UNIVERSIDAD
ANDRES BELLO

UNIVERSIDAD ANDRÉS BELLO

FACULTAD DE INGENIERÍA

INGENIERÍA CIVIL INFORMÁTICA

Reconocimiento de emociones en la educación mediante Deep Learning

BRYAN FABIAN CABEZAS SANDOVAL

Profesores guía:
PABLO HERNÁN SCHWARZENBERG RIVEROS
BILLY MARCK PERALTA MÁRQUEZ

SANTIAGO - CHILE

Diciembre, 2024

1 AGRADECIMIENTOS

Quiero expresar mis más sinceros agradecimientos a todas las personas que me han acompañado a lo largo de este proceso de formación profesional y personal. A mi familia, por ser el pilar fundamental de mi desarrollo y por brindarme siempre su apoyo incondicional. A mi madre y mi hermano, quienes estuvieron a mi lado en los momentos más difíciles, alentándome a dar lo mejor de mí, a alcanzar mis metas y a creer en mis capacidades. También agradezco a mi padre, quien me proporcionó las herramientas necesarias para convertir mis objetivos en realidad.

De manera especial, deseo reconocer al profesor Pablo Schwarzenberg, cuyo apoyo constante durante mi trabajo de título fue invaluable. Su orientación, junto con su disposición para responder dudas y ofrecer consejos en cada etapa del proyecto, fue fundamental para la culminación de este trabajo. También agradezco al profesor Billy Peralta, cuyas observaciones y comentarios constructivos abrieron nuevas puertas, permitiéndome dar inicio al proceso de publicación. Además, agradezco profundamente la confianza que ambos depositaron en mí, la cual ha sido clave para avanzar en los siguientes pasos de mi desarrollo académico y profesional.

A todos ustedes, mi más profundo agradecimiento.

Índice General

1 Agradecimientos 1
Índice General 2
Índice de Figuras 3
Índice de Tablas 3
2 Resumen 4
3 Introducción 4
3.1 Problema 5
3.2 Revisión Bibliográfica 5
3.3 Objetivos 9
4 Metodología 9
4.1 Datasets 9
4.2 Métricas de evaluación 11
5 Desarrollo 12
5.1 Preparación 12
5.2 Arquitectura de modelos 15
6 Resultados y discusión 26
6.1 Resultados de la clasificación de emociones 26
7 Conclusiones 42
8 Limitaciones del trabajo 43
9 Futuras mejoras 44
References 45

Índice de Figuras

1 Ejemplo de imágenes de FER2013 por clases . . . 10
2 Diagrama del primer método para la metodología facial . . . 13
3 Diagrama del segundo método para la metodología facial . . . 13
4 Diagrama de la metodología facial con la concatenación del dataset FER2013 y CK+48 13
5 Número total de imágenes con y sin concatenación FER2013 y CK+48. Antes de la división para el conjunto de validación . . . 14
6 Diagrama de la metodología textual para BERT y RoBERTa . . . 15
7 Arquitectura Propuesta del modelo facial Ensemble usando la concatenación de datasets (FER2013 + CK+48) . . . 18
8 Arquitectura propuesta del modelo textual . . . 23
9 Comparación de matrices de confusión en el conjunto de testeo de los modelos Ensemble usando uno y dos conjuntos de datos en el entrenamiento . . . 30
10 Gráfico de Accuracy y Loss de los conjuntos de entrenamiento y validación . . . 31
11 Comparación de matrices de confusión con otros trabajos . . . 33
12 Ruido existente en las clases tristeza y miedo . . . 34
13 Curva de Precision-Recall para la clase Disgusto . . . 35
14 Curva de Precision-Recall para las clases tristeza y miedo . . . 35
15 Gráfico de Accuracy y Loss de los conjuntos de entrenamiento y validación . . . 39
16 Comparación de matrices de confusión frente a otro trabajo . . . 40
17 Modelo propuesto multimodal . . . 42

Índice de Tablas

1 Número de muestras en el dataset FER2013 . . . 10
2 Número de muestras en el dataset ISEAR . . . 11
3 Detalles de los parámetros ajustados utilizando la técnica de Keras Tuner . . . 17
4 Arquitecturas de modelos testeados 1 . . . 19
5 Arquitecturas de modelos testeados 2 . . . 20
6 Arquitecturas de modelos ensemble testeados . . . 21
7 Arquitectura de modelos textuales testeados 1 . . . 24
8 Arquitectura de modelos textuales testeados 2 . . . 25
9 Resultados de los modelos experimentados con un único dataset FER2013, junto con los modelos ensemble propuestos. Evaluados en el conjunto de testeo de FER2013 . 27
10 Resultados de los modelos experimentados con la combinación de los datasets FER2013 y CK+48, junto con los modelos ensemble propuestos. Evaluados en el conjunto de testeo de FER2013 . . . 28
11 Resultados de modelos faciales con metodologías adicionales desarrolladas . . . 29
12 Comparación de los modelos de modalidad facial frente a otros trabajos en el conjunto de testeo FER2013 . . . 32
13 Resultados de los modelos experimentados con 5 emociones. Evaluados en el conjunto de testeo . . . 37
14 Resultados de los modelos experimentados con 7 emociones. Evaluados en el conjunto de testeo . . . 38
15 Resultados de modelos textuales con metodologías adicionales desarrolladas . . . 38
16 Comparación de los modelos de modalidad textual con otros trabajos utilizando el dataset ISEAR . . . 39
17 Métrica textual perplexity . . . 41

2 RESUMEN

Este trabajo presenta una exhaustiva revisión de múltiples artículos científicos centrados en el reconocimiento de emociones, con el objetivo de explorar, identificar y analizar los diferentes sistemas y modelos que se han implementado en este campo. El enfoque principal es el desarrollo de arquitecturas innovadoras para el reconocimiento de emociones en las modalidades facial y textual. Se proponen nuevas arquitecturas y metodologías para abordar el reconocimiento de emociones, demostrando la eficiencia y la potencia de combinar dos metodologías de balanceo de datos y dos datasets ampliamente utilizados en el reconocimiento de emociones faciales en el conjunto de entrenamiento, FER2013 y CK+48. Además, se aplica un enfoque de modelos ensemble, que mejoran la robustez de las predicciones, logrando un accuracy del 70.36% en el conjunto de testeo predefinido FER2013, superando los resultados de trabajos previos. De manera similar, en la modalidad textual, el uso de modelos transformers de NLP ha permitido superar las métricas actuales. Mediante la arquitectura propuesta RoBERTa-NGram-CNN se alcanzó un accuracy del 79% en el dataset ISEAR y un 80.70% con la integración de un mayor número de emociones, dando un total de 7 emociones al igual que la modalidad facial. Además, como extensión de este trabajo, se propone la integración de los modelos unimodales con mejores resultados en un modelo multimodal mediante la técnica de integración "late fusion" para explorar su capacidad de predecir emociones a partir de entradas simultaneas faciales y textuales. La finalidad del proyecto es desarrollar arquitecturas robustas y precisas unimodales para las modalidades facial y textual que puedan aplicarse en sistemas educativos para proporcionar retroalimentación a los docentes, resaltando la importancia de las emociones en la educación y abordar la problemática crítica en el ámbito educativo, como es la deserción y el desinterés en la educación superior.

3 INTRODUCCIÓN

Las emociones desempeñan un papel esencial en la forma en que las personas se comunican y se relacionan entre sí. Estas emociones se manifiestan a través de diversos medios, desde expresiones del rostro, texto y habla, hasta los gestos y la postura corporal. Adicionalmente, hay ciertos indicadores físicos, como cambios en la presión arterial, variaciones en la temperatura corporal y actividad muscular, entre otros, que pueden servir como indicativos del estado emocional de una persona [20]. Con los avances tecnológicos, el reconocimiento de emociones a través de diversas modalidades ha emergido como una rama fundamental dentro de la inteligencia artificial y la interacción humano-computadora. Las máquinas han desarrollado la capacidad de discernir y clasificar emociones humanas con una precisión creciente. Esta habilidad tiene aplicaciones en diversos campos, desde la publicidad y el entretenimiento hasta áreas esenciales como la medicina y, particularmente, la educación.

La relevancia de las emociones en el ámbito educativo es innegable. Las emociones desempeñan un papel crucial en el rendimiento de los estudiantes, ya que están directamente vinculadas con la cognición, motivación y control [9, 54]. De hecho, estudios como el de D'Errico et al.(2016)[19] han demostrado que cuando los estudiantes experimentan emociones positivas durante sus actividades de aprendizaje, aumentan significativamente su compromiso y participación. Además, según Wang et al.(2020)[53], el reconocimiento y análisis de las emociones de los estudiantes pueden ofrecer perspectivas valiosas que permiten a los educadores ajustar sus metodologías de enseñanza en tiempo real para mejorar la experiencia educativa en el aula.

Con la rápida evolución de la inteligencia artificial y el aprendizaje profundo, las técnicas de reconocimiento de emociones han experimentado avances significativos. Las redes neuronales convolucionales (CNN) han demostrado ser herramientas extremadamente potentes en el reconocimiento de emociones[24], al igual que las redes transformers[52]. A pesar de los éxitos

en la aplicación de CNN para el reconocimiento de emociones faciales en el contexto educativo[55] y los transformers en texto, aún existe una brecha notable de mejora en la literatura en el reconocimiento de las emociones en entornos educativos. Esta limitación impulsa la exploración de modelos que aborden de manera específica las modalidades facial y textual de forma independiente, para ofrecer herramientas más precisas y robustas en la detección emocional de los estudiantes. Para abordar este desafío, se propone el desarrollo de modelos innovadores para el reconocimiento de emociones en las modalidades facial y textual, utilizando arquitecturas avanzadas de ensemble, modelos transformers y técnicas de optimización que mejoren los resultados de estudios previos, determinando la efectividad de estos modelos. Además, se plantea el diseño de una arquitectura multimodal innovadora como una posible dirección para futuras exploraciones. Esta iniciativa representa un paso significativo hacia la transformación de la comprensión y respuesta a las emociones de los estudiantes en distintos entornos educativos.

3.1 Problema

Las emociones desempeñan un papel crucial en la adquisición del conocimiento en el ámbito educativo. Estas pueden potenciar la motivación del estudiante o, en contraparte, desencadenar su desinterés por completo. Según el estudio "Deserción de primer año y Reingreso a la Educación Superior en Chile"[15] del año 2019 y el informe del año 2022 sobre la "Retención de 1er año de pregrado"[46] de la subsecretaría de educación superior, se señala que en el año 2017 un 26% y en el año 2021 un 24.6% de los estudiantes abandonó la educación superior durante su primer año, lo cual, sigue siendo preocupante, debido a que el porcentaje del año 2021 se obtiene con una mayor cantidad de estudiantes que ingresaron a la educación superior con un valor de 1.294.734[47]. Continuando en un ámbito más generalizado y creciente como la educación en línea, esta presenta desafíos aún mayores en términos de deserción, sobre todo en instituciones que priorizan esta modalidad de enseñanza. Solo alrededor del 15% de los estudiantes que cursan programas en línea logran obtener un título o certificación[35]. A pesar de la gravedad de esta situación, aún no contamos con modelos efectivos y replicables que logren representar la realidad de las emociones de los estudiantes en un contexto educativo y que puedan ser implementados en sistemas efectivos que reconozcan las emociones de los alumnos a través de diversas modalidades, con el objetivo de que el docente pueda ajustar sus metodologías de enseñanza a tiempo, brindando apoyo a quienes más lo requieran.

3.2 Revisión Bibliográfica

El reconocimiento de emociones ha experimentado notables avances en la educación, Xu et al.[55], señala la importancia de las expresiones faciales como indicadores de emociones, destacando el papel vital que juega las emociones de los estudiantes a la hora de absorber y procesar la información, proporcionando a los educadores una herramienta para adaptar su enseñanza en tiempo real, destacando que en la realidad las emociones frustración, felicidad y somnolencia sobresalen al momento de ser reconocidas. Aún así, el reconocimiento de emociones en la educación no solo abarca la modalidad facial, sino que también la modalidad del habla. Esto se puede visualizar en el siguiente paper, donde se sigue un enfoque educacional, Abdelhamid[1], plantea mejorar la eficiencia de la educación en línea mediante un sistema basado en mel-espectrogramas y capas Conv1D que envían retroalimentación instantánea al estudiante desde la nube, mostrando el valor de la sincronización en la interacción educador-estudiante. Un enfoque diferente, y donde las redes convoluciones han demostrado ser herramientas eficaces, es el estudio presentado por Meena y Mohbey[32]. Este trabajo se centra en comparar técnicas de aprendizaje por transferencia de emociones faciales, debido al creciente uso de redes sociales. En el estudio se hace uso de tres bases de datos, CK+,

FER2013 y JAFFE. Mediante los resultados se explica que, el modelo VGG-19 obtuvo valores equilibrados en todas las métricas con un 65.41% de accuracy en FER2013. Es importante destacar que, al igual que el trabajo anterior, aunque tuviera valores de accuracy alto, estos modelos sufren un problema de overfitting, debido a la alta complejidad de los modelos. Otro trabajo en el que las redes neuronales convolucionales(CNN) han sido relevantes para el reconocimiento de emociones faciales es el trabajo de Sahoo et al.[43], en el cual, se proponen 2 modelos CNN de 6 y 10 capas que se comparan con VGG16, haciendo uso del dataset FER2013. El objetivo del modelo es su implementación en un sistema instalado en el tablero de un vehículo, de forma que alerte al conductor sobre su estado emocional en tiempo real, considerando como indicadores de desconcentración las emociones negativas. La evaluación de los modelos se hizo mediante las métricas F1, Recall y Accuracy. Los modelos hacen uso de 4 capas primordiales, capa convolucional, normalización, agrupación y regularización. El aporte del estudio muestra que el modelo CNN de 10 capas logra un valor de accuracy superior a ambos modelos a comparar, obteniendo un valor de accuracy de 68.34%, lo cual, sugiere que el modelo podría ser una opción interesante para ser implementado en un sistema al poseer un tiempo computacional menor. Sin embargo, es importante destacar que no superó al mejor modelo comparado FerNet con 69.57%. Siguiendo el uso de los modelos convolucionales para la predicción de las emociones, Oguine et al.[36] desarrollan un modelo convolucional combinado con el detector facial Haar Cascade para ser implementando en una cámara web. El proceso se basa en aplicar el detector facial para obtener únicamente el rostro de la persona, para ser ingresado a un modelo de 7 capas CNN, con capas de agrupación y regularización . Este enfoque logra un accuracy del 70% sin overfitting, pero con sesgo existente hacia la clase mayoritaria debido a la ausencia de técnicas de balanceo de datos. Otro trabajo relevante es de Sadak et al.[42], que emplean el dataset CK+48, utilizando capas de atención junto con capas convolucionales y módulos residuales. Su modelo alcanza un valor de accuracy alto del 99%, lo que demuestra el potencial y la relevancia de las muestras que presenta el dataset CK+48. Además, un punto interesante es el uso de capas de atención resultando ser eficiente. También, Chaudhari et al.[11], que igualmente utiliza el dataset CK+48, plantea la creación de un nuevo dataset combinando datos de FER2013, CK+48 y AffectNet para entrenar un modelo Vision Transformer, obteniendo un accuracy del 53% utilizando datos de testeo de este nuevo dataset. Esta metodología de combinación de muestras podría ser una estrategia prometedora para mejorar los resultados. Por otro lado, un sistema que va más allá del ámbito estrictamente educativo y se enfoca en el reconocimiento de emociones en el habla, se puede ver en el estudio de Lee et al.[29] que propusieron un sistema para la población anciana de Taiwán, que combina Voice Augmentation y Mel-Espectogramas con capas CNN en GoogleNet, alcanzando 79.81% de accuracy. Siguiendo el enfoque es lo que plantea Singh et al.[49] diseñando un modelo SER basado en el género entre hombre y mujer, alcanzando un 72.07% accuracy, Otra forma que se ha abordo esta tarea es como señala Fema y Marquez[24] que subrayan el papel crucial que desempeñan las emociones en la comunicación humana. Se introdujo la combinación de MFCC y CNN en los datasets RAVDESS, TESS y CREMA-D, con grabaciones semánticas, logrando 78% de accuracy. El problema del reconocimiento de emociones mediante la modalidad oral en el ámbito educativo, es que, en las aulas educacionales hay una gran cantidad de alumnos, donde cada uno de ellos puede hablar al mismo tiempo, generando una gran cantidad de ruido en los datos al ingresar al modelo. Una alternativa para suprimir el ruido es el reconocimiento de emociones mediante texto natural, una modalidad muy desafiante y estudiada. Comenzando con modelos simples, como señala Yohanes et al.[57], que propone analizar tres modelos en el dataset ISEAR, logrando un 60.26% de accuracy con GRU, siendo una buen punto de partida para modelos que requieren una

