UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Una aproximación holística para el reconocimiento de las emociones a través de un clasificador multimodal en español

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Elizo Alonso, Mar

Tutor/a: López García, Aarón

Cotutor/a: Taverner Aparicio, Joaquín José

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

valgrAI

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Una aproximación holística para el reconocimiento de las emociones a través de un clasificador multimodal en español

TRABAJO DE FIN DE MÁSTER

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor: Mar Elizo Alonso

Tutores: Aarón López García
Joaquín José Taverner Aparicio

Curso 2023-2024

Agradecimientos

A mi pareja y a mi familia por brindarme su apoyo incondicional sin importar las circunstancias. A mis amigos por estar en cada momento.

A mis tutores Aarón López García y Joaquín José Taverner Aparicio por su ayuda y asesoramiento que han sido determinantes poder completar este trabajo. A Emilio Vivancos Rubio y Ana María García Fornes por su orientación.

A Valgrai por el apoyo académico que me otorgo y por su distinguida labor.

Resum

Una de les àrees de recerca principals en l'àmbit de la computació afectiva és el reconeixement d'emocions. Els esforços multidisciplinaris s'han enfocat a detectar estats d'ànim considerant diferents modalitats d'aprenentatge i diferents fonts d'informació, com ara l'expressió facial, el to de veu o el tipus de llenguatge emprat. Aquests models, però, no són capaços d'identificar totes les subtileses inherents a l'emoció humana. Per aquest motiu, les darreres línies de recerca se centren en la combinació de les diferents modalitats individuals per generar un únic sistema multimodal. Així, el sistema final redueix les limitacions subjacents de cada canal sensorial.

En aquest treball de fi de màster es desenvolupa un classificador multimodal per al reconeixement d'emocions combinant la informació obtinguda a partir de les expressions facials, la veu i l'anàlisi del llenguatge. Com a proposta, a partir dels models entrenats per a cada font perceptiva, hem desenvolupat una estratègia multimodal basada en tècniques estadístiques d'aprenentatge. La part experimental s'ha realitzat considerant subjectes d'estudi tant en espanyol com en anglès.

Els resultats obtinguts ens indiquen que la millor estratègia de fusió, basada en un perceptró multicapa, millora un 4,35% la precisió global de cadascuna de les modalitats. On, a més, presenta una consistència major a l'hora de reconèixer les diferents emocions. Aquests resultats suposen un pas cap al reconeixement de les emocions mitjançant tècniques multimodals.

Paraules clau: Emocions, multimodal, fusió, agreagació, imatge, audi, text

Resumen

Una de las principales áreas de investigación en el ámbito de la computación afectiva es el reconocimiento de emociones. Los esfuerzos multidisciplinares se han enfocado en detectar estados de ánimo considerando diferentes modalidades de aprendizaje y diferentes fuentes de información como pueden ser la expresión facial, el tono de voz o el tipo de lenguaje empleado. No obstante, estos modelos no son capaces de identificar todas las sutilezas inherentes a la emoción humana. Por este motivo, las últimas líneas de investigación se centran en la combinación de las distintas modalidades individuales para generar un único sistema multimodal. De esta manera, el sistema final reduce las limitaciones subyacentes de cada canal sensorial.

En este trabajo fin de máster se desarrolla un clasificador multimodal para el reconocimiento de emociones combinando la información obtenida a partir de las expresiones faciales, la voz y el análisis del lenguaje. Como propuesta, a partir de los modelos entrenados para cada fuente perceptiva, hemos desarrollado una estrategia multimodal basada en técnicas estadísticas de aprendizaje. La parte experimental se ha realizado considerando sujetos de estudio tanto en español como en inglés.

Los resultados obtenidos nos indican que la mejor estrategia de fusión, basada en un perceptrón multicapa, mejora un 4,35 % la precisión global de cada una de las modalidades. Donde además, presenta una mayor consistencia a la hora de reconocer las distintas emociones. Estos resultados suponen un paso hacia el reconocimiento de emociones mediante técnicas multimodales.

Palabras clave: Emociones, multimodal, fusión, agreagación, imagen, audio, texto

Abstract

One of the main areas of research in the field of affective computing is emotion recognition. Multidisciplinary efforts have focused on detecting moods by considering different learning modalities and different sources of information such as facial expression, tone of voice or the type of language used. However, these models are not able to identify all the subtleties inherent to human emotion. For this reason, the latest lines of research focus on combining the different individual modalities to generate a single multimodal system. In this way, the final system reduces the underlying limitations of each sensory channel.

In this master's thesis, a multimodal classifier for emotion recognition is developed by combining information obtained from facial expressions, voice and speech analysis. As a proposal, from the trained models for each perceptual source, we have developed a multimodal strategy based on statistical learning techniques. The experimental part has been carried out considering study subjects in both Spanish and English.

The results obtained indicate that the best fusion strategy, based on a multilayer perceptron, improves the overall accuracy of each of the modalities by 4.35%. It also shows greater consistency when it comes to recognising different emotions. These results represent a step towards emotion recognition using multimodal techniques.

Key words: Emotions, multimodal, fusion, aggregation, image, audio, text

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

1.1 Motivación

Las emociones juegan un papel fundamental en la evolución humana, ya que facilitan la supervivencia y socialización de los individuos. Un aspecto clave de este papel socializador de las emociones es la correcta identificación de las emociones expresadas por el resto de individuos del grupo social. Aunque los humanos desarrollamos desde muy pequeños la capacidad de reconocer las emociones de los demás en las expresiones faciales, los movimientos corporales, o la voz, no ha sido posible incorporar esta capacidad a los sistemas informáticos hasta la última década. Esto es debido principalmente a que los seres humanos utilizan una gran variedad de medios indirectos y no verbales para transmitir sus emociones [67], y solo la utilización de potentes técnicas de inteligencia artificial para analizar señales fisiológicas, visuales y auditivas ha permitido el reconocimiento automático de emociones con suficiente precisión.

Estas técnicas de reconocimiento de emociones mediante inteligencia artificial pueden ser útiles y beneficiosa en diversos ámbitos, tanto para el uso personal como para servir de soporte a diferentes profesionales. Uno de los campos más destacados es el de la salud mental. Poniendo el foco en el cuidado de las personas vulnerables, permitiendo un seguimiento continuado cuando no están en consulta. También puede ser de ayuda para personas mayores y con necesidades especiales, permitiendo una mejor comunicación y vida en sociedad. Además de la salud mental, hay otros ámbitos donde esta tecnología puede ser beneficiosa. En la educación para mejorar los sistemas de aprendizaje y en el ámbito laboral para selección de candidatos. Finalmente, en el ámbito de la seguridad, un reconocedor de emociones puede identificar emociones extremas con el fin de evitar accidentes o situaciones peligrosas.

El desarrollo de sistemas de reconocimiento emocional que no se centren en un único tipo de percepciones, conocidos como sistemas multimodales, ha permitido incrementar la precisión en el reconocimiento automático de emociones. Un sistema de reconocimiento que analiza diferentes modalidades puede combinar la información extraída de cada una de ellas. Por ejemplo, al integrar datos visuales, auditivos y textuales, el sistema puede clasificar de manera más precisa las emociones, ya que cada tipo de dato aporta información diferente. De esta misma manera, se pueden desarrollar modelos más robustos para operar en entornos y/o condiciones donde una modalidad puede estar comprometida, como por ejemplo en entornos oscuros donde no se puede obtener información visual.

Con respecto a bases de datos multimodales para el reconocimiento de emociones, existe un gran desbalanceo entre idiomas. La disponibilidad de conjuntos de datos en otros idiomas es notablemente limitada en comparación con el inglés [79], lo que es una

limitación significativa en este tipo de estudios si se quieren realizar investigaciones en otros idiomas.

Para finalizar, personalmente me gustaría destacar que el campo del reconocimiento de emociones siempre me ha atraído y me ha parecido muy interesante, ya que, creo firmemente que entender mejor las emociones puede mejorar nuestra calidad de vida. Especialmente al ser aplicado al ámbito de la salud, área que me produce un profundo interés y a la que me gustaría dedicar mis conocimientos en inteligencia artificial como carrera profesional. Por ese motivo, cuando me sugirieron diversos temas en los que realizar mi trabajo fin de master, escogí este sin dudarlo, ya que, encontré la oportunidad perfecta para realizar una investigación que pueda ser de ayuda en este campo. Además, me produjo un especial interesante hacerlo sobre un idioma no tan estudiado.

1.2 Objetivos

El objetivo general de este trabajo es la evaluación de la mejora que produce un sistema de reconocimiento de emociones multimodal que fusione los resultados de tres reconocedores: de emociones faciales, de audio y de texto, frente al funcionamiento aislado de los tres reconocedores. El objetivo, por tanto, no es maximizar la precisión de cada reconocedor unimodal individualmente, sino mejorar el funcionamiento combinado como un único sistema de reconocimiento emocional multimodal.

Para alcanzar este objetivo se proponen los siguientes subobjetivos:

- Definir un reconocedor de emociones para imágenes de expresiones faciales, para audio y para texto:
  - Realizar un análisis extenso de los estudios existentes en la bibliografía para cada modalidad.
  - Entrenar un modelo para cada modalidad que permita clasificar emociones tomando como base los conocimientos extraídos de dichos sistemas.
- Estudiar bases de datos multimodales para el reconocimiento de emociones en español:
  - Identificar y analizar bases de datos disponibles que incluyan los datos multimodales necesarios en español.
  - Evaluar la adecuación de las bases de datos para realizar la agregación.
- Definir el reconocedor de emociones multimodal:
  - Realizar un análisis extenso de como se encuentran estos sistemas en la bibliografía.
  - Implementar un sistema multimodal al fusionar los tres anteriores.
- Evaluar los modelos individuales y multimodales al aplicar métricas representativas.
- Comparación de los resultados entre los modelos individuales y la propuesta final del reconocedor multimodal.

1.3 Estructura del trabajo

En el Capítulo 2 se realiza una revisión del estado del arte sobre el reconocimiento de emociones mediante modelos computacionales tanto en inglés como en español. En primer lugar, se introducen las principales teorías sobre el concepto de la emoción desde la perspectiva de la psicología. Seguidamente, se muestra una revisión de las arquitecturas más utilizadas para el reconocimiento de emociones en imagen, texto y audio. A continuación, se describen los principales trabajos identificados en la literatura con propuestas de arquitecturas multimodales para el reconocimiento de las emociones.

En el Capítulo 3 se desarrolla la propuesta para un clasificador de emociones multimodal. Se realiza una descripción de las arquitecturas que se han implementado así como de las bases de datos que se han utilizado para entrenar los modelos y el modelo empleado para fusionar las salidas de los clasificadores.

En el Capítulo 4 se definen y analizan los resultados finales obtenidos. Se analizan los resultados de los clasificadores individuales y se comparan con los resultados obtenidos al aplicar la técnica de fusión para obtener el clasificador multimodal.

Por último, en el Capítulo 5 se presenta la discusión y las conclusiones. Se detallan las conclusiones que se obtienen del trabajo, además introducir las aportaciones más relevantes. También se reconocen las limitaciones del estudio, lo que permite interpretar las conclusiones de manera realista. Finalmente, se identifican las posibles líneas de investigación para futuros trabajos.

CAPÍTULO 2
Estado del arte

En esta sección se realiza una revisión bibliográfica de los temas abordados en este trabajo. En primer lugar, se introduce el concepto de emoción así como las principales teorías desde una perspectiva psicológica. En segundo lugar, se presenta un análisis del estado actual de los reconocedores de emociones. Además, se revisan diversos estudios clasificados según su modalidad (texto, audio e imagen), así como los reconocedores multimodales. En cada caso se destacan los modelos, corpus y sistemas utilizados, con un enfoque especial en aquellos que emplean datos en español. Esta revisión proporciona el contexto e información necesarios para desarrollar este trabajo de manera precisa, estableciendo así una base sólida de conocimientos.

2.1 Emociones

En la actualidad no existe un consenso absoluto sobre la definición de las emociones. Generalmente, se considera que las emociones son respuestas rápidas a modo de reacción frente a un estímulo o evento determinado [81]. Las emociones son un aspecto clave para entender el comportamiento social humano, tanto a nivel interno como externo [57]. En la literatura es posible encontrar distintas teorías que tratan de dar una explicación al fenómeno de las emociones. En un sentido global, es posible clasificar estas teorías en tres categorías: categóricas, dimensionales y de evaluación. Las teorías categóricas se fundamentan en la posibilidad de establecer un número finito de etiquetas para identificar las distintas emociones de forma universal. Una de las teorías más reconocidas en este ámbito es la teoría de las emociones básicas propuesta por Paul Ekman [15, 16]. Esta teoría se basa en la suposición de la existencia de seis emociones básicas: felicidad, enfado, tristeza, ira, miedo, y asco. Para confirmar su teoría, los estudios de Ekman se enfocaron en cómo estas emociones básicas se podían identificar a partir de las expresiones faciales. + En contraposición a los modelos categóricos, los modelos dimensionales sostienen que las emociones no son universales y que no se determinan por una etiqueta específica, sino que se componen de una serie de características denominadas dimensiones del afecto. En este sentido, uno de los autores más reconocidos es James A. Russell [65]. Russell propuso un modelo bidimensional en el que las emociones se expresan a partir de los niveles de valencia y activación. Este modelo se conoce como el *Circumplex Model of Affect* [64] (ver Figura 2.1).

Siguiendo una filosofía similar, los modelos de evaluación sostienen que las emociones son el resultado de evaluar una serie de características que se conocen como variables de evaluación, que incluyen conceptos como las expectativas o la deseabilidad de un evento para explicar la elicitación de emociones [20, 37].

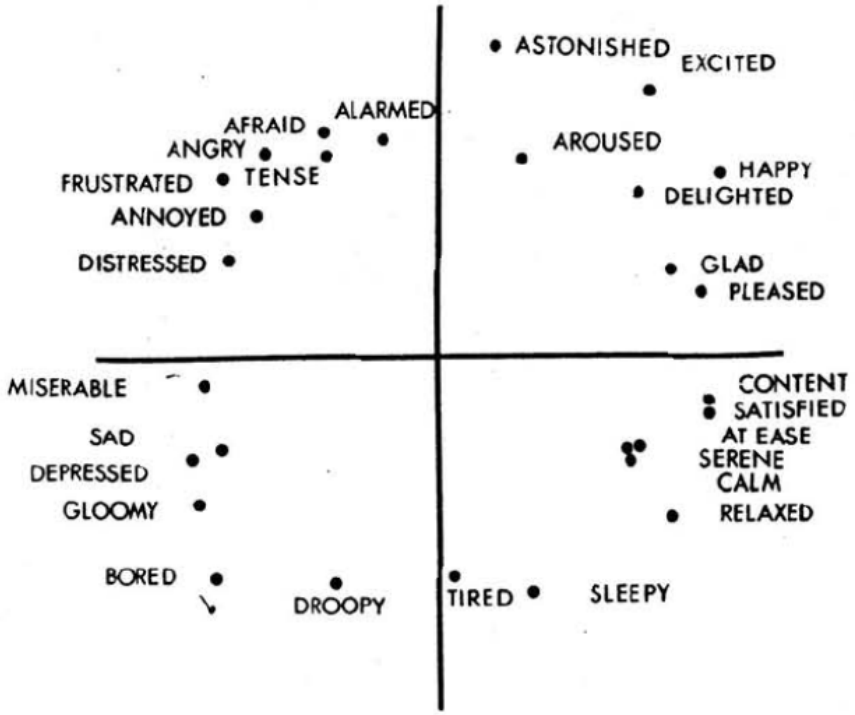

Figura 2.1: Modelo circunflejo del afecto [64]

2.2 Reconocimiento de emociones en inteligencia artificial

El reconocimiento de emociones mediante técnicas de inteligencia artificial es uno de los campos de investigación más populares en la actualidad [67]. Este ámbito de estudio se inscribe dentro de la disciplina conocida como computación afectiva [59]. A lo largo de los años, se han propuesto varios modelos para reconocer las emociones mediante el uso de inteligencia artificial. En [67] se realiza un estudio en el que se analizan más de cien documentos, obteniendo un análisis detallado de las metodologías más utilizadas para la detección de emociones. Los autores identifican cuatro categorías para englobar las propuestas en función de la información utilizada como entrada de los modelos para la detección de emociones. Así tenemos cuatro categorías de clasificadores: faciales, textuales, auditivos y de señales fisiológicas. Los clasificadores faciales son los más estudiados, seguidos por los de audio, los textuales y por último los de señales fisiológicas. Es destacable que los clasificadores faciales son utilizados en más del 50 % de las investigaciones analizadas en el estudio, lo que presenta una gran diferencia en comparación con el resto (ver Figura 2.2).

2.3 Detección de emociones mediante imágenes

2.3.1. Corpus comunes

A continuación se describen los corpus más utilizados en investigaciones sobre el reconocimiento de emociones a partir de imágenes. La mayoría de los corpus analizados incluyen las seis emociones básicas de Ekman junto con el estado neutral:

- **JAFFE** (*Japanese Female Facial Expression*) [44, 45]: es un corpus que incluye 213 imágenes de 10 mujeres japonesas.

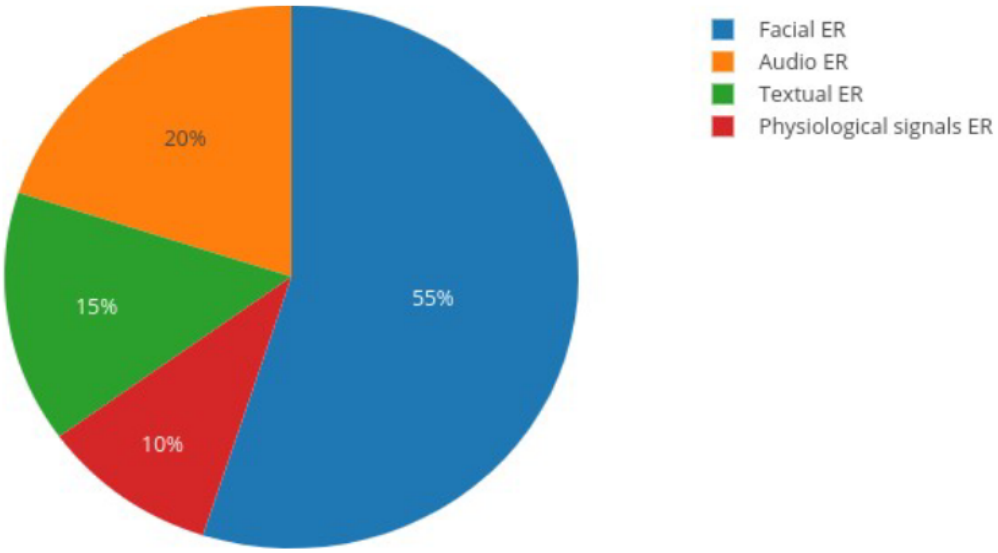

Figura 2.2: Proporción de tipos de datos utilizados en reconocedores de emociones [67]

- **KDEF** (*Karolinska Directed Emotional Faces*) [43]: esta base de datos se crea a través de 70 individuos los cuales representan las emociones desde 5 ángulos diferentes (de frente, a 45 grados y completamente de perfil hacia cada lado). Para cada actor y cada ángulo se recogen dos muestras distintas, haciendo un total de imágenes 4900 imágenes.

- **EDFFE** (*Emotion Detection From Facial Expressions*) [63]: es una competición de Kaggle que recoge 13690 imágenes de diferentes fuentes. En este corpus añaden la emoción de desprecio al conjunto de etiquetas.

- **Yale** (*The Extended Yale Face Database B* [83]): es una base de datos que solo tiene imágenes de la clase neutral, son 16128 en total, tomadas de 28 actores. Esta base de datos se recoge con la idea de proporcionar variabilidad en la iluminación al aplicar 64 condiciones de iluminación diferentes.

- **FER 2013** (*Facial Expression Recognition Challenge*) [14]: se trata de una competición que consta de 35888 imágenes. Las imágenes fueron etiquetadas mediante un proceso automático en el que se usó la API de búsqueda de Google. Posteriormente, Microsoft reetiquetó esta base de datos con 10 anotadores humanos y la denomino como FERPlus [2].

- **CK+** (*Extended Cohn-Kanade dataset*) [42]: Contiene 593 secuencias de video de 123 personas diferentes. De estos videos, 327 están etiquetados con emociones. En este caso no se incluye la emoción neutral pero si se incluye la emoción desprecio.

- **AffectNet** [51]: es un conjunto con aproximadamente 400000 imágenes clasificadas manualmente por anotadores humanos. Este conjunto también incluye la emoción de desprecio.

- **RAF-DB** (*Real-world Affective Faces Database*) [38, 39]: se compone de 30000 imágenes etiquetadas de forma independiente por aproximadamente 40 anotadores. Está divida en dos subconjuntos: un subconjunto que clasifica las emociones con una etiqueta única y un subconjunto en el que las imágenes se clasifican con dos emociones, representando emociones más complejas.

- **ISED** (*Indian Spontaneous Expression Database*) [27]: se trata de una base de datos de 428 videos clasificadas en cuatro clases, las cuales son felicidad, sorpresa, tristeza y asco.

