Framework de Reconocimiento de Emociones para el análisis de salud ocupacional en tiempo real, aplicando un Método Multimodal basado en Deep Learning.

Morejón Cevallos, Mercedes Elizabeth y Noboa Villacís, Andrés Esteban

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Trabajo de titulación, previo a la obtención del título de Ingeniero en Tecnologías de la Información

PhD. Tapia León, Freddy Mauricio

16 de agosto de 2023

Copyleaks

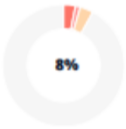
Plagiarism report

MIC-PART2 - TESIS_MOREJON_NOVOA....

Scan details

Scan time:	Total Pages:	Total Words:
August 24th, 2023 at 21:12 UTC	134	33327

Plagiarism Detection



Types of plagiarism		Words
Identical	3.2%	1068
Minor Changes	1.3%	421
Paraphrased	3.5%	1183
Omitted Words	0%	0

AI Content Detection

Text coverage
- AI text
- Human text

Plagiarism Results: (136)

Deep learning con personas | Exámenes de Ciencia... 0.3%

https://www.docsity.com/es/deep-learning-con-personas/81...

Prepara tus exámenes Consigue puntos Orientación Universidad Ven...

1646-9895-rist-40-45.pdf 0.3%

https://scielo.pt/pdf/rist/n40/1646-9895-rist-40-45.pdf

Recebido/Submission: 14/07/2020 Aceitação/Acceptance: 28/10/2020 Revista Ibérica de Sistemas e Tecnologias de Informação Revista Ibérica...

TFG_ALVARO_GONZALEZ_ALMANSA LAREDO.pdf 0.3%

https://oa.upm.es/75003/1/tfg_alvaro_gonzalez_almansa%20...

Ricardo Imbert Paredes

Universidad Politécnica de Madrid Escuela Técnica Superior de Ingenieros Informáticos Grado en Ingeniería Informática Trabajo Fin de Gra...

FREDDY MAURICIO TAPIA LEON

Firmado digitalmente por FREDDY MAURICIO TAPIA LEON
Nombre de reconocimiento (DN): c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000310294, cn=FREDDY MAURICIO TAPIA LEON
Fecha: 2023.09.21 08:22:02 -05'00'

PhD. Tapia León, Freddy Mauricio

Director

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Certificación

Certifico que el trabajo de titulación: "**Framework de Reconocimiento de Emociones para el análisis de salud ocupacional en tiempo real, aplicando un Método Multimodal basado en Deep Learning**" fue realizado por los señores **Morejón Cevallos, Mercedes Elizabeth y Noboa Villacís, Andrés Esteban** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí,21 de septiembre de 2023

FREDDY MAURICIO TAPIA LEON

Firmado digitalmente por FREDDY MAURICIO TAPIA LEON
Nombre de reconocimiento (DN): c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000310294, cn=FREDDY MAURICIO TAPIA LEON
Fecha: 2023.09.21 08:22:30 -05'00'

PhD. Tapia León, Freddy Mauricio

CC: 1714745690

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Responsabilidad de Autoría

Nosotros, **Morejón Cevallos, Mercedes Elizabeth**, con cédula de ciudadanía n°1004373492 y **Noboa Villacís, Andrés Esteban** con cédula de ciudadanía n°1754299749, declaramos que el contenido, ideas y criterios del trabajo de titulación: **"Framework de Reconocimiento de Emociones para el análisis de salud ocupacional en tiempo real, aplicando un Método Multimodal basado en Deep Learning."** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 14 de septiembre de 2023

Morejón Cevallos, Mercedes Elizabeth
C.C.: 1004373492

Noboa Villacís, Andrés Esteban
C.C.: 1754299749

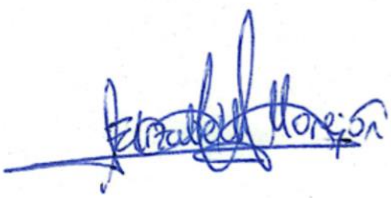
Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Autorización de Publicación

Nosotros **Morejón Cevallos, Mercedes Elizabeth**, con cédula de ciudadanía n°1004373492 y **Noboa Villacís, Andrés Esteban** con cédula de ciudadanía n°1754299749, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Framework de Reconocimiento de Emociones para el análisis de salud ocupacional en tiempo real, aplicando un Método Multimodal basado en Deep Learning."** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 14 de septiembre de 2023



Morejón Cevallos, Mercedes Elizabeth

C.C.: 1004373492

Noboa Villacís, Andrés Esteban

C.C.: 1754299749

Dedicatoria

Dedicado a mis padres por su constante apoyo y por ser mis modelos que seguir en cada aspecto de la vida, por ser el ejemplo de perseverancia y dedicación ha sido mi guía en cada paso que he dado.

Andrés Esteban Noboa Villacís

A mis padres Ramiro y María, por el esfuerzo, sacrificio y apoyo incondicional a lo largo de mi trayectoria académica. A mis hermanos y en especial a mi hermana quien partió de este mundo para guiar mi camino desde el cielo. A cada uno de mis familiares y amigos que de alguna u otra manera estuvieron brindándome apoyo y cariño constante. A Andrés con quien he compartido momentos buenos y malos durante mi vida universitaria bríndame su apoyo incondicional y ayudándome a sobre llevar situaciones difíciles para hoy en día alcanzar este sueño. A mis docentes quienes con palabras de aliento y consejos me supieron guiar para que en el momento oportuno sea una gran profesional sin olvidar la experiencia estudiantil vivida.

Mercedes Elizabeth Morejón Cevallos

Agradecimiento

Agradezco a mi querida madre y a mi valiente padre por ser mis guardianes incansables, por rescatarme de momentos difíciles con su amor incondicional y por estar a mi lado siempre que los he necesitado. A mi hermano Danilo le expreso mi especial gratitud por su inquebrantable amistad y total respaldo. También quiero expresar mi gratitud a mis amigos, quienes han sido pilares fundamentales en este camino, gracias a todos por creer en mí, y por ayudarme a creer en mí mismo.

Andrés Esteban Noboa Villacís

Agradezco infinitamente a la Universidad de las Fuerzas Armadas ESPE, en especial a la carrera de Tecnologías de la Información, a cada uno de sus docentes y amigos por la oportunidad y las enseñas impartidas a fin de formar grandes profesionales.

A mis padres, hermanos y familiares por su apoyo incondicional en todo momento, por sus palabras de aliento para salir a delante y lograr el objetivo propuesto, encaminando este viaje para una vida profesional futura.

Mercedes Elizabeth Morejón Cevallos

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Resumen

El Desarrollo del presente Proyecto contempla desarrollar un Framework de Reconocimiento Emocional para el Análisis de Salud Ocupacional en tiempo real, aplicando un método multimodal basado en Deep Learning. En primera instancia se realiza un exhaustivo proceso de investigación sobre métodos y técnicas de reconocimiento emocional analizando el nivel de afectación dentro del cuadro de salud ocupacional. Posteriormente se trabaja bajo definiciones de reconocimiento, encontrando fuentes de investigación y casos de éxito con relación a procesos similares. Para el desarrollo del framework se emplearon técnicas de reconocimiento emocional basada en rostro (FER), reconocimiento emocional del habla (SER) y sets de datos predefinidos como: MESD, RAVDESS, TESS y SAVEE que permitan la identificación de señales fisiológicas, gestuales, no verbales y vocales a través de la extracción emocional. Dichos conjuntos de datos sirvieron como base para el entrenamiento y validación de un modelo de red neuronal convolucional de capas múltiples, mediante el uso de métricas de evaluación estándar como la matriz de confusión y técnicas de validación por retención "Hold-out", se desarrolla un modelo con una precisión del 92% sobre las 7 emociones universales. El proyecto se lleva a cabo en el entorno de programación de Google Colab bajo lenguaje Python y librerías externas que facilitan la identificación de patrones clave durante el proceso de reconocimiento emocional, empleando técnicas de aprendizaje automático según la sea la necesidad, además se procede con la implementación de los modelos predictivos mediante un aplicativo web de uso sencillo.

Palabras clave: Reconocimiento Emocional, Aprendizaje Profundo, Tecnoestrés, Redes Neuronales, Inteligencia Artificial.

Abstract

The Development of this Project contemplates an Emotional Recognition Framework for the Analysis of Occupational Health in real time, applying a multimodal method based on Deep Learning. In the first instance, an exhaustive research process is carried out on emotional recognition methods and techniques, analyzing the level of involvement in the framework of occupational health. Subsequently, work is done under a similar definition of recognition, finding sources of research and success stories in relation to the processes. For the development of the framework, techniques of emotional recognition based on faces (FER), emotional voice recognition (SER) and predefined data sets were used, such as: MESD, RAVDESS, TESS and SAVEE that allow identifying physiological, gestural, non-verbal and vowels through emotional extraction. These data sets served as the basis for the training and validation of a multilayer convolutional neural network model, using standard evaluation metrics such as the confusion matrix and "Hold-out" validation techniques, a model is developed with an accuracy 92% in the 7 universal emotions. The project is carried out in the Google Colab programming environment under the Python language and external libraries that facilitate the identification of key patterns during the emotional recognition process, using machine learning techniques as necessary, in addition to proceeding with the implementation of predictive models through an easy-to-use web application.

Keywords: Emotional Recognition, Deep Learning, Technostress, Neural Networks, Artificial Intelligence.

Capítulo I

Aspectos Generales

Antecedentes

Las tecnologías relacionadas con Inteligencia Artificial han crecido a medida que el comportamiento humano y su actividad emocional se han visto integradas, esto debido a su valiosa contribución al desarrollo de grandes descubrimientos o aportes que han permitido al ser humano evolucionar constantemente en el desarrollo de una mejor sociedad, además de optimizar la toma de decisiones, siendo el campo de la salud una de estas áreas beneficiadas.

La Universidad Nacional de Córdoba et al., (2016) menciona que gran parte del contexto que aborda el área de la psicología trata acerca del proceso emocional de un individuo, en los últimos años expertos e investigadores han desarrollado interés por la aplicación de Inteligencia Artificial en cuanto al reconocimiento de emociones de manera automatizada abarcando diversos campos tales como: robótica, minería de datos, entretenimiento, educación, automatización de procesos y áreas de salud. Los estudios de comportamiento usual del ser humano permitieron considerar 6 emociones básicas como felicidad, sorpresa, enojo, asco, miedo y tristeza, con el análisis de estas emociones se busca determinar el estado afectivo y su tendencia conductual en el ámbito social.

Un ejemplo concreto de aprendizaje multimodal es el reconocimiento de emociones audiovisuales, que involucra la combinación de información auditiva y visual para identificar las emociones en una situación específica. Hoy en día existen variedad de trabajos enfocados al proceso de reconocimiento emocional, para ello con los recientes desarrollos en arquitecturas de Deep Learning (DL), se ha podido crear avances en el reconocimiento de emociones con Redes de Atención Multimodal (Multimodal Attention Networks - MMAN), mediante el uso de técnicas que incluyen redes de creencias profundas, Redes Neuronales Convolucionales Profundas (Convolutional Neural Network - CNN), Memoria a Largo Plazo (Long Short-Term Memory - LSTM) y Transformadores (Transformers) con el fin de traer características basadas

en imágenes y rostros orientados al aprendizaje en entornos no controlados. (Cabada et al., 2019)

El trabajo propuesto por Luna-Jiménez et al., (2021), evaluó la eficacia de una Arquitectura de Transformación (Transformer) pre entrenada utilizando dos técnicas de aprendizaje por transferencia: Extracción por Incorporación y Sinfonía Fina (Fine-Tuning) para el Reconocimiento Emocional del Habla (Speech Emotion Recognition - SER) y el uso de Unidades de Acción (Action Units - AU) para el Reconocimiento Emocional Facial (Facial Emotion Recognition - FER). Combinando estas dos modalidades mediante un método multimodal, se logra una eficacia de 86.70% en el conjunto de datos audiovisuales de libre acceso (Ryerson AudioVisual Database of Emotional Speech and Song - RAVDESS).

De este modo, se busca implementar un sistema de reconocimiento de emociones en tiempo real que abarque el campo de salud ocupacional bajo un método multimodal, permitiendo identificar rasgos físicos y emocionales de un individuo ya sea a través de facciones faciales, señales emitidas por la voz o puntos clave orientados a solucionar la problemática entorno a la determinación del cuadro emocional de un usuario quien usa tecnología constante y que puede llegar a presentar cuadros críticos de tecnoestrés, cansancio y agotamiento previniendo a futuro accidentes de mayor impacto.

Problema

El habla es una forma crucial en la que el ser humano puede expresar sus emociones, no obstante, se torna un problema complejo identificar el estado emocional a través de señales lingüísticas, vocales, gestuales y visuales. Varios investigadores se han centrado en descubrir nuevos métodos y técnicas que faciliten el resultado óptimo en cuanto a la crisis emocional que el individuo sufre al manejar dispositivos tecnológicos en campos laborales, personales y sociales. (Quiroz Martínez et al., 2020)

Tanto la Inteligencia Artificial como el Aprendizaje Profundo (Deep Learning) han sido considerados campos de investigación importantes en cuanto a modelos de aprendizaje

automático cubriendo diversas áreas y adquiriendo mayor ventaja sobre métodos tradicionales. Sin embargo, el ser humano al mantenerse en constante evolución con dispositivos tecnológicos sufre cambios emocionales constantes entre los que destacan: depresión, ansiedad, fatiga, cansancio, estrés y otros factores.

Según investigaciones realizadas por la Universidad Nacional de Córdoba et al., (2016) han demostrado numerosos estudios indicando que adultos mayores poseen mayor dificultad en reconocimiento de expresiones faciales a diferencia de adultos jóvenes, este tipo de individuos requiere mayor atención para lograr rendimiento cognitivo de acuerdo con lo expresado emocionalmente.

Gracias a sets de datos encontrados en investigaciones anteriores, Redes Neuronales Convolucionales (Convolutional Neural Network - CNN) y Redes Neuronales Recurrentes (Recurrent Neural Network - RNN) estudios han permitido deducir sintomatología emocional a partir de espectros visuales, gestuales y vocales que facilitan la captura y procesamiento de la información emocional, identificando rasgos puntuales del ser humano que ayudan a optimizar el diagnóstico obtenido y permitiendo al personal especializado en salud establecer el tratamiento adecuado bajo el modelo de reconocimiento emocional planteado que integra técnicas multimodales y Aprendizaje Profundo (Deep Learning) para la evaluación del paciente en circunstancias específicas.

Formulación del Problema

Hoy en día los cambios hormonales y emocionales del ser humano generan mayor afectación dentro del cuadro de salud, al no recibir un diagnóstico oportuno y un tratamiento efectivo su cuadro podría verse afectado, es por ello que se ha implementado una estructura (Framework) de reconocimiento de emociones capaz de determinar resultados óptimos bajo las diversas señales fisiológicas que el individuo emite a fin de determinar el estado emocional al momento de hacer uso de dispositivos tecnológicos.

El uso excesivo de la tecnología genera consecuencias negativas dentro del ámbito de salud ocupacional y el campo laboral, el comportamiento inusual del humano con la tecnología se conoce como tecnoestrés o estrés laboral, este proceso se debe a que el paciente sufre momentos de ansiedad, agotamiento, tensión, dolor muscular y falta de concentración al verse en un área con ritmo de trabajo excesivo y bajo herramientas en las cuales carece de experiencia, motivo por el cual al realizar el proceso de investigación se formula la siguiente pregunta:

¿Es posible determinar el estado de sensibilidad mediante reconocimiento de emociones en tiempo real aplicando un método multimodal y la utilización de técnicas de Aprendizaje Profundo (Deep Learning)?

Justificación

En el ambiente laboral o entornos cotidianos el ser humano enfrenta cambios en relación con su comportamiento emocional y de estrés por lo que resulta complejo determinar el estado mental en el que se encuentra, para ello es crucial realizar un análisis emocional basado en técnicas de Deep Learning, reconocimiento y modelos multimodales que ayuden a tratar el déficit presentado. Actualmente el reconocimiento de emociones se identifica por patrones y reacciones de los seres humanos en función de la situación y el medio en el que se encuentran, sin dejar de lado la presencia de medios tecnológicos que causan afectaciones de tecnoestrés interfiriendo en su diario vivir. Sin embargo, la aplicación del Aprendizaje Profundo (Deep Learning - DL) es considerada una tecnología evolutiva de alto impacto, está tecnología será de vital importancia para el entrenamiento del modelo planteado, el cumplimiento de los objetivos y los resultados obtenidos.

Objetivos

General

Desarrollar un prototipo capaz de identificar y analizar las emociones de un individuo dentro del ámbito de salud ocupacional a través de un método multimodal, aplicando

expresiones faciales, aspectos gestuales, no verbales y vocales logrando detectar el nivel de tecnoestrés con mayor precisión y fiabilidad en tiempo real bajo el uso de herramientas de Aprendizaje Profundo (Deep Learning).

Específicos

- Realizar la revisión de literatura sobre el reconocimiento multimodal de emociones, identificando factores de riesgo que contribuyan al tecnoestrés a través de la identificación y clasificación de fuentes primarias.
- Desarrollar un prototipo funcional bajo un método multimodal capaz de reconocer y autenticar el comportamiento de un individuo a través de la combinación de múltiples modalidades como reconocimiento facial, gestual, no verbal y vocal.
- Implementar un sistema basado en computación afectiva en tiempo real, usando técnicas de Aprendizaje Profundo (Deep Learning) y procesamiento de señales fisiológicas mediante el cual se permita detectar, interpretar, procesar y responder ante la toma de decisiones y acciones de un individuo con la manipulación constante de medios tecnológicos.
- Determinar arquitecturas de reconocimiento de Aprendizaje Profundo (Deep Learning) adecuadas para la inferencia de emociones faciales mediante técnicas de Hold-out y Cross Validation , que permitan comprender y modelar patrones de comportamiento humano bajo el uso de la herramienta colaborativa Google Colab .
- Evaluar el prototipo en un entorno laboral no controlado que permita mejorar la calidad de vida digital del personal en cuanto al uso equilibrado de la tecnología bajo contribución de un experto en salud ocupacional.

Alcance

La Computación Afectiva es el término que caracteriza a las diversas emociones emitidas por el ser humano, dando paso a la determinación eficiente del estado mental del individuo, tanto el estrés como la emoción son estados que se presentan en circunstancias no

predeterminadas con alto impacto en cuanto a la regulación y expresión emocional. El presente trabajo se focaliza en realizar un framework de reconocimiento de emociones bajo un método multimodal basado en técnicas y herramientas de Aprendizaje Profundo (Deep Learning) que ayuden a comprender las afectaciones causadas en el individuo.

El método multimodal dentro del área de computación afectiva permite capturar un conjunto determinado de datos bajo un modelo de expresión afectiva, recreando escenas basadas en escenarios reales donde el ser humano exprese estados emocionales a través de señales fisiológicas, comportamiento gestual, no verbal y vocal bajo condiciones críticas o predeterminadas. (Ierache et al., 2020)

Considerando lo mencionado se busca realizar en primera instancia el reconocimiento de expresiones, características faciales y lingüísticas que faciliten la interdependencia entre el aspecto emocional y el tecnoestrés bajo estrategias de muestreo basado en tareas múltiples. Mediante el entrenamiento de modelos de Aprendizaje Profundo (Deep Learning) usando datasets de libre acceso en Reconocimiento Emocional Facial (Facial Emotion Recognition - FER) y Reconocimiento Emocional del Habla (Speech Emotion Recognition - SER), se permitirá la clasificación de las emociones basadas en tiempo real. Finalmente se procederá a la validación de los resultados obtenidos avalados por profesionales en salud ocupacional obteniendo resultados competitivos y dinámicos logrando superar las expectativas de los individuos diagnosticados de forma significativa con relación a otros métodos de detección existentes.

Capítulo II

Marco conceptual y Estado del Arte

Marco Conceptual

Computación Afectiva

En el área de Sistemas la computación afectiva se enfoca en reconocer, procesar y simular el comportamiento humano con el uso de dispositivos tecnológicos como el computador, gracias a técnicas de análisis de sentimientos y emociones se ha podido evaluar la adaptación del ser humano a variedad de procesos y su reacción perceptiva, considerando el análisis de voz como el principal punto que infiere en el campo emocional. (Chanchí-Golondrino et al., 2022)

La computación afectiva tiene una amplia gama de aplicaciones potenciales, que incluyen:

- **Interacción humano-computadora:** los sistemas informáticos afectivos se pueden utilizar para mejorar la experiencia del usuario de computadoras y otros dispositivos. Por ejemplo, un sistema informático afectivo podría usarse para ajustar el brillo de la pantalla de una computadora o el volumen de un altavoz en función del estado emocional del usuario.
- **Atención médica:** los sistemas informáticos afectivos se pueden usar para monitorear y diagnosticar afecciones de salud mental, como depresión y ansiedad. También se pueden utilizar para brindar terapia y apoyo a personas con estas afecciones.
- **Educación:** los sistemas informáticos afectivos se pueden utilizar para personalizar las experiencias de aprendizaje y proporcionar retroalimentación a los estudiantes. También se pueden utilizar para crear entornos de aprendizaje más atractivos e interactivos.

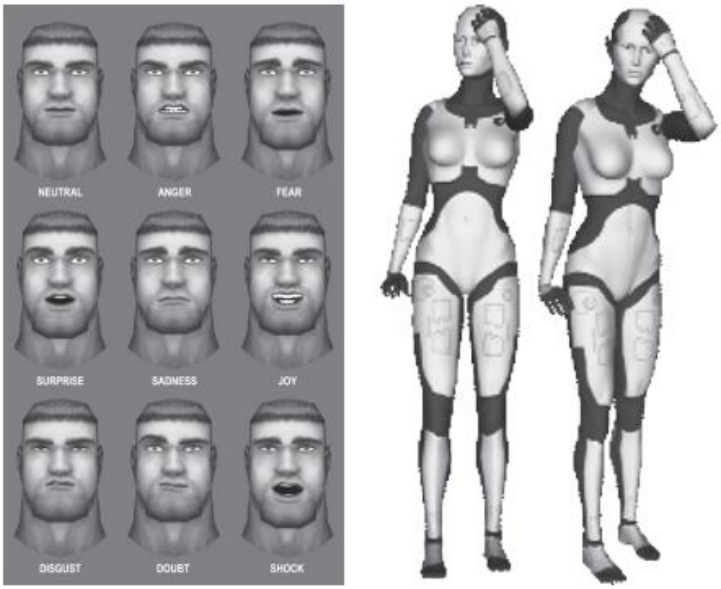
El crecimiento exponencial de la computación afectiva evalúa el comportamiento humano en tiempo real a través de sus emociones. Según lo menciona Baldassarri Santalucía (2016) las emociones expresadas por el ser humano cambian en cualquier momento entre ellas se presentan emociones tales como: estrés, aburrimiento, distracción, déficit de atención entre otras. Hoy en día existen variedad de herramientas que permiten el reconocimiento emocional del individuo según su condición al igual que existen numerosas aplicaciones para dar tratamiento a personas que cursan problemas emocionales avanzados. Gracias a la identificación de patrones fisiológicos se puede dar seguimiento a través de la monitorización de las señales emitidas por el cuerpo, así también por señales visuales que permiten reconocer el estado emocional propio o de otros individuos. Un claro ejemplo del uso y aplicación de la computación afectiva es el desarrollo y evolución de robots emocionales encargados de acompañar a personas de edad avanzada a fin de mitigar problemas de soledad. (Baldassarri Santalucía, 2016)

Sin embargo el potencial alcanzado por la computación afectiva y la interacción emocional puede verse manipulado por grandes compañías en el área de marketing, la cantidad de software existente capaz de detectar emociones positivas y negativas del ser humano han permitido reflexionar en cuanto al uso y manejo de información privada de una persona, la intención de la computación afectiva no es controlar al usuario en cuanto a su comportamiento, sino más bien mejorar su calidad de vida considerando su estado emocional de esta manera se pretende que el enfoque emocional controlado de una persona permita el desarrollo de capacidades y habilidades positivas dentro de la convivencia con otros individuos en momentos adecuados. (Baldassarri Santalucía, 2016)

Como se observa en la Figura 1 las emociones emitidas por el ser humano se expresan a través de expresiones faciales, movimientos, sonidos de voz o gestos corporales. (Baldassarri Santalucía, 2016)

Figura 1

Personajes virtuales bajo expresiones emocionales o corporales comunes



Nota. La figura indica personajes virtuales bajo expresiones emocionales o corporales comunes. Tomado de (Baldassarri Santalucía, 2016)

Reconocimiento emocional

Para realizar reconocimiento emocional basado en algoritmos de Inteligencia Artificial es necesario aplicar técnicas de aprendizaje profundo, procesamiento de lenguaje natural (Natural Lenguaje Processing - NLP), señales fisiológicas emitidas por el individuo, análisis de texto, voz e imágenes que en su conjunto son entrenados constantemente afinando detalles en cuanto a precisión, manejo de datos y de las diversas expresiones corporales.

Por otro lado, el reconocimiento de emociones cubre en su mayor parte el campo de la inteligencia artificial y sus diversas ramas, no obstante, el campo de la psicología influye en los patrones a determinar en un individuo bajo condiciones críticas. Años atrás el ser humano ha tratado de comprender los estados emocionales y el impacto en su comportamiento y en la toma de decisiones, gracias a los avances constantes de la tecnología y a estudios realizados

por investigadores es posible llevar a cabo un análisis emocional con mayor profundidad y precisión.

Los investigadores centraron estudios de reconocimiento emocional basados principalmente en emociones bajo una única modalidad, la mono modalidad de las emociones consiste en identificar expresiones bajo un sólo método sensorial, ayudando a profundizar los resultados dentro del campo de investigación científica, psicología clínica e inclusive dentro de la interacción emitida por el humano-ordenador. Sin embargo, debido a la complejidad del análisis monomodal en cuanto a emociones y cambios humorísticos se requiere integrar múltiples modalidades sensoriales y el contexto en el que se ubican para determinar un resultado más efectivo en menor tiempo de estudio posible. (Mobile Computing, 2023)

No obstante, el reconocimiento de emociones consiste en la capacidad de identificar y clasificar las emociones bajo señales emitidas, ya sea a través de señales de voz, expresiones faciales, gestuales y patrones fisiológicos, el reconocimiento emocional es una técnica usada para estandarizar y fijar patrones que ayuden a determinar el comportamiento del individuo a través del uso de tecnologías que requieren de aprendizaje automático y el entrenamiento de algoritmos que permiten identificar señales de entrada ante las emociones emitidas por el ser humano en campos determinados.

Abdullah et al. (2021) menciona que la aplicación del Aprendizaje Profundo (Deep Learning) en el reconocimiento emocional es considerado como un tema de investigación de gran interés, mostrando variedad de usos en amplias áreas. Para el diseño y revisión de los modelos existentes de reconocimiento emocional es necesario hacer énfasis en técnicas de aprendizaje profundo que incluyan Redes Neuronales Convolucionales (CNN) diseñadas para procesar imágenes e información emitida por el individuo, agrupadas bajo capas a fin de solucionar problemas y comportamientos inusuales, Redes de Creencia Profunda (DBN) constituidas por múltiples capas ocultas cuya funcionalidad depende del aprendizaje no supervisado, este tipo de redes permiten integrar reconocimiento vocal, de patrones y de

contenido bajo imágenes y características extraídas de manera más compleja, memoria a largo plazo y múltiples combinaciones entre redes. (Abdullah et al., 2021)

Inteligencia Artificial (IA)

Chollet (2021) menciona que el proceso de Inteligencia Artificial surgió alrededor de los años 50 abarcando un pequeño grupo de investigadores dedicados al área informática, cada uno de ellos se preguntó si las máquinas podrían pensar como el ser humano. Sin embargo, con el paso del tiempo la Inteligencia Artificial llegó a posicionarse a través de un taller propuesto por un docente matemático llamado John McCarthy. Estudios realizados por el matemático indican intentos por descubrir, abstracciones y conceptos novedosos de forma que permitan resolver a las máquinas problemas comunes ocasionados por el hombre. El constante esfuerzo e investigación por automatizar las tareas del hombre conllevan a definir a la Inteligencia Artificial como un área que engloba la automatización y profundidad del aprendizaje además de otras áreas. (Chollet, 2021)

Hasta los años 80 se creyó que la Inteligencia Artificial podría resolver problemas de la mano de programadores quienes establecieran reglas clave a fin de manejar el conocimiento almacenado bajo base de datos propias de un ambiente tradicional, el enfoque orientado al conjunto de reglas preestablecidas se denomina Inteligencia Artificial Simbólica. La IA simbólica se emplea para dar solución a problemas definidos en primera línea, dando paso al aprendizaje automático y su desarrollo. (Chollet, 2021)

Rouhiainen (2020) en otras investigaciones menciona que la Inteligencia Artificial consiste en la habilidad de los ordenadores de usar algoritmos, mejorarlos y procesarlos a través de la comprensión de los datos para la toma de decisiones tal cual el comportamiento humano. A diferencia del ser humano los procesos automatizados con inteligencia artificial pueden realizar millones de tareas constantes y almacenar grandes volúmenes de datos al mismo tiempo minimizando consumo de recursos y posteriores errores comúnmente cometidos por procesos donde el humano interviene. Hoy en día gracias a la evolución de la IA los

sistemas automatizados pueden realizar tareas poco inusuales que en tiempos pasados requerían de la intervención de coste en infraestructura, mantenimiento y mano de obra ralentizando procesos y dificultando su desarrollo y cumplimiento de actividades dentro de una organización. (Rouhiainen, 2020)

Los procesos de inteligencia artificial están siendo usados para obtención de mejoras significativas en cuanto a eficiencia y rendimiento en los diversos campos donde el ser humano se destaca de manera habitual, alcanzando mayor ventaja bajo técnicas y algoritmos eficientes de Inteligencia Artificial como lo es el aprendizaje automático, la visión artificial, el procesamiento de lenguaje natural y la mejora relacionada con la toma de decisiones. Sin embargo, con el pasar de los años la IA alcanza mayor impacto en diversas áreas tales como: salud, educación, seguridad, negocios, etc. (Rouhiainen, 2020)

Aprendizaje Automático (Machine learning – ML)

Charles Babbage inventor de la máquina analítica junto a su colaboradora Ada Lovelace bajo perspectiva histórica pensaron en si un ordenador sería capaz de ejecutar tareas personalizadas. Años más tarde Alan Turing precursor de la Inteligencia Artificial explicó que las máquinas podrían emular aspectos definidos en cuanto a Inteligencia Artificial. La forma más común de hacer que un ordenador realice tareas programadas es bajo el seguimiento de reglas establecidas a fin de convertir los datos de entrada en salidas oportunas. (Chollet, 2021)

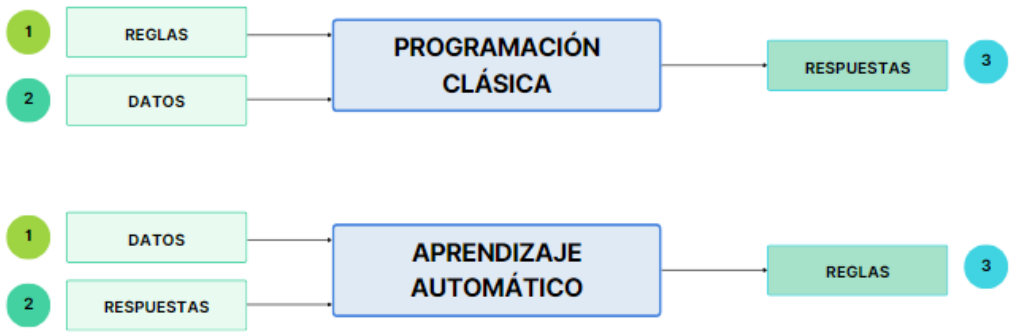
Dentro del área de Machine Learning o aprendizaje automático la computadora es la encargada de analizar los datos de entrada y las salidas obtenidas, un sistema de aprendizaje profundo procede a ejecutar el modelo y entrenarlo constantemente, un claro ejemplo podría ser analizar un conjunto de imágenes etiquetadas por el ser humano a fin de seguir patrones y reglas establecidas asociando las imágenes con sus respectivas etiquetas. (Chollet, 2021)

Como se puede ver en el modelo de programación clásica en la Figura *2*, establece reglas y datos en su entrada para llegar a un resultado, mientras tanto el aprendizaje automático transforma los datos de entrada juntamente con las respuestas en un proceso de

salida donde se obtienen reglas a seguir, acercando al resultado esperado durante el proceso de entrenamiento. (Chollet, 2021)

Figura 2

Paradigma inicial del aprendizaje profundo



Nota. La figura indica el paradigma inicial del aprendizaje profundo. Tomado y traducido de (Chollet, 2021)

Chollet (2021) menciona que el aprendizaje automático se relaciona con el uso y aplicación de estadísticas matemáticas, de esta manera tiende a manipular conjuntos de datos extensos y de complejidad alta, a todo esto, el aprendizaje automático fortalece su dependencia de acuerdo con los constantes avances relacionados a hardware y software en el área informática.

A menudo, los científicos de datos remarcan que no existe un algoritmo predeterminado que pueda resolver cualquier tipo de problema (Mahesh, 2020). El objetivo de dichos algoritmos es el de 'aprender' un modelo o un conjunto de reglas determinadas de un conjunto de datos clasificado, de tal manera que pueda predecir correctamente las etiquetas de datos (Ej.: imágenes que representan alguna emoción) que no estén presentes en el conjunto de datos. El enfoque de Aprendizaje Automático (Machine Learning - ML) permite resolver estos problemas de una manera indirecta, enseñando a un modelo alimentándose con datos de entrada (x) y con datos de salida (y) (correctamente etiquetados previamente) se denomina unsupervised

machine learning o aprendizaje automático no supervisado como se detalla en el Aprendizaje no supervisado, esto es simplemente:

$$y = f(x) \quad (1)$$

Según Chollet (2021) para poder aplicar técnicas de aprendizaje automático es necesario considerar 3 aspectos importantes:

- Entrada de datos: De acuerdo con el escenario es importante tomar puntos clave ya sea a través de señales fisiológicas, voz, texto, imágenes, etc.
- Resultados esperados: Interpretación de los datos obtenidos.
- Análisis de resultados: Determina si las medidas alcanzadas son adecuadas entre la salida actual del algoritmo y su salida.

Método de validación por retención (Hold-out)

Como lo menciona Raschka et al., (2022) durante el proceso de aprendizaje automático, los datos son separados inicialmente en datos de prueba y datos de entrenamiento, de tal manera que para alcanzar un nivel de eficiencia es necesario ajustar las configuraciones pertinentes, el modelo de retención (Hold-out) menciona a la reutilización de datos durante la selección y ejecución del modelo.

Estos datos generalmente se asignan en tres conjuntos o conjuntos de datos diferentes, de esta forma, se dividen de la siguiente manera y como se observa en la Figura 3:

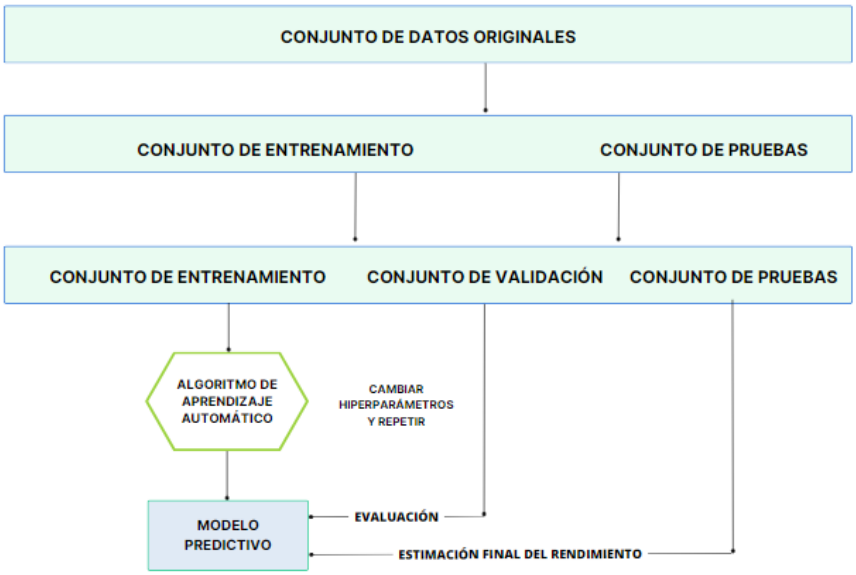
1. Conjunto de entrenamiento: una muestra de datos que se utiliza para ajustar y entrenar el modelo. Estos son los datos reales que se utilizan para entrenar nuestro modelo, esto es de lo que el modelo "aprende".
2. Conjunto de validación: una muestra de datos que se utiliza para proporcionar una evaluación imparcial de un modelo que se ajusta al conjunto de datos de entrenamiento. El modelo utilizará estos datos indirectamente para ajustar sus parámetros, pero nunca directamente para "aprender de".

3. Conjunto de pruebas: una muestra de datos que se utiliza para proporcionar una evaluación imparcial de un ajuste final del modelo en el conjunto de datos de entrenamiento. Los datos de prueba son el conjunto de datos intacto que se utiliza para evaluar el modelo.