menor potencia computacional. Sin embargo, también se han llevado a cabo estudios con algoritmos más potentes para el reconocimiento de emociones, como señala Abas et al."[4]. Que se propone la aplicación de modelos BERT con la agregación de una capa convolucional y max-pooling. Este estudio hace uso del database ISEAR aplicando técnicas como la reducción de palabras repetidas y normalización. El resultado del estudio demuestra que BERT-CNN obtiene una métrica de F1-score de 76% y accuracy de 77%, siendo mayor en comparación con modelos conocidos y variantes. Otra aplicación interesante de una variante de BERT se plantea en el paper de Acheampong et al.[2], donde compara 3 variantes de BERT. Este estudio hace uso del mismo dataset ISEAR, obteniendo que la variante RoBERTa logra un mayor accuracy de 74% al momento de reconocer una emoción. De igual forma, otro estudio de los mismos autores[3], obtiene el mismo valor de accuracy de 74% usando el modelo BERT con capas Bi-LSTM. Esto demuestra que BERT-CNN en comparación a los demás estudio, es un modelo preciso al momento de reconocer emociones. También se han desarrollado arquitecturas que combinan capas de manera variada como lo propuesto por Punnet Kummar et al.[28]. Este estudio implementa una arquitectura dual channel para el reconocimiento de las cuatros emociones Happy, Sad, Hate y Anger en el ámbito textual. La arquitectura combina BERT con capas BiLSTM y Conv1D en dos canales. Esta arquitectura logra alcanzar un 79.17% en accuracy en el dataset ISEAR. Es importante señalar que este ultimo estudio no es directamente comparable con estudios anteriores al utilizar una menor cantidad de clases con una menor complejidad. De igual forma, presenta una arquitectura interesante para analizar. Asimismo, el paper de Balbuena[7], plantea la necesidad de modelos potentes para el reconocimiento de emociones tanto facial como textual. El estudio propone el uso del modelo MiniXception para el reconocimiento facial, y modelo BiLSTM para texto, utilizando los datasets AffectNet y CBET. Los resultados sugieren que MiniXception supera con un 86.11% de accuracy en AffectNet. De manera similar, BiLSTM obtiene un 84.16% de accuracy en CBET, siendo ambos modelos buenas alternativas para desarrollar un modelo multimodal. Adicionalmente, en un enfoque relacionado en la visión, el paper de Chaudhari et al.[12] propone el uso del Vision Tranformers para el reconocimiento de emociones, fusionando tres conjuntos de datos, FER2013, AffectNet y Ck+48. El paper hace uso del modelo Vit-B/16/S, obteniendo un 52.25% de accuracy en tan solo 25 épocas. Finalmente, se puede analizar que existen varias modalidades para el reconocimiento de emociones, por ende la unión de ellas deriva a los modelos multimodales. Este enfoque se analiza en el paper de Huang et al."[26], que propone un modelo multimodal de análisis de sentimientos, que integra el modelo VGG-19 para facial y LSTM para texto. En este estudio se propone el desarrollo de dos modelos con atención unimodales separados para ambas modalidades. Además, se introduce la fusión de manera intermedia con la técnica "late fusion" de los dos modelos unimodales. Los resultados obtenidos en este estudio demuestran que la aplicación multimodal mejora el entendimiento de la emociones. De forma, que estos hallazgos respaldan la eficiencia de una propuesta multimodal. Siguiendo la revisión de los avances del año 2024. En cuanto a la modalidad facial, Ciraolo et al.[13], proponen implementar IA emocional en la medicina, mediante la transformación 3D con Fase Mesh de los datasets FER2013, CK+, AffectNet y Mixed, que combina los 3 datasets anteriores. El modelo SVM superó a los demás modelos, logrando el accuracy mayor de 53.2% para FER2013 y 56.9% en el dataset Mixed. Otra forma de implementación en la educación que proponen Thao et al.[51] es un modelo para ser integrando en un sistema de cámaras para analizar el nivel de concentración del estudiante, experimentando cierta presión psicológica en el estudio. Se propuso un modelo EduVit que es la combinación de MobileVit y un bloque SE, logrando un accuracy del 66% utilizando las emociones happy,sad,surpise y anger del dataset FER2013. Asimismo, es importante destacar

la creciente popularidad de la modalidad de texto, debido a su utilidad. En este sentido, se han realizado diversos estudios recientes como el de Adda et al.[18], que proponen re-entrenar el modelo Mistral7B mediante el dataset ISEAR y compararlo con otros modelos, logrando un accuracy del 76%, siendo elevado en comparación a otros trabajos. Como último, los modelos multimodales no se han quedado atrás, como en el estudio de Almulla et al.[5], que propone un modelo multimodal con tres entradas, texto (ETD), facial (FER2013) y voz. Cada uno de estos modelos se entrenan de forma separada con su conjunto de datos y concatenándose al final para dar una emoción final. Para el texto se emplea LR, para el lado facial hace uso de un modelo CNN. Los resultados interesantes consisten en un accuracy del 64% para la modalidad textual utilizando un modelo LR y un 69% de accuracy para la modalidad facial ocupando CNN. La fusión que se realiza es mediante la concatenación ponderada de los modelos, asignando un peso de importancia diferente del 55%, 38% y 7%.

Al analizar los diferentes papers, se puede concluir que en el panorama actual de investigación sobre el reconocimiento de emociones, existen múltiples estudios que abordan el reconocimiento a través de distintas modalidades, tales como facial, voz o texto. Sin embargo, este reporte introduce una singularidad y un valor agregado al campo de estudio. En primer lugar, se propone la implementación de una nueva arquitectura ensamble en la modalidad facial no vista en las revisiones bibliográficas, permitiendo aprovechar las fortalezas de cada modelo individual. Asimismo, se introduce una nueva arquitectura Fine-Tunning de transformers en la modalidad de texto con la agregación de capas N-grams. Aunque el enfoque principal de este trabajo se limita a los modelos unimodales, se propone también un modelo multimodal para futuras investigaciones, combinando los modelos unimodales que presentaron mejores resultados. Asimismo, se incorpora un análisis de la combinación de dos datasets ampliamente estudiados en el reconocimiento de emociones facial, evaluando si esta implementación mejora los resultados actuales. Como también, en la modalidad textual, se combinarán emociones provenientes de diferentes fuentes para igualar la misma cantidad de emociones en ambos modelos, enriqueciendo la diversidad de datos y mejorando la capacidad predictiva de los sistemas de reconocimiento de emociones.

3.3 Objetivos

(1) **OG** Objetivo general
Desarrollar modelos unimodales faciales y textuales innovadores para el reconocimiento de emociones en el ámbito educativo, con el propósito de ofrecer modelos robustos y precisos, aplicables en sistemas educativos, facilitando la retroalimentación y comprensión de emociones, y estableciendo las bases para una futura integración multimodal.
(2) Objetivos específicos
- **OE1** Diseñar y desarrollar modelos unimodales de reconocimiento de emociones tanto facial como textual.
  Los modelos deben ser capaces de capturar de manera eficiente y precisa las emociones de los estudiantes. Esto incluye replicar modelos de referencia, desarrollar modelos propuestos mediante nuevas arquitecturas, y ajustar hiperparámetros mediante técnicas de optimización para encontrar configuraciones óptimas.
- **OE2** Evaluar y comparar los modelos unimodales con investigaciones previas.
  Se llevará a cabo una evaluación exhaustiva de los modelos unimodales desarrollados, comparando métricas claves definidas para medir su eficiencia. Además, se analizarán los resultados frente a trabajos previos para validar el rendimiento alcanzado y seleccionar los modelos más prometedores en ambas modalidades.
- **OE3** Establecer las bases para una futura integración multimodal a partir de los modelos unimodales seleccionados.
  Se analizarán los modelos unimodales con mejor rendimiento durante la evaluación individual en ambas modalidades, proponiendo un modelo multimodal preliminar que integre dichos modelos, sirviendo como referencia para futuras implementaciones.

4 METODOLOGÍA

4.1 Datasets

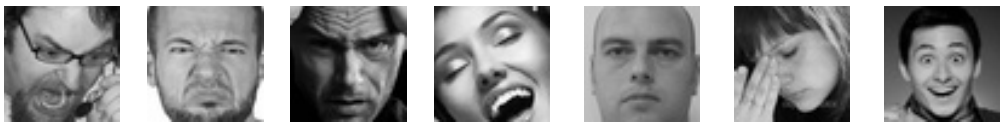
Para capturar los datos se utilizarán los sitios web oficiales de cada base de datos y plataformas web como Data Science Kaggle. Cabe mencionar que los conjuntos de datos a utilizar son los más usados en el análisis de diferentes estudios en ambas modalidades, permitiendo obtener resultados confiables y comparables con otros trabajos de investigación.

- Facial
  - FER2013[21]: Es un conjunto de datos desarrollado por la Universidad de Montreal como parte de un proyecto de investigación y presentando para la Conferencia Internacional sobre Aprendizaje Automático. Este dataset contiene dos conjuntos de datos predefinidos, uno de entrenamiento con 28.709 muestras y otro de testeo con 7.178, sumando un total de aproximadamente 35.887 imágenes de rostros faciales en diferentes expresiones y posiciones, de tamaño 48x48 en escala de grises, etiquetadas en 7 emociones (angry, disgust, fear, happy, sad, surprise y neutral). Además, el dataset presenta un desafío al poseer las clases desbalanceadas en ambos conjuntos, como se puede ver en la Table 1, donde la clase minoritaria Disgust contiene 547 imágenes en total y la clase mayoritaria Happy 8.989.
  - CK+48: Es la versión procesada del conjunto de datos Extended Cohn-Kanade Dataset (CK+)[31], desarrollado por Carnegie Mellon University y University of Pittsburgh. CK+48 es utilizado en tareas de reconocimiento de emociones y contiene 981 imágenes de rostros faciales centrados, extraídos de secuencias de vídeos, con cada rostro repetido tres veces. El dataset se compone de siete clases de emociones desbalanceadas (fear, anger, sadness, happy, contempt, disgust, surprise), donde la clase minoritaria es contempt y la clase mayoritaria

surprise. Las imágenes poseen un tamaño de 48x48 píxeles, están normalizadas y en escala de grises. Este dataset será utilizado para un segundo análisis, concatenándose con el dataset FER2013 en las clases igualitarias.

FER2013 Dataset	Train	Test	Total
Happy	7215	1774	8989
Neutral	4965	1233	6198
Sad	4830	1247	6077
Fear	4097	1024	5121
Angry	3995	958	4953
Surprise	3171	831	4002
Disgust	436	111	547

Table 1. Número de muestras en el dataset FER2013



(a) Angry (b) Disgust (c) Fear (d) Happy (e) Neutral (f) Sad (g) Surprise

Fig. 1. Ejemplo de imágenes de FER2013 por clases

- Texto
  La entrada de texto se realizará principalmente con el dataset ISEAR[45] para el primer análisis. Para el segundo análisis, se utilizarán clases del segundo conjunto de datos MELD[40] para igualar las emociones presentes en el dataset ISEAR con el conjunto de datos facial.
  - International Survey On Emotion Antecedents and Reactions (ISEAR)[45]: De la Universidad de Ginebra, que fue generado a partir de entrevistas realizadas a personas de 37 países, evaluando las emociones al mencionar una situación en particular. El dataset fue etiquetado por la colaboración de psicólogos, estudiantes y participantes, y consta de alrededor de 7.666 oraciones de estudiantes etiquetados en siete emociones (joy, fear, anger, sadness, disgust, shame y guilt) balanceados con alrededor de 1.096 datos por clase como se puede ver en la Table 2.
  - MELD[40]: De la Universidad de Michigan, está compuesto por datos multimodales de la serie de televisión "Friends". Contiene alrededor de 9.989 oraciones de conversaciones, etiquetadas por estudiantes de postgrado con alto dominio del habla y la escritura en inglés, en siete emociones (anger, disgust, fear, joy, neutral, sadness y surprise). La utilización de este dataset tiene la finalidad de entregar las dos emociones faltantes en ISEAR que son Neutral y Surprise.

ISEAR Dataset	Total	Example
Anger	1096	"When the morning newspaper has not arrived."
Disgust	1096	"I found some worms in the food and I had obviously eaten some."
Fear	1095	"When I was involved in a traffic accident."
Guilt	1093	"When my uncle and my neighbour came home under police escort."
Joy	1094	"When I pass an examination which I did not think I did well."
Sadness	1096	"When I lost the person who meant the most to me."
Shame	1096	"When I did not speak the truth."

Table 2. Número de muestras en el dataset ISEAR

El primer análisis consiste en desarrollar una nueva arquitectura para los datasets base FER2013 e ISEAR. Posteriormente, en la modalidad facial, se concatenarán ambos datasets de imágenes para una mayor cantidad de datos de entrenamiento. En la modalidad textual, se añadirán las clases faltantes al dataset ISEAR para igualar ambas modalidades, seleccionando las siete emociones a estudiar: "angry", "disgust", "fear", "happy", "neutral", "sad" y "surprise", con el fin de crear un modelo multimodal en un futuro trabajo.

4.2 Métricas de evaluación

Para la evaluación, se utilizarán métricas de clasificación utilizadas en varios trabajos relacionados en ambas modalidades, como Accuracy, Precision, Recall y F1-Score. El Accuracy es la proporción de predicciones correctas respecto al total de predicciones realizadas. Precision es la proporción de predicciones positivas correctas sobre el total de predicciones positivas realizadas. Recall representa la proporción de instancias positivas que fueron correctamente predichas sobre el total de instancias positivas reales. F1-Score es la media armónica de Precisión y Recall, y evalúa el balance entre la calidad de las predicciones positivas y la proporción de positivos correctamente identificados. Estos valores se obtendrán utilizando la media no ponderada por clase (macro avg), que permite analizar las métricas otorgando la misma importancia a cada clase, independiente de su frecuencia, eliminando sesgos hacia la clase mayoritaria. Las ecuaciones matemáticas se ven a continuación:

Formula Accuracy:

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

Formula Precision:

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

Formula Recall:

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

Formula F1-Score:

$$\text{F1-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

5 DESARROLLO

5.1 Preparación

Con los datos obtenidos, se definieron estrategias específicas para ambas modalidades: facial y textual. Para la modalidad facial, el dataset FER2013 contiene dos conjuntos de datos predefinidos, train y test. El conjunto de entrenamiento se dividirá en un 80% para el entrenamiento del modelo y un 20% para la validación del modelo. El conjunto de testeo no se le aplicará ningún cambio, ya que se ocupará para evaluar el modelo final en datos no vistos y comprender su comportamiento en la realidad. Para el procesamiento, se aplicarán dos métodos de balanceo, debido a que, como menciona Ahmad Khan[27] el balanceo de datos es una buena estrategia para que el modelo aprenda a clasificar cada emoción individual correctamente y eliminar el sesgo hacia la clase mayoritaria. El primer método de la Fig. 2 utiliza ImageDataGenerator para dividir el conjunto de entrenamiento en un 20% para validación y, al mismo tiempo, aplicar Data-Augmentation en el conjunto de entrenamiento, añadiendo ruido y variedad para reflejar un entorno realista. Las dos transformaciones más destacadas incluyen, para la primera: rotación aleatoria entre -10° y 10°, zoom de 20%, desplazamientos horizontales y verticales del 10%, volteo horizontal y relleno cercano. Para la segunda transformación se aplicó: rotación entre -20° a 20° y volteo horizontal aleatorio. En este primer método, se aplicó un balanceo de datos mediante la técnica de ponderación de pesos en el conjunto de entrenamiento, asignando mayores pesos a las clases minoritarias y menores a las mayoritarias, logrando un balance entre las clases durante el entrenamiento. El segundo método de la Fig. 3 implica realizar cambios previos en los datos, dividiendo entrenamiento y validación previamente mediante "train_test_split", manteniendo la misma proporción de 80%-20%. Este método busca balancear las clases a 5.772 imágenes, que es la cantidad de datos de la clase mayoritaria (happy) mediante la técnica de Data-Augmentation. Esto se logrará aplicando transformaciones leves a todas las muestras generadas como rotación entre -2° y 2°, desplazamiento horizontal 2%, desplazamiento vertical 2% y zoom 2%, creando nuevos datos sintéticos y manteniendo la misma cantidad de datos en todas las clases. Luego, se aplicó nuevamente Data-Augmentation en el conjunto de entrenamiento con las mismas transformaciones del primer método. Además, como tercer análisis propuesto con mejores resultados de la Fig. 4, se propuso combinar los datasets FER2013 y CK+48. Se concatenaron las imágenes de las etiquetas igualitarias del conjunto de entrenamiento de FER2013 con todas las imágenes del CK+48, aplicando posteriormente una división del 20% para validación. Esto añadió 927 imágenes adicionales a las clases de FER2013: 135 de "angry", 177 de "disgust", 75 de "fear", 207 de "happy", 84 de "sad" y 249 de "surprise", como se ve en la Fig. 5. La cantidad total de imágenes se incrementó a 29.636, quedando 23.711(80%) para entrenamiento y 5.925(20%) para validación. Este enfoque demostró ser el más efectivo, aumentando la robustez del modelo al proporcionar una mayor variabilidad en los datos. Otro análisis adicional fue concatenar los datos de CK+48 con el conjunto de entrenamiento de FER2013 después de dividir este último para generar el conjunto de validación. Asimismo, se realizó la combinación de FER2013 con CK+ para analizar posibles diferencias en comparación con la combinación con CK+48.