2.3.2. Reconocedor de emociones con imagen

Analizando la bibliografía sobre el reconocimiento de emociones en imagen, es posible observar que la mayoría de los estudios se basan normalmente en redes convoluciones y, más recientemente, en arquitecturas tipo "*transformers*" [78]. Las redes neuronales convolucionales son un tipo particular de red neuronal en la que los datos se estructuran mediante matrices o cuadrículas, lo cual permite trabajar de forma más eficiente con diversos tipos de datos como imágenes o series temporales. Por otro lado, los "*transformers*" son un tipo de arquitectura de red neuronal que se basan en la auto-atención para capturar las relaciones entre diferentes partes de la secuencia de entrada. Los trabajos más recientes están introduciendo el uso de los "*Vision Transformers*", que son un tipo especial de "*transformers*" especializados en tareas de visión computacional [13]. Para ello, las imágenes se dividen en parches con el fin de poder ser tratadas como secuencias. Por ejemplo, en el estudio realizado en [7] los investigadores presentan diferentes modelos para el reconocimiento de emociones en imagen. Entre esos modelos, los autores presentan un modelo denominado ViT-B/16/SAM en el que utilizan "*Vision Transformers*". El modelo combina estos transformers con un optimizador "*Sharpness-Aware Minimizer*" [19] para aplicar una mejor generalización en el proceso de reconocimiento de la emoción en el rostro. Además, aplicaron una disminución gradual del factor de aprendizaje y un "*momentum*" de 0,9 como mecanismo de optimación de la red.

En ese mismo trabajo, los autores proponen tres modelos más:

- Un modelo "*baseline*" en el que se utiliza una red *ResNet-18* con decrecimiento gradual del factor de aprendizaje y "*momentum*" de 0,9 con un optimizador SGD [62], cuyas siglas se refieren a "*Stochastic Gradient Descent*". Esta se utiliza como para ir analizando el resto de modelos.
- El modelo ViT-B/16/S en el que también utilizan un "*Vision Transformer*" básico pre-entrenado con 16 unidades de "*embeddings*" para los parches de las imágenes y adaptación de la capa lineal final para el reconocimiento de la emoción. Los autores aplicaron en este modelo el optimizador SGD con factor de aprendizaje de 0,001 y "*momentum*" de 0,9.
- El modelo ViT-B/16/SG que es igual que el anterior, pero cambia la configuración de la tasa de aprendizaje.

Para entrenar estos modelos se utilizaron las bases de datos FER-2013, CK+ y AffectNet. Un resumen de los resultados obtenidos con los diferentes modelos puede verse en la Tabla 2.1. Como puede verse, el resultado más alto de precisión en el conjunto de validación se encuentra al entrenar con 10 épocas el modelo ViT-B/16/SG con un 56,30 % seguido por el 56,22 % obtenido de entrenar durante 25 épocas el modelo ViT-B/16/SAM. Aunque en este último modelo se obtiene una perdida en el conjunto de validación de 12,93 %, que es el tercero mejor de todas las pruebas, frente al 15,05 % obtenido del modelo con el mejor rendimiento en términos de precisión.

Otro trabajo interesante es presentado en [46]. Los autores implementan una arquitectura para la detección de emociones en imagen que combina el uso dos *ResNet18* [29] y con un "*Multi-Layer Transformer Encoder*" (ver Figura 2.3). Inicialmente se pasan las imágenes de entrada hasta la última capa de convolución de las dos *ResNet18*. Cada *ResNet18*

Tabla 2.1: Resultados obtenidos sobre las 6 emociones básicas y la neutral del sistema propuesto en [7]

Nº etapas	Métricas de ajuste	ResNet-18	ViT-B/16/S	ViT-B/16/SG	ViT-B/16/SAM
5	Precisión entrenamietno	61,38	73,99	74,41	83,88
	Precisión validación	50,43	55,42	55,81	55,14
	Pérdida entrenamietno	10,50	7,09	7,12	7,15
	Pérdida validación	14,70	12,61	13,35	12,82
10	Precisión entrenamietno	66,57	81,42	84,01	87,72
	Precisión validación	52,16	55,71	56,30	55,24
	Pérdida entrenamietno	8,90	5,06	4.,67	6,24
	Pérdida validación	14,90	14,33	15,05	13,39
15	Precisión entrenamietno	68,01	79,07	87,79	87,88
	Precisión validación	52,26	56,01	54,94	55,29
	Pérdida entrenamietno	8,49	4,12	3,63	6,27
	Pérdida validación	14,80	19,47	17,58	13,16
20	Precisión entrenamietno	70,01	81,14	88,37	88,18
	Precisión validación	53,85	54,28	54,65	55,45
	Pérdida entrenamietno	8,35	5,14	3,41	6,20
	Pérdida validación	14,20	15,03	18,76	13,08
25	Precisión entrenamietno	70,02	84,28	88,66	88,64
	Precisión validación	53,89	53,98	55,52	56,22
	Pérdida entrenamietno	8,27	3,97	3,63	6,05
	Pérdida validación	14,39	19,93	17,87	12,93

incluye bloques residuales para hacer frente al desvanecimiento del gradiente y están preentrenadas para extraer características de la imagen. Se utiliza una de ellas para imágenes en blanco y negro y la otra para a color. Luego se agregan los mapas de características resultantes de ambas a través de la capa denominada "*Attentional Selective Fusion*" (ver Figura 2.4). A las características obtenidas se aplica un "*Multi-Layer Transformer Encoder*", esta es la parte del "*transformer*" que se utiliza para, generar representaciones codificadas de las entradas. Finalmente, se utiliza una red completamente conectada para acabar obteniendo las predicciones del modelo. Los resultados que se obtienen, son de 64,80 % de precisión en AffectNet-7 (es decir, sin desprecio), un 88,81 % en FERPlus y un 88,14 % en RAF-DB. Finalmente, aplicaron un estudio de evaluación en el que los resultados oscilan entre 81,88 % y 86,24 % de precisión dependiendo del conjunto de entrenamiento utilizado (ver Tabla 2.2).

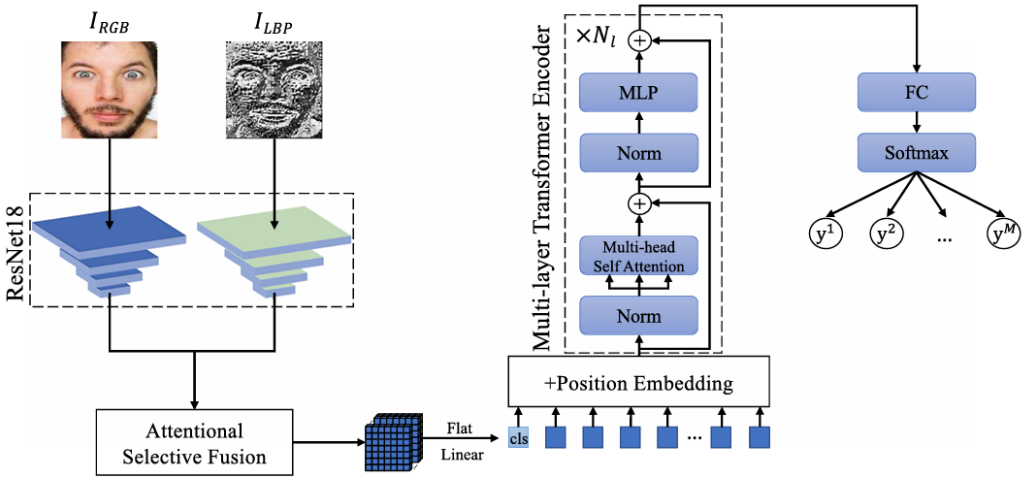

Figura 2.3: Arquitectura propuesta en [46]

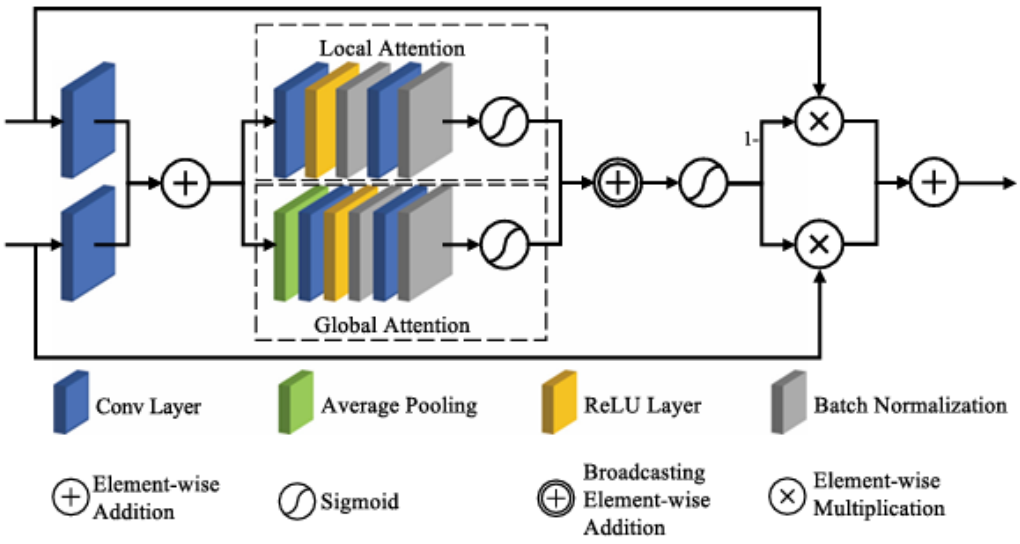

Figura 2.4: "*Attentional Selective Fusion*" propuesta en [46]

Tabla 2.2: Resultados del estudio de evaluación entre bases de datos del trabajo propuesto en [46]

Método	Entrenamiento	Test	Precisión
Nuestro	RAF-DB	CK+	81,88
Nuestro	FERPlus	CK+	83,79
Nuestro	AffectNet-8	CK+	86,24

En [76] los autores proponen una arquitectura que denominan como *EmotionalDAN* la cual se basa en una "*Deep Alignment Network*" que facilita una alineación robusta de rostros [35]. Los autores añadieron además una para los puntos de referencia característicos del rostro, y otra para la clasificación de la emoción (ver Tabla 2.3). La parte de clasificación por emoción en siete clases se ha entrenado con *AffectNet* y se ha probado en las bases de datos CK+, JAFFE y ISED. Los resultados obtenidos en cada uno de los conjuntos probados son de 0,736, 0,465 y 0,62 de precisión respectivamente.

En el siguiente estudio presentado en [33], se propone un modelo tipo *VGGNet*, que es una red convolucional con una arquitectura profunda (ver Figura 2.5). Tras realizar diversos experimentos, los autores concluyeron que se utilizaría el optimizador SGD con "*Nesterov momentum*" (una variante del algoritmo de "*momentum*"), y planificador del factor de aprendizaje "*Reduce Learning Rate on Plateau*" (RLRP). El algoritmo RLRP es una técnica de ajuste del factor de aprendizaje que modifica este valor durante el entrenamiento. El modelo y se entrenó durante 50 épocas más aplicando "*Cosine Annealing*", que permite ajustar la tasa de aprendizaje respecto a una función coseno. El modelo se entrena con la base de datos *FER2013*. Los resultados muestran un precisión de 73,23 %.

Para finalizar, la investigación realizada en [31] aplica un proceso de detección de características en el que se utiliza la desviación local para resaltar áreas faciales importantes para el reconocimiento de emociones. El modelo es tipo *ResNet* (ver Tabla 2.4). Los autores además decidieron incluir una capa de "*global average pooling*" y capas de "*batch normalization*" para mejorar la generalización del modelo y acelerar el proceso de entrenamiento. Aplicaron también diferentes hiperparámetros como el optimizador SGD con "*weight decay*" de 1E-5 para tratar de prevenir el sobreajuste. El factor de aprendizaje al principio fue establecido en 4e-2, y fue disminuyendo en 0,005 cada 10 épocas. El resul-

Tabla 2.3: Arquitectura propuesta en [76]

Nombre	Tamaño de entrada	Tamaño de salida	Kernel
conv1a	224×224×1	224×224×64	3×3, 1, 1
conv1b	224×224×64	224×224×64	3×3, 64, 1
pool1	224×224×64	112×112×64	2×2, 1, 2
conv2a	112×112×64	112×112×128	3×3, 64, 1
conv2b	112×112×128	112×112×128	3×3, 128, 1
pool2	112×112×128	56×56×128	2×2, 1, 2
conv3a	56×56×128	56×56×256	3×3, 128, 1
conv3b	56×56×256	56×56×256	3×3, 256, 1
pool3	56×56×256	28×28×256	2×2, 1, 2
conv4a	28×28×256	28×28×512	3×3, 256, 1
conv4b	28×28×512	28×28×512	3×3, 512, 1
pool4	28×28×512	14×14×512	2×2, 1, 2
fc1	14×14×512	1×1×256	-
fc2 landmark	1×1×256	1×1×136	-
fc2 emotion	1×1×256	1×1×7	-

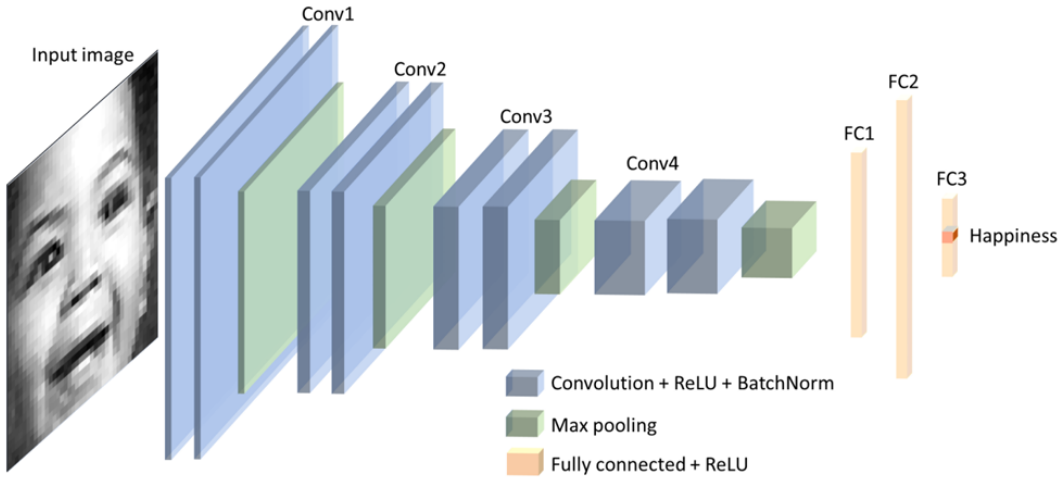

Figura 2.5: Arquitectura propuesta en [33]

tado de esta implementación es de un precisión de 95,23 % en *JAFFE* y 93,24 % en *CK+*. En la Tabla 2.5 ver la división de los conjuntos utilizados.

2.4 Detección de emociones mediante audio

Para el reconocimiento de emociones en el audio se utiliza el procesado de la señal acústica. En concreto se analizan dos tipos de técnicas para la extracción de características. Por un lado, los "*Modulation spectral features*" (MSFs) [26], representan el contenido de modulación en la señal de audio, esta información corresponde a los cambios de la señal a lo largo del tiempo. Por otro lado, el "*Mel-frequency cepstrum coefficient*" (MFCC) [11], facilita una representación del espectro de la potencia del audio basado en la escala de Mel [72]. Esta es una escala de frecuencia perceptual, la idea es que modele como los humanos perciben los cambios en la frecuencia del sonido.

Tabla 2.4: Arquitectura propuesta en [31]

Tipo	Tamaño de filtro / stride	Tamaño de salida
Conv1	5 × 5 / 2	64 × 48 × 32
Maxpool1	2 × 2 / 2	32 × 24 × 32
Conv2	3 × 3 / 1	32 × 24 × 64
Res1	4 Conv	–
Conv3	3 × 3 / 1	32 × 24 × 128
Maxpool2	2 × 2 / 2	16 × 12 × 128
Conv4	3 × 3 / 1	16 × 12 × 128
Res2	4 Conv	–
Conv5	3 × 3 / 1	16 × 12 × 256
Maxpool3	2 × 2 / 2	8 × 6 × 256
Conv6	3 × 3 / 1	8 × 6 × 512
FC1	1024	1024
FC2	512	512

Tabla 2.5: Conjuntos de datos utilizados en [31]

Base	Número de imágenes	Base de datos
Training	8000	CK+
Training	200	JAFFE
Testing	150	CK+
Testing	13	JAFFE

2.4.1. Corpus comunes

A continuación se muestra un resumen de los principales corpus utilizados para la clasificación de emociones en audio:

- **eNTERFACE'05** (*eNTERFACE'05 EMOTION Database*) [49,50]: es un corpus audio-visual, para él se utilizan 42 personas de 14 nacionalidades diferentes. Cada uno escucha 6 historias que representan cada una de las emociones básicas y luego reaccionan a estas en inglés. El corpus fue evaluado por dos anotadores expertos que estudiaron las reacciones que presentaron los participantes y descartaron aquellas que no cuadraban con el contenido mostrado.

- **EMO-DB** (*Berlin Database of Emotional Speech*) [5]: Contiene unos 500 frases en aleman con entonaciones que representan las emociones: felicidad, enfado, tristeza, miedo, aburrimiento, asco, y una neutral.

- **RAVDESS** (*The Ryerson Audio-Visual Database of Emotional Speech and Song* [40]: contiene 7356 videos en inglés realizados por 24 actores profesionales. Los vídeos están etiquetados con las seis emociones básicas más la emoción de calma y el estado neutral.

- **SAVEE** (*Surrey Audio-Visual Expressed Emotion*) [54]: es un corpus audio-visual en el que las grabaciones se realizan en inglés. Utilizaron 4 actores y 408 enunciados distintos. El contenido está etiquetado con las seis emociones básicas y el estado neutral.

- **TESS** (*Toronto emotional speech set*) [55]: en esta base de datos con audios en inglés, dos actrices graban un conjunto de 200 palabras. Cada palabra fue entonada imitando las seis emociones básicas y el estado neutral.

2.4.2. Reconocedor de emociones con audio

En la literatura sobre el reconocimiento de emociones a través de audio, se aplican diversos modelos y técnicas de procesamientos de la señal de audio. Una de las técnicas más utilizadas es la conversión de la señal a imagen para analizar el espectrograma. Por ejemplo, en [18] se propone una arquitectura de una "*Convolutional Neural Network*" (CNN) 3D. Los resultados en términos de precisión promedio de reconocimiento de emociones, utilizando una estrategia independiente del hablante, es de un 90,40 % en *EMO-DB* y 83,20 % en *eINTERFACE05*. Los autores también realizaron un análisis de los resultados al incluir en los modelos la dimensión de género, obteniendo una precisión promedio de 94,42 % en *EMO-DB* y 88,47 % en *eINTERFACE05*.

Otros trabajos también aplican modelos para el procesamiento de secuencias. Por ejemplo, en [70] los autores proponen una arquitectura que incluye capas LSTM ("*Long Short-Term Memory*") [30]. Esta red es un tipo de RNN "*Recurrent neural networks*" que aborda los problemas de dependencia a largo plazo y del desvanecimiento del gradiente. Esta LSTM se combina con un módulo de atención y una CNN 2D. Este modelo se prueba para diferentes conjuntos. La Tabla 2.6 muestra un resumen de los resultados que obtuvieron para los distintos conjuntos de datos. Como puede verse, el valor de precisión oscila entre un 57,50 % y un 99,81 % según el conjunto de datos utilizado.

Tabla 2.6: Resultados de la arquitectura en [70]

Método/Modelo	Corpus usado	Precisión
LSTM + Attention + CNN-2D	RAVDESS	74,44 %
LSTM + Attention + CNN-2D	SAVEE	57,50 %
LSTM + Attention + CNN-2D	TESS	99,81 %
LSTM + Attention + CNN-2D	Customized (RAVDESS + SAVEE + TESS)	90,19 %

2.4.3. Corpus comunes en español

A continuación se presenta un resumen de los principales corpus que existen en el idioma español para la clasificación de emociones en audio:

- **INTER1SP** (*Spanish Emotional speech synthesis database*) [6]: contiene grabaciones de dos actores: un hombre y una mujer. Los actores grabaron el mismo texto entonando para simular las seis emociones básicas más el estado neutral. También grabaron diferentes velocidades y entonaciones en la pronunciación del texto: rápido, lento, suave y fuerte. En conjunto, las grabaciones son un total de 3 horas y 59 minutos de audio grabado para el masculino y 3 horas y 53 minutos para el femenino.

- **EmoMatchSpanishDB** ([21]: contiene 2005 audios interpretados por 50 actores. Los actores reprodujeron 12 oraciones entonadas para simular las seis emociones diferentes y el estado neutral. A los audios se les aplicó un proceso de "*crowdsourcing*" para validar la emoción anotada.

- **EmoFilm** (*Emotional speech from Films*) [56]: es un subconjunto de 1115 audios de 43 películas inicialmente etiquetadas en inglés, pero se recogen los audios doblados tanto en italiano como en español. Los audios están clasificados en enfado, felicidad, tristeza, miedo y desprecio.

