesta y preparado/limpieza de datos. Incluye determinar las librerías necesarias para ello.
3. Llevar a cabo un análisis exploratorio previo de los datos. Se pretende encontrar patrones o indicios de fenómenos como en oversampling o el undersampling.

Desarrollo del clasificador SER:

1. Estudiar las opciones de algoritmos para la tarea de clasificación de emociones. Plantear una serie de pruebas con aquellos considerados más adecuados.
2. Determinar la estrategia de extracción de características de los datos y prepararlos para poder ser ingeridos por el algoritmo elegido. Para ello, se investigarán tareas de escalado de datos y separación entre entrenamiento, validación y test.
3. Iniciar la fase de entrenamiento. Probar con diferentes valores para los hiperparámetros de los modelos y fijar las métricas que se utilizarán para la evaluación.
4. Comparativa de modelos y escoger en base a la eficacia de predicción.

Interfaz de usuario:

1. Estudiar alternativas para el desarrollo de una aplicación web y móvil (multi plataforma). Con un sistema de registro e inicio de sesión personal incorporado y que sea capaz de reconocer las emociones en la voz de un usuario y devolver un resultado en tiempo real.
2. Diseño de la interfaz de usuario.
3. Implementar una arquitectura de cloud computing junto con un framework de programación de aplicaciones móviles para montar una aplicación con alojamiento web y móvil en la nube.
4. Conectar todos los componentes implementando una BBDD adecuada para centralizar los datos que se intercambiaran entre los componentes back-end y front-end.

**Fase 5. Pruebas, conclusiones y análisis de rendimiento**

1. Periodo de pruebas, refuerzo e identificación de disfunciones. Se deberán aplicar los parches y las medidas necesarias para arreglar posibles fallos de funcionamiento.
2. Desarrollar un modelo para la salida a producción de la aplicación. Para ello es necesario un análisis de cómo garantizar la seguridad y confidencialidad de los usuarios a través de protocolos de seguridad de datos (RGPD, LSSI).
3. Redactar la memoria a partir de todos los documentos empleados y redactados durante cada fase de desarrollo.
4. Defensa del TFG.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nombre de la tarea | Mes | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Enero | | | | Febrero | | | | | Marzo | | | | Abril | | | | | Mayo | | | | | Junio | | | | | N.º horas |
| **Investigación previa** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Contexto y estado del arte |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 20 |
| Análisis técnico de las tecnologías empleadas |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| **Planificación y diseño** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Definir objetivos y el alcance |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 30 |
| Definir requisitos y recursos |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Diseño de la arquitectura back-end |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Diseño de la arquitectura front-end |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Diseño del proceso de pruebas |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| **Preparación del entorno** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Instalar y configurar los requisitos computacionales necesarios |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 90 |
| Aprendizaje de nuevas tecnologías |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| **Desarrollo e implantación** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ETL | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Búsqueda de datasets para el entrenamiento |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 50 |
| Desarrollo del pipeline de ingesta y preparado/limpieza de datos |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Análisis exploratorio de los datos limpios |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Desarrollo del Clasificador SER | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Estudiar algoritmos de clasificación |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 150 |
| Extracción de características |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Entrenamiento del modelo y ajuste de hiperparémetros |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Comparativa y elección del mejor modelo |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Interfaz de usuario | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Estudiar opciones para el desarrollo de una aplicación multi plataforma |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 130 |
| Diseño de la interfaz de usuario |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG. Revisión de los algoritmos. |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Montar arquitectura de computación en la nube |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Conectar todos los componentes y administrar la BBDD |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Reunión de seguimiento con la tutora del TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| **Pruebas, conclusiones y análisis de rendimiento** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Pruebas y ajustes de funcionamiento |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | 100 |
| Desarrollar modelo de puesta en producción y políticas de seguridad de datos |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Redacción de la memoria final |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Entrega primera versión completa de la memoria de TFG a la tutora |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |  |
| Realizar cambios sugeridos en el feedback de la tutora de TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| Entrega de la versión final de la memoria de TFG |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |
| **DEFENSA DEL TFG** |  |  |  |  | |  |  |  |  |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |

## Descripción de la solución, metodologías y herramientas empleadas

El proyecto está compuesto por dos grandes componentes. Primero se desarrolla un sistema de aprendizaje automático motorizado por un pipeline de procesado de datos, extracción de características e implementación de un modelo entrenado y capaz de realizar predicciones a partir de nuevas fuentes de datos. Esta primera fase se muestra en la figura 12.

Figura 12 Pipeline principal detrás del sistema de reconocimiento de voz

### Bases de datos

El entrenamiento del algoritmo de clasificación de emociones es el núcleo del presente trabajo. Por lo tanto, encontrar un conjunto de datos adecuado para su entrenamiento era una tarea esencial. Se decide combinar dos datasets diferentes para poder tener un conjunto de entrenamiento grande y variado, pero evitando mezclar datos de diferentes formatos ya que puede dar lugar a incongruencias en el entrenamiento. A su vez, ambos datasets contienen voces categorizadas como masculinas como femeninas, lo cual permite desarrollar algoritmos especializados para cada género, aparte de uno generalizado. Los datasets elegidos son los siguientes:

* **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song) (Livingston & Russo, 2018), datos publicados bajo una licencia de atribución de Creative Commons de uso no comercial [RAVDESS].

La RAVDESS contiene 7356 archivos, en los que participan 24 actores profesionales (12 mujeres, 12 hombres) con un acento norteamericano neutral. El habla incluye expresiones de calma, felicidad, tristeza, ira, miedo, sorpresa y disgusto, y se vocaliza una de dos frases. La canción contiene emociones tranquilas, felices, tristes, enfadadas y temerosas. Cada expresión tiene dos niveles de intensidad emocional (normal, fuerte), con una expresión neutra adicional.

Para este proyecto solo se van a usar archivos de audio y de tipo voz. El archivo de voz (Audio\_Speech\_Actors\_01-24.zip, 215 MB) contiene 1440 archivos: 60 ensayos por actor x 24 actores = 1440. La tabla 1 detalla las características de los archivos.

|  |  |  |
| --- | --- | --- |
| Nº | Característica | Opciones |
| 1 | Modalidad | 01 = AV completo, 02 = solo video, 03 = solo audio |
| 2 | Canal vocal | 01 = habla, 02 = canción |
| 3 | Emoción | 01 = neutral, 02 = calma, 03 = feliz, 04 = triste, 05 = enojado, 06 = temeroso, 07 = disgusto, 08 = sorprendido |
| 4 | Intensidad emocional | 01 = normal, 02 = fuerte (no hay intensidad fuerte para la emoción neutral) |
| 5 | Frase | 01 = “Kids are talking by the door”, 02 = “Dogs are sitting by the door” |
| 6 | Repetición | 01 = 1ª repetición, 02 = 2ª repetición |
| 7 | Actor | 01 a 24. Los actores impares son hombres, los actores pares son mujeres |

Tabla 1 Descripción de la identificación de archivos de audio de la base de datos RAVDESS

Ejemplo de nombre de archivo: 02-01-06-01-02-01-12.mp4

* Solo video (02)
* Discurso (01)
* Temeroso (06)
* Intensidad normal (01)
* Declaración "perros" (02)
* 1a repetición (01)
* 12 ° actor (12)
* Mujer, ya que el número de identificación del actor es par.
* **CREMA-D** (Crowd-sourced Emotional Multimodal Actors Dataset) es un conjunto de datos de 7,442 archivos de 91 actores. Las voces las componen 48 actores masculinos y 43 femeninos entre las edades de 20 y 74, provenientes de una variedad de razas y etnias (afroamericanos, asiáticos, caucásicos, hispanos y no especificados).

Los actores hablan una selección de 12 frases expresadas en una de seis emociones diferentes (ira, disgusto, miedo, feliz, neutral y triste) y cuatro niveles de emoción diferentes (bajo, medio, alto y no especificado).

|  |  |  |
| --- | --- | --- |
| Nº | Característica | Opciones |
| 1 | ID del actor | Número de 4 dígitos. Los IDs 1002-1013,1018,1020,1021,1024,1025,1028-1030,1037,1043,1046,1047,1049, 1052-1056,1058,1060,1061,1063,1072-1076,1078,1079,1082,1084,1089,1091 corresponden al sexo femenino, y el resto al masculino. |
| 2 | Frase | “It's eleven o'clock” (IEO).  “That is exactly what happened” (TIE).  “I'm on my way to the meeting” (IOM).  “I wonder what this is about” (IWW).  “The airplane is almost full” (TAI).  “Maybe tomorrow it will be cold” (MTI).  “I would like a new alarm clock” (IWL)  “I think I have a doctor's appointment” (ITH).  “Don't forget a jacket” (DFA).  “I think I've seen this before” (ITS).  “The surface is slick” (TSI).  “We'll stop in a couple of minutes” (WSI). |
| 3 | Emoción | Enfado (ANG)  Disgusto (DIS)  Miedo (FEA)  Feliz (HAP)  Neutro (NEU)  Triste (SAD) |
| 4 | Nivel de emoción | Bajo (LO)  Medio (MD)  Alto (HI)  Sin especificar (XX) |

Tabla 2 Descripción de la identificación de archivos de audio de la base de datos CREMA-D

Ejemplo de nombre de archivo: 1001\_DFA\_ANG\_XX.wav

* ID del actor (1001), actor masculino
* Frase “Don’t forget a jacket” (DFA)
* Enojado (XX)
* Nivel de emoción no especificado (XX)

### Ingesta, limpieza y análisis exploratorio de datos

Para los procesos de tratado de datos y entrenamiento se implementa una arquitectura en el entorno de trabajo Jupyter Notebook. Una vez importados todos los datos, se organizan de tal manera que se tenga un valor y una etiqueta (emoción) para cada valor, para que el clasificador pueda aprender y diferenciar cada emoción.

Una vez se haya realizado el labeling adecuado para cada tanda de datos, al estar trabajando con dos datasets diferentes, se crea un dataframe de *pandas*, donde se almacenan todos los datos con sus respetivas etiquetas. Este dataframe es el que se va a emplear para las próximas fases de preprocesamiento de datos y de extracción de características.

Un modelo predictivo de machine learning aprende de los datos que ingiere, y las tareas de automatización van a ser guiadas por la calidad de los datos y el nivel de representatividad del grupo de entrenamiento [cleandata2]. La limpieza de datos es un proceso fundamental en toda tarea de machine learning. Permite eliminar errores y datos redundantes para lograr un dataset más fiable [cleandata] lo cual va a mejorar la calidad de entrenamiento del clasificador. Sin embargo, los datasets escogidos ya estan en muy buen estado, al tratarse de fuentes de datos de alta difusión, por lo que no es necesario ningún tipo de transformación para eliminar, modificar o sustituir datos.

Sin embargo, es detecta un problema de desbalance de clases. Es un desafío para las tareas de modelado predictivo trabajar con una distribución de clases severamente sesgada, causando desigualdad en los costos de la clasificación errónea [5]. De tal manera, el algoritmo se obstruye prediciendo las clases mayoritarias y no aprenda sobre las más pequeñas. Una solución es eliminar estas categorías de datos, pero en este proyecto se consideran las ocho emociones igual de importantes, y un aspecto considerado importante es añadir variabilidad al modelo, dotándole de la capacidad de predecir un amplio abanico de emociones, el cual se pretende agrandar en futuras líneas de trabajo. Para problemas en los que estas categorías más escasas son importantes, se aplican técnicas de *oversampling* o *undersampling* [imbclass2]. La primera consiste en generar más datos de las clases minoritarias y la segunda en reducir la cantidad de datos de las clases mayoritarias. En este caso, se aplica *oversampling*, ya que se considera importante trabajar con cantidades de datos elevadas, así el algoritmo tiene más información para el aprendizaje, razón por la que además de hacer un *oversampling* de las clases de ‘calm’ y ‘surprise’, también se hace del resto de clases. De tal manera, se obtiene un dataset con cuantiosos datos y con equilibrio entre clases.