Los tres conjuntos de datos (train, test y val) serán normalizados para que los píxeles estén en un rango de 0 y 1. En los tres conjuntos se conservará el tamaño de las imágenes de 48x48 y se transformarán a escala RGB de tres canales para poder ser utilizados en modelos pre-entrenados.

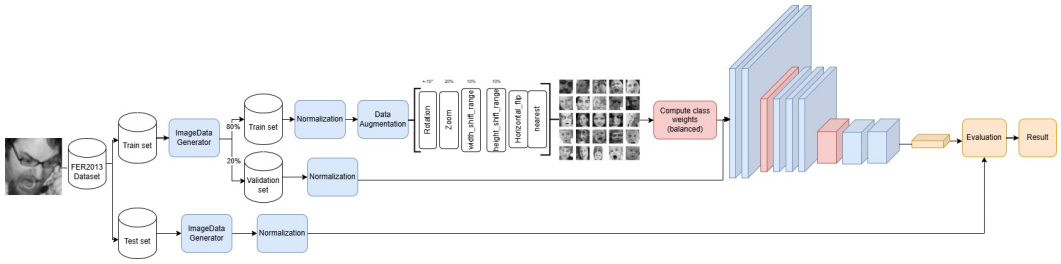


Fig. 2. Diagrama del primer método para la metodología facial

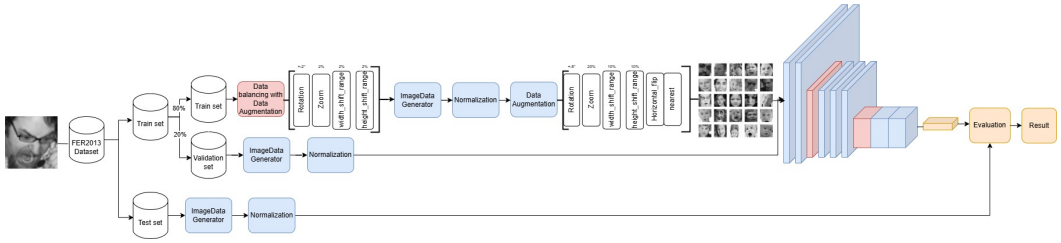


Fig. 3. Diagrama del segundo método para la metodología facial

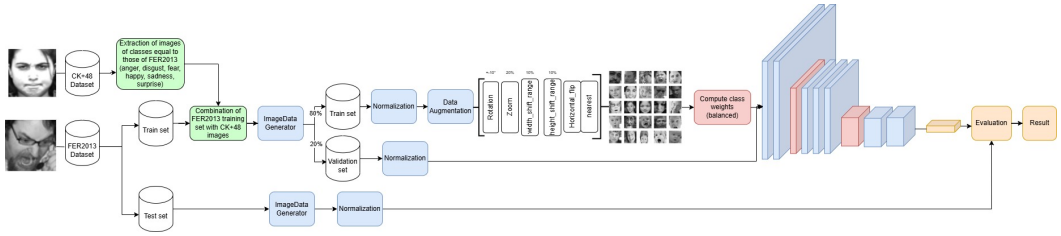


Fig. 4. Diagrama de la metodología facial con la concatenación del dataset FER2013 y CK+48

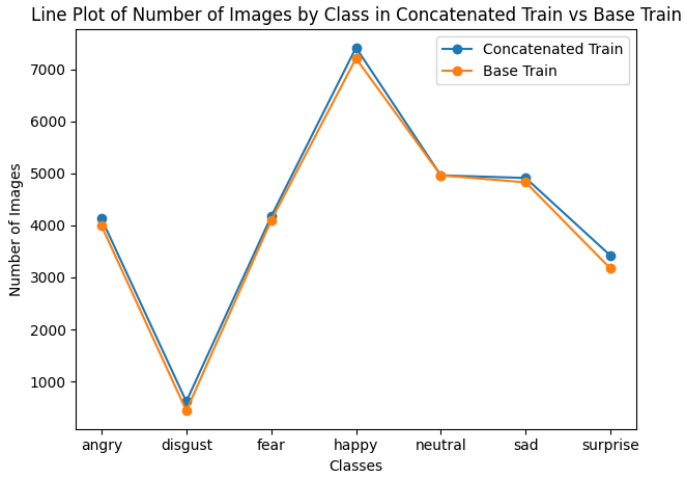


Fig. 5. Número total de imágenes con y sin concatenación FER2013 y CK+48. Antes de la división para el conjunto de validación

En segundo lugar, en la modalidad textual, se procederá a eliminar las emociones 'shame' y 'guilt' del dataset ISEAR, dado que no son representativas en un contexto educativo. En cambio, se centrarán en cinco emociones claves: 'angry', 'disgust', 'fear', 'happy' y 'sad'. Esta decisión se fundamenta en investigaciones que indican que estas emociones son relevantes y comúnmente experimentadas en entornos educativos, lo cual influye en los resultados del aprendizaje y el rendimiento académico[41]. De esta forma, el análisis se llevará a cabo de dos formas. La primera es mediante las cinco emociones mencionadas, y la segunda, agregando las emociones 'neutral', utilizada en la educación para analizar las emociones de los estudiantes en sistemas educativos[55], y 'surprise', para un total de siete emociones, garantizando así la coherencia con la modalidad facial al igualar la cantidad de clases en ambas modalidades para un futuro modelo multimodal. Se añadirán 1.096 datos por cada clase nueva añadida para mantener el equilibrio balanceado del dataset entre las emociones. Las emociones se extraen del conjunto de datos MELD[40]. Posteriormente, se aplicarán diversos métodos de manejo de datos, como se puede ver en la Fig. 6, se utilizarán técnicas de limpieza de datos, como la conversión a minúsculas, eliminación de emoticonos, corrección de errores de ingreso, eliminación de espacios en blanco, signos de puntuación y arrobas. Además, las emociones "joy", "anger" y "sadness" se renombrarán a "happy", "angry" y "sad", para mantener la consistencia con la modalidad facial. Luego, se dividirá el conjunto de datos en 70% para entrenamiento, un 20% para validación y un 10% para testeo, utilizando una semilla específica. Una vez divididos, se aplicará aumento de datos textuales en el conjunto de entrenamiento utilizando el método "EasyDataAugmenter", que introduce variaciones como reemplazo aleatorio de sinónimos, eliminación de palabras al azar, intercambio aleatorio de posiciones de palabras e inserción de nuevas palabras al texto en posiciones aleatorias. Estas variaciones utilizan el diccionario "WordNet"[34]. Esta técnica tiene como objetivo añadir ruido reflejando la variabilidad real. Posteriormente, se aplican las técnicas de limpieza de datos mencionadas previamente al texto generado, y se aplica OneHotEncoder para convertir las clases en una representación binaria. Finalmente, el texto se procesará de dos formas para su ingreso a diferentes embeddings. Para la primera forma, utilizando Word2Vec y GloVe, se tokenizará el texto mediante el tokenizador de Keras y se aplicará padding para que todos

los textos tengan la misma longitud. En cuanto al segundo embedding, como Bert, RoBERTa y otros, se empleará HuggingFace junto con sus tokenizadores respectivos y se aplicará padding. Esto proporcionará los input_ids y attention_mask requeridos por Bert y RoBERTa para su correcto funcionamiento. Es importante mencionar que, en ambas formas se limitará la longitud máxima de texto a 300, considerando las capacidades computacionales disponibles. Además, se llevarán a cabo dos evaluaciones adicionales utilizando siete emociones. El primer análisis consiste en la aplicación de lemmatization al texto de entrada durante el proceso de limpieza de datos, reduciendo palabras a su forma raíz, y el segundo concatenando el dataset IEMOCAP, obteniendo una combinación de tres datasets, asegurando que cada clase tenga 1.311 oraciones. Se comienza con MELD, y si no se alcanza la cantidad esperada, se incorporan oraciones de IEMOCAP. De esta manera, se evaluará si se logra una mejora.

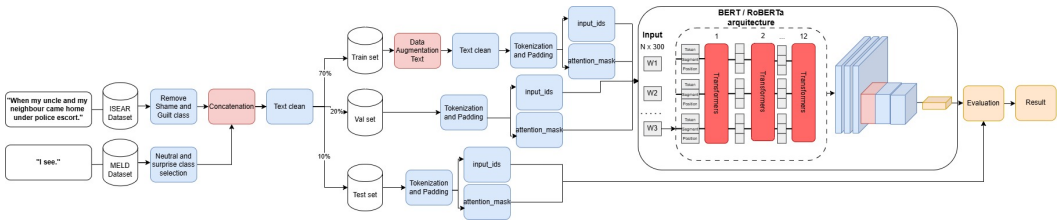


Fig. 6. Diagrama de la metodología textual para BERT y RoBERTa

5.2 Arquitectura de modelos

Para llevar a cabo el entrenamiento de los modelos se va a hacer uso de un equipo con una tarjeta de vídeo NVIDIA RTX 2070.

Para la modalidad facial, se probaron más de 15 variantes de modelos de aprendizaje profundo entre las dos metodologías, incluyendo la replicación de modelos previamente investigados como CNN[24], VGG19[48], Resnet50[23], Mini-Inception[6], ViT Transformers[17], entre otros. Entre todas las variantes probadas con diferentes hiperparámetros, las arquitecturas basadas en redes convolucionales, entrenadas con la concatenación de ambos datasets FER2013 y CK+48, resultaron ser las más efectivas en términos de métricas de rendimiento. Se destacan cinco modelos principales, el primer modelo renombrado como CNN-1 consta de cinco capas convolucionales, con filtros de 32, 64 y 128 en las tres primeras capas y 512 en las dos últimas con regularizador L2 del 0.01. Todas emplean activación ReLU, padding 'same', BatchNormalization, MaxPooling2D y Dropout de 25%, con dos capas densas de 256 y 512 neuronas con regularizador L2 del 0.01 en esta última, BatchNormalization y Dropout del 25% nuevamente en ambas capas, y optimizador Adam con learning rate(lr) del 1e-4. El modelo CNN-2 aumenta su complejidad con una estructura de seis capas convolucionales organizadas en pares, siendo las variantes con mejor resultado. El primer par tiene 32 y 64 filtros, seguido por un segundo y tercer par de 128 y 256. Cada par incluye BatchNormalization, MaxPooling2D y Dropout del 30% , con regularizador L2 del 0.01 en cada capa convolucional del segundo y tercer par. Además, incluye una capa densa de 1024 neuronas con Dropout del 50% y optimizador Adam con lr del 1e-4. El modelo CNN-3 sigue la misma arquitectura que el segundo modelo anterior, pero introduce una capa densa adicional, sumando en total dos capas Fully Connected. La primera capa densa, que sigue después de la capa Flatten, posee 2048 neuronas con activación ReLU, seguido de un Dropout del 50%, y una segunda capa densa con 1024 neuronas con activación ReLU y un Dropout de 50%. Este modelo utiliza el

optimizador Adam con un lr del 1e-4. El cuarto modelo utiliza transferencia de aprendizaje con VGG19 preentrenado con los pesos de ImageNet, añadiendo BatchNormalization, GlobalAveragePooling2D, y dos capas densas de 256 y 128 neuronas con regularizador L2 del 0.001 y Dropout del 40%. El modelo posee optimizador Adamax con lr de 1e-4. El quinto modelo se desarrolló mediante la aplicación de Keras Tuner[37] utilizando la técnica de optimización Bayesian Optimization, basado en la probabilidad y el aprendizaje de la función objetivo. Las opciones de hiperpárametros experimentados se detallan en la Table 3. El modelo de keras tuner obtenido se compone de cinco pares de capas convolucionales, cada uno compuesto por dos capas convolucionales con un tamaño de kernel de 3,3, padding en same para mantener la dimensionalidad, activación ReLU y regularizador L2 con un valor de 0.01. Además, se incluye una capa BatchNormalization, seguido de una capa MaxPooling2D con tamaño de 2,2. El primer par posee 64 y 288 filtros en las capas Conv2D y un Dropout del 20%. El segundo par con 160 y 160 filtros con un Dropout del 40%. El tercer par con 288 y 256 filtros con un Dropout del 20%. El cuarto par con 96 y 224 filtros con un Dropout del 40%. El quinto par con 32 y 160 filtros con una Dropout del 40%. Después de todos, se agregó una capa Flatten y una capa densa de 1024 neuronas con activación ReLU, seguido de una capa Dropout del 50%. Finalmente, se desarrollaron varios modelos ensemble con diferentes combinaciones, siendo tres los propuestos que obtuvieron mejores resultados. El primer modelo ensemble consiste en la combinación de dos arquitecturas de modelos distintos entrenados utilizando solamente el dataset FER2013, mientras que el segundo y tercer modelo ensemble consisten en la combinación de tres arquitecturas de modelos distintos entrenados con la unión de ambos datasets FER2013 + CK+48. Los tres ensembles utilizan el enfoque de promedio para combinar las salidas de los modelos individuales, ya que se observó que esta estrategia no solo mejora los resultados en comparación con métodos como la votación, sino que también favorece una mejora en la generalización del modelo. En los tres casos, los tres primeros modelos del ensemble utilizan la primera metodología de entrenamiento, que aplica el balanceo de datos mediante pesos ponderados, mientras que los últimos tres modelos restantes aplican la segunda metodología de entrenamiento, que emplea el balanceo de datos mediante la generación de datos sintéticos a través de data augmentation, igualando el número de imágenes en cada clase. El primer ensemble se compone de la siguiente manera: Los primeros tres modelos balanceados con pesos ponderados consisten en el Modelo 1 (CNN-1) con la primera transformación de data augmentation, el Modelo 2 (CNN-2) con la primera transformación de data augmentation, y el Modelo 2 (CNN-2) nuevamente, pero esta vez con la segunda transformación de data augmentation, mientras que los últimos tres modelos emplean balanceo a través de la generación de datos sintéticos y la primera transformación de data augmentation, son el Modelo 1 (CNN-1), el Modelo 2 (CNN-2) y el Modelo 4 (VGG19). Por otro lado, el segundo modelo ensemble propuesto, está compuesto por todos los modelos utilizando la primera transformación de data augmentation y los primeros tres modelos balanceados con pesos ponderados, siendo el Modelo 1 (CNN-1), el Modelo 2 (CNN-2), y el Modelo 3 (CNN-3). Estos son seguidos por los últimos tres modelos balanceados mediante la generación de datos sintéticos, consistiendo en el Modelo 1 (CNN-1), el Modelo 2 (CNN-2) y el Modelo 4 (VGG19). El tercer ensemble también fue entrenado con la combinación de los dataset FER2013 y CK+48, obteniendo mejores métricas. Este ensemble utiliza los mismos modelos que el segundo ensemble, pero con una modificación, de los tres primeros modelos que emplean el balanceo mediante pesos ponderados, se reemplaza el modelo CNN-3 por el modelo optimizado mediante Keras Tuner. El ensemble se estructura de la siguiente manera: Los primeros tres modelos utilizan el balanceo de datos mediante pesos ponderados y la primera transformación de data augmentation, siendo el Modelo 1 (CNN-1), el Modelo 2

(CNN-2) y el Modelo 5 Keras Tuner. Estos seguidos por los últimos tres modelos balanceados mediante la generación de datos sintéticos y la primera transformación de data augmentation, consistiendo en el Modelo 1 (CNN-1), el Modelo 2 (CNN-2) y el Modelo 4 (VGG19). De esta forma, se unen estrategias de balanceo y optimización de hiperparámetros, como se muestra en la arquitectura propuesta en la Fig. 7.
Los tres modelos ensembles combinan las salidas de estos seis modelos entrenados para lograr una mayor precisión y robustez en la clasificación de emociones faciales.

Cabe mencionar que los modelos propuestos ensembles se entrenaron durante 200 épocas, con un batch size de 64, un optimizador Adam con learning rate de 1e-4 y shuffle en True.