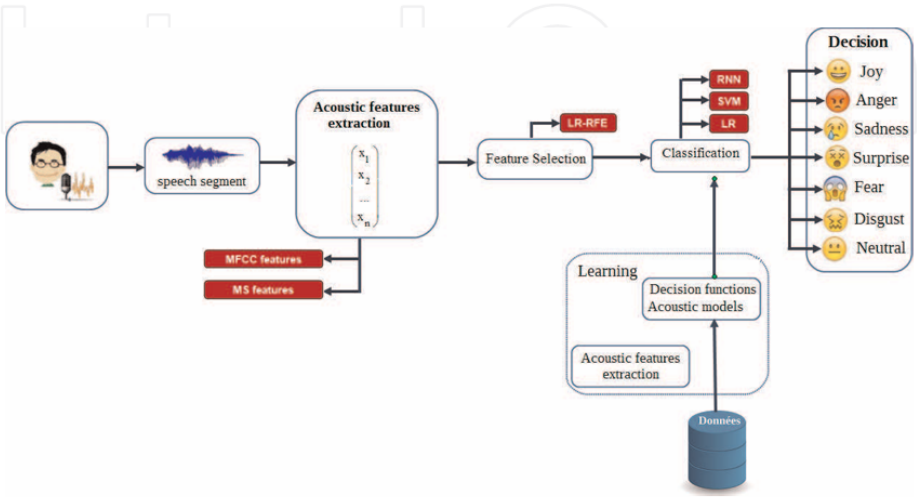

Figura 2.6: Arquitectura propuesta en [32]

2.4.4. Reconocedor de emociones con audio en español

Una vez analizada la literatura sobre las principales contribuciones del estado del arte del reconocimiento de emociones en audio en otros idiomas, realizamos una revisión de las principales propuestas para el idioma español. En primer lugar, en [32] los autores proponen un modelo en el que realizan un preproceso de extracción de características de audio. Para ello, evalúan la posibilidad de utilizar tanto los MSFs como los MFCC como una combinación de ambos. La selección de características se realiza mediante el uso de un "*Linear Regression - Recursive feature elimination*" (LR-RFE), que elimina recursivamente las características menos importantes utilizando regresión lineal. Las características seleccionadas se aplican sobre modelos de "*Multivariate linear regression classification*" (MLR), "*Support Vector Machines*" (SVM) [9] y "*Recurrent neural networks*" (RNN), (ver Figura 2.6). La RNN propuesta consiste de dos capas LSTM consecutivas con activación de tangente hiperbólica, seguidas por dos capas densas. Las características de los datos se escalan con normalización min-max. Este estudio se realiza para alemán y español, la base de datos en español es *INTER1SP*. Los resultados se muestran en la 2.7. Como puede verse, para español, el mejor resultado se obtiene de la aplicación de la RNN, sin aplicar normalización por hablante, concretamente con un 94,01 % de precisión.

Tabla 2.7: Resultados de la combinación de MSFs y MFCC en [32]

SN Clasificador	LR-RFE	Berlin	Español
MLR	No	73,00 (3,23)	83,55 (0,55)
	Yes	79,40 (3,09)	84,19 (0,96)
SVM	No	81,10 (2,73)	89,69 (0,62)
	Yes	80,90 (3,17)	90,05 (0,80)
RNN	No	63,67 (7,74)	90,05 (1,64)
	Yes	78,11 (3,53)	94,01 (0,76)

Otro trabajo interesante podemos encontrarlo en [17]. En este caso los autores utilizan dos corpus: *Emofilm* e *INTERS1P*. Para la clasificación de emociones proponen un modelo en el que se analizan los espectrogramas en escala de Mel, aplicando un tamaño de 256x256. La arquitectura de la red combina una red neuronal convolucional con una capa LSTM. Para incrementar el número de muestras, se aplicaron técnicas de aumentación de datos a las imágenes de los espectrogramas. Los resultados de este modelo se muestran en la Tabla 2.8. Como puede verse, los mejores resultados se obtienen utilizando única-

mente la red convolucional con un precisión de 92,53 % frente a un 82,62 % al combinarla con la LSTM.

Tabla 2.8: Resultados de la arquitectura propuesta en [17]

Modelo	Precisión en el set de prueba
CNN	92,53 %
CNN+LSTM	82,62 %

2.5 Detección de emociones mediante texto

A continuación, se realiza un análisis de los principales trabajos identificados en la literatura que proponen modelos para el reconocimiento de emociones en texto. Es necesario destacar aquí que, tradicionalmente, el reconocimiento de emociones en texto se ha enfocado hacia el análisis de la polaridad, es decir, la clasificación de los textos en categorías de sentimiento positivo, negativo o neutral [60, 74]. Sin embargo, los trabajos más recientes han empezado a explorar técnicas para clasificar las emociones básicas a partir de textos [25, 66].

2.5.1. Corpus comunes

Los principales corpus utilizados en la literatura para la clasificación de emociones en texto son:

- Airlines (*Airline Sentiment*) [10]: consiste en un conjunto de 14640 *tweets* clasificadas en función de la polaridad.
- CrowdFlower (*Apple Sentiment*) [71]: consta de 3804 *tweets* etiquetados con la polaridad.
- Apple texts [69]: se compone de 1630 *tweets* etiquetados con la polaridad.

2.5.2. Reconocedor de emociones con texto

En la literatura sobre el reconocimiento de emociones a través de texto, se aplican diversos modelos y técnicas de procesamientos texto, es decir métodos de procesamiento del lenguaje natural (NLP, por sus siglas en inglés). Por ejemplo, en [74], se propone un modelo para la clasificación de la polaridad en el texto mediante el uso de modelos tipo Bert. El modelo Bert es un modelo de lenguaje desarrollado por Google que fue muy relevante para el contexto del procesamiento de lenguaje natural [12]. Concretamente en este estudio se utilizan RoBERTa y DistilBERT pre-entrenados. A estos modelos se aplican combinaciones de BiLSTMs (Bidirectional Long Short-Term Memory), es decir, un LSTM bidireccional, y BiGRUs (Bidirectional Gated Recurrent Unit), un GRU bidireccional. Las GRU son una variante de las LSTM con una estructura más simple. (ver Figura 2.7). En cuanto a los resultados obtenidos, la Tabla 2.9 muestra un resumen para los diferentes modelos propuestos. Como puede verse, en el corpus Airlines los mejores resultados son los obtenidos con el modelo RoBERTa-3G (3xBiGRU) cuando no se eliminan los emojis en el texto alcanzando un 86 % de precisión. En CrowdFlower el mejor es RoBERTa-3L (3xBiLSTM) sin eliminar emojis con un 82,26 %. Finalmente, para Apple texts el resultado más alto se obtiene con RoBERTa sin combinación, este es de 91,72 % precisión tanto al eliminar o no los emojis.

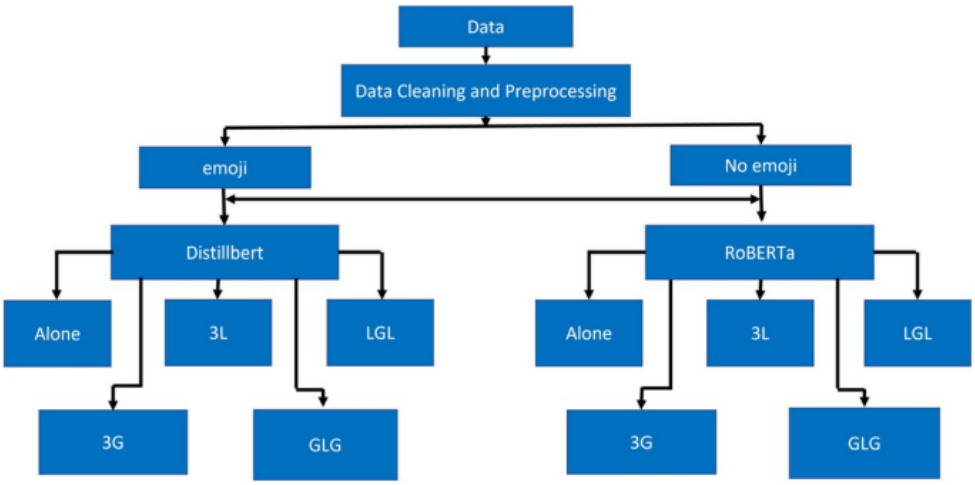

Figura 2.7: Arquitecturas propuesta en [74]

Tabla 2.9: Resultados de la arquitectura propuesta en [74]

Modelo		Con Emojis			Sin Emojis		
		Airlines	CrowdFlower	Apple	Airlines	CrowdFlower	Apple
DistilBERT	-	83,5	78,71	84,97	83,4	77,92	86,2
	3G	81,8	76,48	83,74	82,17	76,61	87,12
	3L	81,9	74,9	83,13	81,83	76,74	82,82
	GLG	83,74	80,42	85,89	83,47	79,24	88,04
	LGL	82,55	78,71	86,81	83,27	78,98	87,42
RoBERTa	-	85,72	82,39	91,72	85,69	79,63	90,18
	3G	86	79,63	91,1	85,72	79,63	91,72
	3L	85,66	82,26	90,18	85,66	81,34	89,57
	GLG	85,28	81,21	91,41	85,93	80,55	90,18
	LGL	85,52	81,34	89,88	84,97	80,16	90,49

En [66] los autores crean de manera independiente una base de datos centrada en *tweets*. Proponen además una arquitectura donde primero se preprocesa el texto para después aplicar dos modelos, uno para una clasificación en 3 clases y otro para 7 que incluye las seis emociones básicas y el estado neutral). Tras ellos se calcula el valor de influencia del "*tweet*" (ver Figura 2.8). Con respecto a la clasificación en siete clases, el mejor valor se obtiene de la aplicación del algoritmo "*Naive Bayes*", este es un algoritmo de clasificación probabilística basado en la aplicación del teorema de Bayes, este es de un 47,34 % deprecisión.

2.5.3. Corpus comunes en español

Se ha realizado una revisión del estado del arte sobre los principales corpus utilizados para la clasificación de las emociones en textos en español. Las bases de datos identificadas se engloban dentro del "*Workshop on Semantic Analysis at SEPLN*" (TASS). Esta competición se celebra desde 2012 anualmente con el objetivo principal de avanzar en la investigación sobre el análisis de sentimientos en español y también se ha ampliado para abarcar diversas tareas relacionadas con el análisis semántico en español [68]. Anualmente se presentan diferentes tareas, por lo que de esta competición los investigadores pueden extraer diferentes corpus.

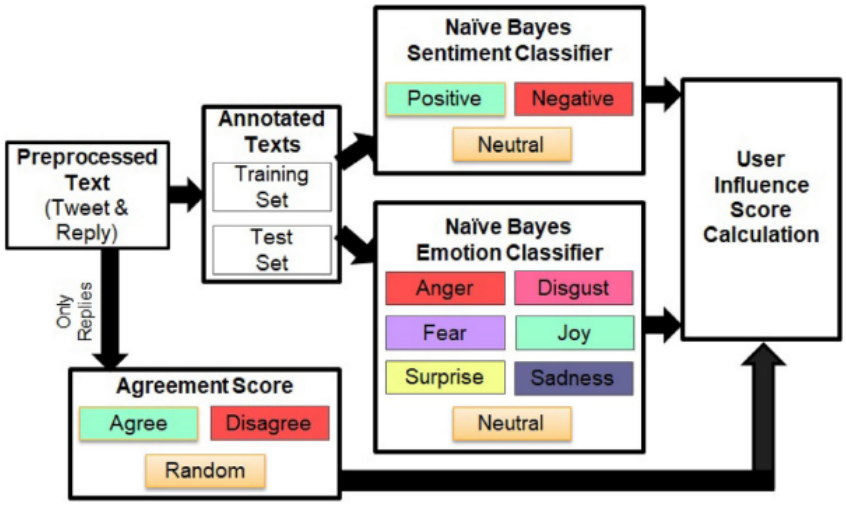

Figura 2.8: Arquitectura propuesta en [66]

2.5.4. Reconocedor de emociones con texto en español

Una vez analizada la literatura sobre las principales contribuciones del estado del arte del reconocimiento de emociones en texto en otros idiomas, a continuación realizamos una revisión de las principales propuestas para el idioma español. Por ejemplo, en [60], se presenta un estudio resultado de la competición TASS del año 2013 [75]. En esta competición se utilizó un corpus que consistía en un conjunto de *tweets* etiquetados con la polaridad. En este corpus propusieron dos tipos de polaridad: *3-level* con positivo, negativo y neutral; y *5-level* con negativo, muy negativo, neutral, positivo y muy positivo. El modelo presentado aplica inicialmente un preprocesado y posteriormente un extractor de características. La salida de estos procesos se une con un clasificador de polaridad para realizar la clasificación. Se implementan diferentes modelos basados en "*Support Vector Machine*" (SVM). El resultado final obtenido del sistema en la competición es de 57,60 % en "*5-level*" y 67,40 % en "*3-level*" de precisión.

En el TASS de 2020 [68] se introdujo una tarea de texto a partir de *tweets*, en la que se debían clasificar las seis emociones básicas más el estado neutral. Los resultados de la competición se recogen en [23]. El grupo que ganó la competición obtuvo un "*F1-score*" de 0,447 y el segundo 0,379 (ver Tabla 2.10). El equipo ganador de la competición, utilizo Bert adaptando el modelo a *tweets* en el idioma de español. Para ello, propusieron un modelo basado en TWilBERT como "*framework*" para el entrenamiento, evaluación y fine-tune de modelos basados en Bert. Tras diferentes experimentos, los autores decidieron utilizar *TWilBERT-large* con una tasa de aprendizaje de 10^{-5} y lotes de 32 muestras sin acumulación de gradientes. Como curiosidad, el segundo equipo en el ranking [22], aplicó un algoritmo de optimización mínimo secuencial que se basa en SVM. La siguiente tabla presenta los resultados considerando precisión y exhaustividad (recall).

Tabla 2.10: Resultados TASS2020 [23]

Equipo	F1-score	Precision	Exhaustividad
ELiRF-UPV	0,447	0,443	0,450
UMUTeam	0,379	0,420	0,345

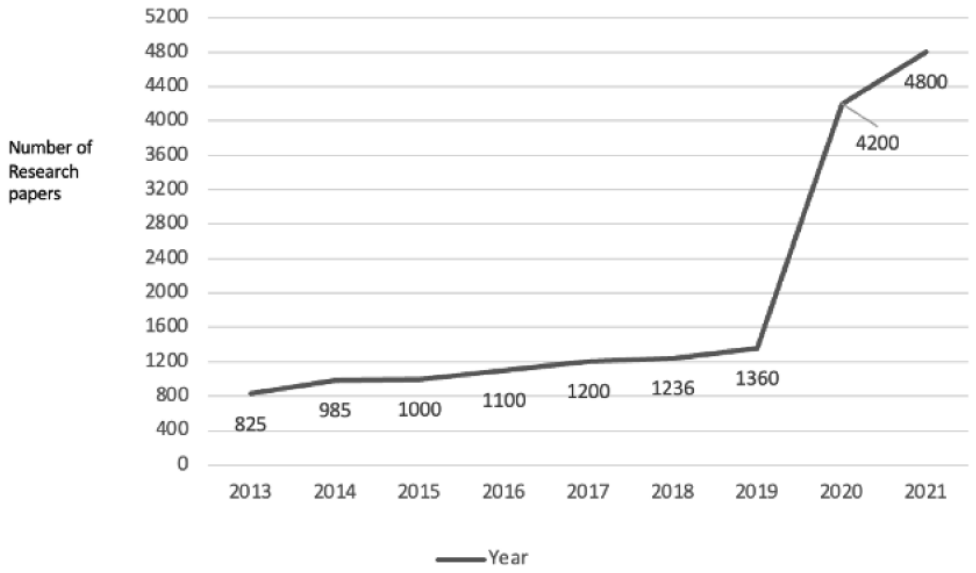

Figura 2.9: Evolución de investigaciones de reconocimiento multimodal [1]

2.6 Detección de emociones mediante técnicas multimodales

El enfoque multimodal se refiere a la combinación de diferentes tipos de datos para enriquecer las entradas que los clasificadores toman con el objetivo de mejorar su precisión. En el contexto del reconocimiento de las emociones, este enfoque implica el uso simultáneo de múltiples canales de información que incluyen elementos verbales y no verbales, como señales de audio, video, expresiones faciales, gestos, lenguaje corporal y datos fisiológicos (por ejemplo, ritmo cardíaco o actividad cerebral) [67]. Integrar estos diversos tipos de datos en un clasificador permite capturar una representación más completa y precisa del estado emocional de una persona. Sin embargo, el enfoque multimodal supone un aumento en la complejidad del modelo, ya que, para que el sistema de reconocimiento emocional funcione eficazmente, es crucial que los datos de las diferentes entradas estén alineados temporalmente. Además, es necesario utilizar algoritmos de fusión de datos capaces de integrar esta información de manera coherente.

El desarrollo de clasificadores multimodales para el reconocimiento de emociones muestra una tendencia al alza en las investigaciones realizadas en los últimos años. El trabajo presentado en [1] evidencia esta tendencia tras analizar la literatura entre los años 2013–2021 sobre las propuestas multimodales para la clasificación de emociones. La Figura 2.9 muestra el número de propuestas que utilizan clasificadores multimodales para el reconocimiento de emociones. Como puede verse, en el año 2019 se produjo un incremento considerable en el número de propuestas que siguió aumentando en los años siguientes.

2.6.1. Corpus comunes

Los corpus comunes para este tipo de sistemas se explican a continuación:

- **CMU-MOSEI** [86]: Es un corpus en inglés formado por 23453 clips extraídos de videos de YouTube.
- **CMU-MOSI** [85]: Es un corpus también en inglés formado por 2199 clips extraídos de videos de YouTube.

Tabla 2.11: Comparación estado del arte multimodal de [1]

Autores	Taxonomía empleada	Método
Alswaidan & Menai (2020)	Text emotion recognition methods	Rule-based, classical learning, deep learning, and hybrid.
Abdullah et al. (2021)	Multimodal emotion recognition methods	Neural network architecture and deep learning technique.
Wu et al. (2014)	Audiovisual emotion recognition techniques	Deep learning, datasets (GEMEP, RML, and VAM) are explained.
Baltrušaitis et al. (2018)	Technical challenges faced by multimodal researchers	Different co-learning methods.
Jam et al. (2021)	Social signals and emotional expressions in real-world human-robot interactions.	Deep learning methods.
Lim et al. (2020)	Taxonomy involves Eye tracking methods	Machine learning algorithms.
Malla et al. (2020)	Speech emotion recognition methods	MFCC, STFT, ECC and Classification is done by combining CNN and LSTM.
Kołakowska et al. (2014)	Sensory data detection available on modern smartphones.	Machine learning approaches implemented to recognize emotional states.
Imani & Montazer (2019)	Various emotions recognition methods have been represented for online learning platforms.	Classification based on feature and machine or deep learning methods.
Sharma & Dhall (2021)	Emotion recognition on visual, speech, text, EEG.	Deep learning classifiers and feature level and decision level fusion.

- **MELD** [8, 61]: Se compone de clips, en inglés, de más de 1400 diálogos de la serie "*Friends*". Cada diálogo está formado por varias muestras, el número total es de más de 13000. Los datos están divididos en diferentes conjuntos.

Adicionalmente, a los corpus ya comentados en [1] se propone una revisión de las bases de datos más utilizadas para el reconocimiento de emociones por clasificadores multimodales, explicando el número de muestras, las modalidades y como han sido anotadas y recogidas (ver Tabla 2.12). Otro estudio interesante se muestra en [36]. Los autores llevan a cabo un análisis de las principales propuestas sobre clasificadores multimodales para la clasificación de emociones. En este estudio, realizan un resumen detallado de diversas bases de datos utilizadas en investigaciones recientes. Se recoge información como el idioma de estas bases, la modalidad, el número de muestras y de donde se extrajeron los datos (ver Tabla 2.13).

2.6.2. Reconocedor de emociones multimodal

Las técnicas utilizadas en tareas de reconocimiento de emociones multimodal son diversas. La estrategia principal consiste en aplicar modelos individuales para cada modalidad, con el objetivo de fusionar los resultados mediante diferentes métodos, ya sea utilizando modelos de aprendizaje automático o técnicas de fusión de medias matemáti-

Tabla 2.12: Bases de datos de reconocimiento multimodal en [1]

Base de datos	Tipo de muestras	Modalidades	Método de anotación	Espontáneo / Posado
emoFBVPD	1380 samples	Audio, video, physiological data	Per stimuli	Spontaneous
DEAP	40 music videos	Face videos, physiological signals, EEG signals	Per stimuli	Spontaneous
SEED	15 film clips	Face videos, physiological signals, EEG signals	Per stimuli	Posed
FER2013	35,887 facial images	Images	Face, gesture, speech, audio, physiological signals	Continuous
K-Emocon	16 sessions	Face, gesture, speech, audio, physiological signals	Per stimuli	Spontaneous
PMEmo	457 subjects	Audio	Continuous	Posed
MAHNOB-HCI	20 film excerpts	EEG, ECG, GSR, RESP, TEMP	Per stimuli	Posed
VREED	34 subjects	Physiological features: ECG and GSR, Behavioural features (eye gaze)	Per stimuli	Posed
MEmoR	5502 video clips	Video, audio, text	Per stimuli	Spontaneous

cas. Además, se proponen distintas arquitecturas que, incluso antes de obtener los resultados de los modelos individuales, aplican ciertas técnicas con la idea de que el modelo final es multimodal.