Uno de los enfoques más empleados para sintetizar datos nuevos es el Synthetic Minority Oversampling Technique (SMOTE) [smote]. Esta técnica lo que hace es seleccionar muestras aleatorias de la clase minoritaria y usa un algoritmo KNN para seleccionar vecinos a los que se dibujan líneas [smotetheory].

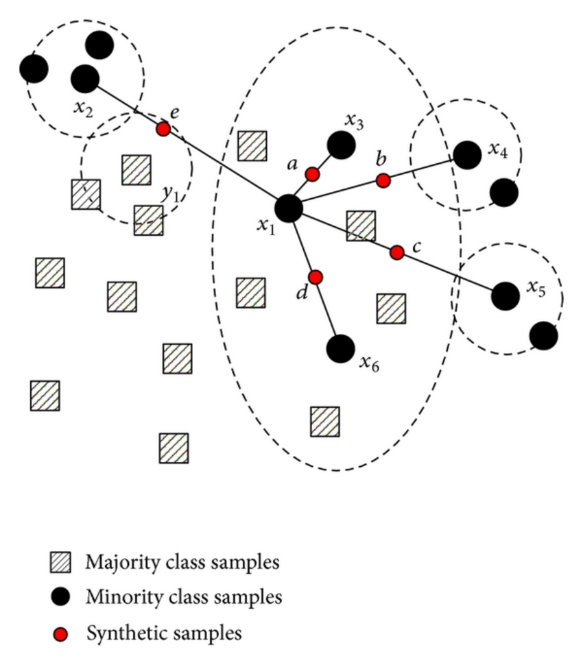


Figura 13 Representación del funcionamiento detrás de la técnica de oversampling SMOTE (Indresh Bhattacharyya, 2018)

Una vez está preparado el conjunto de datos, ya se pueden realizan las operaciones necesarias sobre todo el conjunto.

### Generación de datos artificiales (Data Augmentation)

Las técnicas de Data Augmentation se utilizan para crear nuevos datos sintéticos de entrenamiento añadiendo pequeñas perturbaciones a los datos iniciales. Esto da lugar a una mayor precisión y a una mayor generalización, ya que el objetivo es hacer al modelo invariante a las posibles perturbaciones que pueden encontrarse en muestras de audio [dataAug]. Cuando se trata de una tarea con archivos de audio, se pueden generar datos artificiales con las técnicas especificadas en la tabla x [augAudio].

|  |  |  |
| --- | --- | --- |
| Técnica | Descripción | Librería de Python empleada |
| Noise injection | Elevar el nivel de ruido a partir de un número aleatorio | Numpy |
| Stretch | Transformar la señal utilizando la Transformada de Fourier de Tiempo Reducido (STFT), la estira usando un codificador y usa la inversa de la STFT para reconstruir la señal en el dominio de tiempo | Librosa |
| Shift | Cambiar a el audio hacia la:   * izquierda (avance rápido) x segundos, lo que marca los primeros x segundos como 0 * derecha (rebobinar) x segundos, los últimos x segundos se marcarán como 0 | Librosa |
| Pitch | Cambiar el tono al azar | Librosa |
| Speed | Extiende o disminuye las series de tiempo una tasa fija | Librosa |

Tabla 3 Técnicas de Data Augmentation, descripción y librería de Python utilizada para su aplicación

### Extracción de características (Feature Extraction)

Para este proyecto, se persigue aplicar los beneficios de las características MFCC. Esta va a ser la técnica empleada, ya que se considera que estas características aportan un conocimiento suficientemente alto a los modelos sobre los audios de entrenamiento. Realizar una extracción de características empleando muchas técnicas resulta distorsionar demasiado la información contenida en los audios.

Para extraer los coeficientes MFCC se emplea una vez más la librería *librosa* que ofrece una función para ello y en la que se pueden ajustar distintos parámetros para adecuarla al formato de audio con el que se trabaja. La Figura 3 muestra el proceso detrás de la elaboración de un vector de MFCC.

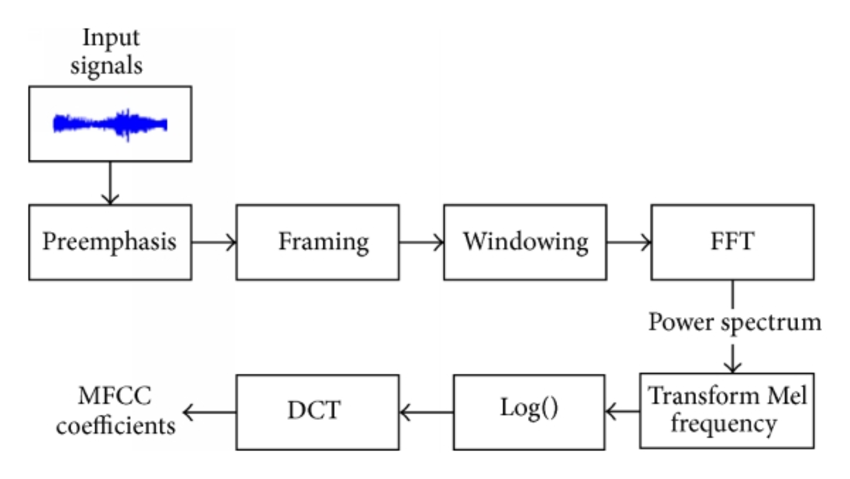


Figura 14 Proceso de extracción de los coeficientes MFCC durante el Feature Extraction (Gong, S., Dai, Y., Ji, J., Wang, J., & Sun, H., 2015)

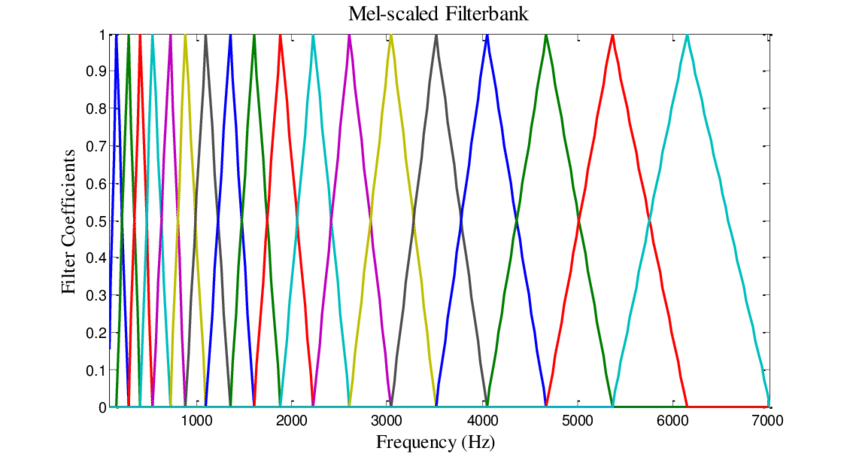
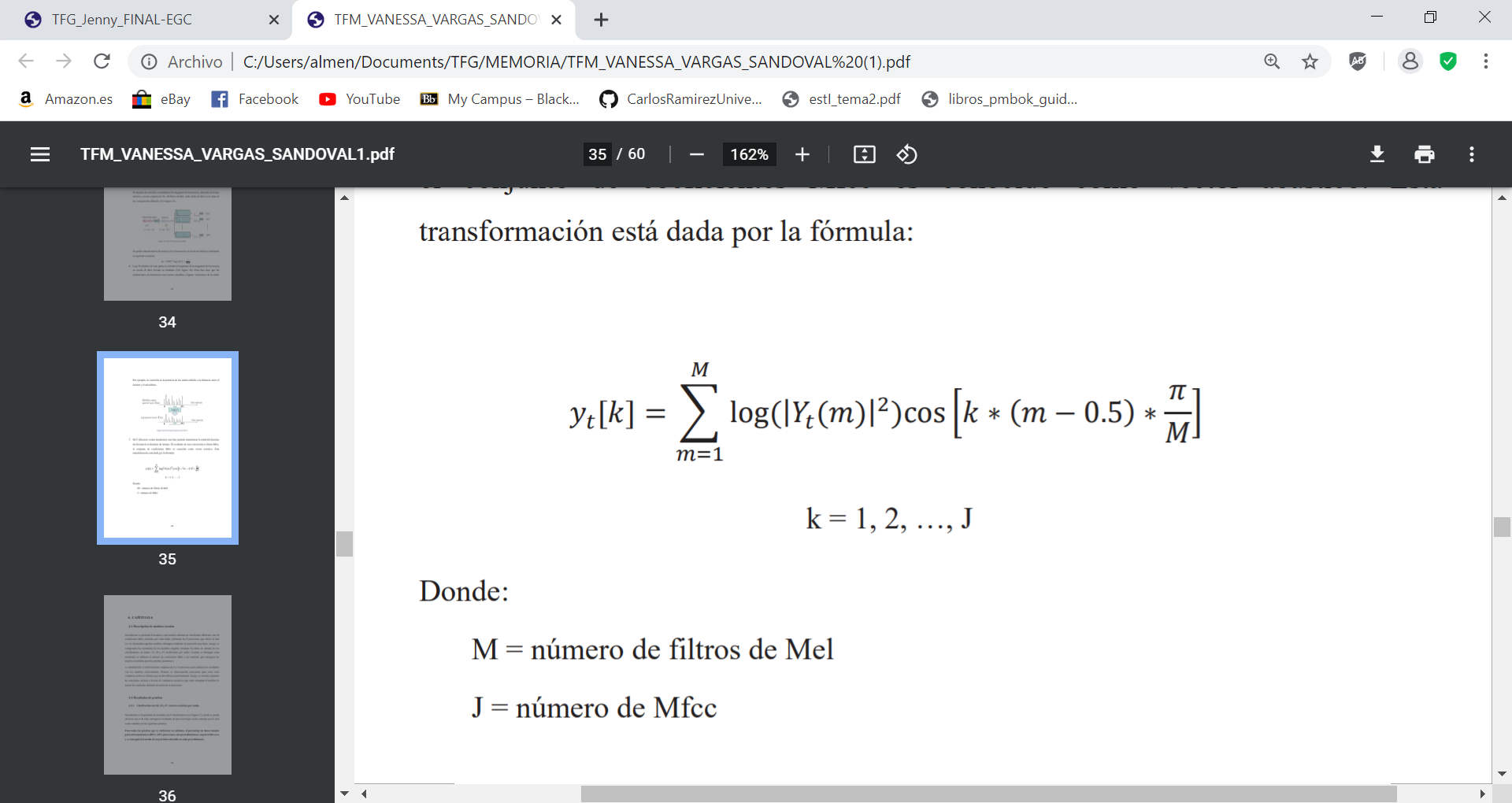
* Primero se aplica un filtro de pre-énfasis que enfatiza las frecuencias de resonancia de la cavidad acústica del tracto vocal.
* Posteriormente se divide la señal en tramos (*framing*) y se le aplica una función de *windowing*, generalmente se trabaja con una ventana de Hamming. El *windowing* sirve para eliminar los bordes de la señal y acentuar la parte central para su análisis.
* Al obtener la Transformada Rápida de Fourier (FFT) se extrae la magnitud de frecuencia de cada tramo y esta información se pasa a escala de Mel mediante el Banco de Filtros (Transform Mel frecuency). MFCC utiliza la escala Mel, que se divide en filtros espaciados linealmente a baja frecuencia (debajo de 1000Hz) y logarítmicos por encima de 1000Hz. Por ello existen más filtros en zonas de baja frecuencia que en las de alta (figura x) [filterbank].