La Figura 3 permite también observar el funcionamiento del método de retención, ocupa el conjunto de datos de validación para evaluar el modelo de entrenamiento bajo parámetros establecidos, una vez que los parámetros estén ajustados acorde al modelo se pone en funcionamiento el conjunto de datos de prueba, este proceso es ejecutado a medida que el modelo es entrenado. (Raschka et al., 2022)

Figura 3

Proceso de selección, verificación y ejecución



Nota. La figura indica el proceso de selección, verificación y ejecución de la información durante el proceso de entrenamiento. Tomado y traducido de (Raschka et al., 2022)

Aprendizaje Profundo (Deep Learning - DL)

El Aprendizaje Profundo se deriva como un subcampo del aprendizaje automático, consiste en comprender un enfoque de forma profunda, la cantidad de capas que conforma el

modelo ayuda a determinar su nivel de profundidad. La capa o capas que en su conjunto representan al Aprendizaje Profundo se caracterizan por aprender automáticamente durante el proceso de entrenamiento, las capas que aprenden del modelo establecido son denominadas redes neuronales, que en su composición están formadas por múltiples capaz una sobre otra. (Chollet, 2021)

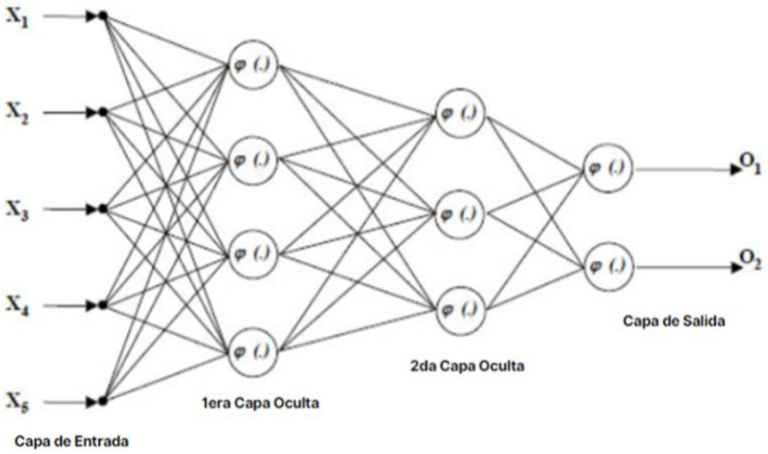
La aplicación de capas almacena pesos denominándose parámetros de una capa, para que el modelo aprenda constantemente es necesario establecer un conjunto de valores asignando pesos por capa a fin de cumplir el objetivo de parametrización, la cual afectará el comportamiento humano.

El término conocido como Deep Learning o Aprendizaje profundo en español, se refiere a Redes Neuronales Artificiales con múltiples capas (Artificial Neural Networks - ANN), denominadas núcleo del Deep Learning, las ANN son una de las técnicas más revolucionarias dentro del Machine Learning en los últimos años, se refiere a una estructura de red de predicción y procesamiento de características no lineal con una sólida capacidad de autoaprendizaje , su estructura básica está compuesta por varias capas denominadas: Capa de Entrada (Input Layer) , Capa Oculta (Hidden Layer) y Capa de Salida (Output Layer). (Montesinos López et al., 2022)

La Figura 4 muestra la estructura de una red neuronal artificial compuesta por una capa de entrada con 5 nodos, dos capas ocultas de activación y finalmente dos capas de salida. Las capas de entrada se forman de acuerdo con la cantidad de información, mientras que las capas de salida corresponden a las variables de respuesta existentes. (Villada et al., 2016)

Figura 4

Arquitectura de redes neuronales artificiales



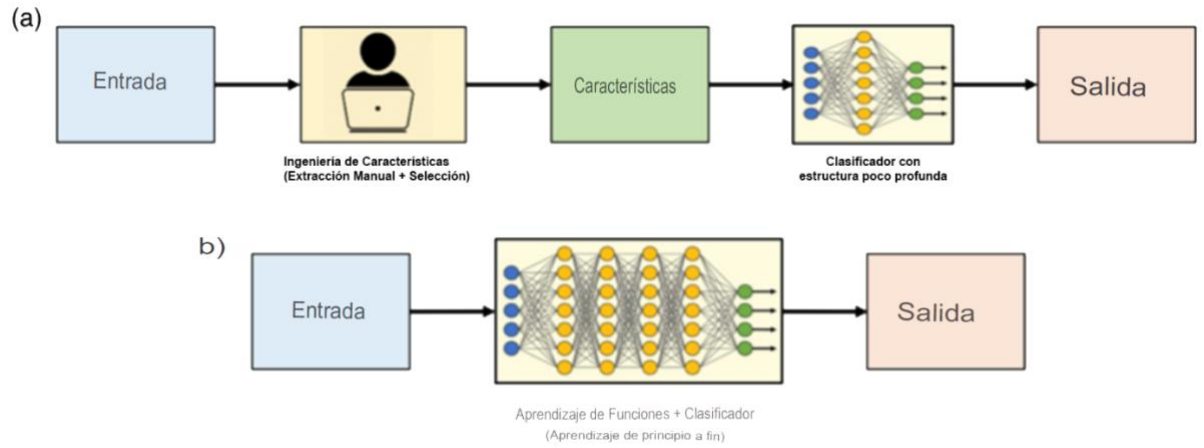
Nota. Tomado y traducido de (Montesinos López et al., 2022)

El Aprendizaje Profundo (Deep Learning) ha evolucionado bastante en los últimos años, superando a los métodos tradicionales con arquitecturas de redes neuronales como las redes neuronales convolucionales (CNN) y varias formas de redes neuronales recurrentes. El flujo del Deep Learning difiere significativamente al ML tradicional, como se observa en la Figura 5, el flujo elimina la necesidad de la extracción manual y selección de características (también conocido como ingeniería de características), este proceso implica la extracción y transformación de distintas variables a partir de los datos en bruto[1] que vayamos a procesar, de forma que se adapten a las características del modelo que vayamos a entrenar, este es proceso bastante largo debido al trabajo manual que este representa, y es necesario para obtener un buen modelo en términos de precisión si hablamos puramente del Aprendizaje Automático (Machine Learning). (Luo et al., 2021)

[1] Datos en bruto o también llamados raw data en inglés, son los datos tal y como vienen de la fuente, sin ninguna transformación ni limpieza.

Figura 5

Flujo tradicional del Aprendizaje Automático vs Lenguaje Profundo



Nota. Tomado y traducido de (Luo et al., 2021)

Chollet (2021) menciona que el Deep Learning se ha popularizado logrando resultados en cuanto a percepción, procesamiento de lenguaje natural, habilidades naturales e intuitivas que han permitido comprender el comportamiento humano frente al desarrollo del computador. Dentro de las áreas que ha logrado fortalecer el aprendizaje profundo se encuentran reconocimiento de voz, clasificación de imágenes, reconocimiento textual, asistentes digitales tales como Google Assistant y Amazon Alexa y sobre ello interpretación y procesamiento de lenguaje natural. (Chollet, 2021)

Aprendizaje Supervisado

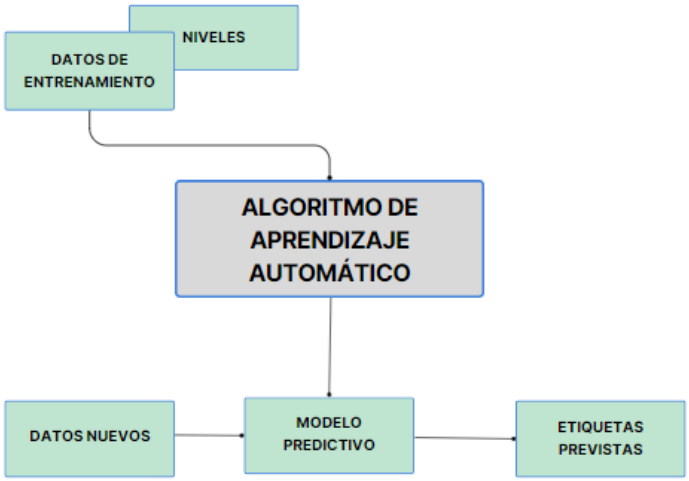
El alcance del objetivo supervisado consiste en el modelamiento a partir de datos de entrenamiento etiquetados por el ser humano a partir de los cuales se realizan predicciones sobre sucesos no previstos. El término supervisado abarca un conjunto de datos de entrada los cuales son procesados para obtener una salida óptima. Existen diferentes tipos de aprendizaje supervisado, entre ellos se encuentra aprendizaje supervisado por clasificación y regresión, durante el proceso de clasificación el aprendizaje supervisado es usado para categorizar patrones a fin de predecir las etiquetas bajo observaciones de sucesos pasados. Las etiquetas visualizadas se presentan en valores discretos y desordenados agrupándose en su conjunto

para el análisis respectivo. Así mismo el aprendizaje que permite la interpretación y comprensión de resultados contiguos se denomina regresión, este tipo contiene como variables de entrada predicciones obteniendo como salidas variables continúas permitiendo la predicción de un resultado. (Raschka et al., 2022)

La Figura 6 indica el comportamiento del aprendizaje supervisado entorno al flujo de trabajo, se observa que los datos de entrenamiento se dirigen a un algoritmo de Aprendizaje Automático establecido en el cual se ajustan las predicciones sobre datos nuevos y etiquetas previstas. (Raschka et al., 2022)

Figura 6

Proceso del aprendizaje supervisado



Nota. La figura indica el proceso del aprendizaje supervisado. Tomado y traducido de (Raschka et al., 2022)

La Southwest Jiaotong University, China et al., (2015) menciona que para que el aprendizaje supervisado tenga mayor eficiencia en resultados requiere de una mejor selección de características. En la mayoría de las investigaciones el uso de la fuerza bruta para la recopilación de datos no es recomendable, en varias ocasiones los datos recolectados presentan ruido y valores incompletos, lo cual conlleva mayor trabajo durante el proceso de reconocimiento de datos.

Aprendizaje no Supervisado

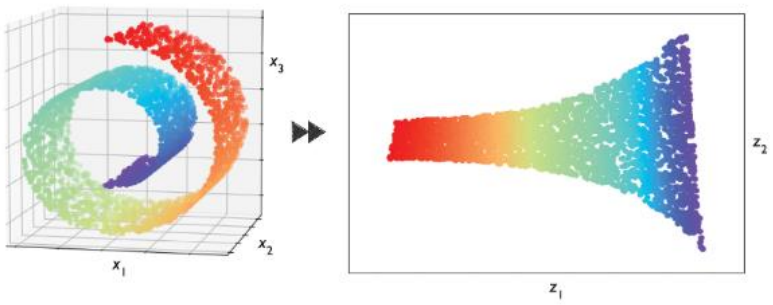
El aprendizaje supervisado abarca un objetivo de reconocimiento basado en etiquetas. Sin embargo, el aprendizaje no supervisado contiene datos no etiquetados o bajo estructura desconocida. Para la extracción de información sin una guía previa de resultados conocidos es necesario aplicar el aprendizaje supervisado. Durante el análisis se busca descubrir patrones sin conocimiento alguno sobre sus datos de origen, otro aspecto que considera el aprendizaje no supervisado durante su desarrollo es la dimensionalidad de la información, en varias ocasiones se integran datos inusuales convirtiéndose en un desafía en cuanto a almacenamiento limitado y rendimiento computacional. La reducción de la dimensionalidad permite abstraer datos de importancia distribuidos en espacios pequeños a fin de conservar la información relevante. (Raschka et al., 2022)

Otro entorno enfocado al aprendizaje no supervisado corresponde a la dimensionalidad, esto hace referencia a la cantidad de datos que en su espacio de almacenamiento limita el rendimiento de algoritmos de aprendizaje predeterminados. En su funcionamiento la dimensionalidad es utilizada a fin de eliminar el ruido de los datos, reduciendo la información sin afectar a la información principal o relevante. (Raschka et al., 2022)

La Figura 7 indica el proceso de reducción de dimensionalidad no lineal basado en un rollo suizo 3D a un espacio dimensional 2D. (Raschka et al., 2022)

Figura 7

Reducción de dimensionalidad de un entorno en 3D a 2D



Nota. Tomado de (Raschka et al., 2022)

Datos Estructurados

La información no procesada en varias ocasiones es considerada relevante para el rendimiento óptimo de un algoritmo de aprendizaje, procesar la información es uno de los pasos principales ante cualquier aplicación de aprendizaje automático. No obstante, las redes neuronales se encuentran diseñadas para desarrollar su funcionamiento bajo datos estructurados, existen redes neuronales que incluyen gráficas las cuales hacen uso del aprendizaje profundo compatible con datos estructurados. (Raschka et al., 2022)

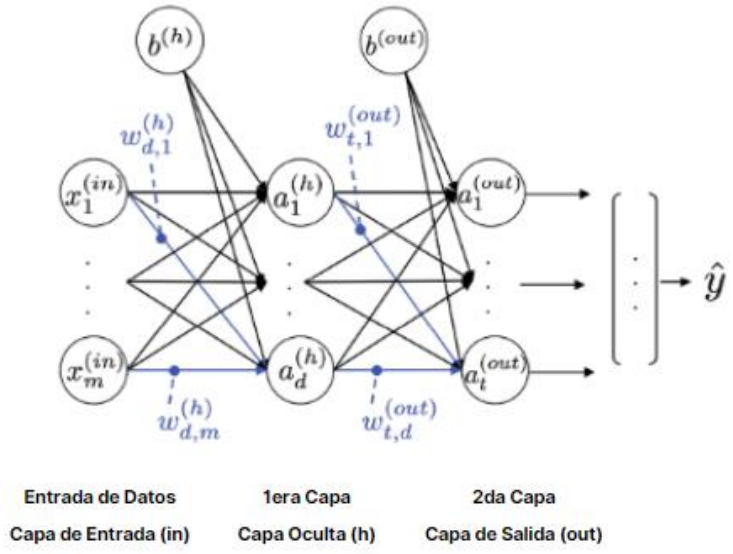
Redes Neuronales de Avance Prealimentada (Feedforward Neural Networks - FNN)

Esta Red Neuronal de Avance Prealimentada (FNN) es un tipo de arquitectura de red neuronal artificial en la cual la información fluye en una única dirección: desde la capa de entrada, a través de una o más capas ocultas, hasta llegar a la capa de salida. Este tipo de red desplaza los datos a través de las capas en forma unidireccional, sin presentar ciclos o conexiones de retroalimentación. La red se compone de nodos (neuronas) dispuestos en capas, y cada nodo de una capa se encuentra conectado a todos los nodos de la capa que le sigue. (Abirami & Chitra, 2020)

Dentro de estas redes neuronales (FNN), una de las principales aplicaciones es el uso del Perceptrón Multicapa (Multi Layer Perceptron - MLP) es una extensión de las FNN. Está compuesto por tres tipos de capas: la capa de entrada, la capa de salida y la capa oculta, como se ilustra en la Figura 8.

Figura 8

Concepto de un MLP



Nota. Ilustración del concepto de un MLP que consta de dos capas. Tomado y traducido de (Raschka et al., 2022)

De acuerdo con la teoría explicada por Raschka et al., (2022), el MLP es un ejemplo característico de una red neuronal artificial de tipo feedforward. El concepto de Feedforward hace referencia a que cada capa proporciona su salida como entrada a la siguiente capa, sin la existencia de bucles o conexiones recurrentes, el proceso de un MLP se detalla como:

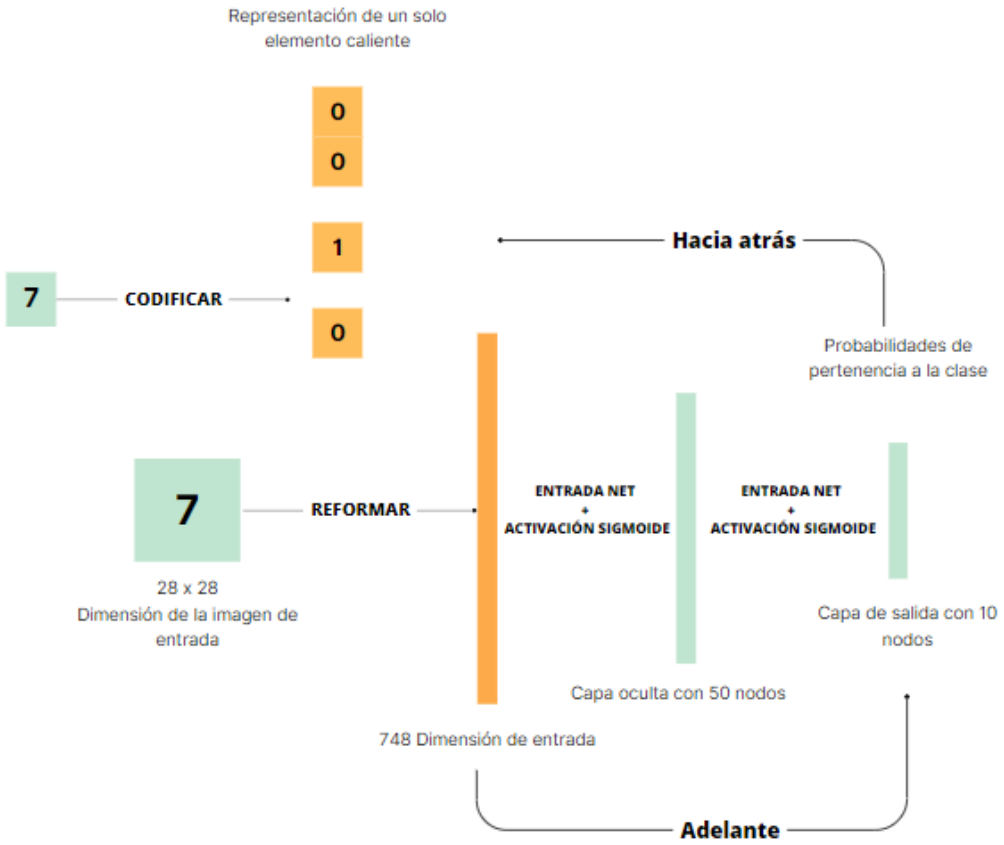
- **Propagación de la capa de entrada:** los patrones de los datos de entrenamiento se introducen en la red desde la capa de entrada y se propagan hacia adelante a través del MLP. Esta propagación implica pasar los datos a través de las capas ocultas, lo que finalmente da como resultado una salida.
- **Cálculo de pérdida: con la salida de la red generada:** El siguiente paso consiste en calcular la función de pérdida que cuantifica la discrepancia entre la salida prevista y los valores objetivo-reales. El objetivo es minimizar esta pérdida para mejorar la precisión del modelo.

- **Retro propagación y actualización del modelo:** al calcular la pérdida, el sistema realiza una retro propagación. Este proceso implica encontrar la derivada de la pérdida con respecto a cada unidad de peso y sesgo en el MLP. Al comprender cómo cambia la pérdida con respecto a estos parámetros, el modelo puede actualizarse y optimizarse para mejorar el rendimiento.

El proceso completo para el reconocimiento de dígitos escritos a mano aplicando los conceptos de Perceptrón Multicapa (MLP) se detalla de la siguiente forma. Ver Figura 9.

Figura 9

Arquitectura MLP



Nota. Arquitectura MLP de 3 capas para el etiquetado de dígitos escritos a mano. Tomado y traducido de (Raschka et al., 2022)

Redes Neuronales Convolucionales (Convolutional Neural Networks)

Según como lo menciona Almabdy & Elrefaei (2019) las Redes Neuronales Convolucionales (CNN) han alcanzado protagonismo dentro del campo de Inteligencia Artificial (IA) debido a la clasificación de patrones, este tipo de red ha logrado superar el alcance máximo del modelo tradicional de visión por computadora para lo que respecta a la clasificación de imágenes.

Una Red Neuronal en su composición se encuentra formada por variedad de neuronas conectadas entre sí a través de un campo receptivo, este campo se encarga de recibir parámetros de entrada a fin de activar una salida como respuesta ante una señal emitida. El principal uso de una red neuronal convolucional consiste en la detección y reconocimiento visual de objetos, dentro de sus principales aplicaciones se encuentran reconocimiento de imágenes y video, Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) y reconocimiento emocional, todo ello bajo el diseño y aplicación de algoritmos que conjuntamente con la data proporcionada por el individuo cruza la fase de entrenamiento constante a fin de fortalecer el nivel de precisión del modelo en cuanto a un diagnóstico preestablecido. (Almabdy & Elrefaei, 2019)

La estructura de composición de una Red Neuronal Convolucional (Figura 10) se centra en una variedad de capas, cada capa es capaz de tomar datos de entrada bajo el manejo de una matriz, de esta manera se obtienen datos de salida, este proceso se puede evidenciar durante el reconocimiento de imágenes, donde la entrada corresponde a las imágenes ingresadas, mientras que en la salida se obtienen datos estadísticos, los cuales indican la probabilidad más eficiente referente al conjunto de imágenes analizado. (Kamencay et al., 2017)

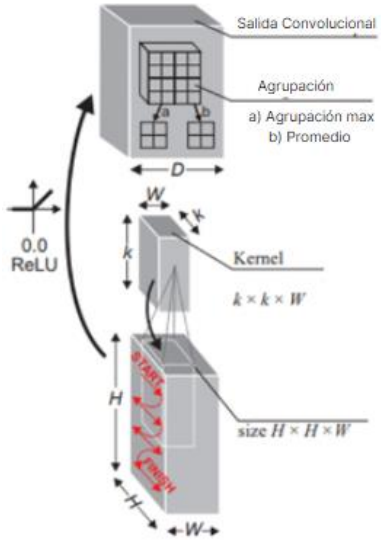
Según lo menciona Kamencay la arquitectura que compone una Red Neuronal Convolucional (CNN) completa está formada por (Kamencay et al., 2017):

- **Capa de entrada:** Almacena los valores de píxeles sin procesar de una imagen.

- **Capa Convolucional:** Calcula la salida de las neuronas que se encuentran conectadas a un punto de entrada conocido como campo receptivo.
- **Capa RELU:** Facilita la activación de los elementos.
- **Capa POOL:** Permite muestrear los datos emitidos entre las neuronas de entrada como las de salida.

Figura 10

Red neuronal convolucional (CNN)



Nota. Proceso de una red neuronal convolucional (CNN). Tomado de (Namatēvs, 2017)

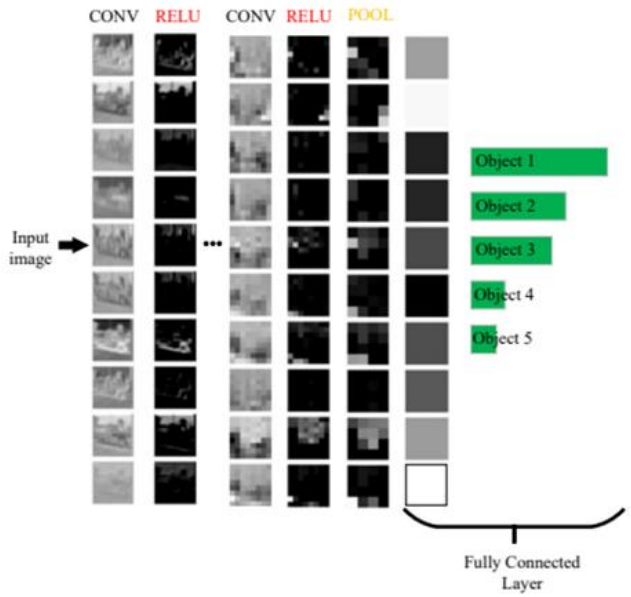
La Figura 11 durante el proceso de una Red Neuronal Convolucional detalla su composición de la siguiente manera, (Kamencay et al., 2017):

- **Entrada (Input 32 x 32):** La imagen sin procesar contiene los valores de los píxeles considerando (32 x 32) al ancho de la imagen.
- **La capa CONV:** Calcula la salida de las neuronas, tomando en consideración el peso y la región a la que se encuentra conectado el volumen de entrada, se observa la presencia de 12 filtros como un volumen establecido (32 x 12).
- **La capa RELU:** Consiste en la función de activación de todos los elementos de la Red Neuronal.

- **La capa POOL:** Aquí se realiza un muestreo descendente dando como resultado un volumen de (16 x 16 x 12)
- **La capa totalmente conectada (FC):** Puntúa a una clase específica, no obstante, al igual que las neuronas ordinarias se observa que cada neurona está conectada a las neuronas referentes al volumen anterior.

Figura 11

Activación de una CNN



Nota. Ejemplo de activación de una CNN. Tomado de (Kamencay et al., 2017)

Memoria a Largo Plazo (LSTMs)

La memoria a largo plazo es considerada como una capa capaz de recordar sucesos pasados, usualmente la información cronológica está destinada al comportamiento de una red neuronal recurrente, su aprendizaje depende de las secuencias de entrada, así también de los diferentes modelos mentales que influyen en el tiempo. Sin embargo, la visión del aprendizaje automático en el mundo real está orientada a mejorar el rendimiento de las tareas comunes. Una red neuronal recurrente bidireccional está orientada a mejorar los resultados obtenidos capturando patrones de orden cronológico. (Chollet, 2021)

La red neuronal recurrente bidireccional se adapta a datos textuales, a partir del año 2016 LSTM bidireccional cubre tareas con enfoque de procesamiento de lenguaje natural, de la misma forma es importante que durante la ejecución de operaciones el ser humano sepa aprovechar datos y fuentes para el entrenamiento entre humano- máquina. (Chollet, 2021)

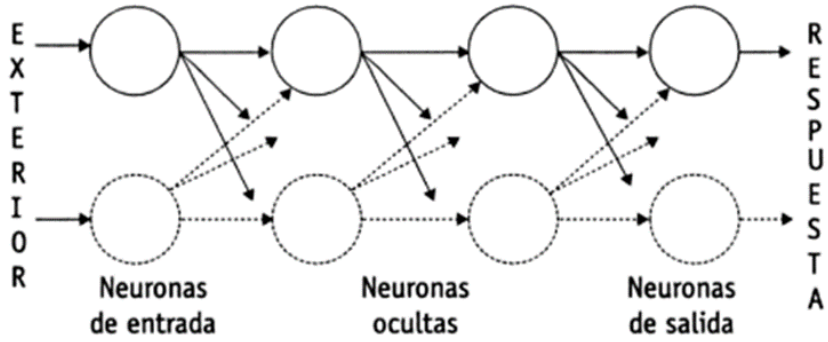
Moghar & Hamiche (2020) mencionan que la memoria a largo plazo (LSTM) es considerada un tipo de red neuronal recurrente (RNN) capaz de capturar información de circunstancias anteriores a fin de usarla en predicciones futuras.

Según menciona Flórez & Fernández (2008), existen 3 tipos de Redes Neuronales, ver Figura 12:

- **Neuronas de entrada:** Reciben señales emitidas por el entorno en el que se encuentra.
- **Neuronas Ocultas:** Reciben y procesan estímulos emitidos por las señales de entrada para obtener como resultado una señal de salida.
- **Neuronas de salida:** Una vez que el entrenamiento haya terminado, la señal de salida deja el sistema por completo.

Figura 12

Tipos de Redes Neuronales



Nota. Tomado de (Flórez & Fernández, 2008)

Cada una de las capas que forman parte de la estructura de una red neuronal cuenta con nodos de entrada dependientes de la dimensión de los datos manejados a fin de en su

entrenamiento minimizar la tasa de error producida por el flujo de información. El sistema LSTM en su desarrollo para almacenar información anterior hace uso de una línea de memoria incorporada. (Moghar & Hamiche, 2020)

Muhuri et al. (2020) indica que para construir un modelo LSTM-RNN es necesario identificar los parámetros que ayudan a determinar el estado de activación durante el entrenamiento, para ello se considera que:

- El tamaño del lote corresponde a la cantidad de registros de entrenamiento desde épocas pasadas hasta la actualidad.
- Una época es definida como una etapa anterior o actual de un entrenamiento determinado.
- La etapa de regularización permite eliminar neuronas poco usuales a lo largo del paso del tiempo.
- La señal de activación recibe las señales de entrada para transformarlas en señales de salida pasando por la capa oculta que conforma la estructura de una red neuronal.

Sobreajuste (Dropout)

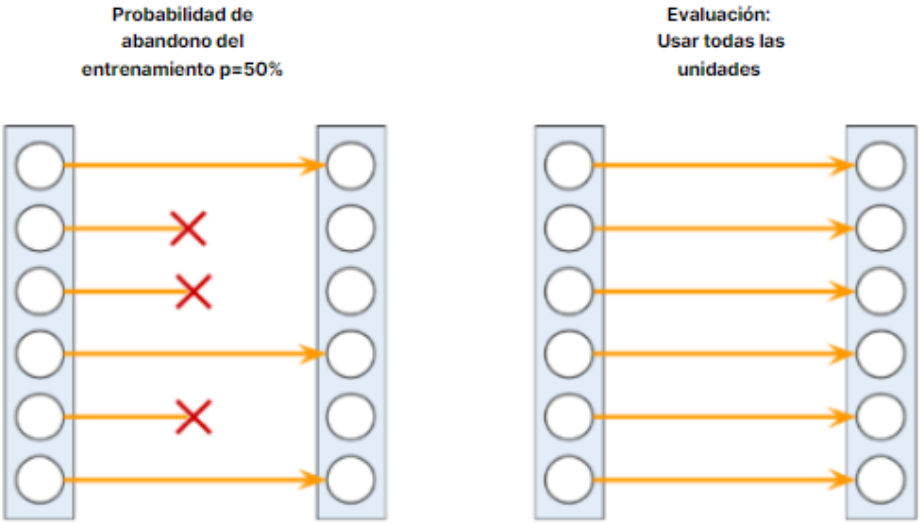
El Sobreajuste o más conocido como Dropout es una técnica usada dentro del aprendizaje automatizado, evita el sobreajuste mejorando la eficacia del modelo, durante la investigación realizada por Labach menciona que a partir del año 2012 el método de deserción omitió neuronas bajo la probabilidad de 0.5 durante cada fase de entrenamiento, a fin de demostrar que la técnica puesta en marcha tenía mejoras significativas en cuando a los resultados obtenidos. Sin embargo, el término Dropout consiste en involucrar parámetros o activaciones neuronales modificadas a lo largo del entrenamiento de una red. (Labach et al., 2019)

En la Figura 13 se puede apreciar un claro ejemplo de probabilidad de abandono con un valor de p=50% durante el proceso de entrenamiento, de esta manera se puede observar la

inactivación de la mitad de las neuronas, no obstante, es necesario la participación de las neuronas en su totalidad durante la fase de preactivación y evaluación. (Raschka et al., 2022)

Figura 13

Aplicación Dropout



Nota. Aplicación Dropout durante entrenamiento neuronal de una red. Tomado y traducido de (Raschka et al., 2022)

Agrupación (Pooling)

El Pooling o más conocido como agrupación, es parte de las técnicas usadas para el procesamiento de datos y aprendizaje automático, dentro del área de las Redes Neuronales Convolucionales (CNN) permite agregar características y reducir a gran escala la carga computacional. A diferencia de otros componentes existentes, la agrupación puede adaptarse a los datos recibidos y a la red dada a fin de mejorar el rendimiento del modelo. (S. Li et al., 2019)

Las Capas de Agrupación o Pooling son capas sin parámetros, a diferencia de las capas convolucionales quienes en su operación emplean funciones matemáticas y algoritmos bidimensionales. (Raschka et al., 2022)

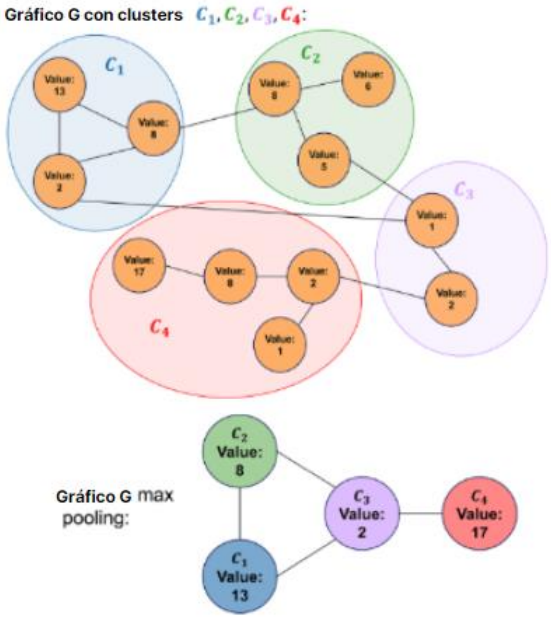
La localidad en modo gráfico se ha desarrollado fácilmente debido a la agrupación de capas de datos por imagen, gracias a la agrupación de nodos es posible determinar capas de agrupación. Se puede definir al Pooling como una agrupación no esclarecida dependiendo del

enfoque en el que se encuentre dirigido, es importante mencionar que la agrupación de nodos da como resultado falta de claridad en su conexión a otros nodos. (Raschka et al., 2022)

En la Figura 14 se observa que el conjunto de nodos conforma el máximo de combinaciones, asignando la unión de todos los índices bajo un clúster designado. Sin embargo, se puede observar cómo los nodos i, j y k son asignados a un clúster C1 comprendiendo que cada nodo que forme parte de un clúster compartirá un borde con un C1. Para entender el comportamiento de la agrupación máxima o Pooling máximo es importante abordar la agrupación de muestreo de forma simultánea, a través de este tipo se distribuyen los nodos bajo clústeres incrustados. (Raschka et al., 2022)

Figura 14

Aplicación máxima de nodos por agrupación.



Nota. Tomado y traducido de (Raschka et al., 2022)

Funciones de activación (Softmax)

La función Softmax se caracteriza por proporcionar probabilidades de clases significativas bajo entornos multinomiales, son usadas comúnmente al momento de normalizar y sumar funciones lineales. Del mismo modo la función Softmax es usada como una salida

normalizada lo cual permite calcular la probabilidad de entrada referente a una clase. (Raschka et al., 2022)

Las funciones de activación Softmax generan una distribución bajo probabilidad de N clases, existen dos formas de clasificar las etiquetas, Chollet, (2021) menciona:

- Codificación categórica conocida como codificación one-hot
- Codificación de etiquetas bajo uso de números enteros y uso de la función de perdida sparse_categorical_crossentropy

Unidad de Activación Lineal Rectificada (ReLU)

La Unidad Lineal Rectificada, comúnmente conocida como ReLU, es la función de activación más utilizada en el campo del Aprendizaje Profundo (Deep Learning) y ha ganado una inmensa popularidad debido a su capacidad para ofrecer resultados de última generación mientras mantiene una alta eficiencia computacional. La función de activación de ReLU es una función matemática simple que introduce la no linealidad en las redes neuronales, lo que les permite aprender y modelar relaciones complejas en los datos. (Agarap, 2018)

Esta misma se define como f(x) = max(0,x), donde "x" es la entrada de la función. Opera estableciendo todos los valores negativos de "x" en cero y pasando los valores positivos sin cambios. Esta característica ayuda a superar el problema del gradiente de fuga, que era un desafío con las funciones de activación tradicionales como sigmoid y tanh. (Oostwal et al., 2021)

Reconocimiento de Emociones Faciales (Facial Emotion Recognition - FER)

Este Reconocimiento de Emociones (FER) es usado para interpretar las emociones emitidas por un individuo, desempeña gran importancia al momento de que un individuo interactúa con un ordenador ya sea a través de actividades físicas, juegos, actividades empresariales o simplemente en circunstancias de la vida diaria.

Gracias al avance constante de la Inteligencia Artificial y el desarrollo de varios algoritmos se ha logrado precisar el reconocimiento emocional bajo alta gama de imágenes

capturadas en tiempo real en condiciones controladas, dando solución a problemas posiblemente complejos. Es importante mencionar que existen cambios constantes en el comportamiento humano motivo por el cual existe variación en cuanto a los resultados. Con el desarrollo en visión por computadora es importante realizar la clasificación de imágenes haciendo uso de las Redes Neuronales Convolucionales (CNN) a fin de mejorar la precisión y con ello solventar condiciones difíciles de identificar a simple vista por el ser humano. (Khaireddin & Chen, 2021)

Gracias al éxito alcanzado por métodos tradicionales en procesos de reconocimiento facial bajo la extracción de facciones gesticulares, visuales y vocales, se ha visto que grandes investigadores han demostrado interés por el aprendizaje profundo. Gracias a estudios de investigación es necesario analizar a través de Reconocimiento Emocional Facial (Facial Emotion Recognition) y el comportamiento humano a fin de obtener mejores resultados en cuanto a detección.

Según Mellouk & Handouzi (2020) en su artículo menciona que Mollahosseini et al. propone una red neuronal convolucional bajo variedad de bases de datos para realizar el proceso de Reconocimiento Emocional Facial (FER), después de extraer puntos de referencia de un conjunto de imágenes se procedió a la aplicación de la técnica basada en datos aumentados. Mellouk & Handouzi, (2020) en su investigación propone estudios de Reconocimiento Emocional Facial, uno de ellos aplica dos capas de agrupación, contenidas en capas convolucionales de tamaños 1x1, 3x3 y 5x5. Sin embargo, la técnica aplicada permite reducir problemas de sobreajuste en las capas convolucionales locales.