En [36] se analizan los modelos que mejores resultados obtienen en este tipo de tareas. Se recogen a modo de resumen algunos de estos modelos, el año en el que se propusieron cada uno de ellos, las bases de datos sobre las que se analizaron y las métricas obtenidas. Los modelos que destacan por obtener buenos resultados en algunos de los conjuntos de datos recogidos en el estudio son el denominado por los autores como DISRFN, que obtiene un 87,50 % en CMU-MOSEI [86], TIMF un 92,28 % en CMU-MOSI [85] y TDET un 86,20 % en CMU-MOSEI (ver tabla 2.14). DISRFN se basa en una red residual profunda que se centra en la estrategia de Red de Fusión de Representación Específica Dinámicamente Invariante, para obtener representaciones conjuntas, separadas por dominios, de cada una de las modalidades y fusionar cada representación a través de una red de fusión de gráficos jerárquica. Esta arquitectura se propone en [28] (ver Figura 2.10).La idea de TIMF es que cada modalidad aprende características por separado para realizar una fusión de tensores de las características de cada modalidad. La fusión de tensores se realiza entre audio-texto y visual-texto. Y la clasificación final se obtiene al aplicar sobre ambos tensores una completamente conectada, para sobre esto aplicar "*soft fusion*", este un método utilizado para mejorar la precisión de las predicciones al combinar las probabilidades (ver Figura 2.11). Este estudio se propone en [73]. TDET es un sistema que propone una red de traducción de "*encoder-decoder*" (estructura de redes neuronales para transformar una secuencia de entrada en una de salida) multimodal con un "*transformer*" para tener en cuenta la mala calidad de las características no lingüísticas (ver Figura 2.12). Los vectores de las probabilidades obtenidas de cada modalidad (representado como U

Nombre	Año	Modalidades	Fuente	Idioma	Número
IEMOCAP	2008	A+V+T	N/A	English	10039
DEAP	2011	A+V+T	N/A	English	10039
CMU-MOSI	2016	A+V+T	YouTube	English	2199
CMU-MOSEI	2018	A+V+T	YouTube	English	23453
MELD	2019	A+V+T	The Friends	English	13000
Multi-ZOL	2019	V+T	ZOL.com	Chinese	5288
CH-SIMS	2020	A+V+T	N/A	Chinese	2281
CMU-MOSEAS	2021	A+V+T	YouTube	Spanish, Portuguese, German, French	40000
MEMOTION	2022	V+T	Reddit	English	10000

Tabla 2.13: Resumen bases de datos de reconocimiento multimodal en [36]

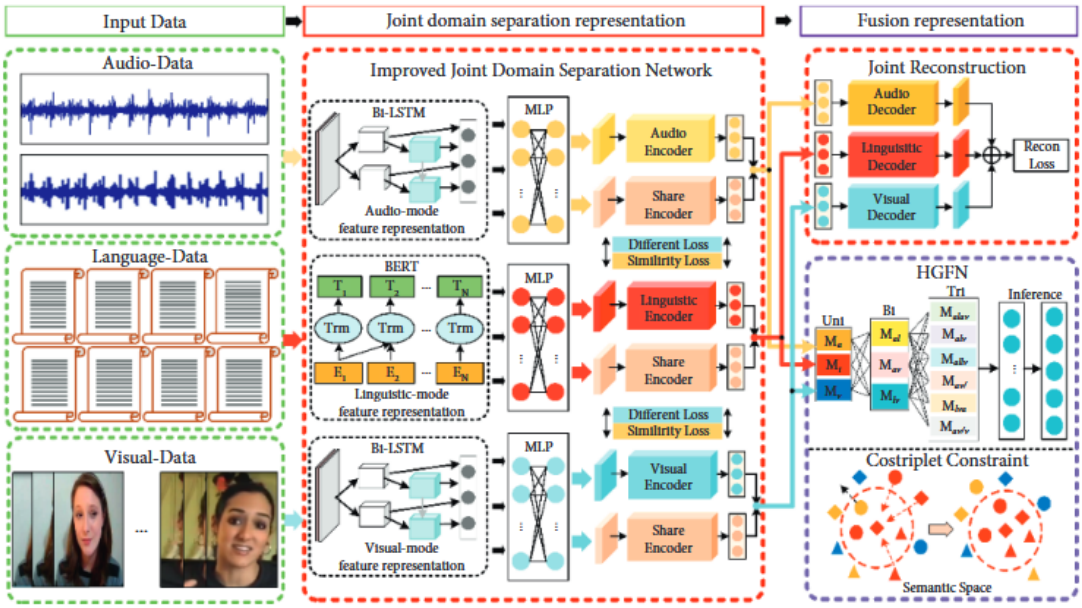

Figura 2.10: Arquitectura propuesta en [28]

en la imagen) se fusionan al aplicar sobre ellas una transformación lineal para determinar la clasificación final. Se propone también una fusión para las muestras por video (ver Figura 2.13). Esta arquitectura se propone en [80].

2.6.3. Corpus comunes en español

Si nos centramos únicamente en el idioma español al analizar la literatura, se puede concluir que hay una parte de trabajos en este lenguaje que se basan en la base de datos que aparece en la Tabla 2.13 denominada como CMU-MOSEAS [84], aunque no hay mucha variedad debido a que la base de datos es bastante reciente (año 2021). Además, en [79] se indica la existencia de la base de datos MOUD [52] que engloba prácticamente el resto de literatura multimodal de reconocimiento de emociones para español.

Por lo tanto, los corpus más comunes en español son:

- **CMU-MOSEAS**: Se tratan de 40000 muestras extraídas de vídeos de YouTube para español, portugués, alemán y francés.

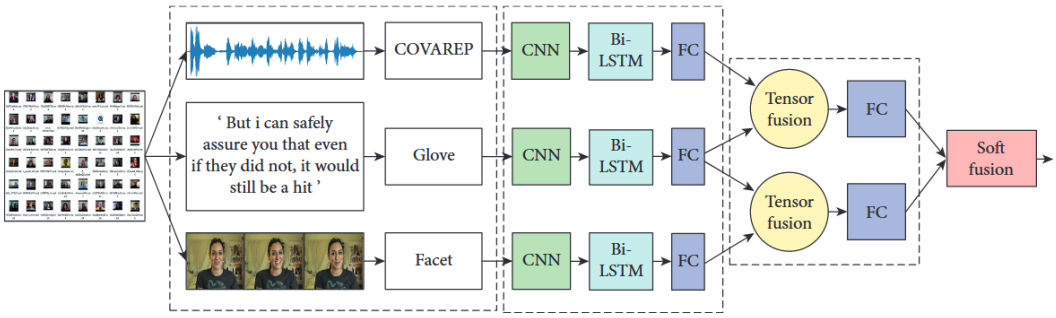

Figura 2.11: Arquitectura propuesta en [73]

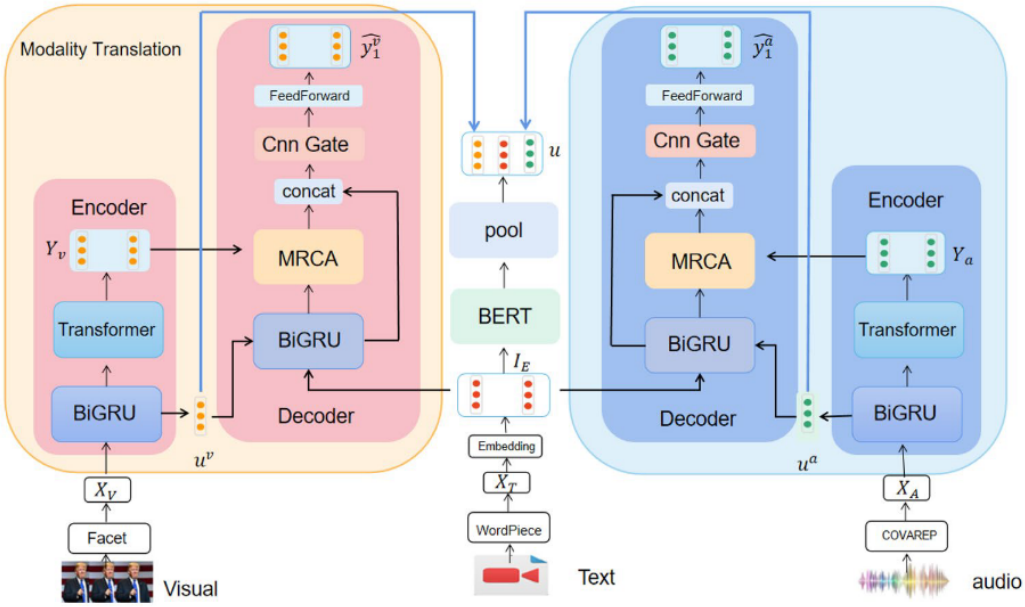

Figura 2.12: Arquitectura propuesta en [80]

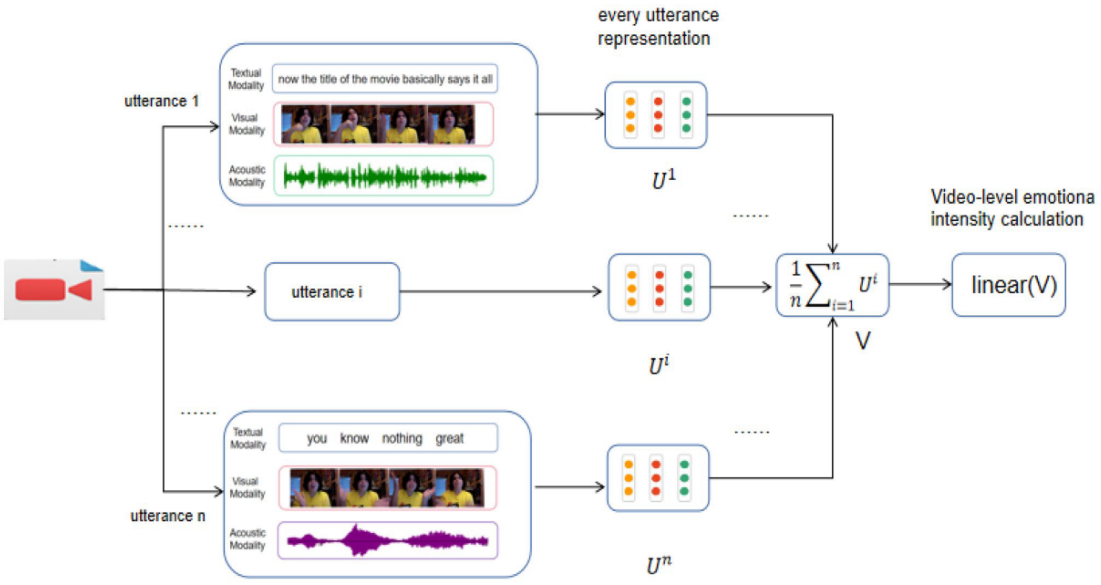

Figura 2.13: Fusión por video de la arquitectura propuesta en [80]

Tabla 2.14: Resumen reconocedores multimodales recogidos en [36]

Modelo	Año	Conjunto de datos	Precisión
MultiSentiNet-Att	2017	MVSA	68.86 %
DFF-TMF	2019	CMU-MOSI	80,98 %
DFF-TMF	2019	CMU-MOSEI	77,15 %
AHRM	2020	Flickr	87,10 %
AHRM	2020	Getty Image	87,80 %
SFNN	2020	Yelp	62,90 %
MISA	2020	MOSI	83,40 %
MAG-BERT	2020	CMU-MOSI	84,10 %
MAG-BERT	2020	CMU-MOSEI	84,50 %
TIMF	2021	CMU-MOSI	92,28 %
TIMF	2021	CMU-MOSEI	79,46 %
Auto-ML based Fusion	2021	B-T4SA	95,19 %
Self-MM	2022	CMU-MOSI	84,00 %
Self-MM	2022	CMU-MOSEI	82,81 %
Self-MM	2022	CH-SIMS	80,74 %
DISRFN	2022	CMU-MOSI	83,60 %
DISRFN	2022	CMU-MOSEI	87,50 %
TETFN	2023	CMU-MOSI	84,05 %
TETFN	2023	CMU-MOSEI	84,25 %
TEDT	2023	CMU-MOSI	89,30 %
TEDT	2023	CMU-MOSEI	86,20 %
SPIL	2023	CMU-MOSI	85,06 %
SPIL	2023	CMU-MOSEI	85,01 %
SPIL	2023	CH-SIMS	81,25 %

- **MOUD**: Consta de clips extraídos de 80 vídeos sobre reseñas de productos, se clasifica en 3 clases con respecto a la polaridad de sentimiento.

2.6.4. Reconocedor de emociones multimodal en español

Una vez analizada la literatura sobre las principales contribuciones del estado del arte del reconocimiento multimodal en otros idiomas, a continuación realizamos una revisión de las principales propuestas para el idioma español. Destacar que las bases de datos comentadas son también aplicadas a tareas no solo de emociones sino a reconocimiento del habla, sobretodos esto ocurre para CMU-MOSEAS la cual, al tratarse también de un corpus multilingüe, aplica un gran valor. Ejemplos de estos trabajos son [47, 48].

En [84] se plantea la aplicación del "*multimodal transformer*" (MulT) [77] (ver Figura 2.14). Este es una extensión para datos multimodales de series temporales. Cada modalidad tiene un "*transformer*" separado con el fin de codificar la información de manera jerárquica. El componente clave de esta arquitectura es un conjunto de bloques de atención cruzada entre los datos de series temporales. Los resultados obtenidos sobre CMU-MOSEAS varían entre 0,57 y 0,69 de "*F1-score*" entre cada una de las clases.

En [82] se propone primero aprender la dependencia de las características de cada una de las modalidades y aprender la relación entre cada una de ellas, esto se realiza a través de una red de autoatención de múltiples cabezas unimodales y una red de atención mutua de múltiples cabezas bimodales. Finalmente, las salidas de tanto las representaciones unimodales y de las bimodales se concatenan y se pasan a través de una capa de sali-

Tabla 2.15: Resultados en [82]

Modalidad	Precisión en MOUD
Visión	87,83 %
Audio	82,61 %
Texto	83,48 %
Visión+Audio	88,70 %
Visión+Texto	89,57 %
Audio+Texto	84,35 %
Visión+Audio+Texto	90,43 %

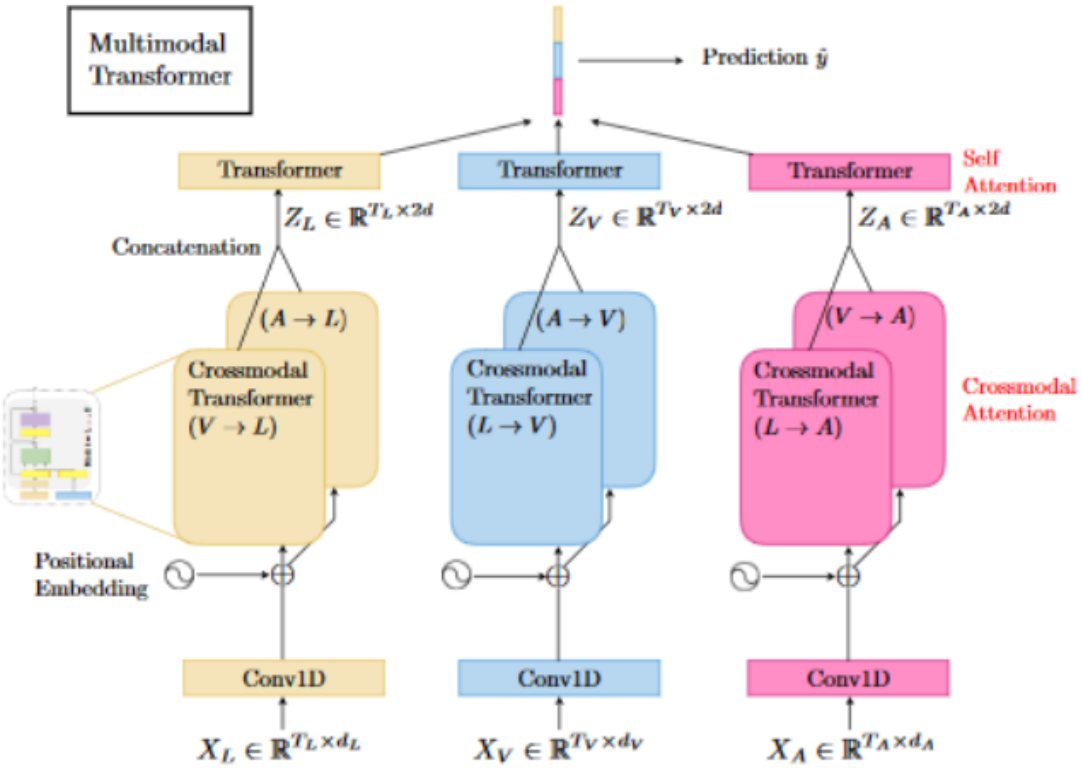

Figura 2.14: Arquitectura propuesta en [77]

da para obtener la clasificación final (ver Figura 2.15). Respecto a MOUD los resultados obtenidos son de 90,43 % al fusionar las tres modalidades (ver Tabla 2.15).

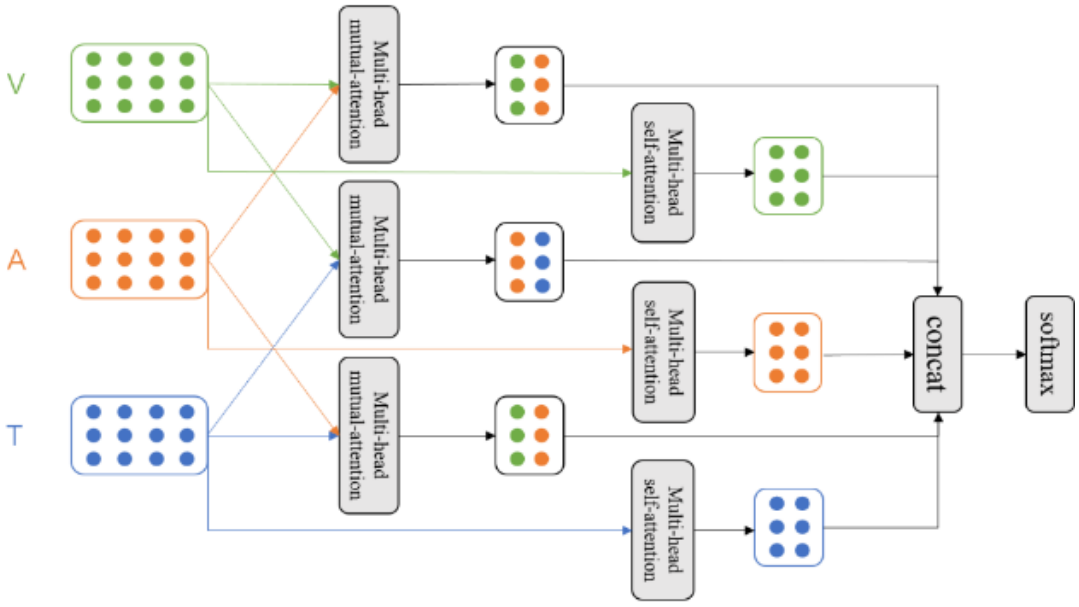

Figura 2.15: Arquitectura propuesta en [82]

CAPÍTULO 3
Implementación

En este capítulo nos centraremos en la implementación que se ha seguido en este trabajo para poder formar nuestro modelo multimodal. Al tratarse de distintos tipos de modelos de clasificación, se detalla cada una de las arquitecturas que componen los modelos de imagen, audio y texto. Una vez se validan estos tres modelos, damos paso a la agregación final que se encarga de fusionar las predicciones individuales en una sola respuesta. Por ello, este capítulo está dividido en dos partes.

La primera parte, se dedica a los modelos individuales (imagen, audio y texto) que sirven como base para la segunda parte, que trata de la fusión multimodal. En ambos casos se explican las bases de datos utilizadas, su tratamiento, las arquitecturas utilizadas, así como los experimentos y resultados obtenidos. Los del multimodal se recogerán en el capítulo siguiente, ya que representa la propuesta final definida para este trabajo.

Es importante destacar que en las bases de datos clasificadas en las seis emociones básicas y la neutral, suele haber desbalanceo en las clases. Siendo asco y miedo las que menos muestras tienen y las que más felicidad y neutral. Lo cual se comprueba en este capítulo al analizar los conjuntos utilizados en cada apartado. Esto se considera una limitación importante a tener en cuenta a la hora de analizar los resultados para el trabajo propuesto.

3.1 Bases de datos de modelos individuales

A continuación se detallan cuáles son las bases de datos que definitivamente se utilizaran en este trabajo, para el entrenamiento de los modelos individuales, y porque se descartaron o no se pudieron utilizar algunas de ellas. Diferenciando por la modalidad a la que fueron aplicadas.

3.1.1. Imagen

Tras analizar que bases de datos eran las más convenientes, se plantearon diferentes corpus para este estudio, para ello se analizan los corpus más comunes mencionados en el Capítulo 2.3.1: JAFFE, KDEF, EDFFE, Yale y FER-2013/FERPlus. Decidimos descartar ISED debido a que no clasifica en todas las clases deseadas. Y se decidió utilizar las bases de datos restantes. Yale se seleccionó, ya que, aunque solo tiene imágenes de la clase neutral, permite incluir variabilidad en la iluminación.

3.1.2. Audio

Para audio las bases de datos utilizadas han sido ya explicadas en el Capítulo 2.4.3, al igual que con imagen se analizaron los corpus más comunes para español y se decidió utilizar INTER1SP y EmoMatchSpanishDB. Se descartó EmoFilm debido a que no clasifica en todas las clases deseadas.

3.1.3. Texto

Para finalizar, para texto se utilizó únicamente un corpus pertenecientes al Worksop de TASS del SEPLN [68] mencionada en el Capítulo 2.5.3. Se utiliza el corpus correspondiente a la tarea 2 (explicado en el Capítulo 2.5.4) debido a que es el único que presenta la división en las clases analizadas en este estudio.

3.2 Modelo de imagen

Empezando con la búsqueda del modelo de imagen se han implementado cuatro arquitecturas diferentes, todas ellas basadas en arquitecturas de la bibliografía. Estas arquitecturas se entrenan y estudian sobre cinco conjuntos diferentes con un preprocesado común. Todos ellos se extraen de las bases de datos explicadas con anterioridad con la idea de buscar el mejor modelo aplicado al mejor conjunto. Al seleccionar estas dos variables se estudia también diferentes técnicas de aumentación de datos. El modelo final se analiza sobre un conjunto de prueba, utilizado únicamente para este análisis, con el fin de estudiarlo sobre datos que no ha visto.

3.2.1. Obtención de los conjuntos

Para entrenar los modelos de imagen decidimos analizar cinco conjuntos de datos distintos (ver Tabla 3.1). En todos los casos, se eliminaron las clases que no correspondan a las seis emociones básicas y a la clase neutral. Los cambios aplicados se mantenían en los siguientes conjuntos si mejoraban los resultados. A continuación se explica cada uno de ellos:

- Conjunto 1: Utiliza el etiquetado original para la base de datos FER-2013. Además de todas las bases de datos seleccionadas anteriormente a excepción de Yale.
- Conjunto 2: Utiliza el etiquetado de FERPlus. Además de todas las bases de datos seleccionadas anteriormente a excepción de Yale.
- Conjunto 3: Utiliza el etiquetado de FERPlus. Además de todas las bases de datos seleccionadas anteriormente a excepción de Yale. Para la base de datos de KDEF se eliminaron los perfiles completos para evitar demasiadas muestras del mismo actor.
- Conjunto 4: Utiliza el etiquetado de FERPlus. Además de todas las bases de datos seleccionadas anteriormente a excepción de Yale. Para la base de datos de KDEF se eliminaron los perfiles completos para evitar demasiadas muestras del mismo actor. Para finalizar, se aplica un mapa de color a las imágenes finales. Esto se realiza para que los tres canales de las imágenes tengan información relevante y no sean una copia uno del otro.

