Facultad de Ingeniería
Escuela de Ingeniería Informática

RECONOCIMIENTO MULTIMODAL DE EMOCIONES UTILIZANDO CONJUNTOS DE DATOS EN AMBIENTES NO CONTROLADOS

Por

Facundo Vicente Martínez Gullé

Trabajo realizado para optar al Título de
INGENIERO CIVIL EN INFORMÁTICA
Prof. Guía: Ana Aguilera Faraco
Prof. Co-Referente: Diego Mellado Carreño
Diciembre 2023

Resumen

El reconocimiento multimodal de emociones se centra en identificar las emociones de un sujeto en situaciones específicas, utilizando inteligencia artificial en diversas modalidades, como imagen, audio y texto. Este Trabajo de Título está basado en el proyecto "An Assessment of In-the-wild Datasets for Multimodal Emotion Recognition", que resalta la necesidad de unificar, limpiar y transformar conjuntos de datos para re-entrenar un modelo de Deep Learning. El objetivo es mejorar las predicciones, generar visualizaciones y realizar un análisis comparativo con los resultados del proyecto base, evaluando así si se logra mejorar el reconocimiento multimodal de emociones en entornos no controlados. Se presenta un marco de trabajo que integra información de las modalidades de imágenes faciales, audio y texto. Esto se logra mediante el uso de técnicas de Deep Learning para adaptar modelos pre-entrenados a conjuntos de datos no controlados. Este enfoque será evaluado mediante experimentos utilizando la base de datos llamada MERDWild, la cual fue creada mediante la unificación de los conjuntos de datos *in-the-wild* llamados AFEW, AffWild2 y MELD, para demostrar su eficacia en el reconocimiento multimodal de emociones, utilizando MAFW como dataset de pruebas. Los resultados finales revelaron los desempeños de los modelos específicos para cada una de las tres modalidades, así como los desempeños de las combinaciones de modalidades en el conjunto de validación y en el conjunto de pruebas MAFW. Las técnicas de Deep Learning utilizadas para los modos de imágenes faciales, audio y texto son VGG19, ResNet50 y DialogXL, respectivamente. Para fusionar las predicciones obtenidas de cada uno de los modelos se utilizó el método de fusión EmbraceNet+. Los resultados evidencian la presencia de sobreentrenamiento y se sugieren posibles soluciones para evitarlo. Adicionalmente, se generó una base de datos homogénea y de alta calidad compuesta por 15.873 archivos de audio, 905.281 imágenes faciales y 15.321 frases. Todos estos elementos superaron los criterios de calidad establecidos, entre ellos se encuentra brillo, contraste y resolución en el caso de las imágenes faciales, nivel peak, distorsión armónica total y promedio de nivel de potencia en audio y en texto el uso de un filtro semántico. Este trabajo proporciona una visión integral del reconocimiento multimodal de emociones, desde la recopilación de datos hasta las visualizaciones de los resultados finales.

Agradecimientos

Primero y ante todo, quiero agradecer a mi profesora guía, Ana Aguilera, por su orientación experta, paciencia y apoyo constante a lo largo de este proceso. Su experiencia y dedicación fueron fundamentales para dar forma y mejorar este trabajo.

Agradezco también a mi profesor co-referente, Diego Mellado, por sus valiosas sugerencias y contribuciones que enriquecieron considerablemente este proyecto.

A Sofía Lavado Zagal, mi pareja, le agradezco por su paciencia, comprensión y permanente apoyo. Tu presencia hizo que este desafío fuera mucho más llevadero.

A mis padres, Solange Gullé y Fernando Martínez, les debo un agradecimiento especial. Su amor incondicional, apoyo financiero y palabras de aliento fueron una parte importante de mi motivación durante este viaje académico.

A mi hermano, Fernando Javier Martínez Gullé. Quiero expresar mi agradecimiento por ser no solo mi hermano, sino también mi mejor amigo y modelo a seguir.

Quiero extender mi agradecimiento a mis suegros, María Zagal y Luis Lavado, por su cálido apoyo y por acogerme en su familia con los brazos abiertos.

Agradezco a mis compañeros y amigos por compartir este viaje académico conmigo. Nuestras discusiones y colaboraciones fueron fundamentales para mi crecimiento y aprendizaje.

A todos mis profesores, quienes compartieron su conocimiento y experiencia a lo largo de mi carrera, siempre con las puertas abiertas para brindar apoyo.

Un agradecimiento especial también a los funcionarios de la escuela que trabajaron incansablemente para proporcionar un entorno propicio para el aprendizaje.

Este logro no habría sido posible sin el apoyo de cada una de estas personas. Estoy profundamente agradecido por las contribuciones de todos.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Para poder entender qué es lo que se desarrollará en este Trabajo de Título primeramente se debe comprender que es una emoción, Bisquerra lo define de la siguiente manera: "La emoción es un estado complejo del organismo caracterizado por una excitación o perturbación que predispone a una respuesta organizada." [2] , es decir, la emoción es un estado en el que se encuentra una persona, la cual es causada por una situación que genera una respuesta tanto física como psicológica en el individuo acorde a una situación dada. Para el desarrollo de este proyecto de título se utilizarán siete categorías de emociones, de las cuales seis de ellas fueron definidas por Eknam [3] como emociones básicas, las cuales son enojo, disgusto, miedo, felicidad, tristeza y sorpresa, agregando la séptima categoría neutral.

Existen diferentes tipos de reconocimiento de emociones (ER) que están enfocados en el mismo fin, este es utilizar diferentes recursos y técnicas para identificar y reconocer las emociones humanas [1]. Uno de estos tipos de reconocimiento es el unimodal, el cual se centra solo en una modalidad de ER, ejemplos de esto son el ER en textos o imágenes. Por otro lado, se encuentra el reconocimiento multimodal de emociones (MER), el cual está formado por más de un modo de reconocimiento, este tipo es precisamente el que se realizará en este proyecto, el cual es más complejo debido a que se contempla para un mismo escenario distintas perspectivas en términos de modalidades de datos (audio, texto, imagen, etc.), para realizar un análisis aún más completo de la emoción que está sintiendo el sujeto evaluado.

Para realizar el reconocimiento de estas emociones se utiliza Deep Learning, el cual es una rama del Machine Learning en la que se intenta simular el comportamiento del cerebro humano [4], este puede analizar expresiones físicas como la postura corporal, expresiones faciales como la mirada y la contracción de los músculos de la cara [5], también puede analizar la voz, entre otros. Esto se ejecuta gracias a una entrada de datos, la cual puede ser trabajada en diferentes formatos, tales como imágenes, texto, audio, vídeo, señales [6], entre otros.

El presente Trabajo de Título está basado en el proyecto “An Assessment of *In-the-wild* Datasets for Multimodal Emotion Recognition” [1], donde se han analizado siete emociones básicas con distintos datasets en ambientes no controlados (también llamado en inglés "*in-the-wild*"), es decir, conjuntos de datos que se encuentran en un contexto real (sin datos preparados en un laboratorio). En dicho proyecto se ha establecido que se deben unificar, limpiar y equilibrar las clases de los datasets utilizados, para reentrenar el modelo de Deep Learning, y así obtener un mejor MER *in-the-wild*, lo que constituye el objetivo fundamental de este trabajo.

El problema está en que la mayoría de las técnicas de MER se entrenan utilizando conjuntos de datos diseñados y construidos en condiciones controladas, lo que dificulta su aplicabilidad en contexto real con condiciones reales. Además, generalmente, los datasets *in-the-wild* no están diseñados para el reconocimiento multimodal [1].

Para superar estos obstáculos, se hará la trasformación, normalización, limpieza y unificación de los conjuntos de datos AFEW [7], AffWild2 [8] y MELD [9] que se encuentran con datos *in-the-wild* [1], de este modo llevar a cabo un proceso de reentrenamiento del modelo utilizado en el proyecto [1], con el objetivo de mejorar el MER de las modalidades audio, texto e imágenes faciales. Se propone evaluar el conjunto de datos resultante sobre tres arquitecturas, una de ellas es VGG19 [10], la segunda arquitectura se llama ResNet50 [11], y la tercera es DialogXL [12]. VGG19, ResNet50 y DialogXL se utilizan para imágenes, audio y texto respectivamente. Estas tres arquitecturas se evaluarán mediante el método de fusión Embracenet+ [13], con el fin de evaluar su desempeño en términos de diferentes métricas de evaluación que serán detalladas a lo largo de informe.

Dentro de las principales contribuciones de este proyecto se encuentra la unificación de los tres conjuntos de datos mencionados generando un único dataset de calidad llamado "MERDWild: Multimodal Emotion Recognition *in-the-wild*", el cual se encuentra correctamente etiquetado con los nombres de las emociones y sus calidades, para luego realizar el entrenamiento de un modelo basado en Deep Learning, y de esta manera aportar en el avance tecnológico del MER *in-the-wild*.

Este documento comienza en el Capítulo 1, donde se presenta una introducción que aborda el contexto y la problemática de este Trabajo de Título. El Capítulo 2 especifica el marco conceptual y el estado del arte. Posteriormente, el Capítulo 3 contiene la definición del problema, la solución propuesta y la importancia del trabajo, tanto en el ámbito científico como socio-económico y se menciona el objetivo general y los objetivos específicos. Luego, en el Capítulo 4 se aborda el Diseño de la solución, modelando la solución del problema y describiendo tanto las técnicas de análisis como los procesos realizados, junto con el diseño de los experimentos y sus variables, además de la descripción de los datasets utilizados y la interpretación que se debe realizar con respecto a los resultados obtenidos agregando las maneras de visualización de los resultados. En el Capítulo 5 se evidencia tanto la estrategia de implementación del proyecto como también la descripción de las etapas del proceso, esto, junto con el software y hardware utilizados, al término de este capítulo

se encuentran las visualizaciones del trabajo realizado. Posteriormente en el Capítulo 6 se muestran las pruebas realizadas. Luego en el Capítulo 7.2 se encuentra el análisis y discusión de los resultados obtenidos. En el Capítulo 8 se haya la conclusión del proyecto, donde se describe el cumplimiento de los objetivos iniciales, las dificultades que se debieron abordar durante el desarrollo del trabajo, limitaciones y proyecciones. Finalmente, se encuentra la bibliografía y el anexo, dentro de éste último se detalla la implantación.

Capítulo 2

Marco Conceptual y estado del arte

En esta sección se abordan dos temas que proporcionan la información necesaria para comprender el contexto teórico y práctico en el cual se desarrolla el presente Trabajo de Título, justificando su relevancia y originalidad.

2.1. Marco conceptual

A continuación se presentan las distintas definiciones de los temas que son fundamentales para la compresión de este trabajo.

2.1.1. Aprendizaje profundo

Aprendizaje profundo o Deep Learning (DL) en inglés, es un subcampo del Machine Learning (ML) que se basa en redes neuronales con tres o más capas. Estas redes intentan emular el comportamiento del cerebro humano y les permiten "aprender" a partir de grandes cantidades de datos [14].

Los modelos de DL son capaces de admitir distintos tipos de aprendizaje, dentro de los cuales se encuentran tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. El aprendizaje supervisado utiliza conjuntos de datos etiquetados para categorizar o realizar predicciones, esto requiere que exista la intervención humana para etiquetar correctamente los datos de entrada. Por otro lado, el aprendizaje no supervisado no necesita de conjuntos de datos etiquetados y, en su lugar, detecta patrones en los datos y los agrupa en función de cualquier característica distintiva que encuentre. El aprendizaje por refuerzo se trata de un proceso en el que un modelo aprende a ser más preciso para realizar una acción en un entorno, basándose en los comentarios que recibe para maximizar el resultado [14].

Existen diversas arquitecturas de redes neuronales, cada una exhibiendo comportamientos distintos entre sí. Esta variabilidad conlleva a un rendimiento diferenciado en una

misma tarea, dado que cada arquitectura adopta enfoques particulares para abordar los problemas específicos. Para ilustrar estas diferencias, consideremos dos ejemplos destacados: las redes neuronales convolucionales, comúnmente empleadas para el reconocimiento de patrones en imágenes, y las redes neuronales recurrentes, utilizadas en el procesamiento de secuencias de datos o información temporal.

Asimismo, existen técnicas avanzadas para afrontar estos desafíos, entre las cuales se destaca el transfer learning (aprendizaje por transferencia). Esta estrategia permite aprovechar la capacidad de reutilizar modelos preexistentes, capitalizando su conocimiento para abordar con éxito nuevos problemas [15].

Existen diferentes arquitecturas de redes neuronales, algunos ejemplos de ellas son las CNN (Convolutional Neural Network) y las RNN (Recurrent Neural Networks) [16]. Estas dos se comportan de maneras distintas y tienen un mejor rendimiento en tareas diferentes debido a su manera de abordar los problemas. CNN se utiliza usualmente para el reconocimiento de patrones en imágenes y RNN se usa en el ámbito de secuencias de datos o secuencias temporales. También existen técnicas para abordar los problemas, una de ellas es el transfer learning (aprendizaje por transferencia), se utiliza para aprovechar la capacidad de reutilizar modelos existentes y gracias a su conocimiento lograr resolver nuevos problemas [15].

2.1.2. Datasets

Los datasets o conjuntos de datos son la materia prima fundamental del sistema de predicción, ya que se utiliza para entrenar, validar y hacer las pruebas al sistema. Este conjunto histórico de datos se compone de elementos individuales que se desean analizar, cada uno con factores, características o propiedades específicas que son relevantes para la tarea de predicción [17]. Existen dos tipos de datasets, uno de ellos que se encuentra con datos obtenidos en ambientes de laboratorio o en ambientes controlados, también existen aquellos conjuntos de datos en los cuales sus datos fueron rescatados en condiciones del mundo real (ambientes no controlados), en inglés son llamados "*in-the-wild*", cuya traducción directa es "en la naturaleza" [18].

2.1.3. Reconocimiento de emociones

El reconocimiento de emociones (ER) es una disciplina dentro de la inteligencia artificial que se enfoca en el reconocimiento de las formas básicas de expresión emocional que manifiestan las personas. Su principal objetivo es detectar y clasificar estas expresiones afectivas para su análisis y uso en diversas aplicaciones [19]. Existen diferentes modalidades o tipos de entradas de datos en las cuales se pueden identificar emociones, dentro de ellas se encuentran las expresiones faciales [8], la modalidad auditiva [20], la modalidad de escritura [9], la postura de las personas [5], entre otras. En el habla y la visión,

las emociones se pueden transmitir a través de énfasis en las palabras, variaciones en el tono, expresiones faciales, entre otros. En cambio, en el reconocimiento de emociones en el texto no es tan explícito. Se debe interpretar el significado del texto, ya que la mayoría de las veces las expresiones emocionales no son directamente inferibles. En lugar de eso, deben ser deducidas a través de la interpretación del significado de los conceptos y la interacción entre ellos, tal como se describen en el documento [21]. Es importante destacar que el ER es fundamental para la interacción entre el humano y el computador, como también en la interacción humano y robot, debido a que permite que los computadores y los robots puedan responder de una manera más sensible a las señales emocionales de los usuarios, logrando interactuar y comunicarse de una manera más natural [22][23].

Reconocimiento de emociones faciales

El reconocimiento de emociones faciales (FER) permite analizar las expresiones faciales en vídeos o en imágenes, su objetivo es lograr obtener información sobre la emoción que se encuentra expresando el individuo analizado [24]. Aunque esta tecnología presenta grandes oportunidades para su uso en diversos contextos, la complejidad de las expresiones faciales y la participación de tecnologías emergentes, como la inteligencia artificial (IA), conllevan importantes riesgos para la privacidad de los individuos [25]. FER se aplica en diversos campos que presentan distintos desafíos [26]. Un ejemplo de esto se puede ver en el campo de la medicina, en el cual se emplea para la detección de depresión o la clasificación del trastorno del espectro autista [27], como también en el campo de la psicología, al detectar la emoción que está expresando el paciente se puede reconocer el estado emocional por el que está pasando la persona en ese momento [28].

Reconocimiento de emociones en texto

El reconocimiento de emociones en texto (TER) es una tarea esencial en la investigación del procesamiento del lenguaje natural (NLP) [21], esto implica obtener las emociones a partir de una entrada de texto. Las soluciones a esta tarea pueden resultar útiles en diversas aplicaciones, como en servicios de redes sociales para conocer a los usuarios que generan los datos y medir los sentimientos de un usuario en particular en el contexto de un producto, empresa u organización específicos [29]. También tiene aplicaciones en chatbots [30], en la interacción humano-computador, psicología, entre otros [28].

Reconocimiento de emociones en audio

Reconocimiento de Emociones en audio (SER) es un sistema que puede detectar las emociones que se presenten en diferentes muestras de audio [31]. El reconocimiento de emociones del habla puede utilizarse para encontrar el rango emocional en varias grabaciones de audio, como entrevistas de trabajo, llamadas de agentes de centros de llamadas,

transmisión de vídeos y canciones. Además, los sistemas de recomendación o clasificación de música también pueden agrupar canciones según el estado de ánimo y recomendar listas de reproducción especializadas para cada usuario [32].

2.1.4. Métricas de evaluación

Las métricas se utilizan para evaluar el rendimiento del modelo entrenado, y la capacidad del modelo para generalizar sobre datos no vistos determina si es adaptable o no [33]. Algunas de las métricas más comúnmente empleadas para medir el rendimiento incluyen la "accuracy" (exactitud), el "F1-score", el "Recall" y la "Precisión". Donde TP representa "True Positive" (verdaderos positivos), FP representa "False Positive" (falsos positivos), FN representa "False Negative" (falsos negativos), y TN representa "True Negative" (verdaderos negativos). Estos datos son fundamentales para el cálculo de las métricas en la evaluación del modelo [33][34][35]. A continuación, se presentan las fórmulas de cada una de las métricas de evaluación:

- Accuracy: También conocido como "Exactitud" en español, es una métrica que cuantifica el porcentaje de casos en los cuales acierta el modelo generado. Este se calcula mediante la siguiente formula [35]:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{2.1}$$

- Recall: Esta métrica es también conocida en español como "Exhaustividad" y proporciona información sobre la cantidad que el modelo de aprendizaje automático es capaz de identificar. Recall se calcula como muestra la siguiente ecuación [35]:

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

- Precisión: La métrica de precisión permite evaluar la calidad del modelo de aprendizaje automático en tareas de clasificación y se calcula de la siguiente manera [35]:

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

- F1-Score: El valor F1 se emplea para combinar las medidas de Precisión y Exhaustividad (Recall) en un único valor. Esta técnica facilita la comparación del rendimiento combinado de la precisión y la exhaustividad entre diversas soluciones [35]. El cálculo de F1-score se realiza mediante la media armónica entre la Precisión y Recall:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

2.2. Estado del arte

En esta sección se tratará de plasmar la realidad actual que existe en el área del reconocimiento multimodal de emociones que utilizan conjuntos de datos *in-the-wild*. Para esto se realizó una búsqueda exhaustiva de documentos actualizados en sitios como Web of Science, IEEE, ScienceDirect, Google Scholar, MDPI, Arxiv y SpringerLink, los cuales son servicios en línea de información científica. Como criterios de inclusión se utilizó el filtrado de años, para no superar los 10 años de antigüedad de los documentos citados, además otro criterio de inclusión es que los proyectos citados en esta sección deben usar conjuntos de datos *in-the-wild*. Al finalizar, se presenta la Tabla 2.1, que resume los aspectos abordados en esta sección. La tabla consta de cuatro columnas: la primera detalla los conjuntos de datos utilizadas, la segunda indica los conjuntos de datos de prueba, la tercera presenta los resultados obtenidos, ya sea en términos de precisión (accuracy) u otros indicadores relevantes, y finalmente, la cuarta columna enumera los métodos o técnicas trabajados en cada artículo. A continuación se presentan los artículos seleccionados para como estado del arte.

2.2.1. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition *in-the-wild*

Este trabajo se basa en crear un conjunto de datos *in-the-wild* para MER, nace desde la necesidad obtener un conjunto de datos que contenga una amplia variedad de emociones, con una gran cantidad de hablantes de distintas nacionalidades y lenguas. Esto fue realizado mediante el recorte automático de diferentes películas, entrevistas, series de TV, entre otros, provenientes de distintas partes del mundo, tales como India, China, Europa y América. El conjunto de datos está orientada al reconocimiento facial de emociones y dentro de ella se encuentran 10.045 videoclips con tres modalidades de anotaciones: de emoción individual, de emoción múltiple y de texto correspondiente a la descripción de los vídeos en idioma Chino e Inglés. Dentro de las emociones individuales que se encuentran en este dataset están las seis emociones básicas definidas por Eknam agregando también las emociones neutral, contempt, anxiety, helplessness y disappointment. Las emociones múltiples son combinaciones de pares de emociones. Se propuso un modelo multimodal que contempla audio, texto descriptivo y vídeo, obteniendo en las anotaciones de emociones individuales un UAR (Unweighted Average Recall) de 31,00 y un WAR (Weighted Average Recall) de 50,29, con una arquitectura conformada por Resnet18_LSTM y C3D_LSTM [36].

2.2.2. Video-based emotion recognition *in-the-wild* using deep transfer learning and score fusion

En este proyecto se realiza un reconocimiento de emociones (seis emociones básicas) multimodales utilizando el conjunto de datos en ambiente no controlado llamado AFEW. Donde se aborda un enfoque que combina datos del tipo imagen y audio, utilizando un conjunto de clasificadores basados en mínimos cuadrados (conocido en inglés como "partial least squares based classifier"), y combina su salida con la fusión de nivel de decisión (weighted score level fusion). Se obtuvo como resultado 54.55 % de accuracy en el conjunto de pruebas utilizando una arquitectura conformada por SIFT-FUN, LPQ-TOP, LGBP-TOP, VGG-Face, openSmile, CNN-FUN, Weighted Score Level Fusion. En el trabajo realizan un preprocesamiento y limpieza solamente para la modalidad de imágenes [37].

2.2.3. Developing crossmodal expression recognition based on a deep neural model

En este trabajo se aborda el reconocimiento multimodal de emociones (seis emociones básicas) utilizando distintos datasets, dentro de ellos se encuentra FABO, que es utilizado para el reconocimiento de expresiones visuales como la expresión facial y corporal, otro set de datos que se utiliza es SAVEE, utilizado para expresiones audiovisuales, el tercer dataset que se utiliza es EmotiW, en su mayoría para realizar el proceso de validación debido a la ausencia de etiquetas en varias muestras del set de prueba, adicionalmente se utiliza el conjunto de datos GTZAN para el reconocimiento de emociones en música. Para poder manejar los datos multimodales se utilizó CCCNN y se logró un 46,10 % de accuracy en MER, con una arquitectura conformada por Adaboost-based detection, MFCC, CCCNN y SOM [38].

2.2.4. Adaptive Multimodal Emotion Detection Architecture for Social Robots

En este proyecto se aborda el ER (cuatro emociones básicas) multimodal, adaptable y flexible para robots sociales. Se trabajó con el conjunto de datos *in-the-wild* llamada IEMOCAP, el cual fue analizado tanto en modalidad de imagen, audio y texto. Para lograr esta detección de emociones se realizó un preprocesamiento de imagen rescatando la cara de la persona, como también el audio para transcribir automáticamente el texto. Se utilizaron métodos de análisis individuales como CNN y RNN (transformers), y el método de fusión llamado EmbraceNet+ para hacer la predicción final de la emoción obteniendo un accuracy de 77,6 %, con una arquitectura conformada por VGG19, MFCC, DialogXL, Embracenet+ [39].

2.2.5. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition *in-the-wild*

En el presente trabajo se realiza MER con datasets en ambientes no controlados. En el cual se identifican seis emociones básicas para los conjuntos de datos: AFEW, AffectNET, AffWild2 bajo el protocolo de de desafío ABAW. Las modalidades que utiliza son: imágenes, audio y texto. Para lograrlo se utiliza las arquitecturas de red neuronal convolucional (CNN) y la red neuronal recurrente de memoria a largo plazo y a corto plazo (LSTM-RNN), en el cual se realizaron predicciones finales con un esquema de fusión de puntuación ponderada. Se realizó un preprocesamiento para poder captar el texto gracias al análisis del audio. Se realizó un preprocesamiento para eliminar los fotogramas de vídeo de AffectNET, se realizó el recorte y detección de rostros en AffWild2 y se utilizó la semejanza del coseno (mayor a 0,5) para poder comprobar que la cara que se muestra en el vídeo sea considerada como el área correctamente detectada y se logró un accuracy de 48,10 % [40].

2.2.6. MuSe 2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress

En este proyecto se realiza el MER *in-the-wild*. En el cual se aborda un desafío que contempla tres temas, detección de humor, MER categóricas y dimensionales, donde cada tema se desarrolla con diferentes fuentes de datos, dentro de ellas se encuentran: MuSe-Humor, Passau-SFCH y MuSe-Reaction, todas ellas realizan MER con audio, texto e imagen y se encuentran separadas en sets de entrenamiento, validación y pruebas. Estos datasets son provenientes de vídeos con audio donde se reconocen siete clases de emociones (Adoración, Diversión, Ansiedad, Disgusto, Dolor empático, Miedo, Sorpresa), donde se logró un accuracy en MER de 44,13 % (sin considerar valencia), con una arquitectura conformada por OpenSmile, eGeMAPS y LSTM-RNN [41].