No cabe duda de que el reconocimiento emocional a través de expresiones faciales es una técnica que constantemente se vuelve desafiante, el comportamiento humano expresa emociones naturales es por ello por lo que a través de técnicas y modelos se busca encontrar un mecanismo computacional que permita resolver un problema. Con el diseño de algoritmos capaces de interpretar el comportamiento humano a través de sus expresiones abre un extenso

camino a lo que comúnmente se conoce como interacción humana – computador. (Canal et al., 2022)

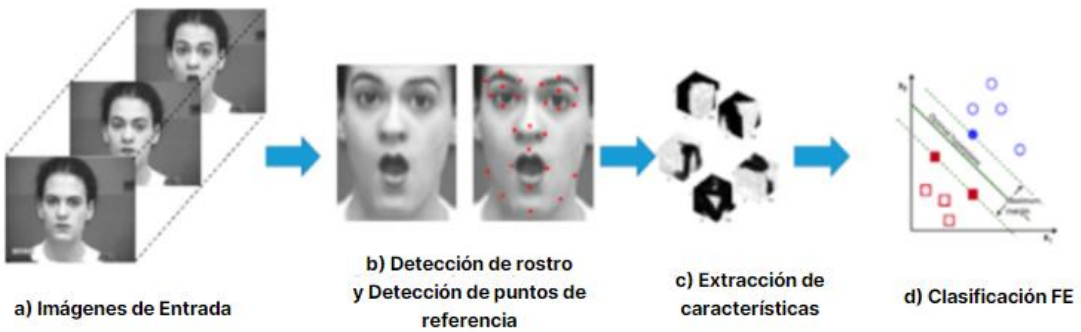
El reconocimiento facial FER cubre grandes áreas como psicología, neurociencia, cognición humana y sobre todo aprendizaje, es importante considerar que el ser humano es un ser cambiante en cuanto a sus expresiones emocionales independientemente de su entorno, es por ello por lo que pueden surgir problemas al entablar un ambiente de comunicación. A fin de solucionar grandes incógnitas expertos, matemáticos e investigadores analizan métodos y algoritmos que controlen el comportamiento humano de manera eficiente. (Canal et al., 2022)

Alrededor de los últimos años las CNN (Convolutional Neuronal Network) han logrado resolver varios problemas bajo señales de entrada tomando en consideración un conjunto de características invariantes, por lo tanto, la CNN ha logrado cubrir áreas que incluyen procesamiento de imágenes y reconocimiento de patrones que ayuden a determinar el estado de un individuo y su condición en un ambiente determinado. (Canal et al., 2022)

En la Figura 15, en primera instancia detecta una imagen partiendo del rostro y sus diversas facciones, posteriormente extrae características o patrones especiales de acuerdo con los componentes faciales y finalmente los datos extraídos son clasificados y entrenados a fin de obtener resultados de reconocimiento en cuanto a lo expresado por un individuo a través del rostro. (Ko, 2018)

Figura 15

Proceso de reconocimiento facial FER



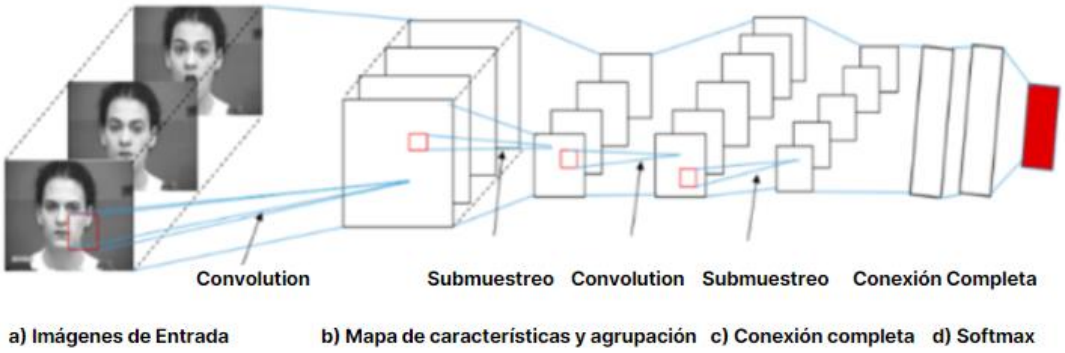
Nota. Tomado y traducido de (Ko, 2018)

Ko (2018) señala el proceso de reconocimiento facial bajo Redes Neuronales Convolucionales (CNN), como se puede apreciar en la Figura 16:

1. El conjunto de imágenes de entrada usa filtro en las capas de convolución.
2. A partir de los resultados obtenidos por las capas de convolución se construyen mapas de entidades y capas bajo agrupamiento máxima reduciendo el conjunto inicial de mapas de entidades.
3. Se aplican capas de redes neuronales cuyo funcionamiento va detrás de las capas convolucionales.
4. Finalmente se observa el reconocimiento de una expresión facial a través de una señal de salida emitida bajo la aplicación de funciones softmax.

Figura 16

Proceso de reconocimiento facial FER enfocado en CNN



Nota. Tomado y traducido de (Ko, 2018)

Reconocimiento de Emociones del Habla (Speech Emotion Recognition - SER)

El Reconocimiento de Emociones a través del habla es un aspecto crucial en las Interacciones Humano Computadora (HCI) en la actualidad. Estos sistemas han logrado mejorar significativamente la interacción natural con las máquinas, permitiendo la comunicación directa a través de la voz en lugar de depender de dispositivos de entrada tradicionales. Esto facilita la comprensión del contenido verbal y la respuesta de los usuarios humanos de manera más eficiente.

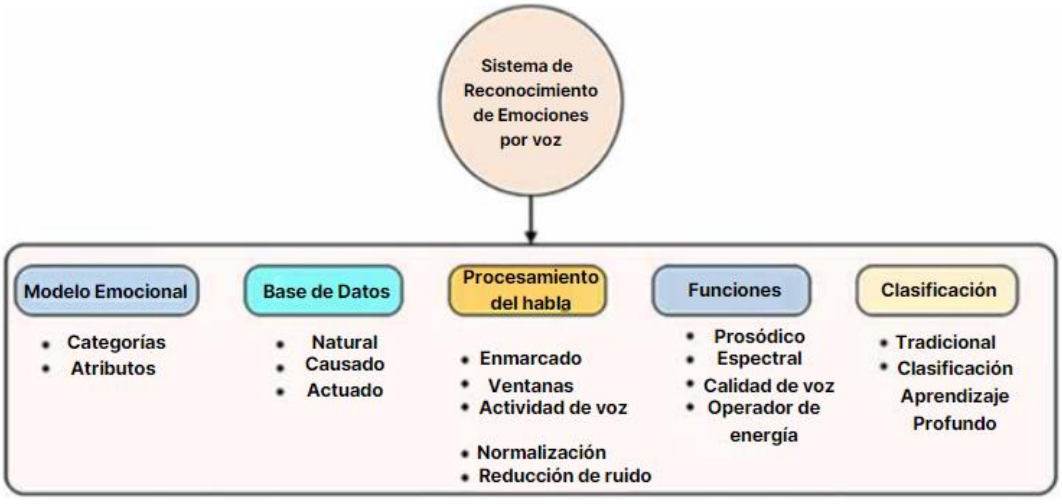
La capacidad de reconocer emociones en la voz humana ha abierto nuevas posibilidades en diversos campos. Por ejemplo, en el ámbito de la atención al cliente, los sistemas de reconocimiento de emociones pueden identificar la frustración o la satisfacción en la voz del cliente, lo que permite a las empresas mejorar la calidad de su servicio y abordar rápidamente los problemas. (Nassif et al., 2019)

Un sistema SER general permite la clasificación de las emociones mediante distintas partes, la clasificación de las emociones humanas varía psicológicamente según la intensidad de la emoción, el tipo de emoción y los límites de diversos parámetros. Estos elementos pueden agruparse y reconocerse en modelos de emociones (Emotional Models).

Estos modelos clasifican múltiples emociones, que se comprenden a través de su nombre, teniendo en cuenta su duración, el impacto en el comportamiento, la sincronización, la velocidad de cambio, la intensidad, la evaluación obtenida y el enfoque del evento. Ver Figura 17:

Figura 17

Sistema de reconocimiento de emociones por voz



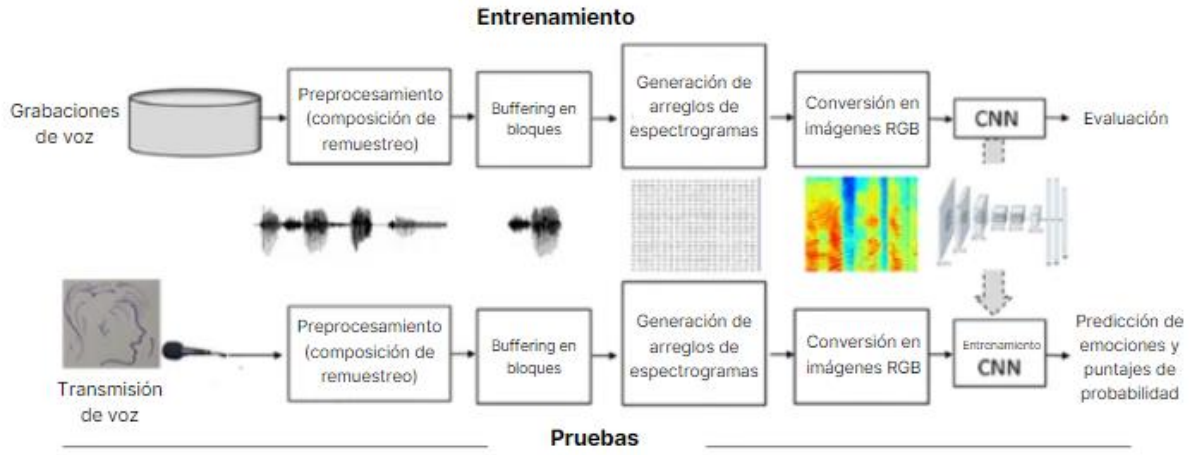
Nota. Tomado de (Wani et al., 2021)

Como se observa en la Figura 18, para poder entrenar dicho modelo es necesario de una base de datos (Database o Dataset), en un contexto de clasificación y entrenamiento de un

sistema de aprendizaje automático (ML), la calidad y flexibilidad de la base de datos es de vital importancia para evaluar una correcta implementación del sistema, para que el proceso de evaluación sea llevado a cabo se requiere el preprocesamiento de los datos a través de bloques. (Wani et al., 2021)

Figura 18

Proceso de Entrenamiento



Nota. Tomado y traducido de (Lech et al., 2020)

Señales de voz en Reconocimiento de Emociones del Habla

Existen distintos tipos de información que podemos obtener a través de las señales del habla. (Wani et al., 2021) Estos tipos de información incluyen:

- Identificación de voz, que nos brinda detalles sobre el contenido de las señales de voz.
- Reconocimiento del hablante, que nos proporciona información acerca de la identidad de la persona que habla.
- Detección de emociones, que nos revela el estado emocional del hablante.
- Evaluación de salud, que nos ofrece información sobre el estado de salud del individuo.
- Identificación de idioma, que nos indica el idioma que está siendo hablado.

- Detección de acento, que nos muestra el acento del hablante.
- Estimación de edad, que nos da información sobre la edad aproximada del hablante.
- Determinación de género, que nos indica el género del hablante.

Coeficiente Cepstrales de Frecuencia de Mel (Mel Frequency Cepstral Coefficients - MFCC)

Estos Coeficientes (MFCC) son un método para extraer características relevantes a partir de audio usado ampliamente en la actualidad en el área del machine Learning (Abdul & Al-Talabani, 2022)

Durante el reconocimiento de sonidos el Coeficiente Cepstral de Mel (MFCC) se enfoca en la potencia a corto plazo emitida por un sonido, La frecuencia de Mel se compone de un cepstrum mapeado[2], indicando que la frecuencia se encuentra espaciada dentro de la escala de Mel dando como resultado una mejor representación del sonido. (González et al., 2019)

Como lo menciona González et al. (2019) el Coeficiente Cepstral de Mel considera:

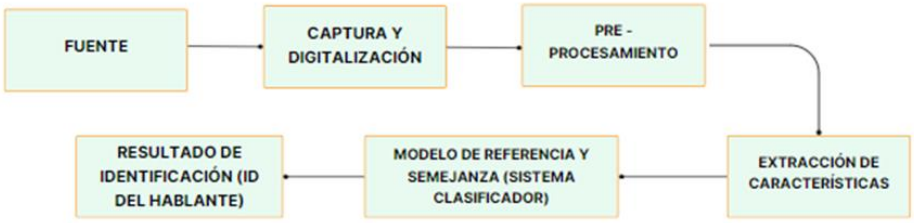
- **Ventaneo y Tramas:** Durante el procesamiento de señales es necesario incorporar la función de ventana definida como, este proceso indica la secuencia real de una longitud usada al percibir una señal y que posee un valor de cero fuera del intervalo al que pertenece.
- **Escala de Mel:** En este punto se puede percibir la frecuencia de un sonido para la realización de futuras comparaciones con otras frecuencias.

La Figura 19 menciona el proceso en bloques de un sistema de reconocimiento de locutor, desde la fuente hasta los resultados del hablante identificado.

[2] Cepstrum se refiere a la modulación del dominio de la frecuencia. Es usado comúnmente en el procesamiento del habla, ya que permite obtener información sólida del hablante.

Figura 19

Proceso de Reconocimiento de Locutor



Nota. Tomado de (González et al., 2019)

IA Multimodal

Jiang et al., (2020) señala que con el avance constante de la Inteligencia Artificial han surgido nuevas interacciones entre el ser humano y el ordenador, considerando a la interacción humana como principal conector entre el hombre y el ordenador. A lo largo del tiempo la evolución de la Inteligencia Artificial ha permitido el reconocimiento emocional, varias investigaciones que abarcan el reconocimiento de emociones mencionan que es necesario la interpretación e implementación centrado en reconocimiento monomodal, incluyendo reconocimiento textual, voz, y sobre todo señales fisiológicas emitidas por el ser humano.

Por otra parte, la aplicación de técnicas multimodales entabladas para el reconocimiento de emociones ha logrado captar la atención de investigadores durante la realización de sus estudios bajo el uso de computación afectiva, orientado un conjunto de datos, investigadores consideran necesario tener en cuenta el monitoreo de la salud en tiempo real, analizando a profundidad el conjunto base de datos bajo la interferencia de emociones multimodales, esto incluye expresiones faciales, extracción de voz y sobre todo características que permitan el reconocimiento humano. (Jiang et al., 2020)

Dentro del campo de Inteligencia Artificial el reconocimiento multimodal es el proceso que permite combinar una o varias etiquetas bajo un banco amplio de información usando entradas de texto, voz, gestos, audio y video. En cuanto al desarrollo de este proceso abarca

múltiples modalidades a fin de combinar diversas técnicas con relación a la variedad de datos manejados. Dentro del campo multimodal existen los mencionados reconocedores multimodales los cuales son usados para procesar y clasificar la información, cabe mencionar que los reconocedores permiten obtener una representación en cuanto al reconocimiento independiente fusionando técnicas directas o indirectas y permitiendo la extracción de información específica para posteriores entrenamientos de modelos existentes dentro del campo de IA. (Jiang et al., 2020)

Ciclo de Vida del Aprendizaje Automático (Machine Learning)

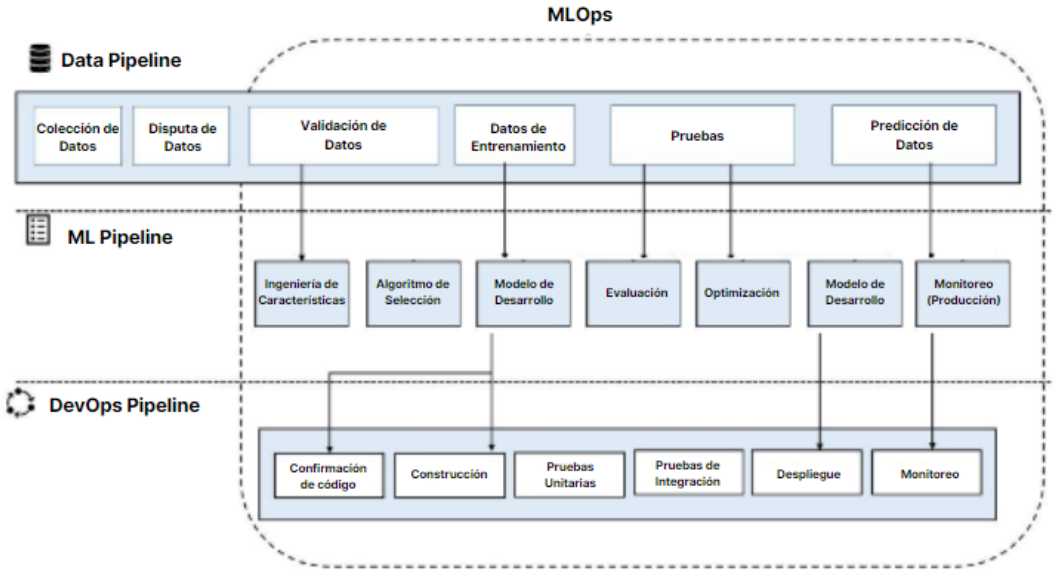
De acuerdo a la definición brindada por Ashmore et al. (2022), el ciclo de vida del Aprendizaje Automático (Machine Learning Operations - MLOps) se refiere al proceso utilizado para el desarrollo y la integración de modelos ML en un sistema completo.

El ciclo de vida consta de varias etapas, detalladas como: Requisitos del modelo, Recopilación y Preprocesamiento de Datos, Diseño y Desarrollo, Evaluación e Implementación y Monitoreo. La efectividad de un modelo de Aprendizaje Automático (ML) depende en gran medida de la calidad de los datos utilizados. Los datos se pueden obtener de fuentes privadas o de código abierto a través de encuestas o experimentos. (Hewage & Meedeniya, 2022)

Antes del entrenamiento, los datos deben depurarse y preprocesarse para abordar imprecisiones y redundancias. Luego se aplican técnicas de ingeniería para identificar características importantes durante el diseño del modelo. Los procesos de ajuste y optimización de Hiperparámetros tienen lugar antes del entrenamiento real. Se mantiene un repositorio para administrar modelos y código base. El código se somete a una etapa de compilación utilizando prácticas de DevOps al confirmarse en el repositorio. Ver Figura 20

Figura 20

Vista de proceso de alto nivel de MLOps



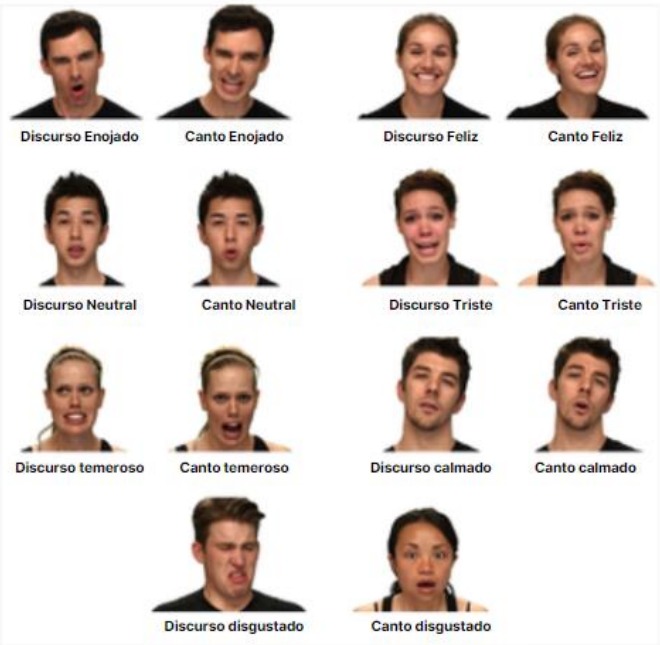
Nota. Tomado y traducido de (Hewage & Meedeniya, 2022)

Base de Datos Audio Visual Ryerson de Canto y Habla Emocional (RAVDESS)

El conjunto de Datos Audiovisuales (Ryerson AudioVisual Database of Emotional Speech and Song - RAVDESS) es un set de datos multimodal de habla y canto emocional, este set de datos cuenta con expresiones emocionales tanto en habla como en canto. El conjunto de datos que compone el Dataset RAVDESS consta de 24 actores profesionales cuyo lenguaje es de origen inglés, cada uno de ellos realiza 104 vocalizaciones únicas con un conjunto de seis emociones consideradas culturalmente universales, entre ellas se detallan: felicidad, tristeza, enojo, temor, sorpresa, disgusto, tranquilidad y neutralidad, como se observa en la Figura 21. (Livingstone & Russo, 2018; Poria et al., 2019)

Figura 21

Emociones encontradas en RAVDESS



Nota. Tomado y traducido de (Livingstone & Russo, 2018)

El set de datos audiovisuales proporcionados por RAVDESS fue validado por un grupo de 247 participantes, muestra propia del set de datos, evaluando cada uno un subconjunto del corpus completo en términos de precisión, intensidad y autenticidad. Además, se contó con la participación de otro grupo de 72 participantes para llevar a cabo pruebas y repeticiones de los datos, con el objetivo de evaluar su confiabilidad. (Siddiqui et al., 2022)

Base de datos de emoción expresada audiovisual de Surrey (SAVEE)

El conjunto de datos de emoción expresada Audiovisual de Surrey (Surrey Audiovisual Expressed Emotion - SAVEE) es una colección de grabaciones de audio y video de expresiones emocionales actuadas por actores profesionales. Está diseñado para facilitar la investigación en el campo del reconocimiento de emociones y el procesamiento del habla. (Avots et al., 2019)

El conjunto de datos captura una amplia gama de emociones representadas por actores masculinos, proporcionando recursos valiosos para estudiar las expresiones emocionales. El set de datos propios de (SAVEE) consta de imágenes de 4 actores masculinos británicos con

seis emociones básicas (disgusto, ira, felicidad, tristeza, miedo y sorpresa), además de un estado neutral. Ver Figura 22.

Figura 22

Características faciales en el conjunto de datos SAVEE

Nota. Ejemplo de extracción de características faciales en el conjunto de datos SAVEE. Tomado de (Wang, 2011)

Base de Datos del Habla Emocional Mexicana (MESD)

La Base de Datos de Expresiones Emocionales Mexicanas (MESD) ofrece enunciados de una sola palabra para los estados emocionales de enojo, disgusto, miedo, felicidad, neutral y tristeza, en español. El MESD ha sido pronunciado tanto por actores no profesionales adultos como por niños: en el corpus existen 3 voces femeninas, 2 voces masculinas y 6 voces infantiles. (Duville et al., 2022)

Conjunto de discurso emocional de Toronto (TESS)

El Conjunto de discurso emocional de Toronto (Toronto emotional speech set; TESS) consta de 7 emociones cardinales para la clasificación por audio, este mismo cuenta con un total de 2.800 estímulos de audio. Cada estímulo corresponde a una palabra pronunciada en un tono emocional específico. Un conjunto total de 200 palabras objetivo fue pronunciado mediante la frase "Di la palabra _____" por dos actrices (de 26 y 64 años) representando las 7 emociones universales (enojo, disgusto, miedo, felicidad, sorpresa agradable, tristeza y neutralidad). (Pichora-Fuller & Dupuis, 2020)

Modelo Transformador (Transformers Model)

Según menciona Raschka et al., (2022) los Transformadores o también conocidos como Transformers han evolucionado encontrándose a la vanguardia de muchas aplicaciones tales como la traducción automatizada de idiomas y procesos de Inteligencia Artificial, abarcando mecanismos básicos de atención y autenticación dentro de su arquitectura.

El modelo de Transformación comprende los siguientes puntos:

- Mecanismos de atención para mejorar el comportamiento de las Redes Neuronales Recurrentes (Recurrent Neuronal Networks - RNN).
- Mecanismos de autoservicio independiente.
- Comprensión de su arquitectura.
- Lenguaje de Transformación a gran escala.
- Clasificación de opiniones.

Considerando los mecanismos de atención que componen a los transformadores se puede establecer el beneficio alcanzado en las redes neuronales convolucionales recurrentes a fin de analizar la secuencia de entrada antes de realizar el proceso de traducción. Es importante mencionar que la descripción del diagrama se realiza automáticamente.

De igual forma Chollet (2021) señala que un modelo transformer abarca un conjunto de vectores capturando atención neuronal, es decir permite transformar cada vector en una representación, el conjunto de vectores representa una secuencia ordenada aprovechando su codificación posicional en lo que respecta al modelo transformer, no obstante, es capaz de procesar párrafos extremadamente grandes indicando mayor eficiencia entorno a las redes neuronales recurrentes o convnets.

Los transformadores son usados para procesamiento de datos incluyendo clasificación textual, se componen de dos partes:

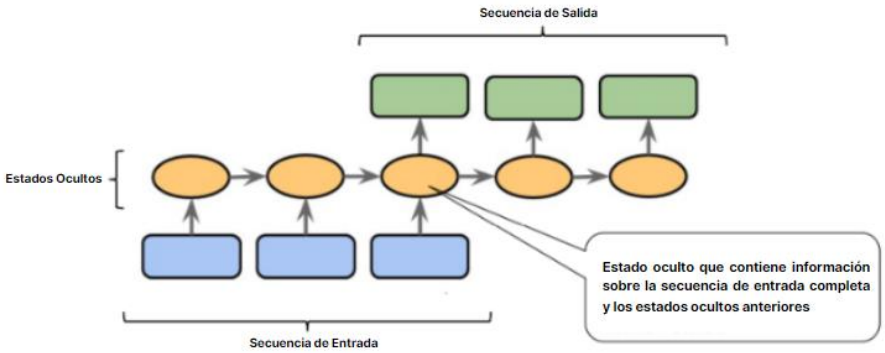
- Transformer Encoder: Transforma un conjunto de vectores de entrada en vectores de salida sensibles a un orden absoluto.

- Transformer Decoder: Permite predecir un suceso a futuro tomando en cuenta el flujo de datos destino.

La Figura 23 muestra el desarrollo de un mecanismo de atención en una Red Neuronal Recurrente tradicional (RNN) para una tarea realizada secuencia tras secuencia, incluye estados ocultos y tareas bajo secuencias de entrada y salida. (Raschka et al., 2022)

Figura 23

Arquitectura Codificador – Decodificador



Nota. Decodificador basado en un modelo secuencia tras secuencia. Tomado y traducido de (Raschka et al., 2022)

En otros términos, la construcción de modelos de aprendizaje profundo requiere mayor complejidad, dado que implica el manejo de datos de entrenamiento a través de cálculos matemáticos y transformación geométrica, los modelos de clasificación y regresión tradicional han llevado a cabo el desarrollo del aprendizaje automático. (Chollet, 2021)

Sin embargo, la asignación de datos permite la realización de predicciones en base a resultados fijos, para eso es necesario realizar el mapeo de datos vectoriales considerando:

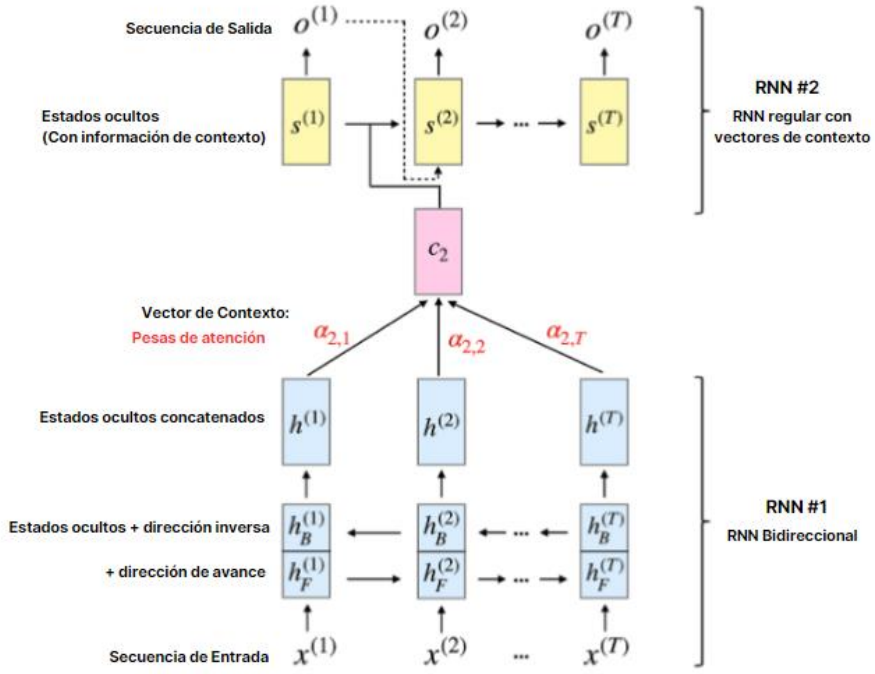
- Asignación de imágenes bajo predicción.
- Asistencia inteligente controladas por Inteligencia Artificial.
- Asignación de datos temporales.

En primera instancia el modelo basado en una Red Neuronal Recurrente (RNN) aborda un conjunto de vectores bidireccionales, considerando a un vector de contexto como una

versión aumentada en cuanto a un vector de entrada, incorporando información bajo un mecanismo de atención. Sin embargo, se puede observar una secuencia de entrada regular hacia adelante como hacia atrás. La captura de este tipo de información depende de las entradas actuales considerando la secuencia antes y después de una oración predeterminada. El segundo elemento de secuencia se puede observar el estado oculto de un pase hacia delante y viceversa, tomando en cuenta que el ejemplo contiene información bajo dos direcciones. Ver Figura 24:

Figura 24

Mecanismo de atención en una Red Neuronal Recurrente (RNN)



Nota. Tomado y traducido de (Raschka et al., 2022)

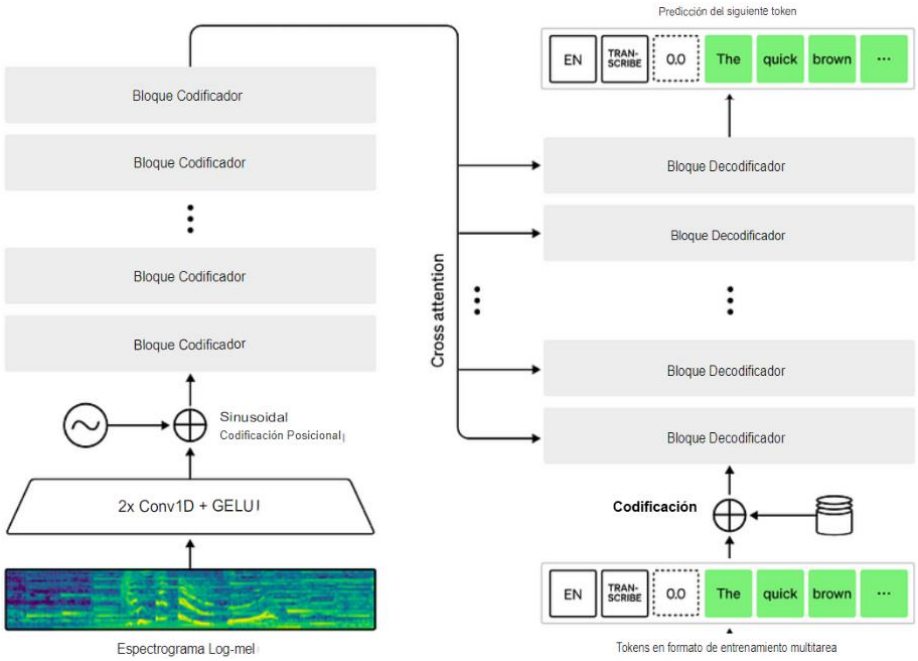
Chollet (2021) indica que a partir del año 2017 la Arquitectura de Transformación (Transformer) ha evolucionado a fin de dar un giro en cuanto a tareas de procesamiento de lenguaje natural. Así mismo apareció la denominada atención neuronal bajo el uso y aplicación de modelos de secuencia que no presentaban redes neuronales recurrentes ni otras capas de convolución.

Susurro (Whisper)

Whisper (Radford et al., 2023), es un modelo de código abierto centrado en la tarea del Reconocimiento Automático de Voz (Automatic Speech Recognition; ASR)[3] desarrollado por OpenAI, está basado en secuencias del modelo Transformador, apilando bloques codificadores y decodificadores basados en el mecanismo de atención el cual permite la propagación de información entre ambos, como se observa en el flujo de la Figura 25, Whisper funciona codificando el audio con la sección del codificador (Encoder), luego guarda la posición de cada trabajo dicho y captura esta información usando el decodificador (Decoder), que predice los denominados tokens[4], que básicamente son las palabras que están siendo dichas, luego se repite este proceso para cada palabra usando la información de las palabras anteriores, ayudando al proceso de reconocimiento de texto.

Figura 25

Arquitectura utilizada en Whisper



Nota. Arquitectura Whisper para el reconocimiento automático de voz. (Radford et al., 2023)

[3] "Reconocimiento Automático de Voz" (ASR) se refiere a la tecnología empleada para convertir las palabras que se dicen en voz alta en texto escrito.

[4] Un token es una secuencia de caracteres de algún documento de texto en particular.

Whisper es una herramienta versátil que se puede utilizar para entrenar en una variedad de actividades relacionadas con el procesamiento del habla. Estas incluyen tareas como el reconocimiento de múltiples idiomas en el habla, la traducción de discursos, la identificación de los patrones lingüísticos en el habla y la detección de la actividad del habla en datos de texto. Esta amplia gama de capacidades hace que Whisper sea especialmente útil para el reconocimiento de emociones. Al combinar datos de texto y señales de audio, es posible detectar las emociones de forma multimodal de forma cohesiva. (Y. Li et al., 2023)

Lenguaje de programación Python

Python es un lenguaje usado dentro del área de la programación, se caracteriza por ser un lenguaje de código abierto analizado e interpretado a medida que va ejecutándose, de esta manera los programas desarrollados dentro del entorno de Python son mucho más rápidos ahorrando tiempo de ejecución. Este lenguaje de programación es compatible con varias plataformas y sistemas operativos debido a la variedad de librerías usadas continuamente por desarrolladores en el campo de Aprendizaje Automático (Machine Learning). (Pineda Pertuz, 2022)

El entorno de Python está comprendido por:

- Variables: Una variable es definida como un espacio de memoria dentro de un programa, adquiere un valor definido por el programador siendo posible su modificación según se requiera.
- Tipos de Datos: Dentro del entorno de desarrollo, Python maneja datos enteros, reales y complejos, así también cadenas o arreglos finitos de elementos, este tipo de datos son manejados de manera dinámica durante el desarrollo del programa.
- Operadores: Son definidos como un conjunto de símbolos usados durante la realización de operaciones incluyendo variables y diversos tipos de datos. Los operadores existentes en Python son de 3 tipos, comprendidos en: Operadores Aritméticos, Relacionales y Lógicos.

Como se observa en la Tabla 1 los operadores aritméticos son aquellos que implican valores o números, su resultado dependen de los operadores que intervengan durante el proceso de programación.

Tabla 1

Operadores Aritméticos

Operador	Descripción	Ejemplos
+	Suma	b = a + 5
-	Resta	b = a - 5
*****	Multiplicación	b = a * 2
/	División Real	c = b / a
//	División Entera	c = b // a
%	Residuo	c = b % 2
°°	Exponenciación	c = b °° 2

Nota. Tomado de (Pineda Pertuz, 2022)

La Tabla 2 comprende operadores relacionales ya sean de tipo número, carácter o datos que represente un valor verdadero o falso.

Tabla 2

Operadores Relacionales

Operador	Descripción
==	Igual que
!=	Diferente que
<	Menor que
>	Mayor que
<=	Menor o igual que
>=	Mayor o igual que

Nota. Tomado de (Pineda Pertuz, 2022)

La Tabla 3 muestra operadores lógicos, este tipo de operadores requieren de la interacción de una tabla de verdad, de esta manera los resultados obtenidos serán comprendidos como valores verdaderos o falsos.

Tabla 3

Operadores Lógicos

Operador	Descripción
Not	Negación
And (y)	Y Lógico
Or (o)	O Lógico

Nota. Tomado de (Pineda Pertuz, 2022)

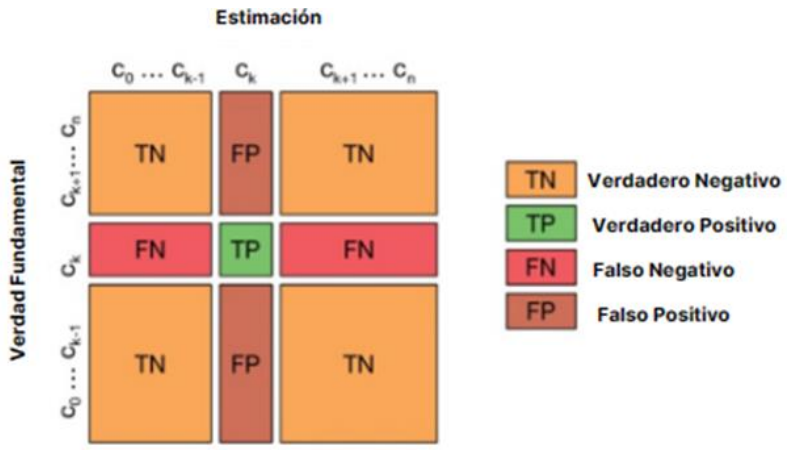
Métricas de Evaluación de modelos de Aprendizaje Automático

Según Griera i Jiménez (2022) menciona que durante el proceso de reconocimiento de emociones se establecen métricas y estándares a seguir a fin de evaluar el rendimiento de un modelo de Aprendizaje Profundo y los problemas presentados durante el proceso de entrenamiento bajo clasificación de los resultados obtenidos. Durante el proceso se establece una matriz de confusión como método de Aprendizaje Supervisado, esta matriz determina un problema con filas las cuales representan la clase real y las predicciones realizadas sobre el modelo definido

En la Figura 26 se puede observar el comportamiento de una matriz de confusión sobre una clasificación multiclase determinada.