- Conjunto 5: Utiliza el etiquetado de FERPlus. Además de todas las bases de datos seleccionadas anteriormente. Para la base de datos de KDEF se eliminaron los perfiles completos para evitar demasiadas muestras del mismo actor. Respecto a Yale, ya que hay un gran número de muestras neutrales (por ejemplo, en el conjunto 3 hay 15575 muestras neutrales frente a un total de 38860 muestras). Se decide incluir una parte de los datos, seleccionando 28 archivos de cada uno de los actores para no sobre ajustarse a ninguno de ellos.

Tabla 3.1: Conjuntos de imágenes

Id	Aplicar RGB	KDEF	FER-2013 Etiquetado	Yale
1	no	todo	original	no
2	no	todo	+	no
3	no	sin perfiles	+	no
4	si	sin perfiles	+	no
5	no	sin perfiles	+	una parte

Para todos los conjuntos se ha reservado parte de prueba que representa el 15 % de los datos. En cuanto a la proporción de clases, las que más muestras tienen, con diferencia, son felicidad y neutral y, por otra parte, las que menos son los de asco y miedo (ver Tabla 3.2 para el conjunto 3).

Tabla 3.2: Distribución de las muestras de conjunto 3 de imagen

Etiqueta	Entrenamiento	Prueba
felicidad	10894	1922
tristeza	2430	428
enfado	2171	382
sorpresa	2909	513
asco	585	103
miedo	806	142
neutral	13239	2336
Total	33034	5826

3.2.2. Preprocesamiento de imagen

Para todos los conjuntos, se aplicó el mismo procesado en las imágenes, excepto al probar la transformación a RGB en el conjunto 4 que se vio empeoraba los resultados. En cuanto al preprocesado aplicado, lo primero que se hace es pasar todas las imágenes por un proceso de reconcomiendo de caras gracias a la herramienta de *face_recoginition* la cual se encuentra en el GitHub [24]. Para tener imágenes significativas se eliminaron en las que no se reconocía ninguna cara y el resto se recorta teniendo en cuenta el recuadro definido por el reconocedor. Tras ellos las imágenes se encuadraban, se pasaban a blanco y negro y se redimensionaban a 224x224.

3.2.3. Arquitecturas

Las arquitecturas implementadas se basan en cuatro estudios que se recogen en [7,31, 33,46] (ver Capítulo 2.3.2). La decisión de implementar estas arquitecturas viene de que presentan buenos resultados y a que permitían implementar diferentes tipos de sistemas

con el fin de buscar cuál podía funcionar mejor, enriqueciendo así el estudio. Concretamente, se aplican diferentes redes convoluciones que implementan técnicas variadas, así como la implementación de sistemas que utilizan componentes de "*transformers*" además de la aplicación de "*vision transformers*". A continuación se detallan las decisiones aplicadas en la implementación de cada una de ellas, en cada título de las siguientes secciones se registra el nombre que se le da a cada una de las arquitecturas en este documento.

Primera arquitectura - *resdob*

En este caso se ha implementado el modelo propuesto en [46]. En cuanto a los conjuntos, decir que el cambio del preprocesado en el conjunto 4 de pasar las imágenes a RGB no muestra diferencias, ya que esta arquitectura tiene entrada para imágenes tanto para RGB como para blanco y negro. Se aplican los hiperparámetros mencionados en el estudio referenciado.

Segunda arquitectura - *vit*

Para este segundo modelo se implementa la arquitectura denominada como ViT-B/16/SAM en [7], se selecciona esta debido a que era la que mejores resultados obtenía si se tiene en cuenta el balance entre precisión y perdida. Para implementarlo se utiliza como ViT pre-entrenado el modelo [1] junto a los hiperparámetros mencionados en el estudio referenciado para esta arquitectura.

Tercera arquitectura - *vgg*

Según como se indica en [33] se aplica un modelo *VGGNet* con los hiperparámetros mencionados en el estudio referenciado. Se decide implementar concretamente una arquitectura VGG 16, ya que, se considera suficiente para capturar características complejas pero sin llegar a producir sobre ajuste.

Cuarta arquitectura - *resn*

La última arquitectura estudiada se presenta en [31]. Se implementa la red *ResNet* propuesta y por ello se redimensiona la imagen a 128x96 para entrenar la red. Se aplican también los hiperparámetros mencionados en el trabajo referenciado.

3.2.4. Entrenamiento de los modelos

En la Tabla 3.3 se recogen los resultados sobre el conjunto de entrenamiento y en la Tabla 3.3 los de validación, el cual representa un 15 % de los datos de entrenamiento. Se ha decidido estudiar sobre diferentes métricas para analizar la clasificación, no solo global, sino de cada una de las clases. Por ello, Cohen κ es la métrica que se tiene en cuenta principalmente. Siguiendo con esta idea y, tras analizar todos los resultados, se concluye que la configuración que mejores resultados ofrece es *vit* sobre el conjunto 3, obteniendo un 0,847 en esta métrica en entrenamiento y 0,7608 en validación. Esta arquitectura es mejor al resto en todos los otros conjuntos, y, de la misma manera, este conjunto también es el mejor conjunto con respecto al resto de modelos.

[1] Hugging Face - modelo ViT utilizado

Tabla 3.3: Resultados de los modelos de imagen (Entrenamiento)

Conjunto	Modelo	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	resdob	0,338	0,1429	0,0722	0,5	0
1	vit	0,7656	0,5903	0,6006	0,774	0,6896
1	vgg	0,5179	0,3781	0,3663	0,8397	0,3553
1	resn	0,6857	0,5603	0,5644	0,9405	0,5867
2	resdob	0,4238	0,1429	0,085	0,5	0
2	vit	0,8789	0,6273	0,655	0,8009	0,8207
2	vgg	0,6571	0,5852	0,4491	0,9001	0,4955
2	resn	0,7578	0,7011	0,5379	0,9552	0,6433
3	resdob	0,4221	0,1429	0,0848	0,5	0
3	**vit**	0,8944	0,7193	0,7254	0,8497	0,847
3	vgg	0,7207	0,6639	0,5302	0,9406	0,5922
3	resn	0,7608	0,7074	0,54	0,9567	0,6484
4	resdob	0,4207	0,1429	0,0846	0,5	0
4	vit	0,8413	0,5678	0,5837	0,7675	0,7644
4	vgg	0,6565	0,5751	0,4405	0,903	0,4907
4	resn	0,7455	0,6852	0,5106	0,9509	0,6253
5	resdob	0,4213	0,1429	0,0847	0,5	0
5	vit	0,8802	0,6621	0,6871	0,819	0,8244
5	vgg	0,6807	0,6124	0,4667	0,43	0,529
5	resn	0,7571	0,6983	0,5396	0,955	0,6416

Aumento de los datos

Teniendo la arquitectura seleccionada y el conjunto de datos que mejores resultados ofrecían, se ha decidido aplicar un pequeño estudio con diferentes funciones para aumentar la variabilidad del conjunto de datos y para ayudar a prevenir el sobre ajuste. La configuración de las transformaciones es:

- **Transformación 1**: Aplicar *RandomAffine*, esta es una transformación afín aleatoria, con parámetros para realizar rotaciones aleatorias entre ±30 grados, traslaciones de hasta 10 % del tamaño de la imagen en dirección tanto horizontal como vertical, así como que se redimensiona aleatoriamente la imagen entre el 80 % y el 120 % del tamaño original. Se decide aplicar estas transformaciones, comunes en el mundo real, para una mejor generalización de la red y para que pueda aprender patrones más diversos al simular variaciones en la pose, posición y tamaño de las caras.
- **Transformación 2**: Implica aplicar *ColorJitter* a la transformación 1. Esta segunda transformación realiza ajustes aleatorios en el color. Para ello, se establece que el brillo de la imagen puede aumentar o disminuir hasta un 20 %. De esta forma se obtienen datos que representen diferentes condiciones de iluminación, haciendo que la red sea más diversa para enfrentar imágenes más variadas en este aspecto.

La Tabla 3.5 muestra los resultados tras aplicar estas transformaciones sobre el conjunto de entrenamiento mientras que la Tabla 3.6 muestra la misma información para el conjunto de validación. La transformación 1 ofrece los mejores resultados, mejorando levemente los del anterior apartado, en un 1 % respecto a Cohen κ en el conjunto de validación. Por ello esta transformación será aplicada para el modelo final.

Tabla 3.4: Resultados de los modelos de imagen (Validación)

Conjunto	Modelo	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	resdob	0,324	0,1429	0,0699	0,5	0
1	vit	0,7045	0,5151	0,5123	0,731	0,6095
1	vgg	0,4595	0,3947	0,2329	0,8063	0,2714
1	resn	0,5818	0,5121	0,3667	0,8595	0,4626
2	resdob	0,4139	0,1429	0,0836	0,5	0
2	vit	0,8354	0,5645	0,5885	0,7649	0,7541
2	vgg	0,597	0,5352	0,3252	0,8615	0,4277
2	resn	0,6432	0,6046	0,3499	0,8695	0,5113
3	resdob	0,4195	0,1429	0,0848	0,5	0
3	**vit**	0,8337	0,6222	0,6231	0,7951	0,7608
3	vgg	0,6373	0,6071	0,3432	0,8582	0,5012
3	resn	0,6608	0,634	0,3709	0,8818	0,4966
4	resdob	0,4247	0,1429	0,0852	0,5	0
4	vit	0,7941	0,5162	0,524	0,7368	0,6952
4	vgg	0,6077	0,5556	0,317	0,8697	0,4468
4	resn	0,4655	0,4126	0,2113	0,781	0,2921
5	resdob	0,4264	0,1429	0,0854	0,5	0
5	vit	0,8351	0,5844	0,6033	0,7756	0,7571
5	vgg	0,6111	0,5742	0,3166	0,8812	0,485
5	resn	0,6136	0,6082	0,2166	0,855	0,3913

Tabla 3.5: Resultados de la aplicación de aumentación de datos para el modelo de imagen (Entrenamiento)

Transformaciones	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	0,8948	0,7169	0,7587	0,8474	0,845
2	0,8866	0,7335	0,7335	0,8445	0,8347

Resultados en conjunto de prueba y configuración final

Teniendo en cuenta lo mencionado anteriormente, la configuración final del modelo se obtiene de entrenar sobre el conjunto 3 realizando un aumento de los datos mediante la transformación 1 y con la arquitectura *vit*. La Tabla 3.7 muestra los resultados obtenidos sobre el conjunto de prueba. Como puede verse, los resultados sobre el conjunto de prueba son ligeramente mejores que los de validación. La matriz de confusión, mostrada en la Figura 3.1, presenta una diagonal relativamente clara, por lo que clasifica correctamente las clases a excepción de asco y miedo.

3.3 Modelo de audio

Para el estudio de diferentes modelos de audio primero se explica el conjunto de datos sobre el que se trabaja, todo ello proveniente de las bases de datos explicadas anteriormente. Teniendo en cuenta la revisión de la bibliografía, al igual que para tratar las imágenes, se toma como base los artículos encontrados para la implementación. A continuación se hace un pequeño resumen como recordatorio de las arquitecturas a utilizar

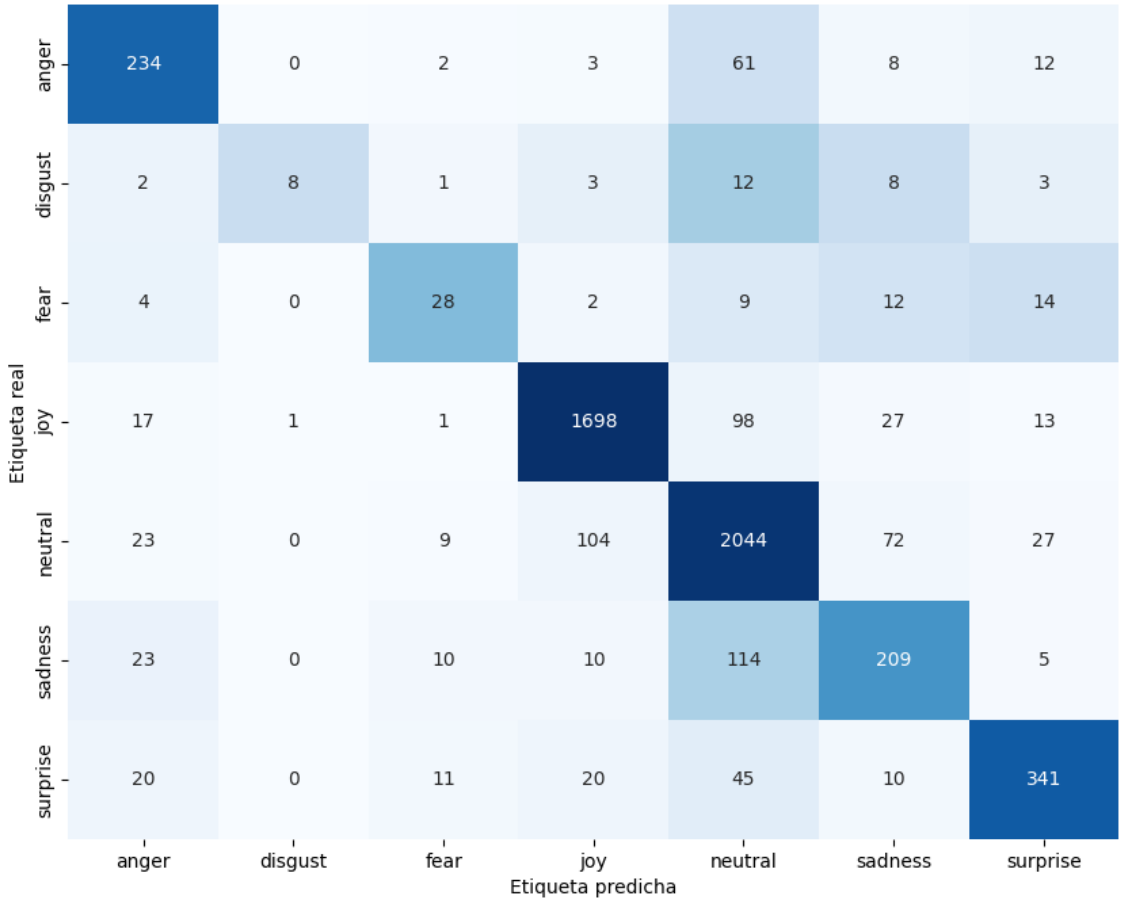

Figura 3.1: Matriz de confusión del modelo de imagen sobre el conjunto de prueba

Tabla 3.6: Resultados de la aplicación de aumentación de datos para el modelo de imagen (Validación)

Transformaciones	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	0,8468	0,629	0,6604	0,7985	0,7719
2	0,8316	0,6004	0,6195	0,7835	0,7544

Tabla 3.7: Resultados del modelo de imagen sobre el conjunto de prueba

Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
0,8483	0,6417	0,6676	0,8054	0,7767

como base, al igual que el preprocesado que se utiliza para cada una, explicando las decisiones tomadas sobre la implementación en este trabajo. Para finalizar se presentan las pruebas realizadas y el estudio proveniente de modificar diferentes hiperparámetros con la idea de poder mejorar los resultados, dando lugar al modelo final. Este se estudia sobre un conjunto de prueba solo utilizado para este análisis.

3.3.1. Explicación del conjunto

Inicialmente, se realizaron todos los experimento sobre el conjunto obtenido de unir las bases de datos seleccionadas. Sin embargo, debido a que los actores que grababan las pruebas de INTER1SP eran solo dos, los modelos se ajustaban demasiado al hablante. Por ello, y teniendo en cuenta que EmoMatchSpanishDB tenía más 50 individuos que graban las muestras, se decidió reducir los datos que se seleccionan de INTER1SP. Se decide que estos constituyan, una tercera parte del conjunto final y que cada uno de los actores tenga la misma proporción, para evitar la dependencia por actor. Por lo que, para cada hablante, se seleccionaron 650 audios de manera aleatoria. En este caso también se reserva un 15 % para prueba. Para este caso la distribución del conjunto final entre clases es más balanceada (ver Tabla 3.8).

Tabla 3.8: Distribución de las muestras de audio en español

Etiqueta	Entrenamiento	Prueba
felicidad	352	61
tristeza	366	64
enfado	401	70
sorpresa	388	68
asco	243	42
miedo	385	63
neutral	707	124
Total	2815	492

3.3.2. Arquitecturas y preprocesado

Las arquitecturas y preprocesados implementados se basan en dos estudios que se recogen en [17, 32] (ver Capítulo 2.4.4). Se decide implementar ambos debido a los resultados obtenidos, y a que permite aplicar un estudio variado, debido a la diferencia en el preprocesado y en el tratamiento de la señal auditiva, y por consiguiente de los sistemas aplicados. A continuación se detallan las decisiones aplicadas en la implementación, en

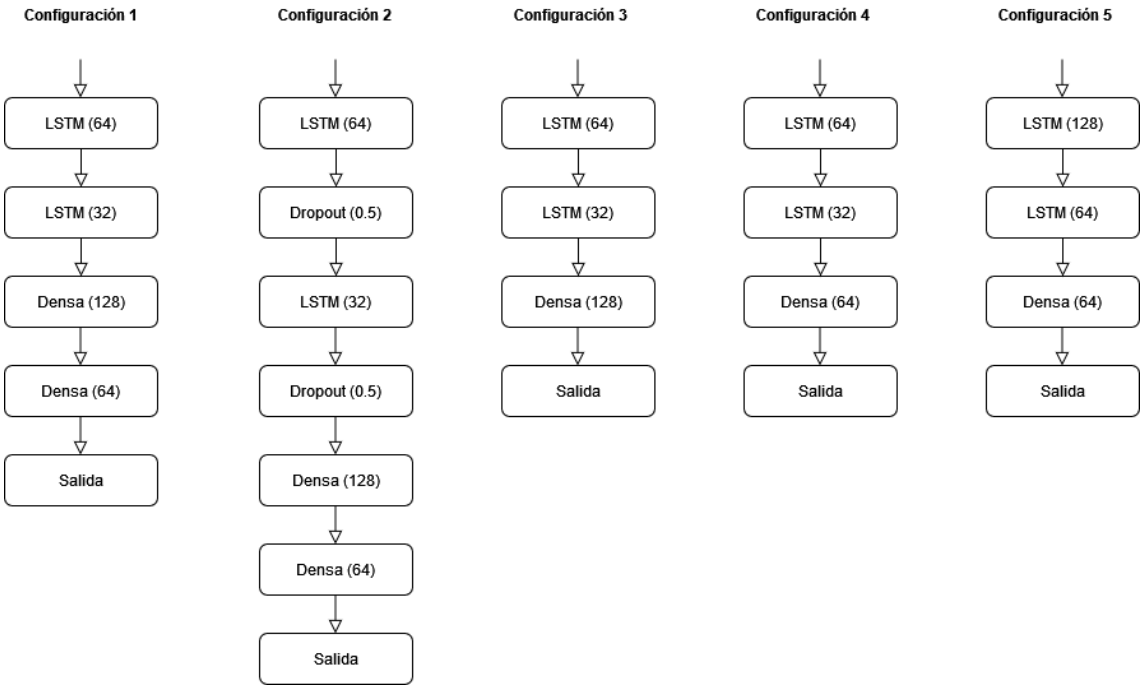

Figura 3.2: Configuración de las arquitecturas LSTM

cada título de las siguientes secciones se registra el nombre que se le da a cada una de las arquitecturas en este documento.

Primera arquitectura y su preprocesado - LSTM

Se implementa el preprocesado y extracción de características que mejores resultados ofrece en [32], es decir, utilizar una LSTM como clasificador y extraer las características al aplicar una combinación de MSFs y MFCC seguido por un LR-RFE. En cuanto a la red LSTM se decide analizar dos configuraciones diferentes con el fin de encontrar la que mejores resultados ofrece. La primera se ajusta a lo indicada en el estudio. La segunda configuración añade más capas, incluyendo capas de " *Dropout*", con el fin de prevenir el sobre ajuste, así como variar el número de neuronas. En cada modificación, se tienen en cuenta los cambios que mejoran los resultados. Estas se definen en la Figura 3.2 junto con el id que se da a cada configuración.

Segunda arquitectura y su preprocesado - CNN_(no)l

En este segundo caso, teniendo en cuenta lo recogido en [17], se implementan las dos arquitecturas explicadas, es decir, una con una capa LSTM y otra sin ella. Se toma esta decisión, ya que, aunque la inclusión de la capa LSTM no ofrezca resultados tan buenos en el estudio, se considera positivo incluir capas que manejen secuencias porque al aplicar datos de audio se pueden capturar mejor las dependencias temporales al igual que su contexto. Además, teniendo en cuenta que parte de los datos provienen de INTERS1P, las redes pueden estar sesgadas por el hablante y los resultados podrían no ser tan dispares realmente. Estas arquitecturas se denominan en este documento como CNN_l y CNN_nol respectivamente, se implementan siguiendo lo especificado en el estudio. La configuración de ambas redes se recoge en la Figura 3.3.