2.2.7. Deep Auto-Encoders with Sequential Learning for Multimodal Dimensional Emotion Recognition

En el presente trabajo se realiza MER con datasets en estado no controlado. Para el desarrollo de este proyecto se propuso una arquitectura de red neuronal profunda que consiste en un codificador automático de dos flujos y una memoria a largo plazo para integrar de manera efectiva flujos de señales visuales y de audio para el ER. El set de datos utilizado para este proyecto es RECOLA, el cual posee 6 emociones básicas. Para el preprocesamiento se utilizó Single Stage Headless para detectar rostros en los cuadros de vídeo, ajustando el color y el tamaño de las imágenes obtenidas para luego ser normalizadas, para el preprocesamiento de audio se segmentaron las señales sin preprocesar las frecuencias

que superaban los 0,2 segundos de cuadros de vídeo para obtener una media cero y una varianza unitaria, de esta manera tener en cuenta la variación en el volumen de diferentes altavoces. En este proyecto se obtuvo un RMSE igual a 0,51, es decir RMSE de 51,00 %, con una arquitectura conformada por LSTM [42].

2.2.8. Audiovisual emotion recognition in wild

En el presente proyecto se realiza MER utilizando datos de laboratorio para entrenar el modelo, y como conjunto de pruebas se utiliza un set de datos en condiciones no controladas. Los datasets de entrenamiento son SAVE, eNTERFACE'05 y RML, para la realización de las pruebas se utiliza el conjunto de datos AFEW. Se identificaron seis emociones básicas y se realizó el reconocimiento en las modalidades de audio y vídeo. La red fue entrenada en Convolution Neural Network (AlexNet) y SVM, por otro lado para el preprocesamiento se rescató desde el vídeo solamente los rostros de Viola-Jones y se extrajeron fotogramas clave para evitar entrenar el sistema con imágenes en las que la expresión facial sigue siendo la misma. En este proyecto se utilizaron dichas fuentes de datos para comparar su rendimiento cada una por si sola para luego juntar las tres conjuntos de datos de entrenamiento y predecir emociones utilizando como conjunto de prueba AFEW y se obtuvo como resultado que para realizar la fusión a nivel de características es necesaria la sincronización de muestras de audio y utilizar cuadros para crear vectores de características apropiados. En este proyecto se obtuvo un accuracy de 27,10 % [43].

2.2.9. HEU Emotion: a Large-scale Database for Multi-modal Emotion Recognition *in-the-wild*

Este trabajo se basa en crear un conjunto de datos *in-the-wild* para MER, que nace desde la necesidad obtener un dataset que contenga una amplia variedad de emociones, con una gran cantidad de hablantes de distintas nacionalidades y lenguas. Posee las modalidades: imágenes faciales, habla y postura. Esto fue realizado mediante una búsqueda planificada dentro de los motores de búsqueda de Tumblr, Google y Giphy con etiquetas relacionadas a las emociones obteniendo 47.450 videoclips utilizando un descargador automático, además se definieron reglas para los vídeos editados manualmente, dentro de ellas se encuentra que en cada vídeo el actor tiene solo una expresión, las tomas de vídeo deben mantenerse en una sola persona para evitar cambios de ida y vuelta, un vídeo largo se puede dividir en varias partes y cada una de esas partes debe expresar solamente una etapa diferente de expresión emocional, agregando que se pueden incluir varios personajes en el mismo marco, pero en la mayoría de los casos todos intentan expresar la misma emoción. En el proceso de limpieza y preparación de los datos se utilizó FFmpeg de OpenCV para obtener la imagen *JPG* de cada vídeo, además se quitaron automáticamente aquellos vídeos que contenían objetos no humanos, esto fue filtrado por YOLO V3, definiendo que

un marco es utilizable si se pueden detectar personas, además al no poder encontrar un tamaño estándar para todas las imágenes, se colocaron en cuadrados y para los vídeos que tenían más de un personaje en el cuadro, se calculó el área del cuadrado donde se ubicaba cada personaje; el personaje que tenía mayor área correspondía a la emoción que expresaba utilizando la posición del bbox dado. Para el etiquetado manual de los datos se utilizaron las seis emociones básicas más otras tres emociones: aburrido, confundido y decepcionado. Finalmente, el conjunto de datos contiene 19.004 videoclips dividida en dos partes HEU-part1 y HEU-part2, donde HEU-part1 se encuentra dividido en 80 % para el entrenamiento, 10 % para la validación y 10 % para testing. HEU-part2 se dividió en en 65 % para el entrenamiento, 35 % para la validación. En este proyecto se utilizaron varios métodos convencionales de aprendizaje automático y aprendizaje profundo para evaluar HEU Emotion y se propuso un módulo de atención multimodal para fusionar características multimodales de forma adaptativa, obteniendo un accuracy de 55,04 %, con una arquitectura conformada por CNN, GRU, 3D- resnetXt, opensmile y MMA [44].

2.2.10. Reconocimiento automático de emociones en condiciones reales a partir de imágenes y audio

En este proyecto se realiza el MER tanto en imágenes como en señales de audio, los datasets utilizados son SFEW y AFEW que se encuentran en estado no controlado. Para esto se implementaron y evaluaron modelos unimodales basados en redes convolucionales, estudiando cada tipo de entrada de manera independiente. Además, se exploró la posibilidad de utilizar transfer learning para este tipo de sistemas. Finalmente, se compararon los resultados obtenidos en unimodal y multimodal para notar si existe una mejora con la incorporación del audio al modelo. En el proceso de preprocesamiento se obtuvo el audio mediante la librería ffmpeg, además se hizo un proceso de estandarización de la longitud de los audios a la longitud del audio más largo y no se realizó un preprocesamiento de imágenes. Se realizó la detección de las seis emociones básicas. En este proyecto se logró un accuracy de 45 %, con una arquitectura conformada por SVM, RBF, fine-tuning, VGG16 y late fusion [45].

2.2.11. Multimodal Embeddings From Language Models for Emotion Recognition *in-the-wild*

En este proyecto se trabaja el reconocimiento de emociones en las modalidades de audio y texto (además solamente para CMU-MOSEI se considera la modalidad visual) a través de BERT y ELMo, representaciones léxicas que modelan la semántica y sintaxis de las palabras con mayor eficacia. Se propuso un método para aprender y extraer incrustaciones multimodales de modelos de lenguaje previamente entrenados. El modelo utilizado por los investigadores empleó capas convolucionales para calcular incrustaciones acústicas

a partir de funciones de audio, las cuales fueron combinadas con incrustaciones de texto en un biLM (modelo de lenguaje bidireccional) utilizando una función de puerta sigmoidea. Para entrenar el modelo, se utilizó audio y texto de vídeos de YouTube en una tarea de modelado de lenguaje basado en audio. Posteriormente, se demostró la efectividad de las incrustaciones de oraciones extraídas de este biLM multimodal en la tarea de reconocimiento de emociones del hablante. Los resultados indicaron mejoras en los datasets de MSP-IMPROV y CMU-MOSEI en comparación con otros proyectos similares, logrando un 62,3 % de WA promedio en CMU-MOSEI en las modalidades visual, léxica y acústica. Este enfoque demuestra los beneficios de utilizar un preentrenamiento no supervisado en modelos multimodales para obtener representaciones efectivas para capturar dinámicas intermodales e intramodales en el lenguaje natural hablado. [20].

2.2.12. An audiovisual and contextual approach for categorical and continuous emotion recognition *in-the-wild*

En este proyecto se realiza MER con datos en ambientes no controlados, se utilizan los conjuntos de datos AffWild2 y ABAW2 para el entrenamiento y AffWild2 para las pruebas. Este proyecto utiliza las modalidades de cara, contexto, cuerpo y audio. Dentro de este proyecto se trabaja con las seis emociones básicas y se utiliza como método CNN-RNN en cascada para poder predecir la emoción final, obteniendo un accuracy de 64,20 %. [20].

2.2.13. Semi-supervised Multi-modal Emotion Recognition with Cross-Modal Distribution Matching

Este proyecto aborda MER en audio, texto e imagen utilizando el aprendizaje semi-supervisado, comparándolo con el aprendizaje supervisado y con diversos proyectos que se encuentran en su estado del arte. El modelo propuesto en este proyecto es Cross-Modal Distribution Matching para el reconocimiento de las seis emociones básicas agregando la emoción neutral. Para el entrenamiento del modelo se utilizaron los conjuntos de datos AMI, IEMOCAP y MELD, para los experimentos se utilizó IEMOCAP y MELD, se obtuvo un F1-Score en el conjunto de pruebas MELD igual a 57,1 %, con una arquitectura conformada por Cross-Modal Distribution Matching [46].

Nombre del artículo	**Datasets Entrenamiento**	**Datasets de Prueba**	**Multimodal Accuracy u otro**	**Método**
MAFW: A Large-scale, Multimodal, Compound Affective Database for Dynamic Facial Expression Recognition *in-the-wild* [36]	MAFW	MAFW	UAR: 31,00 y WAR: 51,29	Resnet18_LSTM y C3D_LSTM
Video-based emotion recognition *in-the-wild* using deep transfer learning and score fusion [37]	AFEW	FER	54,55 %	SIFT-FUN, LPQ-TOP, LGBP-TOP, VGG-Face, openSmile, CNN-FUN, Weighted Score Level Fusion.
Developing crossmodal expression recognition based on a deep neural model [38]	FABO, SAVEE, EmotiW y GTZAN	FABO, SAVEE y GTZAN	46,10 %	Adaboost-based detection, MFCC, CCCNN y SOM.
Adaptive Multimodal Emotion Detection Architecture for Social Robots [39]	IEMOCAP	IEMOCAP	77,60 %	VGG19, MFCC, DialogXL, Embracenet+.
End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition *in-the-wild* [40]	AFEW y AffectNET	AffWild2	48,10 %	1D CNN+LSTM, VGGFace2-EE, y VGGFace2-EE + SVM.
MuSe 2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress [41]	MuSe-Humor, Passau-SFCH y MuSe-Reaction	MuSe-Humor, Passau-SFCH y MuSe-Reaction	44,13 %	OpenSmile, eGeMAPS y LSTM-RNN.
Deep Auto-Encoders with Sequential Learning for Multimodal Dimensional Emotion Recognition [42]	RECOLA	RECOLA	RMSE: 51,00 %	LSTM.

Audiovisual emotion recognition in wild [43]	SAVEE, eNTERFACE'05 y RML	AFEW	27,10 %	SVM y CNN-AlexNet.
HEU Emotion: a Large-scale Database for Multi-modal Emotion Recognition *in-the-wild* [44]	HEU-part1 y HEU-part2	AFEW, HEU-part1 y HEU-part2	55,04 %	CNN, GRU, 3D-resnetXt, opensmile y MMA.
Reconocimiento automático de emociones en condiciones reales a partir de imágenes y audio [45]	AFEW	AFEW	45,00 %	SVM, RBF, fine-tuning, VGG16 y late fusion.
Multimodal Embeddings From Language Models for Emotion Recognition *in-the-wild* [20]	CMU-MOSEI	CMU-MOSEI	Promedio ponderado: 62,30 %	Graph-MFN
An audiovisual and contextual approach for categorical and continuous emotion recognition *in-the-wild* [47]	AffWild2, ABAW2	AffWild2	64,20 %	CNN-RNN cascade.
Semi-supervised Multi-modal Emotion Recognition with Cross-Modal Distribution Matching [46]	IEMOCAP, AMI and MELD	IEMOCAP and MELD	F1-score: 57.1 %	Cross-Modal Distribution Matching.

Tabla 2.1: Estado del arte.

Los trabajos anteriormente descritos no superan los 10 años de antigüedad hasta el 2023, en ellos se pudieron encontrar diferentes técnicas, modalidades y conjuntos de datos para lograr realizar MER *in-the-wild*. Se pudo notar que en "HEU Emotion: a Large-scale Database for Multi-modal Emotion Recognition *in-the-wild*" [44] se abarcó el problema de realizar un conjunto de datos con ciertas reglas para poder escoger cuales eran los datos que se podían utilizar para el desarrollo de un conjunto de datos *in-the-wild* con todo lo que implica el proceso de extracción de los datos y el etiquetado de éstos. Además se lograron identificar diferentes técnicas que son utilizadas para el entrenamiento y validación de los datos, junto con las diferentes datasets que se utilizan en cada trabajo. Por último, se pudo visualizar que en ninguno de los proyectos descritos en esta sección unifica ni utiliza en conjunto los datasets que se plantean en este Proyecto de Título, como lo son AFEW [7], AffWild2 [8] y MELD [9], esto le agrega importancia a la realización de este Trabajo de Título, dado que, como se plantea en el proyecto en el cual está basado este Trabajo de Título “An Assessment of *in-the-wild* Datasets for Multimodal Emotion Recognition” [1] se propone la unificación de datasets para así poder entrenar el modelo de Deep Learning y obtener un mejor MER en condiciones no controladas.

Capítulo 3

Definición del problema y propuesta de solución

En esta sección, se abordan los problemas identificados en el proyecto, así como la propuesta generada para solucionarlos. Además, se destaca la importancia de llevar a cabo este Trabajo de Título, respaldado por la presentación de los objetivos de esta Trabajo de Título, para responder la hipótesis del proyecto: *¿Es la unificación de los datasets suficiente para obtener mejores resultados en MER in-the-wild?*.

3.1. Definición del problema

Existen diferentes técnicas y métodos para poder analizar emociones utilizando Deep Learning de manera multimodal. Para poder entrenar estos modelos de aprendizaje profundo generalmente se utilizan conjuntos de datos que se encuentran en condiciones generalmente ideales, es decir, que no representan la realidad. Para luego utilizar estos modelos con situaciones controladas y obtener resultados deseables, pero estos datasets armados en ambiente controlados no representan las situaciones de la vida cotidiana [45]. La mayoría de las técnicas que se basan principalmente en el aprendizaje profundo, se entrenan utilizando conjuntos de datos elaborados en condiciones controladas, lo que dificulta su aplicabilidad en un contexto *in-the-wild* [1]. Esto se debe a que las características de los conjuntos de datos que se encuentran en condiciones no controladas son distintas a las de los datasets de entrenamiento del modelo, por ejemplo los sets de datos *in-the-wild* poseen distintas poses y movimientos naturales de la cabeza, iluminación cercana a la del mundo real, múltiples sujetos en el mismo marco y oclusiones [7], características que no poseen los conjuntos de datos trabajados en laboratorios. Los datasets que se utilizan en este proyecto son del tipo no controlado y fueron elegidos debido a su estructura y disponibilidad, se solicitaron directamente a los autores por medio de correo electrónico y sus nombres son: Acted Facial

Expression *in-the-wild* (AFEW) [7], Emotion Lines Dataset (MELD) [9] y Affect-in-the-Wild (AffWild2) [8]. El problema que se encuentra en estos conjuntos de datos es que no han sido diseñados para tareas multimodales, por lo tanto surge la necesidad de poder abordar este problema mediante la unificación de los tres conjuntos de datos mencionados, con el fin de mejorar el reconocimiento multimodal de emociones *in-the-wild*. En los siguientes párrafos se detallan los problemas particulares que afectan en el procesamiento multimodal para cada uno de los datasets en condiciones no controladas:

- AFEW [7]: corresponde a imágenes en blanco y negro con actores que gesticulan marcadamente sus emociones, esto podría comprometer su adaptación a la realidad. Además, este set de datos no posee transcripciones de sus audios. Los datos de la transcripción dependen tanto de la calidad de sonido del conjunto de datos como del discurso, agregando que, al hacer la transcripción automática del audio al texto, el texto detectado puede no corresponder a lo que realmente fue pronunciado por el sujeto [1].

- MELD [9]: es un conjunto de datos que está basado en un programa de televisión llamado ”Friends”, fue diseñado para las modalidades de audio y texto. Uno de los problemas detectados es que el número de ejemplos disponibles por clase de emoción está desequilibrado, especialmente en clases menos disponibles como Disgusto o Miedo, lo cual es esperable, ya que la fuente original es un programa de comedia. Además, las imágenes de este dataset son grupales, esto aumenta la complejidad de identificar cuál es el actor que está hablando en la escena para identificar su emoción. De igual modo, este conjunto de datos al ser concebido para ER en conversaciones, requiere más preprocesamiento en la modalidad facial. Sumado a ello, algunos audios tienen una baja calidad de sonido y esto dificulta la comprensión de lo que está diciendo el sujeto, por lo tanto afecta directamente a la detección de la emoción [1].

- MAFW [36]: tiene 10.045 vídeos extraídos de películas, series de TV, entrevistas, entre otros, posee cada vídeo etiquetado con su emoción, posee 11 emociones individuales, 32 clases de emociones múltiples, posee los puntos faciales, anotaciones automáticas y un texto descriptivo de lo que se ve en cada vídeo en idioma Chino e Inglés [36]. Este conjunto de datos se encuentra desbalanceado en su cantidad de muestras de emociones. Este conjunto de datos posee vídeos provenientes de China, Japón, Corea, Europa, América e India, por lo tanto los vídeos poseen diferentes idiomas. Además, no posee ningún diferenciador entre los vídeos en distintos idiomas o procedencias, eso implica que tampoco posee la transcripción del texto, agregando que no se encuentra el audio extraído independientemente y los recortes faciales no fueron proporcionados por los autores. En este proyecto de título este conjunto de datos se utiliza solamente para pruebas debido a que sus videos poseen diferentes lenguas las cuales no se encuentran especificadas por archivo.

- AffWild2 [8]: gran porcentaje de los vídeos que componen este conjunto son reacciones a contenido multimedia o de interacciones con otras personas fuera de cámara, es decir, la emoción etiquetada podría ser diferente a lo que está sucediendo en el audio. Además, el enfoque de este conjunto de datos es en rostros y tiene inconsistencias tanto en las modalidades de audio como de texto. Sumado a ello, existe un desbalanceo en la cantidad de emociones, se encuentra una gran cantidad de clases neutral y felicidad. Agregando que este conjunto de datos no tiene transcripción del audio, además posee audios con una duración de varios minutos con más de una emoción. Para obtener el texto del audio es necesario utilizar técnicas de transcripciones automáticas, estas dependen de la calidad de los audios.

En los datasets mencionados existen imágenes de baja calidad presentes en aquellos datasets que poseen desde el origen recortes faciales, como también aquellas en las que se debe realizar la detección automática de rostros. Además, existen audios con baja calidad, y otros conjuntos de datos sin transcripciones desde el origen, esto afecta directamente a la calidad de la futura transcripción automática del texto.

Otra problema es el etiquetado de las emociones que se tienen en cada una de los datasets, debido a que se debe encontrar un estándar y una manera de poder unificarlas en una única base de datos en términos de clases de emociones, en la Figura 3.1 se muestra un ejemplo de cómo se encuentran las etiquetas para el conjunto de datos AffWild2. En la Tabla 3.1 se muestran las diferentes maneras en las que cada uno de los conjuntos de datos presentan las emociones.

Dataset	**Clases de emociones en su formato de escritura original**	**Formato en el que se encuentra la emoción**	**Conexión con los archivos**
AFEW	Angry, Disgust, Fear, Happy, Neutral, Sad y Surprise.	Nombre de carpeta es la emoción de cada uno de los videos que se encuentran en ella.	Nombre del archivo (sin la extensión).
MELD	anger, disgust, fear, joy, neutral, sadness y surprise.	Archivo *CSV* con la columna Emotion con cada una de las emociones por video.	Concatenación de las columnas "dialogue" y "utterance" conforman el nombre del archivo (sin la extensión).
AffWild2	0 = Neutral, 1=Anger, 2=Disgust, 3=Fear, 4=Happiness, 5=Sadness, 6=Surprise y -1=Other.	Carpeta EXP_CLASSIFICATION _CHALLENGE con un archivo *TXT* por video, con una emoción por recorte facial.	Nombre del archivo (sin la extensión).
MAFW	invalid, anger, disgust, fear, happiness, neutral, sadness, surprise, contempt, anxiety, helplessness y disappointment.	Archivo *CSV* con un valor entre -1 y 1 por cada una de las emociones por vídeo, el valor más alto corresponde a la emoción.	Nombre del archivo (con extensión).

Tabla 3.1: Formatos y conexiones de los archivos de emociones en los diferentes datasets.

Figura 3.1: Muestra de etiquetado de emociones en AffWild2 [1].

3.2. Propuesta de solución

A fin de solventar los problemas descritos anteriormente, se unificarán los tres datasets anteriormente mencionados para formar un conjunto de datos único que deba ser depurado previamente para lograr reajustar una arquitectura y obtener mejores resultados en el MER. Posteriormente, realizar visualizaciones y análisis de los resultados obtenidos, demostrando las fortalezas y debilidades en su comportamiento en el procesamiento multimodal de emociones [1].

En los siguientes párrafos se mencionan las soluciones propuestas para cada uno de los problemas, dichas soluciones se muestran gráficamente en la Figura 3.2:

- Para solventar los problemas de calidad para cada una de las modalidades, es necesaria la utilización de técnicas de control de calidad las cuales serán ajustadas mediante la visualización de gráficas y su relación con los valores límites de dichos métodos para detectar la calidad de los datos. Todos aquellos datos que se determinen de baja calidad se marcarán como no utilizables con un cero (0), y aquellos datos que no sean de baja calidad se marcarán con un uno (1), de esta manera se busca lograr diferenciar los datos de buena calidad y los de baja calidad dentro de un archivo *CSV*, uno para el entrenamiento y otro para la validación de los tres conjuntos de datos unificados. La identificación de las calidades por modalidad se hará mediante el análisis de diferentes técnicas descritas en la Sección 4.4.

- El problema de los audios extensos en AffWild2 se abordará recortando los audios. Se detalla que cada 1 segundo se extrajeron 5 imágenes, por lo tanto se realizará un recorte de los audios según la duración de la emoción.

- Para la transcripción automática del texto se hará un análisis de cuatro técnicas de transcripción, donde se realizan pruebas con respecto a audios que posean música, onomatopeyas y ruido de fondo para identificar cual de las técnicas obtiene mejores resultados en la transcripción de audios *in-the-wild*, los filtros de calidad de texto se aplicarán solamente para aquellos datasets en las que se realizó la transcripción automática.

- Para el caso de discordancia de la emoción que se encuentra en el audio con la emoción que se muestra en la imagen, se utilizará la misma emoción para todas las modalidades, debido a que la solución final a este problema costosa, requiere realizar el etiquetado manual del audio con la emoción en contexto [1] utilizando crowdsourcing, debido a los limitados tiempos de realización de este proyecto no es posible realizar dicho etiquetado manual.

- Para poder solucionar la limitación del desequilibrio de las emociones, se usarán técnicas de aumentación de datos para incrementar las muestras de aquellas emociones que poseen una menor cantidad de ejemplares, de esta manera obtener una mejor detección de la emoción y evitar el sobre-entrenamiento de las clases que poseen una mayor cantidad de muestras en el dataset unificado [1].

- Para resolver el problema de las diferentes construcciones de etiquetas de emociones, se usará un estándar de emoción, el cual será: "Angry" (Enojo), "Happy" (Feliz), "Sad" (Tristeza), "Surprise" (Sopresa), "Fear" (Miedo), "Disgust" (Disgusto) y "Neutral" (Neutral), con el fin de unificar los tres conjuntos de datos conformando un solo dataset con datos limpios y preparados para ingresar a los modelos de DL.

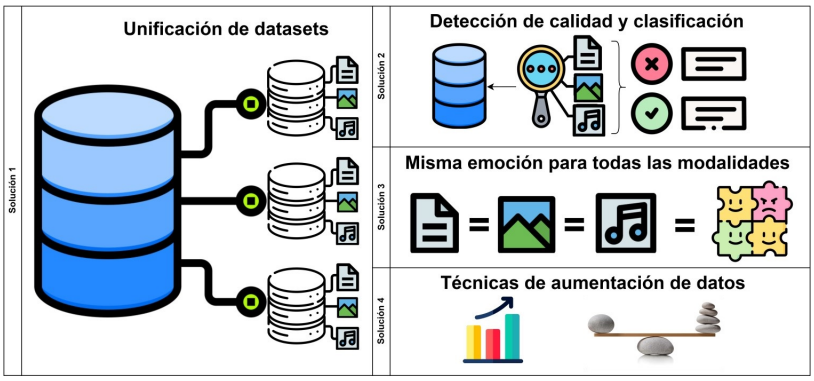


Figura 3.2: Diagrama solución propuesta.

3.3. Objetivos

3.3.1. Objetivo general

Integrar conjuntos de datos *in-the-wild* para el reconocimiento de emociones mediante una arquitectura multimodal basada en técnicas de aprendizaje profundo.