Figura 26

Matriz de confusión bajo clasificación multiclase



Nota. Tomado y traducido de (Griera i Jiménez, 2022)

Una matriz de confusión se caracteriza por manejar 4 tipos de procesos de clasificación, entre ellos se menciona:

- Verdaderos Positivos (TP): Predicciones reales realizadas sobre las clases k.
- Negativos Verdaderos (TN): Predicciones reales no determinadas parte de la clase k.
- Falsos positivos (FP): La predicción realizada sobre una clase k no interfiere en el proceso realizado por una clase real.
- Falsos negativos (FN): La predicción no recae sobre la clase k, pero sí sobre la clase real.

La matriz de confusión durante la realización de un proceso permite considerar métricas de forma cuantitativa y medible. Sin embargo, la precisión, recuperación de datos y predicción de un modelo determinan de manera porcentual la fiabilidad y validez del sistema ante condiciones y espacios predefinidos. (Griera i Jiménez, 2022)

Precisión y Sensibilidad. La sensibilidad durante el proceso de reconocimiento emocional se enfoca en las emociones propias del individuo, fue diseñada para medir e interpretar niveles de personalidad en el ser humano bajo limitaciones psicométricas, emocionales y neuronales. Este proceso logra captar las reacciones positivas y negativas dando como resultado métricas de precisión durante la evaluación de desempeño de un modelo de reconocimiento. (Guarino et al., 2005)

La precisión dentro del área de reconocimiento de emociones se caracteriza por predecir modelos o sistemas diseñados para procesos de reconocimiento, en su definición establece valores positivos y negativos permitiendo la consideración de métricas entre datos a evaluar y sus factores.

Puntación F1 (F1 Score). El Puntaje F1 se refiere al promedio ponderado de los valores de precisión y sensibilidad y se lo calcula también a partir de la matriz de confusión, el criterio

de evaluación del puntaje hace uso de los falsos positivos y negativos como se observa en la fórmula, un buen puntaje F1 nos indica que el modelo clasificador que tenemos tiene una baja cantidad de falsos negativos y positivos. (Alakus & Turkoglu, 2020)

$$Puntaje\ F1 = 2\frac{Presición * Sensibilidad}{Sensibilidad + Presición} \quad (2)$$

Tecnoestrés (Estrés Laboral)

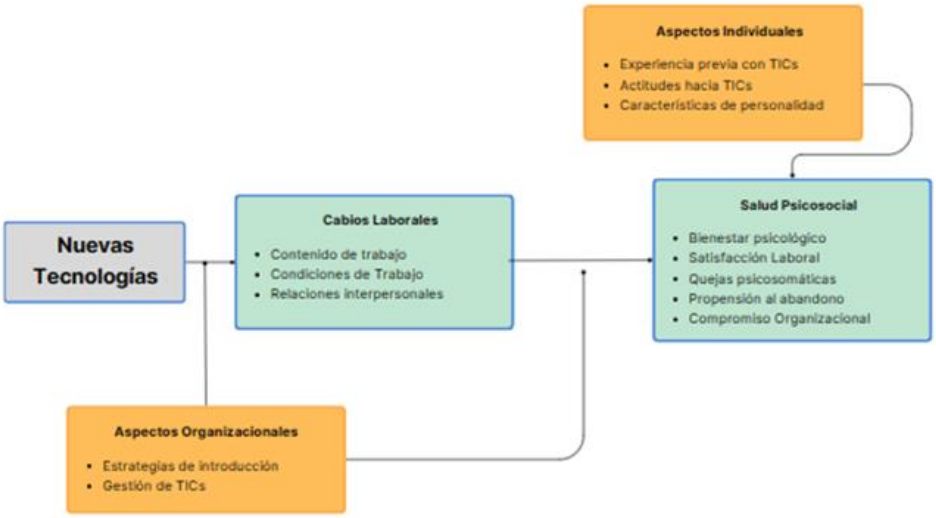
El uso prolongado de tecnologías de la información ha traído como consecuencia grandes afectaciones en torno al ambiente laboral perjudicando cambios emociones durante el proceso y cumplimiento de sus actividades diarias. La Tecnología se ha considerado como una herramienta que facilita y automatiza tareas, minimizando recursos y tiempo de ejecución a medida que los individuos realizan sus actividades mejorar el nivel de productividad incluyendo mayor responsabilidad y desempeño. (Velasquez & Paz, 2020)

En ciertas ocasiones los aspectos no suelen ser únicamente positivos, para muchos individuos el uso y manipulación constante de herramientas tecnológicas podría causar afectaciones en salud, disminuyendo el desempeño de las actividades y causando insatisfacción laboral, dentro de estas consideraciones surge el concepto de Tecnoestrés, concepto ligado directamente a efectos negativos comprendidos por el uso de TICs. (Velasquez & Paz, 2020)

En la Figura 27 se puede apreciar el proceso de Tecnoestrés donde influyen características, aspectos, cambios y afectaciones en el área de salud referente al uso de nuevas tecnologías.

Figura 27

Proceso de Tecnoestrés



Nota. Tomado de (Salanova et al., 1999)

El Tecnoestrés está definido como el proceso ligado a la necesidad en niveles extremistas de estar conectado con el mundo tecnológico, logrando mayor impacto en relación con el cuadro de salud, no obstante, los efectos podrían traer consecuencias positivas como niveles de satisfacción y experiencia al usar herramientas tecnológicas actualizadas, pero a su vez también podrían surgir efectos negativos tales como cansancio, fatiga, ansiedad, debilidad entre otras consecuencias. (Velasquez & Paz, 2020)

Según Velasquez & Paz (2020) menciona que las afectaciones negativas referentes a Tecnoestrés pueden agravar el cuadro psicológico y físico dentro del ambiente laboral. Para comprender mejor el término existen 3 dimensiones que comprende el Tecnoestrés:

- **Dimensión Afectiva:** Comprende emociones o sentimientos ante posibles cambios de entorno, estados de ánimo, cambios de humor, estrés generado por TICs o sintomatología de fatiga y cansancio denominado como tecno ansiedad.
- **Dimensión Actitudinal:** Dentro del área de la salud el Tecnoestrés influye a gran escala en cambios actitudinales de los usuarios, refiriéndose a una actitud escéptica

durante el uso de herramientas tecnológicas trayendo consecuencias negativas durante la realización de sus actividades.

- **Dimensión Cognitiva:** Este tipo de dimensión consiste en la eficiencia o ineficiencia por parte de un individuo al momento de realizar una tarea determinada, es decir durante un proceso de transición los individuos son obligados a ajustarse a cambios no previstos, generando dudas en sus capacidades y desempeño laboral.

Streamlit

Streamlit es considerada como una biblioteca de Python que permite a los usuarios la creación de páginas web interactivas desarrolladas bajo la implementación de código simple. Streamlit cuenta con un servicio gratuito permitiendo alojar el código a través de un repositorio de GitHub, completando sus resultados de manera correcta. (Parker et al., 2021)

Gradio

Gradio es un paquete de Python de código abierto el cual genera una interfaz de usuario orientada a Aprendizaje Automático. Gradio se caracteriza por su facilidad de acceso a cualquier modelo y entorno de Aprendizaje Automático. Sin embargo, Gradio se basa en un sistema de entrevistas dirigido por varios investigadores admitiendo la comparación de una interfaz a través de la entrada del dominio como la interfaz de usuario. (Abid et al., 2019)

Gradio permite compartir información a sus colaboradores de manera segura sin necesidad de herramientas o software externo, cumpliendo el comportamiento del ciclo entre el dominio y sus investigadores. Sin embargo, después del entrenamiento del modelo la interfaz cuenta con 4 parámetros establecidos, 2 de ellos corresponden a los datos de entrada y salida, el siguiente parámetro es el tipo de modelo comprendido entre Keras, Pytorch o Sklearn y finalmente comprende el modelo real a usar durante la función de procesamiento. (Abid et al., 2019)

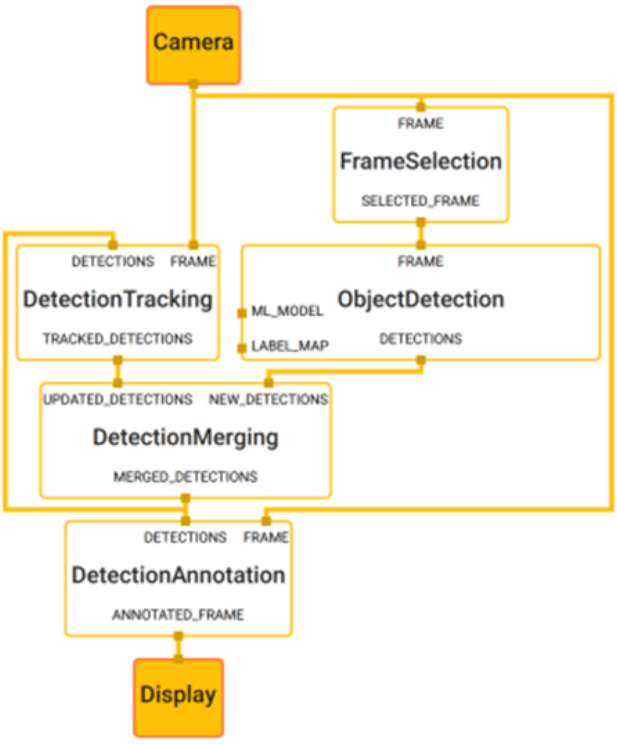
Comparativa de Algoritmos de Reconocimiento Facial (MediaPipe - HaarCascade)

MediaPipe es un framework o marco de trabajo diseñado para construir canalizaciones basadas en datos sensoriales arbitrarios. A través de su modo gráfico se puede incluir modelos, algoritmos de procesamiento, transformación de datos, flujos de audio y video bajo puntos específicos de referencia facial. MediaPipe está diseñado en un entorno de Aprendizaje Automático (Machine Learning) permitiendo la creación automática de prototipos de tubería de percepción. (Lugaresi et al., 2019)

La Figura 28 muestra el proceso de detección de un objeto implementando MediaPipe, cada marco de referencia (fotograma) muestra un nodo, mientras que los datos de entrada y salida están conectados directamente a los nodos a través del flujo de datos. (Lugaresi et al., 2019)

Figura 28

Proceso de Detección usando MediaPipe.



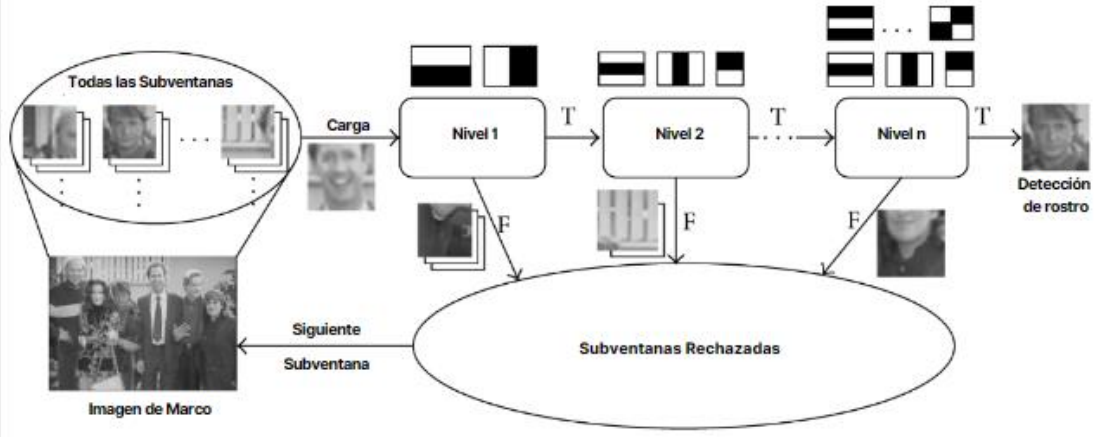
Nota. Tomado de (Lugaresi et al., 2019)

Haar Cascade es un método de Aprendizaje Automático concentrado bajo imágenes positivas y negativas, este algoritmo es desarrollado por Paul Viola y Michael Jones para el proceso de detección de objetos. Dentro de las operaciones que integra la entidad Haar Cascade se encuentra el proceso de clasificación en cascada de imágenes. Aquí se busca detectar rostros y expresiones faciales enfocadas en una imagen determinada. Las imágenes analizadas permiten la extracción de características sumando la composición de pixeles, de esta manera se busca calcular el tamaño de una imagen predeterminada. (Shetty & Rebeiro, 2021)

Como se observa en la Figura 29 el proceso en cascada por clasificadores (Haar Cascade) incluye una imagen de entrada actualizada y un registro que permite almacenar un conjunto de imágenes definidas por pixel, las imágenes son cargadas por niveles de manera que permita la detección de rostros.

Figura 29

Estructura en cascada para clasificadores Haar



Nota. Tomado y traducido de (Kim et al., 2015)

Interfaz de programación de aplicaciones (API) en Aprendizaje Automático (Machine Learning)

Interfaz de Programación de Aplicaciones o más conocido como API permite intercambiar funciones y datos de una aplicación a usuarios, establece un valor, pero durante la

realización de una tarea puede verse propensa a fallas. Es difícil que los desarrolladores de software realicen todas las posibles combinaciones de APIs, es por ello por lo que ha evolucionado empleándose dentro del campo de Aprendizaje Automático (Machine Learning) de manera eficiente, este tipo de tecnología permite desarrollar sistemas bajo Interfaz de Programación de Aplicaciones legibles. (Bahrami et al., 2018)

Una API es un conjunto de funciones y protocolos que permiten a los desarrolladores crear sistemas funcionales bajo estándares predefinidos, de esta manera el sistema desarrollado busca interactuar con el sistema operativo y otros programas compatibles, sin embargo, una API consiste en agilizar un proceso complejo a través de una interfaz de usuario manejable. (Quinaluiza Arias, 2018)

Biblioteca de visión artificial de código abierto (OpenCV)

OpenCV o también conocido como Biblioteca de código abierto de Visión por Computador proporciona infraestructura que permite el desarrollo de aplicaciones bajo el uso de licencias de distribución de software, esta biblioteca integra 2500 algoritmos que incluyen aprendizaje automático. Los algoritmos desarrollados bajo OpenCV buscan detectar movimientos en base al comportamiento real humano. (Morcillo Vizuete, 2020)

OpenCV se caracteriza por emplear en su desarrollo lenguajes de programación tales como C++, Python, JAVA y MATLAB compatible con variedad de sistemas operativos, esta biblioteca está dirigida a entonos de visión basados en tiempo real y puede adaptarse con otras bibliotecas. Una de las ventajas que ofrece OpenCV es optimizar el código a medida que su funcionamiento se de en segundo plano y finalmente ofrecer legibilidad y simpleza durante el proceso de programación. (Morcillo Vizuete, 2020)

Procesamiento del lenguaje natural (NLP)

El Procesamiento de Lenguaje Natural (NLP) se enfoca en procesar la información a partir de una entrada, permite entablar comunicación entre el lenguaje natural con el ordenador a través del procesamiento de textos proporcionados a medida que se desarrollan modelos que

interpretan el comportamiento humano. El área que cubre el Procesamiento de Lenguaje Natural va desde la lingüística computacional a procesos de Inteligencia Artificial simulando la capacidad humana al resolver problemas. (Moreira et al., 2021)

Ventajas del Procesamiento de Lenguaje Natural:

- El Procesamiento de Lenguaje Natural no obliga al sistema a aprender a diferencia de lenguajes de programación que dificultan el uso de sistemas diseñados para procesamiento de información.
- NLP incluye una interfaz amigable, intuitiva y sencilla de manejar.
- Permite resolver las necesidades de los usuarios de forma automatizada optimizando tiempos de ejecución en relación con otro tipo de interfaces.
- Comprende el proceso humano – máquina a fin de interpretar su comportamiento y automatizarlo.

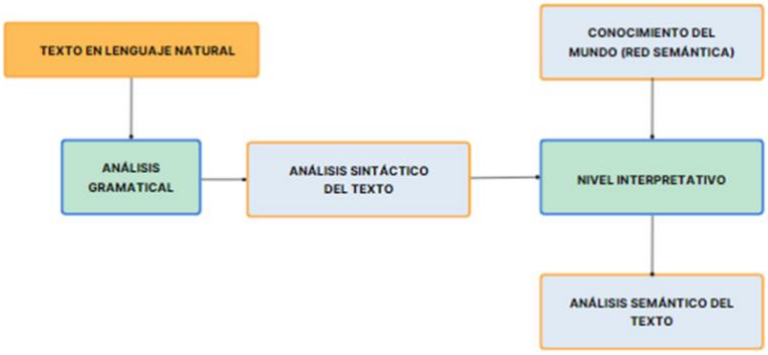
Desventajas del Procesamiento de Lenguaje Natural

- El Procesamiento de Lenguaje Natural tiende a ser limitado en cuanto a entendimiento del comportamiento humano.
- El proceso de comunicación puede verse afectado cuando NLP no logra captar el razonamiento del usuario a fin de entenderlo y satisfacer sus necesidades.
- Es necesario entender el Lenguaje Natural para mejorar el proceso de toma de decisiones.

En la Figura 30 se puede apreciar un modelo de Procesamiento de Lenguaje Natural el cual incluye un texto de entrada, depende de su nivel gramatical y análisis sintáctico para que el texto sea interpretado y comprendido. (Moreira et al., 2021)

Figura 30

Procesamiento de Lenguaje Natural



Nota. Procesamiento de Lenguaje Natural Tomado y Adaptado de (Moreira et al., 2021)

Librería React

React es una librería propia de JavaScript la cual permite crear interfaces que interactúan con el usuario, se inició en el año 2011 por el ingeniero de software Jordan Walke, siendo de código abierto y ganando popularidad debido a su versionamiento a lo largo del tiempo. (Saks, 2019)

En el 2015 apareció una biblioteca funcional para dispositivos móviles denominada React Native y dos años después una biblioteca funcional en el campo de realidad virtual denominada React360. React en su trayectoria ha sido desarrollada por Facebook, incluye una interfaz interactiva de fácil uso. Sin embargo, dentro de la programación React consiste en dividir sus componentes creando espacios o interfaces más complejas bajo lenguaje de programación JavaScript permitiendo el flujo de información de forma más rápida y eficiente. (Saks, 2019)

Modelos de Representaciones de codificador bidireccional (BERT)

Dentro del Procesamiento de Lenguaje Natural se diseñó un modelo de Representación de Codificador Bidireccional (BERT) el cual permite solucionar problemas presentados en un conjunto de datos escasos. Al manejar corpus con datos extensos durante el desarrollo de una tarea específica se requiere de la reducción de manera notable de los datos de entrenamiento

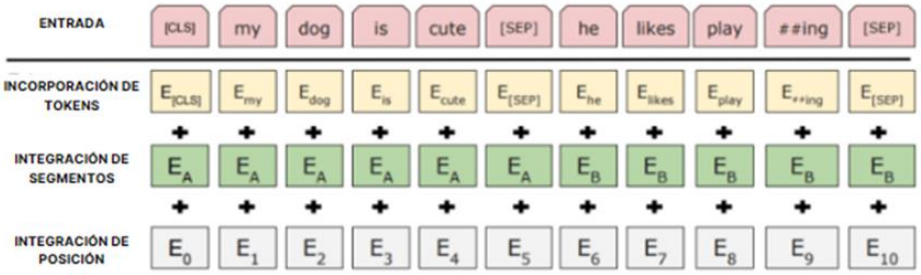
generando dificultad en modelos manejados bajo Aprendizaje Profundo para llegar al resultado esperado. (Auquilla Vicuña & Mora Álvarez, 2022)

BERT es considerado como un mecanismo de pre-procesamiento manejado bajo un lenguaje general, cumpliendo grosso modo los requerimientos establecidos durante la transferencia de aprendizaje. El proceso de entrenamiento que incurre en el Modelo de Representación de Codificador Bidireccional (BERT) permite la extracción profunda de un contexto determinado. (Auquilla Vicuña & Mora Álvarez, 2022)

En la Figura 31 se puede apreciar el texto de entrada quien posteriormente es transformado bajo una secuencia de tokens añadiendo etiquetas y metadatos de acuerdo con su segmentación y posición.

Figura 31

Secuencia de entrada de datos para BERT.



Nota. Tomado de (Auquilla Vicuña & Mora Álvarez, 2022)

PyTorch

Pytorch es una biblioteca de código libre cuya funcionalidad consiste en eliminar obstáculos que impiden a investigadores desarrollar algoritmos basados en métricas de Aprendizaje Profundo. La biblioteca Pytorch está desarrollada de manera flexible permitiendo a los usuarios realizar variedad de combinaciones de algoritmos, el flujo de trabajo incluye datos de entrenamiento y de prueba dando resultados más eficientes. (Musgrave et al., 2020)

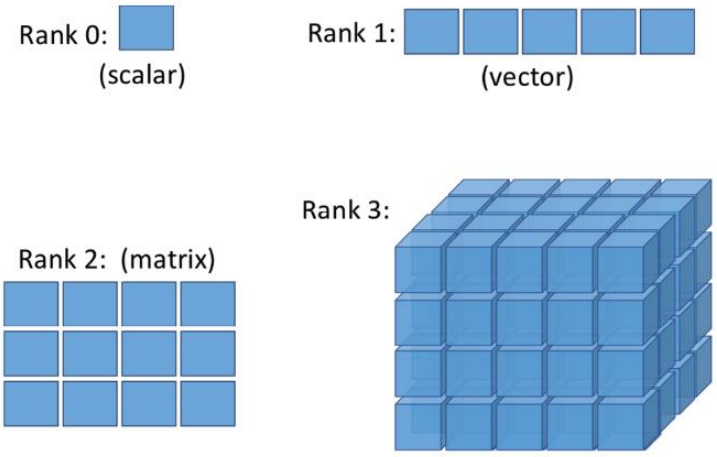
Paszke et al., (2017) menciona que Pytorch distingue dos características esenciales, comprendidas en:

- Ejecución dinámica: Ejecuta un cálculo deseado dentro de un entorno dinámico. Pytorch se caracteriza por trabajar bajo marcos dinámicos a diferencia de la biblioteca de TensorFlow.
- Ejecución inmediata: Este proceso ejecuta cálculos sin materializar un gráfico directo, este gráfico es creado durante el preso de entrenamiento por cada interacción.

En la Figura 32 se puede contemplar rangos de tensores comprendidos entre 0, 1,2 y 3.

Figura 32

Tensores en Python.



Nota. Tomado de (Raschka et al., 2022)

TensorFlow

A partir del año 2015 Google lanzó TensorFlow de código abierto, es considerado como una biblioteca de software la cual facilita el Aprendizaje Automático al momento de definir, entrenar y desplegar los resultados de un modelo de aprendizaje predefinido. Durante el desarrollo del Aprendizaje Profundo y su interés por manejar grandes conjuntos de datos se descubren métodos que implican una Función de Activación Lineal Rectificada (ReLU) y una serie de bibliotecas que facilitan a desarrolladores crear software basado en Aprendizaje Automático, el conjunto de bibliotecas fue ampliado a partir del año 2015 por TensorFlow,

siendo este una interfaz que maneja algoritmos de aprendizaje de sistemas distribuidos. (Goldsborough, 2016)

TensorFlow es una biblioteca flexible al momento desarrollar, siendo una herramienta compatible con variedad de aplicaciones e infiriendo en el desarrollo de redes neuronales profundas y sistemas distribuidos heterogéneos. (Abadi et al., 2017)

El lenguaje de programación que maneja TensorFlow está basado en C++ permitiendo la comparación de datos de manera estadística y bajo patrones técnicos. TensorFlow representa algoritmos de Aprendizaje Automático a través de gráficos computacionales. Sin embargo, un gráfico computacional constituye un flujo de datos contemplados por vértices o nodos durante una operación, estos se denominan tensores. Un tensor depende de su tamaño y dimensión para mejor ejecución de un modelo. (Goldsborough, 2016)

Google Colaboratory

Google Colaboratory (también conocido como Colab) es un servicio en la nube basado en Jupyter Notebooks [5] para difundir la educación y la investigación del aprendizaje automático. Proporciona un tiempo de ejecución completamente configurado para el aprendizaje profundo y el acceso gratuito a una Unidad de procesamiento gráfico (Graphics processing unit; GPU)[6] robusta. (Carneiro et al., 2018)

Transformada de Fourier

Es una herramienta usada para para el procesamiento de señales de audio, está definida bajo parámetros de tiempo y frecuencia, se ajusta a las necesidades del usuario modificando el tono de un sonido determinado. Esta transformada se basa en una ecuación que indica la señal de entrada en relación con el tiempo, seguido de la longitud M de la ventana de Hamming y R considerado como el tamaño de salto de la muestra, durante el cálculo, una vez

[5] Jupyter Notebooks es una aplicación web de código abierto que permite crear y compartir documentos que contienen código en vivo

[6] Procesador formado por muchos núcleos más pequeños y especializados en el área de machine Learning y procesamiento de gráficos

obtenida la venta se fija la frecuencia, de esta manera se puede extraer características espectrales de una frecuencia determinada. (Chamba & Jiménez, 2022)

$$X_m(w) = \sum_{n=-\infty}^{\infty} X|n|w(n - mR)e^{-jwm} \quad (3)$$

Pysentimiento

Pysentimiento es un conjunto de herramientas y modelos basados en transformaciones para el análisis de sentimientos y emociones, se puede encontrar en idioma español e inglés. Así también esta biblioteca es de código abierto, trabaja bajo entorno gratuito abarcando gran cantidad de datos utilizados, para realizar el proceso de entrenamiento y la obtención de resultados es necesario incluir dentro de la codificación la biblioteca Pysentimiento de esta manera se pretende obtener resultados seguros y confiables. (Pérez et al., 2021)

Dentro del modelo de Transformadores (Transformers) Pysentimiento es usado para analizar y comprender texto bajo el conjunto de un set de datos determinado, de esta manera se permite entrenar al modelo desarrollado bajo técnicas de Aprendizaje Supervisado, todo esto dirigido al análisis de sentimientos, así también Pysentimiento permite etiquetar de manera automática los datos durante el proceso de entrenamiento. (Carbajal Bacilio & Suarez Mariscal, 2023)

Entropía Cruzada y Funciones de Pérdida (Loss) en Machine Learning

El término Entropía Cruzada o también conocido como Cross Entropy fue propuesto a partir del 2004 por Rubinstein dando solución a un problema de estimación de probabilidades, a lo largo de su evolución llegó a constituirse como una herramienta importante para la solución de problemas de optimización complejos. (Helene Bischel, 2015)

Durante la identificación de problemas de optimización el algoritmo de Entropía Cruzada realiza una búsqueda aleatoria bajo probabilidades de ocurrencia para cada variable

determinada. Los casos de uso que contempla la Entropía Cruzada facilitan optimizar problemas de producción, es decir que si se requiere dentro de una organización determinar un conjunto de factores óptimos es recomendable aplicar el algoritmo, este proceso es ampliamente utilizado en áreas estadísticas, Aprendizaje Automático (Machine Learning) y Procesamiento de Lenguaje Natural (NLP). (Sánchez & Ramírez, 2013)

De forma matemática la entropía cruzada entre distribuciones de probabilidad m y n se define bajo la siguiente fórmula:

$$H(\mathrm{m},\mathrm{n}) = \sum_i m(i)\log(n(i)) \qquad (4)$$

Aquí el índice (i) es el encargado de recorrer todos los posibles eventos, m(i) indica la probabilidad asociada al evento determinado (i) y n(i) consiste en la predicción realizada por el evento (i).

La función de pérdida también denotada como función de costo o de error está orientada a la pérdida de un conjunto de datos, identificando el costo de pérdida total, es decir el promedio sobre todo el conjunto. (Raschka et al., 2022)

Optimizador Adam

El Optimizador Adam es un algoritmo diseñado para realizar búsquedas en menor tiempo bajo una ruta de conexión con aplicaciones desarrolladas por Inteligencia Artificial (IA). Abarca un conjunto de funciones reales o vectoriales de forma aleatoria, durante el proceso de ejecución se debe llevar a cabo un proceso continuo que permita llegar a la optimización de los resultados. (Aguilar et al., 2021)

Este tipo de algoritmos emplea un conjunto de estrategias heurísticas dando solución a un problema basado en 4 pasos que son: Compresión del problema, Plan de solución, Ejecución del Plan y Análisis de Resultados. Entre sus ventajas se menciona simplicidad, eficiencia, consumo menor de memoria, fácil calibración y aplicación ante problemas complejos,

así también se indica que el algoritmo ha sido desarrollado bajo lenguaje de programación Python ocupando el entorno de desarrollo de Google Colab dando éxitos en sus resultados y solución ante problemas complejos. (Aguilar et al., 2021)

Estado del arte

Para lograr analizar el estado del arte basado en reconocimiento emocional es necesario realizar la revisión de literatura respectiva. Paul Ekman en sus investigaciones propone 7 expresiones emocionales universales (Figura 34), dentro de su teoría Paul Ekman relaciona 8 emociones básicas, tales como enojo, desprecio, asco, felicidad, tristeza, miedo y sorpresa, adicional menciona una emoción más considerada como neutral (Figura 33). El análisis de estas emociones es realizado bajo un ambiente controlado y bajo ángulos diferentes. Una vez obtenido el conjunto de imágenes, se determina en un 90% la realización del entrenamiento y en un 10% las pruebas necesarias logrando como resultado captar expresiones únicas que determinen como se encuentra un individuo emocionalmente.

Figura 33

Codificación de emociones

Etiqueta	Traducción	Código
Angry	Enojo	1
Contemptuous	Desprecio	2
Disgusted	Asco	3
Fearful	Miedo	4
Happy	Feliz	5
Neutral	Neutral	6
Sad	Tristeza	7
Surprise	Sorpresa	8

Nota. Tomado de (Barrionuevo et al., 2020)

Figura 34

Estados emocionales propuestos por Paul Ekman

Nota. Tomado de (Barrionuevo et al., 2020)

Planificación de la revisión

Identificación de la necesidad de una revisión

Para la realización del fundamento teórico, es necesario definir el problema central del presente estudio, para esto se analiza las posibles brechas que permitan identificar patrones y puntos críticos de una persona durante sus actividades diarias. Sin embargo, es necesario realizar una exhaustiva investigación previa de artículos relacionados a reconocimiento emocional, de esta manera se pretende extraer información necesaria que ayude a entender y alcanzar los objetivos plateados inicialmente en este documento.

Especificación de las preguntas de investigación

Se plantean tres interrogantes, mismas que serán usadas durante el proceso de revisión sistemática de literatura:

RQ1. ¿Qué técnicas multimodales están siendo utilizadas actualmente para reconocer y analizar emociones a partir de datos combinados de audio, video y texto?

RQ2. ¿Cuáles son las principales ventajas y desafíos de aplicar el reconocimiento multimodal de emociones en contextos del mundo real, como en interacciones sociales o en sistemas de asistencia inteligente?

RQ3. ¿Cómo se comparan los enfoques de reconocimiento unimodales (por ejemplo, solo audio o solo video) con los enfoques multimodales en términos de precisión y robustez para identificar y comprender las emociones de manera más efectiva?

Desarrollo de un protocolo de revisión

Criterios de inclusión y exclusión

Los artículos analizados fueron comprendidos a partir del año 2018, de esta manera se buscó dar el enfoque primordial a investigaciones que contemplen dentro de su contexto análisis y reconocimiento emocional, bajo la implementación y uso de técnicas multimodales.

El proceso de revisión de literatura fue realizado bajo los siguientes puntos:

- Artículos cuyo contenido se encuentre inmerso en el proceso de reconocimiento emocional bajo técnicas multimodales.
- Artículos que permitan clasificar, comparar y evaluar las diversas técnicas y herramientas utilizadas para el proceso de reconocimiento.
- Artículos cuya revista de publicación o conferencias se encuentren en un cuartil Q3 o superior.

Selección de estudios primarios

Para la selección de artículos de investigación relacionados al tema propuesto, se opta por considerar los siguientes criterios de inclusión:

- Idioma: Español – Inglés
- Año: 2018 – 2023
- Tipo de publicación: Revistas y Conferencias

En base a los puntos mencionados anteriormente y considerando los criterios de los diversos investigadores, se procede al análisis respectivo de 5 artículos primarios, los cuales permiten realizar la sustentación respectiva para realizar el estado del arte. Los estudios considerados se muestran en la siguiente Tabla 4:

Tabla 4

Artículos Primarios

Artículos Publicados	Título	Cita
AP1	Framework multimodal emocional en el contexto de ambientes dinámicos.	Ierache, J., Sattolo, I., & Chapperón, G. (2020). Framework multimodal emocional en el contexto de ambientes dinámicos. Revista Ibérica de Sistemas e Tecnologias de Informação, 40, 45–59.
AP2	Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educación emocional.	Ierache, J. S., & Elkfury, F. (2021). Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educción emocional. In XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja).
AP3	Sistema integrado de reconocimiento de emociones para interacción hombre – máquina.	Martínez, F. H., Hernández, C. A., & Giral, D. A. (2023). Sistema integrado de reconocimiento de emociones para interacción hombre-máquina. Información tecnológica, 34(1), 117-128.
AP5	Use of the Student Engagement as a Strategy to Optimize Online Education, Applying a Supervised Machine Learning Model Using Facial Recognition	Noboa, A., Gonzalez, O., & Tapia, F. (2022). Use of the Student Engagement as a Strategy to Optimize Online Education, Applying a Supervised Machine Learning Model Using Facial Recognition. 283–295.

Revisión de la Documentación

Ierache, J., Sattolo, I., & Chapperón, G. (2020). Framework multimodal emocional en el contexto de ambientes dinámicos. Revista Ibérica de Sistemas e Tecnologias de Informação, 40, 45–59.

Este estudio presenta métodos, modelos e instrumentos en lo que respecta a computación afectiva, permitiendo diseñar un framework multimodal basado en la captura y combinación de diversas fuentes de datos, a través del modelo diseñado se busca ofrecer mayor fidelidad y precisión en sus resultados. Su enfoque va dirigido a entornos simulados o virtuales recreando escenarios basados en hechos de la vida diaria. Durante el desarrollo del framework se consideró la captura de ondas cerebrales bajo un modelo BCI (Neurosky / Emotiv), adicionalmente consideraron tomar capturas de pantalla a rostros de individuos determinados a fin de analizar su estado emocional bajo el establecimiento de las 7 emociones universales propuestas por Paul Ekman, de esta manera en su propuesta y durante la realización de las pruebas los investigadores buscan llegar a establecer autocorrección y equilibrio en los individuos testeados en lo referente a su comportamiento emocional.

Ierache, J. S., & Elkfury, F. (2021). Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educción emocional. In XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja).

El siguiente artículo busca desarrollar un clasificador de emociones que permite identificar el estado emocional de un individuo a través del análisis de voz. Debido a la falta de implementación en lo referente a reconocimiento emocional los investigadores buscan desarrollar una API a través de la cual se cargue la data proporcionada por el usuario, bajo el uso y manipulación de redes neuronales recurrentes RRN y redes neuronales convolucionales CNN, a medida que el modelo es entrenado las redes neuronales son ejecutadas y testeadas con muestras de audios extraídos en diferentes conjuntos de datos, tales como: INTERS1P (ELRA) y EMOFILM. El presente artículo se enfoca en 2 ambientes basados inicialmente en

descartar emociones negativas para obtención de mayor probabilidad. Sin embargo, el clasificador fue desarrollado bajo el modelo de Russel el cual permite obtener las coordenadas necesarias para la integración a las ocho emociones propuestas dentro de un modelo categórico.

Martínez, F. H., Hernández, C. A., & Giral, D. A. (2023). Sistema integrado de reconocimiento de emociones para interacción hombre-máquina. Información tecnológica, 34(1), 117-128.

El presente estudio comprende un modelo entrenado para la detección de los estados emocionales de una persona través de la voz bajo la implementación en un sistema integrado humano-robot. Para la realización del modelo el enfoque principal va dirigido a una red neuronal convolucional denominada Dense Convolutional Network (DenseNet) la cual permite clasificar imágenes y procesar archivos audiovisuales. Al mismo tiempo para el entrenamiento del modelo el artículo menciona el uso del conjunto de datos denominado The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) compuesto por 7356 archivos dirigidos a 24 actores clasificados en 12 hombres y 12 mujeres. Los archivos son la extracción de la voz bajo acento americano, gracias al conjunto de datos implementado se puede analizar 8 estados emocionales, los audios son transformados a imágenes bajo el formato MEL de manera que permitan la percepción humana basadas en patrones específicos.

Cordero, J., Aguilar, J., & Aguilar, K. (2019). Enfoques Inteligentes para Identificar Estilos de Aprendizaje de los estudiantes mediante las Emociones en un salón de clases. Revista Ibérica de Sistemas e Tecnologias de Informação, (E17), 703-716.

El siguiente artículo está dirigido a estudiantes enfocado en métodos inteligentes que permitan analizar los diferentes estilos de aprendizaje a través de su comportamiento emocional. El artículo se enfoca en el análisis multimodal a fin de modelar y reconocer el estilo de aprendizaje visual, auditivo, kinestésico o de lectura. Su arquitectura se compone de un patrón bajo el modelo VARK dirigido a un ambiente multimodal (facial, gestual, acústico y

vocal). A través de la realización de un cuestionario compuesto por 16 preguntas se busca definir el estilo de aprendizaje de los estudiantes encuestados, considerando sus reacciones emocionales de acuerdo con sus necesidades.