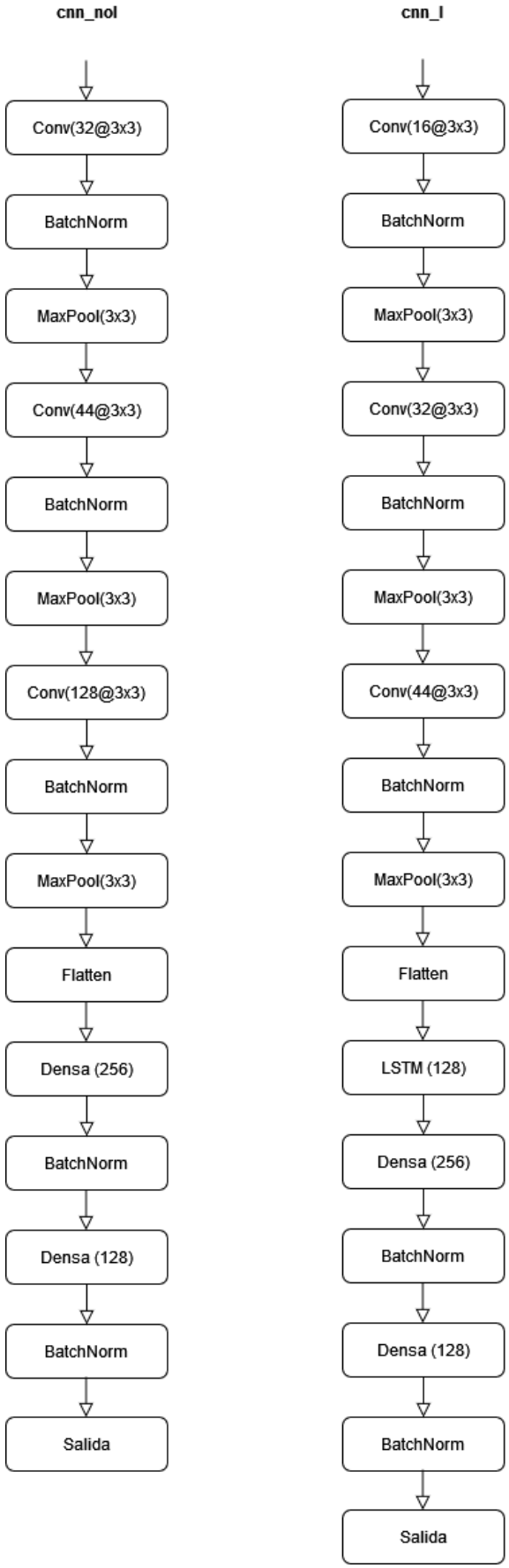

Figura 3.3: Configuración de las arquitecturas CNN_(no)l

3.3.3. Entrenamiento de los modelos

Una vez definidas tanto las arquitecturas como el preprocesado utilizado se aplican los experimentos para cada configuración. Se realizan todas las pruebas con el optimizador Adam [34], abreviatura para "*Adaptive Moment Estimation*", debido a la popularidad y reconocimiento en la comunidad. Los resultados obtenidos sobre el conjunto de entrenamiento se encuentran en la Tabla 3.9. La Tabla 3.10 muestra los resultados con el conjunto de validación, compuesto por un 15 % del total de muestras. Estos resultados recogen las mismas métricas que el anterior apartado, siendo la más relevante, de nuevo, Cohen κ. Para CNN_l y CNN_nol los resultados son para un número de épocas muy bajo, ya que, empieza a sobre entrenar rápidamente. Aunque, en cualquier caso, en validación, clasifica un gran número de muestras a la misma clase. El mejor resultado proviene de aplicar LSTM concretamente la configuración 4 propuesta en este trabajo, este es de un valor de la métrica principal de 0,3616 en validación.

Tabla 3.9: Resultados de los modelos de audio en español (Entrenamiento)

Modelo	Configuración	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
CNN_l	-	0,2987	0,1373	0,2545	0,6786	0,1644
CNN_nol	-	0,4031	0,2202	0,3581	0,7668	0,2885
LSTM	1	0,488	0,2584	0,454	0,8537	0,3796
LSTM	2	0,4578	0,198	0,406	0,8242	0,3422
LSTM	3	0,508	0,282	0,4672	0,8599	0,4046
LSTM	**4**	0,5133	0,2842	0,4781	0,8607	0,4132
LSTM	5	0,5067	0,2744	0,4772	0,8562	0,4046

Tabla 3.10: Resultados de los modelos de audio en español (Validación)

Modelo	Configuración	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
CNN_l	-	0,1429	0	0,0357	0,4518	0
CNN_nol	-	0,1261	0,1261	0,032	0,4902	0
LSTM	1	0,4298	0,2398	0,3753	0,7937	0,3052
LSTM	2	0,4103	0,1563	0,3729	0,7965	0,2958
LSTM	3	0,444	0,2185	0,4034	0,8193	0,3327
LSTM	**4**	0,4618	0,2185	0,4419	0,8278	0,3616
LSTM	5	0,4014	0,1936	0,351	0,776	0,2783

Ajuste de hiperparámetros

Con el fin de mejorar los resultados se decide modificar algunos hiperparámetros para optimizar la convergencia y mejorar la capacidad del modelo. Estos son el "*scaler*" utilizado el optimizador y la tasa de aprendizaje. El "*scaler*" se decide modificar, ya que, aunque en el estudio en el que se basa la implementación se indica la utilización de minmax. Esta elección puede provocar un gran impacto para este tipo de sistemas. Los resultados obtenidos sobre el conjunto de entrenamiento se pueden analizar en la Tabla 3.11 y los de validación en la Tabla 3.12. No hay mejora al realizar este estudio, por lo que se mantendrán los hiperparámetros utilizados anteriormente Adam con la tasa de aprendizaje por defecto, es decir, 0,001 y "*scaler*"minmax).

Tabla 3.11: Resultados de la optimización de hiperparámetros en el modelo de audio en español (Entrenamiento)

Scaler	Tasa	Optimizador	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
robust	0,001	Adam	0,4343	0,1434	0,3825	0,8143	0,3118
standard	0,001	Adam	0,4205	0,1421	0,3646	0,7993	0,2393
norm	0,001	Adam	0,2351	0	0,0577	0,585	0
minmax	0,01	Adam	0,409	0,1101	0,3259	0,7928	0,2687
minmax	0,0001	Adam	0,3606	0,0413	0,2574	0,7488	0,2043
minmax	0,001	rmsprop	0,4774	0,2473	0,4371	0,8409	0,2973
minmax	0,001	namdam	0,5013	0,2735	0,454	0,8468	0,3977

Tabla 3.12: Resultados de la optimización de hiperparámetros en el modelo de audio en español (Validación)

Scaler	Tasa	Optimizador	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
robust	0,001	Adam	0,3837	0,1332	0,3648	0,7809	0,255
standard	0,001	Adam	0,3712	0,1155	0,3079	0,7785	0,2334
norm	0,001	Adam	0,2433	0	0,0559	0,587	0
minmax	0,01	Adam	0,3393	0,0924	0,2599	0,7541	0,1956
minmax	0,0001	Adam	0,3499	0,0231	0,2793	0,7296	0,2082
minmax	0,001	rmsprop	0,4121	0,2078	0,3794	0,7908	0,2973
minmax	0,001	namdam	0,4405	0,1812	0,4111	0,8174	0,3359

Resultado sobre el conjunto de prueba y configuración final

La configuración final se obtiene de aplicar la configuración 4 del modelo LSTM utilizando el "*scaler*" minmax y el optimizador Adam con una tasa de aprendizaje de 0,001. Las métricas obtenidas sobre el conjunto de prueba están en la Tabla 3.13 se empeora cerca de un 20 % los valores de Cohen κ respecto a los de validación. En cuanto a la matriz de confusión obtenida, muestra que no clasifica correctamente las clases y clasifica gran parte de las muestras a enfado, asco y neutral (ver la Figura 3.4).

Tabla 3.13: Resultados del modelo de audio sobre el conjunto de prueba en español

Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
0,3923	0,1321	0,2152	0,6907	0,1858

3.4 Modelo de texto

En cuanto al modelo de texto, al igual que en los apartados anteriores, primero se explican los conjuntos que se estudian, los cuales vienen derivados de la base de datos ya comentada. Basándose otra vez en la bibliografía se extrae la arquitectura a estudiar, al igual que las decisiones tomadas para su implementación. También se aplican distintos tipos de preprocesado. Al definir el mejor conjunto y preprocesado se prueba a modificar la tasa de aprendizaje. Por último, se analiza el modelo sobre un conjunto de prueba solo

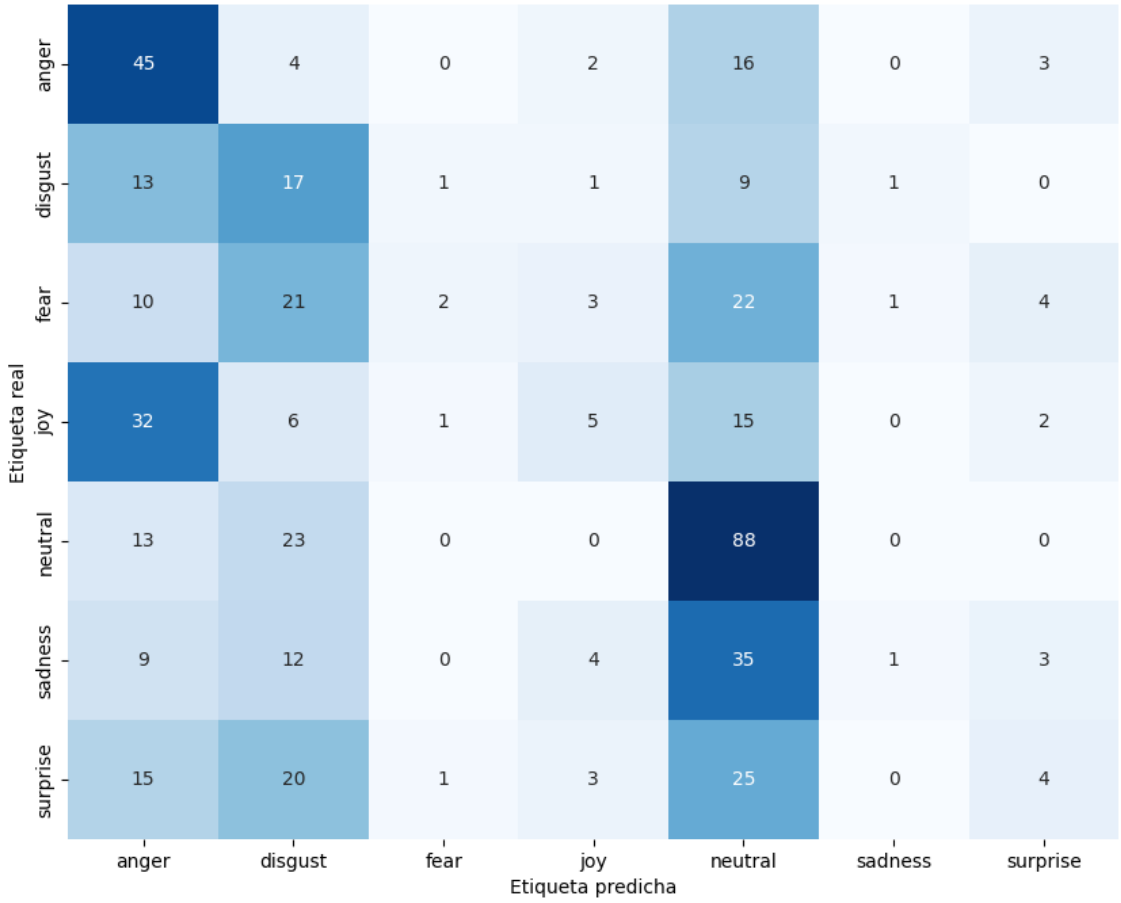

Figura 3.4: Matriz de confusión del modelo de audio en español sobre el conjunto de prueba

utilizado para este análisis con el fin de estudiar el funcionamiento del modelo sobre datos que no ha visto.

3.4.1. Obtención de los conjuntos

El corpus utilizado en este caso presenta un gran desbalanceo en los datos, concretamente asco y miedo son los que menos muestras tienen, y neutral y felicidad los que más (ver Tabla 3.14). Debido a que el conjunto de datos es reducido y a esta diferencia significativa de muestras entre cada una de las clases, se decidió crear diferentes conjuntos al aplicar submuestreo. Para ello se prueba a reducir los datos, fijando el número de muestra a cierto valor, de las clases felicidad y/o neutral sobre el conjunto de entrenamiento y se crean cuatro conjuntos diferentes (ver la Tabla 3.15). En este caso se definen los conjuntos de validación, entrenamiento y prueba siguiendo la división definida en la propia competición de la que se extrae el corpus.

Tabla 3.14: Distribución de las muestras de *tweets* [68]

Etiqueta	Entrenamiento	Validación	Prueba
joy	1270	185	360
sadness	706	103	200
anger	600	87	170
surprise	241	35	68
disgust	113	16	32
fear	67	10	19
others	2889	421	817
Total	5886	857	1666

Tabla 3.15: Conjuntos de datos para modelo de texto

Conjunto	Reducción de neutral	Reducción de felicidad
1	No	No
2	800	800
3	1000	1000
4	1000	No

3.4.2. Preprocesamiento de texto

En cuanto a los diferentes preprocesados que se aplican sobre los *tweets*, hemos adjuntado una definición junto al nombre con el que se referencian en este documento. Se aplican de base preprocesados típicos de tareas de procesamiento del lenguaje natural para trabajar con datos estructurados y limpios y además se prueban diferentes tratamientos del texto, teniendo en cuenta que se tratan de *tweets*:

- **Preprocesado base**: Se "*tokeniza*" el texto, es decir, se divide en unidades individuales como palabras. Cada palabra se convierte a minúsculas, se eliminan las palabras vacías y los signos de puntuación.
- **Preprocesado 1**: Se aplica el preprocesado base y tras ello se traducen los emojis a su representación textual en español, con el fin de utilizar la información lingüística que puede ofrecer.

- **Preprocesado 2**: Se aplica el preprocesado 1 además de eliminar los enlaces y las palabras "HASHTAG" y @USER para comprobar si el modelo extrae información de ello o lo considera ruido.

- **Preprocesado 3**: Se aplica el preprocesado 2 y se eliminan los caracteres repetidos, por ejemplo "holaaa"se modifica a "hola". Se aplica esto debido a que este tipo de comportamientos son comunes en este tipo de datos y no ofrece información.

3.4.3. Arquitectura

Se implementa únicamente un modelo. Este modelo ha sido desarrollado tomando como punto de partida una arquitectura introducida en [25] y revisada en mayor profundidad en el Capítulo 2.5.4. En este caso solo se implementa un modelo, ya que este gano la competición de la que se extrae el corpus estudiado. Respecto a la aplicación del sistema, se decidió que el modelo sobre el que se aplica el "*fine-tune*" fuera un Bert base multilingüe. Ya que la intención era no ajustarse solo a *tweets* y que el modelo final pueda generalizar para los datos de la parte multimodal.

3.4.4. Entrenamiento de los modelos

Se entrenan todas las pruebas con una tasa de aprendizaje 3e-5 y el optimizador AdamW [41], variante del optimizador Adam con"*weight decay*". Debido a que estos parámetros son comúnmente utilizados y se consideran una buena práctica al aplicar modelos de tipo Bert. Los resultados de los diferentes experimentos se recogen en la Tabla 3.16 para el conjunto de entrenamiento y en la Tabla 3.17 para validación. De nuevo se estudian diferentes métricas priorizando la clasificación en todas las clases, por ello, la métrica usada para tomar las decisiones será la de Cohen κ. Al analizar los resultados se concluye que el mejor modelo es el obtenido de entrenar sobre el conjunto 3 con el preprocesado con la configuración 3, siendo el valor de Cohen κ de 0,5337.

Tabla 3.16: Resultados de los modelos de texto en español (Entrenamiento)

Conjunto	Preproceso	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	base	0,7292	0,4643	0,46	0,7022	0,5966
2	base	0,7892	0,6274	0,6363	0,7948	0,7336
3	base	0,7897	0,6039	0,6148	0,7828	0,7503
4	base	0,7651	0,5494	0,5412	0,7527	0,6951
3	1	0,777	0,6047	0,5913	0,7819	0,7138
3	2	0,7626	0,6219	0,6016	0,7898	0,6988
3	**3**	0,7809	0,566	0,5768	0,7628	0,7171

Ajuste de hiperparámetros

Se estudia la posible optimización de la tasa de aprendizaje, con el fin de optimizar la convergencia del modelo, manteniendo los valores bajos tal y como se espera para esta implementación. Los resultados obtenidos se recogen en la Tabla 3.18 para el de entrenamiento y en la Tabla 3.19 para validación. No hay mejora por lo que se seguirá utilizando un factor de $3 \cdot 10^{-5}$.

Tabla 3.17: Resultados de los modelos de texto en español (Validacion)

Conjunto	Preproceso	Precisión	Exhaustividad	F1-score	AUC	Cohen κ
1	base	0,6116	0,3644	0,3456	0,6398	0,4232
2	base	0,546	0,4359	0,4424	0,6779	0,441
3	base	0,6018	0,4433	0,4382	0,6852	0,4887
4	base	0,5733	0,3996	0,3984	0,6594	0,4369
3	1	0,6125	0,4485	0,435	0,6885	0,4998
3	2	0,6036	0,4796	0,4386	0,7045	0,4987
3	**3**	0,6411	0,4385	0,4343	0,6859	0,5337

Tabla 3.18: Resultados de la optimización de hiperparámetros en el modelo de texto en español (Entrenamiento)

Tasa de Aprendizaje	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
$3 \cdot 10^{-6}$	0,7045	0,4586	0,4522	0,7016	0,6149
$3 \cdot 10^{-4}$	0,2687	0,1429	0,0605	0,5000	0,0000

Resultado sobre conjunto de prueba y configuración final

Finalmente, teniendo definida la estructura a seguir, se analiza este modelo en el conjunto de prueba. Los resultados obtenidos se recogen en la Tabla 3.20. Empeoran los resultados un 10 % respecto al conjunto de validación al estudiar la métrica principal. La matriz de confusión no representa resultados alentadores, aunque son ligeramente mejores que para el caso de audio (ver Figura 3.5).

3.5 Base de datos modelo multimodal

A continuación se detallan los problemas y limitaciones encontrados al buscar una base de datos multimodal en español para el reconocimiento de emociones que cumpla con los requisitos necesarios, al igual que las decisiones tomadas para continuar con el estudio.

3.5.1. Estudio, decisiones y problemas encontrados

En un principio, se decide utilizar la base de datos CMU-MOSEAS, mencionada en el Capítulo 2.6.3. Sin embargo, tras pedir acceso a la base a través de los diferentes canales disponibles no fue posible obtener respuesta de los autores. Finalmente nos informaron de que hubo un problema al realizar el respaldo de los datos y que estos se perdieron.

Aunque la base de datos MOUD, que también aparece en el Capítulo 2.6.3, no se tenía en cuenta inicialmente, al contener 3 clases de polaridad: neutral, positivo y negativo, se decidió utilizarla al no haber otra base de datos en español. Para ello se dividen las clases de la siguiente manera: neutral=neutral, positiva=felicidad, se elimina sorpresa y el resto se clasifican como negativas. Se entrenaron los mejores modelos de cada modalidad con la misma arquitectura y configuración para esta nueva clasificación. Esto se realiza con el fin de predecir los datos de MOUD y sobre ello aplicar técnicas de fusión. Para las modalidades de audio y texto, los resultados obtenidos sobre validación son ligeramente superiores al de la clasificación con siete clases. En el caso de audio, la Kappa de Cohen aumenta algo más de un 2 % y para texto un 3 %. Al predecir los datos de audio y texto,

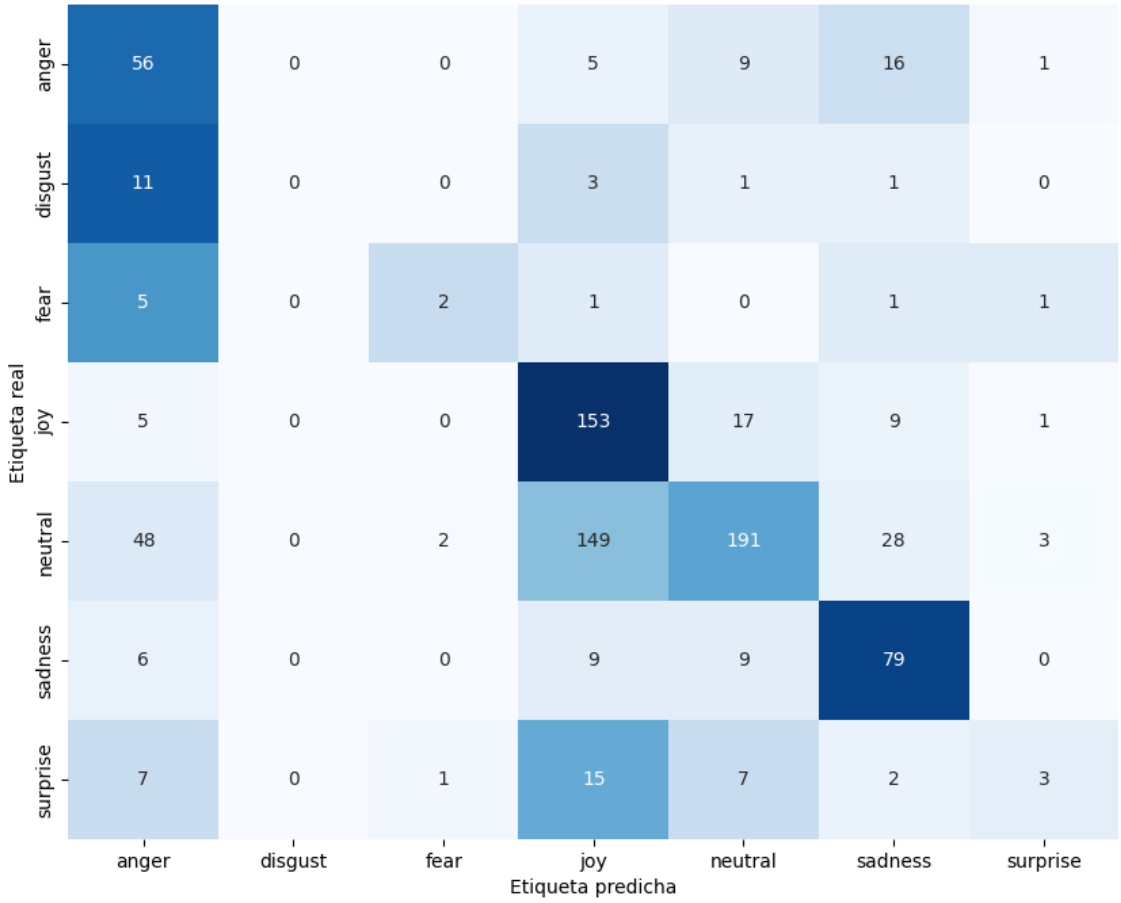

Figura 3.5: Matriz de confusión del modelo de texto en español sobre el conjunto de prueba

Tabla 3.19: Resultados de la optimización de hiperparámetros en el modelo de texto en español (Validación)

Tasa de aprendizaje	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
$3 \cdot 10^{-6}$	0,6179	0,4037	0,4000	0,6656	0,4983
$3 \cdot 10^{-4}$	0,2621	0,1429	0,6000	0,5000	0,0000

Tabla 3.20: Resultados del modelo de texto sobre el conjunto de prueba

Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
0,5648	0,4253	0,3905	0,6725	0,4162

los resultados obtenidos destacan por empeorar en gran medida, en el caso de audio el valor de κ llega incluso a ser negativo (ver Tabla 3.21 para audio y 3.22 para texto). Si se analizan las matrices de confusión en ambos casos sobre MOUD se concluye que ambos modelos clasifican gran parte de muestras a la misma clase. El de audio clasifica mayormente a negativo, mientras que el de texto a neutral (ver Figura 3.6 para audio y Figura 3.7 para texto). No se analizan los de imagen debido a que, al obtener estos resultados en audio y texto, se decidió estudiar la base de datos y el proceso seguido para obtenerlos. Tras realizar un análisis de los vídeos y las etiquetas proporcionadas, se encontraron numerosas inconsistencias. Por ejemplo, el clip que contiene la siguiente transcripción: "Sobre todo y- como gente como yo que tengo el pelo largo, es que es imposible este producto, pero y mira-.^{es}tá clasificado con un sentimiento positivo, cuando el texto tiene una clara polaridad negativa y el audio y el vídeo también muestran un sentimiento negativo.