3.3.2. Objetivos específicos

1. Desarrollar un conjunto de datos unificado de alta calidad, integrando datos relevantes para el reconocimiento multimodal de emociones en ambientes no controlados.

2. Implementar y entrenar una arquitectura basada en técnicas de aprendizaje profundo con el fin de llevar a cabo el reconocimiento multimodal de emociones en ambientes de contexto real.

3. Realizar visualizaciones y análisis de los resultados obtenidos por el modelo entrenado, evaluando su rendimiento a través de métricas específicas.

3.4. Importancia del trabajo

Las emociones son una parte esencial de la vida y la comunicación entre seres humanos. Una parte muy importante del mensaje que comunicamos depende de los factores emocionales como la expresión, los gestos, el tono o la forma de articular el mensaje. Todo esto hace que la identificación de las emociones sea fundamental para la interacción y la comprensión de la comunicación interpersonal [45]. Es por esto que se requiere llevar a cabo este proyecto de MER, para lograr implementar este tipo de reconocimiento en un

aspecto más cercano a la realidad de las interacciones sociales. Este proyecto es relevante en diferentes ámbitos científicos y socio-económicos, en la evaluación de experiencia de usuario, evaluación del estado emocional de las personas en la rama de psicología, en el aumento de las ventas en el caso del marketing [48], en mejorar la interacción humano-computador en la robótica [1], en el análisis de sentimientos para predecir los movimientos de la bolsa en redes sociales [49], en la detección de la emoción en chatbots para la generación de respuestas emocionalmente relevantes para el usuario [50], para el uso de ER en la mejora de la experiencia de los usuarios en los videojuegos [51], entre otros. Para realizar esto de la manera correcta es importante el uso de conjuntos de datos adecuados y con buena calidad en todas sus modalidades, junto con del desarrollo de algoritmos robustos y eficientes para obtener resultados precisos en el ER. Este proyecto generará un conjunto de datos unificado con la característica de poseer una buena calidad en las modalidades de texto, audio e imágenes faciales, el cual no se encuentra actualmente en la literatura, por lo tanto será un nuevo y extenso recurso para futuros proyectos.

Capítulo 4

Diseño

En el presente capítulo se aborda el diseño de la solución en la cual explica la modelación del problema junto con la descripción de las técnicas que se utilizan para el análisis de los datos, además de exponer la descripción de los procesos necesarios para explicar cómo se recolectaron los datos, donde se almacenarán, cuál será el análisis y cómo se realizará la visualización de ellos.

4.1. Descripción de procesos

En la siguiente lista se encuentra la descripción de los procesos de los tres datasets que se utilizan en la tarea de unificación de datos.

- La recolección de los conjuntos de datos AFEW, AffWild2 y MELD lo facilitó Ana Aguilera (Profesora Guía) y Diego Mellado (Profesor Co-Referente), las cuales fueron solicitadas directamente a los autores y se otorgó el acceso con fines académicos [1]. MAFW fue el último dataset conseguido y también fue solicitada a los autores mediante correo electrónico.

- Para el almacenamiento de los datos, se crearon dos repositorios en GitHub: uno que contiene los códigos correspondientes al preprocesamiento, entrenamiento, validación y pruebas, llamado "MultimodalEmotionRecognitionInTheWild-Thesis"; y otro denominado "MERDWild", en el cual se encuentra una descripción de la base de datos y las instrucciones para su obtención. El conjunto de datos MERDWild se encuentra almacenado en OneDrive.

- Con respecto al análisis de los datos, primeramente se realiza un preprocesamiento de los datos para poder mejorar la calidad de estos, limpiarlos, normalizarlos y transformarlos, luego se utilizarán métodos supervisados de aprendizaje profundo para identificar patrones y lograr hacer el ER individuales para cada modalidad (audio,

rostro y texto), y luego calcular una emoción final multimodal mediante el método de fusión Embracenet+ [1].

- En cuanto a la visualización de datos, se lleva a cabo mediante las métricas de evaluación junto con la visualización de la matriz de confusión, curvas ROC y las curvas correspondientes al accuracy frente a las épocas y la pérdida (loss) frente a las épocas, con el fin de analizar los resultados obtenidos de las predicciones del modelo, para luego tomar decisiones y extraer conclusiones.

4.2. Modelación de la solución del problema

El modelo recibe como entrada los tres datasets, los datos de entrada se someten a un procesamiento individual en el que se extraen los archivos correspondientes a las diferentes modalidades de datos (rostro, audio y texto). Posteriormente, se evalúa la calidad de los datos, se limpian, transforman, se agrupan y se establece un estándar conformando un único conjunto que comprende datos clasificados como buena calidad. Luego, se aplican métodos de aprendizaje profundo para cada modalidad y, finalmente, se fusionan los resultados mediante el método Embracenet+ para obtener una única emoción final. La Figura 4.1 muestra de manera esquemática la solución que se aplicará al problema.

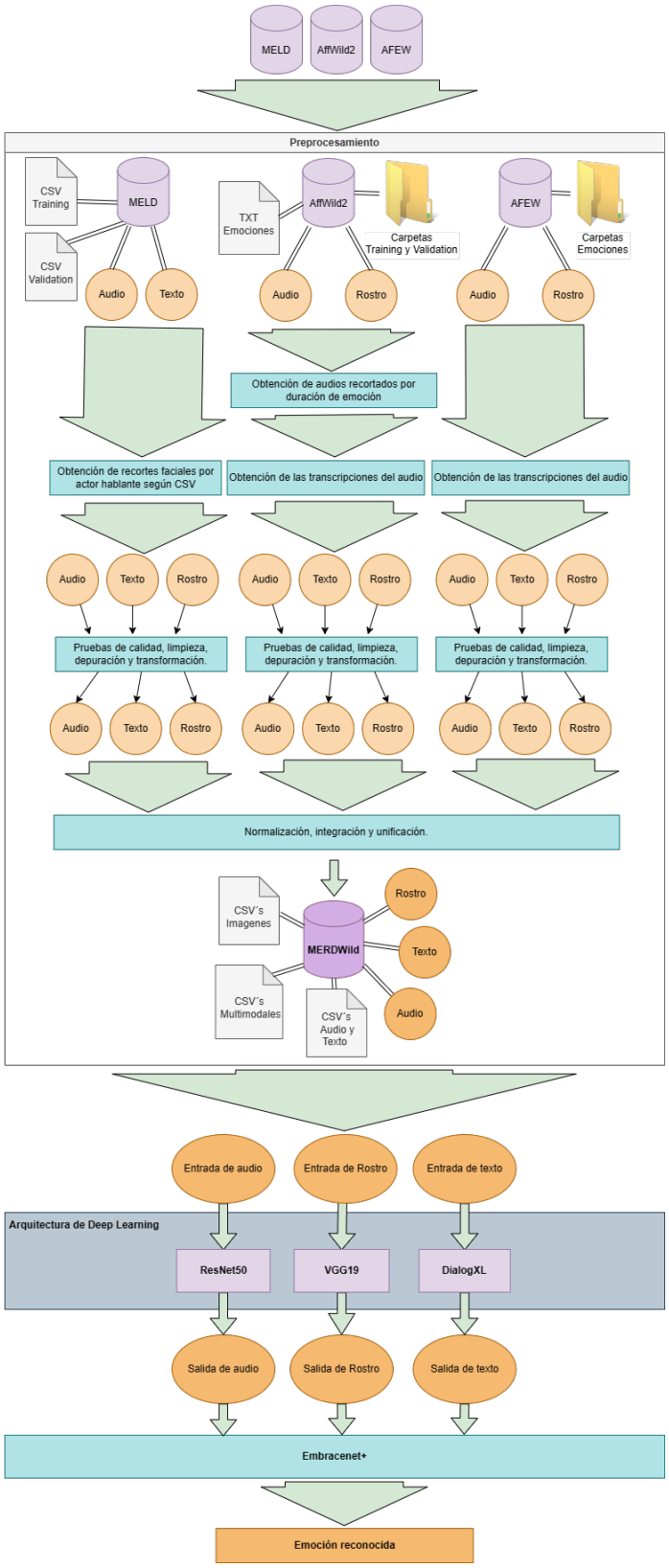


Figura 4.1: Modelo de la solución.

4.3. Datasets

En el presente Trabajo de Título se utilizan cuatro datasets como base para la investigación, los cuales son descritos en los siguientes párrafos, en donde exhibirá para cada uno de ellos una descripción, junto con la Tabla 4.1 que muestra los datos de origen, la recolección de registros, el formato en el que se encuentra el dataset y donde se encuentra almacenado para cada uno de los conjuntos de datos utilizados.

1. AFEW: Este conjunto de datos está basada en películas y programas de TV con un total de 1.809 videoclips, está enfocada en rostros y posee las modalidades facial, audio y postura. El rango de edad de las personas que se encuentran en los vídeos se encuentra entre 1 y 70 años, con aproximadamente 330 personas identificadas. En este conjunto de datos se encuentran los datos correspondientes al nombre, edad del actor, edad del personaje, genero, postura y expresiones faciales individuales, además se puede encontrar a más de una persona por escena y cada persona identificada tiene su expresión facial definida. Las expresiones faciales que se encuentran en este dataset son las seis básicas más la "neutral". Los vídeos están en formato *AVI*, y varían sus calidades [7].

2. AffWild2: Este dataset se basó en vídeos reaccionando a contenido de Youtube, posee 258 personas, 30 horas de vídeo en formato *MP4* y posee las modalidad de rostro y audio. En este conjunto de datos se manejan las seis emociones básicas más neutral y "otra". Posee más de una persona por escena y se encuentra cada rostro etiquetado con la emoción correspondiente [8]. En la Figura 4.2 se puede ver un extracto de este conjunto de datos, específicamente en la modalidad de rostros.

Figura 4.2: Muestras de rostros recortados de AffWild2 [1].

3. MELD: Este set de datos contiene alrededor de 13.000 declaraciones de 1.433 diálogos de la serie de TV "Friends". Posee las modalidades de audio, texto e imagen. Este conjunto de datos está enfocada en diálogos (audio y texto), por lo que no posee rostros recortados. Se manejan las seis emociones básicas agregando la emoción

neutral, además posee rango de sentimiento. El dataset posee en varias escenas más de una persona. Se encuentra organizada en tres conjuntos de carpetas, una para el entrenamiento, otra para la validación del modelo y una tercera para el proceso de pruebas. Este conjunto de datos posee un archivo *CSV* con información sobre el nombre del actor, la emoción que está expresando, el sentimiento, la transcripción el inicio y el final en la duración del vídeo, entre otros datos [9], tal como se muestra en la Figura 4.3.

Sr No	Utterance	Speake	Emotion	Sentiment	Dialogue_ID	Utterance_ID	Season	Episode	StartTi	EndTi
1	also I was the point person on my company□s transition from the KL-5 to GR-6 system.	Chandler	neutral	neutral	0	0	8	21	0:16:16	0:16:22
2	You must□ve had your hands full.	The Intervie	neutral	neutral	0	1	8	21	0:16:22	0:16:23
3	That I did. That I did.	Chandler	neutral	neutral	0	2	8	21	0:16:23	0:16:26
4	So let□s talk a little bit about your duties.	The Intervie	neutral	neutral	0	3	8	21	0:16:27	0:16:30
5	My duties? All right.	Chandler	surprise	positive	0	4	8	21	0:16:34	0:16:41
6	Now you□ll be heading a whole division, so you□ll have a lot of duties.	The Intervie	neutral	neutral	0	5	8	21	0:16:41	0:16:44
7	I see.	Chandler	neutral	neutral	0	6	8	21	0:16:49	0:16:52
8	But there□ll be perhaps 30 people under you so you can dump a certain amount on them.	The Intervie	neutral	neutral	0	7	8	21	0:16:49	0:16:55
9	Good to know.	Chandler	neutral	neutral	0	8	8	21	0:16:59	0:17:00
10	We can go into detail	The Intervie	neutral	neutral	0	9	8	21	0:17:00	0:17:03
11	No don□t I beg of you!	Chandler	fear	negative	0	10	8	21	0:17:03	0:17:05
12	All right then, we□ll have a definite answer for you on Monday, but I think I can say with some confidence, you□ll fit in well here.	The Intervie	neutral	neutral	0	11	8	21	0:17:05	0:17:13
13	Really?!	Chandler	surprise	positive	0	12	8	21	0:17:13	0:17:17
14	Absolutely. You can relax	The Intervie	neutral	neutral	0	13	8	21	0:17:18	0:17:21
15	But then who? The waitress I went out with last month?	Joey	surprise	negative	1	0	9	23	0:36:40	0:36:43
16	You know? Forget it!	Rachel	sadness	negative	1	1	9	23	0:36:44	0:36:47
17	No-no-no-no, no! Who, who were you talking about?	Joey	surprise	negative	1	2	9	23	0:36:44	0:36:49
18	No, I-I-I-I don't, I actually don't know	Rachel	fear	negative	1	3	9	23	0:36:49	0:36:52
19	Ok!	Joey	neutral	neutral	1	4	9	23	0:36:52	0:36:54
20	All right, well...	Joey	neutral	neutral	1	5	9	23	0:36:54	0:36:55

Figura 4.3: Datos en modalidad de texto MELD.

4. MAFW: El origen de este dataset es de extractos de videoclips de películas, series de TV, entrevistas, entre otros, con hablantes de distintas lenguas. Posee 10.045 vídeos, provenientes de diferentes lugares del mundo, por ejemplo India, Europa, China, América, entre otros. Los temas que se tratan en los vídeos son ciencia ficción, familiares, suspenso, amor, comedia, entrevistas, etcétera. Este conjunto de datos posee 11 emociones individuales anotadas por 11 anotadores capacitados, dentro de ellas se encuentran las seis emociones básicas más Neutral, Contempt, Anxiety, Helplessness y Disappointment, además se tienen 32 emociones múltiples para algunos vídeos específicos. Posee 4 archivos de Excel, los cuales corresponden a las anotaciones de las emociones individuales, múltiples y generales, y otro para la descripción para cada uno de los vídeos en los idiomas Inglés y en Chino [36]. En la Figura 4.4 se encuentra un ejemplo de los textos descriptivos de MAFW junto con a la imagen del video al que corresponde la descripción y su emoción [36].

Figura 4.4: Ejemplo imagen, emoción y textos descriptivos de MAFW.

Dataset	**Origen**	**Recolección**	**Formato**	**Almacenamiento**
AFEW [7]	Películas y programas de TV.	Solicitado por correo electrónico a los autores.	Imágenes faciales, audio e imágenes de postura.	Computador perteneciente a la Universidad de Valparaíso.
AffWild2 [8]	Contenido de Youtube.		Imágenes faciales y audio.	
MELD [9]	Serie de TV "Friends".		Visual, audio y texto.	
MAFW [36]	Videoclips de películas.		Visual, audio y texto descriptivo.	

Tabla 4.1: Descripción de datasets.

4.4. Preprocesamiento

Para el preprocesamiento de los datos de cada uno de los conjuntos de datos se realizó un proceso independiente en cada una de sus modalidades, para lograrlo se utilizaron diferentes medidas para la detección de calidad.

4.4.1. Preprocesamiento por modalidad

Las aplicaciones de las técnicas y medidas anteriormente descritas para las tres modalidades se encuentran descritas en los siguientes puntos:

Dataset	Brillo	Contraste	Resolución	Entropía	SIFT	HOG con Similitud del coseno
AFEW	$15 \leq X \leq 140$	$5 < X$	<40x40	$4 \leq X$	$3 < X$	$0,8 \leq X$
AffWild2	$15 \leq X \leq 235$	$5 < X \leq 850$	<40x40	$2,5 \leq X$	$3 < X$	$0,85 \leq X$
MELD	$28 \leq X \leq 180$	$5,7 < X \leq 475$	<40x40	$4 \leq X \leq 6,6$	$3 < X$	$0,8 \leq X$
MAFW	$15 \leq X \leq 235$	$5 < X \leq 850$	<40x40	$6 \geq X$	$3 < X$	$0,8 \leq X$

Tabla 4.2: Tabla de valores límites de variables de modalidad facial.

Modalidad de imágenes faciales

en el caso de MELD se realizaron los recortes automáticos de los rostros utilizando OpenCV por medio de ”Haar Cascade Classfiers” que detecta las caras de las personas en cada video [52]. Luego se realizó un script para identificar a los actores, una vez identificados, se seleccionaron solamente las caras de los actores que se encontraban hablando según la columna “Speaker” de los archivos *CSV*. Para AFEW y AffWild2 se proporcionan los recortes de las caras desde el origen. Una vez obtenidos todos los rostros se realiza el análisis de la calidad de las imágenes de manera independiente para cada dataset. De esta manera se filtró y se realizó un proceso de selección de datos según parámetros obtenidos de técnicas, tal como se muestra en la Tabla 4.2. Esta tabla tiene los valores límites de aceptación ajustados mediante el filtrado visual humano y por medio de gráficas de visualización del comportamiento de los datos por técnica. Entre las cuales se encuentran la detección de brillo [53], contraste [53], resolución [54], entropía [55], además del análisis de los puntos faciales detectados por imagen llamado "Scale Invariant Feature Transform" (SIFT) [56], esto junto con la utilización de la similitud del coseno [40] entre las imágenes y el Histograma de gradientes orientados (HOG) [57] para refinar la detección de rostros de cada dataset. Para finalmente ajustar las imágenes a un tamaño estándar de ancho y largo correspondiente a 64 píxeles.

Modalidad de audio

Primeramente fue necesario extraer los audios del conjunto de datos MELD, dicha extracción fue realizada mediante el uso de la librería moviepy editor [58]. En AFEW y AffWild2 ya se encontraban extraídos desde el origen, pero para AffWild2 fue necesario realizar un recorte de los audios debido a que su duración es más extensa respecto a los demás datasets. Se detalla que cada 1 segundo se extrajeron 5 imágenes, realizando un recorte de los audios según la duración de la emoción. Luego, para cada uno de los conjuntos de datos se realizó de manera independiente la detección de la calidad del audio, como se muestra en la Tabla 4.3. En esta tabla se detallan los valores límites ajustados por medio del filtrado humano auditivo, dicho filtrado humano se complementó con la visualización de gráficas tal como muestra la Figura 6.3. Adicionalmente se encuentran diferentes técnicas como el promedio del nivel de potencia [59], nivel peak [59], la distorsión armónica total [60] y la relación señal y ruido [61]. Finalmente, los audios de los datsets se guardaron

Dataset	Promedio nivel de potencia	Nivel Peak	Distorsión armónica total	Relación señal ruido
AFEW	-	$X \geq -38$	$X < 43,5$	$X \leq -3,1$
AffWild2	$-45 < X < -12$	$X \geq -27$	$19,7 < X < 35,2$	$-27,3 \leq X \leq -4$
MELD	$-50 < X < -15$	-	$X > 24$	-
MAFW	$-55 < X < -5$	$X \geq -35$	$29 < X < 49$	$X \leq -3$

Tabla 4.3: Tabla de valores límites de variables de modalidad de audio.

en una carpeta en formato *WAV*. Es importante destacar que se realizó la investigación y aplicación de filtrado por voz, la cual tenía como objetivo separar al hablante de la música utilizando como parámetro el rango de frecuencia de la voz humana, esto no fue posible debido a que la música también oscila dentro del rango de ese rango, por lo tanto se descartó dicha técnica.

Modalidad de texto

Solamente MELD presenta las transcripciones de los audios desde el origen. Para AFEW y AffWild2 se realizó un experimento utilizando cuatro algoritmos de transcripción de texto para evaluar cual cumplía con el requerimiento de ser más completo y obtener mejores resultados. Dentro de los algoritmos que se utilizaron se encuentran: SpeechBrain [62], WAV2Vec2 realizado por Jonatas Grosman [63], Speech Recognition [64] y WAV2Vec2 realizado por Vitou Phy [65]. Los resultados del experimento demostraron que el algoritmo que mejor se adaptaba para realizar la transcripción de audios *in-the-wild* es SpeechBrain. Para dicha transcripción se estableció como límite los primeros 7 segundos de audio, se rellenó con ceros si el audio tenía menos duración, tal como se realizó en el proyecto en el que se basa este Trabajo de Título [1]. Finalmente, una vez transcrito el texto se realizó un nuevo paso en el preprocesamiento, utilizando diferentes técnicas, dentro de ellas está transformar el texto a letras minúsculas [66] [67] [68], tokenización [66] [67], eliminación de espacios en blanco [67], eliminación de stopwords [67] [69], eliminación de caracteres especiales [67] y se utiliza la lematización [66] [67] [69] utilizando un listado de palabras en inglés como referencia de palabras existentes. Como filtro de calidad de texto se utilizó primeramente el filtro de audio, estableciendo que si el audio es de baja calidad no se transcribirá el texto, además se utilizó la base de la detección semántica usada en otro proyecto [70], ajustando como valor límite de exclusión los datos menores o iguales a cero, además se marcaron como datos de texto no usables aquellos donde su audio posee baja calidad, solamente para los conjuntos de datos AFEW y AffWild2, para MELD no se aplicó ningún filtro de calidad de texto debido a que posee los datos transcritos manualmente. El último filtro aplicado a la modalidad de texto consiste en que si el registro de la columna "Lemmatized_Tokens" se encuentra vacía luego del preprocesamiento aplicado, el texto se marca como no utilizable. Es importante destacar que sumado al preprocesamiento de texto anteriormente descrito se evaluó la calidad del texto en términos de coherencia y cohesión,

se realizó una investigación y se aplicaron dichos filtros desde los cuales no se logró obtener un valor límite determinante para filtrar la calidad de las transcripciones en ninguna de las dos técnicas analizadas, es por esto que se descartó su aplicación en el proyecto.

Aumento de datos

Para resolver este problema se usan técnicas de aumentación de datos para cada una de las modalidades. Para la modalidad de imágenes faciales se aplicaron operaciones de aumento aleatorias a los datos de entrenamiento, dentro de ellas se encuentran las transformaciones, las cuales tenían un 50 % de posibilidades de aplicarse, en cada iteración se modifica solamente una imagen. Dentro de las transformaciones se incluye la operación de voltear con una rotación que se encuentra en el rango de 10 grados, además de aplicar un contraste y brillo aleatorios, o también la adición de Poissonnoise, es importante destacar que cada una de las técnicas se aplican de forma independiente. Adicionalmente se requiere un tamaño de entrada definido para el correcto funcionamiento del modelo, fue necesario redimensionar las imágenes originales a 64 × 64 píxeles, utilizando la biblioteca scikit-image. Para el entrenamiento de la modalidad de audio se realizó el aumento de datos para cada muestra de audio, en el cual se usó un desplazamiento aleatorio desde el centro del audio, sumado a ello se usó un ruido gaussiano con $\sigma = 1 \times 10^{-5}$. En la modalidad de texto no se aplicó ninguna técnica de aumentación de datos [1].

En la Tabla 4.4 se muestran las frecuencias de las muestras existentes por emoción en cada una de las modalidades, en ella se puede observar que la base de datos MERDWild se encuentra con un desequilibrio en la cantidad de archivos por emoción, debido a que las técnicas de aumentación de datos solamente se realizaron para el entrenamiento de los modelos de manera independiente.

Emoción	**Filtro de Imagen**		**Filtro de Audio**		**Filtro de Texto**	
	No aptas	**Aptas**	**No aptas**	**Aptas**	**No aptas**	**Aptas**
Angry	2.997	82.937	18	1.523	69	1.472
Disgust	1.012	25.402	9	433	16	426
Fear	1.167	32.802	15	471	27	459
Happy	5.534	167.584	66	2.717	120	2.663
Neutral	9.550	402.808	198	7.934	445	7.687
Sad	5.656	115.510	29	1.110	58	1.081
Surprise	2.445	78.238	33	1.685	185	1.533

Tabla 4.4: Frecuencia de emociones por modalidad y filtro.

Estructura unificada

Se estableció que se utilizarán los datos que corresponden a entrenamiento y validación tal como se utilizan en los conjuntos de datos originales. Para AFEW y AffWild2 se encuentran establecidas las separaciones entre conjuntos de entrenamiento y de validación. Para el caso de MELD, donde se encuentra separado por conjunto de entrenamiento ("train") y desarrollo ("dev"), se utilizó el conjunto de desarrollo (dev) que posee una menor cantidad de muestras como conjunto de validación. El conjunto de datos MAFW será utilizado solamente para pruebas. Aquellas separaciones de conjuntos de datos se encuentran establecidas tanto en los archivos *CSV* unificados, los cuales poseen las siete emociones, como también en la estructura de la unificación por carpetas que se detalla en el próximo párrafo.

Para el proceso de unificación se estableció una jerarquía de carpetas, las cuales contemplan tanto las etiquetas de emociones, como los resultados de los filtros de calidad para las modalidades de rostro y audio, dentro del filtrado de calidad de los audios se encuentran también las transcripciones de los audios. Además, cada carpeta de recortes de imágenes faciales y cada archivo de audio posee el nombre del vídeo original al que pertenece. Esta estructura se encuentra plasmada en la Figura 4.5.