Noboa, A., Gonzalez, O., & Tapia, F. (2022). Use of the Student Engagement as a Strategy to Optimize Online Education, Applying a Supervised Machine Learning Model Using Facial Recognition. 283–295.

El presente artículo se enfoca en desarrollar un modelo bajo aprendizaje supervisado que permita el reconocimiento facial a través de emociones, a fin de medir el Student Engagement. Este artículo dirige su investigación a un entorno de aprendizaje online, durante el desarrollo usa imágenes de rostros bajo la manipulación de conjuntos de datos de libre acceso, clasificando las emociones y validando los resultados bajo la implementación de redes neuronales convolucionales CNN.

Características del estado del Arte

Los estudios realizados de acuerdo con los diferentes artículos analizados mencionan métodos, técnicas y modelos diversos, enfocados al análisis y comportamiento emocional de un ser humano. Además, hacen uso de un conjunto de datos predeterminados los cuales son extraídos y comprendidos durante el entrenamiento del modelo. Los investigadores en sus artículos dirigen sus estudios a diversos ambientes y consideran técnicas multimodales que permitan determinar resultados óptimos a través de reconocimiento facial, gestual o vocal. Todos los artículos analizados hacen énfasis en lo que respecta al uso de redes neuronales, de esta manera y bajo patrones específicos se busca la implementación del reconocimiento emocional.

Capítulo III

Marco Metodológico

Marco Metodológico

El enfoque del tema propuesto se centra en el reconocimiento emocional a través del diseño de un framework bajo métodos multimodales basados en Deep Learning. El sistema se encarga de combinar un conjunto de datos y procesar la información obtenida de un entorno controlado predeterminado. La metodología usada consiste en analizar varios tipos de sets de datos que permitan la extracción de características emocionales claves como la voz, el video y el texto derivado del audio como datos de entrada, no obstante gracias a la interpretación del modelo bajo aprendizaje profundo se permite comprender a los datos mejorando la precisión de reconocimiento. La estructura (Framework) diseñada permite adentrarse en el campo amplio del reconocimiento emocional abarcando diversas áreas, en este caso orientado al campo de salud ocupacional a fin de ofrecer un resultado que permita mejorar la calidad de vida digital del individuo quien hace uso de herramientas tecnológicas constantes en su diario vivir.

Mediante el desarrollo de la estructura (Framework) propuesta se realizará el proceso de reconocimiento emocional a través de archivos de audio y video, lo archivos serán cargados en el modelo cuyo desarrollo implica la combinación de sets de datos que facilitan el reconocimiento emocional basado en aspectos faciales, vocales, gestuales y no verbales. Se hará uso de herramientas tales como Reconocimiento Emocional Facial (Facial Emotion Recognition - FER) para el reconocimiento del rostro, Faster Whisper para el análisis de sentimientos a través del texto extraído del audio y finalmente la combinación de sets de datos predefinidos como: RAVDEES, SAVEE, MESD y TESS para detección a través del audio. Sin embargo, los datos extraídos serán analizados y validados bajo criterio de un profesional de salud psicológica y ocupacional a fin de determinar la condición del ser humano en cuanto al manejo de herramientas tecnológicas y su afectación emocional en un entorno controlado.

Investigación en Ciencia de Diseño (Design Science Research - DSR)

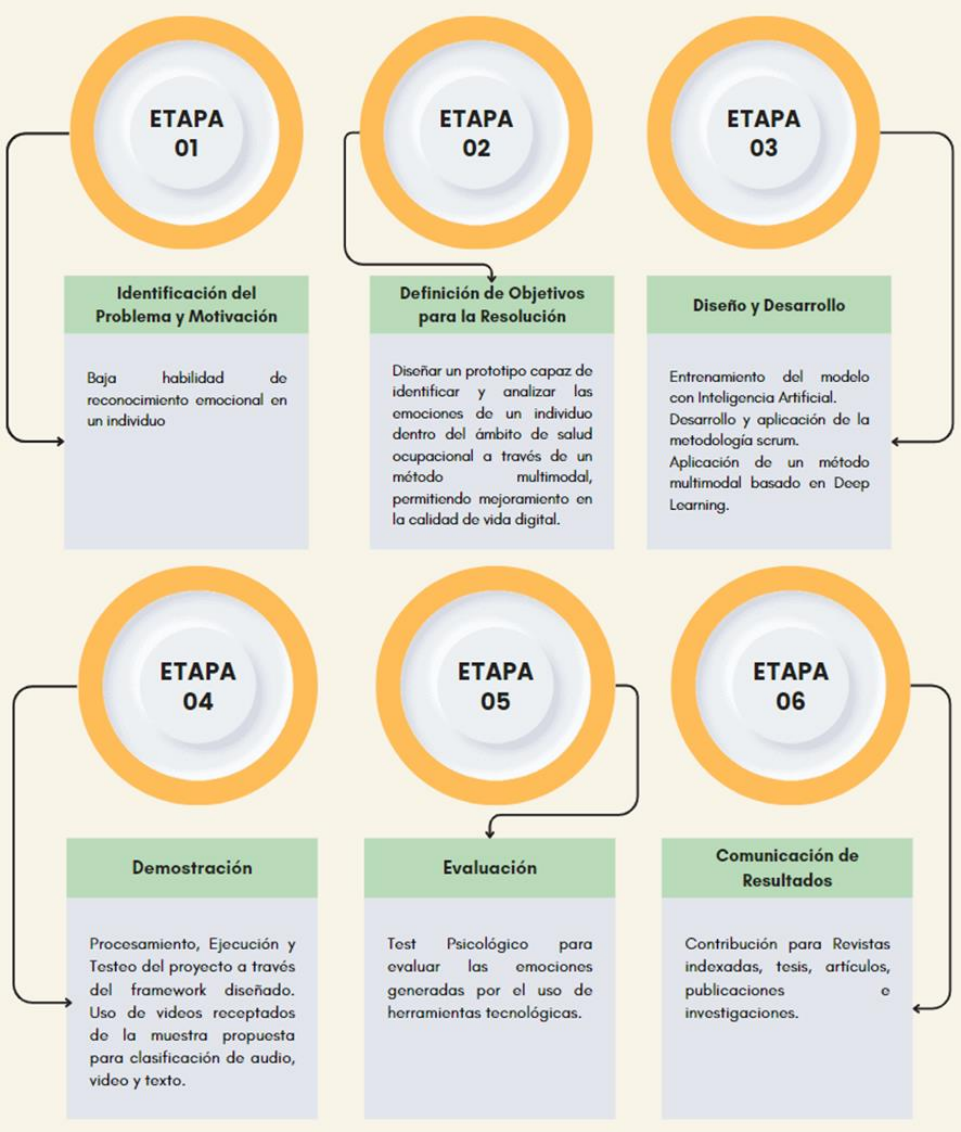
Investigaciones basadas en Ciencias de Diseño (Design Science Research - DSR) ha permitido abarcar áreas de soporte, estrategias, toma de decisiones, modelos y algoritmos computacionales facilitando a los investigadores a ampliar conocimientos bajo la evolución de la ciencia. La investigación basada en diseño incluye metodologías y aportes en cuanto al establecimiento de una planificación a través de la construcción de preguntas que permitan recopilar datos concretos y estructurados. (Hoang Thuan et al., 2019)

El ciclo de vida de la Investigación en Ciencia de Diseño (Design Science Research - DSR) se basa en la construcción, formulación y respuesta de argumentos clave. Sin embargo, el diseño de framework de reconocimiento emocional propuesto en este trabajo pone a consideración 6 etapas basadas en DSR, como se observa en la Figura *35*:

- Identificación del Problema y Motivación
- Definición de Objetivos para la Resolución
- Diseño y Desarrollo
- Demostración
- Evaluación
- Comunicación de Resultados

Figura 35

Etapas de desarrollo del proyecto bajo metodología DSR



Metodología de Desarrollo

A fin de aprovechar el desarrollo de la estructura (Framework) de reconocimiento emocional, se optó por la aplicación de la metodología ágil, este tipo de metodología permite a personas que se encuentran en etapa de investigación enfocar sus resultados de manera rápida y garantizada. A diferencia de metodologías tradicionales que implican mayor tiempo de desarrollo en proyectos de investigación con resultados poco favorables para el investigador.

Metodología Scrum

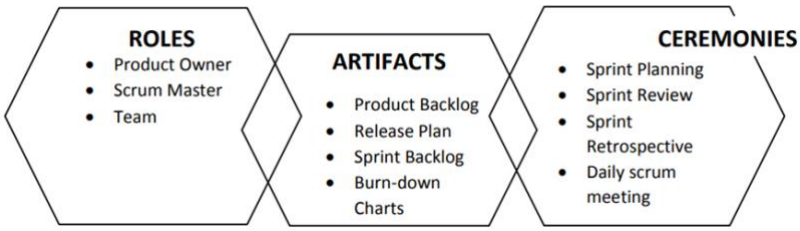
Scrum es una de las metodologías ágiles más comunes y conocidas por investigadores para el desarrollo de proyectos, es usada bajo un amplio entorno, estudios y revisiones previas sistemáticas. Fue propuesto por Schwaber a partir de 1995 incorporando en su desarrollo procedimientos y etapas donde interviene la literatura académica. Es así como Scrum es una técnica versátil usada más allá de sus fines, esta herramienta se caracteriza por usar plataformas adaptadas a diversos contextos y garantizar que los resultados obtenidos sean los esperados. (Hron & Obwegeser, 2022)

La estrategia Scrum permite desafiar entornos tradicionales bajo la intervención de todos sus colaboradores en diferentes disciplinas, esta metodología ágil permite reconocer procesos durante la ejecución de un proyecto bajo planificación y definición del problema, así mismo Scrum se encuentra orientado al desarrollo de proyectos de software o proyectos que requieran complejidad alta, en varias ocasiones Scrum es usado en áreas de programación proporcionando mejorar los resultados obtenidos. (Sachdeva, 2016)

En varias ocasiones la metodología de Scrum se adapta a las necesidades del proyecto, esta adaptación se enfoca en base al entorno original, dentro de sus componentes Scrum engloba una serie de sprints con longitud fija, este tipo de sprints facilitan la construcción del proyecto previa a una retroalimentación. (Sachdeva, 2016)

Figura 36

Componentes Scrum



Nota. Tomado de (Sachdeva, 2016)

Como lo menciona Sachdeva (2016) en su artículo, Scrum es una metodología ágil común en el desarrollo de proyectos bajo un equipo previamente estructurado el cual cumple reglas y planificaciones establecidas, menciona también que durante el ciclo de desarrollo de un proyecto determinado la organización da seguimiento a las etapas que componen este tipo de metodología, poniendo a consideración los componentes mencionados.

Roles de Scrum

En su estructura Scrum no requiere de la intervención de un administrador o un usuario quien lidere la ejecución del proyecto, más bien hace uso de tres roles:

- Product Owner o Propietario del proyecto
- Scrum Master
- Team o Equipo de desarrollo

En la Figura *37* se puede apreciar que el propietario se destaca por ser la persona responsable de levantar y poner en ejecución el proyecto bajo los requerimientos propuestos, el Scrum Master es responsable de cumplir las reglas establecidas y entrenar al resto del equipo y finalmente el Team en su conjunto garantiza la calidad del producto diseñado y sus operaciones.

Figura 37

Roles y relaciones que compone Scrum

Nota. Tomado y traducido de (Sachdeva, 2016)

Artefactos de Scrum

La Figura 36 engloba al conjunto de artefactos que la metodología Scrum usa en el diseño y desarrollo de proyectos, para comprender su funcionalidad, se detallan los siguientes puntos:

- Product Blocklog: Se encarga de enumerar los requisitos del producto desarrollado. En este espacio el equipo analiza las funcionalidades que el producto tendrá cuando se encuentre terminado.
- Release Plan: Abarca las funcionalidades que el usuario espera del producto desarrollado. El plan de lanzamiento incluye los riesgos y costos del proyecto.
- Sprint Blocklog: Se encarga de reunión al equipo durante el proceso de planificación, además cada espacio longitudinal requiere de una descripción y prioridad durante el desarrollo, es importante la realización del seguimiento a las tareas y actividades asignadas a cada miembro del equipo.
- Burn – down: Corresponde a la cantidad de trabajo pendiente presentado en modo gráfico. En su contenido se registra la suma del esfuerzo del equipo durante el desarrollo y ejecución del proyecto.

Ceremonias Scrum

La Figura 36 engloba al conjunto de ceremonias que la metodología Scrum usa durante el proceso de planeación, verificación y puesta en marcha de un proyecto, detallándose de la siguiente manera:

- Sprint Planning: Durante la fase de planificación se estima un tiempo de duración del proyecto, dando cumplimiento a los objetivos propuestos.
- Sprint Review: El proceso de revisión se lleva a cabo bajo una calendarización previa.
- Sprint: El dueño del producto se encarga de realizar la retroalimentación respectiva a fin de solventar dudas existentes por parte del equipo.

- Retrospective: Aquí se entabla una reunión de aproximadamente 2 horas de duración en la cual se proponen puntos en función y mejoramiento del desarrollo del producto.
- Daily Scrum Meeting: En esta etapa cada miembro del equipo es responsable de explicar y cumplir las tareas asignadas durante la fase de diseño y desarrollo del producto.

Capítulo IV

Desarrollo

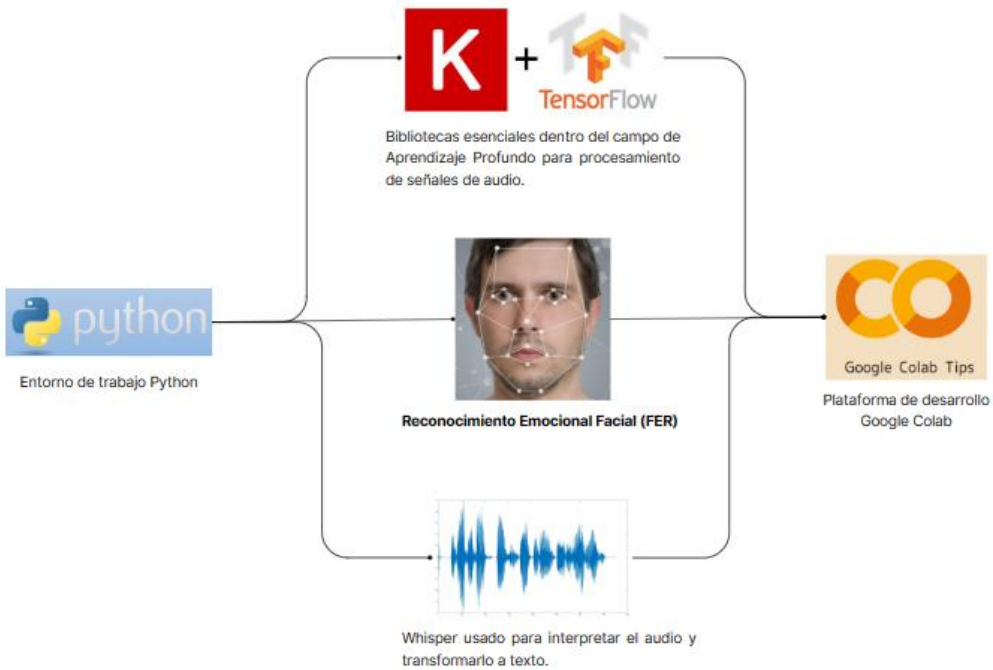
Desarrollo del sistema

Descripción del sistema

El presente proyecto se llevó a cabo en un entorno configurado con Python 3.9.7 y las bibliotecas esenciales, que incluyen TensorFlow, Keras, Librosa (para el manejo de audio), DeepFace (para el reconocimiento de emociones faciales), Whisper (para la traducción de voz) y la biblioteca de Transformadores (Transformers) para los modelos BERT. El entrenamiento del modelo se realizó en el entorno de Google Colab, específicamente utilizando una instancia NVIDIA A100. Esta GPU de alto rendimiento agilizó significativamente el proceso de entrenamiento del modelo, como se ilustra en el diagrama de entrenamiento presentado en la Figura 38.

Figura 38

Descripción del Sistema de Reconocimiento emocional.



El desarrollo del sistema multimodal de Aprendizaje Profundo para el reconocimiento de emociones involucró varios pasos claves. En primer lugar, se recopilaron cuatro conjuntos de datos diferentes: RAVDESS, TESS, SAVEE y MESD, cada uno de los cuales contenía grabaciones de audio y video de actores que expresaban diversas emociones bajo diferentes entornos. Los datos de audio manejados entre los diferentes conjuntos se procesaron con Librosa[7] para extraer características esenciales como los Coeficientes Cepstrales de Frecuencia Mel (MFCC) y los espectrogramas Mel.

Al aprovechar los datos de múltiples fuentes e integrar modelos de Aprendizaje Profundo, el sistema multimodal desarrollado puede estimar una probabilidad de las emociones tanto del audio como de las expresiones faciales, proporcionando una solución integral para las tareas de reconocimiento emocional.

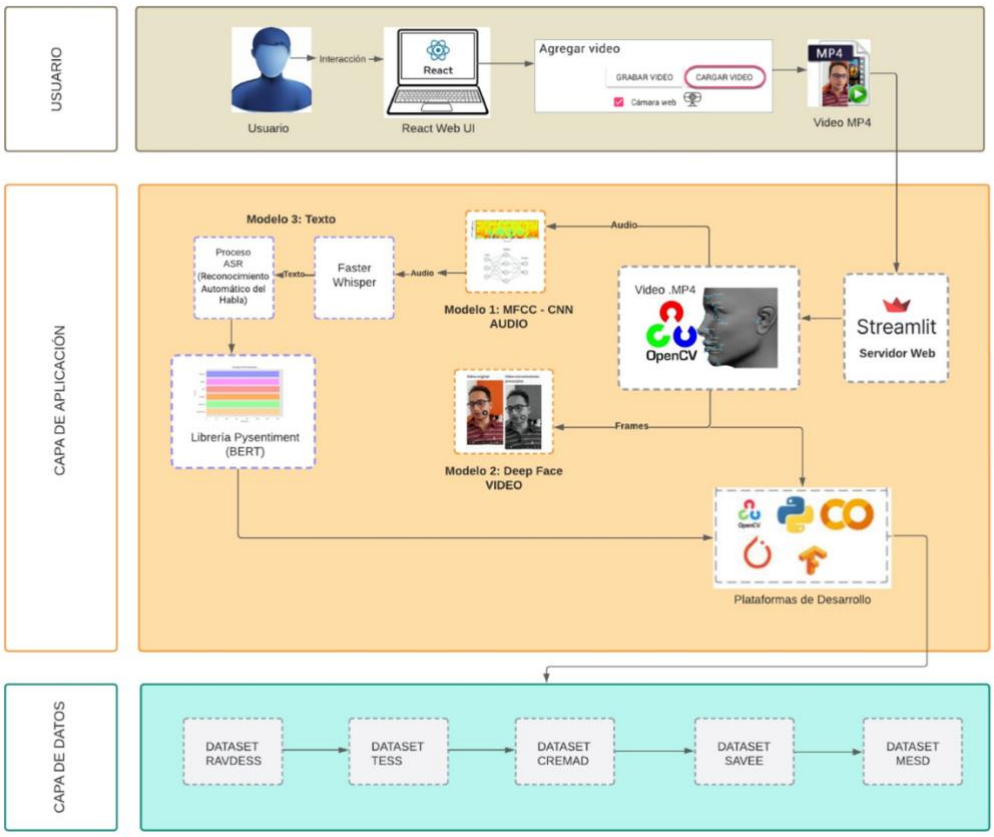
Para integrar la transcripción de voz, se incorporó Faster Whisper, lo que habilita la capacidad de transcripción de voz a texto. Posteriormente después de entrenar y evaluar el modelo de fusión multimodal, integramos un modelo BERT ajustado y entrenado con datos en español para manejar entradas basadas en texto en español obtenidas por Whisper mediante transcripción. Finalmente, implementamos el sistema con una interfaz de usuario, lo que permite a los usuarios ingresar un video para el reconocimiento de emociones, así como usar su cámara web para obtener las emociones.

[7] Librería en Python diseñada para el análisis de música y sonido.

Arquitectura del sistema

Figura 39

Arquitectura de capas del sistema propuesto



Nota. Arquitectura de capas del sistema propuesto

Como se observa en la Figura 39 la arquitectura del sistema se encuentra dividida en 3 capas:

Capa de usuario. En la capa de usuario, React es usado para crear la interfaz de usuario que permite interactuar con la aplicación Streamlit. Esto facilita la integración entre la vista frontal (frontend) de React y el backend creado en Streamlit. El Uso de Streamlit y React en conjunto permite crear una aplicación completa de Aprendizaje Profundo (Deep Learning) para el reconocimiento de emociones utilizando al máximo las ventajas de cada uno, permitiendo la simplicidad para una rápida visualización de datos y la flexibilidad de React para crear interfaces de usuario interactivas.

El componente clave a considerar en la capa de usuario son:

- Diseño de una interfaz amigable mediante el uso e implementación de componentes React, el aplicativo cuenta con 3 opciones dentro de uso, así también existen 3 rutas principales, las cuales permiten: “Subir Video” donde el usuario puede cargar un video en formato mp4, “Live” donde se puede visualizar las emociones en tiempo real y una opción de configuración para configurar opciones respecto a la vista presentada por el modelo predictivo, donde se incluyen opciones de activación o desactivación.

Capa de Aplicación. El backend de Streamlit se encarga principalmente de dividir el video en diversos fotogramas separando los rostros de las personas, para que dichos rostros sean alimentados al modelo de reconocimiento emocional, el cual obtendrá el estado emocional de los mismos. Posteriormente se obtiene el audio a partir del video cargado y se realiza la detección de características MFCC para la alimentación del modelo del reconocimiento emocional de voz, finalmente se utiliza la integración con Whisper para obtener transcripciones del audio cargado referente a un video específico en formato .mp4. Las transcripciones obtenidas son posteriormente procesadas para detectar emociones del texto.

Elementos clave del Backend de Streamlit:

- **Opción de carga y procesamiento de Video:** El backend de Streamlit se encarga del procesamiento un video leyendo sus cuadros y envinándolos al Modelo de MediaPipe para que posteriormente se obtengan sus emociones con DeepFace, convirtiendo cuadros a escala de grises y creando un nuevo video de salida con los cuadros procesados. Es una serie de pasos que transforman el video original en una nueva versión basada en emociones.
- **Integración fácil entre Modelos IA:** Streamlit permite una integración fluida con modelos Tensorflow o Pytorch, permitiendo cargar y realizar predicciones con los mismos, se puede mostrar los resultados de las predicciones del modelo de IA en la

interfaz de usuario utilizando componentes de texto, gráficos o imágenes para visualizar y presentar los resultados.

- **Muestra de resultados:** Streamlit puede funcionar con bibliotecas de visualización de datos populares como Matplotlib, Seaborn y Plotly para crear gráficos y cuadros interactivos y visualmente atractivos de los resultados obtenidos en las predicciones de los estados emocionales de la persona.

Capa de datos. La capa de datos de la aplicación con modelos entrenados en conjuntos de datos RAVDESS, TESS, SAVEE y MESD implicaría la gestión, el preprocesamiento y la organización de los datos de estas diferentes fuentes.

Modelos IA

Los siguientes modelos fueron implementados para realizar la predicción completa de forma multimodal de una persona a través de la combinación de múltiples modalidades como reconocimiento facial, gestual, no verbal y vocal.

Modelo 1 (MFCC – CNN) – Vocal. El Modelo principal para el reconocimiento de emociones por audio se encuentra conformado por un modelo de Redes Neuronales Convolucionales (CNN) entrenado en múltiples conjuntos de datos a partir de los Coeficientes Cepstrales de Mel (MFCC), dichas características extraídas del audio permiten al modelo identificar una representación bastante precisa del discurso en sí, la CNN acepta una entrada de tensores con dimensiones (236,40) y su salida representada por la capa final correspondiente a las 7 emociones propuestas en el modelo contemplando: ira, miedo, neutral, triste, asco, feliz y sorpresa.

Modelo 2 (Deep Face) – Facial. Se refiere al modelo utilizado para el reconocimiento de emociones faciales, se utiliza Deep Face para procesar los fotogramas de los rostros en un formato de 128x128 pixeles y obtener la emoción más significativa a través de fotogramas, el modelo en particular permite la inferencia de 7 distintos estados emocionales: ira, miedo,

neutral, triste, asco, feliz y sorpresa, estas emociones son dibujadas en la cara de la persona mediante el uso de Open CV.

Modelo 3 Whisper. – Verbal. El modelo Whisper permite la transcripción de audio a texto de una manera rápida, este mismo ha sido configurado mediante una API con distintos parámetros, particularmente se especifica el uso del lenguaje español y de una cantidad de parámetros “large-v2” que se refiere al modelo más avanzado disponible, mediante este modelo se logra una gran exactitud a un costo de ejecución bastante bajo.

Modelo 4 Pysentimiento. – Verbal. El texto obtenido por Whisper es enviado al modelo del procesamiento de lenguaje natural, Pysentimiento suministra evaluaciones de sentimientos mediante el empleo de modelos pre entrenados de Transformers para tareas de Procesamiento de Lenguaje Natural (NLP), se obtiene el resultado sentimental de las frases obtenidas en distintas líneas de tiempo para el texto transcrito mediante Whisper.

Esta arquitectura garantiza una funcionalidad completa al permitir que el backend de Streamlit integre servicios de transcripción, análisis de emociones tanto vocal como verbal y detección de rostros, enriqueciendo la experiencia de usuario y proporcionando resultados más precisos.

Implementación de modelo para el reconocimiento de emociones por cara

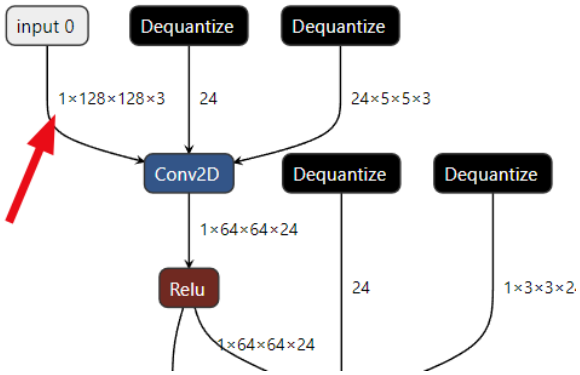
Implementando Reconocimiento Facial con Mediapipe

Mediapipe es un detector Facial de código abierto desarrollado por Google y basado en el Modelo BlazeFace[8], este mismo permite la detección de caras con una precisión medida del 98.61%, el modelo acepta imágenes de la forma de tensores (128,128,3) de 3 dimensiones como se observa en la entrada del diagrama de la red neuronal Figura 40. (Bazarevsky et al., 2019)

[8] BlazeFace es un modelo entrenado en un corpus de imágenes de 66 mil imágenes de personas geo diversas

Figura 40

Diagrama de la entrada del modelo BlazeFace



Nota. Diagrama de la entrada del modelo BlazeFace, donde se tiene que aceptar imágenes de dimensión 128 x 128.

Para realizar la implementación de este modelo, como se observa en el código de la Figura 41, mediante DrawingSpec, dibujamos los puntos faciales de la persona, se utiliza además MediaPipe para detectar rostros en imágenes, dibujar cajas alrededor de las caras detectadas y mostrar las imágenes resultantes con las detecciones dibujadas.

Figura 41

Código - Función para implementar mediapipe

```
import mediapipe as mp

mp_face_detection = mp.solutions.face_detection
mp_drawing = mp.solutions.drawing_utils
drawing_spec = mp_drawing.DrawingSpec(thickness=1, circle_radius=1)

with mp_face_detection.FaceDetection(
    min_detection_confidence=0.5, model_selection=0) as face_detection:
  for name, image in short_range_images.items():
    # Conviert la imagen BGR a RGB y procésela con MediaPipe Face Detect
    results = face_detection.process(cv2.cvtColor(image, cv2.COLOR_BGR2R

    # Dibuja detecciones de cara con MediaPipe.
    print(f'Detecciones de cara del archivo {name}:')
    if not results.detections:
      continue
    annotated_image = image.copy()
    for detection in results.detections:
      mp_drawing.draw_detection(annotated_image, detection)
    resize_and_show(annotated_image)
```

Nota. Función para implementar mediapipe.

El resultado de la detección se muestra en la Figura 42. la imagen obtenida indica la detección facial de una persona mediante Media pipe, mostrando tanto la referencia de los puntos faciales como la representación de su cara.

Figura 42

Detección de puntos de referencia faciales

Nota. Detección de puntos de referencia faciales y rectángulo que representa la cara de una persona correspondiente al set de datos RAVDESS.

Para realizar la implementación de BlazeFace en el aplicativo Streamlit, se realiza una división de Fotogramas o “frames”[9] mediante la utilidad de manejo de videos Open CV VideoCapture, se recuperan el ancho, la altura y los FPS de los cuadros de video y se genera un código para el códec de video MP4. El Ciclo While permite iterar en cada frame del video para su análisis posterior.

[9] Un “Frame” en el contexto de un archivo de video, es una imagen estática que cuando se reproduce en secuencia con otros cuadros del video, crea movimiento.

Figura 43

Código - Función de procesamiento de video en Streamlit.

```
def process_video(self, video_data):
    self.write_bytesio_to_file(self.temp_file_to_save, video_data)
    self.cap = cv2.VideoCapture(self.temp_file_to_save)
    width = int(self.cap.get(cv2.CAP_PROP_FRAME_WIDTH))
    height = int(self.cap.get(cv2.CAP_PROP_FRAME_HEIGHT))
    frame_fps = self.cap.get(cv2.CAP_PROP_FPS)
    fourcc_mp4 = cv2.VideoWriter_fourcc(*'mp4v')
    self.out_mp4 = cv2.VideoWriter(self.temp_file_result,
                                   fourcc_mp4,
                                   frame_fps,
                                   (width, height),
                                   isColor=False)

    while True:
        ret, frame = self.cap.read()
        if not ret:
            break
        emotion_dominant = self.mediapipe_face_detection(frame)
        emotion_label = self.detect_emotion(frame)
        gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
        self.out_mp4.write(gray)
    self.cap.release()
    self.out_mp4.release()
```

Estas propiedades son útiles cuando se trabaja con captura de video y se escriben cuadros con OpenCV, el objetivo es procesar los cuadros de la persona y obtener una representación mediante un rectángulo, para esto se utiliza el FaceDetection de Mediapipe como se observa en el código de la Figura 44.

Figura 44

Código - Uso de mediapipe para obtener la imagen

```
mp_face_detection = mp.solutions.face_detection.FaceDetection(min_detection_confidence=0.5)
results = mp_face_detection.process(image)
if results.detections:
    ih, iw, _ = image.shape
    for detection in results.detections:
        bboxC = detection.location_data.relative_bounding_box
        bbox = int(bboxC.xmin * iw), int(bboxC.ymin * ih), int(bboxC.width * iw), int(bboxC.height * ih)
        x, y, w, h = bbox
        x = max(0, x)
        y = max(0, y)
        w = min(iw - x, w)
        h = min(ih - y, h)
        if w > 0 and h > 0:
            face_image = image[y:y + h, x:x + w]
```

Ciclo de Detección de Caras

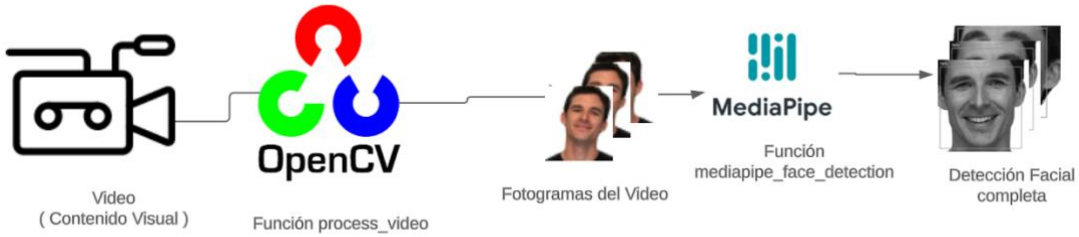
Separar la imagen de cara

Nota. Uso de mediapipe para obtener la imagen de una cara y separar la misma para el posterior análisis emocional.

Se observa la ejecución de este proceso de una forma más visual en la Figura 45:

Figura 45

Flujo de ejecución para la detección facial mediante MediaPipe



Nota. Flujo de ejecución para la detección facial mediante MediaPipe.

Implementando DeepFace

DeepFace es un framework ligero de código abierto para Python que permite el análisis de atributos faciales como son la edad, el género, la emoción y la raza de una persona. Para la imagen cargada Figura 46, se obtendrá la información respecto a la emoción. (Serengil & Ozpinar, 2021)

Figura 46

Código - Imagen correspondiente al set de datos RAVDESS

El modelo usado por DeepFace para la detección facial de emociones se encuentra especificado a detalle en la figura el código define a una arquitectura de red neuronal convolucional con 3 capas de convolución seguidas de capas de agrupamiento (pooling), capas densas con funciones de activación ReLU y capas de Dropout. La red está diseñada para clasificar imágenes de entrada en diferentes clases (emociones en este caso), y la arquitectura se compone de capas que extraen y aprenden características de las imágenes en distintos niveles de abstracción, la última capa densa tiene como salida la probabilidad de las 7 emociones: ira, miedo, neutral, triste, asco, feliz y sorpresa.

Figura 47

Código - Uso de DeepFace

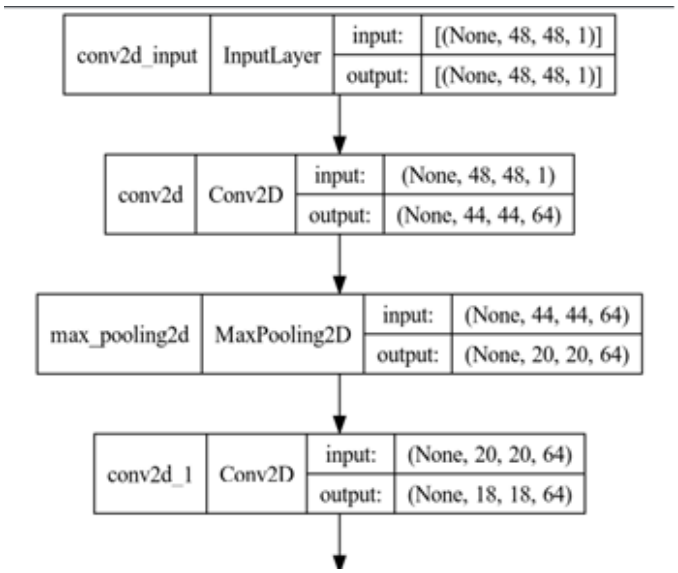
```
# Primera Capa de Convulución
model.add(Conv2D(64, (5, 5), activation="relu", input_shape=(48, 48, 1)))
model.add(MaxPooling2D(pool_size=(5, 5), strides=(2, 2)))
# Segunda Capa de Convolución
model.add(Conv2D(64, (3, 3), activation="relu"))
model.add(Conv2D(64, (3, 3), activation="relu"))
model.add(AveragePooling2D(pool_size=(3, 3), strides=(2, 2)))
# Tercera Capa de Convolución
model.add(Conv2D(128, (3, 3), activation="relu"))
model.add(Conv2D(128, (3, 3), activation="relu"))
model.add(AveragePooling2D(pool_size=(3, 3), strides=(2, 2)))

model.add(Flatten())

model.add(Dense(1024, activation="relu"))
model.add(Dropout(0.2))
model.add(Dense(1024, activation="relu"))
model.add(Dropout(0.2))

model.add(Dense(num_classes, activation="softmax"))
```

Nota. Uso de DeepFace para la detección de la emoción más predominante.

Se observa de manera gráfica el flujo de entrada de la red CNN en la Figura 48 obtenida mediante la librería de VisualKeras, se muestra cómo se conforman las 3 capas de convolución de la red pre entrenada.

Figura 48

Arquitectura usada por Deep Face



Nota. Gráfico representativo de la arquitectura usada por Deep Face para obtener emociones Faciales

Para usar redes pre-entrenada, se puede obtener la emoción dominante mediante el uso de DeepFace.analyze como se muestra en el código de la Figura 49, esta función obtiene las emociones y se opta por no forzar la detección de caras mediante enforce_detection, debido al uso de Mediapipe como algoritmo preferido para encontrar caras, se opta por dejar este parámetro en Falso, se muestra el formato de salida del modelo DeepFace en la Figura 50.

Figura 49

Código - Uso de DeepFace para la detección de la emoción más predominante

```
emotions_list = DeepFace.analyze(
    check_img,
    actions=['emotion'],
    enforce_detection=False
    )
```

Figura 50

Código - Formato de Salida del Modelo Deep Face

```
[{'emotion': {'angry': 1.2228238421840842e-07,
   'disgust': 1.89416665403834e-11,
   'fear': 2.5402655978723487e-05,
   'happy': 99.98505115509033,
   'sad': 2.8090974524275225e-05,
   'surprise': 0.0003444183448664262,
   'neutral': 0.014548888429999352},
  'dominant_emotion': 'happy',
  'region': {'x': 21, 'y': 14, 'w': 78, 'h': 78}}]
```

Para implementar el modelo DeepFace dentro del aplicativo de reconocimiento en tiempo real, se utilizará una función que acepte las imágenes de cara detectadas mediante Mediapipe como se observa en la Figura 51.