Teniendo esto en consideración, se decidió descartar también la base de datos MOUD para la validación del modelo multimodal. De esta forma afrontamos uno de los principales retos de este trabajo, puesto que no contábamos con ninguna base de datos en español que nos permitiera validar el modelo propuesto. La solución sería crear una nueva base de datos y clasificarla o no habría manera de realizar ningún experimento para aplicar técnicas de fusión este idioma. Por lo tanto, debido a que no se dispone del tiempo ni de las herramientas para poder realizarlo, y con el fin de poder llegar a los objetivos plateados en este trabajo, se decidió entrenar los modelos propuestos para el español con datos en inglés. De esta forma, pudimos validar la propuesta de fusión de las redes al utilizar la base de datos MELD, que contiene vídeos etiquetados con emociones en idioma inglés. A continuación se detalla el proceso de entrenamiento y de fusión de los datos.

Tabla 3.21: Resultados del modelo de audio para MOUD

Conjunto	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
Entrenamiento	0,762	0,7746	0,7071	0,9224	0,5671
Validación	0,6728	0,6481	0,5718	0,8498	0,3887
Prueba (MOUD)	0,4144	0,4044	0,2857	0,6157	-0,0680

3.5.2. Base de datos seleccionada y limitaciones encontradas

Teniendo en cuenta que, como se explica en el anterior apartado, no ha sido posible encontrar una base de datos fiable en español, se analizan diferentes bases de datos multimodales en otros idiomas, teniendo en cuenta los corpus más comunes explicados en el

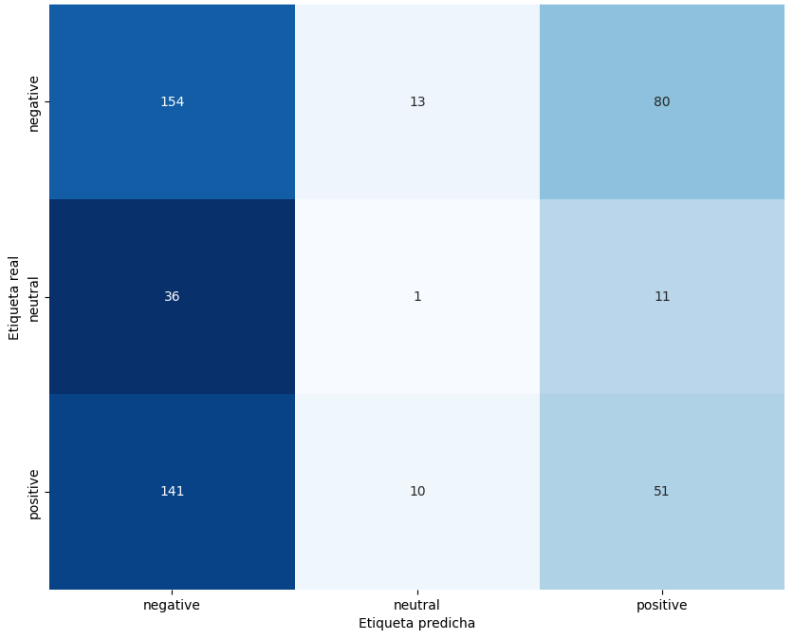

Figura 3.6: Matriz de confusión del modelo de audio sobre el conjunto de MOUD

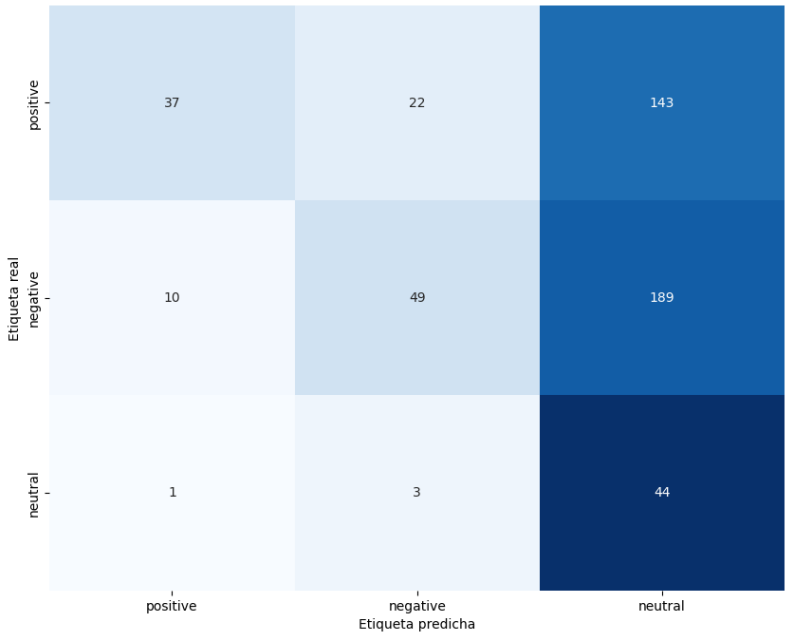

Figura 3.7: Matriz de confusión del modelo de texto sobre el conjunto de MOUD

Tabla 3.22: Resultados del modelo de texto para MOUD

Conjunto	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
Entrenamiento	0,8516	0,8752	0,8497	0,9018	0,7682
Validación	0,7226	0,7539	0,7216	0,8083	0,5767
Prueba (MOUD)	0,2610	0,4325	0,2693	0,5704	0,0923

Capítulo 2.6.1. Se acabó eligiendo la base de datos MELD, ya que proporcionaba todas las modalidades necesarias con los datos clasificados en las 6 emociones básicas y la neutral. Sin embargo, esta base de datos también presenta limitaciones, ya que la distribución por clases está, de nuevo, desbalanceada, siendo miedo y asco las que menos tienen (ver Tabla 3.23). Dado que será necesario volver a entrenar los modelos en este conjunto para audio y texto, y que se utilizará como base para la fusión de datos, es importante tener en cuenta esta limitación al realizar el análisis. En la siguiente sección ampliaremos los detalles del entrenamiento de los modelos de texto y audio en inglés.

Tabla 3.23: Distribución de clases en MELD [61] [8]

Emoción	Entrenamiento	Validación	Prueba
anger	1109	153	345
disgust	271	22	68
fear	268	40	50
joy	1743	163	402
neutral	4710	470	1256
sadness	683	111	208
surprise	1205	150	281
Total	9989	1109	2610

3.6 Fusión multimodal

Con el fin de generar un vector de probabilidad agregado que resuma la información de cada uno de los modelos ya entrenados, es necesario generar una estrategia de fusión que combine los tres modos y que asegure un rendimiento superior a cada uno de los clasificadores en individual. En primer lugar, hay que procesar las muestras de la base de datos y entrenar los nuevos modelos de audio y texto. Estos serán luego utilizados junto al de imagen, ya obtenido, para predecir instancias de la base de datos multimodal y sobre ello aplicar diferentes técnicas de fusión con el fin de mejorar las predicciones de los modelos por sí solos.

3.6.1. Procesado de datos

El procesamiento que se ha llevado a cabo para cada una de las modalidades de las bases de datos ha sido la siguiente:

- **Imagen**: Teniendo los vídeos se obtienen todos los "*frames*" de cada uno de ellos gracias a la librería *cv2*. Hay que destacar que, debido al procesado de eliminar donde no se detectan caras, hubo muestras que se acabaron eliminando y no se tendrán en cuenta para la agregación.

- **Audio**: Los vídeos de cada muestra se procesan para extraer el audio. Y se clasifican según como aparecen en la base de datos.
- **Texto**: En este caso no hay que aplicar ningún procesado, ya que se tiene el texto transcrito.

En cuanto a la partición de los conjuntos, seguimos las directrices definidas por los propietarios de este conjunto de datos. El conjunto de prueba es clave para realizar los experimentos para la fusión de los modelos.

3.6.2. Entrenamiento de los nuevos modelos

Los modelos para audio y texto derivan de entrenar los mejores modelos de audio y texto con la misma arquitectura y configuración definidas en cada uno de sus apartados para esta nueva base de datos. Esto se realiza debido a que, en el caso del texto, se está aplicando un modelo BertT multilingüe, lo que permite al modelo ajustarse correctamente a este cambio en el idioma. En cuanto a audio, el extenso sistema de extracción de características se cree que podrá hacer frente a esta modificación. Es importante recordar que el objetivo principal de este trabajo es mejorar la precisión de los modelos a través de la fusión. Los datos que se utilizaron para entrenar estos modelos se han procesado de la misma manera que para los mejores modelos individuales. En cuanto a texto, habrá procesados que no afectan debido a que no son *tweets*.

Los nuevos modelos de audio y texto empeoran los resultados en audio en un 30 % y en texto un 20 % en la Kappa de Cohen si se comparan respecto a los obtenidos en validación en los modelos individuales en español. Esta diferencia puede deberse al desbalanceo en la base de datos, ya que, para audio este desabalnceo no se presentaba al realizar el estudio inicial y, en texto, se aplicaban técnicas para hacerle frente. Adicionalmente, en audio puede deberse a que la arquitectura creada funciona correctamente para audios limpios, sin ruidos de fondo y que son actuados a conciencia para la emoción, pero no para audios que pueden tener ruidos de todo tipo y que no están modelizados con la idea de mostrar la emoción. En el caso del texto, el modelo no empeora tanto los resultados como el de audio, esto puede deberse a que se utilizó un modelo pre-entrenado no especificado únicamente para *tweets*, pero, por otra parte, se pierde precisión debido al preprocesado para *tweets*. (ver Tabla 3.24 para audio y Tabla 3.25 para texto).

Tabla 3.24: Resultados del modelo de audio en MELD

Conjunto	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
Entrenamiento	0,481	0,253	0,1324	0,8010	0,0456
Validación	0,4197	0,0921	0,123	0,7418	0,0351

Tabla 3.25: Resultados del modelo de texto en MELD

Conjunto	Precisión	Exhaustividad	F1-Score	AUC	Cohen κ
Entrenamiento	0,6524	0,3577	0,3719	0,6389	0,4591
Validación	0,5681	0,3097	0,308	0,6086	0,36

Los resultados obtenidos sobre el conjunto reservado para la fusión en cada uno de los modelos, se recogen en el siguiente capítulo, junto con los de los sistemas de fusión planteados. Para el caso de imágenes, se analizan dos maneras diferentes de obtener la probabilidad y predicción de cada muestra:

- Máximo: Se obtienen las del "*frame*" que presenta el máximo valor en cuanto a probabilidad de predicción de alguna clase.
- Promedio: Calcular la probabilidad promedio de todas las clases de cada "*frame*" del video.

Se obtienen mejores resultados al aplicar el promedio, por ello se calcularon las probabilidades que se utilizaron para los diferentes sistemas de fusión de esta manera.

3.6.3. Sistemas de fusión

La agregación encargada de generar el sistema de fusión tiene como entrada los tres vectores de probabilidad asociados a las predicciones de cada uno de los tres modelos. Para facilitar la notación, consideraremos los vectores (p_i^I, p_i^A, p_i^T) como la salida de los modelos de imagen (I), audio (A) y texto (T) de la muestra i del conjunto de datos. Queda claro que, por el tipo de tarea de clasificación al que nos enfrentamos, se tiene que $p_i^I, p_i^A, p_i^T \in [0,1]^7$ para cada entrada i del conjunto. A partir de cada uno de estos vectores, el sistema de fusión será el encargado de generar el vector de probabilidad agregado p_i^F que emplearemos para validar el sistema multimodal.

Para la fusión, hemos tenido en cuenta varios tipos de funciones y enfoques para generar el sistema. Estos se dividen en dos tipos. Por una parte, los enfoques basados en una agregación estadística invariante y, por otra parte, un enfoque alimentado mediante un algoritmo de aprendizaje automático que ofrece una perspectiva inteligente al sistema.

En cuanto al enfoque estadístico base, hemos decidido aplicar las conocidas medias pitagóricas como función de agregación. Estas son tres y se conocen como las medias armónica, geométrica y aritmética. La justificación sobre su uso se basa en su simplicidad y las propiedades de invariancia (valor e intercambio) y homogeneidad que se satisfacen [4]. Asimismo, cada una de ellas ofrece una característica diferente que resulta útil al combinar probabilidades. Cada una de estas medias se define de la siguiente manera, donde hemos añadido también su justificación, para cada muestra i del conjunto.

- Media armónica: Sirve para considerar todos los valores de la distribución de probabilidad de cada modalidad de forma equilibrada, llegando a ser más representativo para los casos que presenten una falsa predicción por su carácter divisivo.

$$3 \cdot \left(\frac{1}{\frac{1}{p_{1i}^I} + \frac{1}{p_{1i}^A} + \frac{1}{p_{1i}^T}}, \quad \ldots, \quad \frac{1}{\frac{1}{p_{7i}^I} + \frac{1}{p_{7i}^A} + \frac{1}{p_{7i}^T}} \right).$$

- Media geométrica: Sirve para combinar probabilidades con valores extremos, es decir 0 y 1, ya que o bien anula o bien mantiene la probabilidad al fusionarla debido a su carácter multiplicativo.

$$\left(\sqrt[3]{p_{1i}^I \cdot p_{1i}^A \cdot p_{1i}^T}, \quad \ldots, \quad \sqrt[3]{p_{7i}^I \cdot p_{7i}^A \cdot p_{7i}^T} \right)$$

- Media aritmética: También conocida como promedio, es el tipo de agregación más empleada para los sistemas de fusión multimodal. Es la media con interpretación más intuitiva y su fusión resulta ser la más conveniente para que todas las modalidades contribuyan con el mismo impacto gracias a su carácter aditivo lineal.

$$\frac{1}{3} \cdot \left(p_{1i}^I + p_{1i}^A + p_{1i}^T, \quad \ldots, \quad p_{7i}^I + p_{7i}^A + p_{7i}^T \right)$$

Respecto a los algoritmos de aprendizaje automático, se implementan diversos tipos de clasificadores básicos para cubrir diferentes tipos de enfoques. Donde además, resultan ser también muy utilizados en este tipo de tareas por su efectividad y bajo coste computacional. Se decidió utilizar los siguientes algoritmos: KNN, SVM, Naive Bayes, Decision tree, Random Forest [3], AdaBoost y MLP. En [53], se explican cada uno de los algoritmos detallados anteriormente, a excepción de Random Forest. Para la implementación de los algoritmos de aprendizaje automático se aplican inicialmente los parámetros por defecto de cada uno para la librería Sklearn [58]. Después de realizar toda la fase experimental, el clasificador MLP, conocido como perceptrón multicapa, es el que obtuvo el mejor rendimiento. Cabe dstacar que este modelo se compone de una capa oculta y una de salida. Finalmente, se aplicó una validación cruzada de 5 pliegues y una optimización de hiperparámetros a través de una malla de valores:

- **Lambda**: es un parámetro de regularización y para el que se prueba los valores de 1, 0,1, 0,01, 0,001, 0,0001 y 0,00001. Para generar una maya distribuida logarítmicamente.
- **Número de Neuronas**: Se estudian todos los valores enteros entre 20 y 30 puesto que la entrada del modelo está compuesta por 21 elementos. De este modo estudiamos desde el caso base hasta casos de mayor complejidad.

Los hiperparámetros que ofrecen los mejores resultados son de una lambda 0,01 y 25 neuronas. De modo que esta es la estrategia de fusión final que se considera para este trabajo.

CAPÍTULO 4
Resultados

En este capítulo se presentan los resultados finales de este trabajo. Estos son los obtenidos de los modelos finales aplicados al conjunto de datos multimodal en el que se realiza la agregación. Y los resultantes tras aplicar los sistemas de fusión. Para las métricas, no se estudian ni "*precision*" ni "*recall*", ya que no están bien definidas en el caso en que un modelo discrimina automáticamente una clase, es decir, cuando no predice una clase en particular. Debido a esto, se ha decidido mostrar el resto de las métricas previamente estudiadas en este trabajo, además del "*Cross-Entropy Error*" (CCE), referido a la función de pérdida utilizada.

En la Tabla 4.1, se plasman los resultados de los modelos individuales y los distintos tipos de agregación considerados para la estrategia de fusión multimodal. En cuanto a los resultados del modelo de imagen, se observa una notable disminución en el rendimiento al aplicarlos sobre estos datos. Esto puede deberse a que, tratándose de una serie, es posible que las caras reconocidas y obtenidas no correspondan al hablante, además de que presentan más oclusiones y pueden ser imágenes con mayor variabilidad en cuanto a contexto y escenario. Los mejores resultados de los modelos individuales los engloba el modelo de texto con una diferencia de más de 30 % de Cohen κ con respecto al de imagen y audio. Uno de los sistemas que consigue mejorar lo obtenido por el modelo de texto es KNN, que mejora en menos de un 1 % el valor de Cohen κ aunque el CCE empeora en más de dos puntos. Por otra parte, Naive Bayes mejora ligeramente el CCE, pero empeora cerca de un 3 % Cohen κ respecto a lo propuesto por texto. Finalmente, la estrategia final mejora cerca de un 6 % el valor de Cohen κ y también mejora CCE comparándolo con el modelo de texto. Se ha cumplido de esta manera el objetivo principal de este trabajo que es, a través de un sistema de fusión, mejorar el funcionamiento de modelos individuales.

En cuanto a la matriz de confusión de la estrategia final, nos muestra que asco y miedo no son clasificadas en absoluto y qué neutral clasifica muestras no correspondientes (ver Figura 4.1). Aunque esto era esperable si se tiene en cuenta los modelos iniciales y el desbalanceo en la base de datos. Al comparar esta matriz de confusión con respecto a las de cada uno de los modelos individuales de imagen, audio y texto (Figuras 4.2, 4.3 y 4.4 respectivamente) se comprueba esto claramente, ya que las matrices, tanto de imagen, pero sobre todo de audio, no dejan lugar a una gran mejora. Al comparar la de texto con respecto a la propuesta final se observa que esta última se adapta mejor a las clases, concretamente la mejora principal se obtiene para la clase de enfado.