Figura 15 Mel filter banks basis functions using 20 Mel-filters in the filter bank (Yusnita, M. A., Paulraj, M. P., Yaacob, S., Yusuf, R., & Shahriman, A. B., 2013)

* Después se obtiene el logaritmo de la señal, lo que hace que las estimaciones de frecuencia no sean tan sensibles a variaciones en la señal.
* Finalmente se aplica la Transformada de Coseno Discreta (DCT) a los coeficientes del espectro mel para convertirlos al dominio del tiempo y obtener los coeficientes cepstrales (MFCC). Esta transformación viene dada por la fórmula:



M = Nº filtros Mel

J = número de MFCC

Se decide extraer 50 MFFCs por audio, ya que la tarea de reconocimiento de emociones es compleja, y al estar usando tan solo una técnica de extracción de características, se pretende explotar la máxima riqueza de estos coeficientes.

Este proceso se aplica por un lado a los audios de voz masculina y por otro a los de voz femenina, por lo tanto, el resultado final son dos datasets que comparten la misma estructura: una columna con el nombre de la etiqueta, y el resto son 50 columnas con los 50 coeficientes extraídos para cada audio.

### Modelo y Entrenamiento

*Preparado de datos*

Una vez tenemos los datos en formato adecuando, se deben normalizar y dividir entre conjunto de entrenamiento y de pruebas. El proceso previo al entrenamiento se detalla en la figura 15.

Figura 16 Secuencia del preprocesamiento de datos

Figura 15.

Al tratarse de un problema de clasificación multi clase, y que los datos son categóricos, lo cual confunde al modelo. Por eso se aplica One-Hot Encoding a la etiqueta Y (las emociones), de tal manera que se crea una columna para cada valor distinto de las emociones que se están codificando y, para cada registro, marca con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0. Este es el formato adecuado para el entrenamiento [7].

Los datos se dividen en conjuntos de entrenamiento y test (80%-20%), repartiendo las emociones en cada grupo de manera aleatoria. Las dimensiones de los conjuntos de datos en este punto se muestran en la tabla 4.

|  |  |
| --- | --- |
| **Conjunto de datos** | **Dimensiones** |
| Ambos generos | |
| x\_train  y\_train | (68107,58)  (68107,58) |
| x\_test  y\_test | (17027, 58)  (17027, 8) |
| Femenino | |
| x\_train  y\_train | (39379, 58)  (39379, 8) |
| x\_test  y\_test | (9845, 58)  (9845, 58) |
| Masculino | |
| x\_train  y\_train | (28728, 58)  (28728, 8) |
| x\_test  y\_test | (7182, 58)  (7182, 8) |

Tabla 4 Dimensiones de los diferentes conjuntos de datos utilizados

Para el escalado de datos, se utiliza el módulo StandarScaler(), que escala los datos siguiendo una distribución normal (distribución Gaussiana con media 0 y varianza unitaria):

.

Al trabajar con modelos de Deep Learning, si las características no siguen esta distribución, el modelo puede resultar inestable. Esto afecta de manera negativa al aprendizaje y a la sensibilidad hacia los valores entrantes, desembocando en un error de generalización mayor [5]. Para trabajar con los modelos descritos a continuación, también es necesario expandir la dimensión (+1) de los datos de entrada, para ello se utiliza la librería *numpy*.

*Arquitectura del modelo: CNN+LSTM*

Este proyecto aplica una solución hibrida entre una CNN (Convolutional Neural Network) y un tipo de RNN (Recurrent Neural Network), el LSTM (Long-Short Term Memory). Esta arquitectura 1D CNN-LSTM se resume en la figura x.

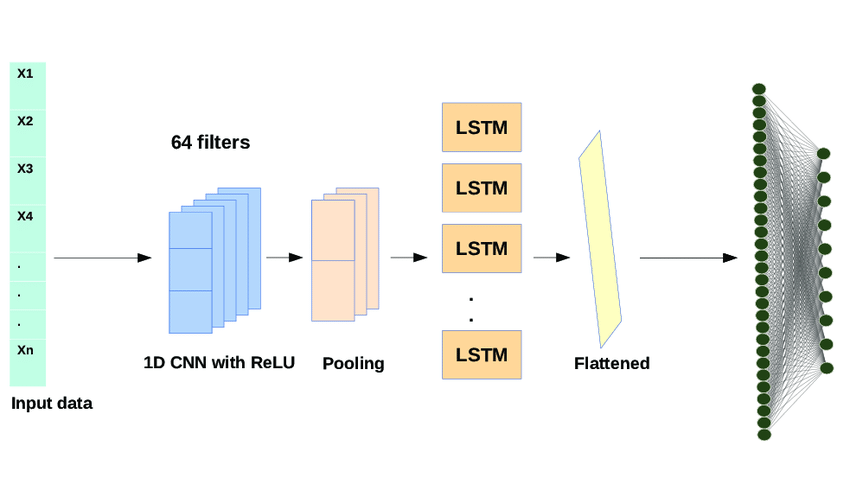
Tras probar diferentes opciones, detalladas en el apartado de resultados, se pudieron comprobar los beneficios de esta combinación entre ambas redes. En contraste con los modelos actuales que asumen un campo receptivo espacio-temporal fijo o un promedio temporal simple para el procesamiento secuencial, los modelos convolucionales recurrentes son "doblemente profundos" en el sentido de que pueden estar compuestos por "capas" espaciales y temporales. Estos modelos pueden tener ventajas cuando los conceptos son complejos y / o los datos de entrenamiento son limitados, ambos factores del presente trabajo. Esta arquitectura define dos submodelos: el modelo CNN para la extracción de características y el modelo LSTM para interpretar las características a lo largo del tiempo [cnnlstm].

Figura 17 Arquitectura de un modelo híbrido 1D CNN + LSTM (Hamad, R. A., Yang, L., Woo, W. L., & Wei, B., 2020)

Para el desarrollo de las redes neuronales artificiales se trabaja de la mano de la librería Keras, una de las librerías más populares de Redes Neuronales en Python. Es capaz de correr sobre Tensorflow, otra librería destacada de Machine Learning. Keras es ideal para experimentación con Deep Learning [8].

Se probaron diferentes modelos probando entre estructuras CNN, CNN-LSTM y experimentando con los optimizadores y los ‘callbakcs’, además de probar extrayendo diferentes características.

El modelo final consiste en una arquitectura 1D CNN-LSTM con los parámetros plasmados en la tabla 5.

|  |  |
| --- | --- |
| Nº epochs | 75 |
| Batch size | 48 |
| Optimizer | Adam |
| loss | Categorical crossentropy |
| Metrics | Accuracy |
| Callbacks | ReduceLROnPlateau |
| Mínimum learning rate | 0.000001 |
| Monitor metric | loss |

Tabla 5 Hiperparámetros del modelo final 1D CNN + LSTM

Este modelo se aplica a un conjunto de datos con 50 coeficientes MFCC extraídos de cada audio, para detectar una de las ocho emociones ofrecidas por los datasets utilizados.

El método de aprendizaje transversal implementado da lugar a un modelo de 13 capas, descritas en la tabla 6.

|  |  |  |
| --- | --- | --- |
| **Nº Capa** | **Nombre** | **Descripción** |
| **1** | Conv1D | Filter size= 256  Kernel size= 6  Strides = 1  Padding = same  Activation = relu  Al ser la primera capa, se establecen las dimensiones del audio de entrada |
| **2** | AveragePooling1D | Pool size = 4  Strides = 2  Padding = same |
| **3** | Conv1D | Filter size= 128  Kernel size= 6  Strides = 1  Padding = same  Activation = relu |
| **4** | AveragePooling1D | Pool size = 4  Strides = 2  Padding = same |
| **5** | Conv1D | Filter size= 128  Kernel size= 6  Strides = 1  Padding = same  Activation = relu |
| **6** | AveragePooling1D | Pool size = 4  Strides = 2  Padding = same |
| **7** | Dropout | 20% droput |
| **8** | LSTM | 128 hidden units  Activation = relu |
| **9** | Dropout | 20% dropout |
| **10** | Flatten |  |
| **11** | Dense | Units = 32  Activation = relu |
| **12** | Dropout | 30% dropout |
| **13** | Dense | Units = 8  Activation = softmax |

Tabla 6 Arquitectura seleccionada para la red neuronal artificial detrás del sistema final

El modelo CNN-LSTM lo componen cuatro tipos de capas importantes:

* **Conv1D**: al estar trabajando con señales de audio se emplean filtros convolucionales de 1 dimensión. Esta capa identifica regiones destacadas entre intervalos y activa ciertas características del audio, generando el mapa de características.
* **AveragePooling1D**: se utiliza como técnica de ‘downsampling’; simplifica la salida disminuyendo la tasa de muestreo tomando el valor promedio sobre la ventana definida por pool size. Esto reduce el número total de parámetros que la red necesita para aprender. Esta capa ayuda a evitar el problema de overfitting.
* **Unidad lineal rectificada (ReLU)**: da lugar a un entrenamiento más rápido y eficaz induciendo no linealidad.
* **LSTM**: aprende las dependencias a largo plazo en series de tiempo y datos secuenciales. Esto añade un componte que puede “recordar” estados previos y utilizar esta información para decidir cuál será el siguiente.
* **Dropout**: técnica de regularización que reduce la complejidad del modelo eliminando unidades tanto ocultas como visibles antes de pasar a la siguiente capa. Esto ayuda al modelo a generalizar y también reduce el overfitting.
* **Flatten:** se emplea antes de las fully-connected layers del final para aplanar los datos y reducir sus dimensiones para poder ser recogidos por las capas Dense.
* **Dense**: esta es la capa que realiza la clasificación. Todos los nodos están conectados con todos las de la capa anterior.
* **Softmax**: devuelve la probabilidad de que cada una de las clases sea cierta, y es la función de activación adecuada para problemas de multi clasificación.

La figura 17 a continuación muestra la dimensión de los datos a medida que transcurren a través de la arquitectura.