Parameter	Choices	Step	Value
number of Conv2D layers	2, 3, 4, 5, 6	max value: 6, random	5
number of filters 1 (each Conv2D layer)	32, 64, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448, 480, 512	step: 32	64, 160, 288, 96, 32
number of filters 2 (each Conv2D layer)	32, 64, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448, 480, 512	step: 32	288, 160, 256, 224, 160
Dropout range (Fraction of the input units to drop) (each Conv2D layer)	0.1, 0.2, 0.3, 0.4, 0.5	step: 0.1	20%, 40%, 20%, 40%, 40%
Max trials	-	-	6
Executions per trial	-	-	3
Total execution time	-	-	18h:11m

Table 3. Detalles de los parámetros ajustados utilizando la técnica de Keras Tuner

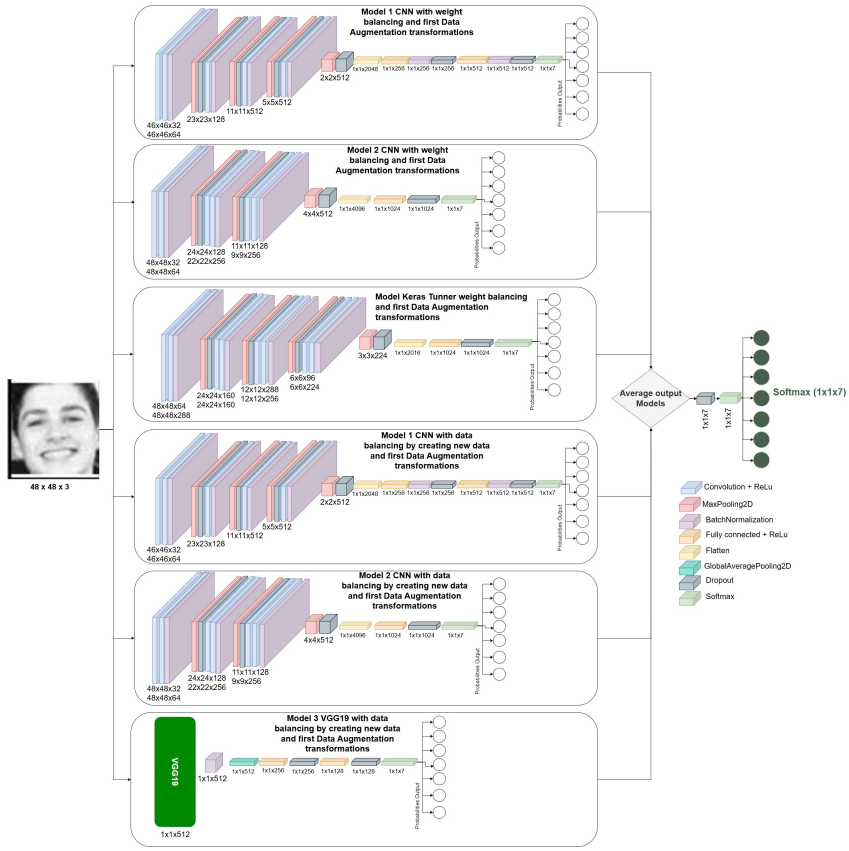


Fig. 7. Arquitectura Propuesta del modelo facial Ensemble usando la concatenación de datasets (FER2013 + CK+48)

En las siguientes tablas se presentarán las arquitecturas que obtuvieron resultados razonables y elevados, seleccionados por razones de espacio en la hoja. En la Tabla 4 y Tabla 5, se muestran las arquitecturas de los modelos experimentados individualmente, y en la Tabla 6, las arquitecturas de los modelos ensemble. Los dos primeras tablas detallan las capas utilizadas, los valores de filtros en cada capa convolucional, el número de neuronas en las capas densas, el optimizador empleado y las técnicas de regularización aplicadas durante el entrenamiento. La tercera tabla detalla los modelos individuales utilizados para el desarrollo de los ensambles, el tipo de ensamble utilizado para la unión, el número de neuronas en las capas densas, el optimizador utilizado y las técnicas de regularización aplicadas durante el entrenamiento. Los modelos incluyen variaciones de arquitecturas como CNN, ResNet, VGG19, ViT, MiniXception, XceptionNet, entre otros. Además, se presentan parámetros claves como el learning rate, los hiperpárametros aplicados en las capas convolucionales, el valor de regularización, entre otros. Esto para facilitar la comparación entre ellos. En el caso de los modelos pre-entrenados, únicamente se indicarán las capas agregadas adicionalmente, omitiendo aquellas que son parte de la arquitectura por defecto. Todos los modelos CNN utilizan un tamaño de kernel de 3x3, ya que este tamaño ha demostrado ofrecer mejores resultados.

Models	Capas	Filtros	Neuronas	Optimizador y regularización
Model ViT Transformer base	Dense	-	7(ouput)	Adam (1e-4)
Model MobileNet	Dense	-	64	Adam (1-e3), Dropout (25%, 50%)
Model XceptionNet	Resizing, Dense	-	2048	Adam(1e-4), Dropout (50%)
Model MiniXceptioNet	Conv2D, SeparableConv2	8x2, 16x3, 32x3, 64x3, 128x3	-	Adam (1e-4), Dropout (25%), L2(0.01)
Model VGG19	Dense	-	256, 128	Adam (1e-4), Dropout (40%), L2 (0.001)
Model Keras Tunner	Conv2D, Dense	64, 288, 160x2, 288, 256, 96, 224	1024	Adam (1e-4), Dropout (20%, 40%, 50%))
CNN-1 (5 CNN + 2 Dense) with weight balancing	Conv2D, Dense	32, 64, 128, 512x2	256, 512	Adam (1e-4), Dropout(25%), L2(0.01)
CNN-2 (6 CNN + 1 Dense)	Conv2D with and without padding (same), Dense	32, 64, 128, 256, 128, 256	1024	Adam(1e-4), Dropout (30%, 50%), L2 (0.01)
CNN-3 (6 CNN + 2 Dense)	Conv2D, Dense	32, 64, 128, 256, 128, 256	2048, 1024	Adam (1e-4), Dropout (30%, 60%), L2 (0.01)
CNN-4 (10 CNN + 1 Dense)	Conv2D with padding(same), Dense	64x4, 128x6	2048	Adam (1e-4), Dropout (25%, 50%), L2 (0.001)

Table 4. Arquitecturas de modelos testeados 1

Models	Capas	Filtros	Neuronas	Optimizador y regularización
CNN-5 (4 CNN + 1 Dense)	Conv2D with padding(same), Dense	256x4	1792	Adam (1e-4), Dropout (25%, 50%)
CNN-6 (4 CNN + 1 Dense)	Conv2D, Dense	32, 64, 128, 256	1024	Adam (1e-4), Dropout (30%, 50%), L2 (0.01)
CNN-7 (8 CNN + 1 Dense	Conv2D with padding(same), Dense	32, 64, 32, 64, 128, 256, 128, 256	1024	Adam (1e-3), Dropout (30%, 50%), L2 (0.01)
CNN-8 (8 CNN + 1 Dense)	Conv2D with padding(same), Dense	64x2, 128x2, 256x2, 512x2	1024	Adam (1e-4), Dropout (30%, 50%), L2 (0.01)
CNN-9 (8 CNN + 1 Dense)	Conv2D, Dense	64x2, 128x2, 256x2, 128x2	512	Adam (1e-4), Dropout (25%, 50%), L2 (0.001, 0.01)
CNN-10 (8 CNN + 1 Dense)	Conv2D with padding (same), Dense	32, 64, 128x2, 256, 128x2, 256	1024	Adam (1e-4), Dropout (30%, 50%)
CNN-11 (8 CNN + 1 Dense)	Conv2D with padding (same), Dense	64x2, 128x2, 256x2, 512x2	1024	Adam (1e-4), Dropout (30%, 50%), L2 (0.01)
CNN-12 (19 CNN + 1 Dense)	Conv2D with padding (same), Dense	64, 32x6, 64x6, 128x6	1024	Adam (1e-4), Dropout (30%, 50%), L2 (0.01)
CNN-13 (4 CNN + 1 Dense)	Conv2D with padding (same), Dense	32, 64, 128, 256	512	Adam (1e-4), Dropout (25%, 50%), L2 (0.01)
CNN-14 (6 CNN + 2 Dense)	Conv2D, Dense	32, 64, 128x2, 256x2	256, 128	Adam (1e-2), Dropout (25%, 50%)

Table 5. Arquitecturas de modelos testeados 2

Ensemble models	Concatenated ensemble models	Unión	Neuronas	Optimizador
Model 1 ensemble	CNN-1, VGG19, CNN-6 (weight balancing) falta	Average	7(output)	Adam (1e-4), Dropout (10%)
Model 2 ensemble	CNN-1, VGG19, CNN-6 (weight balancing)	Average	128	Adam (1e-4), Dropout (50%)
Model 3 ensemble	CNN-1, VGG19, CNN-2 (weight balancing)	Average	7(output)	Adam (1e-4), Dropout (10%)
Model 4 ensemble	CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation)	Voting(0.1, 0.3,0.6)	7(output)	Adam (1e-4), Dropout (10%)
Model 5 ensemble	CNN-1, CNN-2, VGG-19 (data balancing)	Average	7(output)	Adam(1e-4), Dropout (30%)
Model 6 ensemble	CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation) (weight balancing) and CNN-1, VGG19 (data balancing)	Average	7(output)	Adam(1e-4), Dropout (10%)
Model 5 ensemble	CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation) (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	Average	7(output)	Adam(1e-4), Dropout (5%)
Model 6 ensemble	CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-2 (data balancing)	Average	7(output)	Adam(1e-4), Dropout (5%)
Model 7 ensemble	CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, Keras-Tunner (data balancing)	Average	7(output)	Adam(1e-4)
Model 8 ensemble	CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, Keras-Tunner, CNN-2 (data balancing)	Average	7(output)	Adam(1e-4)
Model 9 ensemble	CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, VGG19 (data balancing)	Average	7(output)	Adam(1e-4), Dropout (5%)
Model 10 ensemble	CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing	Average	7(output)	Adam(1e-4)
Model 11 ensemble	CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	Average	7(output)	Adam(1e-4)

Table 6. Arquitecturas de modelos ensemble testeados

Por otro lado, en la modalidad de texto, se experimentó con diversas arquitecturas, incluyendo la replicación y mejora de los modelos previamente investigados. Se probaron modelos utilizando LSTM[25], BiLSTM[22], Embeddings preentrenados como Word2Vec[33] y GloVe[39], así como modelos más avanzados como BERT[16], BERT-CNN (combinación de BERT con capas CNN), BERT-BiLSTM-CNN (combinación de BERT con capas BiLSTM y CNN), DistilBERT[44], XLNet[56], RoBERTA[30], RoBERTA-CNN (combinación de RoBERTA con capas CNN) y RoBERTA-BiLSTM (combinación de RoBERTA con capas BiLSTM). Se destacan cuatro modelos principales. Para el desarrollo del Modelo uno se utilizó el embedding preentrenado de GloVe obtenido oficialmente de su sitio web. Se creó la matriz de embedding mapeando cada palabra tokenizada al correspondiente vector de embedding de GloVe y se agregó el parámetro Trainable en False en la capa de embedding. La salida de esta se conecta a una capa BiLSTM de 64 units con activación tanh, seguida de una capa MaxPooling1D. Posteriormente, se ingresó a una capa Dense de 64 neuronas con activación ReLU, con un Dropout del 50% y un optimizador Adam con learning rate del 0.001. El modelo dos realiza los mismos pasos del modelo uno, pero con la diferencia de ocupar el embedding de Word2Vec preentrenado con "GoogleNews" para ingresar a una arquitectura N-Gram de igual forma que los modelos tres y cuatro. El Modelo tres utiliza el modelo preentrenado de BERT, con sus capas configuradas en trainable False y su respectiva entrada de lista de tensores para obtener la primera salida de BERT, la cual ingresa a una arquitectura N-Gram compuesta por tres capas convolucionales Conv1D. Cada capa recibe como entrada la salida de BERT y posee 128 filtros con activación ReLU y padding de "same", y utiliza tamaños de kernel size de 2 para la primera capa, 3 para la segunda capa y 4 para la tercera capa respectivamente, permitiendo así capturar distintos contextos gramaticales. Posteriormente, se aplica GlobalMaxPooling1D a cada capa convolucional para mantener la dimensionalidad de las salidas. Las salidas de estas capas se concatenan e ingresan a un Dropout de 50%, seguido de una capa Dense de 512 neuronas con activación ReLU, y nuevamente a una capa Dropout de 50%. El modelo utiliza el optimizador Adam con learning rate 1e-4. Finalmente, el modelo propuesto cuatro, como se ve en la Fig. 8, sigue el mismo formato que el Modelo tres, pero empleando RoBERTa como modelo base y tokenizador, que es una versión mejorada y más robusta de BERT. De esta forma se desarrolla la arquitectura RoBERTa-CNN, la cual recibe dos entradas: input_ids y attention_mask del tokenizador RoBERTa, que son ingresadas como lista de tensores al modelo. La salida de RoBERTa se conecta a una arquitectura N-Gram, que consta de tres capas convolucionales Conv1D. Cada capa convolucional recibe la salida del modelo RoBERTa y posee 128 filtros con activación ReLU y padding de "same". Los tamaños del kernel size utilizados son de 2 para la primera capa, 3 para la segunda capa y 4 para la tercera capa, respectivamente. Posteriormente, a cada capa se le aplica GlobalMaxPooling1D para mantener la dimensionalidad de las salidas. Las salidas de estas capas se concatenan e ingresan a un Dropout del 50%, seguido de una capa Dense de 512 neuronas con activación ReLU. Finalmente, antes de la capa de salida, se aplica nuevamente un Dropout del 50%. El modelo utiliza el optimizador Adam con learning rate de 1e-4. Este modelo se desarrolló para predecir cinco y siete salidas respectivamente, permitiendo analizar el rendimiento y predicción de forma diferenciada. Cabe mencionar que este modelo se entrenó durante 20 épocas con un batch size de 128, capas de RoBERTa base configuradas en trainable False y shuffle en True.

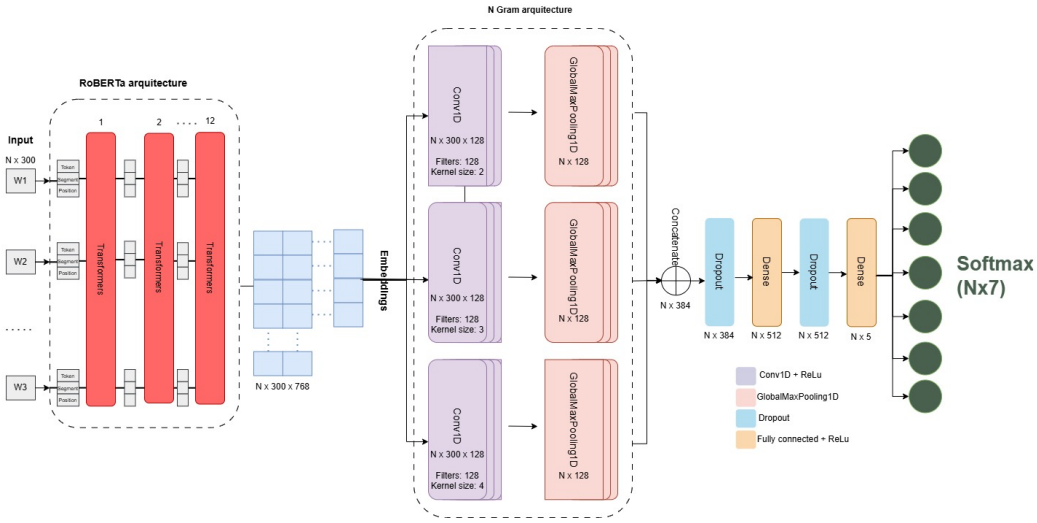


Fig. 8. Arquitectura propuesta del modelo textual

La fórmula de la arquitectura n-grams[14] utilizada en los modelos sería la siguiente:

$$G_h = CNN^h * (T) = [g_1^{\rightarrow h}, ...g_m^{\rightarrow h}]$$

Donde:

– T: es el texto de entrada al cual se le aplica la capa convolucional CNN^h para crear los n-grams. CNN^h: es una capa de convolución específica para n-grams de longitud h, donde h puede ser 1(unigram 1G), 2(bigram 2G), 3(trigram 3G), 4(fourgram 4G) u otra longitud de n-gram según se configure la capa convolucional. G_h: representa el resultado total de aplicar las capas CNN específicas para n-grams de longitud h al texto de entrada T. $[g_1^{\rightarrow h}, ...g_m^{\rightarrow h}]$: representa la lista de embeddings (resultados) individuales generados por capa aplicación específica de CNN^h a diferentes partes del texto T, de forma que cada g tiene un subíndice i y un superíndice $\rightarrow h$ para indicar el n-gram y su longitud.

En las siguientes tablas se presentan las arquitecturas que obtuvieron resultados razonables y elevados en la modalidad textual, siguiendo el mismo enfoque utilizado previamente. En la Table 7, se muestran las arquitecturas de los modelos experimentados utilizando cinco emociones, mientras que en la Table 8 se presentan las arquitecturas de los modelos entrenados con siete emociones, empleando la combinación de los datasets ISEAR y MELD. Cada tabla detalla aspectos claves, como el tipo de capa utilizada (convolucional, recurrente, dense, etc.), el número de filtros, las unidades ocultas de las redes recurrentes, las neuronas de las capas densas, el optimizar seleccionado con su learning rate y las técnicas de regularización empleadas. La referencia "ks" en la columna del número de filtros, se refiere al tamaño de kernel size para cada capa convolucional 1D.