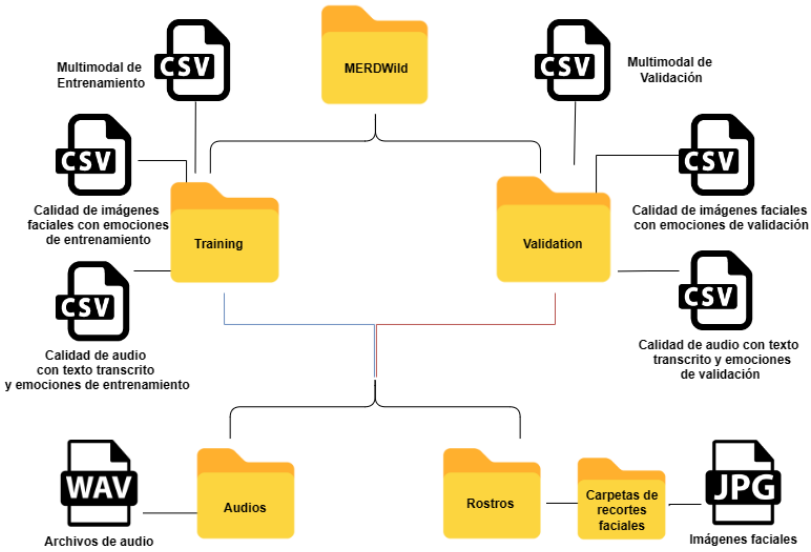


Figura 4.5: Estructura de MERDWild.

Para estandarizar las categorías emocionales en los diversos conjuntos de datos, se implementó un procedimiento que definió un conjunto consistente de siete emociones para este proyecto. Estas etiquetas se seleccionaron debido a su presencia común en los conjuntos de datos originales, en cada dataset se tenían diferentes formas de etiquetar una misma emoción. De esta manera se generalizaron las clases otorgándoles los siguientes nombres:

Emoción / Dataset	**Frecuencia porcentual**						
	Angry	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
AFEW	1.766	1.000	0.989	1.903	1.795	1.480	0.983
AffWild2	0.560	0.308	0.468	4.339	16.916	1.097	1.549
MELD	7.123	1.617	1.732	10.479	28.704	4.630	7.397

Tabla 4.5: Comparativa de emociones por conjunto de datos (%) con alineación por uterancia para datos de buena calidad.

Dataset	**Filtro de Imagen**		**Filtro de Audio**		**Filtro de Texto**	
	No aptas	**Aptas**	**No aptas**	**Aptas**	**No aptas**	**Aptas**
AFEW	3.446	59.754	29	1.031	43	1.017
AffWild2	7.255	448.860	280	3.901	296	3.885
MELD	17.660	396.667	59	10.941	581	10.419

Tabla 4.6: Frecuencia de archivos en datasets por modalidad y filtro.

"Neutral", "Angry", "Disgust", "Fear", "Happy", "Sad" y "Surprise". Por consiguiente, se procedió a organizar las etiquetas emocionales existentes en cada conjunto de datos a estas categorías emocionales estandarizadas. En la Tabla 4.5 se muestran los porcentajes por uterancia de las emociones disponibles por dataset que poseen buena calidad en la base de datos MERDWild.

La estructura de las etiquetas emocionales se detalla en la Tabla 3.1, para realizar esta estandarización se generó un archivo *CSV* de calidades, se destaca la columna "Images_KEYs", "Audio_KEY" y "Image_KEY", las cuales facilitan el acceso a los datos de audio e imagen, además se resalta la columna "Multimodal_Connection", que cumple la función de vincular los datos de texto, audio e imágenes. La columna "Emotion" contiene la emoción de la modalidad y "Source_DB" se utiliza para identificar el conjunto de datos de origen de cada archivo. De la misma manera, se encuentra un *CSV* multimodal por set de datos, el cual está ajustado para las tres modalidades, posee una columna indicando las modalidades disponibles por uterancia denominada "Modalities", la cual indica si se encuentra presente la modalidad con la letra inicial (en orden: I, A, T), si una modalidad no se encuentra disponible hay una X en su posición. Este archivo multimodal no posee los datos por técnica de filtrado, y solamente contiene datos de buena calidad. En las Figuras 4.6 y 4.7, se pueden observar algunas las columnas anteriormente mencionadas junto con las correspondientes a los filtros de imagen, audio y texto, las cuales se encuentran con el nombre "Image_Filter", "Audio_Filter" y "Text_Filter" respectivamente.

Audio_KEY	Emotion	Multimodal_Connection	Duration	Audio_Filter	Text_Filter	Transcription	Lowercase_Transcription	Text_Without_Stopwords
Training\Audios\1-30-1280x720_1.wav	Happy	1-30-1280x720_1	5.0	1	1	YOU'VE BEEN WAITING HALF LIKE THAT AND LET'S B...	you have been waiting half like that and let u...	waiting half like let us begin
Training\Audios\1-30-1280x720_11.wav	Neutral	1-30-1280x720_11	5.0	1	1	IMAGINE YOU'RE FLOATING LIKE A BIRD FLYING IN ...	imagine you are floating like a bird flying in...	imagine floating like bird flying sky
Training\Audios\1-30-1280x720_13.wav	Neutral	1-30-1280x720_13	5.0	1	1	FLOATING LIKE A BIRD FLYING IN THE SKY BREATHING	floating like a bird flying in the sky breathing	floating like bird flying sky breathing
Training\Audios\1-30-1280x720_15.wav	Neutral	1-30-1280x720_15	5.0	1	1	IN THE SKY BRAZEN AMBER	in the sky brazen amber	sky brazen amber
Training\Audios\1-30-1280x720_17.wav	Neutral	1-30-1280x720_17	5.0	1	1	BRUIN EMBRYOED	bruin embryoed	bruin embryoed
...	...	...	...	...	...	...	...	...
NaN	Surprise	dia861_utt4	NaN	0	0	NaN	NaN	NaN
NaN	Neutral	dia873_utt3	NaN	0	0	NaN	NaN	NaN
NaN	Surprise	dia884_utt7	NaN	0	0	NaN	NaN	NaN
NaN	Angry	dia887_utt11	NaN	0	0	NaN	NaN	NaN
NaN	Surprise	dia889_utt5	NaN	0	0	NaN	NaN	NaN

Figura 4.6: Archivo *CSV* multimodal (Parte 1).

Tokens	Source_DB	Set	Lemmatized_Tokens	Image_Filter	Images_KEYs	Modalities
['waiting', 'half', 'like', 'let', 'us', 'begin']	AffWild2	train	['wait', 'half', 'like', 'let', 'we', 'begin']	1	[Training/Rostros/1-30-1280x720_1/00001.jpg, T...	IAT
['imagine', 'floating', like', 'bird', 'flyin...	AffWild2	train	['imagine', 'float', 'like', 'bird', 'fly', 's...	1	[Training/Rostros/1-30-1280x720_11/03404.jpg]	IAT
['floating', like', 'bird', 'flying', 'sky', ...	AffWild2	train	['float', 'like', 'bird', 'fly', 'sky', 'breat...	1	[Training/Rostros/1-30-1280x720_13/03406.jpg, ...	IAT
['sky', 'brazen', 'amber']	AffWild2	train	['sky', 'brazen', 'amber']	1	[Training/Rostros/1-30-1280x720_15/03595.jpg, ...	IAT
['bruin', embryoed']	AffWild2	train	['bruin', 'embryoe']	1	[Training/Rostros/1-30-1280x720_17/03656.jpg, ...	IAT
...	...	...	...	...	...	...
NaN	MELD	train	NaN	1	[Training/Rostros/dia861_utt4/result_dia861_ut...	IXX
NaN	MELD	train	NaN	1	[Training/Rostros/dia873_utt3/result_dia873_ut...	IXX
NaN	MELD	train	NaN	1	[Training/Rostros/dia884_utt7/result_dia884_ut...	IXX
NaN	MELD	train	NaN	1	[Training/Rostros/dia887_utt11/result_dia887_u...	IXX
NaN	MELD	train	NaN	1	[Training/Rostros/dia889_utt5/result_dia889_ut...	IXX

Figura 4.7: Archivo *CSV* multimodal (Parte 2).

4.4.2. MERDWild

Se generó una base de datos íntegra y organizada etiquetada para cada una de las tres modalidades de datos. Esta combina información de imágenes faciales, audio y texto transcrito, obtenida de los conjuntos de datos anteriormente mencionados. La base de datos presenta múltiples registros de diferentes variaciones en iluminación, ángulo y rotación de los registros faciales, así como registros de audio con condiciones similares al mundo real, conteniendo música, ruido y conversaciones de fondo, junto con la transcripción del texto presente en estos.

Contiene un total de 15.873 registros de audio en formato *WAV* considerados de buena calidad con duraciones que oscilan entre 0.064–41 segundos.

Se dispone además de más de 905 mil imágenes consideradas de buena calidad disponibles en formato *JPG*, las cuales corresponden a cortes de rostros de tamaño 64×64 píxeles, y agrupadas en alrededor de 13.971 carpetas, asociadas a cada uterancia presente. Y finalmente, se cuenta con 15.321 transcripciones de audio consideradas de buena calidad.

La base de datos fue estructurada en un conjunto de Entrenamiento, que abarca el 78,1 % de los datos; y un conjunto de Validación con el 21,9 % restante. Cada uno de estos conjuntos contienen tres archivos *CSV*, dos de ellos son dedicados a la calidad de los archivos, donde se detalla la etiqueta, calidad de las imágenes, transcripción y calidad del audio de cada evento, y un archivo orientado a la multimodalidad, en el cual se encuentran los datos alineados por uterancia.

4.4.3. Técnicas de detección de calidad en el preprocesamiento

Dentro de las medidas de calidad utilizadas por modalidad se encuentran:

Medidas modalidad facial

Para la detección de la calidad de imagen es necesario comprender cuales son aquellos datos significativos que influyen en la calidad de una imagen, para esto se realizó el análisis de las características de los rostros en escala de grises, dentro de las técnicas usadas se encuentran:

- Se calcula el **brillo promedio** en cada una de las imágenes de la siguiente manera:

```
def evaluate_brillo(a):
    img = cv2.imread(a)
    hsv = cv2.cvtColor(img, cv2.COLOR_BGR2HSV)
    h, s, v = cv2.split(hsv)
    mean_brightness = v.mean()
    return mean_brightness
```

- Para el cálculo del **contraste promedio** se utiliza el operador Laplaciano el cual resalta los bordes en una imagen al calcular la segunda derivada de la intensidad de los píxeles, tal como muestra el siguiente código:

```
def evaluate_contraste(a):
    img = cv2.imread(a)
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    mean_contrast = cv2.Laplacian(gray, cv2.CV_64F).
       var()
    return mean_contrast
```

- Para el calculo de la **resolución** se realizaron dos funciones, una entrega el ancho y la otra el alto de la imagen:

```
def evaluate_resolucionH(a):
    img = cv2.imread(a)
    height, width, channels = img.shape
    return height
def evaluate_resolucionW(a):
    img = cv2.imread(a)
    height, width, channels = img.shape
    return width
```

- La **entropía** se calcula para medir que tan variables son los píxeles de la imagen en un radio de 15 píxeles, esto se realiza de la siguiente manera:

```
def evaluate_entropia(a):
    im1 = img_as_float(io.imread(a, as_gray=True))
    radius = 15
    selem = disk(radius)
    entropy = rank.entropy(im1, selem=selem)
    mean_entropy = np.mean(entropy)
    return mean_entropy
```

- **Scale Invariant Feature Transform (SIFT)**, se usa para obtener la cantidad de puntos faciales de una imagen para determinar si existe o no un rostro en ella.

```
sift = cv2.xfeatures2d.SIFT_create()
def detect_faces(r):
    image = cv2.imread(r)
    keypoints, descriptors = sift.detectAndCompute(
       image, None)
```

```
        keyp=len(keypoints)
        return keyp
```

- La **Similitud del Coseno** se utilizó de manera combinada con el **Histograma de Gradientes Orientados (HOG)**, para obtener la relación de semejanza entre dos imágenes consecutivas fructificando su análisis mediante las características obtenidas de **HOG**.

```
def sim_coseno(ruta_imagen1, ruta_imagen2):
    image1 = skimage.io.imread(ruta_imagen1, as_gray=
       True)
    image2 = skimage.io.imread(ruta_imagen2, as_gray=
       True)
    fd1 = hog(image1, orientations=8, pixels_per_cell
       =(16, 16), cells_per_block=(1, 1), block_norm='L2
       -Hys', feature_vector=True)
    fd2 = hog(image2, orientations=8,
    pixels_per_cell=(16, 16), cells_per_block=(1, 1),
       block_norm='L2-Hys', feature_vector=True)
    similitud = cosine_similarity([fd1], [fd2])
    return similitud[0][0]
```

Medidas modalidad audio

Para la evaluación de la calidad de un audio existen diferentes técnicas que se pueden utilizar para detectar la calidad. Entre las utilizadas en este proyecto se encuentran:

- El **promedio de nivel de potencia (PNP)**, se calcula mediante el uso de las librerías soundfile y numpy, primeramente se estima la energía como la suma de los cuadrados de las muestras de audio, luego se aplica un filtro de respuesta finita al impulso a la señal de audio, para finalmente dividir la energía en el ruido, y calcular el logaritmo de base 10 del resultado y multiplicarlo por 10 para obtener el PNP en decibelios (dB) [71].

$$\text{PNP (dB)} = 10 \cdot \log_{10}\left(\frac{\text{Energía}}{\text{Número de Muestras}}\right)$$

- El calculo del **nivel peak**, está dado por la siguiente fórmula, en la cual se puede observar que "máx" busca el valor absoluto máximo en el vector de audio y el uso del logaritmo corresponde a la conversión a decibelios (dB) [72].

$$\text{Nivel peak (dB)} = 20 \cdot \log_{10}\left(\text{máx}\left(|\text{audio}|\right)\right)$$

- Para el cálculo de la **distorsión armónica total (DAT)** se utiliza la librería soundfile para obtener frecuencia de muestreo del audio, donde "Fundamental" es el valor máximo en el vector de audio y los armónicos se calculan como la suma de todas las componentes de audio. Al igual que en las ecuaciones anteriores se utiliza el logaritmo para convertir a decibelios [73]:

$$\text{DAT (dB)} = 20 \cdot \log_{10}\left(\frac{\sqrt{\sum \text{audio}^2}}{\text{Fundamental}}\right)$$

- La **relación señal a ruido (RSR)** se mide en decibelios y para su cálculo es necesario obtener la "Potencia de Señal" mediante la suma de los cuadrados de todas las muestras de audio y la "Potencia de Ruido" como la suma de los cuadrados de las muestras de audio filtradas para estimar el ruido, para finalmente dividirlas y recibir como resultado la RSR, tal como se muestra en la siguiente ecuación [74]:

$$\text{RSR (dB)} = 10 \cdot \log_{10}\left(\frac{\text{Potencia de Señal}}{\text{Potencia de Ruido}}\right)$$

Medidas modalidad texto

En esta detección de la calidad se utilizó como base el filtro de audio, y específicamente como filtro de texto se usaron los resultados de la semántica solamente para los conjuntos de datos AffWild2 y AFEW, debido a que MELD posee las transcripciones desde el origen.

Para el **filtro semántico** se toma la transcripción del audio como entrada, luego se corrigen los posibles errores de ortografía y gramática, después se extrae la información sobre entidades en el texto utilizando "DBpedia", después se analiza el contexto semántico de las palabras en el texto y se calcula un puntaje basado en la similitud semántica, esta similitud se calcula por medio de la librería SpaCy [70] y se obtiene un valor entre -1 y 1, el cual refiere a la similitud que hay entre las palabras de la misma frase transcrita. Finalmente dicho puntaje se utiliza para filtrar si se utiliza o no el texto originalmente transcrito. Para determinar si la transcripción posee una calidad usable o no, se tomó como valor límite los valores menores o iguales a cero, por lo tanto si el valor del texto transcrito es menor o igual a cero el texto posee una baja calidad y se descarta (se le asigna el valor "0"), de lo contrario se marca como usable (se le asigna el valor "1").

4.5. Técnicas de análisis

En el siguiente listado se presentan diferentes métodos utilizados en el modelo para realizar predicciones de emociones en las tres modalidades. Estas arquitecturas fueron

seleccionadas de manera idéntica a las del proyecto [1]. Se facilitaron los códigos para adaptarlos y comprobar la hipótesis de este Trabajo de Título.

- VGG19: su sigla se debe al grupo de investigadores que la creó (Visual Geometry Group de la Universidad de Oxford), el número 19 se debe a las 19 capas de profundidad que posee esta red neuronal convolucional (CNN) [10]. Esta red intenta simular las conexiones que existen entre las neuronas de los seres humanos, las CNN son utilizadas para el reconocimiento de patrones en imágenes por medio de la extracción de características [75]. En este proyecto se aplican las 19 capas convolucionales con filtros con un tamaño de 3x3, sumado a ello, 2 capas en las cuales cada una de ellas posee 64 canales, este modelo entrega como salida el valor máximo mediante la operación max pooling la cual tiene un tamaño de 2x2. Luego, se alterna el anterior proceso con grupos de 2 capas con 128 canales, 4 capas de 256 canales, 4 capas de 512 canales y 4 capas de 512 canales. Finalmente se realiza la operación max pooling, dicha salida entra a una red MLP con 3 capas densas de tamaños 4.096, 4.096 y 1.000, luego una capa final con una función de activación Softmax [1].

- ResNet50: su sigla se debe a que es una red neuronal residual, el número 50 corresponde a la cantidad de capas de profundidad que posee el modelo. Las redes residuales también pertenecen a las CNN y se utilizan para el reconocimiento de patrones en imágenes. Su funcionamiento se basa en saltos entre capas intermedias para lograr una mejora en los resultados y obviar ciertas capas que generan pérdidas de datos relevantes [11]. Como entrada al modelo se espera una imagen del espectrograma del audio con un tamaño correspondiente a 224x224, la cual corresponde solamente a los primeros 7 segundos del audio. Luego de una capa convolucional de tamaño de filtro 7x7 y 64 canales. Luego la imagen pasa por múltiples bloques residuales, los cuales se componen de 3 capas convolucionales de tamaño de filtro 1x1, 3x3 y 1x1, después la entrada del bloque se suma a la salida del bloque. Después de haber pasado por varios grupos, la salida se agrupa seleccionando el valor máximo del área (max pooling), de esta manera se reduce su tamaño. Esta operación sucede en los bloques 3, 4, 6 y 3 bloques de ResNet. Finalmente, la salida se promedia, creando un vector de características de longitud igual a 2.048. Este vector se pasa por una capa densa y una capa de salida con una activación Softmax. Esta última capa de salida tiene una cantidad de neuronas correspondiente a la cantidad de emociones (7) [1].

- DialogXL: este componente se basa en XLnet, se utiliza para el reconocimiento de emociones en conversaciones y busca almacenar un contexto histórico prolongado y autoatención consciente del diálogo para lidiar con las estructuras que contienen varias partes [12]. Consiste en una capa de incrustación, 12 capas de transformación y una red neuronal de retroalimentación. En ella se capturan dependencias útiles entre hablantes, donde cada oración dicha por un hablante se enlaza a través de una capa

de incrustación, la cual tokeniza la oración en una serie de vectores. Luego, la representación se introduce a una pila de redes neuronales, en cada capa de la pila tiene un componente de autoatención consciente del diálogo y un componente de recurrencia de expresiones. El estado oculto del token de categorización y el contexto histórico se alimentan a través de una red neuronal de retroalimentación que se encuentra al final de la última capa para producir la emoción reconocida [1].

- Embracenet+: es el método de fusión que se utiliza para poder combinar los resultados de las diferentes modalidades. Consiste en una arquitectura que se bifurca e integra Embracenet con otros métodos de fusión [13]. La arquitectura involucra tres modelos simples de Embracenet que trabajan para mejorar el aprendizaje de correlación de las modalidades, así como los resultados finales. Está conformado por una capa lineal de 32 neuronas (D1,1), una capa de abandono con una probabilidad de desintegración de 0,5 y otra capa lineal de 16 neuronas (D1, 2) componen cada una de las capas de acoplamiento alteradas. Además como técnicas de fusión se utiliza una suma ponderada, cuya salida es un vector de n probabilidades, una por cada emoción (7), y una concatenación, cuya salida es un vector de 3n, una por cada modalidad (3). Posteriormente, otro Embracenet recibe tres vectores de valores 16, n y 3n, los cuales funcionan como modalidades. Estos vectores se manejan acoplando capas de una capa lineal de 16 neuronas cada una, lo que lleva a una última cada lineal adicional de n neuronas, que genera la predicción final [1].

Las variables de entrenamiento de los modelos anteriormente descritos para las modalidades de audio, texto e imagen, se asignaron respectivamente los valores de batch size: 16, 32 y 32, épocas (epoch): 20, 40 y 20. Además, para el método de fusión se utilizaron 20 épocas y 16 de batch size. Para cada una de las modalidades y también para Embracenet+ se usó un learning rate de 0,0001. Estos valores son exactamente los mismos que se usaron en el proyecto [1], debido a que como se mencionó anteriormente se desea comparar los resultados obtenidos.

4.6. Diseño de experimentos

El diseño del proceso de experimentación comienza con la preparación de los datos incluye limpieza y selección de características para optimizar su uso en modelos de Deep Learning, tal como se describe en la Sección 4.3. Luego se realiza la fase correspondiente al modelado, consiste en la selección y aplicación de modelos de Deep Learning para el MER. En la evaluación se llevarán a cabo las pruebas de los modelos construidos para determinar si se logra cumplir con lo esperado, junto con ello, se plantearon tres preguntas de investigación:

1. *¿Cuál es la emoción que se encuentra con mayor frecuencia en la totalidad de modalidades?*

   Para responder la primera pregunta se debe realizar un análisis descriptivo de los conjuntos de datos con el preprocesamiento ya realizado, para obtener la emoción que se encuentra en una mayor cantidad en el dataset unificado.

2. *¿Es posible realizar mejoras en los conjuntos de datos y en la arquitectura utilizada?*

   Para lograr responder esta pregunta primeramente es necesario haber realizado el entrenamiento, validación y pruebas para obtener las métricas de evaluación del modelo, y con respecto a los resultados extraer conclusiones que propongan posibles mejoras a los conjuntos de datos y a la arquitectura utilizada.

3. La anteriormente mencionada hipótesis de este proyecto: *¿Es la unificación de los datasets suficiente para obtener mejores resultados en MER in-the-wild?*

   Para responder la hipótesis del proyecto se requiere obtener los resultados finales del MER *in-the-wild* para poder comparar con los resultados obtenidos en el proyecto [1].

Dentro de las variables que son necesarias para el desarrollo de los experimentos se encuentran en las variables utilizadas para medir la calidad de las distintas modalidades, entre las cuales se encuentra para la modalidad de imágenes faciales el brillo, contraste, entropía, entre otros. Para la modalidad de audio se encuentra la distorsión armónica total, el nivel peak, la relación señal ruido, entre otras. Además se tiene la semántica en el modalidad de texto, todas las variables anteriormente mencionadas en la Sección 4.4. Sumado a las anteriores variables, se tienen los hiper-parámetros de los modelos de DL, entre ellos se encuentra tamaño del lote (batch size), tasa de aprendizaje (learning rate) y las épocas (epoch). Es importante destacar que dentro de las variables necesarias para poder medir el desempeño del modelo son F1-Score, Recall, Precisión y Accuracy, las cuales se encuentran detalladas en la Sección 4.7. La última variable a contemplar y no menos importante son las emociones que se encuentran en cada uno de los datasets, los cuales se deben utilizar para hacer la estandarización y unificación de las emociones.

4.6.1. Caso de estudio

El caso de estudio que se utiliza en este proyecto corresponde al proceso de entrenamiento y validación del modelo usando la base de datos unificada llamada MERDWild, una vez realizado esto, se evaluará el modelo entrenado por medio de la utilización del conjunto de datos MAFW para comprobar su funcionamiento con datos que no pertenecen a los conjuntos de datos con los que se entrenó el modelo. En este caso se analizarán los

resultados por medio de la métricas de evaluación anteriormente mencionadas. Particularmente se evaluarán las diferentes combinaciones de modalidades, es decir, se evaluará cada modalidad individualmente, luego los pares de modalidades (audio y texto, imagen y audio, imagen y texto), y finalmente el caso donde las tres modalidades se encuentran activas.

4.7. Interpretación de resultados

Para la interpretación de los resultados se utilizarán cuatro métricas de evaluación de desempeño del modelo generado, dichas métricas están descritas en la Subsección 2.1.4 y sus nombres son: accuracy, recall, precisión y F1-Score.