Figura 51

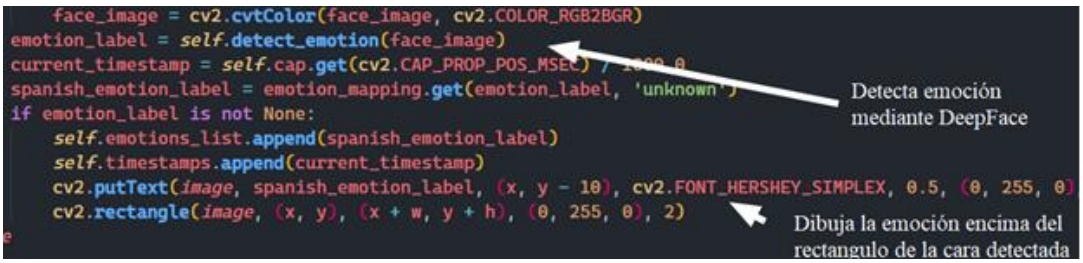
Código - Función usada para detectar las emociones predominantes

```
def detect_emotion(self, face_image):
    if face_image is None or face_image.size == 0:
        return None
    emotions_list = DeepFace.analyze(
        face_image,
        actions=['emotion'],
        detector_backend="skip",
        enforce_detection=False)
    if not emotions_list:
        return None
    emotions = emotions_list[0]
    emotion_label = emotions['dominant_emotion']
    return emotion_label
```

El resultado de la emoción obtenida, se agrega una anotación de texto al cuadro de imagen de entrada. La etiqueta de emoción en español se muestra justo encima de la esquina superior izquierda del cuadro delimitador alrededor de la cara detectada, el código de la Figura 52 detalla los métodos usados para lograr esto.

Figura 52

Código - Detección continua en Tiempo Real de emociones mediante Deep Face



El resultado del ciclo mencionado anteriormente se observa en la Figura 53, donde se obtiene un video completo con las emociones obtenidas mediante las expresiones faciales.

Figura 53

Detección continua en Tiempo Real de emociones mediante Deep Face

Desarrollo de modelo para el reconocimiento de emociones por audio

Recolección de datos iniciales. Se consideraron diversos conjuntos de datos para llevar a cabo la recopilación de información, incluyendo TESS, RAVDESS, SAVEE y MESD. Todos estos conjuntos de datos están vinculados al reconocimiento de expresiones emocionales del audio y consisten en muestras de voz en las que se pronuncian diferentes

expresiones emocionales. En la siguiente Tabla 5, se proporciona una descripción detallada de los conjuntos de datos empleados:

Tabla 5

Descripción detalla del corpus de datos a utilizar para el análisis, validación y el entrenamiento del modelo SER.

Set de Datos	Año	Contenido	Emociones	Formatos	Tamaño	Lenguaje	Libre Acceso
The Ryerson Audio-Visual Database of Emotional Speech and Song (**RAVDESS**)	2018	7356 grabaciones de 24 actores M = 26,0 años; SD = 3,75; rango de edad = 21–33; 12 hombres y 12 mujeres	8 emociones enojo tranquilidad miedo tristeza sorpresa felicidad disgusto neutral	Audio y Video	24.8 GB	Inglés	Si
MESD Mexican Emotional Speech Database	2021	864 archivos de audio en WAV, 3 voces femeninas, 2 masculinas y 6 infantiles	6 emociones: ira Disgusto Miedo felicidad, neutralidad tristeza	Audio	89.74 MB	Español Mexicano	Si
TESS Toronto emotional speech set	2010	2800 archivos, dos actrices femeninas (26 y 64 años)	7 emociones Ira Disgusto Miedo Felicidad Sorpresa Agradable Tristeza Neutral	Audio	281.33 MB	Inglés	Si
SAVEE Surrey Audio-Visual Expressed Emotion	2011	480 expresiones en inglés británico interpretadas por 4 actores masculinos	6 emociones: Ira Asco Felicidad Tristeza Sorpresa Neutral	Audio y Video	65 MB	Inglés	Si

El objetivo es desarrollar un modelo de reconocimiento de emociones que pueda generalizar correctamente, por esta razón los conjuntos de datos vienen de grabaciones de la mayor cantidad de gente posible, más datos ayudan al modelo a aprender una gama más amplia de patrones y características de diferentes escenarios.

Formateo de datos

Para el entrenamiento y validación del modelo se utilizarán principalmente archivos .WAV y .MP4 para el manejo de audio, y además se usarán archivos .CSV para guardar información respecto al corpus combinado de datos.

WAV (Waveform Audio File Format): Es considerado como un formato de archivo el cual almacena audios sin descomprimir, convirtiéndose en el formato más utilizado que permite el almacenamiento de grabaciones de audio en CD. WAV contiene muestras de audio sin procesamiento previo capturando de forma directa la onda del sonido sin necesidad de afectar la calidad.

MP4 (MPEG-4 Part 14): MP4 es un formato de archivo que permite almacenar audio, video, imágenes fijas, durante el proceso de compresión usa códecs de video de H.265 o H.265 manteniendo la calidad del video a medida que su tamaño es reducido.

CSV (Comma-separated values): CSV es un formato de archivo definido para almacenar y tabular información de manera sencilla, en este formato los datos son organizados por filas y columnas separados por una coma, las filas representan un registro y las columnas corresponden a los atributos.

Análisis de RAVDESS

El conjunto de datos de Habla y Canto Emocional de Ryerson (RAVDESS) constituye un conjunto de datos multimodal que engloba emociones expresadas emociones expresadas a través de la voz y el canto. La estructura de los archivos encontrados en RAVDESS se encuentra separada por el número de actores como se observa en la Figura 54.

Figura 54

Estructura de las carpetas y videos de RAVDESS

Nota. Se denota la separación de carpetas del actor y el formato de nombre de los archivos.

En el archivo, el identificador numérico presente en el nombre de archivo es crucial para categorizar y diferenciar las muestras de audio. Cada identificador está compuesto por varios dígitos que transmiten información relevante sobre la grabación en cuestión como se observa de forma detallada en la Tabla 6.

Tabla 6

Descripción detallada de los identificadores en los archivos

Identificador	Descripción de niveles
Modalidad	(01 = full-AV, 02 = solo video, 03 = solo audio).
Canal	Canal vocal (01 = habla, 02 = canción).
Emoción	(01 = neutral, 02 = tranquilo, 03 = feliz, 04 = triste, 05 = enojado, 06 = temeroso, 07 = asco, 08 = sorprendido).
Intensidad	(01 = normal, 02 = fuerte).
Declaración	(01 = “Los niños hablan junto a la puerta”, 02 = “Los perros están sentados junto a la puerta”).
Repetición	(01 = 1ra repetición, 02 = 2da repetición)
Actor	(01 a 24. Los actores impares son hombres, los actores pares son mujeres)

Por ejemplo, para un archivo con un nombre de archivo 01-01-01-01-01-01-01.wav, se observa el detalle en la Tabla 7.

Tabla 7

Información representativa del archivo a partir de los identificadores numéricos

Modalidad	01	Audio-Video
Canal	01	Habla
Emoción	01	Alegría
Intensidad	01	Baja
Declaración	01	“Los niños hablan junto a la puerta”
Repetición	01	Primera repetición
Actor	01	Primer Actor

A partir del uso de estos identificadores, se procedió con el análisis exploratorio de datos, este mismo es un paso inicial crítico en el proceso de desarrollo de modelos de Aprendizaje Profundo (Deep Learning). Implicando el examen y la visualización exhaustivos de un conjunto de datos para obtener información, descubrir patrones, identificar anomalías y comprender las relaciones entre las variables, antes de cargar estos datos, creamos una

función para mapear los identificadores antes mencionados, se observa la misma en la Figura *55*.

Figura 55

Ejemplo de mapeo de información

```python
def load_rav_data(RAV):
    emotion = []
    voc_channel = []
    full_path = []
    modality = []
    intensity = []
    actors = []
    phrase = []

    for root, dirs, files in tqdm(os.walk(RAV)):
        for file in files:
            try:
                modal = int(file[1:2])
                vchan = int(file[4:5])
                lab = int(file[7:8])
                ints = int(file[10:11])
                phr = int(file[13:14])
                act = int(file[19:20])
                modality.append(modal)
                voc_channel.append(vchan)
                emotion.append(lab)
                intensity.append(ints)
                phrase.append(phr)
                actors.append(act)
                full_path.append((root, file))
            except ValueError:
                continue
```

Nota. Ejemplo de mapeo de información de los identificadores numéricos para poder cargar el conjunto de datos.

Para realizar el análisis de estos datos, se utiliza la combinación de las librerías Matplotlib, seaborn y librosa en Python. "librosa" para el análisis de audio, "Matplotlib" para el trazado básico y "Seaborn" para una estética de visualización mejorada, por ejemplo, para

obtener el resultado de la distribución de emociones según el identificador número, usamos el fragmento de código Python observado en la Figura 56.

Figura 56

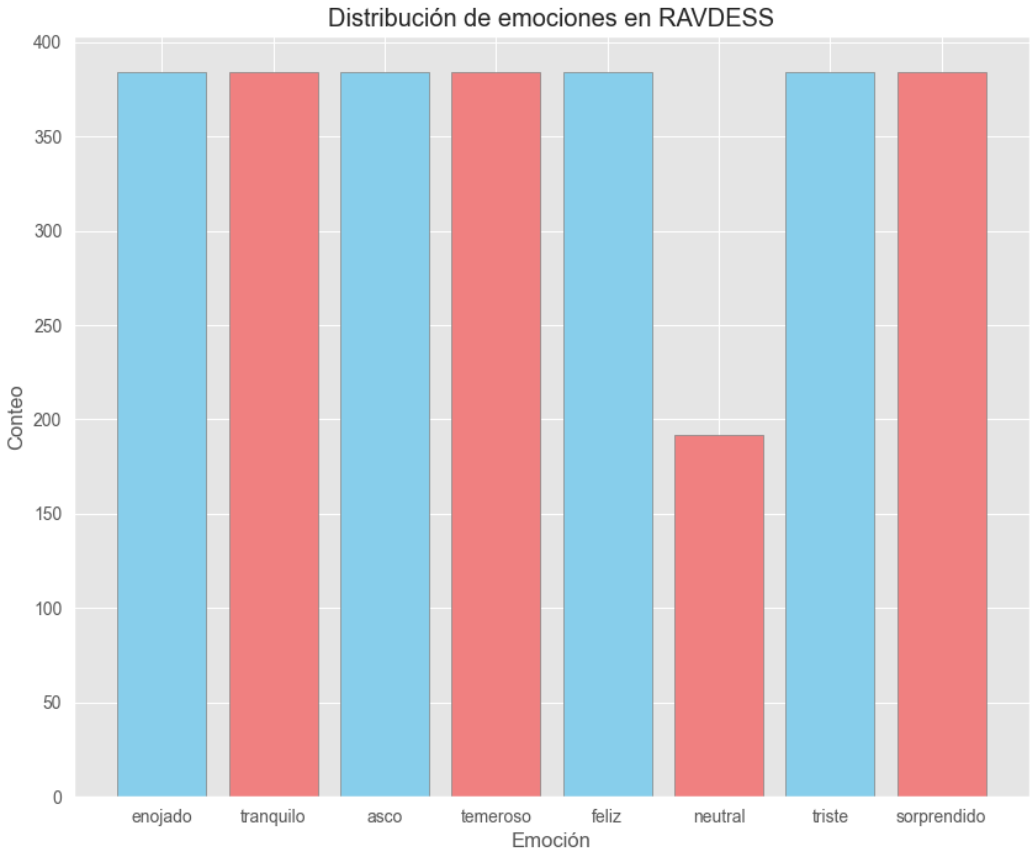
Código matplotlib para la generación de gráficas

```
emotion_labels_dict = {
    'angry': 'enojado',
    'calm': 'tranquilo',
    'disgust': 'asco',
    'fearful': 'temeroso',
    'happy': 'feliz',
    'neutral': 'neutral',
    'sad': 'triste',
    'surprised': 'sorprendido'
}

plt.figure(figsize=(10, 8))
emotion_counts = df_rav.emotion.value_counts().sort_index()
emotion_labels_list = [emotion_labels_dict[label] for label in emotion_counts.index]
bars = plt.bar(emotion_labels_list, emotion_counts, color=['skyblue', 'lightcoral'])

for bar in bars:
    bar.set_edgecolor('gray')

plt.title('Distribución de emociones en RAVDESS')
plt.xlabel('Emoción')
plt.ylabel('Conteo')
plt.savefig('bar-00.png')
plt.close()
with open("bar-00.png", "rb") as img_file:
    img_data = img_file.read()
```

El código crea un gráfico de barras que muestra la distribución de emociones en un conjunto de datos utilizando nombres de emociones en español. Se observa de la Figura 57 obtenida que existe un conteo bastante uniforme de las distintas emociones en RAVDESS, de esta manera se puede evitar el desbalanceo de clases, esto sucede cuando algunas clases en un conjunto de datos tienen mayor cantidad de muestras que otras. Esto puede ser problemático porque los algoritmos de aprendizaje tienden a sesgarse hacia las clases más numerosas

Figura 57

Distribución emocional por el set de datos RAVDESS

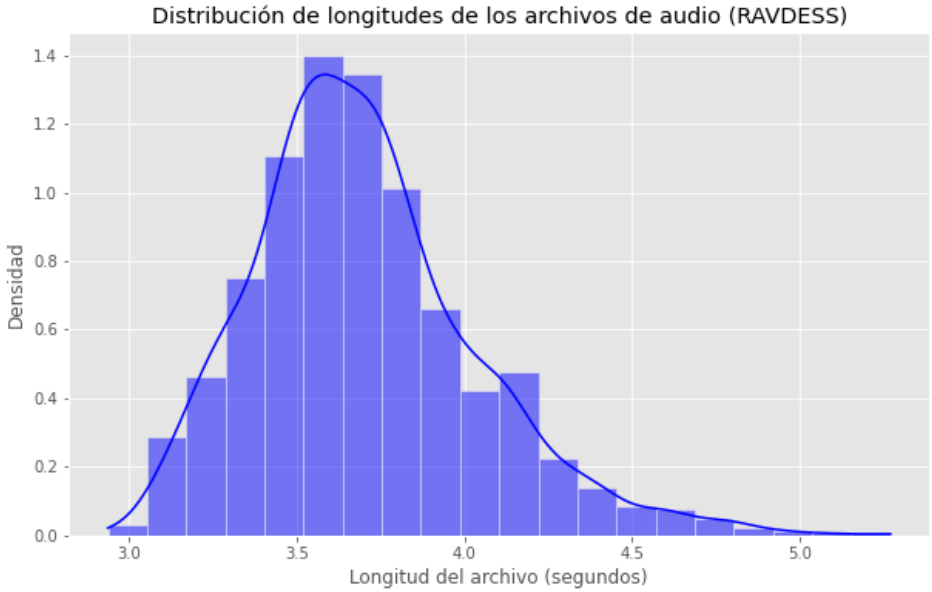


Para obtener una mejor idea de los archivos de audio, obtendremos la longitud promedio que estos representan, en la siguiente figura se muestra la carga archivos de audio, calcula sus duraciones, crea un histograma con una curva de densidad estimada[10] para visualizar las longitudes de los archivos de audio, y ajusta una distribución normal a los datos para comprender mejor su distribución, se observa la misma en la Figura 58.

[10] La "curva de densidad" generalmente se refiere a un gráfico que representa la distribución de la densidad de una variable en un conjunto de datos

Figura 58

Distribución de longitudes de los archivos de audio en RAVDESS



Observamos también las diferencias en los gráficos generados por librosa en la Figura 59, mediante el tratamiento del audio en el conjunto de datos, esto se utiliza para crear un gráfico de forma de onda básico a partir de una señal de audio. Este gráfico representa visualmente la amplitud de la señal de audio a lo largo del tiempo como se observa en la Figura 60.

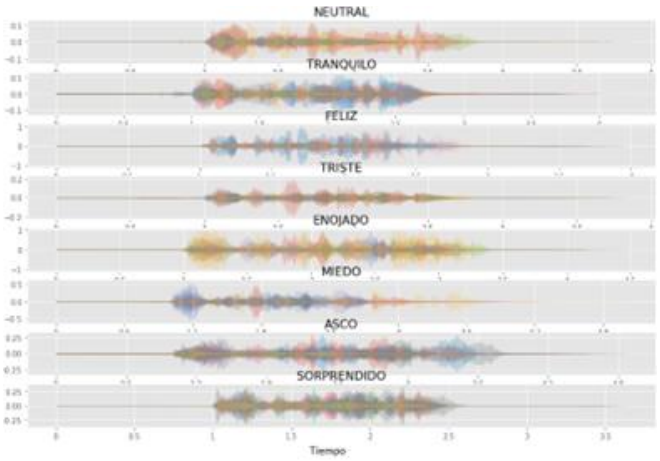
Figura 59

Uso de librosa para mostrar la forma de onda de datos de audio

```
def plot_speech(fname, ind, axis):
    data, sampling_rate = librosa.load(fname)
    plt.figure(figsize=(15, 5))
    librosa.display.waveshow(data, sr=sampling_rate, alpha=.2,
```

Figura 60

Emociones en RAVDESS representadas en formas de onda



SAVEE. En el conjunto de datos SAVEE, cada archivo de audio está etiquetado con un código de dos caracteres que indica la emoción expresada en la grabación. Los códigos corresponden a las emociones enumeradas anteriormente. Por ejemplo, un nombre de archivo como "03A01.wav", Figura 61. indica que el tercer orador expresa enojo. Los primeros dos caracteres denotan al hablante y los siguientes dos caracteres denotan la emoción, se observa el detalle de los identificadores en la Tabla 8, es importante también mencionar que todas las grabaciones son de actores masculinos.

Figura 61

Nombre del archivo y datos especificados en formato .wav

Tabla 8

Detalle de identificadores de SAVEE

Identificador	Emoción	Género
(A)	Ira	Masculino
(D)	Disgusto	Masculino
(F)	Miedo	Masculino
(H)	Felicidad	Masculino
(S)	Tristeza	Masculino
(SU)	Sorpresa	Masculino

Cargamos los datos de SAVEE aplicando los identificadores mencionados anteriormente, se muestra esto en el código Figura 62:

Figura 62

Cargando Datos de SAVEE

```
def load_savee_data(SAVEE):
    dir_list = os.listdir(SAVEE)
    emotion = []
    path = []

    for i in dir_list:
        if i[-8:-6] == '_a':
            emotion.append('angry_male')
        elif i[-8:-6] == '_d':
            emotion.append('disgust_male')
        elif i[-8:-6] == '_f':
            emotion.append('fear_male')
        elif i[-8:-6] == '_h':
            emotion.append('happy_male')
        elif i[-8:-6] == '_n':
            emotion.append('neutral_male')
        elif i[-8:-6] == 'sa':
            emotion.append('sad_male')
        elif i[-8:-6] == 'su':
            emotion.append('surprise_male')
        else:
            emotion.append('Unknown')
        path.append(SAVEE + i)

    SAVEE_df = pd.DataFrame(emotion, columns=['emotion_label'])
    SAVEE_df['source'] = 'SAVEE'
    SAVEE_df = pd.concat([SAVEE_df, pd.DataFrame(path, columns=['path'])], axis=1)

    return SAVEE_df
```

MESD. La Base de Datos del Habla Emocional Mexicana (MESD por sus siglas en inglés) es una colección de grabaciones que consisten en frases de una sola palabra que

capturan emociones como ira, disgusto, miedo, felicidad, neutral y tristeza. Los formatos obtenidos del conjunto de datos se detallan en la Figura 63.

Figura 63

Datos pertenecientes al conjunto de datos MESD

A continuación, se presenta una tabla representativa que describe los identificadores encontrados en MESD:

Tabla 9

Detalle de Identificadores de MESD

Identificador	Detalle
Emoción	Anger (ira), Disgust (disgusto), Fear (miedo), Happy (felicidad), Neutral (neutral) y Sadness (tristeza)
Actor	C (Niño) , F (Mujer) y H (Hombre)
Frase	Ejemplo: “Antes”
Corpus	A y B

Se muestra el código que permite cargar el conjunto de datos en la Figura 64.

Figura 64

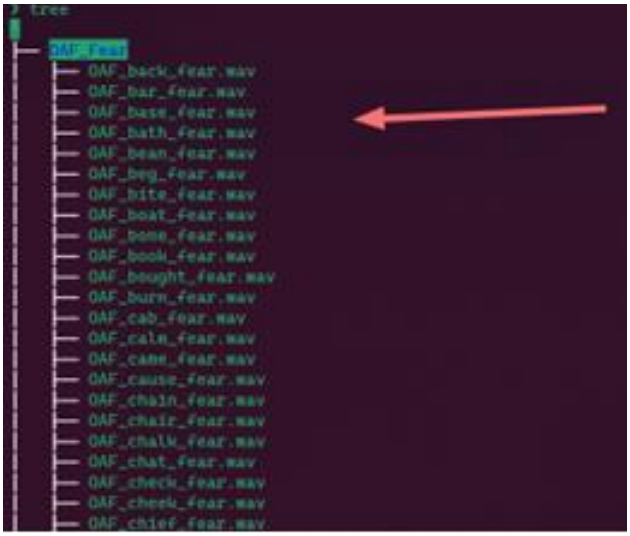
Cargando el conjunto de datos MESD

```
load_dotenv()
def load_mesd_data(MESD):
    dir_list = os.listdir(MESD)
    info_list = []
    info_tmp = []

    for s in dir_list:
        info = s.split('_')
        info[-1] = info[-1].replace('.wav', '')
        info_list.append(info)

    df_mesd = pd.DataFrame(info_list, columns=['Emotion', 'InfoActor
    infoactor_map = {
        'C': 'child',
        'F': 'female',
        'M': 'male'
    }


    df_mesd['InfoActor'] = df_mesd['InfoActor'].map(infoactor_map)
    df_mesd['Path'] = [os.getenv('DATA_PATH') + 'MESD/' + filename
    df_mesd['Phrase2'].fillna("", inplace=True)


    df_mesd['Phrase'] = df_mesd[['Phrase1', 'Phrase2']].apply(lambda
```

TESS. El Conjunto de Datos de Expresiones Emocionales de Toronto (TESS) es una base de información ampliamente empleada y de acceso público, que incluye grabaciones de actores que representan una variedad de expresiones emocionales tanto en formato de audio como visual, se observa el formato del conjunto de Datos en la Figura 65.

Figura 65

Listado de datos pertenecientes a TESS



Se muestra el código que nos permite cargar todo el conjunto de datos en la Figura 66, aquí se muestra la carga total de los datos pertenecientes a TESS.

Figura 66

Implementación de función para carga de datos de TESS

```
def load_tess_data(TESS, dir_list):
    path = []
    emotion = []

    for i in dir_list:
        fname = os.listdir(TESS + i)
        for f in fname:
            if i == 'OAF_angry' or i == 'YAF_angry':
                emotion.append('angry_female')
            elif i == 'OAF_disgust' or i == 'YAF_disgust':
                emotion.append('disgust_female')
            elif i == 'OAF_Fear' or i == 'YAF_fear':
                emotion.append('fear_female')
            elif i == 'OAF_happy' or i == 'YAF_happy':
                emotion.append('happy_female')
            elif i == 'OAF_neutral' or i == 'YAF_neutral':
                emotion.append('neutral_female')
            elif i == 'OAF_Pleasant_surprise' or i == 'YAF_pleasant_surprised':
                emotion.append('surprise_female')
            elif i == 'OAF_Sad' or i == 'YAF_sad':
                emotion.append('sad_female')
            else:
                emotion.append('Unknown')
            path.append(TESS + i + "/" + f)
```

Combinación de los distintos conjuntos de datos

Para la combinación de los conjuntos de datos, se utilizó las funciones que cargan datos de diferentes conjuntos de datos. Se llama a estas funciones con argumentos específicos relacionados con los respectivos conjuntos de datos y se asigna los marcos de datos resultantes a las variables para su posterior uso, se observa esto en la Figura 67.

Figura 67

Cargando los datos para su posterior combinación

```
from utils.classes.loaders.MesdLoader import load_mesd_data
from utils.classes.loaders.SaveeLoader import load_savee_data
from utils.classes.loaders.TessLoader import load_tess_data
from utils.classes.loaders.RavdessLoader import load_rav_data

df_rav = load_rav_data(RAV)
df_tess = load_tess_data(TESS,dir_list)
df_savee = load_savee_data(SAVEE)
df_mesd = load_mesd_data(MESD)
```

Mapeamos de igual forma en un diccionario las emociones negativas y positivas para el posterior entrenamiento en el modelo, se observa el proceso para realizar conjunto de datos combinado mediante el código en la Figura 68.

Figura 68

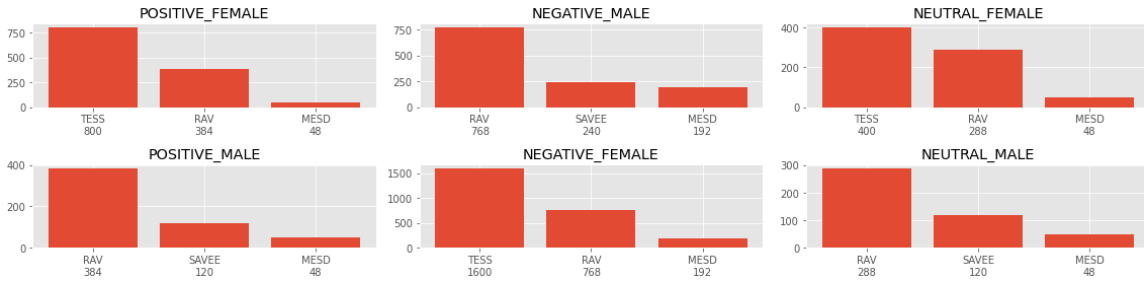
Conjunto de datos combinado

```
new_labels_dict_comb = {'angry_male':'negative_male', 'angry_female':'negative_female',
                        'calm_male':'neutral_male', 'calm_female':'neutral_female',
                        'disgust_male':'negative_male', 'disgust_female':'negative_female',
                        'fearful_male':'negative_male','fearful_female':'negative_female',
                        'fear_male':'negative_male', 'fear_female':'negative_female',
                        'happy_male':'positive_male', 'happy_female':'positive_female',
                        'neutral_male':'neutral_male', 'neutral_female':'neutral_female',
                        'sad_male':'negative_male', 'sad_female':'negative_female',
                        'surprised_male':'positive_male', 'surprised_female':'positive_female',
                        'surprise_male':'positive_male', 'surprise_female':'positive_female',
                        'Unknown': 'unk'}

df_combined['emotion2'] = df_combined['emotion_label'].map(new_labels_dict_comb)
df_combined.head()
```

Observamos la distribución de emociones en la gráfica de barras, donde “RAV” representa el conjunto de datos RAVDESS, se observa que la distribución de datos según su emoción Negativa y Positiva, y el género de la persona en la Figura 69.

Figura 69

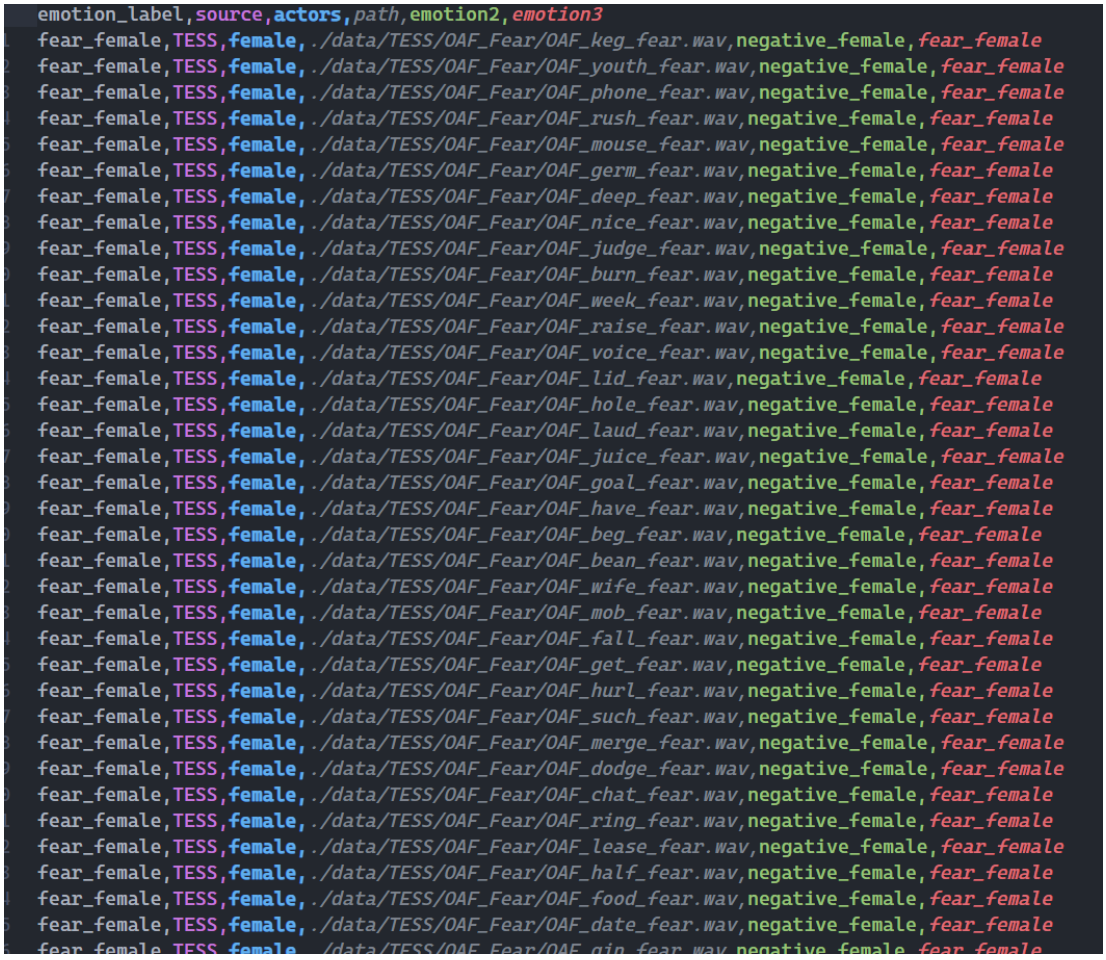
Distribución de datos en el conjunto combinado



Dicho conjunto los guardamos en el formato CSV, se puede observar el conjunto completo en la siguiente Figura 70. El archivo CSV organiza información sobre grabaciones de discursos emocionales con un total de 6737 registros, incluido el tipo de emoción, el género del orador, la fuente y las rutas de archivo a las grabaciones de audio, con múltiples columnas que brindan diferentes aspectos del contexto y el etiquetado emocional.

Figura 70

Archivo combined.csv que representa todos los datos del corpus.



```
emotion_label,source,actors,path,emotion2,emotion3
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_keg_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_youth_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_phone_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_rush_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_mouse_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_germ_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_deep_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_nice_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_judge_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_burn_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_week_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_raise_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_voice_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_lid_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_hole_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_laud_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_juice_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_goal_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_have_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_beg_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_bean_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_wife_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_mob_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_fall_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_get_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_hurl_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_such_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_merge_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_dodge_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_chat_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_ring_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_lease_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_half_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_food_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_date_fear.wav,negative_female,fear_female
fear_female,TESS,female,./data/TESS/OAF_Fear/OAF_gin_fear.wav,negative_female,fear_female
```

Extracción de Características

La extracción de características es un paso crucial en el Deep Learning, en paso donde los datos sin procesar se transforman en un conjunto de características relevantes e informativas, es esencial identificar, debido a que estamos hablando puramente de Deep Learning, no se aplica el paso de reducción de dimensionalidad.

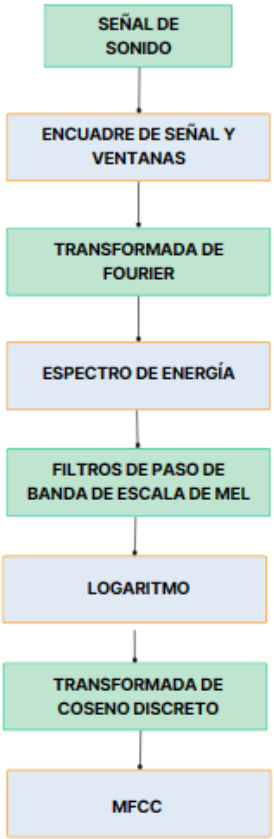
Para empezar, recordando las características de señales usadas principalmente en el reconocimiento de emociones por voz, en primera instancia el componente del habla, mismo que resulta de la combinación entre la respuesta de frecuencia del tracto vocal y el pulso glotal.

El interés principal radica en comprender el tracto vocal, que contiene el conocimiento esencial que ayuda al modelo aprender de estos datos.

Así que, el objetivo es desvincular el tracto vocal de la onda de habla, excluyendo el pulso glotal. Esto se consigue al modificar la amplitud en una forma logarítmica (dado que la percepción del sonido sigue una escala logarítmica), emplear el escalado Mel y luego aplicar una transformación adicional (la Transformada de Coseno Discreto). El resultado de este proceso son los coeficientes de onda (MFCCS), y el procedimiento para obtener estas características se detalla en la Figura 71.

Figura 71

Extracción de características MFCC de una señal de sonido.



Se ha implementado una función en Python que nos permite extraer las características MFCC de una señal de sonido, como se observa en la Figura 72, esta función permite además generar un gráfico a partir de dichas características.

Figura 72

Función que realiza la extracción de MFCC y realiza el gráfico

Algorithm 1: Función `sampleAudioData`

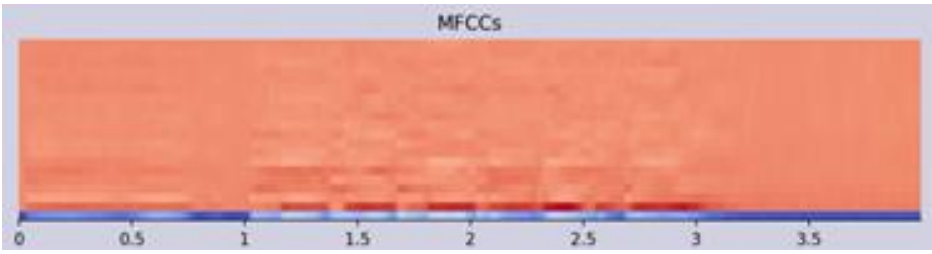
Data: La instancia de la clase `AudioProcessing`
Entrada: `findDB` (bool, Encontrar Dataset) `findmfcc` Encontrarcaracteristicas MFCC?
`plots` (bool, Cargar Plots?
Salida : Datos de audio (`data`), espectrograma escalado en decibelios (`Xdb`) y MFCCs (`mfccs`)

```
sampleAudioData()
Xdb ← None;
if self.findDB then
    X ← lib.stft(data);
    Xdb ← lib.amplitude_to_db(abs(X));
end
mfccs ← None;
if self.findmfcc then
    mfccs ← lib.feature.mfcc(data, sr=sampleRate, n_mfcc=40);
end
if self.findDB and self.plots then
    fig, ax ← plt.subplots(1, 2, figsize=(16, 3));
    plt.subplot(121);
    lib.display.waveplot(data, sr=sampleRate);
    plt.subplot(122);
    lib.display.specshow(Xdb, sr=sampleRate, x_axis='time',
     y_axis='hz');
    plt.show();
end
else if self.plots then
    lib.display.waveplot(data, sr=sampleRate);
    return (data, Xdb, mfccs);
end
```

Se observa el grafico característico de los MFCC en la Figura 73. El gráfico en sí consta de una serie de líneas o curvas, cada una de las cuales corresponde a un coeficiente MFCC diferente. La forma de estas curvas proporciona información sobre cómo cambia el contenido espectral de la señal de audio con el tiempo.

Figura 73

Características MFCC obtenidas a partir de un audio de 4 segundos.



Para el procesamiento del conjunto de datos combinado anterior, se ejecuta el código mostrado en la Figura 74, después de que se ejecute este código, la lista contendrá tuplas donde el primer elemento de cada tupla son los datos de audio cargados y el segundo elemento es la frecuencia de muestreo correspondiente.

Figura 74

Carga y procesamiento de audios combinados.

```
from tqdm import tqdm
import resampy

X = []
for i in tqdm(mydf['path']):
    X.append(librosa.load(i, res_type='kaiser_fast', sr=44000))
```

2.7s

```
3%|▌         | 181/6736 [00:02<01:24, 77.70it/s]
```

División en conjuntos de datos de entrenamiento, validación y pruebas (Hold-out validation)

El uso de divisiones de entrenamiento, validación y prueba es una práctica fundamental en el desarrollo de modelos. Estas divisiones tienen distintos propósitos y ayudan a garantizar la confiabilidad y la capacidad de generalización de un modelo entrenado.

De acuerdo a lo especificado por (Saunshi et al., 2021) una proporción muy utilizada es una división 80:20, donde el 80% es asignado para entrenamiento/validación y el 20% para pruebas. Se usará este mismo tipo de separación para el entrenamiento del modelo, esto se lo hace mediante la función de sklearn train_test_split, Figura 75, el parámetro random_state se establece en 12 para garantizar la reproducibilidad en el proceso.