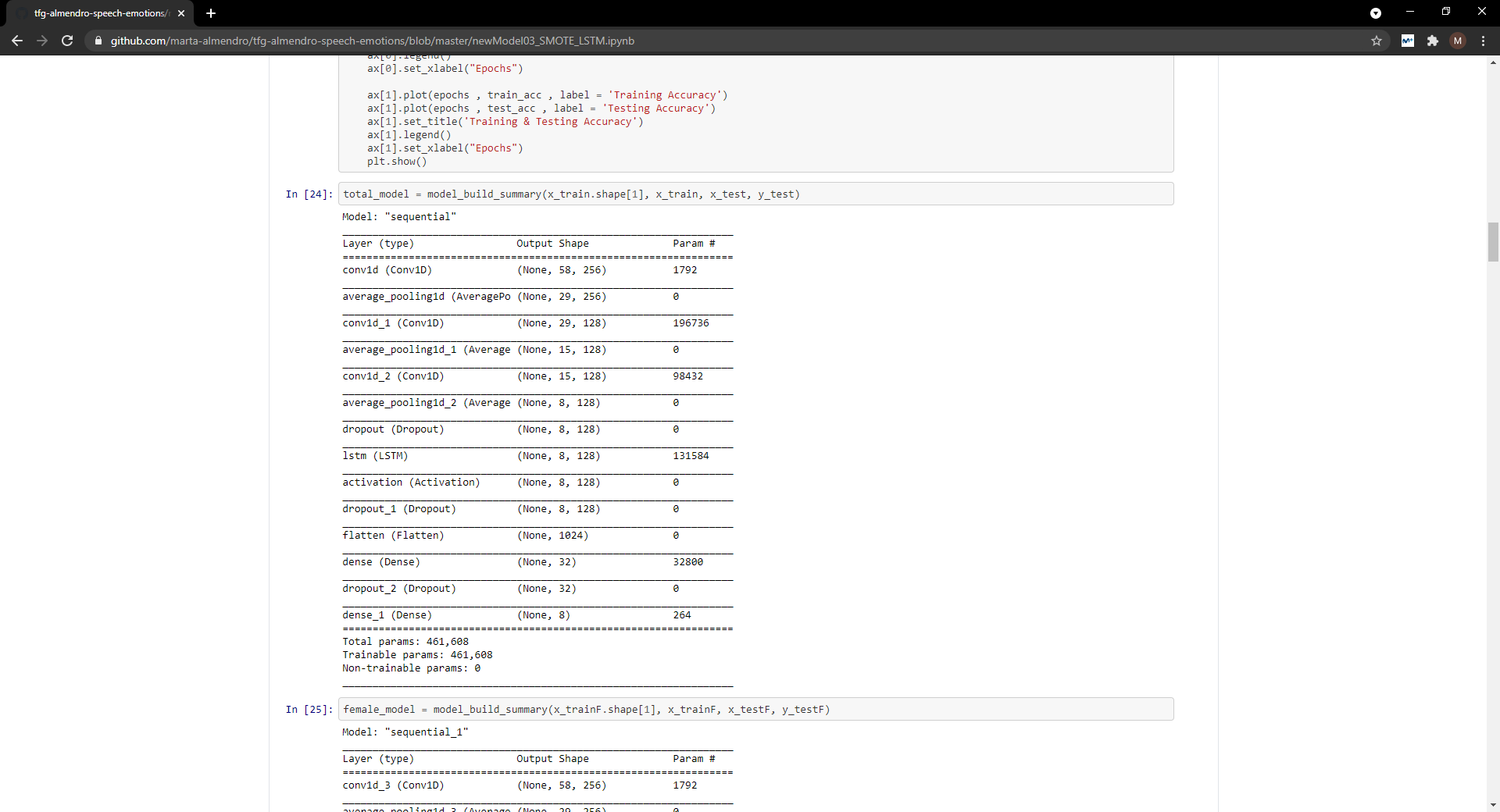


Figura 18 Arquitectura de la estructura seleccionada y las dimensiones de los datos de salida de cada capa

Ya que los datos están distribuidos de manera equilibrada, una métrica adecuada para comparar la destreza de los distintos modelos es la eficacia (*model accuracy*). Un análisis de los resultados obtenidos y la comparativa entre modelos se detalla en la sección 4.6 Resultados del Proyecto.

La figura 18 muestra la estructura del código usado para el entrenamiento.

**//Anaconda Jupyter Notebook con Lenguaje Python**

Figura 19 Esquema del script desarrollado para la fase de entrenamiento

1. Ingesta de audio

2. Extracción del espectograma y la forma de onda

3. Creación de datos sintéticos (Data Augmentation)

4. Extracción de características MFCC utilizando la librería *librosa* (Feature Extraction)

**// Software de procesamiento**

5. Mezcla aleatoria de los datos, división entre entrenamiento y test

6. Preparado de datos para ingesta del algoritmo

7. Estructura de la red neuronal artificial

8. Entrenamiento del algoritmo con los datos

9. Predicción de la emoción humana a partir de los datos de entrenamiento

|  |  |
| --- | --- |
| Librería | Descripción |
| os | Librería de Python que permite interactuar con el sistema operativo |
| pandas | Librería de Python especializada en el manejo y análisis de estructuras de datos |
| matplotlib | Librería de visualización de datos y trazado gráfico para Python |
| numpy | Librería de Python para trabajar con arrays y en el dominio del algebra lineal |
| seaborn | Librería de Python para visualización de datos basada en matplotlib. Especializada en gráficas y representaciones estadísticas |
| librosa | Librería de Python para el análisis de audio y música. Incluye capacidades de extracción de datos/características de los archivos de audio. |
| sklearn | Librería open source para aprendizaje automático en Python |
| Imblearn | Librería open source basada en sklearn que provee herramientas para tareas de clasificación con datos desbalanceados |
| tensorflow | Librería open source para aprendizaje automático en Python |
| keras | Librería open source de redes neuronales escrita en Python |

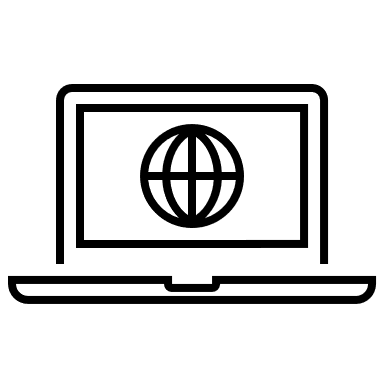
Tabla 7 Listado y descripción de las librerías empleadas en Python durante el proyecto

### Aplicación móvil

Una vez se ha completado el entrenamiento y se obtiene un modelo final, se plantea una arquitectura front-end que pueda recoger datos de entrada y recibir las predicciones del modelo sobre esos datos de entrada, con capacidad de respuesta en tiempo real. La arquitectura propuesta se muestra en la figura 20.

Imagen que contiene Icono

Descripción generada automáticamente



Servidor web

Figura 20 Estructura y flujo de datos de la aplicación propuesta

*Tecnologías empleadas*

**Lenguajes de programación:**

Figura 21 Logos de las tecnologías empleadas

* Typescript
* HTML

**Frameworks**

* Ionic
* Angular

**Control de versiones**

* Git

**Base de datos**

* Firebase

**Servidor web**

* Google Firebase

**Gestor de paquetes**

* Npm

## Recursos requeridos

Los recursos requeridos para este proyecto son únicamente computacionales. Estos se dividen en componentes hardware y software:

* Ordenador MSI (Windows) para desplegar la arquitectura. Aquí se instalan todos los programas necesarios para poder realizar todas las partes del proyecto:
  + Anaconda con Jupyter Notebook para guardar los datos de entrenamiento y entorno de programación para el preprecesado de datos, entrenamiento de los modelos y predicciones.
  + Se emplea el Anaconda prompt para descargar todas las librerías necesarias, usando el comando pip. Debido a la naturaleza más compleja de las dependencias de tensoflow y keras, se crea un environment especial donde se desarrollan los modelos de aprendizaje Deep Learning con estas librerías.
  + La aplicación móvil se desarrolla en Visual Studio Code. Aquí se trabaja con los frameworks Angular y Ionic.
* Dispositivos iPhone y Android: para probar la grabación de voz ya que con el emulador del ordenador no es posible comprobar su funcionamiento.
* Suscripción a la plataforma Google Cloud Platform: proveen el, hosting y conexión directa y fácil a la BBDD (Firebase).

## Presupuesto

|  |  |  |
| --- | --- | --- |
| **Tipo de coste** | **Valor** | **Comentarios** |
| Horas de trabajo en el proyecto | 470 horas | Duración aproximada del proyecto de 4 meses |
| Equipo técnico utilizado | | |
| Ordenador portátil- MSI Prestige 14 A10RB-020ES | 1349€ | Recurso ya adquirido. Precio actual en el mercado si se tuviese que adquirir nuevo. |
| Monitor- Asus 23.8 pulgadas | 129,99€ | Recurso ya adquirido. Precio actual en el mercado si se tuviese que adquirir nuevo. |
| Software utilizado | | |
| Jupyter notebook (Ipython Notebook) | 0€ | Software libre |
| Google Cloud Platform | 0€ | Se ha trabajado bajo la prueba gratuita de 300USD.  Una vez esta expira sigue un modelo de pago por uso. |
| Visual Studio | 0€ | Software libre |
| Frameworks y librerías | 0€ | Software libre |
| Fuentes de investigación | | |
| Estudios e informes | 0€ | Todos los informes y estudios de investigación empleados para realizar este proyecto eran gratuitos. |
| Materiales empleados | | |
| Smartphone – Xiaomi Mi 9 SE | 269€ | Recurso ya adquirido. Precio actual en el mercado si se tuviese que adquirir nuevo. |

Tabla 8 Estimación del presupuesto del trabajo realizado

## Viabilidad

Este apartado es opcional, pero es interesante que incluyas un breve análisis de viabilidad económica del proyecto (relación coste / beneficio), y un análisis de sostenibilidad a futuro del resultado de tu proyecto.

## Resultados del proyecto

### Ingesta, limpieza y análisis exploratorio de datos

Parte del proceso de verificación de los datos para el entrenamiento es que haya equilibrio entre las clases. Para ello se hace un conteo de ellas utilizando elementos visuales mostrados en la figura 21. Se hace un conteo tanto para las emociones de los hombres como de las mujeres, y se descubre que ambas se comportan igual; se identifica un destacado desbalance de clases que puede obstaculizar el proceso de entrenamiento.