Para la visualización de los resultados se utilizarán las métricas descritas en la sección 4.7 para realizar gráficas de evolución del proceso de entrenamiento contra el proceso de validación de los resultados. Además de la utilización de la matriz de confusión, conocida en inglés como "Confusion Matrix" la cual es una herramienta esencial para evaluar el rendimiento de un algoritmo de clasificación, ya que proporciona una visión más precisa de cómo se están clasificando las distintas emociones, esta se basa en el recuento de aciertos y errores en cada una de las categorías de clasificación. Esto nos permite identificar visualmente si el algoritmo está realizando clasificaciones incorrectas y en qué medida lo está haciendo [76], tal como se muestra en la Figura 4.8, en la cual se muestran los falsos positivos (FP), falsos negativos (FN), verdaderos positivos (TP) y verdaderos negativos (TN). La Figura 4.9 muestra una representación de una configuración deseada de la matriz de confusión para las siete emociones.

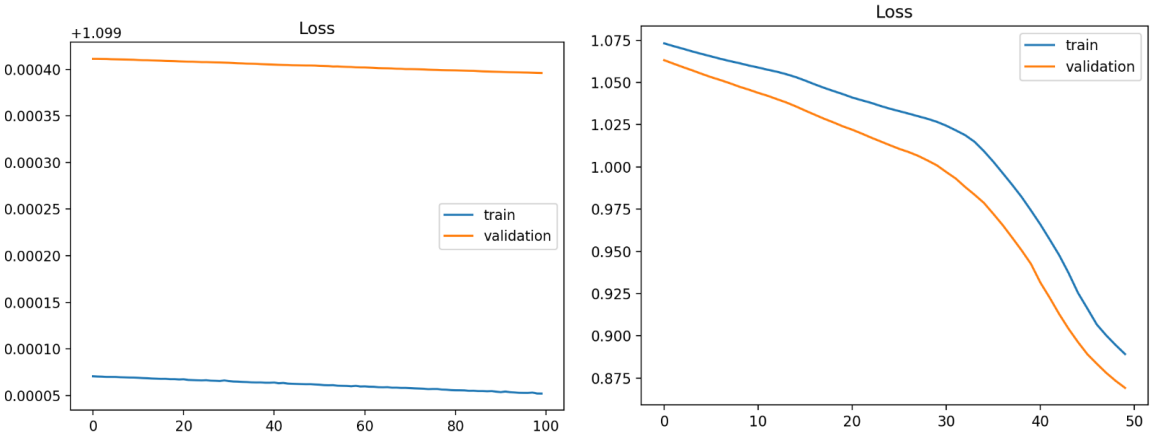
		PREDICTED	
		0	1
TRUE LABEL	0	TN	FP
	1	FN	TP

Figura 4.8: Estructura matriz de confusión.

Figura 4.9: Ejemplo de configuración ideal de matriz de confusión.

Existen tres dinámicas que son comunes de observar en las curvas de aprendizaje:

- Underfitting: También conocido en español como subajuste, este se refiere a un modelo que no logra capturar los patrones presentes en los datos de entrenamiento ni generalizar adecuadamente a nuevos datos. Cuando un modelo presenta underfit su rendimiento será deficiente en los datos de entrenamiento [77]. En la Figura 4.10, se muestran dos casos de Underfitting.



(a) Ejemplo 1 de subajuste (Underffiting) [78]. (b) Ejemplo 2 de subajuste (Underffiting) [78].

Figura 4.10: Ejemplos de subajustes.

- Overfitting: En español es conocido como "Sobre-ajuste". Se produce cuando un modelo aprende los detalles y el ruido presentes en los datos de entrenamiento, lo cual afecta negativamente su rendimiento en nuevos datos, debido a que el modelo aprende de memoria, esto implica que el modelo adquiere conceptos a partir del ruido o las fluctuaciones aleatorias presentes en los datos de entrenamiento. El problema radica en que estos conceptos no son aplicables a nuevos datos y tienen un impacto negativo en la capacidad de generalización del modelo [77]. En la Figura 4.11 obtenida de la fuente [79], se ve un caso de Overfitting, donde se visualiza la curva de la pérdida (loss) y la exactitud (accuracy) en el entrenamiento y validación.

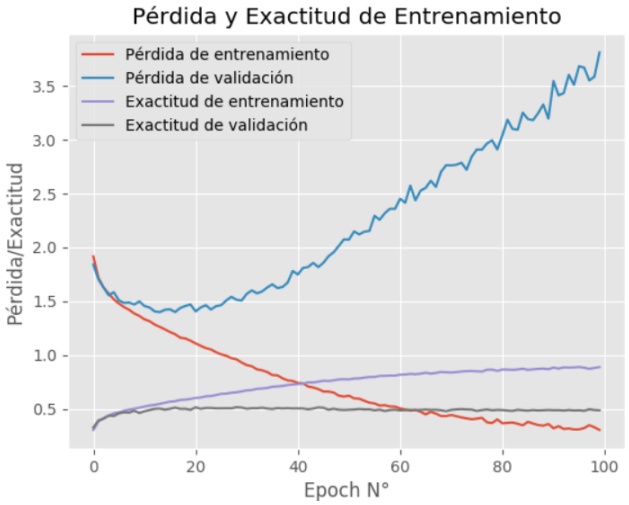
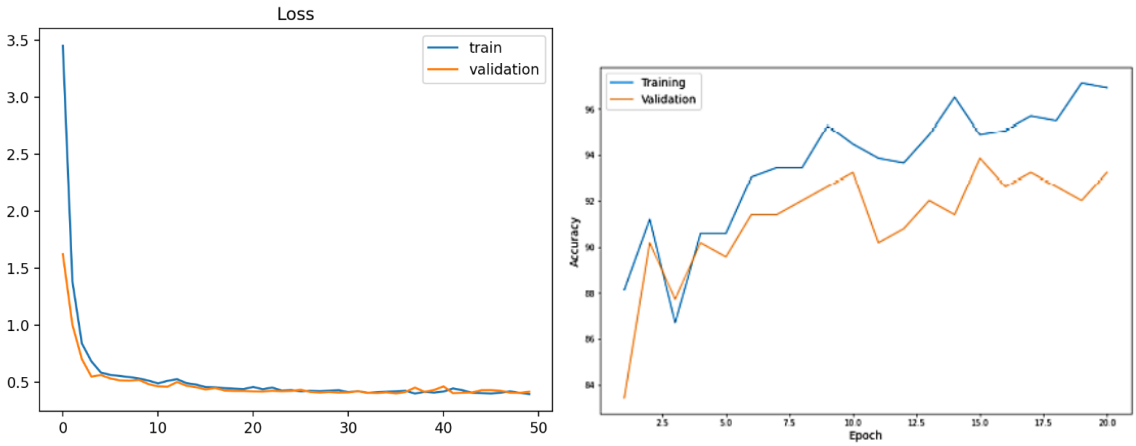


Figura 4.11: Ejemplo de sobre-ajuste (Overffiting).

- Good Fit: Sucede cuando se encuentra un punto medio en el aprendizaje del modelo en el que no ocurre underfitting ni overfitting [77]. Un ejemplo de un buen ajuste se encuentra en la Figura 4.12 (a), donde se muestra una gráfica de la pérdida (loss) con respecto a las épocas (epoch) sobre el conjunto de entrenamiento y de validación, en dicha figura en la parte (b) se observa un ejemplo de un buen ajuste con respecto al accuracy y las épocas.

Las curvas ROC (Receiver Operating Characteristic curve), se utiliza para evaluar el rendimiento de un clasificador, su propósito es mostrar la relación entre la sensibilidad y la especificidad de los modelos de DL. Cada punto en la curva representa un umbral de probabilidad diferente para clasificar una entrada según una emoción específica. La sensibilidad indica la capacidad del modelo para detectar correctamente esa emoción, mientras que (1 - especificidad) refleja su propensión a hacer falsos positivos al clasificar incorrectamente. Un punto ideal en la curva se encuentra cercano a la esquina superior izquierda, indicando



(a) Good fit con respecto a la pérdida. [78]. (b) Buen ajuste con respecto al accuracy [80].

Figura 4.12: Ejemplos de un buen ajuste.

un alto rendimiento del modelo en sensibilidad y especificidad. La diagonal de referencia representa un rendimiento aleatorio [81].

En la Figura 4.13 se presenta una curva ROC que ilustra tres de los casos posibles. La curva de color rosa, que representa la emoción "Neutral", muestra un rendimiento consistente en la clasificación de dicha emoción. En contraste, la curva de color azul, correspondiente a la emoción "Happy", revela un rendimiento deficiente en la clasificación, mientras que las demás curvas asociadas a las restantes emociones presentan un rendimiento intermedio.

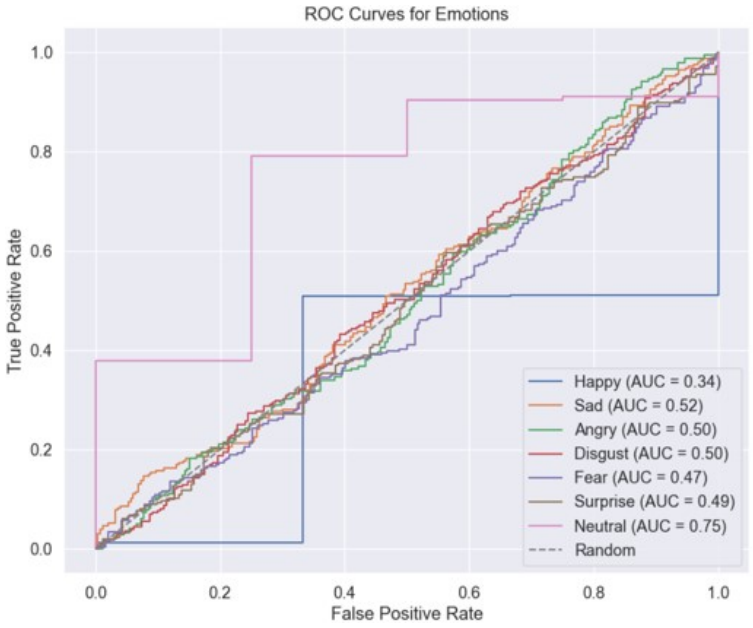


Figura 4.13: Ejemplo de curvas ROC.

Capítulo 5

Implementación

En esta sección se presenta la implementación del trabajo realizado revelando tanto el Software como los componentes de Hardware utilizados para el desarrollo de este proyecto de título, explicando la estrategia de implementación utilizada.

5.1. Software utilizado

Anaconda [82] es una distribución de software que se utiliza ampliamente en el campo del DL, este utiliza el lenguaje de programación llamado Python [83], el cual se usa frecuentemente para proyectos de DL y ML, por su potencia y efectividad, además de poseer extensas librarías gratuitas y una sintaxis fácil de aprender [84]. Anaconda consiste en un entorno completo, el cual es fácil de usar y que incluye paquetes y herramientas que facilitan el proceso de trabajo. Dentro de ella se encuentra Jupyter Notebook [85], el cual es una de las varias aplicaciones que posee Anaconda. Jupyter Notebook es una aplicación web de código abierto que permite crear documentos interactivos que contienen código, visualizaciones y texto explicativo. Dentro de las ventajas de utilizar Jupyter Notebook de Anaconda es que posee una fácil instalación, creación de entornos virtuales para trabajar en múltiples proyectos, es posible crear y ejecutar código de manera interactiva en celdas individuales, se puede incluir texto explicativo al código, junto con ecuaciones y visualizaciones para la creación de informes directamente en el entorno de desarrollo. Algunas de las librerías relevantes más utilizadas para el desarrollo de proyectos de DL:

- Scikit-learn: es una biblioteca de aprendizaje automático, esta proporciona diferentes algoritmos de clasificación, regresión, agrupación, entre otros. Además, posee herramientas para el proceso de evaluación de modelos, selección de características y preprocesamiento de datos [86].
- Librosa: es una librería usada para el análisis de música y audio [87].

- Matplotlib: es una librería de visualización, posee gráficas estáticas, animadas e interactivas en Python [88].
- Seaborn: una librería especializada en la visualización de datos, basada en matplotlib. Ofrece interfaces y gráficos de alto nivel para crear representaciones atractivas e interactivas, especialmente útiles en el ámbito estadístico [89].
- PyTorch: es una biblioteca de Python de ML de código abierto, la cual se basa en tensores y proporciona una forma flexible de construir y entrenar modelos de aprendizaje profundo [90].
- NumPy: es una biblioteca fundamental para la computación numérica en Python. Proporciona estructuras de datos de matriz multidimensional eficientes, así como funciones para realizar operaciones matemáticas y manipulaciones de matrices [91].
- Pandas: es una biblioteca utilizada para el análisis y manipulación de datos. Ofrece estructuras de datos flexibles, como por ejemplo los llamados DataFrames, que permiten trabajar con datos tabulares de forma eficiente. Pandas también proporciona funciones para filtrar, limpiar y transformar datos, entre otras funciones [92].
- CV2 (OpenCV): es una biblioteca de visión por computadora de código abierto. Proporciona herramientas y funciones para procesar y analizar imágenes y videos. OpenCV se utiliza frecuentemente para tareas de detección de objetos, reconocimiento facial, seguimiento de objetos y más [93].

En la Figura 5.1 se puede ver el esquema de alto nivel que describe los componentes de Software que son utilizados en este proyecto de título.

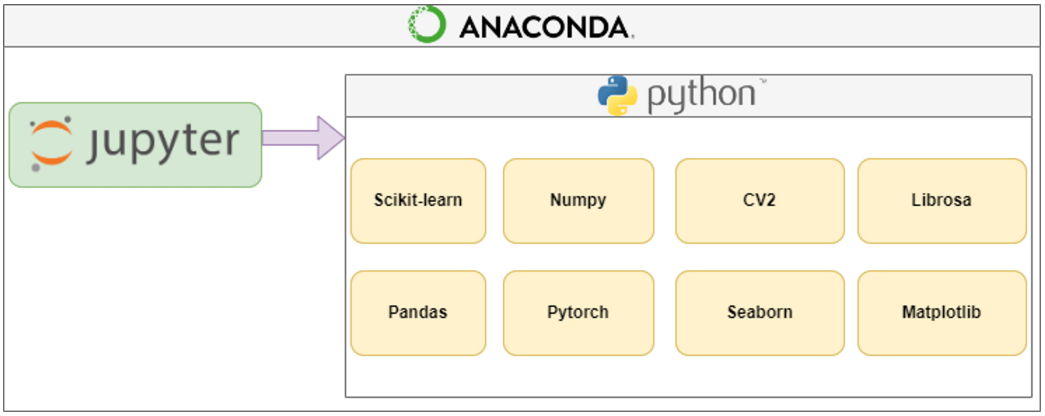


Figura 5.1: Esquema de alto nivel de integración de componentes del sistema.

5.2. Hardware utilizado

Para el desarrollo del presente Trabajo de Título se utilizó una computadora proporcionada por la Universidad de Valparaíso para realizar la fase de preprocesamiento. Para la ejecución del algoritmo para la clasificación de emociones multimodales se usó un computador personal, en el cual se procesó la arquitectura propuesta junto, para ejecutar el proceso de entrenamiento, validación y pruebas MERDWild. En la Figura 5.2 se muestra un esquema e información necesaria para el desarrollo del proyecto en términos de Hardware, en la cual se muestra el acceso a Internet, el computador, los datos y el algoritmo.
Las especificaciones del computador proporcionado por la universidad para el preprocesamiento de los datasets para el desarrollo del proyecto son:

- Procesador: Intel(R) Core (TM) i7-8700 CPU @ 3.20GHz 3.19 GHz basado en x64
- Memoria RAM: 64 GB 4200 MHz
- Disco Duro: 3,63 TB
- GPU: NVIDIA Quadro P1000 4 GB GDDR5 (128 bit)
- Sistema operativo: Windows 10 Pro de 64 bits

Las especificaciones del computador personal usado para el entrenamiento, validación y pruebas son:

- Procesador: Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz (8 CPUs) en x64
- Memoria RAM: 16 GB 2667 MHz
- Almacenamiento: 512 GB SSD (M.2)
- GPU: NVIDIA GeForce GTX 1650 4 GB GDDR5 (128 bit)
- Sistema operativo: Windows 11 Home Single Language 64 bits

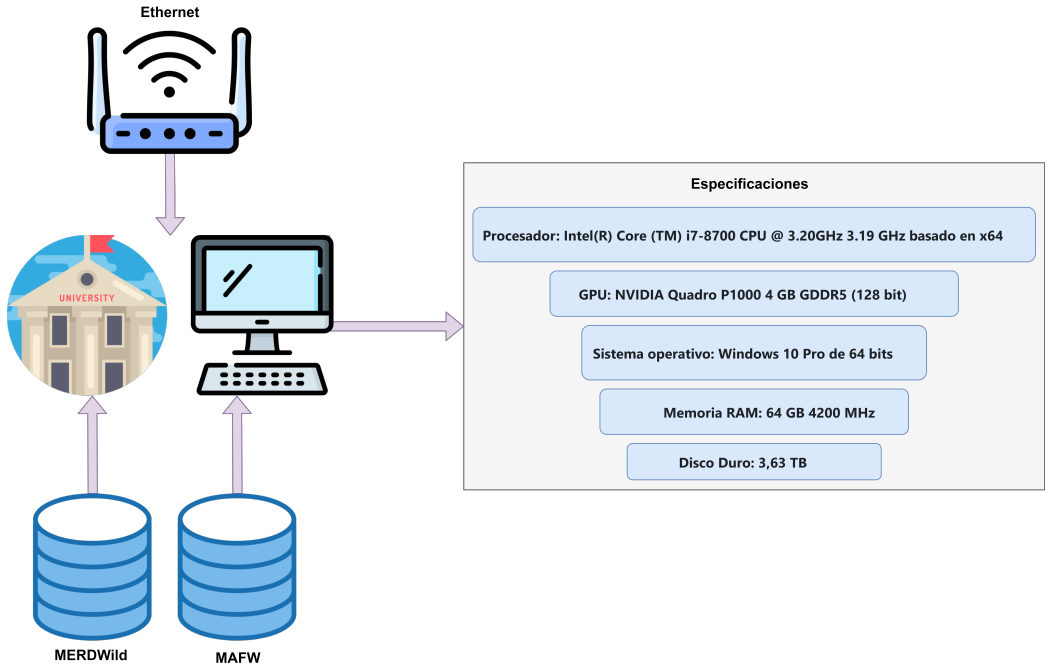


Figura 5.2: Esquema de alto nivel del sistema en términos de Hardware.

5.3. Estrategia de implementación

En esta sección se describe la etapa correspondiente a la estrategia utilizada para realizar el proceso de implementación. En la cual se mostrará el trabajo realizado para obtener los resultados, mostrando el patrón utilizado, las etapas de este proceso y un esquema de alto nivel que representa la estrategia de implementación.
Las diferentes etapas del proceso corresponden a cada uno de los pasos realizados para la obtención de la emoción final, estas etapas se encuentran descritas en el siguiente listado:

- Obtención de datos: la totalidad de los conjuntos de datos utilizados en este proyecto fueron solicitados directamente a los autores vía correo electrónico.
- Almacenamiento de datos: la base de datos MERDWild se encuentra almacenada en OneDrive, las instrucciones para obtenerla se encuentra en un GitHub llamado "MERDWild" y los scripts de entrenamiento y preprocesamiento se encuentran en el repositorio de GitHub llamado "MultimodalEmotionRecognitionInTheWild-Thesis".
- Preprocesamiento de los datos: en esta etapa se trataron los datos de cada una de los conjuntos de datos de manera individual, separando cada una de las modalidades presentes. Donde se realizó la extracción de los datos de audio, texto e imágenes

faciales en el caso de no existir por cada uno de los datasets. En esta etapa se realiza la limpieza, transformación, integración, estandarización y unificación de los datos.

- Procesamiento de los datos: una vez que los datos ya se encuentran unificados, transformados y estandarizados se ingresan a la arquitectura basada en técnicas de DL descritas en la Sección 4.5, en las cuales se procesan las tres modalidades anteriormente descritas, para luego fusionar los pesos obtenidos, validar el modelo y calcular la emoción final.

- Análisis de los datos: en esta etapa se utilizan las métricas de evaluación descritas en la Sección 4.7, aplicando los criterios desarrollados en dicha Sección para obtener una visión de los resultados con respecto a las curvas de aprendizaje como también respecto a las métricas de evaluación.

- Resultados: se obtienen los resultados de los modelos, los cuales son utilizados para sacar conclusiones, tomar decisiones y mejorar los modelos de ser necesario.

Se requirió utilizar el patrón de implementación "Extract, Transform, Load" (ETL) [94] en la etapa de preprocesamiento de los tres conjuntos de datos utilizados en este proyecto. El objetivo principal de ETL es extraer los datos de los diversos datasets en su estado original sin procesar, con el fin de obtener las tres modalidades necesarias. Luego, se llevó a cabo la transformación de los datos, donde se realizaron acciones como limpieza, normalización y estructuración para lograr una uniformidad entre los diferentes conjuntos de datos en términos de clases de emociones y modalidades. Por último, los datos preprocesados fueron cargados en los modelos correspondientes para su posterior procesamiento y análisis.

En la Figura 5.3 se puede ver el esquema de alto nivel de estrategia de implementación utilizado, en el cual se observan las diferentes etapas de la implementación que fueron descritas anteriormente.

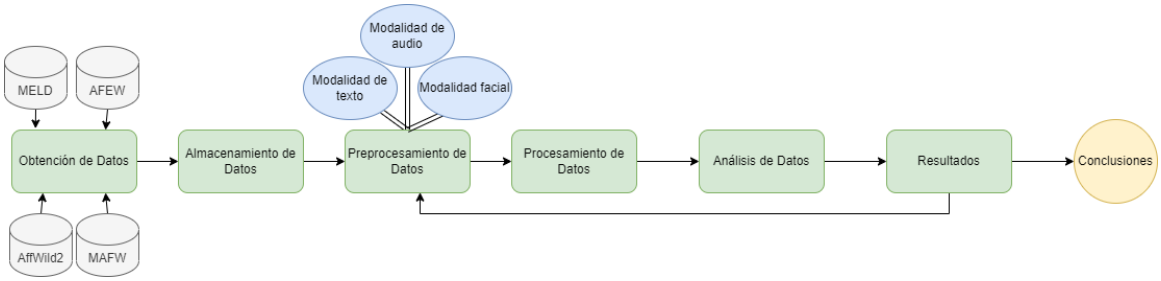


Figura 5.3: Esquema de alto nivel de estrategia de implementación.

5.4. Visualizaciones

Dentro de los resultados obtenidos, se incluye el trabajo realizado en el preprocesamiento, tal como se describe en la Sección 4.3. Se encuentran ejemplos de imágenes que ilustran los resultados de la exclusión e inclusión de datos por modalidad y la unificación de las etiquetas de emociones. Para las modalidades de audio y texto se crearon dos archivos *CSV* genéricos que contienen a ambas modalidades (uno para entrenamiento y otro para la validación), lo mismo se realizó para la modalidad de imágenes faciales. Finalmente se encuentran dos archivos *CSV* orientados a la multimodalidad. Esto se debe a que es fundamental para la unificación de los datos una estandarización y transformación del formato de los archivos para una correcta integración y homogeneidad de la información.

5.4.1. Modalidad de rostros

Tal como se explicó en la fase de preprocesamiento, en los datasets MAFW y MELD se realizaron recortes automáticos de las imágenes faciales para cada uno de los vídeos, la estructura que se le asignó a los recortes fue la misma que se utiliza en AffWild2 y AFEW, esta consiste en carpetas, en cada una de las carpetas se encuentran los recortes faciales del vídeo, dicha carpeta posee el nombre del vídeo al que corresponden los recortes faciales. La estructura dada para las imágenes faciales se muestra en la Figura 5.4 y en la Figura 5.5 se puede ver un ejemplo de los recortes realizados para uno de los videos de MELD.

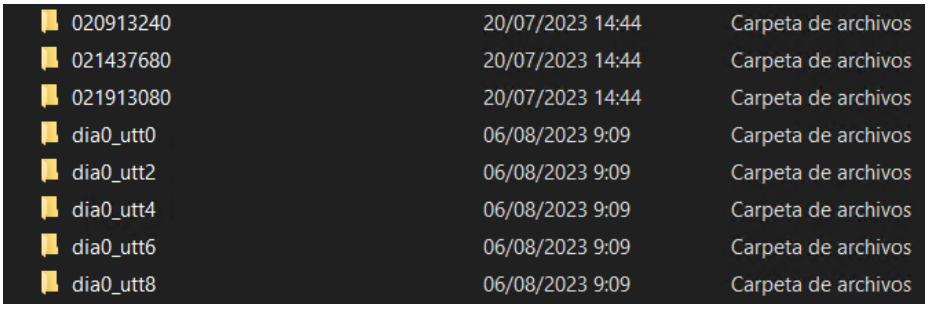


Figura 5.4: Estructura de las carpetas que poseen los recortes faciales.

Figura 5.5: Recortes automáticos de rostros realizado para MELD.

Una vez que se tienen cada uno de los recortes faciales para cada una de los conjuntos de datos se realizó el proceso de detección de calidad de imágenes faciales, cuyos valores límites se encuentran en la Tabla 4.2. A continuación, en la Figura 5.6 se muestran algunas de las muestras que fueron descartadas por su baja calidad. Se generó un archivo *CSV* genérico para cada uno de los datasets, filtrando de esta manera los datos que se deben utilizar para MER.