Figura 75

Uso sklearn para realizar divisiones al conjunto de datos

```
from sklearn.model_selection import train_test_split

X_train, X_val, y_train, y_val = train_test_split(mfccs, y, test_size=0.2, random_state=12)
X_val, X_test, y_val, y_test = train_test_split(X_val, y_val, test_size=0.5, random_state=15)

X_train.shape, X_val.shape, X_test.shape, y_train.shape, y_val.shape, y_test.shape
```

Se puede verificar que el conjunto de datos este separado de manera correcta mediante shape() como se observa en la Figura 76, el conjunto de datos X_train (que corresponde al entrenamiento) contiene 5386 registros y recordando que el total de datos que contiene el conjunto de datos combinado original es de 6737 se obtiene que aproximadamente un 0.79 porciento de datos para entrenamiento, se observa también la distribución 50-50 entre validación y pruebas.

Figura 76

Uso de shape() para verificar la división correcta.

```
4 X_train.shape, X_val.shape, X_test.shape, y_train.shape, y_val.shape, y_test.shape
✓ 0.0s
((5386, 236, 40), (673, 236, 40), (674, 236, 40), (5386,), (673,), (674,))
```

Selección del Modelo Deep Learning

Se reconoce que las redes neuronales recurrentes (RNN) tienen un desempeño favorable en la realización de tareas de identificación de emoción por voz. No obstante, existe un cuerpo robusto de investigación que indica que, en numerosos escenarios, las redes neuronales convolucionales (CNN) pueden superar a las RNN. En esta instancia, se emplear por usar redes CNN, empleando capas de convolución unidimensionales y capas de pooling unidimensionales. Esta elección se basa en que nuestros datos de entrenamiento están conformados por tres dimensiones.

Se observa la construcción del modelo mediante Keras/Tensorflow en la Figura 77.

Figura 77

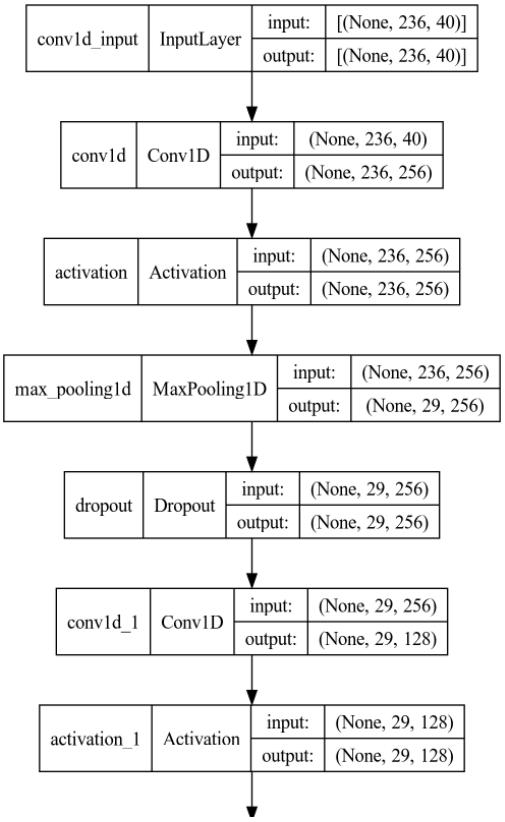
Construcción de bloques del modelo con Keras/Tensorflow.

```
model3 = Sequential()

model3.add(layers.Conv1D(256, 5,padding='same',
                         input_shape=(236,40)))
model3.add(layers.Activation('relu'))
model3.add(layers.MaxPooling1D(pool_size=(8)))
model3.add(layers.Dropout(0.2))

model3.add(layers.Conv1D(128, 5,padding='same'))
model3.add(layers.Activation('relu'))
model3.add(layers.MaxPooling1D(pool_size=(4)))
model3.add(layers.Dropout(0.1))

model3.add(layers.Flatten())
model3.add(layers.Dense(64))
model3.add(layers.Dense(7))
model3.add(layers.Activation('softmax'))

model3.summary()
```

Adicional, la Figura 78 muestra la parte inicial de la arquitectura del modelo de forma gráfica, la red CNN se compuso de dos bloques, cada uno conformado por una capa de convolución unidimensional, una función de activación ('ReLu'), una capa de agrupamiento unidimensional y una capa de eliminación ('dropout'). A estos dos bloques les siguieron dos capas densas totalmente conectadas y una función de activación 'SoftMax', ya que estamos tratando con un problema de clasificación multiclase (7 emociones distintas).

Figura 78

Entrada de Arquitectura representativa al modelo para reconocimiento de emociones SER



Entrenamiento del Modelo Deep Learning

En la búsqueda del mejor modelo posible, se recurre al proceso conocido como ajuste de Hiperparámetros[11]. El fragmento de código en la Figura 79 detalla la configuración de tres tipos de "callbacks" en TensorFlow/Keras, que son funciones que se activan durante el entrenamiento de una red neuronal con el propósito de realizar tareas específicas. Por ejemplo, el ajuste de la tasa de aprendizaje se logra a través de ReduceLROnPlateau, que persigue determinar la tasa óptima para mejorar el entrenamiento. Asimismo, mediante EarlyStopping, se interrumpe el entrenamiento cuando se identifica el modelo que logra la mayor precisión en el conjunto de datos de validación.

[11] Hiperparámetros se refieren a parámetros que podemos ajustar en entrenamiento del modelo de machine Learning, se busca mejorar la precisión de este cambiándolos.

Figura 79

Búsqueda de la mejor taza de aprendizaje para el modelo CNN

```
weight_path2 = './best_weights2.hdf5'

reduce_lr = tf.keras.callbacks.ReduceLROnPlateau(monitor='val_accuracy',
                                                 factor=0.5, patience=4,
                                                 verbose=1, mode='max',
                                                 min_lr=0.00001)

early_stop = tf.keras.callbacks.EarlyStopping(monitor='val_loss', patience=45,
                                              verbose=1)

model_checkpoint2 = tf.keras.callbacks.ModelCheckpoint(filepath=weight_path2,
                                                       save_weights_only=True,
                                                       monitor='val_accuracy',
                                                       mode='max',
                                                       save_best_only=True)
```

Para el entrenamiento del modelo, es utilizó la función de perdida de “sparse_categorical_crossentropy” esto en Keras se refiere a la pérdida de entropía cruzada entre las etiquetas y las predicciones, se usa el optimizador “Adam”, el cual tiene como objetivo proporcionar una convergencia más rápida y un mejor rendimiento al adaptar la tasa de aprendizaje para cada parámetro, se observa estos parámetros en la Figura 80.

Figura 80

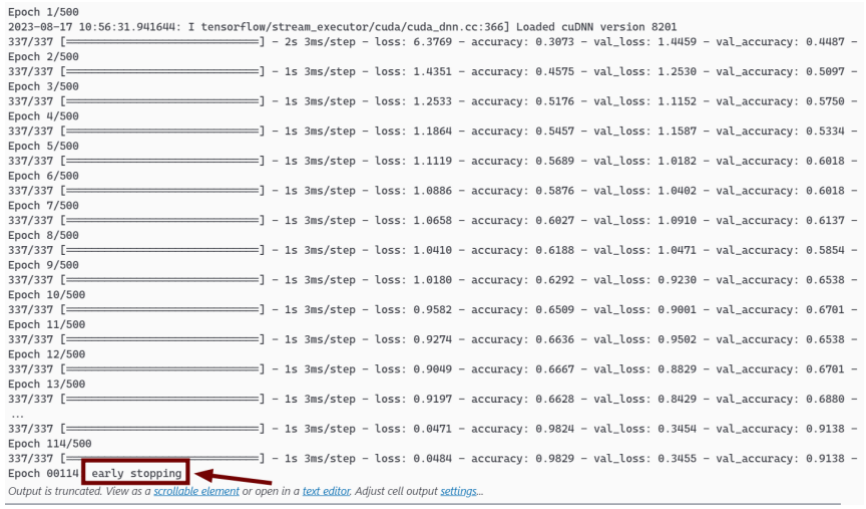
Parámetros para entrenamiento del modelo

```
model3.compile(loss='sparse_categorical_crossentropy',
               optimizer='adam',
               metrics='accuracy')

history = model3.fit(X_train, y_train,
                     batch_size=16,
                     epochs=500,
                     validation_data=(X_val, y_val),

                callbacks=[reduce_lr, early_stop,
                           model_checkpoint2])
```

Ahora se realiza el entrenamiento del modelo mediante la función fit(), que busca reducir el nivel de perdida en el conjunto de datos de validación y a su vez mejorar el “accuracy” o precisión del modelo , con los callbacks definidos anteriormente se observa que el entrenamiento se detiene en la época 114 (Figura 81), alcanzando su pico de precisión.

Figura 81

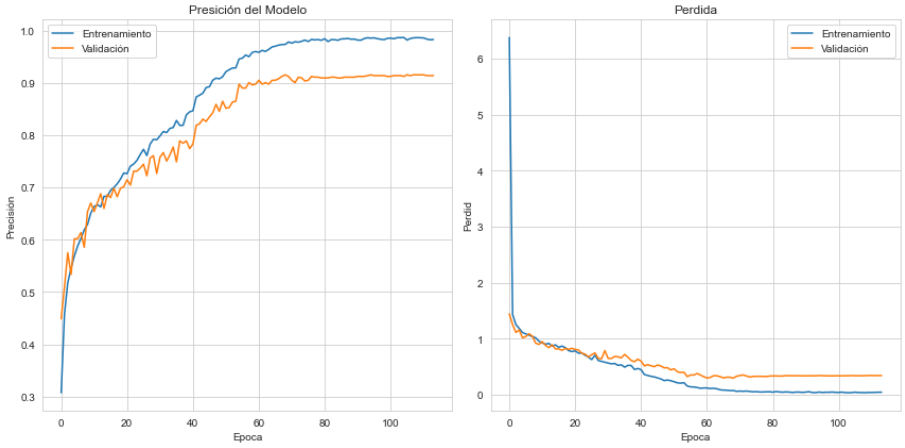
Entrenamiento del modelo en 500 épocas



Todas las iteraciones que fueron corridas durante el entrenamiento se observan de forma gráfica en la Figura 82. Se observa que existe la curva de precisión crece a menudo que las épocas progresan, indicando que el modelo si está aprendiendo realmente de los datos, igualmente la función de perdida decrece, indicando un buen rendimiento del modelo.

Figura 82

Graficas de perdida y precisión del modelo para 114 épocas



Pruebas de validación del Modelo

Para las pruebas de validación del modelo entrenado, se generó una matriz de confusión como métrica de evaluación, la función de la Figura 83 tiene la finalidad de generar

un informe con métricas de evaluación y representar gráficamente una matriz de confusión a partir de las predicciones y las etiquetas verdaderas de un conjunto de datos de prueba.

Figura 83

Generación de matriz de confusión como reporte para la validación del modelo

```
def report_res_and_plot_matrix(y_test, y_pred, plot_classes, save_path=None)
    # Reportar métricas
    acc = accuracy_score(y_test, y_pred)
    print(f"Precisión: {acc:.4f}")
    # Configurar el estilo ggplot
    plt.style.use('ggplot')
    # Graficar la matriz
    cnf_matrix = confusion_matrix(y_test, y_pred)
    plt.figure(figsize=(20, 20))  # Establecer el tamaño del gráfico
    fig, ax = plt.subplots()
    tick_marks = np.arange(len(plot_classes))
    plt.xticks(ticks=tick_marks, labels=plot_classes, rotation=90)
    plt.yticks(ticks=tick_marks, labels=plot_classes, rotation=0)  # Sin rotac
    group_counts = [f'{value:0.0f}' for value in cnf_matrix.flatten()]
    group_percentages = [f'{100 * value:0.1f} %' for value in
                         cnf_matrix.flatten()/np.sum(cnf_matrix)]
    labels = [f'{v1}\n({v2})' for v1, v2 in
              zip(group_counts, group_percentages)]
    n = int(np.sqrt(len(labels)))
    labels = np.asarray(labels).reshape(n, n)
    sns.heatmap(cnf_matrix, annot=labels, fmt='', cmap='YlGnBu',  # Cambiar la
                xticklabels=plot_classes, yticklabels=plot_classes)
    ax.xaxis.set_label_position("bottom")
    plt.tight_layout()
    plt.title('Matriz de confusión', y=1.1)
    plt.ylabel('Etiqueta actual')
    plt.xlabel('Etiqueta predicha')
    if save_path:
        plt.savefig(save_path)  # Guardar la figura en el archivo especificad

    plt.show()
```

El código genera la siguiente Tabla 10.

Tabla 10

Detalle de rendimiento del modelo con métricas de evaluación

Emoción	Precisión	Sensibilidad	Puntuación F1	Soporte
Miedo	0.96	0.83	0.89	113
Asco	0.91	0.96	0.93	90
Neutral	0.95	0.98	0.97	125
Feliz	0.96	0.87	0.91	85
Tristeza	0.89	0.97	0.93	91
Sorpresa	0.92	0.98	0.95	88
Enfado	0.94	0.96	0.95	82
Exactitud	**0.93**			**674**
Promedio ponderado	**0.94**	**0.93**	**0.93**	**674**

A partir de dicha Tabla 10, se obtienen los siguientes resultados:

Precisión (Precisión). Los valores de precisión son bastante altos, oscilando entre 0.89 y 0.96. Esto indica que el modelo tiene una alta proporción de predicciones correctas en relación con las predicciones positivas que hizo para cada clase. Es un indicativo positivo de la capacidad del modelo para evitar falsos positivos.

Sensibilidad (Recall). Los valores de sensibilidad también son altos, con rangos entre 0.83 y 0.98. Esto significa que el modelo está capturando eficazmente la mayoría de los ejemplos de cada clase en el conjunto de datos. Es una señal positiva de su capacidad para evitar falsos negativos.

Puntuación F1 (F1-Score). La puntuación F1 combina precisión y sensibilidad y es útil cuando se busca un equilibrio. Los valores son altos, cerca de 0.90 o más para cada clase. Esto indica un buen equilibrio entre precisión y sensibilidad.

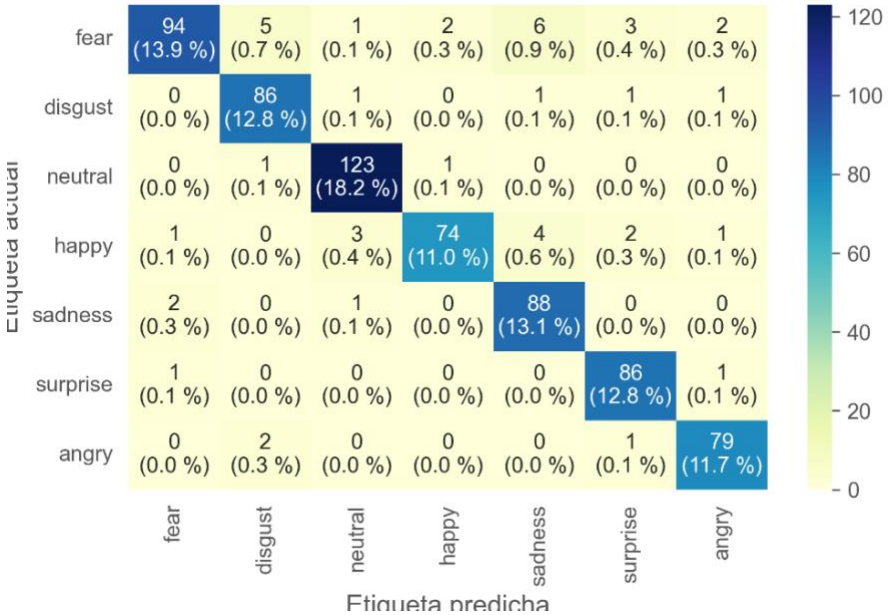
Exactitud (Accuracy). La exactitud general del modelo es del 0.93, lo que significa que aproximadamente el 93% de las predicciones son correctas en todo el conjunto de datos. Esto es un indicio de que el modelo está funcionando bien en general.

Promedio ponderado (Weighted Average). Los valores de promedio ponderado están cerca de 0.93 en términos de precisión, sensibilidad y puntuación F1. Esto sugiere un buen rendimiento general del modelo, especialmente considerando el desequilibrio de clases en el conjunto de datos.

Se muestra la matriz de confusión para todas las clases en la Figura 84, se muestra que la gran mayoría de nuestras clases de emociones fueron clasificadas correctamente, la mayoría de los datos se encuentran en la diagonal que representan los Verdaderos Positivos y Verdaderos Negativos , los demás números representan los Falsos positivos y Falsos negativos, que son muy pocos, se observa también que el modelo logra predecir de una manera más certera emociones neutrales debido a su porcentaje elevado.

Figura 84

Matriz de Confusión para las 7 clases (emociones) a predecir



Desarrollo de Procesamiento Natural de Lenguaje con Reconocimiento Automático del Habla (ASR)

Implementando Whisper. Para la implementación de Whisper, se buscó la implementación de un modelo pre-entrenado que tenga la mayor precisión para la menor cantidad de tiempo de ejecución posible, para esto, se probó distintas implementaciones de Whisper como se observa en la Tabla 11.

Tabla 11

Comparando el tiempo de ejecución de las distintas implementaciones de Whisper

Implementación	Precisión	Tiempo	Uso de memoria máxima del GPU	Uso de memoria máxima del CPU
openai/whisper	fp16	90s	11325MB	9439MB
faster-whisper	fp16	54s	4755MB	3244MB
faster-whisper	int8	59s	3091MB	3117MB

Nota. Tomado de (Ghillaume, 2023)

Faster whisper, es una reimplementación del modelo Whisper desarrollado por OpenAI, pero esta versión utiliza CTranslate2[12] como motor de inferencia. Este enfoque busca acelerar el procesamiento de modelos, permitiendo una inferencia más rápida y efectiva al permitir seleccionar el tipo de modelo a utilizar. Existen distintos modelos como los mencionados anteriormente según su cantidad de parámetros. Para realizar las pruebas de rendimiento, se utilizó el comando de linux time[13] y nvidia-smi [14] como se detalla en la Figura 85, para poder medir la memoria usada en la GPU, con un audio de una duración de 23.42 segundos, se observa el detalle completo de la prueba en la Tabla 12.

Tabla 12

Detalle de las diferencias entre los modelos Whisper.

Modelo	Lenguaje	Tiempo (en segundos)	Frase	Memoria de GPU utilizada
tiny	Español	0.156766 segundos	Bueno, y yo me siento bien, creo que es una herramienta superfundamentada de V, que nos ayudó de un montón para de alguna forma facilitar la comunicación	16 mb
base	Español	0.165180 segundos	Bueno, yo me siento bien, creo que es una herramienta super fundamental y que nos ayuda un montón.	16 mb
medium	Español	0.581997 segundos	Bueno, yo me siento bien, creo que es una herramienta súper fundamental y que nos ayuda un montón	2912 mb
large-v1	Español	0.556571 segundos	Yo me siento bien, creo que es una herramienta super fundamental y que nos ayuda un montón	416 mb
large-v2	Español	0.600834 segundos	Bueno, yo me siento bien. Creo que es una herramienta súper fundamental y que nos ayudó un montón	1536 mb

[12] CTranslate2 es conocido por su alta velocidad y eficiencia en la ejecución de modelos basados en Transformer, este mismo implementa un tiempo de ejecución personalizado que aplica muchas técnicas de optimización del rendimiento.
[13] El comando time se utiliza para determinar cuánto tiempo tarda en ejecutarse un comando determinado.
[14] El comando nvidia-smi informa la utilización de la memoria GPU como un porcentaje

Figura 85

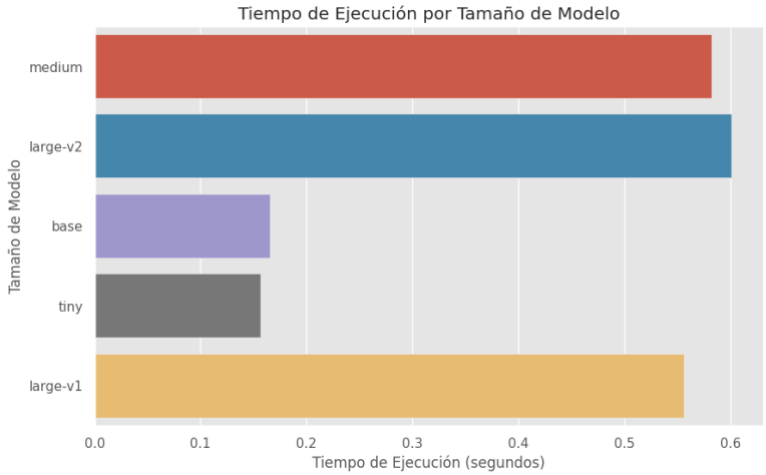
Uso de time y nvidia-smi para las pruebas de rendimiento para Faster Whisper

```
# Itera a través de cada tamaño de modelo
for model_size in model_sizes:
    # Crea la instancia del modelo Whisper
    model = WhisperModel(model_size, device="cuda", compute_type="float16")
    # Verifica la utilización de memoria de la GPU antes de ejecutar el modelo
    pre_memory_info = subprocess.check_output(["nvidia-smi", "--query-gpu=memory.used",
                                               "--format=csv,noheader,nounits"],
                                              universal_newlines=True)
    pre_memory_usage = [int(x) for x in pre_memory_info.strip().split("\n")]
    # Comienza a medir el tiempo de ejecución
    start_time = time.time()
    # Ejecuta el modelo
    segments, info = model.transcribe("./faster-whisper-webui/audio_video_test.wav",
                                      beam_size=5)
    # Finaliza la medición del tiempo de ejecución
    end_time = time.time()
    # Verifica la utilización de memoria de la GPU después de ejecutar el modelo
    post_memory_info = subprocess.check_output(["nvidia-smi", "--query-gpu=memory.used",
                                                "--format=csv,noheader,nounits"],
                                               niversal_newlines=True)
    post_memory_usage = [int(x) for x in post_memory_info.strip().split("\n")]
    # Calcula el uso de memoria (diferencia entre el uso posterior y anterior de memoria)
    memory_usage = [post - pre for pre, post in zip(pre_memory_usage,
                                                    post_memory_usage)]
    # Libera la caché de memoria de la GPU
    torch.cuda.empty_cache()
    # Itera a través de cada segmento y almacena los resultados en el DataFrame
    for segment in segments:
```

Se observan las diferencias de rendimiento en la gráfica de barras obtenida en la Figura 86, se muestra el tiempo de ejecución en segundos comparado con el tamaño del modelo, no existen muchas diferencias en los tiempos de ejecución entre “medium”, “large-v2” y “large-v1”, mientras que los modelos “base” y “tiny” tienen un tiempo de ejecución mucho menor.

Figura 86

Diferencias en Tiempos de ejecución en segundos para distintos modelos de Whisper



Debido a que la diferencia entre tiempos de ejecución no es tan significativa, y la diferencia entre la calidad de resultados entre “base” y “medium” es bastante elevada, se opta

por usar el modelo large-v2 para obtener los mejores resultados. Para obtener resultados de Faster Whisper, se realizó la implementación de una interfaz de programación de aplicaciones (Application Programming Interface; API) mediante Gradio, con el método observado en la Figura 87 la función gr.Interface permite crear una GUI basada en la web en torno a un modelo de aprendizaje automático, en este caso la transcripción creada por Whisper.

Figura 87

Método que permite la implementación de una REST API para Whisper

```
simple_transcribe = gr.Interface(fn=ui.transcribe_webui_simple_progress
        if is_queue_mode else ui.transcribe_webui_simple,
            description=ui_description, article=ui_article, inputs=simple_inputs(),
                outputs=[
        gr.File(label="Descargar"),
        gr.Text(label="Transcripción"),
        gr.Text(label="Segmentos")
])
```

Esta función nos permite asignar una serie de “inputs” o entradas que permiten especificar qué atributos vamos a asignar al momento de realizar la transcripción con el modelo de whisper, para llamar a la API existen distintos parámetros entre los cuales se detalla en la siguiente Tabla 13:

Tabla 13

Parámetros enviados desde el backend a la API

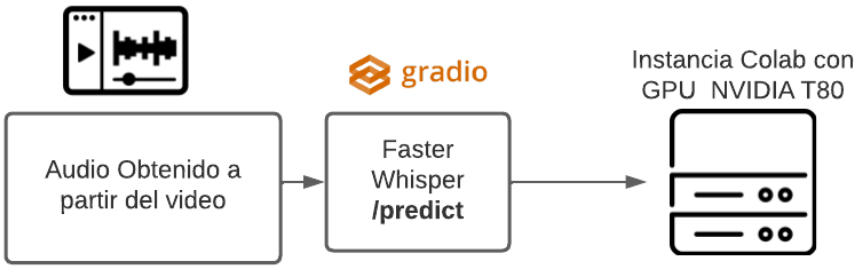
Parámetro	Descripción	Parámetro enviado	Tipo de Dato
Model	Representa la elección de modelo a partir de los explicados anteriormente	large-v2	String
Lenguage	El lenguaje del texto de transcripción.	Spanish	String

Task	Identificador de la tarea a realizar en la secuencia de audio, en este caso “transcribe” - transcribir	transcribe	String
Word Timestamps	Representa el estado verificado del parámetro 'Word Timestamps’ o Marcas de tiempo en las palabras, se envía como Falso	False	Boolean
Word Timestamps - Highlight Words	Representa el estado verificado del parámetro que nos indica si resaltamos las palabras, se envía como Falso.	False	Boolean
Audio File	Lista de ruta(s) de archivo o URL(s) a archivos en el componente de audio, se envía dentro el nombre del archivo de audio a la API	[f"./audio/{audio_filename}"]	{ name: string }

Se describe el flujo implementado en la aplicación en la gráfica, el audio de la persona se obtiene mediante el backend en streamlit, dicho audio se envía al modelo alojado dentro de la instancia GPU con una NVIDIA T80 proporcionada por Google colab, de esta forma se obtiene el audio a partir del video.

Figura 88

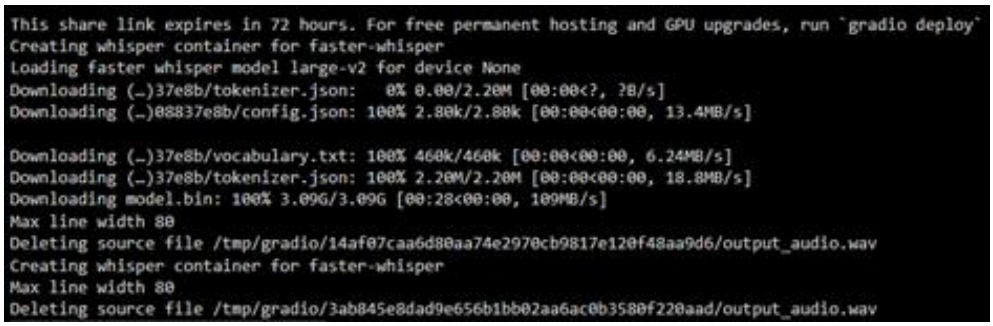
Flujo y llamada de API Gradio a Streamlit



El aplicativo toma un archivo de audio, transcribe su contenido, segmenta el texto transcrito en función de las marcas de tiempo y presenta los resultados a través de una interfaz de usuario basada en Streamlit y React, el backend se encarga de obtener el audio .wav a partir del video y posteriormente enviarlo al API de Faster Whisper, el procesamiento de la transcripción y el flujo se observa en la Figura 89.

Figura 89

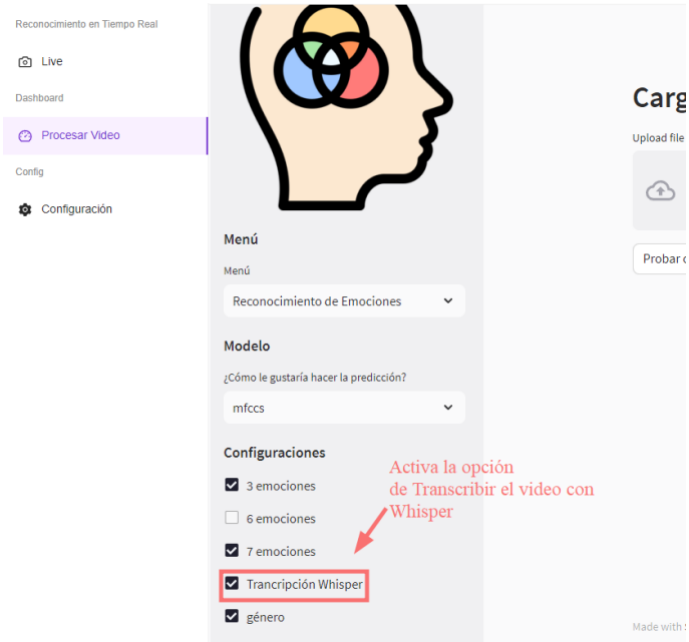
El API recibe el archivo de audio enviado desde Streamlit como "output_audio.wav".



Dentro del aplicativo, se puede especificar como opción dentro de las configuraciones si se desea obtener la transcripción, el menú desplegable permite activar o desactivar la transcripción del video mediante Whisper, como se observa en la Figura 90.

Figura 90

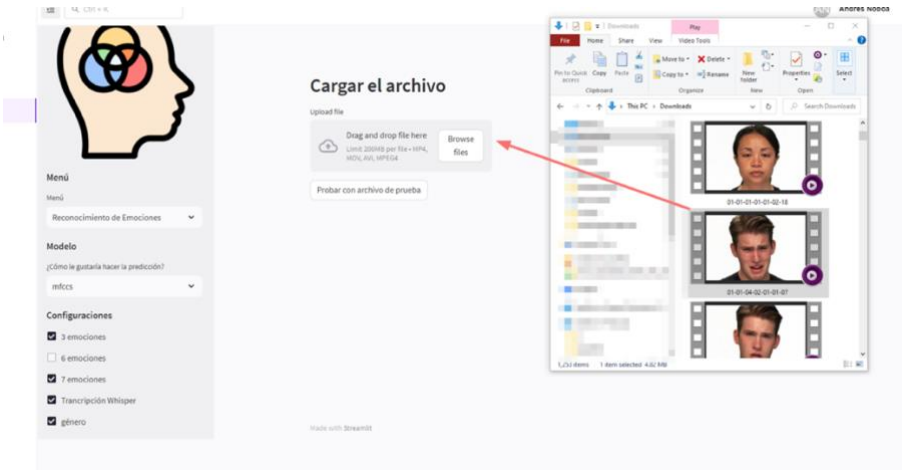
Configuraciones del sistema para obtener transcripción de texto



Para demostrar el funcionamiento del sistema ASR propuesto en el aplicativo, se procede a cargar un video al set de datos RAVDESS en el cual el actor pronuncia la frase "Los niños están hablando por la puerta", Una vez que haya elegido el archivo de video, la aplicación comenzará a procesar y cargar automáticamente el video, Figura 91.

Figura 91

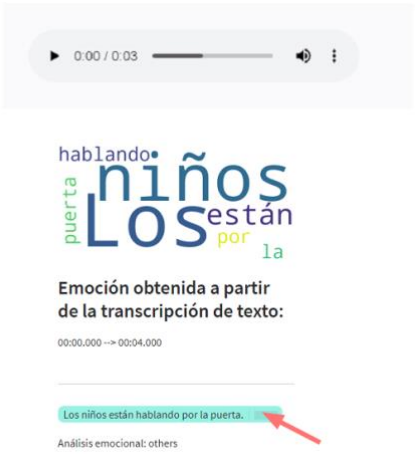
Proceso de cargar el archivo de video en el sistema



El video es procesado en el aplicativo, y se observa la transcripción obtenida del mismo en la Figura 92, la transcripción de un video se lo realiza con marcas de tiempo y se anota los puntos de tiempo específicos en los que ocurre cada segmento del texto, adicional se muestra una nube de palabras del texto obtenido en toda la longitud del video.

Figura 92

Transcripción obtenida a partir del video mediante Whisper



Modelo de Análisis Sentimental. El modelo de análisis sentimental corresponde al utilizado en la librería Py sentimiento, esta librería se encarga de realizar predicciones

emocionales al texto mediante el uso de modelos BERT, para cargar el detalle de los puntajes F1 para cada modelo BERT en Pysentimiento en la siguiente Tabla 14.

Tabla 14

Puntajes F1 para los modelos encontrados en Pysentimiento

Modelo	Emoción
BERTin	50.2 +- 2.9
BETO	52.2 +- 1.4
Electricidad	46.3 +- 2.3
RoBERTa-es	53.1 +- 2.2
RoBERTuito	55.3 +- 0.8

Nota. Tomado de (Pérez et al., 2021)

Se implementa la librería de Pysentimiento mediante la función analyze_emotion (Figura 93). El fragmento de código define una función que toma un texto como entrada (en este caso el texto obtenido mediante la transcripción por Whisper) y utiliza un analizador definido por pysentimiento, este analizador se cargó con la tarea de predicción de emociones en lenguaje español (Figura 94).

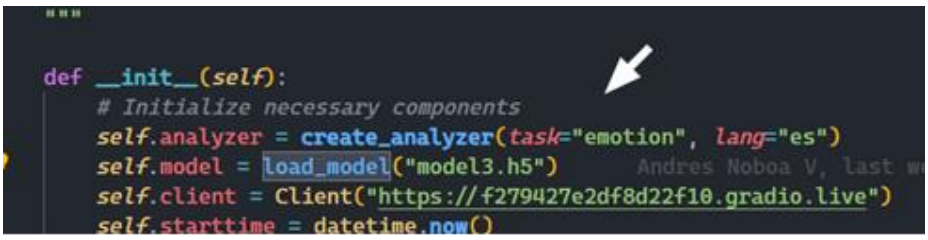
Figura 93

Implementación de pysentimiento para realizar predicciones de texto

```
def analyze_emotion(self, text):
    emotion_result = self.analyzer.predict(text)
    emotion = emotion_result.output
    probabilities = emotion_result.probas
    emotion_mapping = {
        'others': 'neutral',
        'joy': 'feliz',
        'sadness': 'triste',
        'fear': 'miedo',
        'anger': 'enojado',
        'surprise': 'sorpresa',
        'disgust': 'asco'
    }

    spanish_emotion_label = emotion_mapping.get(emotion, 'unknown')
    return spanish_emotion_label, probabilities
```

Figura 94

Instanciación del analizador de pysentimiento con el objetivo de obtener emociones del texto



Funciones del Aplicativo de predicción de emociones multimodal

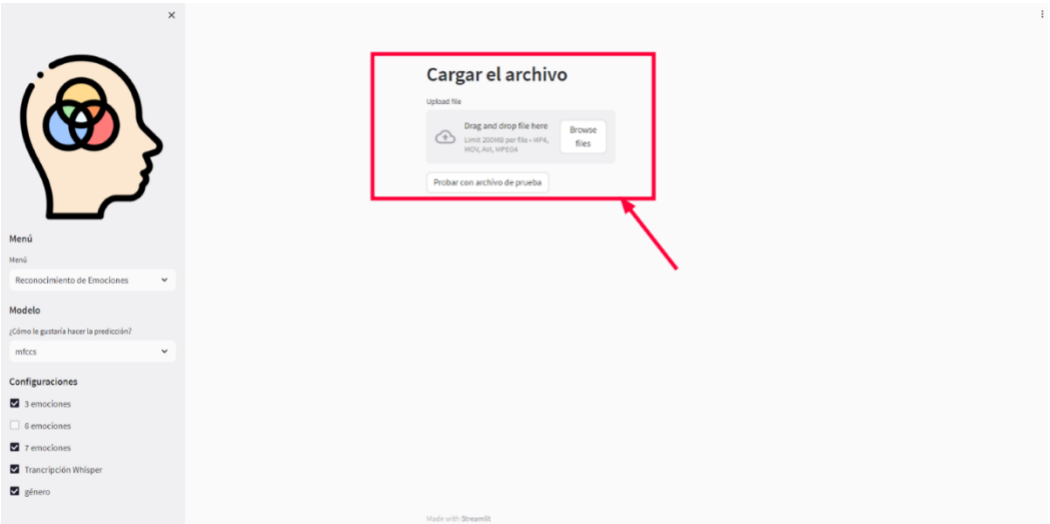
Después de completar el proceso de entrenamiento e implementación de los modelos de inteligencia artificial, se avanzó en la integración de las tres funciones individuales para detectar emociones (imagen facial, audio y texto) en la aplicación desarrollada en Streamlit.

Reconocimiento de emociones a partir de Video

El sistema facilita la subida de vídeos en formato .mp4, como se ilustra en la Figura 95. El procedimiento para cargar los vídeos se lleva a cabo a través de la función de carga de archivos. Además, se admiten vídeos de hasta 200 MB en formato .mp4 para garantizar una experiencia de usuario cómoda y fluida al interactuar con la aplicación.

Figura 95

Proceso de Carga de videos en el aplicativo



El video es procesado y se obtienen las distintas emociones, se muestra dentro del aplicativo tanto el video original que ingreso el usuario como el video con las emociones procesadas en la parte derecha (Figura 96).

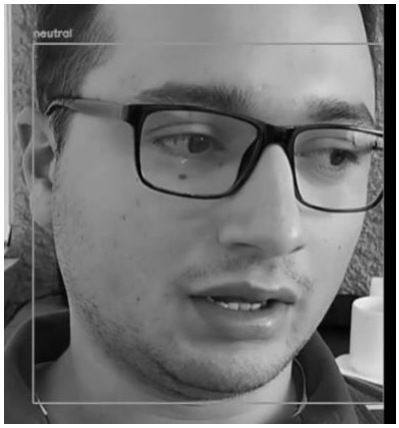
Figura 96

Video Original y video Procesado con emociones faciales obtenidas.