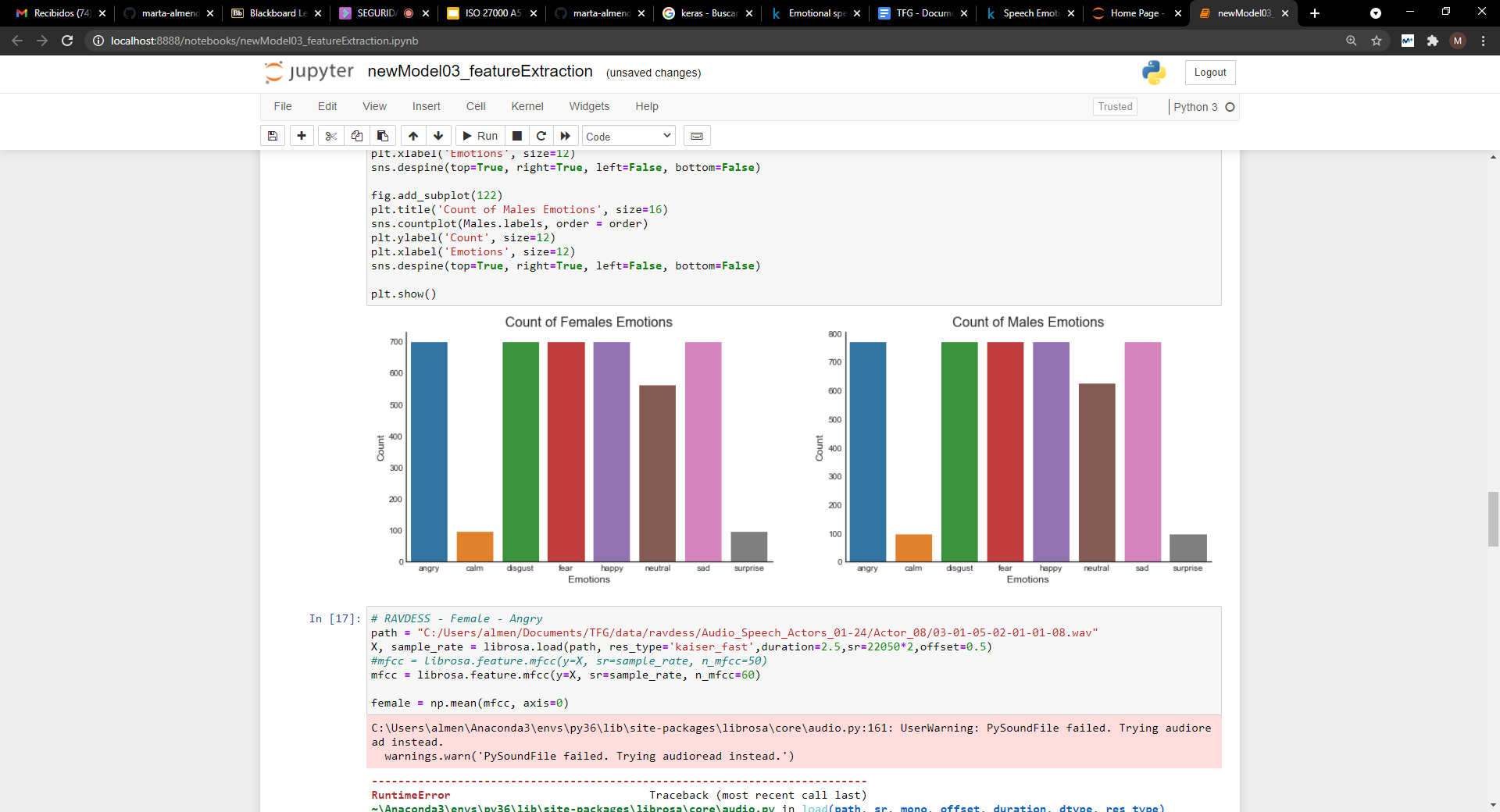


Figura 22 Diagrama de barras mostrado el conteo de cada emoción antes de aplicar oversampling a los datasets femenino (izquierda) y masculino (derecha)

Existe una carencia de emociones de *calm* y de *surprise*. Es un desafío para las tareas de modelado predictivo trabajar con una distribución de clases severamente sesgada, causando desigualdad en los costos de la clasificación errónea [6].

Tras haber realizado el preprocesamiento al completo de los datos, incluyendo las técnicas de *oversampling*, generación de datos sintéticos y extracción de características, el dataset ya está preparado para poder pasar a la fase de entrenamiento. En este punto se trabaja con un dataset que contiene la misma cantidad de datos (figura 22) para cada emoción, integrado con nuevos datos sintéticos, y al conjunto entero se le han extraído las características que van a permitir al modelo a continuación entender los datos.

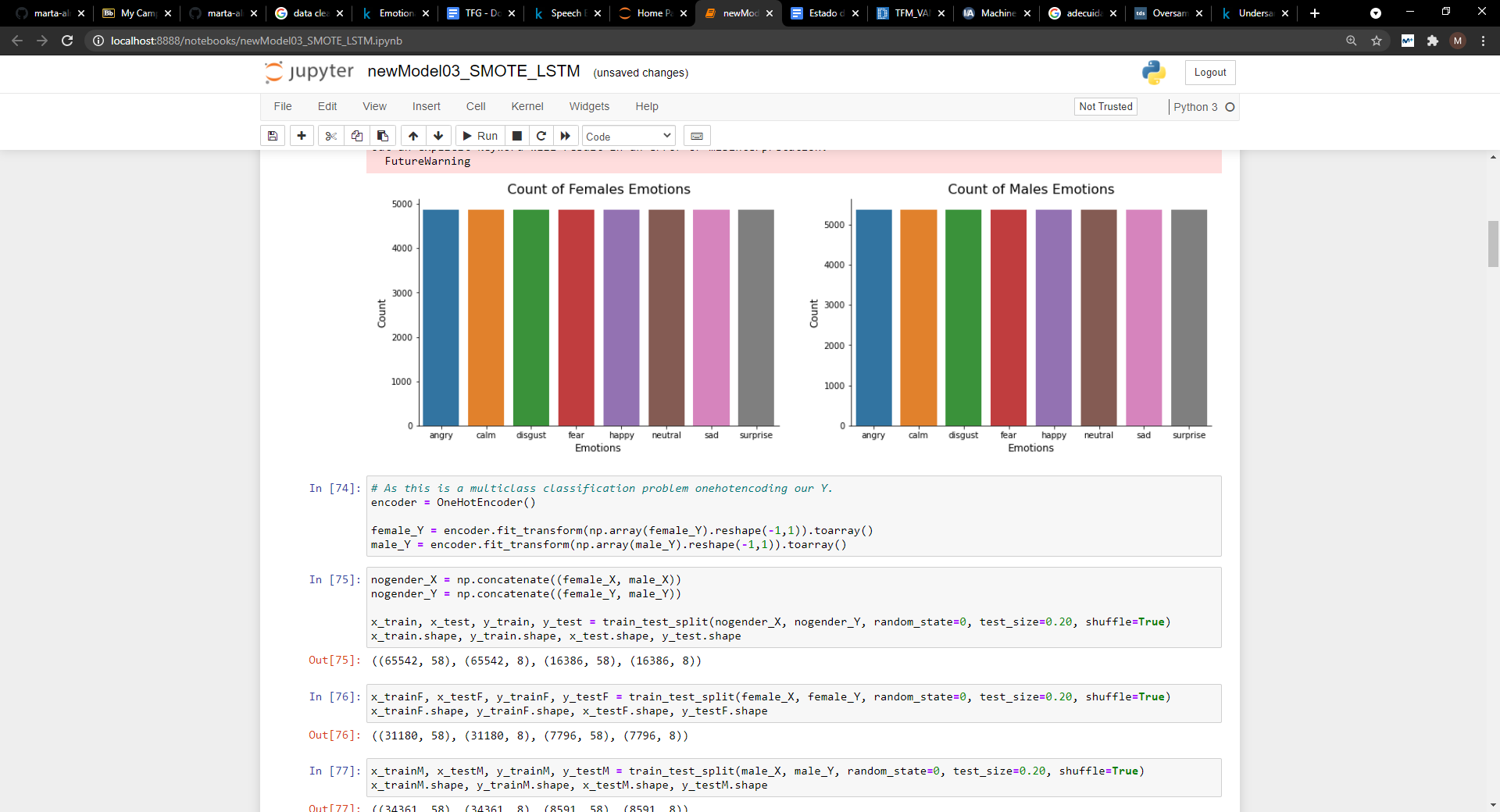


Figura 23 Diagrama de barras mostrado el conteo de cada emoción después de aplicar oversampling a los datasets femenino (izquierda) y masculino (derecha)

### Extracción de características (Feature Extraction)

Para determiner el número de MFCC que seleccionar de los datos, se hace un análisis previo de los espectrogramas tras haber aplicado la extracción de características MFCC. La figura x muestra el resultado tras extraer 20 coeficientes, y la figura x con 50.

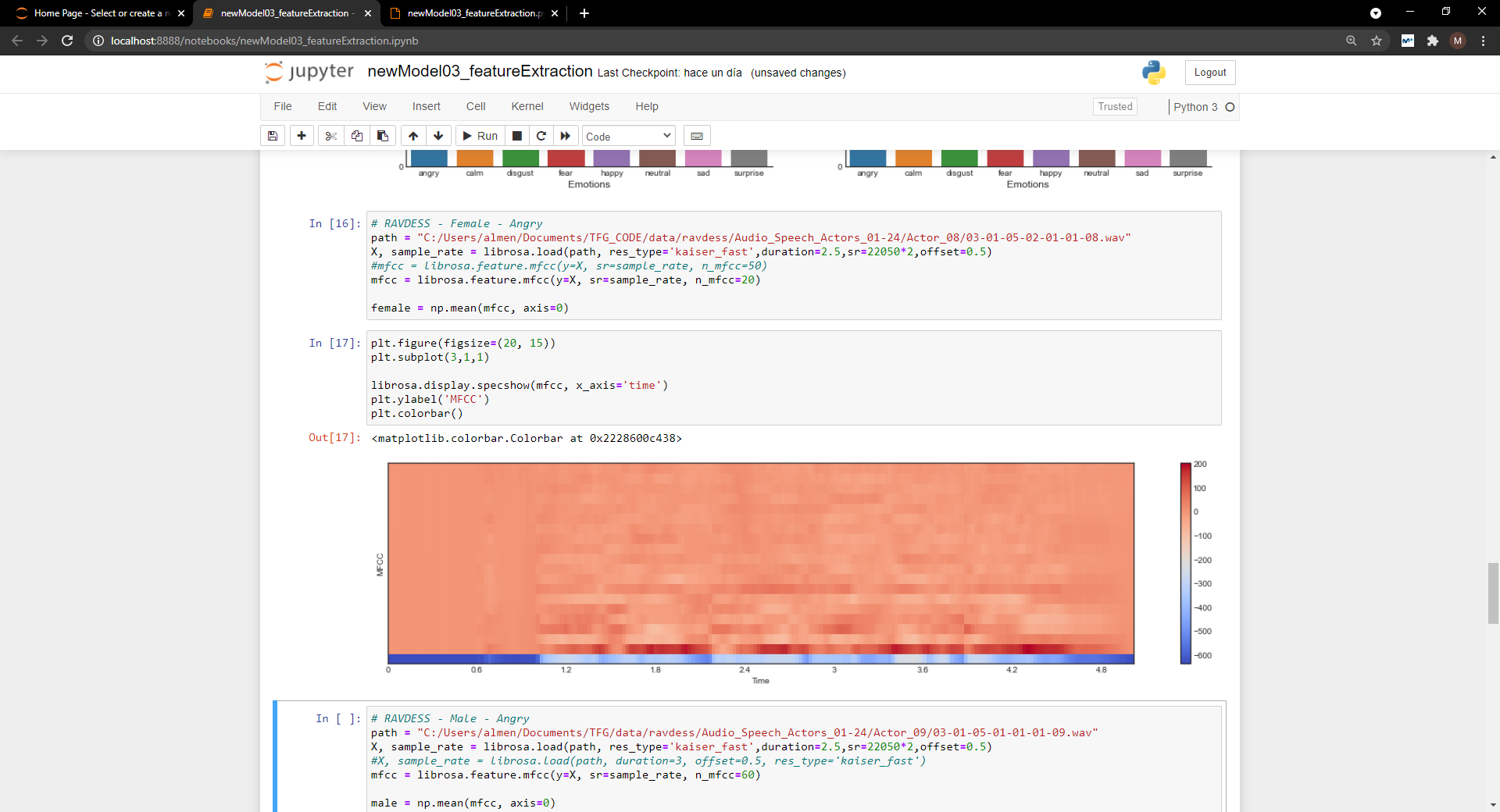


Figura 24 Espectograma de MFCCs de un audio de muestra: 03-01-05-02-01-01-08.wav, emoción de enfado de una mujer. Extracción de 20 coeficientes