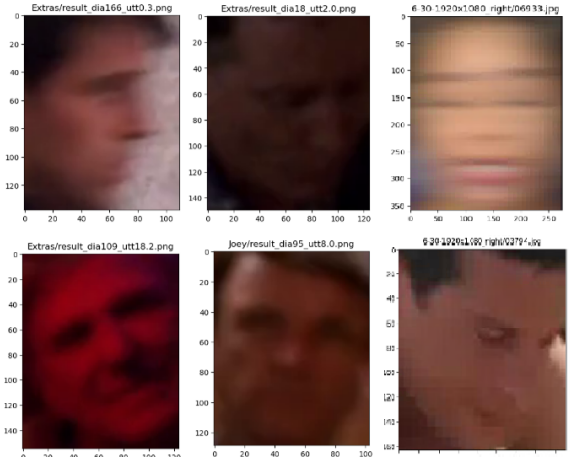


Figura 5.6: Ejemplos de imágenes faciales descartadas por baja calidad.

5.4.2. Modalidad de audio

Para esta modalidad se realizó el preprocesamiento que se describe en la Sección 4.3. Fue necesario primeramente obtener cada uno de los audios separados de los vídeos, este proceso se realizó de manera automática para los conjuntos de datos MAFW y MELD mediante la librería VideoFileClip, para AFEW y AffWild2 ya se encontraban extraídos y fueron facilitados por los autores del proyecto [1], pero en AffWild2 fue necesario recortarlos tal como se explicó en la Sección 4.4. La Figura 5.7 muestra un espectrograma de uno de los audios de MERDWild, en el cual fue necesario realizar un rellenado con ceros a la derecha del audio para obtener los siete segundos necesarios para el entrenamiento, y en la Figura 5.8 se observa el mismo audio pero en forma de onda.

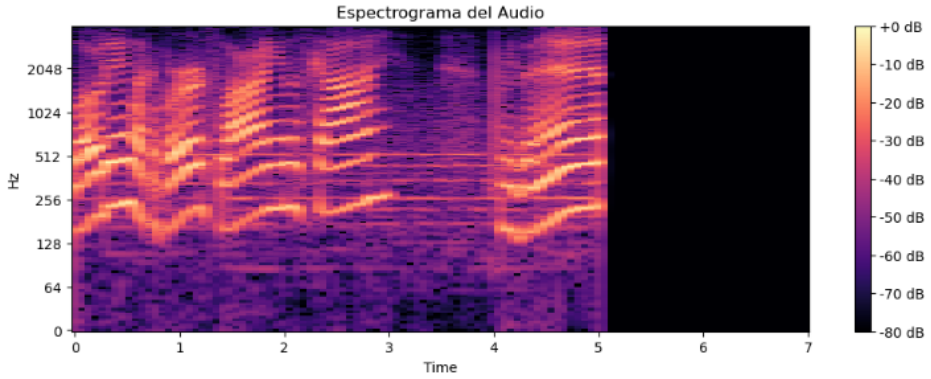


Figura 5.7: Ejemplo de un audio presentado como espectrograma.

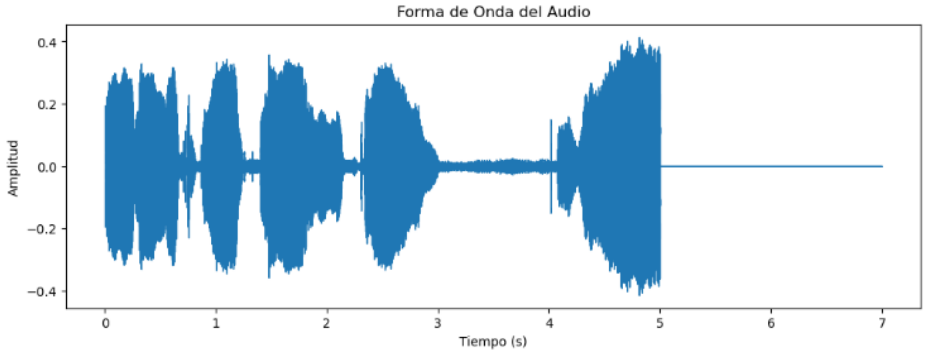


Figura 5.8: Ejemplo de un audio presentado en forma de onda.

Una vez obtenidos los audios para cada uno de los datasets se realizó la revisión de calidad de ellos, en los cuales se determinó la calidad con los valores límites que se muestran en la Tabla 4.3, luego de filtrar los audios según su calidad se guardó un archivo *CSV* genérico para cada uno de los conjuntos de datos.

5.4.3. Modalidad de texto

Para la presente modalidad se realizó la transcripción de los conjuntos de datos AFEW y AffWild2, debido a que el conjunto de datos MELD posee las transcripciones desde su origen. Para realizar la transcripción del texto se utilizó SpeechBrain. Una vez transcritos los audios se realizaron las técnicas de detección de calidad descritas en la Sección 4.4.

En la Figura 5.9 se encuentra una muestra del análisis de las técnicas de transcripción automática de audio a texto que se analizaron para seleccionar cual era la mejor para resolver este problema *in-the-wild*.

filename	Duration	Vitouphy	SpeechBrain	jonatasgrosman	SpeechRecognition
r:\sentirobc	5,5	anburm fuk lke all de don j curior stilon on blost	AND THEN IF YOU'D LIKE I'LL INDULGE YOUR CURIOSITY ALI	ande feed like all indulge you curiou steel not longst	and then if you'd like Allendale curiosity all night
r:\sentirobc	2,2	i dont now before this i was ha	I DON'T KNOW BEFORE THIS I WAS	now before this i was halp	no before this
r:\sentirobc	3,3	i need to get into grangolds into one of the vols s	I NEED TO GET INTO GRIMAUD'S INTO ONE OF THE VAULTS	i need to get into grangos into one of the volds	I need to get into Gringo's into one of the vaults
r:\sentirobc	2,8	you are not dancin you are not dansing im a	IF YOU ARE NOT DANCING YOU ARE NOT DANCING IN A	you are not dancing you are not dancing ino	you are not dancing you are not dancing
r:\sentirobc	1,8	how be derbye a	YOU'RE RIGHT IN THIRTY EIGHT	happy thirty et	happy
r:\sentirobc	3,0	oh no diany what is it	OH NO DANNY WHAT IS IT	oh no dne wher is it	
r:\sentirobc	2,7	nes but i defended you i said na	BUT I DEFENDED YOU I SAID NO	idness but i defended you i said na	but I defended you I said
r:\sentirobc	1,8	a spelr aera	AND CYRIL A	spiner a	Sarah
r:\sentirobc	8,0	good morning its another lovely day the rizon sh life folivy	GOOD MORNING AND ANOTHER LOVELY DAY AND THE WRI	good morning it's another lovely day the rizon shy life for livi	good morning it's another lovely day The Rise and S
r:\sentirobc	4,1	to geda welwe weter is kancus bet	DON'T GET AT HIM WE WILL MAKE IT HIS COUNTRY BET	together wil the mitores country bet	together we will make this country
r:\sentirobc	4,9	wa you te caereginanes for m comstuborcange	WELL YOU ARE A KIND OF GENIUS WHEN IT COMES TO MAC	well youow that rogenous when it comes to machines	well you're kind of a genius when it comes to machi
r:\sentirobc	3,7	h how well i kan scarcely imagind i egreater endorsemen	OH WELL I CAN SCARCELY IMAGINE A GREATER INDORSEME	how well i can scarcely imagine a greater endorsement	
r:\sentirobc	1,7	u	ANTONYM	t	
r:\sentirobc	4,6	a great hahaai most the dimast disastrous and o quietly crak	AGREED BUT I'M FORCED THE DIAMONDS DISASTROUS AND	grat we most thedirmas disastrous or quietly crie	
r:\sentirobc	4,6	atherewas ram avapn	BUT WHEN THERE IS A RAIN AYE	newsram	
r:\sentirobc	4,3	i wanted to tell you that im im realy	I WANTED TO TELL YOU THAT I'M I'M WELL	i wanted to tell you that i'm really	I wanted to tell you that I'm
r:\sentirobc	3,5	o no gond be	OH NO GOD	oh no god	oh no God
r:\sentirobc	5,0	and im symmilly i reallised that it was actually right	AND THEN SUDDENLY I REALIZED THAT IT WAS ACTUALLY RI	and then simily i realized that it was actually right	and then suddenly I realized that it was actually righ
r:\sentirobc	1,7	squ betin you want tok bras	HIS GREY FUMES RAISED THE FACE	scrog too you want to pack	
r:\sentirobc	2,0	ran t onot	REMNANTS	room	
r:\sentirobc	1,7	nev ofde	NON AYE	les nd t	
r:\sentirobc	1,2	e ka jo	HM HM		
r:\sentirobc	1,7	teto a hn apoc a	A LITTLE LONGER BUT A LITTLE LONGER	edo oven hom h e profoxet beter	
r:\sentirobc	1,7	hehahatha ii	HM HM	hpe	
r:\sentirobc	2,1	aa	HY A A		
r:\sentirobc	3,9	a a a ampa	HM HM HM HM HM	dag	

Figura 5.9: Transcripciones de audio a texto utilizando cuatro técnicas diferentes.

5.4.4. Archivos con etiquetas de emociones

Para integrar los datasets es primordial lograr unificar y estandarizar las etiquetas pertenecientes a las emociones de cada uno de los datasets, las cuales poseen diferentes maneras de presentarse por cada uno de los conjuntos de datos, tal como se muestra en la Tabla 3.1. En las Figuras 5.10 y 5.11 se puede observar un extracto del archivo *CSV* correspondiente a la validación de audio y texto, junto con cada una de sus columnas y datos de las 3 primeras filas. En la Figura 5.11 se puede ver en la columna "Lemmatized_Tokens", la cual almacena la fase final de todo el preprocesamiento de texto y en la Figura 5.10 se observa la columna "Emotion", que contiene la emoción expresada.

Unnamed: 0	Emotion	Set	Multimodal_Connection	Source_DB	Audio_KEY	Audio_Filter	Transcription	Lowercase_Transcription
0	Neutral	val	118-30-640x480_1	AffWild2	Validation/Audios/118-30-640x480_1.wav	1	OH MY NAME IS GUSHING LITTLE FEATHER I'M APACH...	oh my name is gushing little feather i am apac...
1	Neutral	val	118-30-640x480_11	AffWild2	Validation/Audios/118-30-640x480_11.wav	1	IS ANY MARYLAND BRENDA THIS EVENING AND HE HAS...	is any maryland brenda this evening and he has...
2	Neutral	val	118-30-640x480_13	AffWild2	Validation/Audios/118-30-640x480_13.wav	1	AND HE HAS ASKED ME TO TELL YOU IN A VERY LONG...	and he has asked me to tell you in a very long...

Figura 5.10: Parte 1 del archivo *CSV* (visualizado como DataFrame) de Validación de audio y texto.

Text_Without_Stopwords	Tokens	Lemmatized_Tokens	Text_Filter	Semantic_Filter	Duration	Average_Power_Level	Peak_Level	Total_Harmonic_Distortion	Signal_to_Noise_Ratio
oh name gushing little feather apache present	['oh', 'name', 'gushing', 'little', 'feather',...	['oh', 'name', 'gush', 'little', 'feather', 'a...	1	0.133424	5.0	-27.763789	-8.322293	27.712709	-12.834611
maryland brenda evening asked tell little	['maryland', 'brenda', 'evening', 'asked', 'te...	['maryland', 'brenda', 'evening', 'ask', 'tell...	1	0.118613	5.0	-24.989113	-7.741013	31.618778	-13.112777
asked tell long speech cannot show	['asked', 'tell', 'long', 'speech', 'can', 'no...	['ask', 'tell', 'long', 'speech', 'can', 'not'...	1	0.145056	5.0	-25.925499	-7.741013	30.656403	-12.590220

Figura 5.11: Parte 2 del archivo *CSV* (visualizado como DataFrame) de Validación de audio y texto.

Capítulo 6

Pruebas

En el presente capítulo se muestran las pruebas que fueron realizadas tanto para las calidades de las tres modalidades de manera individual como también los resultados obtenidos del entrenamiento del modelo planteado.

6.1. Pruebas de calidad por modalidad

En esta sección se presentan los resultados de las pruebas de calidad para las tres modalidades de datos, donde los límites de los valores fueron ajustados específicamente para cada variable. Esto se realizó en función de la evaluación humana de las gráficas, acompañada de la escucha y/o visualización de los datos filtrados.

6.1.1. Modalidad de imágenes faciales

En la siguiente imagen se pueden observar las gráficas de las técnicas utilizadas para captar la calidad de las imágenes faciales de la base de datos MELD. Cada uno de los valores límites utilizados se pueden ver en la Tabla 4.2. En las Figuras 6.1 y 6.2 se muestran las gráficas por cada técnica de detección de calidad, cada uno de los puntos corresponde a una imagen en particular, las que poseen el color azul son las que superaron los filtros de imagen y fueron marcadas como utilizables, las que tienen baja calidad se encuentran con color rojo. Cada imagen posee un valor distinto por técnica, esto se debe a que cada imagen es única y tiene un valor diferente a las demás debido a que poseen diferentes características, como el brillo, contraste, entre otras.

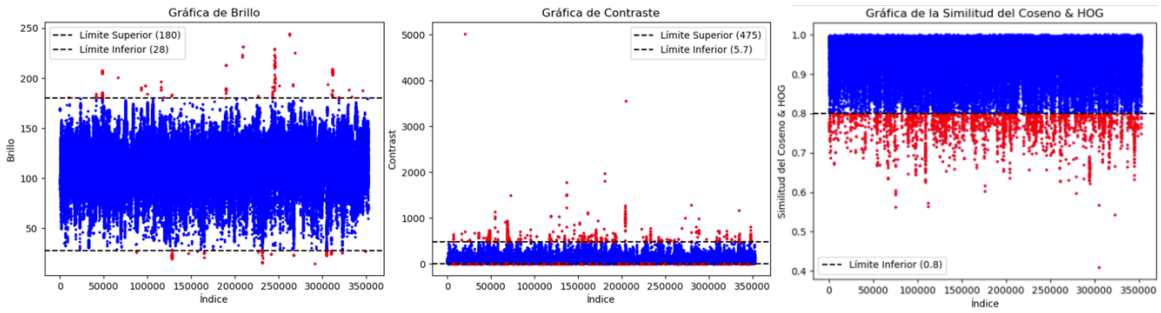


Figura 6.1: Técnicas de detección de calidades de imágenes faciales en MELD, específicamente para la detección de brillo, contraste y la similitud del coseno utilizando HOG.

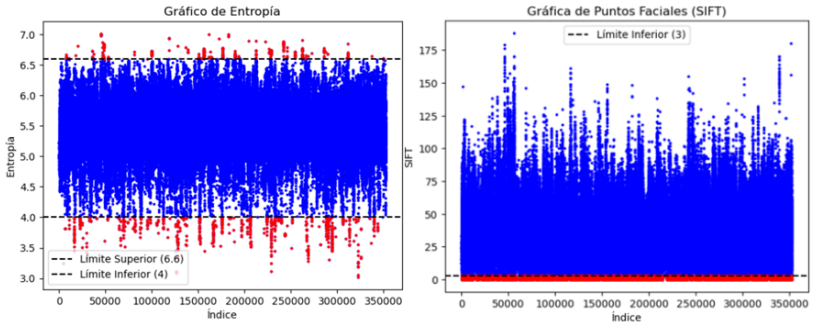


Figura 6.2: Técnicas de detección de calidades específicamente entropía y puntos faciales (SIFT), utilizadas para el filtrado de las calidades de imágenes faciales en MELD.

6.1.2. Modalidad de audio

En la Figura 6.3 se muestran las gráficas por cada técnica de detección de calidad, cada uno de los puntos corresponde a un audio en particular, los que poseen el color azul son los que superaron los filtros de audio y fueron marcados como utilizables, los que tienen baja calidad se encuentran con color rojo. Cada uno de los valores límites se pueden ver en la Tabla 4.3. Cada audio posee un valor diferente por técnica, esto se debe a que cada audio es único y tiene un valor distinto a los demás debido a que poseen diferentes características, como la distorsión armónica total, promedio de nivel de potencia, entre otros.

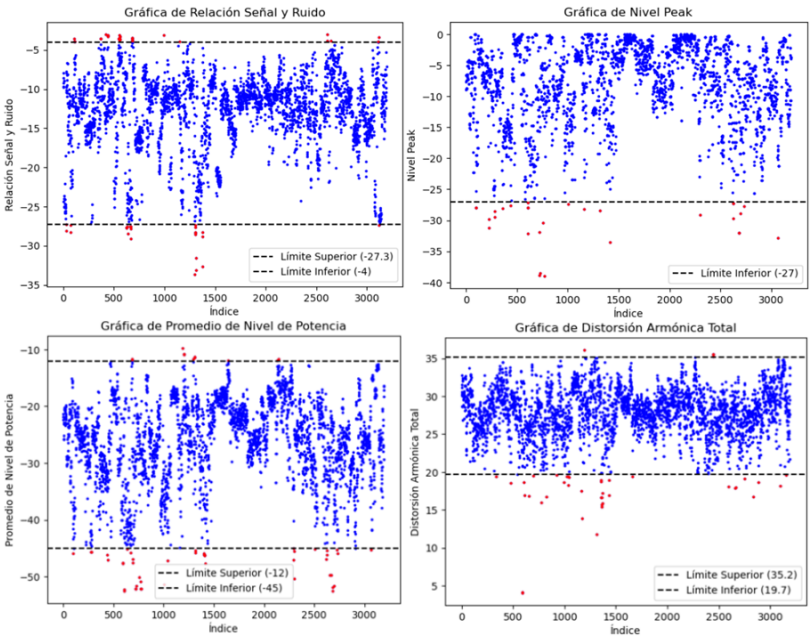


Figura 6.3: Técnicas de detección de calidades en audios de AffWild2.

6.1.3. Modalidad de texto

La Figura 6.4 muestra la gráfica del filtrado realizado mediante el filtro semántico, cada punto en la gráfica representa a una frase transcrita de un video, los de color azul son aquellos que superaron el filtro semántico y los de color rojo poseen baja calidad. En la Figura 6.5 muestra el archivo *CSV* del conjunto de datos AFEW presentado como Dataframe, en el cual se pueden observar los valores del filtro semántico (Semantic_Filter), filtro de texto (Text_Filter) y el dataset de origen (Source_DB), junto con las fases del preprocesamiento de texto. Cada texto posee un valor distinto en el filtro semántico, esto se debe a que cada texto es único y tiene un valor diferente a las demás debido a que poseen diferentes valor en el filtro de semántica.

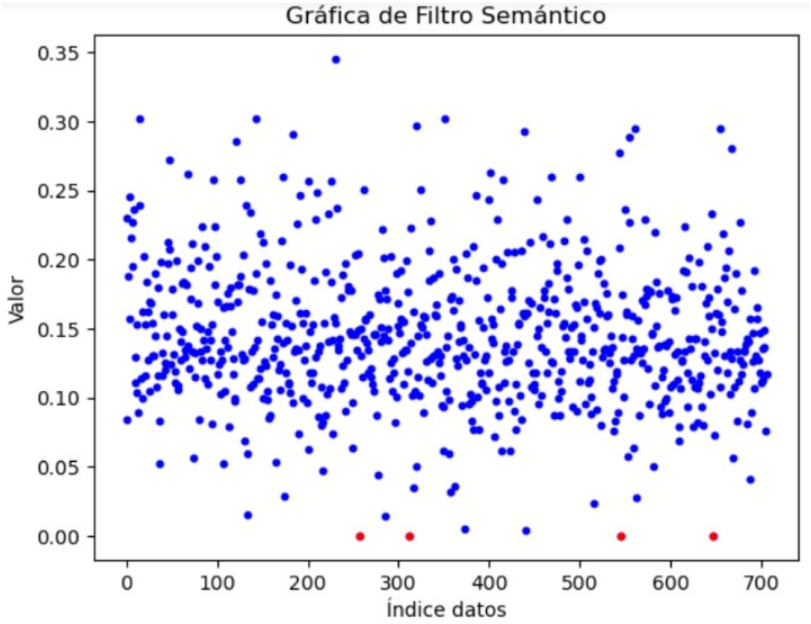


Figura 6.4: Técnica de detección de calidad semántica en el conjunto de datos AFEW en la modalidad de texto.

Transcription	Lowercase_Transcription	Text_Without_Stopwords	Tokens	Semantic_Filter	Text_Filter	Source_DB	Set	Lemmatized_Tokens	Emotion
NYM	nym	nym	['nym']	0.0	0	AFEW	train	['nym']	Surprise
THERE'S WHAT I WROTE	there is what i wrote	wrote	['wrote']	0.0	0	AFEW	train	['write']	Angry
A LITTLE LONGER BUT A LITTLE LONGER	a little longer but a little longer	little longer little longer	['little', 'longer', 'little', 'longer']	0.0	0	AFEW	train	['little', 'long', 'little', 'long']	Happy
NYM	nym	nym	['nym']	0.0	0	AFEW	train	['nym']	Fear

Figura 6.5: Fracción de archivo *CSV* de entrenamiento que muestra los valores de baja calidad semántica del conjunto de datos AFEW.

6.2. Resultados caso de estudio

Tal como se mencionó en la Sección 4.6.1 el caso de estudio es el reentrenamiento del modelo mediante las fases de entrenamiento y validación utilizando la base de datos MERDWild, para luego evaluar su desempeño en el dataset MAFW. Esto, con el objetivo de evaluar la predicción de las emociones para datos con los que no fue entrenado el modelo. El objetivo es comparar el desempaño del modelo con los resultados obtenidos en

el proyecto [1]. Dichos resultados serán evaluados mediante las métricas y gráficas descritas en la Sección 4.7. Para la evaluación del modelo generado se usarán las diferentes combinaciones de modalidades presentes.

Una vez ejecutados los modelos y evaluado su desempeño se obtuvo los siguientes resultados evaluados en MERDWild y MAFW:

En la Figura 6.6 (a) se encuentran las curvas ROC de emociones para la modalidad facial en el conjunto de datos MERDWild. La emoción con mejor desempeño fue el disgusto (Disgust) con un AUC de 0,65, mientras que la más baja fue la sorpresa (Surprise) con un AUC de 0,49. Además en la Tabla 6.1 se encuentra el reporte de clasificación en el cual se destacan medidas como Accuracy, Recall, F1-Score y Support, en donde destaca la emoción Neutral con mejor puntaje en todas las métricas, y Disgust fue la emoción con peores resultados, lo cual coincide con ser la emoción con menor cantidad de ejemplares.

En la Figura 6.6 (b) se muestra la curva ROC para la modalidad de audio. La emoción enojo (Angry) alcanzó un AUC de 0,71, mientras que la felicidad (Happy) obtuvo la puntuación más baja con un AUC de 0,51. En la Tabla 6.2 se observan los resultados de las predicciones de las emociones por métrica, donde la emoción Miedo (fear) es la que tiene el peor resultado con cero en todas las métricas y la emoción neutral obtuvo los más altos resultados en este reporte de clasificación.

En la Figura 6.6 (c) se muestra la curva ROC para la modalidad de texto. La felicidad (Happy) alcanzó un AUC de 0,67, mientras que el disgusto (Disgust) obtuvo la puntuación más baja con un AUC de 0,54. Además se encuentra la Tabla 6.3, en la cual se muestran los resultados de las predicciones de las emociones por métrica, donde se puede observar que la emoción Miedo (Fear) es la que tiene peores resultados con el valor cero en todas las métricas, excepto en Support el cual cuenta la cantidad de ejemplares por emoción.

En la Figura 6.7 se encuentra la matriz de confusión de MERDWild correspondiente a la fusión de las modalidades de audio, texto e imágenes faciales por medio de Embracenet+. En dicha imagen se puede observar que no existen aciertos en las emociones Disgust, Fear ni Surprise, la emoción que más fue correctamente reconocida fue la Neutral. Por otro lado, en la Figura 6.8, la cual corresponde a MAFW, suceden las mismas situaciones que en matriz correspondiente a MERDWild pero con distintas frecuencias de datos.

En la Tabla 6.4 se tiene un resumen y comparación de los resultados obtenidos en este proyecto para MERDWild y MAFW, comparados con el proyecto base [1], donde se puede observar que MERDWild posee un mejor accuracy en todas las modalidades en comparación con AFEW. También se nota que la modalidad de audio de MERDWild es la que obtuvo la mejor puntuación de todos los datasets en la modalidad de audio. En MAFW el accuracy en la modalidad de audio es la más baja, esto genera una discordancia, pero a la vez puede deberse a que MAFW posee diversos idiomas, lo cual afecta directamente en las predicciones de audio.