Models	Embedding	Capas	filtros	hidden units	neuronas	optimizar and regularization
GloVe-BiLSTM	glove.6B.300d	Bidirectional, Dense	-	64	64	Adam (1e-3), Dropout (50%)
BERT-BiLSTM-CNN	BERT	BERT, Bidirectional, Conv1D, Dense	32x4 ks:(3,5,5,3)	32x2	512	Adam (1e-3), Dropout (50%)
Word2Vec-CNN	GoogleNews-vectors-negative300	Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v1	BERT	BERT, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v2	BERT	BERT, Conv1D, Dense	64x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v3	BERT	BERT, Conv1D, Dense	64x3 ks: (3,4,5)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v4	BERT	BERT, Conv1D, Dense	64x4 ks: (3,4,5,6)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v5	BERT	BERT, Conv1D, Dense	32x3 ks: (200, 300, 400)	-	512	Adam (1e-4), Dropout (20%, 50%)
BERT-CNN v6	BERT	BERT, Conv1D, Dense	32x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
BERT-CNN v7	BERT	BERT, Conv1D, Dense	32x3 ks: (2,3,4)	-	512	Adam (1e-5), Dropout (50%)
BERT-CNN v8	BERT	BERT, Conv1D, Dense	128x4 ks: (2,3,4,5)	-	512	Adam (1e-4), Dropout (20% each layer, 40%, 50%)

Table 7. Arquitectura de modelos textuales testeados 1

Models	Embedding	Capas	filtros	hidden units	neuronas	optimizar and regularization
RoBERTa-CNN	RoBERTa	RoBERTa, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
RoBERTa-CNN	RoBERTa	RoBERTa, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (20% each layer, 50%)
DistilBERT-CNN	DistilBERT	DistilBERT, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
XLNet-CNN	XLNet	XLNet, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
ALBERT-CNN	ALBERT	ALBERT, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
DeBERTa-CNN	DeBERTa	DeBERTa, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)
ELECTRA-CNN	ELECTRA	ELECTRA, Conv1D, Dense	128x3 ks: (2,3,4)	-	512	Adam (1e-4), Dropout (50%)

Table 8. Arquitectura de modelos textuales testeados 2

Las arquitecturas presentadas se optimizaron con varias configuraciones de hiperparámetros, se optó por el uso de padding en "same" en las capas Conv1D, para conservar las dimensiones originales de las secuencias en tareas de procesamiento de texto, manteniendo así su longitud. Por otro lado, en los modelos que combinan capas BiLSTM y CNN dentro de la misma arquitectura, el padding no mostró mejoras en el rendimiento, ya que las capas recurrentes(como BiLSTM) ya procesan secuencias temporales manteniendo información contextual a lo largo de la secuencia.

Para el entrenamiento de los modelos, se configuraron 300 y 200 épocas para la modalidad facial y entre 10-20 épocas para la modalidad textual. Cada modelo unimodal requirió aproximadamente 3 horas de entrenamiento, utilizando callbacks estratégicos, como EarlyStopping con una paciencia variable entre 10 y 15 épocas, ModelCheckpoint para guardar los mejores modelos, y TensorBoard para monitorear accuracy y loss durante el entrenamiento. Además, se empleó ReduceLROnPlateau exclusivamente para los modelos faciales, reduciendo la tasa de aprendizaje si la pérdida en validación no mejoraba. En ambas modalidades, se probaron configuraciones de batches(32, 64, 128), valores de dropout(0.1, 0.005, 0.001) y optimizadores(Adam, SGD, RMSProp, Adamax) para optimizar el rendimiento.

Para ambas modalidades los modelos utilizaron la función de pérdida categorical cross-entropy:

$$CCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij})$$

La función de activación a ocupar en la salida es Softmax[38] al tener una salida multiclase.

$$Softmax(z_{ij}) = \frac{e^{z_{ij}}}{\sum_{k=1}^{C} e^{z_{ik}}}$$

6 RESULTADOS Y DISCUSIÓN

6.1 Resultados de la clasificación de emociones

En esta sección se van a evaluar los resultados obtenidos por los modelos de clasificación unimodales en ambas modalidades. Las métricas a evaluar son aquellas utilizadas en otros papers relacionados, incluyendo Accuracy, Precision, Recall y F1-Score, calculadas mediante la media no ponderada por etiqueta (macro avg), que permite analizar cada clase con la misma importancia, independiente de su frecuencia, proporcionando una evaluación equilibrada del rendimiento. Además, se incluirán los resultados de la matriz de confusión. También, se presentarán la curva Precision-Recall para la modalidad facial y Perplexity para la modalidad textual, como contribuciones novedosas de este estudio.

A continuación, en la Table 9 se presentan los resultados de los modelos experimentados, incluyendo aquellos modelos con las mejores métricas explicados previamente, así como los modelos propuestos Ensemble, utilizando únicamente el dataset FER2013. Por otro lado, en la Table 10, se muestran los resultados de los mismos modelos que obtuvieron los mejores resultados en la experimentación anterior, con la adición del modelo CNN-3 y de los modelos Ensemble propuestos con mejores métricas, pero esta vez entrenados con la combinación de ambos datasets FER2013 y CK+48. Además, en la Table 11 se presentan los resultados utilizando las dos metodologías adicionales explicadas anteriormente. Asimismo, se incluyen los resultados obtenidos del modelo experimental visual transformer "vit-base-patch16-224" en ambas tablas, Table 9 y Table 10 , dado que estos modelos están siendo ampliamente utilizados en la actualidad, junto con algunos experimentos realizados con modelos pre-entrenados que obtuvieron un desempeño inferior en comparación con los modelos mencionados, así como el modelo obtenido mediante la técnica de Keras Tuner. Los valores representan el promedio de la media no ponderada por etiqueta (macro avg). Todos los resultados fueron evaluados utilizando el conjunto de testeo predefinido del dataset FER2013, ya que son datos no vistos al no ser utilizados en el entrenamiento, no presentan ninguna modificación y son empleados para la evaluación de modelos en tareas de clasificación de emociones.

Model FER2013	Accuracy	Precision	Recall	F1-Score
Model ViT Transformer with weight balancing	51.67%	45%	49%	45%
Model MobileNet with weight balancing	59.50%	56%	59%	57%
Model XceptionNet with weight balancing	60.04%	60%	59%	59%
Model MiniXceptioNet with weight balancing	24.81%	63%	63%	62%
Model VGG19 with weight balancing	64.46%	63%	63%	62%
Model Keras Tunner with weight balancing	66.94%	65%	65%	65%
CNN-1 (5 CNN + 2 Dense) with weight balancing	66.49%	64%	66%	64%
CNN-2 (6 CNN + 1 Dense) with weight balancing and first Augmention	66.56%	64%	65%	64%
CNN-2 (6 CNN + 1 Dense) with weight balancing and second Augmention	66.56%	64%	64%	64%
CNN-3 (6 CNN + 2 Dense) with weight balancing	65.82%	63%	64%	64%
CNN-4 (10 CNN + 1 Dense) with weight balancing	58.24%	53%	58%	53%
CNN-5 (4 CNN + 1 Dense) with weight balancing	54.26%	48%	56%	48%
CNN-6 (4 CNN + 1 Dense) with weight balancing	65.28%	63%	64%	63%
CNN-7 (8 CNN + 1 Dense) with weight balancing	52.50%	45%	53%	45%
CNN-8 (8 CNN + 1 Dense) with weight balancing	52.74%	45%	54%	44%
CNN-9 (8 CNN + 1 Dense) with weight balancing	65.50%	63%	64%	63%
CNN-10 (8 CNN + 1 Dense) with weight balancing	63.80%	59%	64%	60%
CNN-11 (8 CNN + 1 Dense) with weight balancing	64.40%	60%	64%	61%
CNN-12 (19 CNN + 1 Dense) with weight balancing	51.75%	46%	53%	46%
CNN-13 (4 CNN + 1 Dense) with weight balancing	52.28%	47%	53%	47%
CNN-14 (6 CNN + 2 Dense) with weight balancing	57.30%	51%	58%	52%
CNN-1 (5 CNN + 2 Dense) with data balancing	66.99%	69%	64%	65%
CNN-2 (6 CNN + 1 Dense with data balancing	67.62%	67%	65%	66%
VGG19 + 2 Dense with data balancing	66.03%	69%	62%	65%
Model 1 ensemble with CNN-1, VGG19, CNN-6 (weight balancing)	68.15%	67%	67%	66%
Model 2 ensemble with CNN-1, CNN-2, VGG-19 (data balancing)	69.61%	71%	67%	68%
Model 3 ensemble with CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation) (weight balancing) and CNN-1, VGG19 (data balancing)	69.94%	69%	68%	68%
Model 4 ensemble with CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation) (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	70.20%	70%	68%	69%

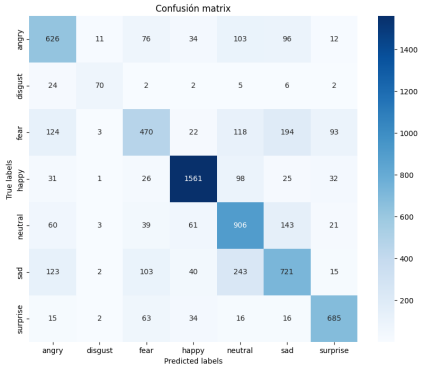
Table 9. Resultados de los modelos experimentados con un único dataset FER2013, junto con los modelos ensemble propuestos. Evaluados en el conjunto de testeo de FER2013

Model FER2013 + CK+48	Accuracy	Precision	Recall	F1-Score
Model ResNet50V2 with weight balancing	54.34%	49%	55%	51%
Model ViT Transformer with weight balancing	56.44%	50%	54%	50%
Model VGG19 with weight balancing	57.82%	54%	57%	54%
Model MobileNet with weight balancing	58.95%	55%	60%	56%
Model XceptionNet with weight balancing	60.35%	59%	60%	59%
Keras Tunner (8 CNN + 1 Dense) with weight balancing	67.62%	65%	67%	66%
Keras Tunner (8 CNN + 1 Dense) with data balancing	67.72%	70%	64%	66%
CNN-1 (5 CNN + 2 Dense) with weight balancing	66.55%	64%	67%	65%
CNN-2 (6 CNN + 1 Dense) with weight balancing	66.08%	63%	66%	64%
CNN-3 (6 CNN + 2 Dense) with weight balancing	65.88%	63%	65%	64%
CNN-1 (5 CNN + 2 Dense) with data balancing	68.75%	71%	66%	67%
CNN-2 (6 CNN + 1 Dense) with data balancing	67.46%	69%	63%	65%
VGG19 + 2 Dense with data balancing	66.14%	68%	63%	65%
Model 1 ensemble with CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-2 (data balancing)	68.65%	67%	68%	67%
Model 2 ensemble with CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, Keras-Tunner (data balancing)	69.82%	69%	69%	69%
Model 3 ensemble with CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, Keras-Tunner, CNN-2 (data balancing)	69.90%	70%	69%	69%
Model 4 ensemble with CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, VGG19 (data balancing)	70.13%	70%	69%	69%
Model 5 ensemble with CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	70.26%	70%	69%	69%
Model 6 ensemble with CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	70.36%	70%	69%	70%

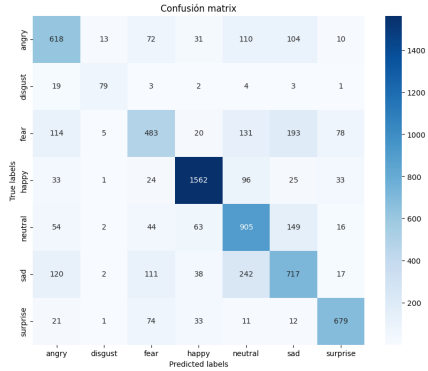
Table 10. Resultados de los modelos experimentados con la combinación de los datasets FER2013 y CK+48, junto con los modelos ensemble propuestos. Evaluados en el conjunto de testeo de FER2013

Additional methodologies developed	Accuracy	Precision	Recall	F1-Score
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. CNN-1 (5 CNN + 2 Dense) with weight balancing	65.47%	63%	65%	63%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. CNN-2 (6 CNN + 1 Dense) with weight balancing	66.98%	65%	66%	65%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. CNN-3 (6 CNN + 2 Dense) with weight balancing	66.23%	64%	66%	64%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. CNN-1 (5 CNN + 2 Dense) with data balancing	66.55%	64%	67%	65%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. CNN-2 (6 CNN + 1 Dense) with data balancing	66.55%	63%	66%	64%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. VGG19 + 2 Dense with data balancing	65.88%	63%	65%	64%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. Keras Tunner (8 CNN + 1 Dense) with data balancing	67.72%	70%	64%	66%
FER2013 + CK+48 with train-validation split, along with data cleaning and clean. Keras Tunner (8 CNN + 1 Dense) with weight balancing	67.62%	65%	67%	66%
Model FER2013 + CK+48 with train-validation split, along with data cleaning. Ensemble 1 with CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	69.54%	70%	67%	68%
Model FER2013 + CK+48 with train-validation split, along with data cleaning. Ensemble 2 with CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	69.70%	70%	68%	68%
Model FER2013 + CK+. CNN-1 (5 CNN + 2 Dense) with weight balancing	66.88%	65%	66%	65%
Model FER2013 + CK+. CNN-2 (6 CNN + 1 Dense) with weight balancing	66.21%	64%	65%	64%
Model FER2013 + CK+. CNN-3 (6 CNN + 2 Dense) with weight balancing	66%	63%	65%	64%
Model FER2013 + CK+. CNN-1 (5 CNN + 2 Dense) with data balancing	67.87%	70%	65%	66%
Model FER2013 + CK+. CNN-2 (6 CNN + 1 Dense) with data balancing	67.23%	69%	64%	66%
Model FER2013 + CK+. VGG19 with data balancing	65.28%	69%	61%	63%
Model FER2013 + CK+. Ensemble 1 with CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing)	69.16%	70%	67%	68%

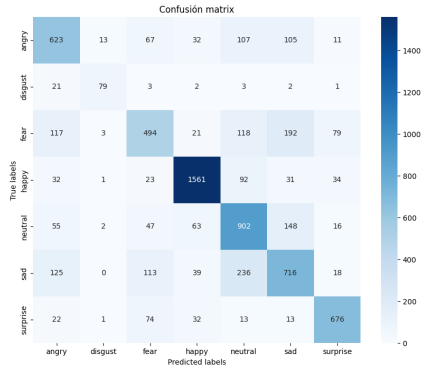
Table 11. Resultados de modelos faciales con metodologías adicionales desarrolladas



(a) Model 4 ensemble - CNN-1, CNN-2 (first augmentation), CNN-2 (second augmentation) (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing) with one dataset



(b) Model 5 ensemble - CNN-1, CNN-2, CNN-3 (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing) with two dataset



(c) Model 6 ensemble - CNN-1, CNN-2, Keras-Tunner (weight balancing) and CNN-1, CNN-2, VGG19 (data balancing) with two datasets

Fig. 9. Comparación de matrices de confusión en el conjunto de testeo de los modelos Ensemble usando uno y dos conjuntos de datos en el entrenamiento

Como se puede observar en la Table 9 y en la Table 10, los modelos ensemble lograron mejores resultados en comparación con los modelos individuales experimentados, destacando la potencia de la combinación de diferentes modelos y métodos de entrenamiento, reduciendo las limitaciones individuales. Por otro lado, en la Table 11, se evidencia que las metodologías adicionales, como la integración de CK+48 posterior a la división de los conjuntos y la experimentación con CK+ en la combinación, no superaron los resultados obtenidos con las dos metodologías propuestas inicialmente. Además, un análisis más detallado que compara las matrices de confusión entre el modelo ensemble propuesto entrenado con un solo dataset, y los dos siguientes modelos ensemble propuestos entrenados con la combinación de dos datasets en el conjunto de entrenamiento, seleccionados por ser los que obtuvieron los mejores resultados, se puede observar en la Fig. 9. Los resultados evidencian que los modelos entrenados con dos datasets (FER2013 + CK+48) obtuvieron mejores resultados en todas las métricas claves. En

particular, el tercer modelo ensemble denominado "Modelo 6" en la Table 10, que incorpora un modelo individual Keras Tuner en la unión y se entrena con la combinación de los dos datasets, demostró una mejor generalización en el conjunto de testeo FER2013. Este modelo alcanzó una mejora en accuracy del 0.10%, logrando un 70.36%, y un aumento del 1% en F1-Score, logrando un 70%, respecto al segundo mejor modelo ensemble obtenido. Además, el Modelo ensemble 6 mostró una mejora en la matriz de confusión, con un total de 12 predicciones correctas adicionales en comparación con el modelo propuesto entrenado con un único dataset, y un aumento de 8 predicciones correctas adicionales en comparación con el segundo modelo ensamble propuesto (denominado "Modelo 5 ensemble" en la Table 10), que también utilizó la combinación de dos datasets para entrenar. Estos resultados demuestran el potencial de la integración del modelo Keras Tuner en el modelo ensamble, lo que permitió mejorar las métricas generales y la capacidad de generalización. Es importante señalar que esta mejora gracias a Keras Tuner solo se observa en los modelos que emplean la combinación de dos datasets para entrenar. Asimismo, la técnica Keras Tuner y el optimizador Bayesian Optimization permitió ajustar los hiperparámetros y encontrar un modelo óptimo para ser introducido en el ensemble, complementándose adecuadamente y consiguiendo una mejora en las métricas. Por ende, analizando las métricas calculadas en el macro avg y el aumento en las predicciones correctas, el modelo Ensemble 6 analizado y entrenado con la combinación de los datasets FER2013 y CK+48, muestra mejor generalización. Además, el progreso del modelo entre épocas y cómo aprende, se puede visualizar en la Fig. 10, donde se evidencia cierto overfitting en ambos conjuntos de validación y entrenamiento, debido a la diferencia entre los resultados obtenidos en el conjunto de validación y entrenamiento en comparación con las métricas resultantes del conjunto de testeo. Esto puede deberse a que, al combinar ambos datasets, algunos datos de CK+48 ingresan en el conjunto de validación al realizar la división del conjunto de entrenamiento para generar los conjuntos de train y val, respectivamente. Sin embargo, incluso al realizar la división antes de combinar los datasets, no se observa una mejora en comparación, posiblemente porque la combinación previa ayudó a evitar que cierto ruido ingrese al conjunto de entrenamiento. Además, dado que los modelos no son deterministas, es posible que los resultados puedan variar en cada ejecución. Aún así, la curva de pérdida disminuye de manera constante, lo cual indica que el modelo está aprendiendo. Es importante señalar que, en comparación con los modelos existentes actuales investigados, el problema del overfitting es común en aquellos modelos con mejores métricas y que no presentan sesgo hacia la clase mayoritaria. Esto sugiere que la robustez necesaria para que los modelos generalicen adecuadamente los datos conlleva inevitablemente a cierto overfitting, como se refleja en los resultados sobre el conjunto de testeo y validación.