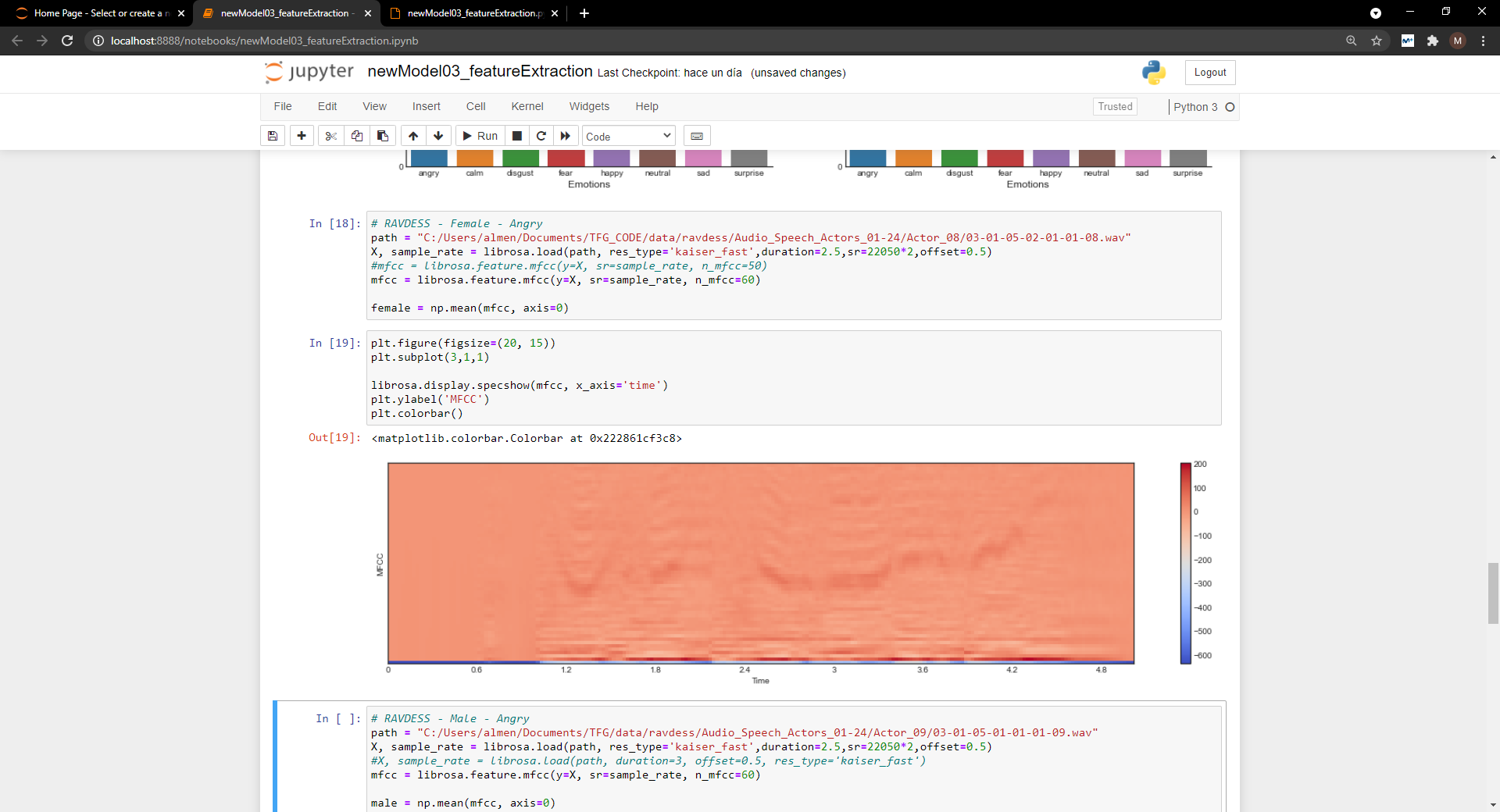
Extrayendo 50 coeficientes (figura 23) se distinguen más formantes. Las zonas oscuras son las que muestran formantes en el espectro; los sonidos se pueden identificar mucho mejor a través de los formantes y sus transiciones. Por lo tanto, se opta por la extracción de 50 MFCCs de cada muestra de audio.

Figura 25 Espectograma de MFCCs de un audio de muestra: 03-01-05-02-01-01-08.wav, emoción de enfado de una mujer. Extracción de 50 coeficientes

### Modelo y Entrenamiento

Se probaron diferentes estructuras de redes neuronales, combinando diferentes técnicas de extracción de características, tipos de redes neuronales artificiales, y combinación de estas. Los detalles de esta experimentación, incluyendo los resultados de cada prueba, se resumen en las tablas a continuación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features extraídos | Arquitectura | Optimizador | Validation Acc. | Validation Loss |
| ZCR, Chroma Shift, MFCC, RMSV, Mel Spectogram | 2 layers Conv1D + MaxPooling + Dense + Flatten | Adam | 0.68 | 1.03 |
| ZCR, Chroma Shift, MFCC, RMSV, Mel Spectogram | 4 layers Conv1D + MaxPooling +  Flatten | Adam | 0.71 | 1.32 |
| ZCR, Chroma Shift, MFCC, RMSV, Mel Spectogram | 2 layers Conv1D+ BatchNormalisation + 1 layer LSTM + MaxPooling + Dense + Flatten | Adam | 0.72 | 0.97 |
| ZCR, Chroma Shift, MFCC, RMSV, Mel Spectogram | 2 layers Conv1D + BatchNormalisation + 1 layer LSTM + MaxPooling + Dense + Flatten | Adam + ReduceLROnPlateau | 0.77 | 0.82 |
| MFCC | 4 layers Conv1D + 1 layer LSTM + AvgPooling + Flatten + 2 layers Dense | Adam + ReduceLROnPlateau | 0.87 | 0.51 |
| MFCC | **3 layers Conv1D +**1 layer LSTM **+ AvgPooling + Flatten + 2 layers Dense** | **Adam + ReduceLROnPlateau** | **0.89** | **0.37** |

Tabla 9 Arquitectura de los modelos entrenados para el conjunto de datos de género mixto y los resultados de validation accuracy y los para cada etapa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features extraídos | Arquitectura | Optimizador | Validation Acc. | Validation Loss |
| MFCC | 4 layers Conv1D + AvgPooling + Flatten + 2 layers Dense | Adam + ReduceLROnPlateau | 0.92 | 0.36 |
| MFCC | **3 layers Conv1D + AvgPooling + Flatten + 2 layers Dense** | **Adam + ReduceLROnPlateau** | **0.93** | **0.35** |

Tabla 10 Arquitectura de los dos últimos modelos entrenados para el conjunto de datos de género femenino y los resultados de validation accuray y los para cada etapa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features extraídos | Arquitectura | Optimizador | Validation Acc. | Validation Loss |
| MFCC | 4 layers Conv1D + AvgPooling + Flatten + 2 layers Dense | Adam + ReduceLROnPlateau | 0.91 | 0.51 |
| MFCC | **3 layers Conv1D + AvgPooling + Flatten + 2 layers Dense** | **Adam + ReduceLROnPlateau** | **0.91** | **0.4** |

Tabla 11 Arquitectura de los modelos entrenados para el conjunto de datos de género masculino y los resultados de Validation Accuracy y Loss para cada etapa

A su vez se visualizan las gráficas de las métricas de Accuracy y Loss extraídas durante las fases de entrenamiento y de validación. La figura 24 a continuación muestra los resultados obtenidos para estas métricas durante para el entrenamiento de la entrada (2) de la tabla 9:

* **Features extraídos**: ZCR, Chroma Shift, MFCC, RMSV, Mel Spectogram
* **Arquitectura**: 4 layers Conv1D + MaxPooling + Flatten
* **Optimizador**: Adam
* **Validation Accuracy**: 0.71
* **Validation Loss**: 1.32

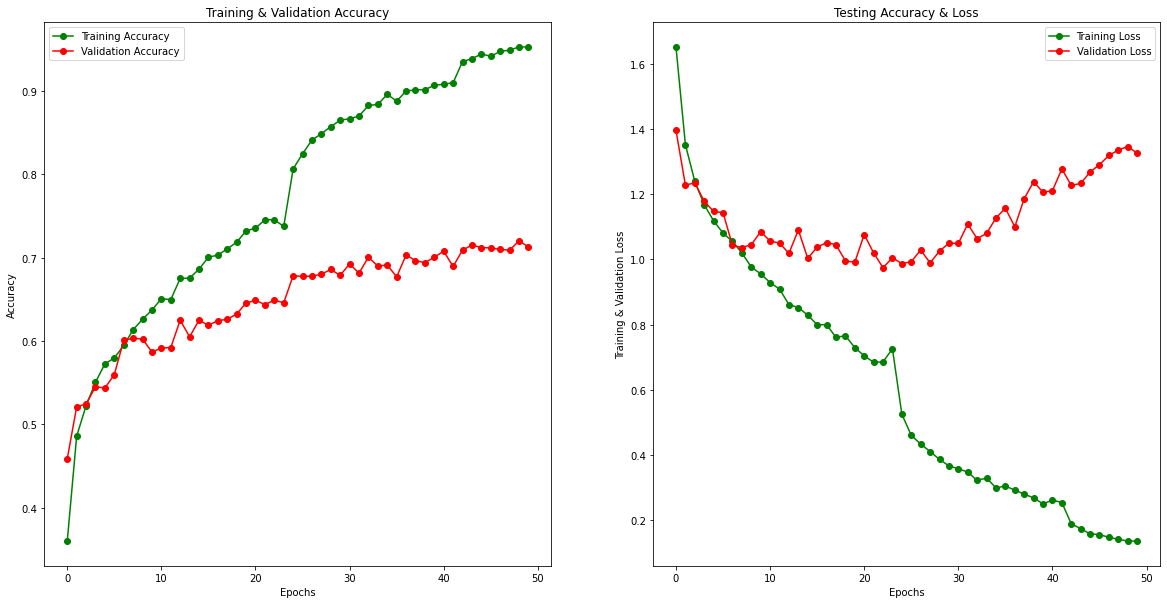


Figura 26 Gráficas de Accuracy y Loss durante el entrenamiento del segundo modelo de la tabla 9.

El patron mostrado en las gráficas lo adoptan los primeros modelos, hasta que se decide probar un enfoque de extracción de características más simple. Lo que más preocupaba era el comportamiento del Loss, ya que sugería la posibilidad de que el modelo estaba *overfitting*. El *overfitting* es una de las mayores amenazas para estos algoritmos de clasificación, por lo tanto, se consideraba importante tenerlo en el punto de mira.

Para solucionar este problema se puso el foco en probar diferentes modificaciones en la estructura y en el ajuste de hiperparámetros.

*Cambios en los métodos y en la estructura:*

1. Cambiar la estrategia de Feature Extraction, considerando que la extracción de los MFCC va a ser suficiente para este trabajo. Esto reduce la carga computacional y preserva mejor la forma original de los datos, lo cual resulta ser la mejor opción que seguir.
2. Cambiar las Pooling Layers de Max a Avg, la primera lleva a cabo agrupaciones más discretas, al estar realizando la operación máxima, extrae las características más destacadas, como los bordes, de cada audio. Average Pooling realiza cálculos más generalizados. Las características MFCC aportan mucha información sobre las características más destacadas, y durante las primeras fases de entrenamiento se puede comprobar que una extracción de características excesiva puede dañar los datos y, por lo tanto, el modelo.
3. Aplicar una arquitectura hibrida CNN-LSTM, sacando partido de los beneficios de ambas modalidades.

*Ajuste de hiperparámetros (Hyperparemeter Tuning)*

**Learning rate.** La tasa de aprendizaje es un hiperparámetro muy importante y, a menudo, requiere algo de experimentación. Con una tasa de aprendizaje demasiado grande, puede rebotar alrededor de un óptimo, o puede comenzar disparando hacia una parte del espacio de parámetros donde los gradientes desaparecen. Con un valor demasiado pequeño, es posible que tarde demasiado en converger a un óptimo, o puede que encuentre un óptimo local deficiente. Estos efectos se ven atenuados por el efecto impulso del optimizador Adam.