Emotion	**Accuracy**	**Recall**	**F1-Score**	**Support**
Neutral	0.536	0.297	0.382	102519
Angry	0.154	0.160	0.157	17676
Disgust	0.094	0.125	0.107	4771
Fear	0.072	0.254	0.113	12552
Happy	0.246	0.403	0.305	31058
Sad	0.092	0.035	0.051	18907
Surprise	0.047	0.059	0.052	15755
accuracy			0.252	203238
macro avg	0.177	0.190	0.167	203238
weighted avg	0.340	0.252	0.271	203238

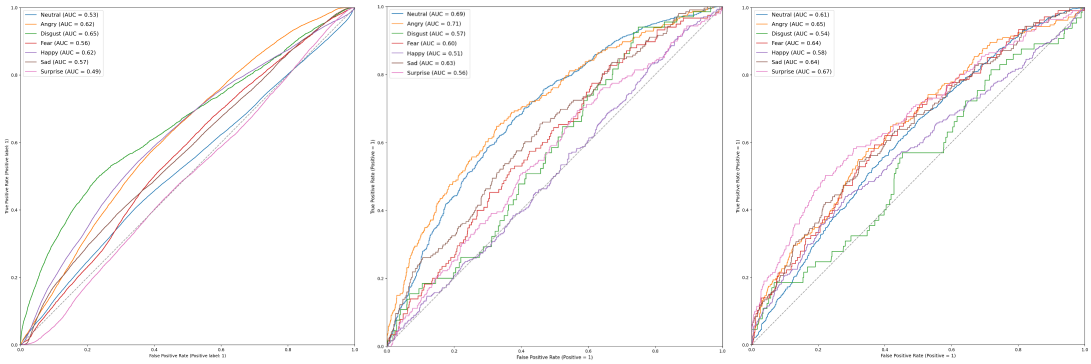
Tabla 6.1: Reporte de clasificación de imágenes faciales de MERDWild.

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.555	0.911	0.690	1152
Angry	0.366	0.118	0.178	221
Disgust	0.333	0.015	0.029	65
Fear	0.000	0.000	0.000	115
Happy	0.245	0.071	0.110	354
Sad	0.211	0.138	0.167	188
Surprise	0.124	0.067	0.087	225
accuracy			0.493	0.493
macro avg	0.262	0.189	0.180	2320
weighted avg	0.386	0.493	0.399	2320

Tabla 6.2: Reporte de clasificación de audio de MERDWild.

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.532	0.920	0.674	1119
Angry	0.256	0.047	0.079	213
Disgust	0.500	0.046	0.085	65
Fear	0.000	0.000	0.000	108
Happy	0.308	0.106	0.158	339
Sad	0.375	0.083	0.136	180
Surprise	0.333	0.144	0.201	201
accuracy			0.505	2225
macro avg	0.329	0.192	0.191	2225
weighted avg	0.414	0.505	0.402	2225

Tabla 6.3: Reporte de clasificación de texto de MERDWild.



(a) Curva ROC para imágenes. (b) Curva ROC para audio. (c) Curva ROC para texto.

Figura 6.6: Curvas ROC por modalidad de imágenes faciales, audio y texto.

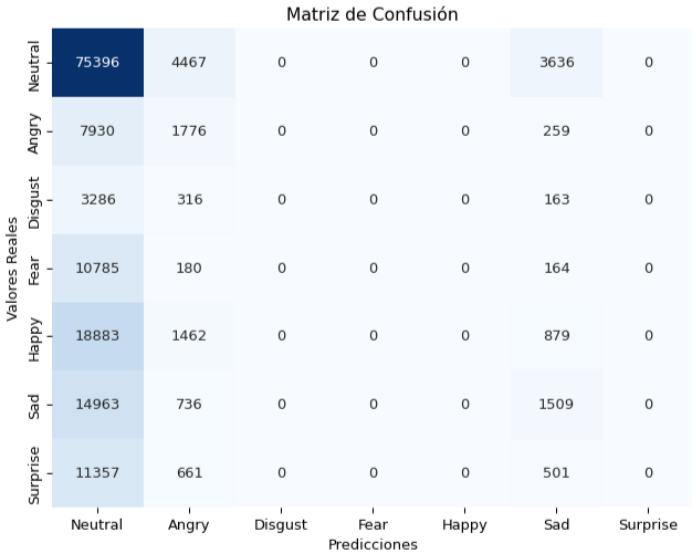


Figura 6.7: Matriz de confusión para la fusión de las tres modalidades en MERDWild.

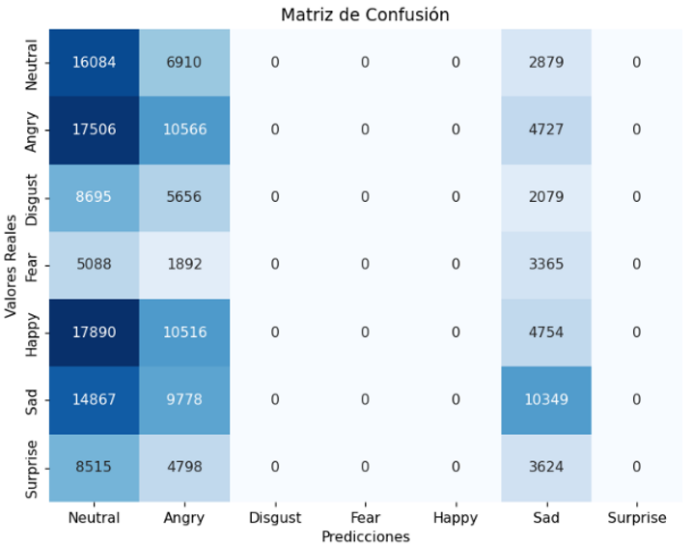


Figura 6.8: Matriz de confusión para la fusión de las tres modalidades en MAFW.

Datasets evaluados							
Modalidades			**Proyecto [1]**			**Este proyecto**	
						Train and Validation	Test
Imágenes	**Audio**	**Texto**	**AFEW**	**MELD**	**AffWild2**	**MERDWild**	**MAFW**
x			19,58 %	45,63 %	62,01 %	**36.16 %**	**22.01 %**
	x		16,17 %	45,63 %	40,52 %	**52.42 %**	**19.52 %**
		x	14,82 %	46,60 %	61,05 %	**43.48 %**	**20.50 %**
x	**x**		18,06 %	45,63 %	48,91 %	**38.58 %**	**25.31 %**
	x	**x**	15,90 %	20,39 %	52,52 %	**47.61 %**	**20.86 %**
x	**x**	**x**	18,87 %	45,63 %	58,94 %	**49.39 %**	**21.70 %**

Tabla 6.4: Resultados de MERDWild comparados con el proyecto [1] mediante la métrica accuracy.

Capítulo 7

Análisis y discusión de resultados

En la presente sección se abordará un análisis de los resultados obtenidos del entrenamiento y pruebas de MERDWild, con el objetivo de proponer mejoras tanto al dataset como a la arquitectura, discutiendo y analizando los resultados obtenidos, para luego proponer mejoras y responder a las preguntas de investigación expuestas en la Sección 4.6. En el siguiente listado se encuentran las preguntas de investigación:

1. *¿Cuál es la emoción que se encuentra con mayor frecuencia en la totalidad de modalidades?*
2. *¿Es posible realizar mejoras en los conjuntos de datos y en la arquitectura utilizada?*
3. Hipótesis: *¿Es la unificación de los datasets suficiente para obtener mejores resultados en MER in-the-wild?*

7.1. Análisis de resultados

En la presente sección se abordará el análisis de los resultados obtenidos para cada una de las preguntas de investigación.

Para lograr responder a la primera pregunta es necesario haber realizado la unificación de los datos para obtener de manera final cual es la emoción que presenta una mayor frecuencia en la totalidad de modalidades. Este proceso implica el preprocesamiento de los datos para obtener los archivos alineados por uterancia. Una vez alineados los archivos se puede generar una visualización de los resultados.

Para responder la segunda y tercera pregunta de investigación es necesario haber obtenido los resultados las pruebas en el conjunto MAFW y MERDWild, para luego sacar conclusiones y dar respuesta a las preguntas de investigación.

Para analizar los resultados de una manera objetiva es necesario observar las curvas de pérdida (loss) frente a epochs (épocas), y las curvas epoch frente a accuracy. En los siguiente párrafos se analizarán estas gráficas para cada una de las tres modalidades (imagen, audio y texto) para descifrar que es lo que está sucediendo con las predicciones.

En la Figura 7.1 se muestra la curva correspondiente a loss frente a epochs, en la Figura 7.2 se encuentra la gráfica de la relación entre accuracy y las epochs correspondientes a la modalidad de imágenes faciales. En ambas imágenes se muestra una relación inversa entre las curvas de entrenamiento y validación.

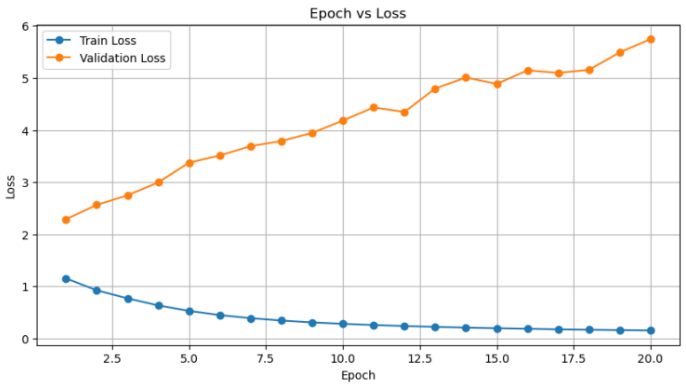


Figura 7.1: Curva de epochs frente a loss para la modalidad de imágenes.

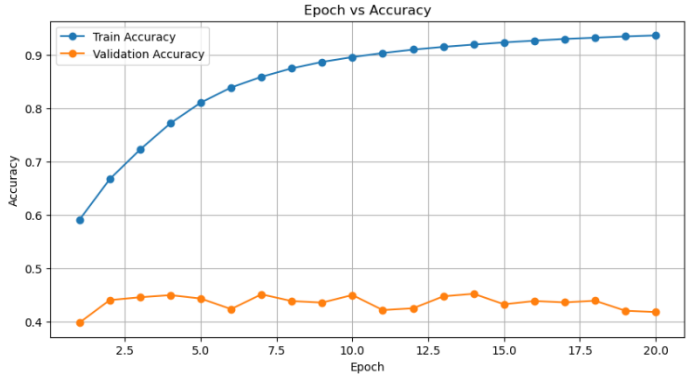


Figura 7.2: Curva de epochs frente a accuracy para la modalidad de imágenes.

En las Figuras 7.3 y 7.4 se observan las curvas de la relación entre epochs y accuracy, y epochs frente a loss para la modalidad de audio. En la primera imagen se muestra que la pérdida en el entrenamiento al transcurrir las épocas se encuentra oscilando constantemente entre valores altos de pérdida y bajos, esto se puede deber a una inestabilidad de convergencia del modelo causada por un learning rate inadecuada (0,0001) o una alta complejidad del modelo de DL utilizada (ResNet50). Además, en la segunda imagen se

muestra un decrecimiento en el accuracy de la validación, el cual inicia desde la época 14 en adelante.

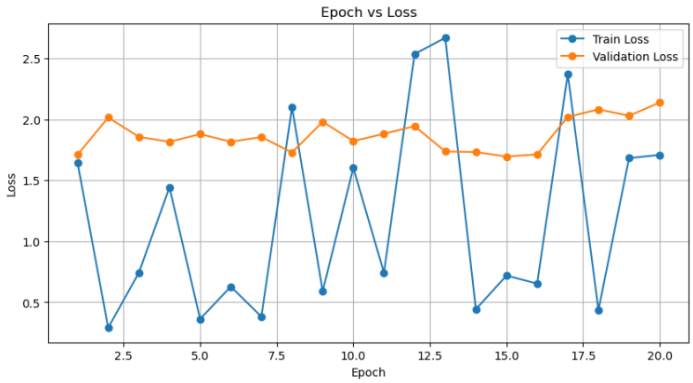


Figura 7.3: Curva de epochs frente a loss para la modalidad de audio.

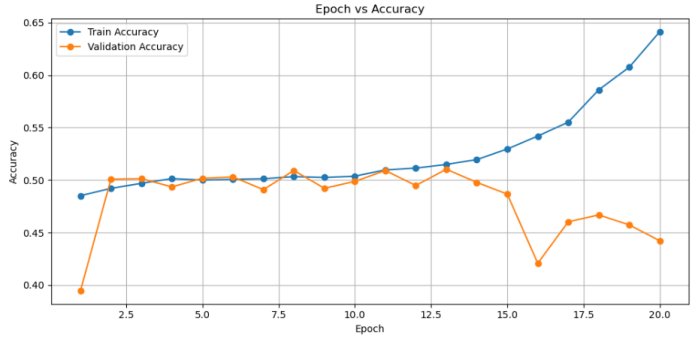


Figura 7.4: Curva de epochs frente a accuracy para la modalidad de audio.

Para la modalidad de texto, en las Figuras 7.5 y 7.6, se tienen las gráficas correspondientes a epochs frente a loss y epochs frente a accuracy respectivamente. En la Figura 7.5 se muestra que la pérdida va disminuyendo con el pasar de las épocas, esto es un buen indicador, pero en la segunda gráfica se encuentra un incremento de accuracy desde la primera época a la sexta época, luego comienza a decrecer el valor de accuracy. Las curvas de entrenamiento, tanto de accuracy como loss no se encuentran en las imágenes debido a que el modelo llamado "MLPClassifier" utilizado tanto en el proyecto [1] como en el presente trabajo, no proporciona directamente una función para calcular la pérdida y el accuracy en el conjunto de entrenamiento. En las gráficas correspondientes a la modalidad de texto solamente se tienen resultados hasta la época 16 debido a que MLPClassifier tiene early stopping implementado.

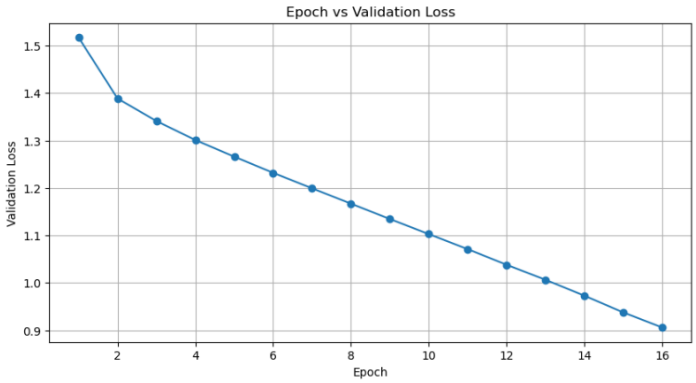


Figura 7.5: Curva de epochs frente a loss para la modalidad de texto.

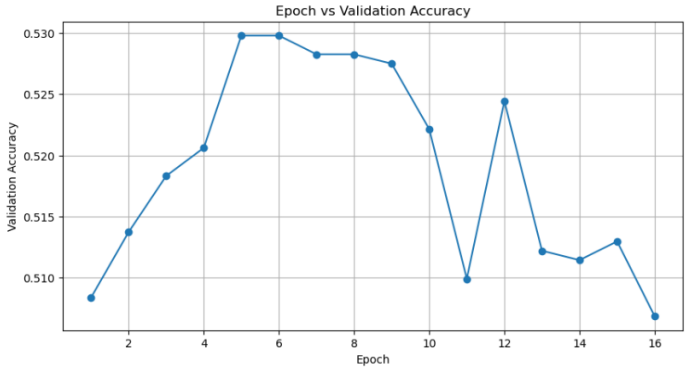


Figura 7.6: Curva de epochs frente a accuracy para la modalidad de texto.

7.2. Discusión de resultados

En esta sección se presentarán los resultados obtenidos desde el análisis de cada una de las preguntas de investigación planteadas en este proyecto.

1. La respuesta a la primera pregunta de investigación se responde con la visualización de la Figura 7.7, donde se puede observar que la emoción que se encuentra con mayor frecuencia en los tres conjuntos de datos utilizados para conformar la base de datos MERDWild es la "Neutral".

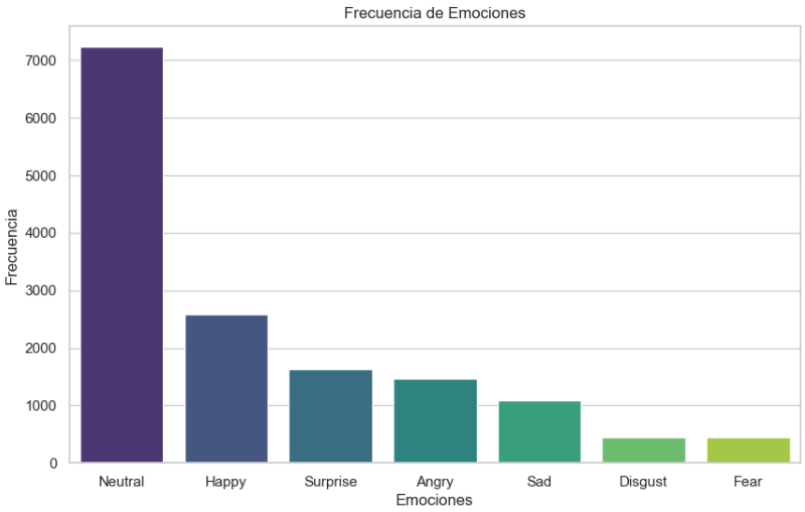


Figura 7.7: Emociones unificadas de MERDWild.

2. La respuesta a la segunda pregunta de investigación es: si, una de las posibles mejoras que puede resultar útil para aumentar la correcta detección multimodal de emociones *in-the-wild* es la etiquetación de las emociones por modalidad de manera individual. Otra posible mejora es la detección del hablante para los datos provenientes de los conjuntos de datos AffWild2 y AFEW, debido a que existen diferentes ejemplos en los que la persona que está siendo enfocada en cámara no es la que está hablando y genera una discordancia de emociones y aumentando la confusión del modelo. Otro de los aspectos que se puede mejorar en esta base de datos integrada es la eliminación del ruido de fondo y música en los audios, lo cual beneficiaría al aprendizaje de los datos de audio, como también mejoraría la transcripción automática del texto. Además se puede seguir investigando sobre técnicas de eliminación del fondo de la imagen facial, para causar el mismo efecto descrito anteriormente. También se pueden cambiar los valores límites de las variables de calidad de imágenes faciales, junto con la utilización de otra fase del preprocesamiento de texto. Otra de las posibles mejoras es cambiar los hiper-parámetros de los modelos (epochs, batch size, learning rate, cantidad de capas, entre otros), como también modificar la arquitectura utilizada para así buscar mejorar MER *in-the-wild*.

3. La respuesta a la hipótesis de este Proyecto de Título es que la unificación de los datasets no es suficiente para obtener mejores métricas en MER *in-the-wild* con exactamente la misma arquitectura que fue utilizada en el proyecto [1].

   En el caso de las imágenes fáciles, en las Figuras 7.1 y 7.2 se muestra una relación inversa entre las curvas de entrenamiento y validación, lo que indica un sobre-entrenamiento (overfitting) del modelo de imágenes faciales. Para solucionar este

problema, se puede utilizar la reducción de la complejidad del modelo, reduciendo el número de capas. Además, el uso de dropout y la modificación de los hiper-parámetros pueden ser estrategias efectivas [95].

En cuanto a la modalidad de audio, representada en la Figura 7.3, se observa la curva de epoch frente a loss. Para abordar el fenómeno que se muestra en la gráfica, se sugiere probar diferentes tasas de aprendizaje, estrategias de optimización, simplificar la arquitectura del modelo, aplicar técnicas de regularización como dropout o implementar la normalización de lotes (batch normalization). En la Figura 7.4, se muestra un overfitting de la desde la época 14 en adelante, para abordar este problema es necesario utilizar una técnica de early stopping para mitigar este sobreentrenamiento.

En el caso de la modalidad de texto, en la Figura 7.3, se muestra que la pérdida disminuye con el pasar de las épocas, lo cual es un buen indicador. Sin embargo, en la Figura 7.6 gráfica se evidencia un aumento de accuracy desde la primera época hasta la sexta, seguido de un decrecimiento. Esto sugiere que el modelo se está sobreentrenando a partir de la época 6 en adelante. Para solucionar este problema, se pueden aplicar estrategias similares a las propuestas para las imágenes faciales, como la reducción de complejidad, el uso de dropout o ajustes en los hiper-parámetros.

A pesar de los desafíos de overfitting en las diferentes modalidades, la modalidad de audio muestra un mejor accuracy, con un valor igual a 52,42 %, superando los resultados de los tres conjuntos de datos utilizados en el proyecto original [1], como se detalla en la Tabla 6.4.

Capítulo 8

Conclusión

En el presente Trabajo de Título, se abordó el reconocimiento multimodal de emociones en entornos no controlados, destacando un enfoque especial en el preprocesamiento de los datasets. Este preprocesamiento fue crucial para la unificación de los conjuntos de datos y la creación de una nueva base de datos integrada llamada "MERDWild", la cual posee datos para las modalidades de texto, audio e imágenes faciales. Para ello, se llevó a cabo la extracción de las modalidades anteriormente mencionadas para reconocer la emoción de los archivos que ingresan a los modelos. Se utilizó una arquitectura con técnicas como VGG19, ResNet50 y DialogXL para las modalidades imagen, audio y texto respectivamente. De esta manera, mediante el método de fusión Embracenet+ obtener el resultado final de la emoción del vídeo ingresado.

En este proyecto, uno de los objetivos iniciales ha sido construir un set de datos unificado que incorpore datos de buena calidad. Para lograrlo, se realizó un preprocesamiento de los datasets AFEW, MELD y AffWild2 definiendo los valores límites de exclusión de datos cuidadosamente, aplicando rigurosos criterios de calidad obtenidos de diversas fuentes. Se llevó a cabo un proceso exhaustivo de limpieza, normalización y estandarización de los datos, asegurando su calidad para su uso en el desarrollo de modelos de DL. Logrando entrenar la arquitectura propuesta, junto con visualizar y analizar los resultados obtenidos. Dentro de los resultados obtenidos en el entrenamiento, validación y pruebas se llegó a la conclusión de volver a entrenar los modelos utilizando otra arquitectura para evitar el sobre-entrenamiento que se presentó.

Dentro de las dificultades que se presentaron en el transcurso del proyecto de título se encuentra la imposibilidad de utilización de uno de los conjuntos de datos que se planteó inicialmente, llamado CMU-MOSEI, este conjunto de datos que ya no se encuentra en el proyecto, debido a que el formato en el que se encontraban sus datos era diferente al de los demás datasets, por lo tanto, no permitía realizar la unificación. Por esto, ingresó al proyecto el dataset MAFW, para agregar valor y enriquecer la base de datos unificada. Finalmente se utilizó MAFW como conjunto de pruebas debido a que este dataset posee

diferentes idiomas y su utilización en el entrenamiento sumaría confusión al modelo de audio y texto transcrito. Otra dificultad fue realizar el preprocesamiento de cada una de las modalidades en cada uno de los datasets, esto implicó una extensa investigación, la realización de múltiples pruebas visuales y audibles para poder filtrar de la mejor manera los datos y obtener un set de datos de buena calidad, esto junto con la estandarización de las etiquetas de las emociones. Por último se encontró la dificultad de realizar el entrenamiento y validación del modelo para obtener resultados que logren responder la hipótesis de este proyecto.

Una de las limitaciones de este proyecto de título es la etiquetación de los datos en los tiempos establecidos, debido a que cada uno de los datasets posee una modalidad preferida para el cual fue creado. Para poder etiquetar cada uno de los tres datasets para aquellas modalidades para las que no fue diseñado se debe realizar crowdsourcing, lo cual requiere la creación de una aplicación web para que las personas voluntariamente etiqueten los datos, ese proceso implica extensos tiempos de etiquetación, por lo tanto, se acordó utilizar la misma etiqueta de emoción para todas las modalidades, y de esta manera, evitar el uso de crowdsourcing debido a los limitados tiempos de realización de este proyecto de título. La última limitación que se presenta es el no uso de aquellos datos que no pertenecen a las siete emociones que se utilizan en este proyecto, como lo es la emoción "Otra" en AffWild2 y "contempt", "dissapointment", "helpplessness" y "anxiety" en MAFW.

Como parte de las perspectivas futuras, se sugiere llevar a cabo la etiquetación de diversas modalidades de datos en la base de datos MERDWild. Para los datos provenientes del conjunto de datos MELD, es necesario etiquetar la modalidad de imágenes faciales. Además, se recomienda realizar la etiquetación de datos de audio y texto en los datasets AFEW y AffWild2. El propósito de esta acción es reducir la confusión en los modelos de predicción. Para lograr una mejora en los resultados de MER *in-the-wild*, se recomienda adoptar un enfoque sistemático. Inicialmente, se sugiere ajustar los hiper-parámetros del modelo. Posteriormente, se deben explorar técnicas como batch normalization y dropout para mitigar posibles problemas de sobreentrenamiento. En caso de que estas estrategias no generen mejoras, se propone la posibilidad de modificar la arquitectura de la red de aprendizaje profundo.