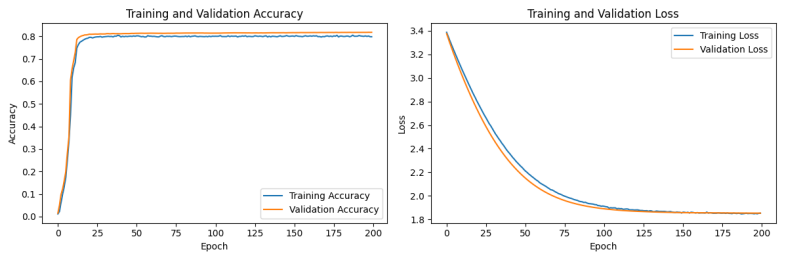


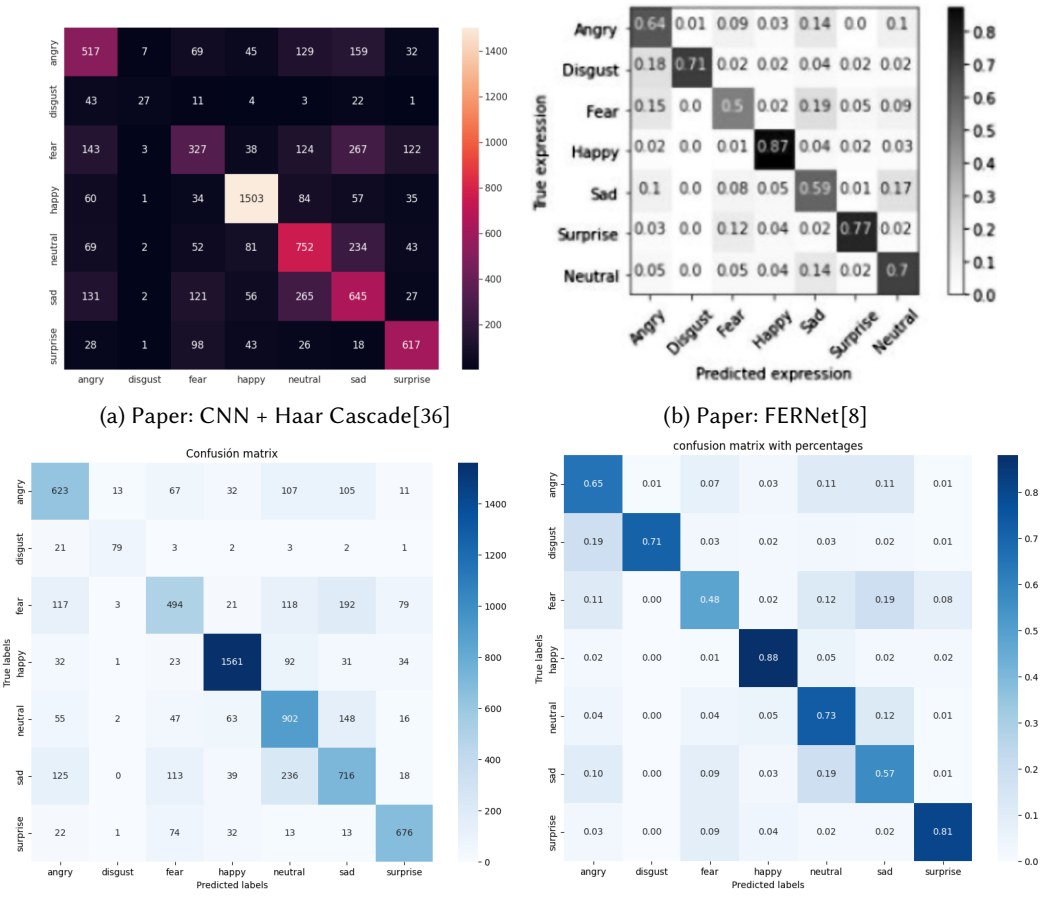
Fig. 10. Gráfico de Accuracy y Loss de los conjuntos de entrenamiento y validación

Actualmente los estudios realizados en el campo de la modalidad facial, utilizando el mismo conjunto de testeo del dataset FER2013, se presentan en la Table 12 con sus respectivas métricas obtenidas.

Models	Accuracy	Recall	F1-Score
VIT-B/16/S[12]	54.84%	52.25%	51.84%
SVM[13]	56.9%	56.9%	56.6%
VGG19[32]	65.41%	65%	62%
EduVit[51]	66.51%	-	54.60%
CNN-10 layers[43]	68.34%	-	-
FerNet[8]	69.57%	-	-
CNN + Haar Cascade[36]	70%	65%	66%
Proposed ensemble model 1	**70.20%**	**68%**	**69%**
Proposed ensemble model 2	**70.26%**	**69%**	**69%**
Proposed ensemble model 3	**70.36%**	**69%**	**70%**

Table 12. Comparación de los modelos de modalidad facial frente a otros trabajos en el conjunto de testeo FER2013

En comparación con otros trabajos, los tres modelos propuestos superan todas las métricas en relación al mejor modelo actual, CNN + Haar Cascade. Específicamente, al analizar las métricas macro avg, que permite evaluar cada clase con la misma importancia independiente de su frecuencia. El tercer modelo ensemble propuesto (Model 6 ensemble en la Table 10) logra un incremento del 0.36% en Macro Accuracy, 4% en Macro Recall y 4% en Macro F1-Score, lo cual permite una comparación más justa contra modelos que presentan métricas sesgadas hacia la clase mayoritaria. Además, un análisis más detallado se puede observar en la matriz de confusión, la cual permite visualizar cómo el modelo clasifica cada clase de manera individual.



(a) Paper: CNN + Haar Cascade[36]

(b) Paper: FERNet[8]

(c) Proposed ensemble model

(d) Proposed ensemble model with percentages

Fig. 11. Comparación de matrices de confusión con otros trabajos

Como se puede visualizar en la matriz de confusión de la Fig. 11 del tercer modelo ensemble, que obtuvo las mejores métricas, donde las filas corresponden a la clases reales y las columnas a la clases predichas. Se evidencia que el modelo propuesto logra una mejora en comparación a los dos paper previos con resultados más altos. Por un lado, se observa que el modelo propuesto consigue una mejora en todas las clases con respecto al modelo con mejor accuracy CNN + Haar Cascade, aumentando 106 predicciones correctas adicionales en la clase Angry, 52 en Disgust, 167 en Fear, 58 en Happy, 150 en Neutral, 71 en Sad y 69 en Surprise. Además, el modelo propuesto también mejora en comparación con el modelo FERNet que tenía la mejor matriz de confusión, mejorando en cuatro clases, un 1% en Angry y Happy, 3% en Neutral y 4% en Surprise. Se mantiene en la clase Disgust con un 71%, pero disminuye un 2% en las clases Fear y Sad. De manera que, el modelo propuesto demuestra una mayor capacidad de generalización con datos reales, lo cual se refleja en las mejoras obtenidas en las métricas. Esta mejora se logró mediante el entrenamiento con las clases balanceadas, asegurando que el modelo pueda generalizar correctamente cada una de ellas, a diferencia de los modelos de trabajos anteriores, donde no se específica el uso de técnicas de balanceo y que podría haber

generado sesgos hacia la clase mayoritaria. Además, la incorporación de una mayor cantidad de datos de entrenamiento mediante la combinación de dos datasets demuestra que aún existe espacio de mejora en la clasificación de emociones faciales. Como también, el desarrollo de un modelo ensemble más complejo y robusto ha sido clave en los resultados positivos.

Al analizar la matriz de confusión, se observa que las emociones "Fear" y "Sad" son las que presentan una menor precisión a la hora de clasificar, a menudo confundiéndose con emociones como "Sad" y "Neutral", lo que indica que estas emociones representan un desafío para el modelo, posiblemente debido al ruido presente en las muestras y a etiquetas que no reflejan correctamente la emoción real. En contraste, las emociones "Happy", "Neutral" y "Surprise" son las clases clasificadas con mayor precisión en la realidad. Además, se puede observar un claro desbalance en el conjunto de testeo, donde la emoción "Disgust" cuenta con la menor cantidad de datos en comparación a la emoción "Happy", que presenta una volumen considerablemente mayor, lo cual subraya la importancia del balanceo propuesto durante el entrenamiento para mejorar la generalización del modelo. El ruido presente en el dataset se logra visualizar en la Fig. 12, obtenido a través de un análisis individual de cada imagen en el conjunto de testeo del dataset FER2013, donde se confirma que las emociones con menores predicciones correctas por parte del modelo presentan problemas en la calidad de las etiquetas. Como se evidencia en la matriz de confusión del modelo propuesto (Fig. 11), se observa que la emoción "Sad" se confunde mayormente con "Neutral". Esto se corrobora en la Fig. 12, que muestra imágenes etiquetadas como "Sad" dentro del dataset, pero que son visualmente muy similares a la emoción "Neutral". Este mismo patrón se observa en la emoción "Fear" que se confunde en su mayoría con "Sad" y "Neutral". Al examinar las imágenes de esta clase, se evidencia que existen rasgos característicos de personas llorando, con expresiones tristes y neutrales, lo que implica que el modelo confunda dichas categorías. Este análisis refuerza la idea de que existe un cierto nivel de ruido en el conjunto de datos FER2013, un problema que ya ha sido señalado en otros estudios. Este ruido impacta en los resultados obtenidos tanto en este trabajo como en investigaciones previas, y es un factor clave que influye en las métricas de rendimiento de los modelos evaluados.

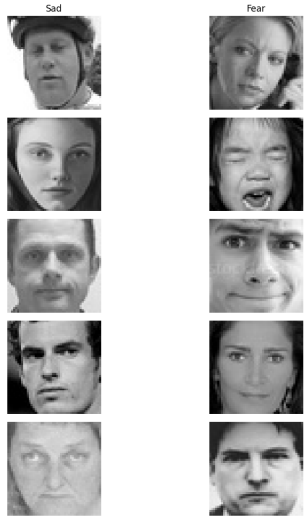


Fig. 12. Ruido existente en las clases tristeza y miedo

Además, dada la naturaleza desbalanceada del conjunto de datos, especialmente con la clase minoritaria "Disgust", que representa un porcentaje muy reducido de las instancias, se incluye en la Fig. 13 una curva precisión-recall. Esta curva permite una evaluación más detallada del rendimiento del modelo en esta clase específica, junto con la métrica de precisión promedio que resume su desempeño global.

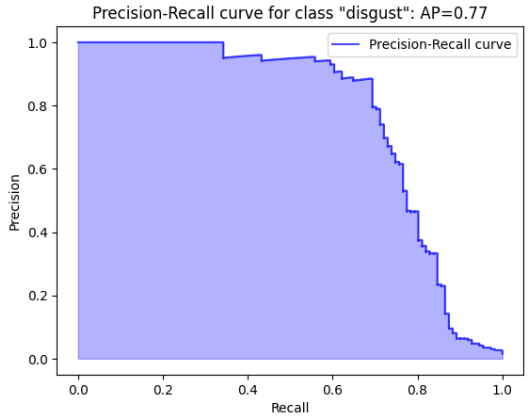
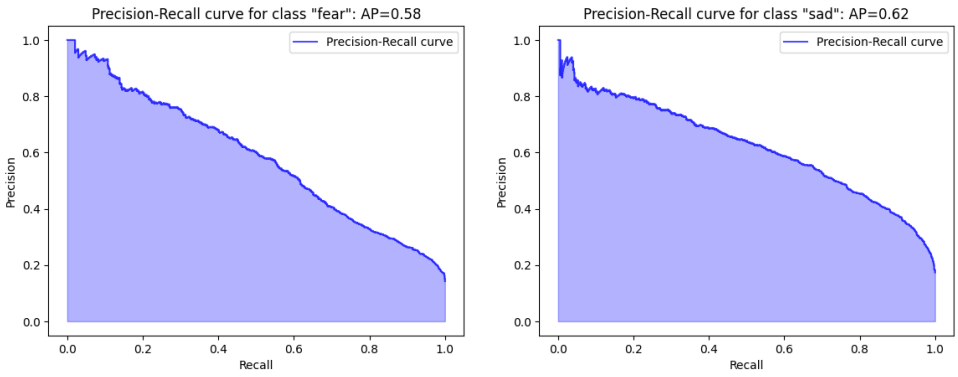


Fig. 13. Curva de Precision-Recall para la clase Disgusto

La curva muestra que el modelo tiene una precisión promedio de 0.77 para la clase "Disgust". Se observa que el modelo comienza con una alta precisión en el extremo izquierdo, lo cual indica una alta confianza en sus predicciones iniciales y a medida que el recall aumenta y el modelo empieza a incluir más falsos positivos, como es esperado, la precisión disminuye gradualmente, demostrando que existe un buen equilibrio entre precisión y recall.

También, se incluyen las curvas de precision-recall para las clases "Fear" y "Disgust", ya que, como se mencionó anteriormente, son las clases con menor precisión a la hora de predecir datos no vistos pertenecientes a estas categorías. Para profundidad en el análisis de su rendimiento, se incluyen ambas figuras en la Fig. 14.



(a) Precision-Recall "Fear"

(b) Precision-Recall "Sad"

Fig. 14. Curva de Precision-Recall para las clases tristeza y miedo

Para la clase "Fear", el modelo presenta una precisión promedio de 0.58. Al analizar la curva correspondiente, se observa una tendencia similar a la clase previamente analizada "Disgust", mostrando que el modelo es más preciso al identificar un número reducido de ejemplos. Sin embargo, a medida que el recall aumenta, la precisión disminuye, generando un mayor número de falsos positivos. Una situación similar ocurre con la clase "Sad", la cual, obtuvo una precisión promedio de 0.62. Aunque esta clase tiene una precisión ligeramente superior, de igual forma disminuye a medida que el recall incrementa. Estos comportamientos están relacionados con dicho ruido mencionado y analizado previamente, donde al aumentar la cantidad de muestras, el modelo se comienza a confundir con las emociones "Neutral" y "Sad", debido a la similitud visual entre emociones y a cierto problema de calidad en las etiquetas, clasificando correctamente las emociones más representativas. Una posible solución para abordar este problema es ajustar el umbral de decisión (threshold) del modelo, lo cual puede contribuir a reducir el número de falsos positivos y, a su vez, mejorar la precisión. Sin embargo, es importante tener en cuenta que puede disminuir el recall, ya que el modelo podría clasificar incorrectamente más ejemplos como negativos, omitiendo algunas instancias verdaderas de las emociones "Fear" o "Sad".

Para la modalidad de texto, en la Table 13, se presentan los resultados obtenidos de los modelos experimentados utilizando 5 emociones. De manera similar, en la Table 14 se muestran los resultados utilizando 7 emociones, incluyendo los mejores modelos obtenidos previamente, así como los destacados en ambas tablas mencionados anteriormente por separado. También se incorpora el modelo propuesto RoBERTa-CNN. Además, en la Table 15 se presentan los resultados de las dos metodologías experimentadas adicionalmente, siendo la primera aplicando lematización a los datos, y la segunda, implica la agregación de un mayor volumen de datos con el dataset IEMOCAP[10]. Los valores representan el promedio de la media no ponderada por etiqueta (macro AVG) del reporte de clasificación, el cual representa el desempeño del modelo considerando todas las clases por igual, independiente de su frecuencia.