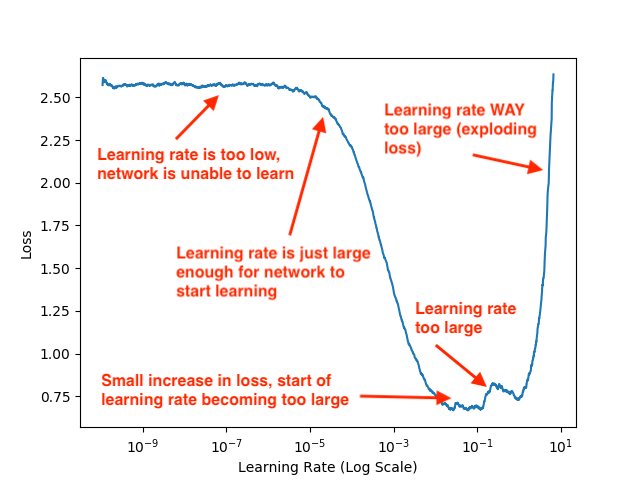


Figura 27 Comportamiento del Loss en función del Learning Rate (Rosebrock, 2021)

Es posible que la causa de que el Loss incremente al final durante el entrenamiento venga de que el learning rate sea demasiado alto (figura 25). Dicho esto, se decide trabajar con el learning rate por defecto que establece Adam, y luego implemantar el callback ReduceLROnPlateau, que soluciona este problema reduciendo el learning rate cuando la métrica objetivo deja de mejorar.

Aplicar estos cambios tiene un impacto notable en las curvas de resultados del entrenamiento. Los mejores resultados pertenecen el modelo femenino, seguido por el masculino y el mixto. Esto sugiere que una división por sexo es beneficioso para esta tarea de clasificación de emociones.

Figura 28 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género mixto

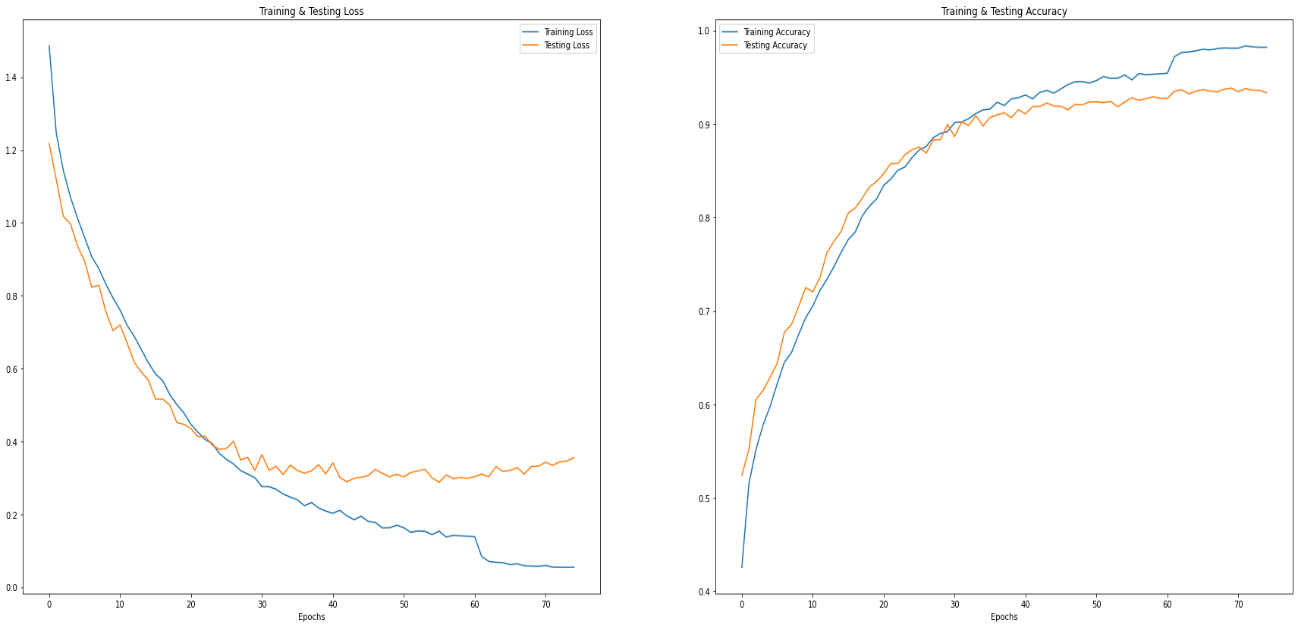
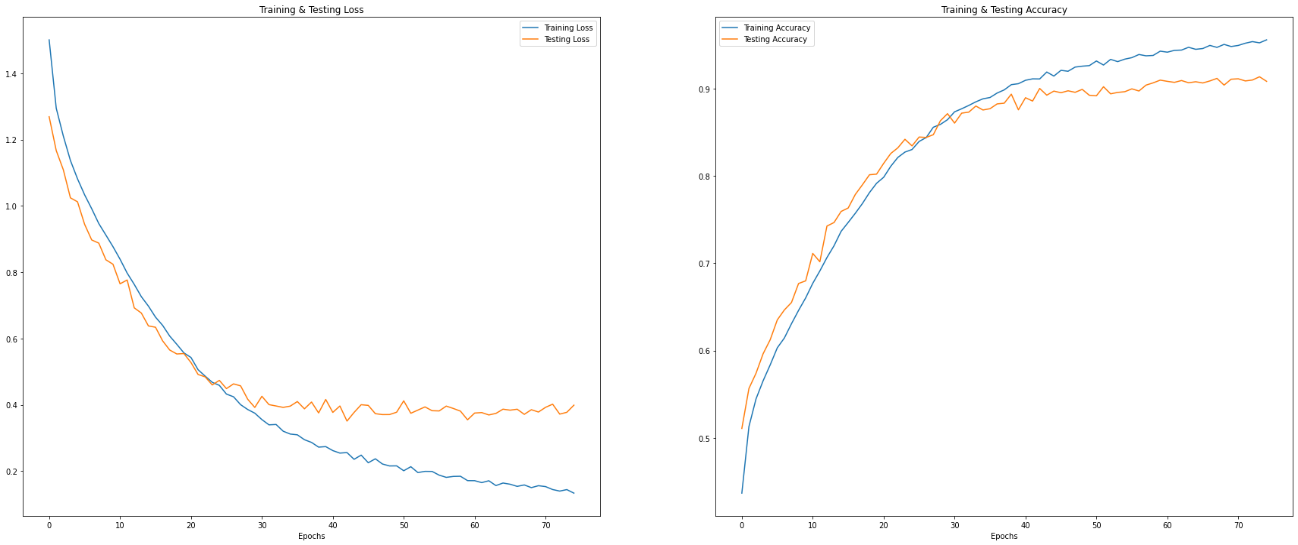


Figura 29 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género masculino

Figura 30 Curvas de Accuracy y Loss para el entrenamiento y validación del modelo final. Género femenino

Una vez se da con la arquitectura que ofrece mayor eficacia, se pasa a la fase de testeo, durante la cual se analiza en detalle la capacidad de predicción del modelo seleccionado. Para ello se analizan las matrices de confusión, a través de las cuales se identifica si existen confusiones específicas en la detección de clases.

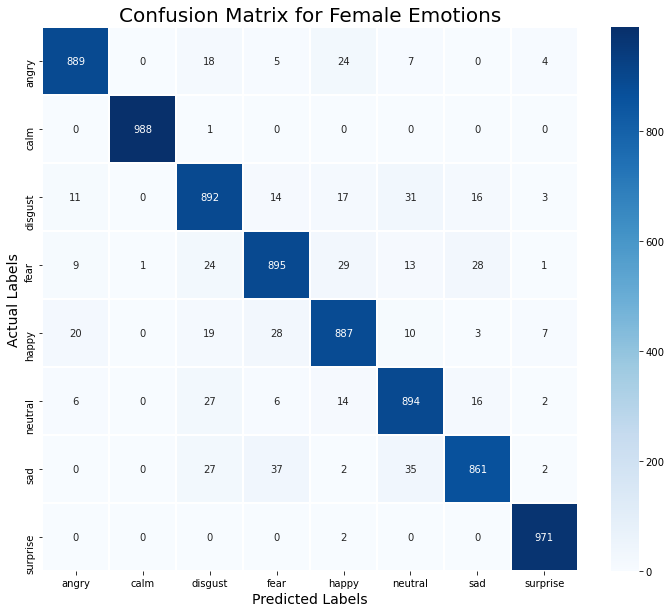
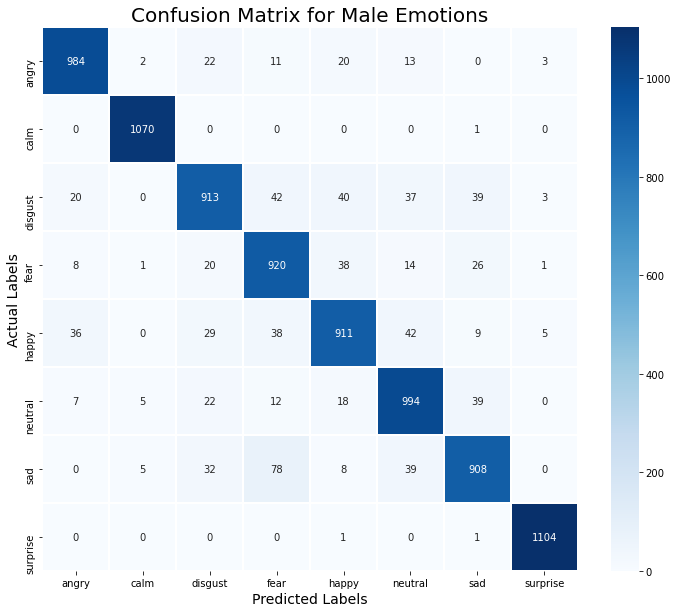


Figura 31 Matriz de confusión para las predicciones de emociones del dataset femenino

Por lo general, el clasificador realiza predicciones exitosas. Se observa que las emociones que más confunde son *fear* con *sad,* lo cual resulta una confusión razonable. El patrón es muy similar para ambas matrices de confusión, lo que indica que el desempeño del algoritmo no va a ser condicionado por el género del interlocutor.

Figura 32 Matriz de confusión para las predicciones de emociones del dataset masculino

Las emociones que mejor se clasifican son *surpise* y *calm,* pero no existe una diferencia muy notable entre estas y el resto. El ligero aumento en estas dos categorías puede ser debido a que son las dos emociones a las cuales se le aplica el oversampling al comienzo. El oversampling produce muestras menos variadas de los datos, lo que hace que estas dos categorías tengan datos a partir de los cuales es más fácil generalizar.

La infraestructura final del sistema es capaz de recibir peticiones de múltiples usuarios. Se tiene que replicar el proceso realizado de extracción de característica y procesado de datos para poder ser servidos al modelo entrenado. Una imagen global del sistema lo muestra la figura 31.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 33 Sistema de reconocimiento de emociones

# DISCUSIÓN

Esta sección es más habitual en trabajos de tipo "científico" o de "investigación", donde uno presenta brevemente los resultados principales y los discute, pero también puede utilizarse en otro tipo de trabajos.

Puedes incluir secciones específicas para discutir cuestiones como: limitaciones del estudio, limitaciones de la tecnología empleada, cambios respecto a objetivos planteados inicialmente.