Tabla 4.1: Resultados finales sobre el conjunto destinado para realizar la fusión

	Fusión	CCE	Precisión	AUC	Cohen κ
Modos	Imagen	2,1798	0,4421	0,5782	0,0685
	Audio	1,5682	0,4793	0,5250	0,0015
	Texto	1,2000	0,6126	0,7935	0,4109
Medias	Armónica	1,4998	0,5599	0,6994	0,2705
	Geométrica	1,3166	0,5618	0,7727	0,2405
	Aritmética	1,2708	0,5658	0,7957	0,2369
Modelos ML	KNN	3,8737	0,6306	0,7387	0,4182
	SVM	1,1348	0,6110	0,7843	0,3621
	Naive Bayes	3,5936	0,5697	0,7474	0,3815
	Decision Tree	18,8361	0,4774	0,5960	0,2732
	Random Forest	2,4915	0,6110	0,7458	0,3984
	AdaBoost	1,8612	0,5913	0,6841	0,3976
	Estrategia final	1,0868	0,6561	0,8238	0,4744

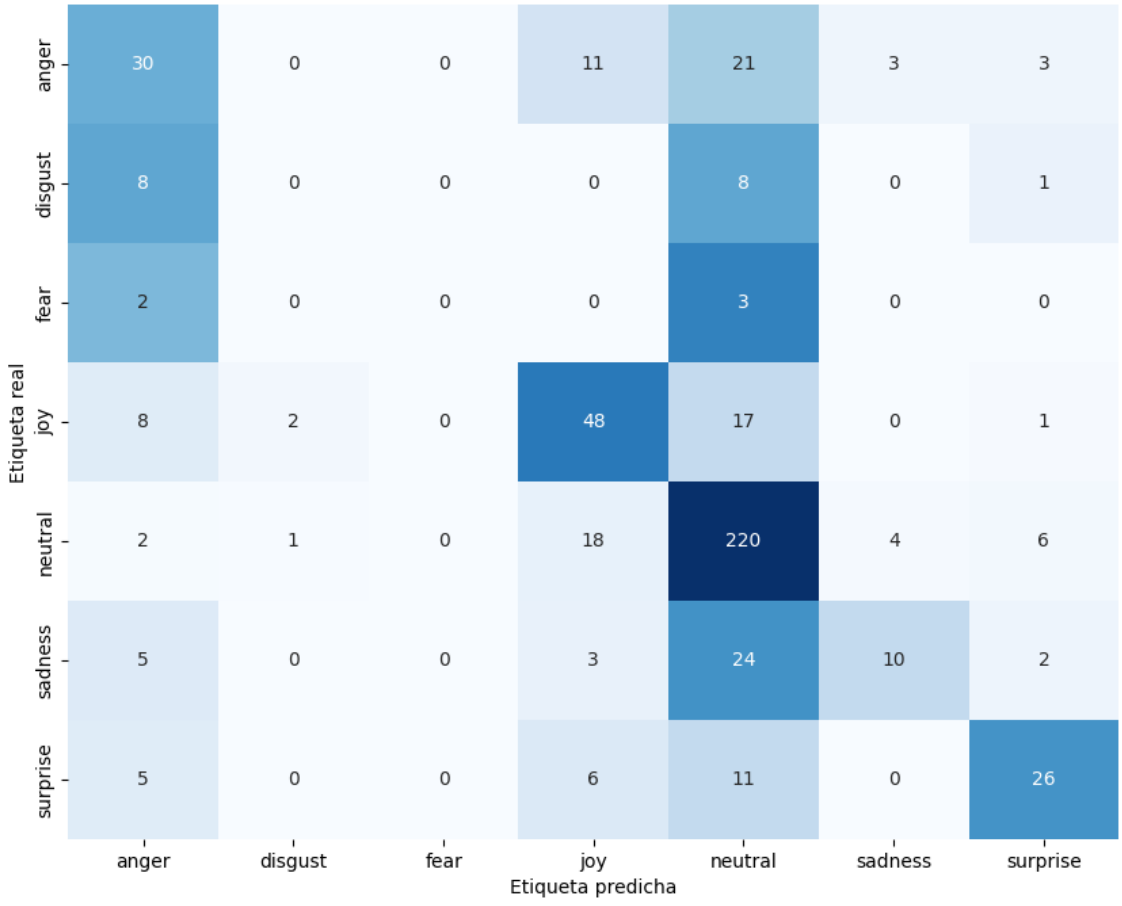

Figura 4.1: Matriz de confusión estrategia final

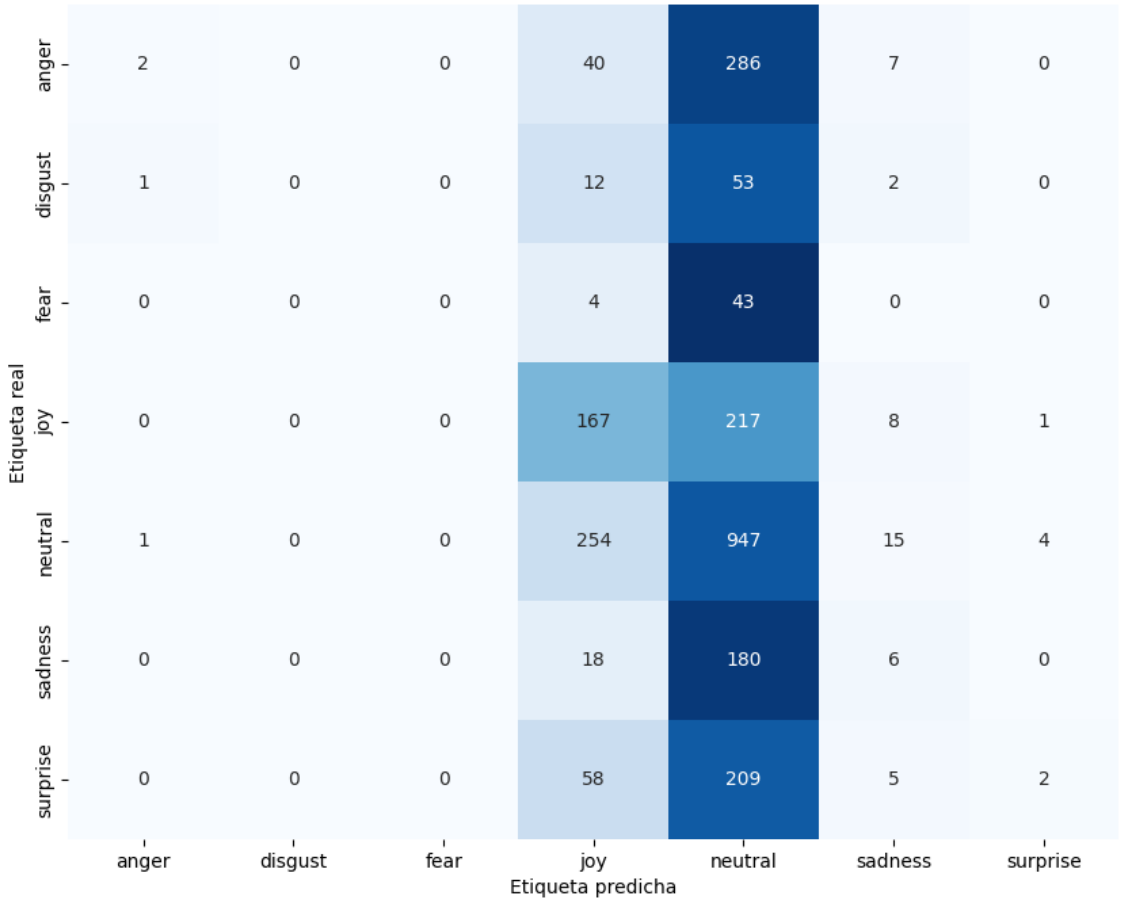

Figura 4.2: Matriz de confusión modelo de imagen en MELD

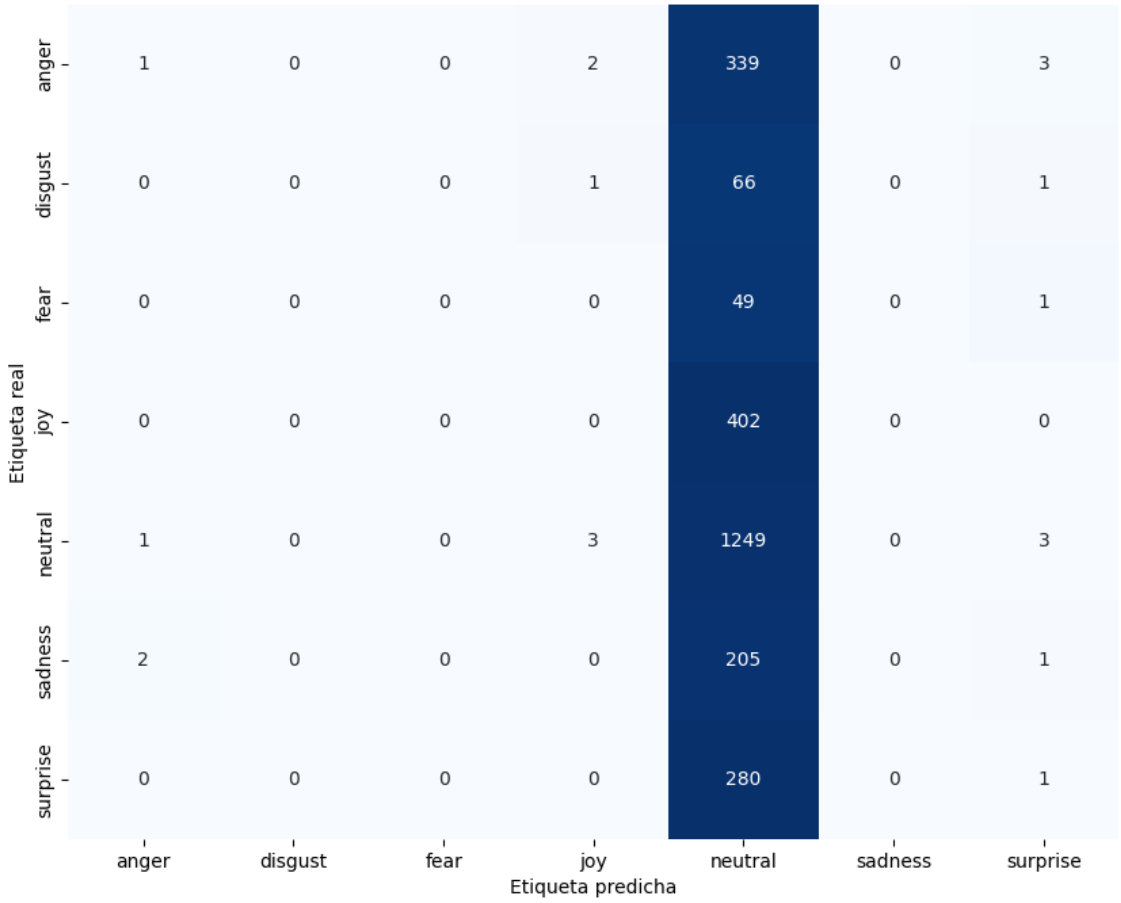

Figura 4.3: Matriz de confusión modelo de audio en MELD

Figura 4.4: Matriz de confusión modelo de texto en MELD

CAPÍTULO 5
Conclusiones y trabajos futuros

5.1 Conclusiones

En este trabajo se ha desarrollado un clasificador multimodal para el reconocimiento de emociones. Para ello, se ha realizado un análisis en profundidad de la literatura existente sobre las técnicas de inteligencia artificial utilizadas para el reconocimiento de emociones en las distintas modalidades. De este análisis se extrajeron los modelos de clasificación de emociones en texto, voz e imagen, que posteriormente fueron fusionados para completar la propuesta del modelo multimodal.

Para entrenar los modelos, se llevó a cabo un análisis de los principales corpus existentes tanto en inglés como en español. En una primera aproximación al clasificador multimodal, se entrenaron los modelos con bases de datos en español. Sin embargo, a la hora de validar la propuesta multimodal mediante la fusión de los modelos, se hacía necesario contar con una base de datos multimodal etiquetada en español. Para ello, se revisaron las principales bases de datos multimodales y no se pudo encontrar ninguna que cubriera de forma adecuada los requerimientos necesarios, ya fuera por dificultades a la hora de acceder a los datos o el etiquetado de los mismos. Teniendo esto en cuenta, se decidió recurrir a una base de datos multimodal que contenía vídeos etiquetados en inglés. Se realizó un entrenamiento de los sistemas usados para el español con la base de datos en inglés y se procedió a realizar la validación de la fusión de los clasificadores de texto, audio e imagen. Los resultados mostraron una mejora en la clasificación de las emociones al aplicar el modelo de fusión frente a los modelos individuales. Sin embargo, hubo emociones como el miedo o el asco que no obtuvieron buenos resultados en la clasificación. Esto se debió en gran parte a la estructura del corpus, ya que las muestras estaban muy desbalanceadas.

A pesar de estas limitaciones, los resultados son prometedores y marcan una línea para las futuras investigaciones sobre la clasificación de emociones mediante el uso de clasificadores multimodales. Hemos conseguido superar estos obstáculos y completar todos los objetivos y demás hitos planteados al inicio de este trabajo. Aportando un gran conocimiento y análisis en el campo, además de haber sido capaces de detectar las limitaciones y problemas aún existentes dentro del mismo.

5.2 Trabajos futuros

Este trabajo presenta una sólida base de investigación que se puede continuar desarrollando. Teniendo en cuenta también las limitaciones especificadas, el trabajo futuro podría llevarse a cabo a partir en diferentes líneas. En primer lugar, una de las principa-

les áreas de mejora es la creación de una base de datos multimodal en español. Este es un objetivo complejo, ya que requiere la recolección de datos en español, su etiquetado y validación. La disponibilidad de una base de datos de este tipo sería fundamental para mejorar y validar los modelos desarrollados.

En segundo lugar, sería interesante analizar y aplicar técnicas de balanceo de datos para mejorar los modelos actuales. El entrenamiento con conjuntos de datos más equilibrados puede conducir a un rendimiento significativamente mejorado de los modelos.

En tercer lugar, para los modelos de audio, sería conveniente explorar nuevas arquitecturas y sistemas, además de mejorar los métodos especificados en este documento con diferentes técnicas y preprocesado diferentes. También sería beneficioso encontrar o crear bases de datos con una mayor variedad de hablantes.

Por otro lado, para el modelo de texto, sería interesante investigar métodos que mejoren el modelo actual mediante la inclusión de técnicas de procesamiento no basadas en tweets. Es crucial encontrar o crear bases de datos más extensas, con más muestras y mejor balanceadas. Del mismo modo, el modelo de imagen también puede beneficiarse de la experimentación con nuevas arquitecturas y sistemas, así como de la mejora de las técnicas y preprocesamiento especificados en este documento. Además, sería útil encontrar o crear bases de datos que estén mejor balanceadas. Finalmente, para mejorar el modelo multimodal, se deben explorar más técnicas de agregación para la realización del modelo. Además, es importante investigar diferentes arquitecturas multimodales que no se basen únicamente en la agregación de las predicciones de los modelos individuales.

Bibliografía

[1] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023.

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[3] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[4] Peter S Bullen, Dragoslav S Mitrinovic, and Means Vasic. *Means and their inequalities*, volume 31. Springer Science & Business Media, 2013.

[5] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[6] ELRA catalogue. Emotional speech synthesis database. http://catalog.elra.info, ISLRN: 477-238-467-792-9, ELRA ID: ELRA-S0329. Accessed: febrero 2024.

[7] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. Vitfer: Facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, Aug 2022.

[8] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[10] CrowdFlower. Twitter us airline sentiment. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment, 2015. Accessed: abril 2024.

[11] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Dumitru, Ian Goodfellow, Will Cukierski, and Yoshua Bengio. Challenges in representation learning: Facial expression recognition challenge. https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge, 2013. Accessed: febrero 2024.

[15] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[16] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003.

[17] Fernando Elkfury and Jorge Ierache. Clasificación y representación de emociones en el discurso hablado en español empleando deep learning. *RISTI-Revista Ibérica de Sistemas e Tecnologías de Información, versión impresa ISSN*, pages 1646–9895, 2021.

[18] Mohammad Reza Falahzadeh, Edris Zaman Farsa, Ali Harimi, Arash Ahmadi, and Ajith Abraham. 3d convolutional neural network for speech emotion recognition with its realization on intel cpu and nvidia gpu. *IEEE Access*, 10:112460–112471, 2022.

[19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[20] Nico H Frijda. The place of appraisal in emotion. *Cognition & Emotion*, 7(3-4):357–387, 1993.

[21] Esteban Garcia-Cuesta, Antonio Barba Salvador, and Diego Gachet Pãez. Emomatchspanishdb: study of speech emotion recognition machine learning models in a new spanish elicited database. *Multimedia Tools and Applications*, 83(5):13093–13112, 2024.

[22] José Antonio García-Díaz, Angela Almela, and Rafael Valencia-García. Umuteam at tass 2020: Combining linguistic features and machine-learning models for sentiment classification. In *IberLEF@ SEPLN*, pages 187–196, 2020.

[23] Manuel García-Vega, Manuel Carlos Diáz-Galiano, MA García-Cumbreras, Flor Miriam Plaza Del Arco, Arturo Montejo-Raéz, Salud María Jiménez-Zafra, E Martínez Cámara, César Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, et al. Overview of tass 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain*, pages 163–170, 2020.

[24] Adam Geitgey. face_recognition: The world's simplest facial recognition api for python and the command line. https://github.com/ageitgey/face_recognition, 2024. Accessed: febrero 2024.

[25] José Ángel González-Barba, José Arias-Moncho, Lluis Felip Hurtado Oliver, and Ferran Pla Santamaría. Elirf-upv at tass 2020: Twilbert for sentiment analysis and emotion detection in spanish tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, pages 179–186. CEUR, 2020.

[26] Steven Greenberg and Brian ED Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *1997 IEEE international conference on acoustics, speech, and signal processing*, volume 3, pages 1647–1650. IEEE, 1997.

[27] SL Happy, Priyadarshi Patnaik, Aurobinda Routray, and Rajlakshmi Guha. The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, 8(1):131–142, 2015.

[28] Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, and Yifu Xua. Dynamic invariant-specific representation fusion network for multimodal sentiment analysis. *Computational Intelligence and Neuroscience*, 2022(1):2105593, 2022.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[31] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, Apr 2019.

[32] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. Automatic speech emotion recognition using machine learning. *Social Media and Machine Learning [Working Title]*, 2019.

[33] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on FER2013. *CoRR*, abs/2105.03588, 2021.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 88–97, 2017.

[36] Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. Multimodal sentiment analysis: A survey. *Displays*, page 102563, 2023.

[37] Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[38] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[39] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[40] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[42] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[43] Daniel Lundqvist, Anders Flykt, and Arne Öhman. Karolinska directed emotional faces. *Cognition and Emotion*, 1998.

[44] Michael J Lyons. .[ex]cavating aire-excavated: debunking a fallacious account of the jaffe dataset. *arXiv preprint arXiv:2107.13998*, 2021.

[45] Michael J Lyons, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets (ivc special issue). *arXiv preprint arXiv:2009.05938*, 2020.

[46] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, Apr 2023.

[47] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages.

[48] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939, Nov 2022.

[49] Olivier Martin, J Adell, A Huerta, Irene Kotsia, Arman Savran, and Raphael Sebbe. Multimodal caricatural mirror. In *eINTERFACE'05-Summer Workshop on Multimodal Interfaces*, 2005.

[50] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[51] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[52] Multimodal Multimedia and Machine Learning Lab. Moud dataset. http://multicomp.cs.cmu.edu/resources/moud-dataset/, 2024. Accessed: abril 2024.

[53] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[54] University of Surrey. Speech and audio visual emotion (savee). http://kahlan.eps.surrey.ac.uk/savee/. Accessed: abril 2024.

[55] University of Toronto. Toronto emotional speech set (tess). https://tspace.library.utoronto.ca/handle/1807/24487. Accessed: abril 2024.

[56] Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Alice Baird, and Björn Schuller. Categorical vs dimensional perception of italian emotional speech. 2018.

[57] Brian Parkinson. Emotions are social. *British journal of psychology*, 87(4):663–683, 1996.

[58] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[59] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[60] Ferran Pla and Lluís-F Hurtado. Sentiment analysis in twitter for spanish. In *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19*, pages 208–213. Springer, 2014.

[61] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[62] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[63] Prof. Brian Rowe. Emotion detection from facial expressions. https://kaggle.com/competitions/emotion-detection-from-facial-expressions, 2016. Accessed: febrero 2024.

[64] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[65] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[66] Kashfia Sailunaz and Reda Alhajj. Emotion and sentiment analysis from twitter text. *Journal of computational science*, 36:101003, 2019.

[67] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[68] SEPLN. Tass 2020 - taller de análisis de sentimientos en la sepln. http://tass.sepln.org/2020/#tasks, 2020. Accessed: febrero 2024.

[69] Seriousran. Apple twitter sentiment texts. https://www.kaggle.com/datasets/seriousran/appletwittersentimenttexts, 2021. Accessed: abril 2024.

[70] Jagjeet Singh, Lakshmi Babu Saheer, and Oliver Faust. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6):5140, 2023.

[71] Slythe. Apple twitter sentiment from crowdflower. https://www.kaggle.com/datasets/slythe/apple-twitter-sentiment-crowdflower?select=Apple-Twitter-Sentiment-DFE.csv, 2016. Accessed: abril 2024.

[72] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[73] Jianguo Sun, Hanqi Yin, Ye Tian, Junpeng Wu, Linshan Shen, and Lei Chen. Two-level multimodal fusion for sentiment analysis in public security. *Security and Communication Networks*, 2021(1):6662337, 2021.

[74] Amira Samy Talaat. Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, 10(1):110, 2023.

[75] TASS 2013. Tass 2013 - about. http://tass.sepln.org/2013/about.php, 2013. Accessed: febrero 2024.

[76] Ivona Tautkutė and Tomasz Trzciński. Classifying and visualizing emotions with emotional dan. *Fundamenta Informaticae*, 168(2-4):269–285, 2019.

[77] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[79] V Vetriselvi et al. Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning. *Applied Soft Computing*, 157:111553, 2024.

[80] Fan Wang, Shengwei Tian, Long Yu, Jing Liu, Junwen Wang, Kun Li, and Yongtao Wang. Tedt: transformer-based encoding–decoding translation network for multimodal sentiment analysis. *Cognitive Computation*, 15(1):289–303, 2023.

[81] Richard Wollheim. *On the emotions*, volume 288. Yale University Press, 2008.

[82] Chen Xi, Guanming Lu, and Jingjie Yan. Multimodal sentiment analysis based on multi-head attention mechanism. In *Proceedings of the 4th international conference on machine learning and soft computing*, pages 34–39, 2020.

[83] Yale. The extended yale face database b. 2001.

[84] Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1801. NIH Public Access, 2020.

[85] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[86] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.