Models 5 emotions	Accuracy	Precision	Recall	F1-Score
Model 1 GloVe-BiLSTM - 64 units - 64 batch	68.61%	69%	69%	69%
Model 2 BERT-BiLSTM-CNN - 32 filters and units - 128 batch	75%	75%	75%	75%
Model 3 Word2Vec-CNN 128 filters - 2,3,4 kernel - 128 batch	71.53%	72%	72%	71%
Model 4 BERT-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	78.10%	78%	78%	78%
Model 5 BERT-CNN 64 filters - 2,3,4 kernel	75.91%	76%	76%	76%
Model 6 BERT-CNN 64 filters - 3,4,5 kernel	77.18%	77%	77%	77%
Model 7 BERT-CNN 32 filters - 200,300,400 kernel	71.89%	72%	72%	72%
Model 8 BERT-CNN 32 filters - 2,3,4 kernel	76.82%	77%	77%	77%
Model 9 BERT-CNN 64 filters - 3,4,5,6 kernel	78.10%	78%	78%	78%
Model 10 BERT-CNN 128 filters - 2,3,4,5 kernel	77.73%	77%	78%	78%
Model 11 BERT-CNN 64 filters - 2,3,4,5 kernel - dropout 0.2 each layer	77.55%	77%	78%	77%
Model 12 BERT-CNN 128 filters - 2,3,4,5 kernel - dropout 0.2 each layer	78.10%	78%	78%	78%
Model 13 BERT-CNN 32 filters - 2,3,4,5 kernel - dropout 0.2 each layer	78.10%	78%	78%	78%
Model 14 BERT-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-5 lr	60.58%	62%	61%	60%
Model 15 BERT-CNN 32 filters - 2,3,4 kernel - 1e-5 lr	64.05%	64%	64%	64%
Model 16 BERT-CNN 128 filters - 2,3,4 kernel - 64 batch - 1e-4 lr	78.28%	78%	78%	78%
Model 17 BERT-CNN 128 filters - 2,3,4 kernel - 32 batch - 1e-4 lr	77.18%	77%	77%	77%
Model 18 BERT-CNN 256 filters - 2,3,3 kernel - 64 batch - 1e-4 lr	77.73%	78%	78%	78%
Model 19 BERT-CNN 256 filters - 2,3,4 kernel - 64 batch - 1e-4 lr	76.45%	76%	76%	76%
Model 20 RoBERTa-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	78.64%	79%	79%	79%
Model 21 RoBERTa-CNN 128 filters - 2,3,4 kernel - 64 batch - 1e-4 lr	77.73%	78%	78%	78%
Model 22 RoBERTa-CNN 128 filters - 2,3,4 kernel - 32 batch - 1e-4 lr	77%	77%	77%	77%

Table 13. Resultados de los modelos experimentados con 5 emociones. Evaluados en el conjunto de testeo

Models 7 emotions	Accuracy	Precision	Recall	F1-Score
Model 1 GloVe-CNN - 128 filters - 2,3,4 kernel - 128 batch	69.23%	69%	69%	69%
Model 2 Word2Vec-CNN 128 filters - 2,3,4 kernel - 128 batch	71.96%	72%	72%	72%
Model 3 RoBERTa-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	79.92%	81%	80%	80%
Model 4 BERT-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	80.18%	80%	80%	80%
Model 5 RoBERTa-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - dropout 0.2 each layer	79.79%	80%	80%	80%
Model 6 RoBERTa-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	80.70%	81%	81%	81%
Model 7 DistilBERT-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	79.79%	80%	80%	80%
Model 8 XLNet-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	72.88%	73%	73%	73%
Model 9 ALBERT-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	72.75%	73%	73%	73%
Model 10 DeBERTa-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	80.18%	80%	80%	80%
Model 11 ELECTRA-CNN 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr - doble clean	67.66%	68%	68%	68%

Table 14. Resultados de los modelos experimentados con 7 emociones. Evaluados en el conjunto de testeo

Additional methodologies developed	Accuracy	Precision	Recall	F1-Score
7 emotions with Lemmatization. RoBERTa-CNN - 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	79.40%	79%	79%	79%
7 emotions with 3 datasets(ISEAR,MELD,IEMOCAP). RoBERTa-CNN - 128 filters - 2,3,4 kernel - 128 batch - 1e-4 lr	70.80%	71%	71%	71%

Table 15. Resultados de modelos textuales con metodologías adicionales desarrolladas

Como se puede ver en la Table 13 y Table 14, el modelo propuesto RoBERTa-CNN obtiene mejores resultados en todas las métricas en comparación con los demás modelos experimentados, alcanzando un accuracy del 78.64% con 5 emociones y 80.70% con 7 emociones. Estos resultados destacan la potencia del modelo preentrenado RoBERTa junto con la arquitectura N-gram, una combinación que ha demostrado ser muy potente y efectiva para la clasificación de emociones en la modalidad textual. Además, se observa que las metodologías adicionales, como la lematización y la combinación de un tercer modelo, no superan los resultados obtenidos con la clasificación de 7 emociones. En la Fig. 15, se muestra el progreso del modelo propuesto RoBERTa-CNN para 7 emociones a lo largo de las épocas. Para la selección del mejor modelo, se utilizó un checkpoint en la época 14, donde se observa una diferencia mínima entre el conjunto de validación y entrenamiento, eliminando posibles problemas de overfitting en épocas posteriores. Además, se refleja en la curva de pérdida cómo el modelo efectivamente sigue aprendiendo correctamente.

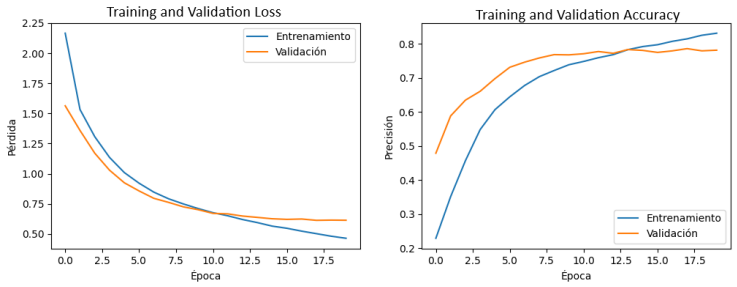


Fig. 15. Gráfico de Accuracy y Loss de los conjuntos de entrenamiento y validación

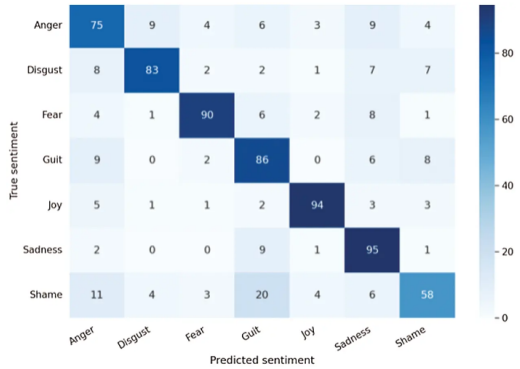
Actualmente, los estudios realizados en el campo de la modalidad textual con el mismo dataset ISEAR se pueden ver en la Table 16, junto con sus respectivas métricas obtenidas.

Models	Accuracy	Recall	F1-Score
GRU[57]	60.26%	60.26%	59.87%
BERT-BiLSTM[3]	74%	73%	73%
RoBERTa Base[2]	74.31%	-	-
Mistral 7B Fine-tuned[18]	76%	-	-
BERT-CNN[4]	77%	76%	76%
Proposed model RoBERTa + Ngram-CNN 5 emotions	**79%**	**79%**	**79%**
Proposed model RoBERTa + Ngram-CNN 7 emotions	**80.70%**	**81%**	**81%**

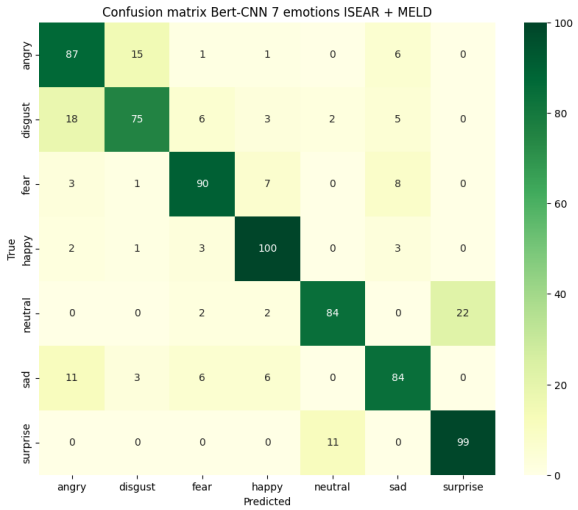
Table 16. Comparación de los modelos de modalidad textual con otros trabajos utilizando el dataset ISEAR

En comparación con otros trabajos, se observa que los dos modelos propuestos lograron una mejora. En particular, el modelo propuesto textual, utilizando 7 emociones, logra mejorar en todas las métricas claves, superando en un 3.70% el Macro Accuracy, 5% el Macro Recall y 5% el Macro F1-Score al mejor modelo anterior, BERT-CNN[4]. Esto evaluando las métricas de la media no ponderada por etiqueta (macro avg), que permite analizar cada clase con la misma importancia, independiente de su frecuencia. Además, dado que existen dos clases distintas, una forma de visualizar las predicciones del modelo es mediante la matriz de confusión, que se puede ver en la Fig. 16.

Como se puede visualizar en la matriz de confusión de la Fig. 16, del modelo propuesto, donde las filas corresponden a las clases reales y las columnas a las clases predichas, se evidencia que el modelo propuesto logra una mejora y una igualación en 3 de las 5 clases igualitarias, aumentando 12 predicciones correctas en la clase "Angry", igualando la cantidad predicha en la clase "Fear" y aumentando 6 predicciones correctas en la clase "Happy". Sin embargo, para las clases "Disgust" y "Sad" se observa una disminución en las predicciones correctas. Estas distribuciones resaltan las diferencias en las predicciones incorrectas entre ambos modelos, mostrando cómo cada uno maneja las clases de manera distinta, siendo este un aporte hacia una nueva arquitectura en el ámbito educativo.



(a) BERT-CNN



(b) Proposed RoBERTa-NGram-CNN model

Fig. 16. Comparación de matrices de confusión frente a otro trabajo

Por otro lado, al analizar la matriz de confusión, se observa que la emoción "Disgust" presenta la menor cantidad de predicciones correctas, confundiéndose sobre todo con las emociones "Angry" y "Fear", lo que indica que esta emoción presenta un desafío para el modelo, posiblemente debido al ruido existente en el dataset. Dado que los datos fueron etiquetados a partir de entrevistas en diferentes partes del mundo, es posible que no todas las personas se expresen de la misma manera. Por otro lado, al analizar las emociones incorporadas como "Neutral" y "Surprise", que son relevante en el contexto educativo, se observan buenos resultados, sobre todo en la emoción "Surprise" demostrando el potencial del modelo para ser incorporado en el contexto educativo.

Este rendimiento superior en el modelo propuesto puede atribuirse a la capacidad de RoBERTa para capturar contextos más complejos y matices en el lenguaje, combinado con la arquitectura N-Gram-CNN, que permite capturar diferentes niveles de información gramatical y semántica.

Asimismo, el uso de capas convolucionales con distintos tamaños de kernel proporciona una mayor flexibilidad para identificar patrones relevantes en los datos textuales.

Además de métricas estándar, este estudio se distingue por integrar Perplexity como medida adicional para evaluar la coherencia y calidad de las predicciones del modelo de reconocimiento de emociones basado en texto. En este contexto, un valor de Perplexity más cercano a 0 indica una mejor capacidad del modelo para generar secuencias de texto gramaticalmente correctas y coherentes.

Perplexity	Value
Proposed model RoBERTa-NGram-CNN 7 emotions	4.402052

Table 17. Métrica textual perplexity

La Perplexity indica la incertidumbre del modelo al predecir secuencias de texto, reflejando su capacidad para generar expresiones gramaticalmente correctas y emocionalmente coherentes. Como se observa en la Table 17, el modelo propuesto, RoBERTa-NGram-CNN, logra un valor de Perplexity de 4.402, lo que sugiere una buena fluidez en la generación de texto emocionalmente relevante y precisa, indicando que las predicciones del modelo están muy cercanas a las secuencias reales.

La fórmula de Perplexity se expresa como:

$$\text{PPL}(X) = \exp\left(-\frac{1}{T}\sum_{i=1}^{T}\log p_\theta(x_i \mid x_{<i})\right)$$

Donde $\text{PPL}(X)$ es la perplexity para la secuencia, T es la longitud de la secuencia, x_i representa el token en la posición i de la secuencia X, $x_{<i}$ representa todos los tokens anteriores al token x_i y $p_\theta(x_i \mid x_{<i})$ es la probabilidad predicha por el modelo para el token x_i dada la secuencia anterior $x_{<i}$.

Finalmente, habiendo logrado resultados favorables en ambas modalidades (facial y textual). Se propone como futura mejora la creación de un modelo multimodal, mediante la integración del modelo facial ensemble propuesto con accuracy de 70.36% y el modelo propuesto RoBERTa-CNN con accuracy del 80.70% en la modalidad textual. Esta combinación se realizaría utilizando la concatenación de ambos modelos con la técnica "late fusion"[50], en la cual las salidas de los modelos unimodales preentrenados se combinarán de forma igualitaria antes de realizar la predicción final. La unión de ambos modelos se llevaría a cabo de manera que el modelo multimodal cuente con tres entradas, una para la modalidad facial, con un tamaño de 48x48x3, correspondiente a las imágenes de entrada, y dos para la modalidad textual, input_ids y attention_mask con una longitud de secuencia de 300 tokens. Posteriormente, las salidas de ambos modelos se concatenan dentro de la arquitectura del modelo, combinando las características extraídas por cada modelo antes de realizar la predicción final. Además, se aplicaría el preprocesamiento necesario para que las imágenes y texto puedan ser ingresados al modelo. La propuesta de la arquitectura se muestra en la Fig. 17. Este modelo generaría una única salida de siete emociones, que representa la emoción experimentada, integrando el rostro facial y el contenido textual. La creación del modelo multimodal, permite aprovechar las fortalezas de ambos modelos unimodales, aportando un modelo innovador particularmente en el ámbito educativo, permitiendo a los docentes obtener una visión más

completa y detallada de los estudiantes. Como también, la posibilidad de utilizar dos modelos propuestos de forma individual, en caso de requerir un análisis específico por modalidad.

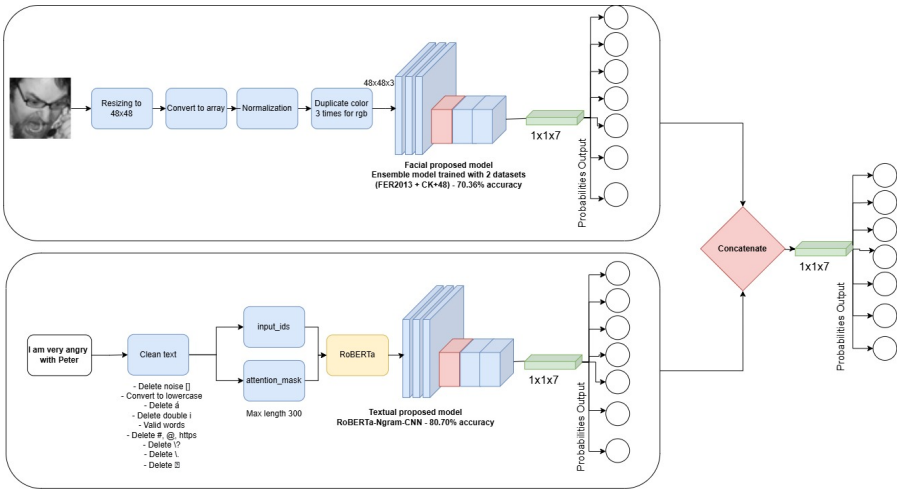


Fig. 17. Modelo propuesto multimodal

7 CONCLUSIONES

En conclusión, en este trabajo se han propuesto dos arquitecturas innovadoras para la modalidad textual y facial, cumpliendo así el primer objetivo específico de diseñar y desarrollar modelos unimodales de reconocimiento de emociones tanto facial como textual, los cuales han superado las métricas actuales en ambas modalidades. Para la modalidad facial, se emplearon redes ensemble, que demostraron ser eficaces y poderosas al complementar las fortalezas de cada modelo individual y mitigar sus desventajas. Además, se propuso el uso del balanceo de datos, una técnica no abordada en estudios previos, la cual se implementó con la combinación de dos metodologías diferentes de balanceo de datos, pesos ponderados y balanceo con Data Augmentation. Se demostró que el balanceo de datos mediante el ingreso con Data Augmentation mejora la capacidad predictiva, aunque aumenta el riesgo de overfitting, mientras que el balanceo de datos mediante pesos ponderados reduce esta posibilidad a costa de un menor rendimiento. La unión de ambas estrategias permitió crear un modelo ensemble que superó las métricas de los estudios anteriores y eliminó sesgos hacia las clases mayoritarias. Por su parte, en la modalidad textual, el modelo propuesto RoBERTa con arquitectura N-Gram-CNN mostró un rendimiento superior en comparación a otros modelos, aprovechando las capacidades de las redes preentrenadas como RoBERTa y la arquitectura N-gram para capturar diversos contextos gramaticales. Este desarrollo supone un avance significativo hacia el contexto educativo, introduciendo al dataset nuevas emociones comúnmente experimentadas en entornos educativos, permitiendo entregar un enfoque más detallado al docente. Posteriormente, al comparar los modelos, se cumplió con el segundo objetivo específico, que consiste en evaluar y comparar los modelos unimodales con investigaciones previas. Los resultados en la modalidad facial demostraron que la combinación de los dos datasets comúnmente utilizados en el reconocimiento de emociones faciales, FER2013 y CK+48 en los datos de entrenamiento, permitió que el modelo aprenda de un mayor número de muestras y reconocer expresiones de distintos rostros y escenarios, lo que añade un mayor nivel de robustez en

comparación con entrenarlo con un solo dataset. Como resultado, se logró un aumento en la matriz de confusión de 655 predicciones correctas en total en comparación al mejor estudio anterior, obteniendo en el conjunto de testeo predefinido de FER2013 un Macro Accuracy del 70.36%, un Macro Recall del 69% y un Macro F1-Score del 70%, destacando la potencia del modelo y la capacidad de generalización a nuevos datos. Además, se aportó el gráfico de Precision-Recall, que permite evaluar de manera más detallada la clase minoritaria "Disgust" en el conjunto de testeo, alcanzando un promedio de 0.77, lo que demuestra que el modelo posee un buen equilibrio entre precisión y recall para dicha clase. En cuanto a la modalidad textual, el modelo propuesto RoBERTa-NGram-CNN logró un aumento del 4% en métricas claves en comparación a trabajos anteriores, obteniendo un 80.70% en Macro Accuracy, 81% en Macro Precision, 81% en Macro Recall y 81% en Macro F1-Score, mejorando e igualando el desempeño en 3 de las 5 clases equitativas, demostrando que existe una brecha de mejora para la modalidad textual. Como también, se incorporó la métrica Perplexity, no analizada anteriormente, consiguiendo un valor de 4.402, lo que subraya el potencial del modelo para aplicaciones que requieren una interpretación precisa de las emociones expresadas en texto, como la evaluación emocional en entornos educativos y sociales. Finalmente, en este estudio se cumplió el tercer objetivo específico, al establecer las bases para la creación de una arquitectura multimodal mediante la concatenación de los modelos unimodales con los mejores resultados desarrollados en este trabajo. La combinación de ambos modelos, junto con sus respectivas metodologías de preprocesamiento, representa una oportunidad para futuras investigaciones, mejorando no solo los modelos unimodales, sino también potenciando el desempeño multimodal, lo que podría optimizar aún más el rendimiento en el reconocimiento de emociones. De este modo, se cumplieron todos los objetivos planteados en este estudio, incluido el objetivo general, al proponer dos modelos unimodales que superan las métricas actuales, representando un aporte significativo al ámbito educativo, no abordado previamente en los trabajos analizados. Estos nuevos modelos proporcionan herramientas efectivas y útiles para los educadores y las instituciones educativas, facilitando un análisis más detallado y permitiendo la toma de decisiones estratégicas para abordar problemas críticos, como la deserción en la educación superior.