Puedes incluir respuestas a preguntas del tipo: ¿la metodología inicialmente pensada ha sido útil?, ¿he tenido que adaptarme a cambios a lo largo del proyecto?, ¿qué cambios han sido y cómo he adaptado el proyecto para poder manejar esos cambios?, ¿qué impacto ha tenido el resultado de mi proyecto?

Este proyecto comenzó con el objetivo de entender como implementar redes de aprendizaje profundo y como conseguir una arquitectura que realice las tareas de predicción de manera mas eficaz. Sin embargo, a medida que transcurría el curso del proyecto, se volvió predominante la importancia de la ingeniería de características. Se observa que, en el reconocimiento de emociones a través de la voz, la elección de las características extraídas del audio tiene un impacto más eminente que la complejidad del modelo. Esto lleva a la conclusión de que esta última no compensa con los resultados obtenidos; incluso son los modelos más simples los que devuelven resultados más satisfactorios.

# CONCLUSIONES

## Conclusiones del trabajo

Breve descripción objetiva del resultado en relación al objetivo general de tu proyecto.

## Conclusiones personales

Describe tus impresiones y experiencia personal durante el desarrollo del proyecto, o destacar la importancia que tiene el tema para ti, lo que has aprendido, o la trascendencia que ha tenido para ti o para otros este proyecto.

# FUTURAS LÍNEAS DE TRABAJO

Es importante destacar en este apartado las líneas con las que se podría continuar tu trabajo. Indica todo lo que has anotado como futuro trabajos durante el desarrollo del proyecto, o aquellos aspectos que estaban fuera del alcance, pero que son interesantes para desarrollar a futuro el valor de los resultados de tu trabajo.

# REFERENCIAS

[1] Adam, F: The Genuine Works of Hippocrates, Translated from the Greek with a Preliminary Discourse and Annotations. New York, W Wood and Co (1886)

[2] Affective Computing Market | 2020-2027 | Industry Report | Covid Insights. (2021). Retrieved 21 May 2021, from <https://www.mordorintelligence.com/industry-reports/affective-computing-market>

[3] Affective Computing Market Outlook with COVID-19 Impact Analysis & Opportunities, Future Challenges, Growth Statistics and Forecast to 2023 – Factory Gate. (2021). Disponible en: from https://factorygate.co.uk/uncategorized/997096/affective-computing-market-outlook-withcovid-19-impact-analysis-opportunities-future-challenges-growth-statistics-and-forecast-to2023/ (Consultado el 14 de enero, 2021).

[4] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.

[feature4] Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155-177.

[5] Brownlee, J. (2021). How to use Data Scaling Improve Deep Learning Model Stability and Performance. Retrieved 21 May 2021, from https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

[6] Brownlee, J. (2021). Why Is Imbalanced Classification Difficult?. Retrieved 21 May 2021, from <https://machinelearningmastery.com/imbalanced-classification-is-hard/>

[7] Brownlee, J. (2021). Why One-Hot Encode Data in Machine Learning?. Retrieved 21 May 2021, from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[8] Brownlee, J. (2021). Your First Deep Learning Project in Python with Keras Step-By-Step. Retrieved 21 May 2021, from https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

[SMOTE] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

[7] Clynes, M. (1977). *Sentics: The touch of emotions*. Anchor Press.

[4] Cohen, S., Tyrrell, D. A., & Smith, A. P. (1991). Psychological stress and susceptibility to the common cold. *New England journal of medicine*, *325*(9), 606-612.

[augAudio] Data Augmentation for Audio. (2021). Retrieved 21 May 2021, from <https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6>

[16] de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI. Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems. Int J Environ Res Public Health. 2020 Jul 31;17(15):5542

[buddist] de Silva, P. (2020). Health and Emotional Experience. In *Mindfulness-based Emotion Focused Counselling* (pp. 171-176). Palgrave Macmillan, Cham.

[17] Dileep, A. D., & Sekhar, C. C. (2013). GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(8), 1421-1432.

[cnnlstm] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

[COLORS] Donaldson, M. (2017). Plutchik's wheel of emotions—2017. Update.

[intsim] Fabien, M. (2021). Emotion Recognition WebApp. Retrieved 21 May 2021, from <https://maelfabien.github.io/project/poleemploi/#real-time-multimodal-emotion-recognition>

[expt] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, *47*(7), 829-837.

[dataAug] franspg, V. (2021). Generación de datos artificiales (Data Augmentation). Retrieved 21 May 2021, from <https://franspg.wordpress.com/2020/01/27/generacion-de-datos-artificiales-data-augmentation/>

[Dalai lama] Goleman, D. (Ed.). (2003). *Healing emotions: Conversations with the Dalai Lama on mindfulness, emotions, and health*. Shambhala publications.

[MFCCDiagrama] Gong, S., Dai, Y., Ji, J., Wang, J., & Sun, H. (2015). Emotion analysis of telephone complaints from customer based on affective computing. *Computational intelligence and neuroscience*, *2015*.

[15] Google Health. (2021). Disponible en: https://health.google/health-research/ (Consultado el 8 de enero, 2021).

[3] Gordon, J. S. (1981). Holistic medicine: toward a new medical model. *The Journal of clinical psychiatry*, *42*(3), 114-119.

[cnnlstm] Hamad, R. A., Yang, L., Woo, W. L., & Wei, B. (2020). Joint Learning of Temporal Models to Handle Imbalanced Data for Human Activity Recognition. *Applied Sciences*, *10*(15), 5293.

[1] Havariyoun G, Christofi N. Coding with Pediatric Patients at King's College Hospital: Experience from a London Hospital. J Med Imaging Radiat Sci. 2016 Dec;47(4):294-297.

[smotetheory] How to Effortlessly Handle Class Imbalance with Python and SMOTE. (2021). Retrieved 21 May 2021, from <https://towardsdatascience.com/how-to-effortlessly-handle-class-imbalance-with-python-and-smote-9b715ca8e5a7>

[SERCNN2] Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014, November). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 801-804).

[SER+sectors] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 235-238.

[covidmodel] Islam, M. Z., Islam, M. M., & Asraf, A. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in medicine unlocked*, *20*, 100412.

[cleandata] Jesmeen, M. Z. H., Hossen, J., Sayeed, S., Ho, C. K., Tawsif, K., Rahman, A., & Arif, E. M. H. (2018). A survey on cleaning dirty data using machine learning paradigm for big data analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, *10*(3), 1234-1243.

[9] Jung, H. W., Seo, Y. H., Ryoo, M. S., & Yang, H. S. (2004, November). Affective communication system with multimodality for a humanoid robot, AMI. In *4th IEEE/RAS International Conference on Humanoid Robots, 2004.* (Vol. 2, pp. 690-706). IEEE.

[DIBUJOS] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327-117345.

[Feature1] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, *15*(2), 99-117.

[SER4] Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, *13*(2), 293-303.

[MFCC] Logan, B. (2000, October). Mel frequency cepstral coefficients for music modeling. In *Ismir* (Vol. 270, pp. 1-11).

[10] Luneski, A., Konstantinidis, E., & Bamidis, P. (2010). Affective medicine: a review of affective computing efforts in medical informatics. *Methods of information in medicine*, *49*(3), 207-218.

[cleandata2] Machine learning demystified: the importance of data. (2021). Retrieved 21 May 2021, from <https://www.information-age.com/machine-learning-demystified-importance-data-123466738/>

[mercado2] Market, A. (2021). Affective Computing Market Size, Share and Global Forecast to 2024 | COVID-19 Impact Analysis | MarketsandMarkets. Retrieved 21 May 2021, from <https://www.marketsandmarkets.com/Market-Reports/affective-computing-market-130730395.html>

[14] Mesko, B. Artificial Intelligence Will Redesign Healthcare. 2016. Disponible en: https://www.linkedin.com/pulse/artificial-intelligence-redesign-healthcare-bertalanmesk%C3%B3-md-phd (Consultado el 8 de enero, 2021)

[SERCNN] Murugan, Harini. (2020). Speech Emotion Recognition Using CNN. International Journal of Psychosocial Rehabilitation. 24. 10.37200/IJPR/V24I8/PR280260.

[SER5] Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech communication*, *41*(4), 603-623.

[imbclass2] Oversampling and Undersampling. (2021). Retrieved 21 May 2021, from <https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf>

[DIBUJOHMI] Papetti, S. (2013). Design and Perceptual Investigations of Audio-Tactile Interactions. *Proceedings of AIA DAGA*.

[6] Picard, R. W. (2000). *Affective computing*. MIT press.

[BasicEmotions] Piórkowska, Magda & Wrobel, Monika. (2017). Basic Emotions. 10.1007/978-3-319-28099-8\_495-1.

[lr] Rosebrock, A. (2021). Keras Learning Rate Finder - PyImageSearch. Retrieved 21 May 2021, from https://www.pyimagesearch.com/2019/08/05/keras-learning-rate-finder/

[Ems] Ruiz, P., Apud, I., Maiche, A., González, H., Pires, A. C., Carboni, A., ... & González Perilli, F. (2016). Manual de introducción a la psicología cognitiva.

[OPS] Salud Mental y COVID-19 - OPS/OMS | Organización Panamericana de la Salud. (2021). Retrieved 21 May 2021, from https://www.paho.org/es/salud-mental-covid-19

[DL+SERCNN3] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

[11] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90-99.

[OMSComite] Statement on the fifth meeting of the International Health Regulations (2005) Emergency Committee regarding the coronavirus disease (COVID-19) pandemic. (2021). Retrieved 21 May 2021, from https://www.who.int/news/item/30-10-2020-statement-on-the-fifth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic

[5] Stone, A. A., Cox, D. S., Valdimarsdottir, H., Jandorf, L., & Neale, J. M. (1987). Evidence that secretory IgA antibody is associated with daily mood. *Journal of personality and social psychology*, *52*(5), 988.

[2] The World Health Organisation (WHO) (1948). Preamble to the constitution of the World Health Organisation as adopted by the International Health Conference. Disponible en: https://apps.who.int/gb/bd/PDF/bd47/SP/constitucion-sp.pdf?ua=1 (Consultado el 5 de enero, 2021).

[8] Unbehauen, H. (Ed.). (2009). *CONTROL SYSTEMS, ROBOTICS AND AUTOMATION-Volume III: System Analysis and Control: Classical Approaches-III*. EOLSS Publications.

[SER2] Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*.

[vmote] ‎Vmote - Voice Messaging. (2021). Retrieved 21 May 2021, from <https://apps.apple.com/gt/app/vmote-voice-messaging/id1273369846>

[empath] Vocal Emotion Recognition Test by Empath. (2021). Retrieved 21 May 2021, from <https://webempath.net/lp-eng/>

[MFCC2] Xie, L., & Liu, Z. Q. (2006, August). A comparative study of audio features for audio-to-visual conversion in mpeg-4 compliant facial animation. In *2006 International Conference on Machine Learning and Cybernetics* (pp. 4359-4364). IEEE.

[filterbank] Yusnita, M. A., Paulraj, M. P., Yaacob, S., Yusuf, R., & Shahriman, A. B. (2013). Analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English. *International Journal of Automotive and Mechanical Engineering*, *7*, 1053.

# ANEXOS

Sirven para incluir documentación complementaria (planos, circuitos, código, ficheros de configuración, especificaciones técnicas y hojas de características, fichas explicativas, resultado de encuestas, reglamentación y normativas requeridas, etc.).

[PÁGINA INTENCIONADAMENTE EN BLANCO]