Bibliografía

[1] Ana Aguilera, Diego Mellado, and Felipe Rojas. An Assessment of In-the-Wild Datasets for Multimodal Emotion Recognition. *Sensors*, 23(11):5184, May 2023.

[2] Rafael Bisquerra Alzina. Educación emocional y competencias básicas para la vida. *Revista de Investigación Educativa*, 21(1):7–43, ene. 2003.

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

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[5] Alfredo Miguel Aguado, Lourdes Nevares Heredia, et al. La comunicación no verbal. *Tabanque: revista pedagógica*, 1(10):141–154, 1995.

[6] Luis Alberto Pérez Gaspar et al. Sistema de reconocimiento multimodal de emociones para interacción humano-robot. *REPOSITORIO NACIONAL CONACYT*, 2015.

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimedia*, 19(3):34–41, May 2012.

[8] Dimitrios Kollias and Stefanos Zafeiriou. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *arXiv*, November 2018.

[9] 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*, October 2018.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Team. Redes neuronales residuales - Lo que necesitas saber (ResNet) — DATA SCIENCE. *Data Sci.*, December 2020.

[12] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797, 2021.

[13] Juanpablo Andrew Heredia Parillo. A multi-modal emotion recogniser based on the integration of multiple fusion methods, 2021.

[14] What is Deep Learning? | IBM, May 2023. [Online; accessed 1. May 2023].

[15] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

[16] Patel Dhruv and Subham Naskar. Image Classification Using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A Review. In *Machine Learning and Information Processing*, pages 367–381. Springer, Singapore, March 2020.

[17] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Comput. Methods Programs Biomed.*, 213:106504, January 2022.

[18] CERN: Compact facial expression recognition net | Elsevier Enhanced Reader, May 2023. [Online; accessed 1. May 2023].

[19] Adrian Saez De La Pascua. *Deep learning para el reconocimiento facial de emociones basicas*. PhD thesis, Universitat Politècnica de Catalunya, January 2019.

[20] Shao-Yen Tseng, Shrikanth Narayanan, and Panayiotis Georgiou. Multimodal Embeddings From Language Models for Emotion Recognition in the Wild. *IEEE Signal Process. Lett.*, 28:608–612, March 2021.

[21] Puneet Kumar and Balasubramanian Raman. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks*, 150:392–407, June 2022.

[22] Nicolás Eduardo Grágeda Ushak. Reconocimiento de emociones utilizando la voz en ambientes dinámicos de interacción humano-robot. *Repositorio Académico - Universidad de Chile*, 2023.

[23] Tecnología de reconocimiento emocional: una nueva frontera en la interacción humano-computadora, August 2023. [Online; accessed 5. Nov. 2023].

[24] Konstantina Vemou and Anna Horvath. Techdispatch# 1/2021-facial emotion recognition, 2021. [Online; accessed 4. Nov. 2023].

[25] Corporate-body. Edps:european Data Protection Supervisor. EDPS TechDispatch : facial emotion recognition. Issue 1, 2021. *Publications Office of the European Union*, May 2021.

[26] Nicolás Mastropasqua and Daniel Acevedo. Reconocimiento de expresiones faciales con redes profundas livianas usando Label Distribution Learning y el espacio de Action Units. *1.*, 8(10):23–28, December 2022.

[27] Beibin Li, Sachin Mehta, Deepali Aneja, Claire Foster, and Linda Shapiro. A Facial Affect Analysis System for Autism Spectrum Disorder. *ResearchGate*, pages 4549–4553, September 2019.

[28] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowl. Inf. Syst.*, 62(8):2937–2987, August 2020.

[29] Jitendra Kumar Rout, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena, and Karen L Williams. A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 18:181–199, 2018.

[30] Mounika Karna, D. Sujitha Juliet, and R. Catherine Joy. Deep learning based Text Emotion Recognition for Chatbot applications. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 988–993. IEEE, June 2020.

[31] Victor Emilio Hernández Leal. Emociones en señales de voz: reconocimiento con redes neuronales profundas. B.S. thesis, Universitat Politècnica de Catalunya, 2021.

[32] ProjectPro. Speech Emotion Recognition Project using Machine Learning. *ProjectPro*, April 2023.

[33] Nagesh Singh Chauhan. Métricas De Evaluación De Modelos En El Aprendizaje Automático. *DataSource*, September 2020.

[34] Precision, Recall, F1, Accuracy en clasificación - IArtificial.net, October 2020. [Online; accessed 2. May 2023].

[35] Precision, Recall, F1, Accuracy en clasificación - IArtificial.net, October 2020. [Online; accessed 29. May 2023].

[36] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022.

[37] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vision Comput.*, 65:66–75, September 2017.

[38] Pablo Barros and Stefan Wermter. Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior*, October 2016.

[39] Juanpablo Heredia, Edmundo Lopes-Silva, Yudith Cardinale, Jose Diaz-Amado, Irvin Dongo, Wilfredo Graterol, and Ana Aguilera. Adaptive Multimodal Emotion Detection Architecture for Social Robots. *IEEE Access*, 10:20727–20744, February 2022.

[40] Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, and Wolfgang Minker. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. *Multimodal Technologies and Interaction*, 6(2):11, 2022.

[41] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. *arXiv*, June 2022.

[42] Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son N. Tran, Thin Nguyen, Sridha Sridharan, and Clinton Fookes. Deep Auto-Encoders With Sequential Learning for Multimodal Dimensional Emotion Recognition. *IEEE Trans. Multimedia*, 24:1313–1324, March 2021.

[43] Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 30(5):975–985, July 2019.

[44] Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. HEU Emotion: A Large-scale Database for Multi-modal Emotion Recognition in the Wild. *arXiv*, July 2020.

[45] Caterina Elionor Muntaner González. Reconocimiento automático de emociones en condiciones reales a partir de imágenes y audio. 2021.

[46] Jingjun Liang, Ruichen Li, and Qin Jin. Semi-supervised Multi-modal Emotion Recognition with Cross-Modal Distribution Matching. *arXiv*, September 2020.

[47] Panagiotis Antoniadis, Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3645–3651, 2021.

[48] Luisa A SIMó. Emociones del consumidor: componentes y consecuencias de marketing. *Estudios sobre consumo*, 64:9–26, 2003.

[49] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *arXiv*, October 2016.

[50] Ghazala Bilquise, Samar Ibrahim, Khaled Shaalan, et al. Emotionally intelligent chatbots: A systematic literature review. *Human Behavior and Emerging Technologies*, 2022.

[51] Guanglong Du, Shuaiying Long, and Hua Yuan. Non-Contact Emotion Recognition Combining Heart Rate and Facial Expression for Interactive Gaming Environments. *IEEE Access*, PP(99):1, January 2020.

[52] OpenCV 3 Object Detection : Face Detection using Haar Cascade Classfiers - 2020, June 2023. [Online; accessed 20. Jun. 2023].

[53] Fayaz Ali Dharejo, Yuanchun Zhou, Farah Deeba, and Yi Du. A Color Enhancement Scene Estimation Approach for Single Image Haze Removal. *IEEE Geoscience and Remote Sensing Letters*, 17(9):1613–1617, November 2019.

[54] Miguel Ángel Antúnez Galindo. *Algoritmos de detección de objetos para la detección y seguimiento de ojos*. PhD thesis, Universitat Politècnica de Catalunya, January 2019.

[55] Shuihua Wang, Xiaojun Yang, Yudong Zhang, Preetha Phillips, Jianfei Yang, and Ti-Fei Yuan. Identification of Green, Oolong and Black Teas in China via Wavelet Packet Entropy and Fuzzy Support Vector Machine. *Entropy*, 17(10):6663–6682, September 2015.

[56] Zahra Hossein-Nejad, Hamed Agahi, and Azar Mahmoodzadeh. Image matching based on the adaptive redundant keypoint elimination method in the SIFT algorithm. *Pattern Anal. Applic.*, 24(2):669–683, May 2021.

[57] Histogram of Oriented Gradients — skimage v0.20.0 docs, May 2023. [Online; accessed 25. May 2023].

[58] moviepy, June 2023. [Online; accessed 20. Jun. 2023].

[59] Jhi Hoon Joo, Parque Myung Chul, Dong Seog Han, and Veljko Pejovic. Deep Learning-Based Channel Prediction in Realistic Vehicular Communications. *IEEE Access*, 7:27846–27858, February 2019.

[60] Christopher Johann Clarke, Balamurali B T, and Jer-Ming Chen. Characterising nonlinear behaviour of coupling capacitors through audio feature analysis and machine learning. *Audio Engineering Society*, May 2021.

[61] Peng He, Yang Zhang, Xinyue Yang, Xiao Xiao, Haolin Wang, and Rongsheng Zhang. Deep Learning-Based Modulation Recognition for Low Signal-to-Noise Ratio Environments. *Electronics*, 11(23):4026, December 2022.

[62] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A General-Purpose Speech Toolkit. *arXiv*, June 2021.

[63] Jonatas Grosman. Fine-tuned XLS-R 1B model for speech recognition in English. `https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-english`, 2022. [Online; accessed 27. Jun. 2023].

[64] SpeechRecognition, February 2022. [Online; accessed 27. Jun. 2023].

[65] vitouphy/wav2vec2-xls-r-300m-english · Hugging Face, June 2023. [Online; accessed 27. Jun. 2023].

[66] José Camacho-Collados and Mohammad Taher Pilehvar. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *ArXiv*, abs/1707.01780, 2017.

[67] Saurav Pradha. Effective text data preprocessing technique for sentiment analysis in social media data. *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019.

[68] Shahab Raji and Gerard de Melo. What sparks joy: The affectvec emotion database. In *Proceedings of The Web Conference 2020*, WWW '20, page 2991–2997, New York, NY, USA, 2020. Association for Computing Machinery.

[69] Lukas Povoda, Radim Burget, Jan Masek, Vaclav Uher, and Malay Kishore Dutta. Optimization Methods in Emotion Recognition System. *Radioengineering*, 25(3):565–572, September 2016.

[70] Pamela Quinteros. *Modelo de credibilidad extendido sobre T-CREO*. PhD thesis, Universidad de Valparaíso, 2022. [Online; accessed 21. Aug. 2023].

[71] Generalidades, April 2007. [Online; accessed 10. Jun. 2023].

[72] Aguante de Potencia (potencia admisible) de cajas acústicas : RMS, pico y programa, PMPO, June 2023. [Online; accessed 10. Jun. 2023].

[73] Distorsión Armónica, August 2017. [Online; accessed 11. Jun. 2023].

[74] Relación señal/ruido: La guía definitiva, April 2023. [Online; accessed 10. Jun. 2023].

[75] Soad Almabdy and Lamiaa Elrefaei. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*, 9(20):4397, 2019.

[76] Matrices de confusión - EcuRed, May 2023. [Online; accessed 30. May 2023].

[77] Jason Brownlee. Overfitting and Underfitting With Machine Learning Algorithms - MachineLearningMastery.com. *MachineLearningMastery*, August 2019.

[78] Sitio big Data. Diagnosticar el rendimiento del modelo de aprendizaje automático - sitiobigdata.com. *Sitiobigdata*, September 2022.

[79] ¿Qué es Overfitting y Cómo lo Detectamos? - DataSmarts Español, July 2020. [Online; accessed 30. May 2023].

[80] Daniel Alexis Pérez-Aguilar, Redy Henry Risco-Ramos, and Luis Casaverde-Pacherrez. Transfer learning en la clasificación binaria de imágenes térmicas. *ings*, (26):71–86, June 2021.

[81] Jaime Cerda and Lorena Cifuentes. Uso de curvas roc en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, 29(2):138–141, 2012.

[82] Anaconda software distribution, 2020.

[83] Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019.

[84] El tutorial de Python, June 2023. [Online; accessed 26. Jun. 2023].

[85] Project Jupyter Documentation — Jupyter Documentation 4.1.1 alpha documentation, November 2023. [Online; accessed 12. Dec. 2023].

[86] Tokio School. ¿Qué es Scikit-Learn? Guía completa de esta librería de Python. *Tokio School*, April 2023.

[87] Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter,

Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad, Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef van der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, and Waldir Pimenta. librosa/librosa: 0.10.1, August 2023.

[88] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[89] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[90] Jonathan Quiza. *Deep Learning con Pytorch - Ciencia y Datos - Medium*. Ciencia y Datos, August 2018.

[91] Alfredo Sánchez Alberca. La librería Numpy | Aprende con Alf. *Aprende con Alf*, May 2022.

[92] alexandre. Pandas : La biblioteca de Python dedicada a la Data Science. *Formation Data Science | DataScientest*, December 2022.

[93] Sitio big Data. OpenCV Python: Face Detection Neural Network - sitiobigdata.com. *Sitiobigdata*, June 2019.

[94] Paseo por las Arquitecturas de Integración, June 2023. [Online; accessed 26. Jun. 2023].

[95] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, feb 2019.

Capítulo 9

Anexo

En el presente capítulo se muestran los anexos correspondientes a este trabajo de título, esto incluye la implantación del proyecto en dos repositorios de GitHub.

9.1. Implantación

Tal como se mencionó anteriormente, se crearon dos repositorios de GitHub para realizar la implantación.

El primer repositorio fue creado especialmente para mostrar guardar toda la estructura del preprocesamiento por modalidad y por dataset, la cual fue organizada mediante diferentes ficheros, los cuales poseen los scripts correspondientes a cada una de estas fases. Además, en otra carpeta se guardaron los tres códigos orientados al entrenamiento, validación y pruebas de los modelos generados. En la Figura 9.1 se muestra el repositorio llamado "MultimodalEmotionRecognitionInTheWild-Thesis", desde este repositorio, específicamente en la parte final, al hacer click en el vínculo llamado "aquí" y que se encuentra en color azul, redirige al repositorio de la base de datos MERDWild.

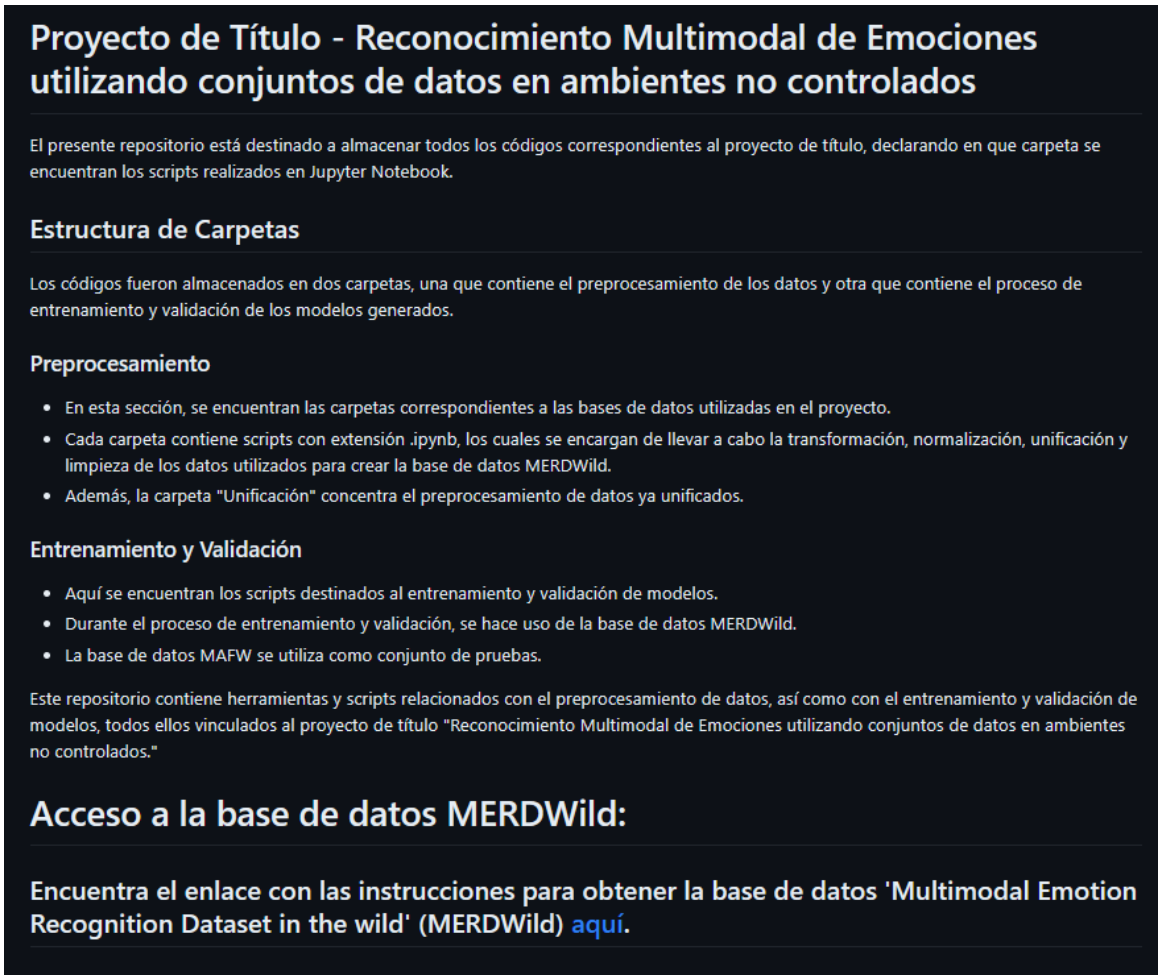
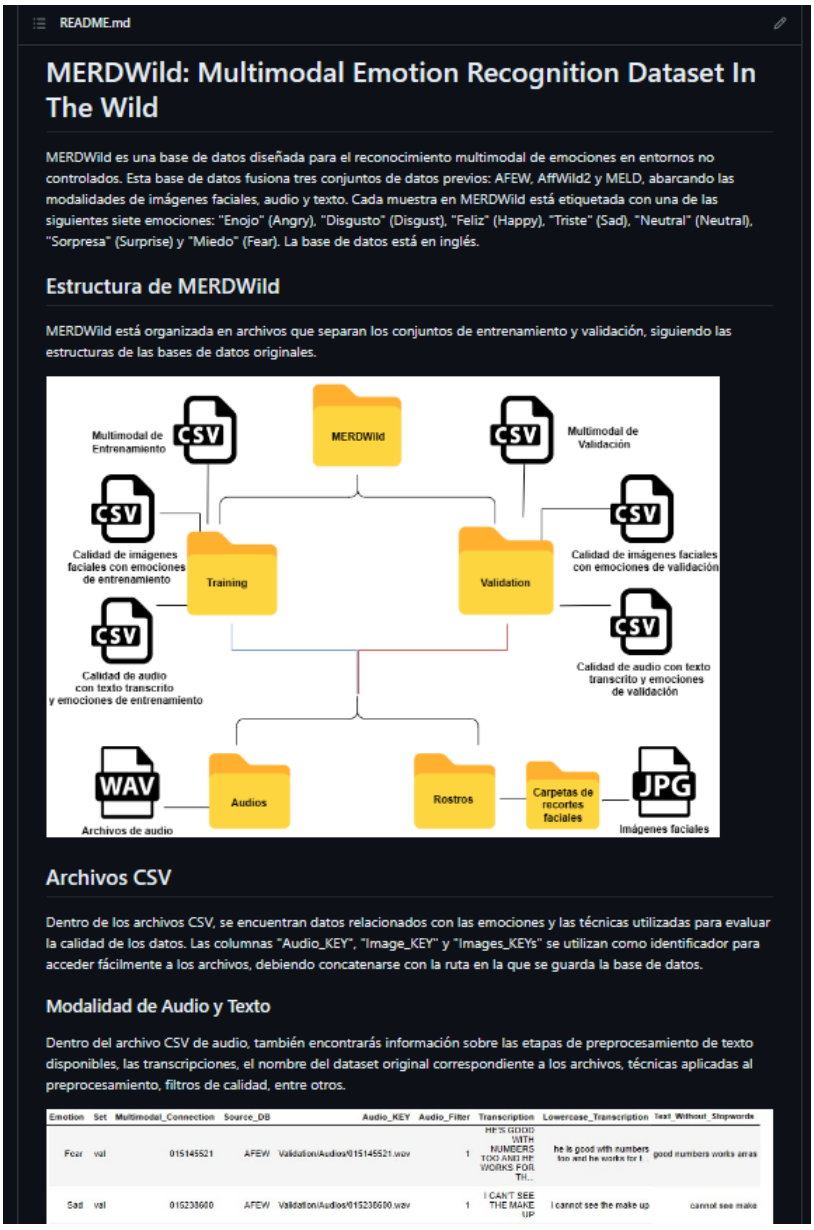


Figura 9.1: Vista principal del repositorio MultimodalEmotionRecognitionInTheWild-Thesis

En la Figura 9.2 se encuentra el esquema principal del repositorio llamado "MERD-Wild", el cual está destinado en la descripción general de la base de datos unificada y manera de obtener los datos es por medio de la solicitud al correo electrónico: ana.aguilera@uv.cl. Al recibir el correo electrónico se envía el enlace de la base de datos almacenada de manera privada en OneDrive. En la Figura 9.3 se observa el almacenamiento en la nube del conjunto de datos MERDWild.



README.md

MERDWild: Multimodal Emotion Recognition Dataset In The Wild

MERDWild es una base de datos diseñada para el reconocimiento multimodal de emociones en entornos no controlados. Esta base de datos fusiona tres conjuntos de datos previos: AFEW, AffWild2 y MELD, abarcando las modalidades de imágenes faciales, audio y texto. Cada muestra en MERDWild está etiquetada con una de las siguientes siete emociones: "Enojo" (Angry), "Disgusto" (Disgust), "Feliz" (Happy), "Triste" (Sad), "Neutral" (Neutral), "Sorpresa" (Surprise) y "Miedo" (Fear). La base de datos está en inglés.

Estructura de MERDWild

MERDWild está organizada en archivos que separan los conjuntos de entrenamiento y validación, siguiendo las estructuras de las bases de datos originales.

Archivos CSV

Dentro de los archivos CSV, se encuentran datos relacionados con las emociones y las técnicas utilizadas para evaluar la calidad de los datos. Las columnas "Audio_KEY", "Image_KEY" y "Images_KEYs" se utilizan como identificador para acceder fácilmente a los archivos, debiendo concatenarse con la ruta en la que se guarda la base de datos.

Modalidad de Audio y Texto

Dentro del archivo CSV de audio, también encontrarás información sobre las etapas de preprocesamiento de texto disponibles, las transcripciones, el nombre del dataset original correspondiente a los archivos, técnicas aplicadas al preprocesamiento, filtros de calidad, entre otros.

Emotion	Set	Multimodal_Connection	Source_DB	Audio_KEY	Audio_Filter	Transcription	Lowercase_Transcription	Text_Without_Stopwords
Fear	val	015145521	AFEW	Validation\Audios\015145521.wav	1	HE'S GOOD WITH NUMBERS TOO AND HE WORKS FOR TH...	he is good with numbers too and he works for t...	good numbers works arras
Sad	val	015238600	AFEW	Validation\Audios\015238600.wav	1	I CAN'T SEE THE MAKE UP	I cannot see the make up	cannot see make

Figura 9.2: Vista principal del repositorio MERDWild.

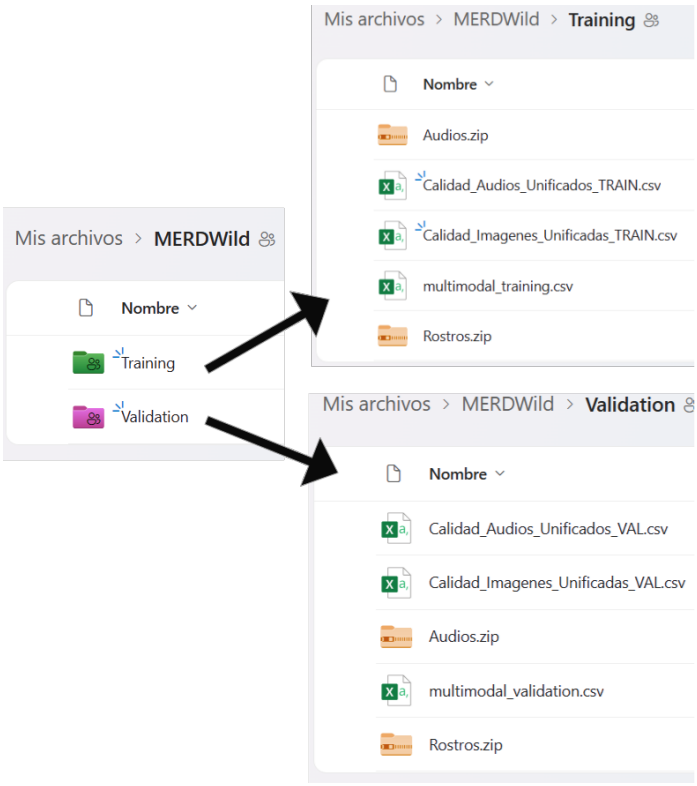


Figura 9.3: Almacenamiento en la nube de la base de datos MERDWild.