8 LIMITACIONES DEL TRABAJO

Al desarrollar el trabajo, se encontraron diversas limitaciones que influyeron en los resultados de los modelos. Para la modalidad facial, una de las limitaciones encontradas es la calidad del dataset FER2013, ya que, como se analizó previamente en los resultados, existe una gran similitud entre clases, especialmente entre las emociones "Sad" y "Fear", que tienden a confundirse con "Neutral". Como segunda limitación es el fuerte desbalance que existe entre las clases en el dataset FER2013, en particular en la clase "Disgust". Tanto en el conjunto de entrenamiento como en el de testeo predefinidos, se observa una gran disparidad, con solo 111 muestras para la clase minoritaria "Disgust" y 1.774 muestras para la clase mayoritaria "Happy" en el conjunto de testeo. Esta disparidad es aún mayor en el conjunto de entrenamiento, con 436 muestras en "Disgust" y 7.215 en "Happy". Este desbalance afecta el aprendizaje del modelo, favoreciendo las clases mayoritarias. En cuanto a la modalidad textual, se encontró la limitación de la baja cantidad de muestras en el dataset ISEAR, que cuenta con solo 7.666 datos con alrededor de 1.096 muestras por cada clase, lo cual podría ser insuficiente para que los modelos aprendan de manera efectiva. Otra limitación identificada es el ruido presente en las muestras textuales. A pesar del esfuerzo por limpiar los datos en el preprocesamiento, algunos textos contenían tanto ruido que, tras la limpieza, quedaron reducidos a una o dos frases, lo cual puede limitar la capacidad para representar adecuadamente dichas muestras. Por otro

lado, para la modalidad multimodal, se encontró la limitación de la complejidad necesaria en el preprocesamiento para extraer los rostros y transcripciones del dataset multimodal, ya que los rostros no mantienen una posición fija en cada escena y las transcripciones textuales carecen de estandarización, lo que dificulta tanto la creación como la evaluación del modelo. Por último, una limitación técnica clave en ambas modalidades fue la potencia computacional. Cada entrenamiento demoró aproximadamente 3 horas con una tarjeta de video RTX 2060, lo que limitó la posibilidad de experimentar con una mayor cantidad de hiperparámetros más complejos y modelos preentrenados de mayor tamaño.

9 FUTURAS MEJORAS

Para abordar las limitaciones encontradas, se pueden implementar mejoras para futuros trabajos. Para mejorar la calidad de los datos del dataset FER2013 en la modalidad facial, se recomienda explorar la combinación del dataset FER2013 con datasets con acceso privado de mayor cantidad de muestras, como AffectNet, que cuenta con aproximadamente 440.000 imágenes clasificadas, lo cual puede permitir obtener una mayor variedad y representatividad en los datos de entrenamiento. Respecto al gran desbalanceo existente entre la clase minoritaria "Disgust" y la clase mayoritaria "Happy", se recomienda explorar otras técnicas de balanceo adicionales a las experimentadas en este estudio, como el balanceo mediante redes generativas adversariales, o agregar más muestras en las clases desbalanceadas desde el dataset AffectNet, para analizar si el riesgo de overfitting disminuye y mejora la capacidad de generalización. Para la modalidad textual, una posible mejora es aportar mediante el desarrollo de un nuevo dataset que abarque las siete emociones estudiadas y que presente un etiquetado estandarizado y coherente. Asimismo, se podría probar con datasets principales menos estudiados, pero con mayor cantidad de muestras, como MELD, que cuenta con 13.000 muestras o GoEmotions con 58.009 datos. De esta forma, también se podría disminuir la limitación del ruido en los datos textuales, ya que al presentar mayor volumen de datos, el preprocesamiento disminuirá las muestras con un menor impacto al conjunto global. Para superar las limitaciones computacionales, se recomienda utilizar tarjetas de video especializadas, como NVIDIA A100 o V100, o servicios en la nube como AWS EC2 para experimentar con una mayor cantidad de hiperparámetros, implementar ensambles más complejos como Stacking, y probar con mayor cantidad de capas y modelos preentrenados de mayor tamaño. Finalmente, como una futura mejora y extensión de este trabajo, se propone la creación y experimentación del modelo multimodal propuesto que integre los modelos unimodales presentados en este estudio. Este modelo multimodal podría combinar los resultados de los modelos facial y textual para proporcionar una visión más completa y robusta en el reconocimiento de emociones, especialmente en entornos educativos. Este enfoque posiblemente potenciaría la precisión en la detección de emociones. Como también, se recomienda implementar los modelos en un sistema que pueda ser probado en entornos reales en el ámbito educativo para validar su efectividad en la práctica.

REFERENCES

[1] Abdelaziz A. Abdelhamid. 2022. Speech Emotions Recognition for Online Education. (November 2022). https://www.researchgate.net/publication/365490150_Speech_Emotions_Recognition_for_Online_Education

[2] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. 117–121. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379

[3] Acheampong Francisca Adoma, Nunoo-Mensah Henry, Wenyu Chen, and Niyongabo Rubungo Andre. 2020. Recognizing Emotions from Texts using a Bert-Based Approach. (2020), 62–66. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317523

[4] Mahinda Zidan Mahmoud Othman Ahmed R. Abas, Ibrahim Elhenawy. 2022. BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Computers, Materials & Continua* 71, 2 (2022), 2943–2961. http://www.techscience.com/cmc/v71n2/45793

[5] Mohammed A. Almulla. 2024. A multimodal emotion recognition system using deep convolution neural networks. *Journal of Engineering Research* (2024). https://doi.org/10.1016/j.jer.2024.03.021

[6] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time Convolutional Neural Networks for Emotion and Gender Classification. arXiv:cs.CV/1710.07557

[7] José Guillermo Balbuena Galván. 2022. *Modelos de Detección de Emociones en Texto y Rostros para Agentes Conversacionales Multimodales.* Master's thesis. Pontifica universidad católica del perú, Lima.

[8] J.D. Bodapati, U. Srilakshmi, and N. Veeranjaneyulu. 2022. FERNet: A Deep CNN Architecture for Facial Expression Recognition in the Wild. *J. Inst. Eng. India Ser. B* 103 (2022), 439–448. https://doi.org/10.1007/s40031-021-00681-8

[9] M. Boekaerts. 2007. *Understanding Students' Affective Processes in the Classroom.* Academic Press, 37–56. https://www.sciencedirect.com/science/article/abs/pii/B9780123725455500046?via%3Dihub

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42, 4 (December 2008), 335–359.

[11] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation* 5, 4 (2022). https://doi.org/10.3390/asi5040080

[12] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation* 5, 4 (2022). https://www.mdpi.com/2571-5577/5/4/80

[13] Davide Ciraolo, Maria Fazio, Rocco Salvatore Calabrò, Massimo Villari, and Antonio Celesti. 2024. Facial expression recognition based on emotional artificial intelligence for tele-rehabilitation. *Biomedical Signal Processing and Control* 92 (2024), 106096. https://doi.org/10.1016/j.bspc.2024.106096

[14] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 126–134.

[15] Ministerio de Educación SIES. 2019. *Deserción de primer año y Reingreso a la Educación Superior en Chile. Análisis cohorte 2015.* Informe. Servicio de Información de Educación Superior. https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/4599/Desercionreingreso_2015_2019.pdf?sequence=1&isAllowed=y Subsecretaría de Educación Superior, Gobierno de Chile.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018). https://doi.org/10.48550/arXiv.1810.04805 Submitted on 11 Oct 2018 (v1), last revised 24 May 2019 (v2).

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:cs.CV/2010.11929

[18] Seyed Hamed Noktehdan Esfahani and Mehdi Adda. 2024. Classical Machine Learning and Large Models for Text-Based Emotion Recognition. *Procedia Computer Science* 241 (2024), 77–84. https://doi.org/10.1016/j.procs.2024.08.013 19th International Conference on Future Networks and Communications/ 21th International Conference on Mobile Systems and Pervasive Computing/14th International Conference on Sustainable Energy Information Technology.

[19] D'Errico Francesca, Paciello Marinella, and Cerniglia Luca. 2016. When emotions enhance students' engagement in e-learning processes. *Journal of E-Learning and Knowledge Society* 12, 4 (2016). https://doi.org/10.20368/1971-8829/1144

[20] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. 2016. Analysis of Emotional Speech—A Review. In *Toward Robotic Socially Believable Behaving Systems - Volume I*, A. Esposito and L. Jain (Eds.). Intelligent Systems Reference Library, Vol. 105. Springer, Cham, 205–238. https://doi.org/10.1007/978-3-319-31056-5_11

[21] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. *Challenges in Representation Learning: A report on three machine learning contests.* Technical Report 1307.0414. arXiv. arXiv:stat.ML/1307.0414 https://arxiv.org/abs/1307.0414

[22] Alex Graves and Jürgen Schmidhuber. 2005. 2005 Special Issue: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18 (2005), 602–610. https://api.semanticscholar.org/CorpusID:1856462

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:cs.CV/1512.03385

[24] C. Hema and Fausto Pedro Garcia Marquez. 2023. Emotional speech Recognition using CNN and Deep learning techniques. *ScienceDirect* (August 2023). https://www.sciencedirect.com/science/article/abs/pii/S0003682X23002906?fr=RR-2

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[26] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* 167 (2019), 26–37. https://doi.org/10.1016/j.knosys.2019.01.019

[27] Azal Ahmad Khan. 2022. Balanced Split: A new train-test data splitting strategy for imbalanced datasets. arXiv:cs.LG/2212.11116

[28] Puneet Kumar and Balasubramanian Raman. 2022. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks* 150 (2022), 392–407. https://doi.org/10.1016/j.neunet.2022.03.017

[29] Ming-Che Lee, Shu-Yin Chiang, Sheng-Cheng Yeh, and Ting-Feng Wen. 2020. Study on emotion recognition and companion Chatbot using deep neural network. *SpringerLink* (27 March 2020). https://link.springer.com/article/10.1007/s11042-020-08841-6

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:cs.CL/1907.11692 https://arxiv.org/abs/1907.11692

[31] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[32] Gaurav Meena and Krishna Kumar Mohbey. 2023. Sentiment analysis on images using different transfer learning models. *ScienceDirect* (2023). https://www.sciencedirect.com/science/article/pii/S1877050923001424?ref=pdf_download&fr=RR-2&rr=8063070218850dd1

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. arXiv:cs.CL/1310.4546

[34] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[35] Sanjaya Mishra. 2017. Open universities in the Commonwealth: At a glance. (2017). http://hdl.handle.net/11599/2786

[36] Ozioma Collins Oguine, Kanyifeechukwu Jane Oguine, Hashim Ibrahim Bisallah, and Daniel Ofuani. 2022. Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction. arXiv:cs.CV/2206.09509

[37] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. 2019. KerasTuner. https://github.com/keras-team/keras-tuner.

[38] Tim Pearce, Alexandra Brintrup, and Jun Zhu. 2021. Understanding Softmax Confidence and Uncertainty. arXiv:cs.LG/2106.04972

[39] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[40] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Rada Mihalcea, and Erik Cambria. 2018. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[41] Wolfram Titz Reinhard Pekrun, Thomas Goetz and Raymond P. Perry. 2002. Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist* 37, 2 (2002), 91–105. https://doi.org/10.1207/S15326985EP3702_4

[42] Hany M. Sadak, Ashraf A. M. Khalaf, and Gerges M. Salama. 2024. DANA: Deep Attention Network Architecture for Facial emotions Recognition using limited resources. In *2024 International Telecommunications Conference (ITC-Egypt)*. 44–49. https://doi.org/10.1109/ITC-Egypt61547.2024.10620536

[43] Goutam Kumar Sahoo, Jayakrishna Ponduru, Santos Kumar Das, and Poonam Singh. 2022. Deep Leaning-Based Facial Expression Recognition in FER2013 Database: An in-Vehicle Application. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 1–6.

[44] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:cs.CL/1910.01108 https://arxiv.org/abs/1910.01108

[45] Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning: Correction. *Journal of Personality and Social Psychology* 67, 1 (1994), 55. https://doi.org/10.1037/0022-3514.67.1.55

[46] Servicio de Información de Educación Superior. 2022. *Informe 2022 Retención de 1er año de pregrado Cohorte 2017 - 2021*. Technical Report. Servicio de Información de Educación Superior. https://www.mifuturo.cl/wp-content/uploads/2022/09/Retencion_Pregrado_2022_SIES.pdf

[47] Servicio de Información de Educación Superior. 2022. *Matrícula en Educación Superior en Chile*. Technical Report. Servicio de Información de Educación Superior. https://educacionsuperior.mineduc.cl/wp-content/uploads/sites/49/2022/07/2022_MATRICULA.pdf

[48] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:cs.CV/1409.1556

[49] Vandana Singha and Swati Prasad. 2023. Speech emotion recognition system using gender dependent convolution neural network. *ScienceDirect* (2023). https://www-sciencedirect-com.recursosbiblioteca.unab.cl/science/article/pii/S1877050923002272

[50] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 399–402.

[51] Le Quang Thao, Do Trung Kien, Ngo Chi Bach, Dang Thi Thanh Thuy, Luong Thi Minh Thuy, Duong Duc Cuong, Nguyen Ha Minh Hieu, Nguyen Ha Thai Dang, Pham Xuan Bach, and Le Phan Minh Hieu. 2024. Monitoring and improving student attention using deep learning and wireless sensor networks. *Sensors and Actuators A: Physical* 367 (2024), 115055. https://doi.org/10.1016/j.sna.2024.115055

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:cs.CL/1706.03762

[53] Weiqing Wang, Kunliang Xu, Hongli Niu, and Xiangrong Miao. 2020. Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation. (September 2020). https://www.hindawi.com/journals/complexity/2020/4065207/

[54] Eric Hsiao-Kuang Wu, Chun-Han Lin, Yu-Yen Ou, Chen-Zhong Liu, Wei-Kai Wang, and Chi-Yun Chao. 2020. Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot. (April 2020). https://ieeexplore.ieee.org/document/9069183

[55] Yaping Xu, Yanyan Li, Yunshan Chen, Haogang Bao, and Yaqian Zheng. 2023. Spontaneous visual database for detecting learning-centered emotions during online learning. *Image and Vision Computing* 136 (2023), 104739. https://doi.org/10.1016/j.imavis.2023.104739

[56] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:cs.CL/1906.08237 https://arxiv.org/abs/1906.08237

[57] Daniel Yohanes, Jessen Surya Putra, Kenneth Filbert, Kristien Margi Suryaningrum, and Hanis Amalia Saputri. 2023. Emotion Detection in Textual Data using Deep Learning. *Procedia Computer Science* 227 (2023), 464–473. https://doi.org/10.1016/j.procs.2023.10.547 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).