Se puede maximizar el video dentro del aplicativo para observar con mayor detalle las emociones que presenta la persona en ese momento (Figura 97), en este caso la persona presenta una emoción neutral según sus movimientos faciales.

Figura 97

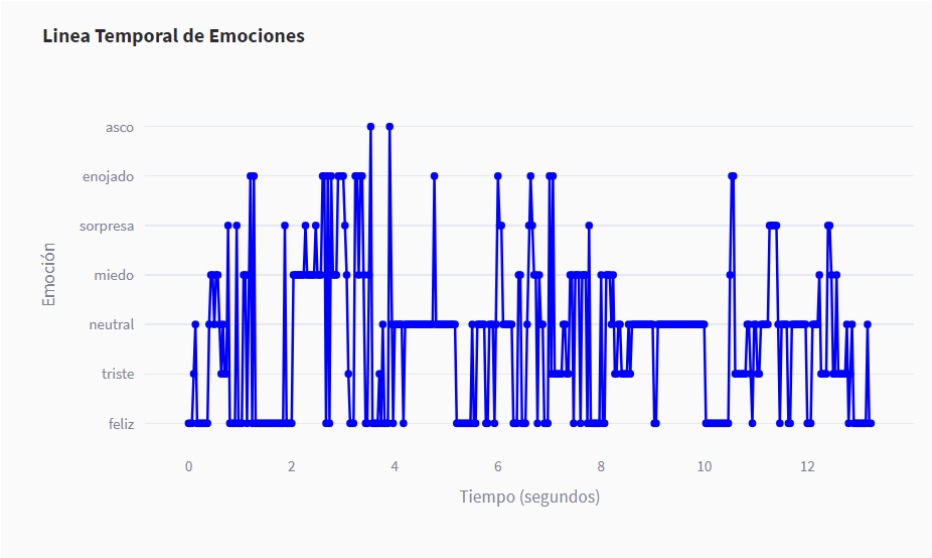
Rectángulo dibujado alrededor de la Cara, adicional a la emoción obtenida.



Además, se obtiene una línea temporal de emociones como se observa en la Figura 98, en los picos se observan las emociones que experimento la persona según el modelo predictivo facial.

Figura 98

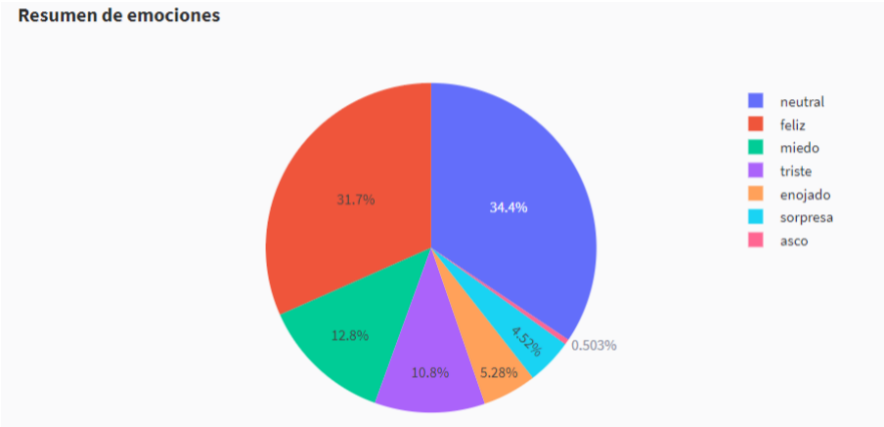
Línea temporal de emociones generada en el aplicativo



También se genera un resumen completo de las emociones que tenía la persona según el modelo predictivo, de esta forma se puede obtener la emoción general que experimento la persona según el modelo (Figura 99)

Figura 99

Resumen interactivo generado en el aplicativo



Reconocimiento de emociones a partir de Audio

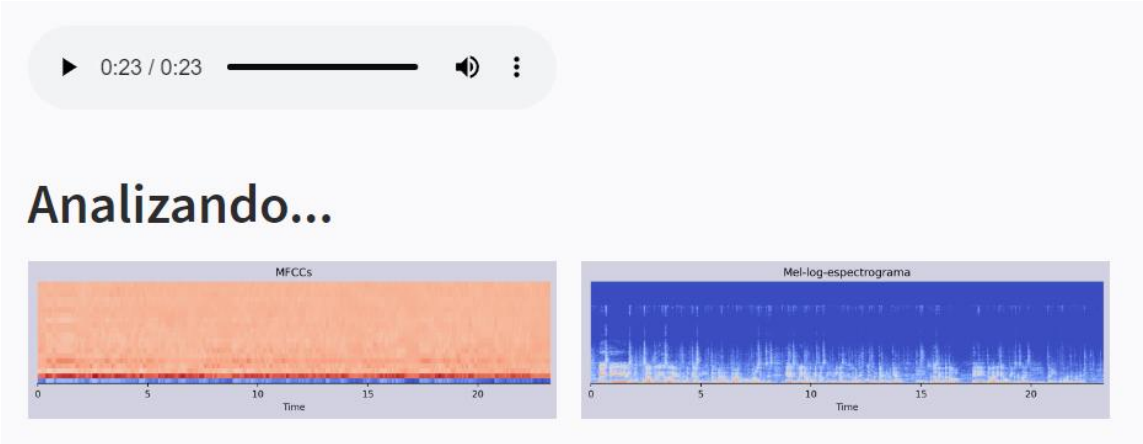
Se genera a partir del video ingresado de forma automática la forma de onda representativa del audio de la persona en dicho video (Figura 100), adicional se genera también un reproductor exclusivo de audio (Figura 101).

Figura 100

Forma de onda obtenida a partir del audio

Figura 101

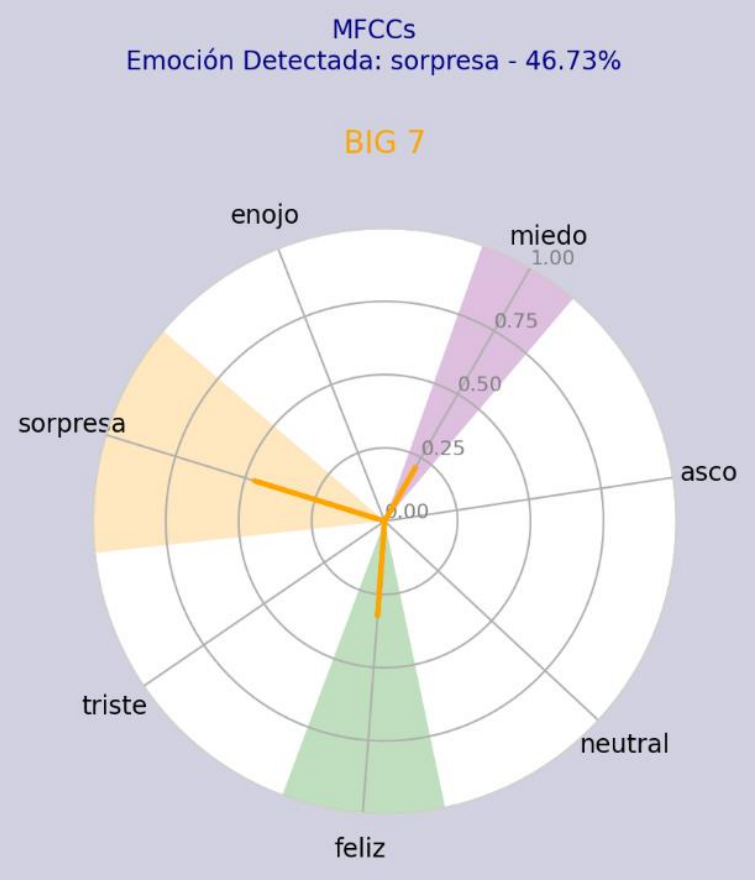
Análisis de Características pertenecientes al audio generadas por el sistema



Finalmente, mediante un gráfico polar, se muestran las probabilidades de emociones a partir solamente del audio de la persona detectadas por el modelo de reconocimiento de emociones del habla (Figura 102).

Figura 102

Gráfico Polar que representa las emociones según su porcentaje

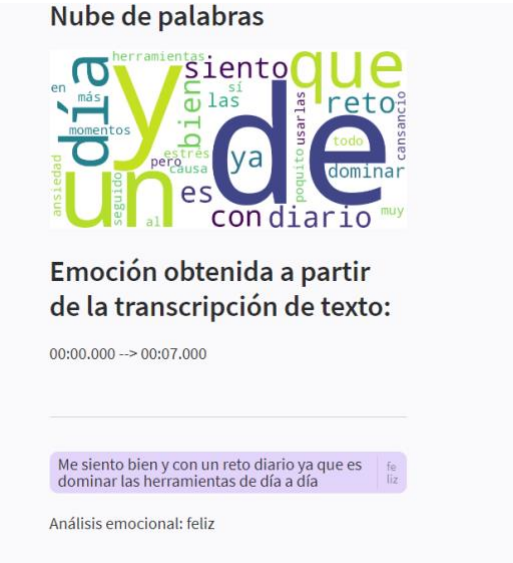


Reconocimiento de emociones a partir de Texto

Dentro del reconocimiento de emociones del texto, se obtiene una nube de palabras representativa de todo lo que la persona ha dicho, y se separa en marcas de tiempo, luego de esto se obtiene la emoción y se anota en el análisis emocional obtenido por el modelo Pysentimiento

Figura 103

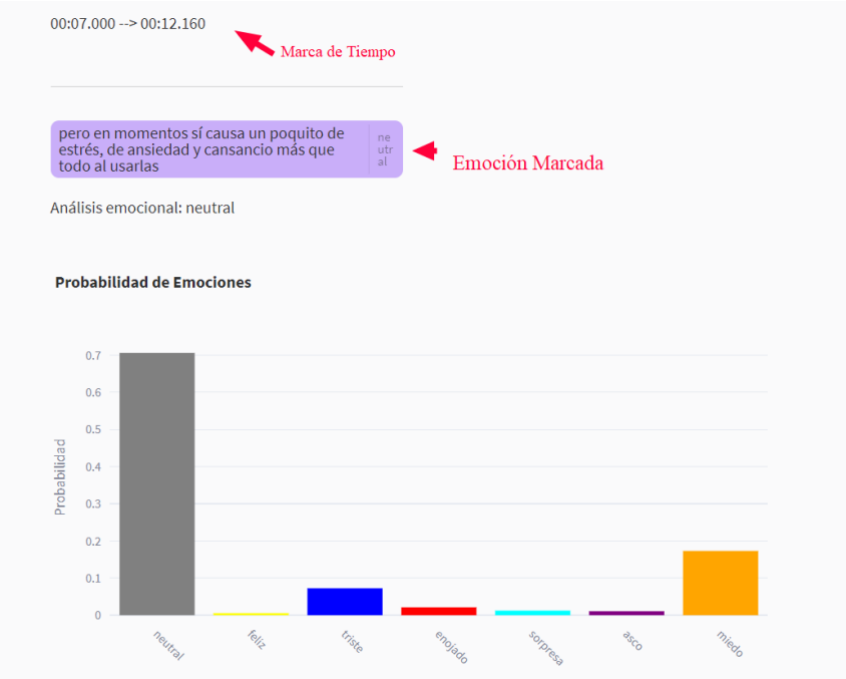
Nube de palabras obtenida mediante la transcripción de texto.



Adicional, en cada una de las oraciones recortadas mediante Whisper, se crea la emoción marcada y un gráfico de barras que muestra la probabilidad de las distintas emociones del texto (Figura 104).

Figura 104

Gráfico de probabilidad de emociones según el texto.



Para demostrar el flujo completo de la aplicación, se muestran todos los resultados generados en el aplicativo en la Figura 105, Figura 106 y Figura 107

Figura 105

Video Cargado en el sistema y resultados obtenidos.

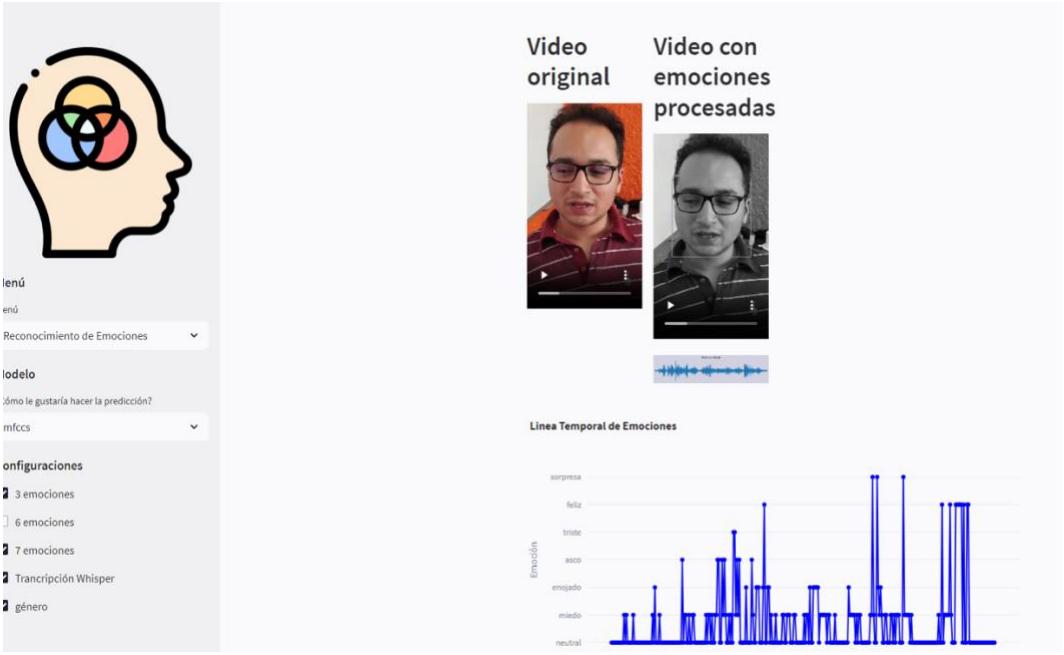


Figura 106

Video Cargado en el sistema y resultados obtenidos

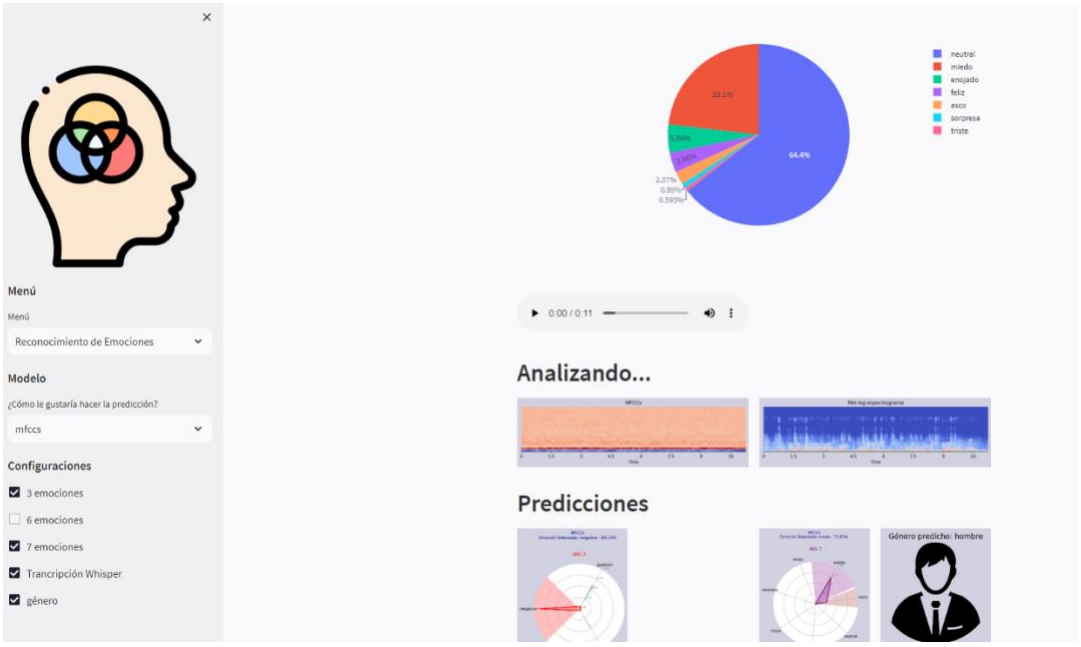


Figura 107

Video Cargado en el sistema y resultados obtenidos.

Capitulo V

Evaluación de Resultados

Descripción de los Datos

El conjunto de datos manejados durante el proceso de reconocimiento emocional implica la realización de una entrevista en un entorno laboral no controlado, bajo la aplicación de un test que permite identificar sintomatología de tecnoestrés en los usuarios. Se observa señales fisiológicas, gesticulaciones, tono de voz y puntos clave que ayudan a determinar el comportamiento bajo condiciones relacionadas al manejo de las herramientas tecnológicas. Cada muestra tomada es extraída en formato .mp4 y particionada bajo una línea de tiempo donde se identifica cada pregunta de la entrevista por usuario. Adicional, cada partición es transformada en frames o más conocidos como fotogramas por segundo, de esta manera el sistema permite identificar de manera más eficiente las emociones expresadas durante la realización de sus funciones dentro del área laboral.

Recolección de datos iniciales

Para el análisis multimodal de emociones se receptaron archivos de audio y video, se tomó una muestra total de 20 personas “in the wild” o en la naturaleza, esto significa que se graba en entornos no controlados del mundo real, capturando las expresiones faciales naturales y las emociones de las personas durante sus actividades cotidianas. A diferencia de los entornos de laboratorio controlados, donde la iluminación, los ángulos de la cámara y los sujetos se controlan cuidadosamente, los videos tomados en la naturaleza suelen ser espontáneos e impredecibles, lo que puede presentar varios desafíos para el reconocimiento de emociones.

En primera instancia se estableció el análisis del tema propuesto en esta investigación bajo el conocimiento y experiencia de la Psicóloga Clínica Betsy Morales, durante el proceso de comunicación se mencionaron temas en cuanto al área de salud ocupacional misma que tiene gran impacto al momento de realizar el proceso de reconocimiento emocional. Gracias al

conocimiento de la experta en el tema se plantearon puntos clave a considerar durante el proceso de reconocimiento emocional y bajo el modelo emocional universal propuesto por Paul Ekman quien maneja siete emociones a través de las cuales se identifica a un individuo y se determina su estado emocional, en este caso el framework a desarrollar se encuentra dirigido a un entorno de trabajo no controlado. Ver Apéndice A

Para la recolección de datos iniciales se realizó una investigación previa de test validado que permiten identificar patrones en cuanto a reconocimiento emocional, una vez identificada la muestra se procedió con la aplicación del test obteniendo como resultado un video en formato mp4, mismo que será cargado y entrenado en el modelo. Las entrevistas a aplicar se encuentran revisadas y validadas bajo consentimiento de la psicóloga especializada en área. Ver Apéndice B

En la Figura 108 se puede apreciar que los videos recolectados fueron aplicados a la muestra total de la organización, clasificados por género y área de trabajo, siendo 6 personas de género masculino y 14 de género femenino, adicional nos enfocamos en áreas de contabilidad y sistemas ya que el entorno laboral se caracteriza por manejar las áreas mencionadas.

Figura 108

Clasificación de la muestra por género y área de acuerdo con el personal encuestado.

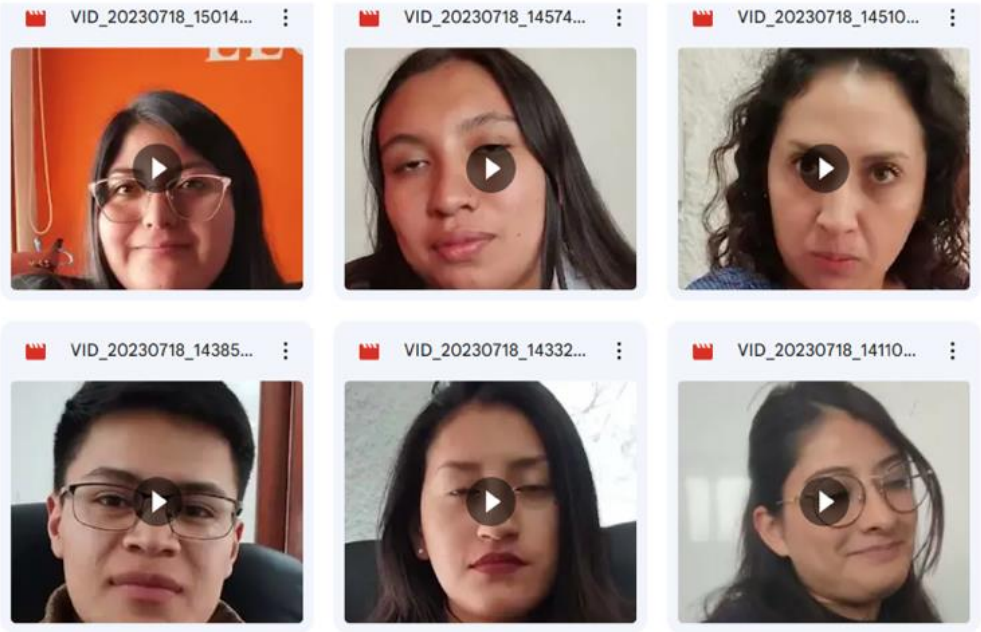
Nombre Video	Genero	Encuestado/a	Área
VID_20230718_134816.mp4	Masculino	Jean Pierre Toapanta	Sistemas
VID_20230718_140611.mp4	Femenino	Ivone Ushiña	Contabilidad
VID_20230718_141109.mp4	Femenino	Karen Guerra	Contabilidad
VID_20230718_143327.mp4	Femenino	Melisa Peralta	Contabilidad
VID_20230718_143858.mp4	Masculino	Alex Carrillo	Sistemas
VID_20230718_145108.mp4	Femenino	Sara Chiriboga	Contabilidad
VID_20230718_145745.mp4	Femenino	Shyrley Castañeda	Contabilidad
VID_20230718_150147.mp4	Femenino	Karina Ochoa	Contabilidad
VID_20230718_150523.mp5	Femenino	Vanesa Gualoto	Contabilidad
VID_20230718_151118.mp6	Femenino	Joselyn Mantilla	Contabilidad
VID_20230718_151723.mp7	Femenino	Abigail Guerra	Contabilidad
VID_20230718_142409.mp8	Femenino	Karla Ciro	Contabilidad
VID_20230718_143632.mp9	Femenino	Fernanda Gualacata	Contabilidad
VID_20230718_144236.mp10	Femenino	Samanta García	Contabilidad
VID_20230718_144813.mp11	Femenino	Lizeth Maza	Contabilidad
VID_20230718_145352.mp12	Masculino	Nicolás Villavicencio	Sistemas
VID_20230718_145945.mp13	Masculino	Edwin Bravo	Sistemas
VID_20230718_231529.mp14	Femenino	Elizabeth Morejón	Contabilidad
VID_20230718_233802.mp15	Masculino	Carlos Quiroz	Sistemas
VID-20230721-WA0001.mp4	Masculino	Steeven Vargas	Sistemas

El framework permite cargar los videos recolectados analizando facciones faciales, gestuales y vocales respectivamente, de esta manera interpreta el estado emocional de la persona analizada.

Para el desarrollo del framework de reconocimiento emocional se utilizaron plataformas en la nube como Google Drive, el cual, permitió alojar los videos receptados de la muestra para el análisis emocional respectivo en el sistema desarrollado. Ver Figura 109.

Figura 109

Material a ser analizado en el sistema desarrollado



Verificación de la calidad de los Datos

Los datos extraídos son previamente analizados por el profesional especializado en el área de psicología para identificar el comportamiento de los individuos y las características en torno a sintomatología relacionada con el tecnoestrés o estrés laboral, durante la realización del presente trabajo el sistema se encarga de procesar y automatizar los datos a medida que se identifiquen bajo un método multimodal comprendido entre reconocimiento facial, auditivo y textual obteniendo como resultado gráficas estadísticas que distingue el comportamiento emocional del individuo durante la aplicación del test.

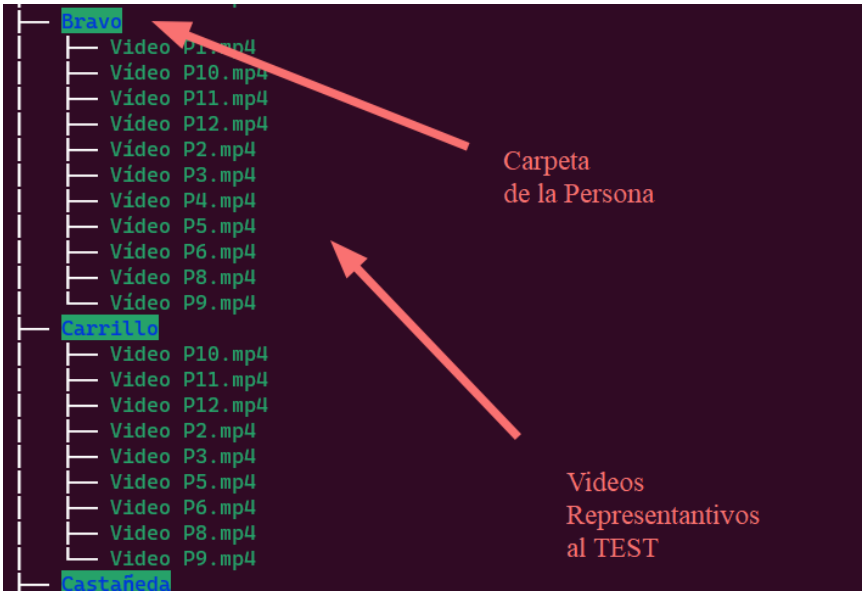
Análisis de los Datos

La muestra recolectada fue analizada previamente por el especialista en el campo psicológico. Antes de ser procesada por el sistema se identificaron patrones de comportamiento relacionado con el tema propuesto, de esta manera se aplicaron modelos de reconocimiento facial, del habla y textual bajo codificación en lenguaje de programación Python con la intervención de sets de datos para el proceso de entrenamiento del sistema, de esta manera permite identificar a través de las 7 emociones predefinidas el estado emocional del individuo.

La Figura 110 comprende la muestra empleada para el proceso de reconocimiento emocional extraído en formato .mp4:

Figura 110

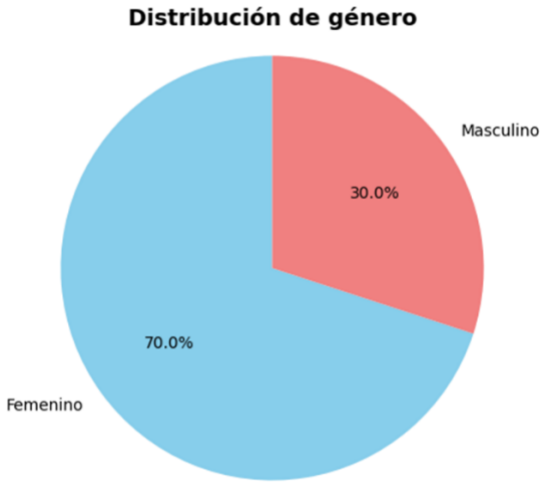
Extracción de la muestra en formato .mp4



La Figura 111 permite identificar que la mayoría de la muestra analizada son de género femenino, esto se debe a que el entorno laboral utilizado se enfoca en áreas de administración:

Figura 111

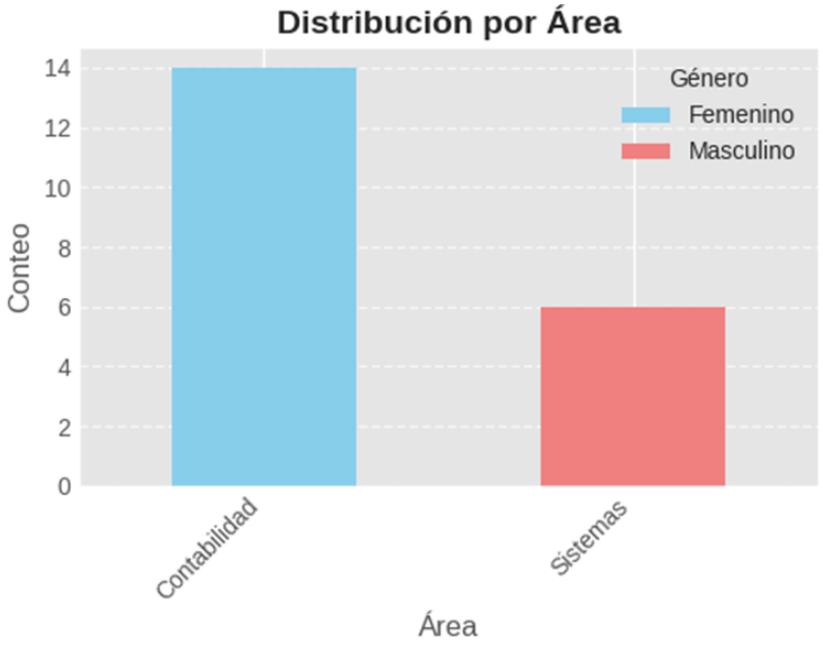
Detalle de longitud de los videos muestrales.



La cantidad de participantes son un total 20 individuos, donde la mayoría son mujeres y pertenecen al área de contabilidad y un 30% pertenecen al área de sistemas. Ver Figura 112.

Figura 112

Detalle de las personas participantes en la prueba



Proceso de Evaluación

La Tabla 15 muestra el proceso funcional de evaluación del sistema al identificar las emociones de un individuo:

Tabla 15

Proceso funcional de evaluación del sistema

Proceso	Descripción
Ejecución del Sistema	El usuario debe regirse a contestar el test, de esta manera se recolecta un video en formato .mp4 el cuál será procesado por el sistema diseñado.
Vista previa del Sistema	El sistema presenta la opción de cargar un video, durante este proceso se observa una barra de progreso que determina la carga del video al sistema, una vez que el video ha sido completado se procesa la información enfocada en el método multimodal facial, auditivo y textual.
Carga de video	Cuando la información del video ha sido procesada correctamente, se despliega una serie de gráficos, interpretando las emociones transmitidas por un individuo, mismas que serán analizadas por un profesional en salud ocupacional y psicológica.
Análisis de video	Al procesar el video las emociones que predominan se muestran en tonos más fuertes, permitiendo al profesional experto en salud ocupacional y psicológica enfocarse y determinar el estado emocional y el comportamiento en torno al uso de herramientas tecnológicas dentro del área laboral.
Fin	El proceso es ejecutado una única vez al cargar un video predeterminado, para terminar con su ejecución se procede a cerrar la ventana flotante presentada a la vista del usuario.

Resultados Obtenidos

Resultados de Datos Generales

Sujeto de Prueba 1

- Género: Masculino
- Área Laboral: Sistemas

Figura 113

Video 1

Análisis Psicológico. Gesticulaciones neutrales. Reconoce las TICs como una herramienta que ayuda en varias áreas y tener el máximo provecho. Denota gran satisfacción con el uso y la presencia de las TICs. No ha experimentado estrés ansiedad o malestar al usar por tiempo prolongado las TICs. Experimenta niveles de ansiedad normales y esperables en el trabajo cuando hay pendientes o información importante que puede dañarse o eliminarse. Sí ha experimentado ansiedad si no está conectado al internet, menciona que no puede estar sin esa conexión; menciona que es importante poder administrar mejor el tiempo que usamos para la tecnología. También siente que pierde el sentido del tiempo. Logra identificar sus logros con satisfacción.

Sujeto de Prueba 2

- Género: Femenino
- Área Laboral: Contabilidad

Figura 114

Video 2

Análisis Psicológico. Inquietud psicomotora. (luz en el video no se puede ver con claridad). Ha sentido ansiedad al empezar a utilizar nuevas tecnologías. Adecuadas estrategias de afrontamiento para el estrés. Siente niveles de ansiedad y estrés esperables debido a la naturaleza de su trabajo. No experimenta estrés al no estar conectada a internet, salvo cuando tiene que realizar algún trabajo. Reconoce sus logros con satisfacción. Reconoce que las TICs ayudan en su trabajo a ser más automático. Menciona satisfacción con el uso del as TICs en su trabajo porque va a aprendiendo como utilizar las nuevas tecnologías para su beneficio.

Sujeto de Prueba 3

- Género: Femenino
- Área Laboral: Contabilidad

Figura 115

Video 3

Análisis Psicológico. Debido al ángulo y toma de la cámara es dificultoso reconocer adecuadamente las expresiones faciales. Mantiene buenas estrategias de afrontamiento, entiende que el aprendizaje puede ser frustrante en inicio. Cuando ha sentido malestar por acumulación de trabajo o uso prolongado de las TICs se toma 5 minutos para despejarse. Al trabajar con información importante ha desarrollado estrategias para resguardar esa información. No siente estrés al no tener conexión a internet, salvo cuando tiene un trabajo que realizar. Reconoce las TICs como herramientas que permiten ser más eficientes en el trabajo y en general

Sujeto de Prueba 4

- Género: Femenino
- Área Laboral: Contabilidad

Figura 116

Video 4

Análisis Psicológico. Participante con gesticulación acorde a sus respuestas. Muestra satisfacción y tranquilidad durante la entrevista. Con adecuadas estrategias de afrontamiento y consciente de que usar nuevas tecnologías en un inicio generan frustración y posteriormente aportan en el desempeño laboral y el ahorro de tiempo. Sí ha experimentado ansiedad en situaciones en las que tiene que manjar información importante y por un error puede dañarse o perderse. Reconoce las TICs como una herramienta que porta en varios sentidos.

Sujeto de Prueba 5

- Género: Masculino
- Área Laboral: Sistemas

Figura 117

Video 5

Análisis Psicológico. Facies neutrales. Mantiene un adecuado interés en el aprendizaje de las nuevas tecnologías. Entiende que al inicio se puede sentir frustración, pero después se da un interés para poder usarlo de manera adecuada. Sí ha experimentado niveles altos de ansiedad en 4 ocasiones, menciona que es demandante el trabajo. Tiene adecuadas estrategias de afrontamiento al estrés. Reconoce las TICs como herramientas que ayudan a agilizar el trabajo y muestra flexibilidad en continuar usando nuevas tecnologías. Expresa satisfacción media en el uso de las TICs.

Sujeto de Prueba 6

- Género: Femenino
- Área Laboral: Contabilidad

Figura 118

Video 6

Análisis Psicológico. Debido al ángulo y toma de la cámara es dificultoso reconocer adecuadamente las expresiones faciales. La participante entiende el uso del as TICs como un aprendizaje que puede ayudar a agilizar el trabajo y a la organización, recalca que dependiendo de las circunstancias puede ser más o menos beneficioso (internet lento o rápido). Menciona que experimenta cierta preocupación o ansiedad cuando no se tiene la capacitación adecuada para la utilización de una nueva herramienta, pero con la ayuda necesaria se pueden lograr muchos avances. Estrategias de afrontamiento adecuadas para el estrés.

Sujeto de Prueba 7

- Género: Femenino
- Área Laboral: Contabilidad

Figura 119

Video 7

Análisis Psicológico. Neutralidad en gesticulación. Reconoce las TICs como una ayuda para el trabajo que realiza y tiene interés por aprender a usarlas. Ha experimentado falta de concentración por la cantidad de estímulos que existen en este medio. Experimenta niveles de ansiedad esperables para el trabajo que desarrolla. Menciona que ha tenido mucho cansancio después del uso prolongado de las TICs. No siente que las TICs interfieren en su planificación diaria, sino más bien como una ayuda para agilizar los procesos.

Sujeto de Prueba 8

- Género: Femenino
- Área Laboral: Contabilidad

Figura 120

Video 8

Análisis Psicológico. Tiene facies de felicidad durante toda la entrevista. Experimenta preocupación y frustración cuando hay algo que no puede usar respecto a las TICs, pero cuando logra entender cómo funciona se siente motivada a continuar con su trabajo. Tiene adecuadas estrategias de afrontamiento al estrés. Expresa sentirse conforme con el uso y la implementación de las TICs en su lugar de trabajo.

Sujeto de Prueba 9

- Género: Femenino
- Área Laboral: Contabilidad

Figura 121

Video 9

Análisis Psicológico. Inquietud psicomotora. Reconoce como aprendizajes y oportunidades el uso de nuevas tecnologías en el trabajo. Reconoce las actividades frente al uso prolongado de TICs. Experimenta niveles de estrés normales frente al daño o mal uso de datos en el trabajo. Cuando se trata del trabajo sí experimenta ansiedad cuando no hay acceso a internet. Logra reconocer sus logros y reconoce que las TICs facilitan el trabajo y ahorra tiempo. Expresa sentirse más productiva en su trabajo al usar las TICs

Sujeto de Prueba 10

- Género: Femenino
- Área Laboral: Contabilidad

Figura 122

Video 10

Análisis Psicológico. Facies neutrales. Experimenta niveles de ansiedad y estrés normales, tiene estrategias de afrontamiento adecuadas al estrés. Si experimenta que el uso de las TICs le hace perder el sentido del tiempo. Reconoce que las TICs ayudan a automatizar el trabajo. No expresa desagrado alguno por el uso de las TICs.

Sujeto de Prueba 11

- Género: Femenino
- Área Laboral: Contabilidad

Figura 123

Video 11

Análisis Psicológico. Facies neutrales durante la entrevista. Experimenta niveles de estrés normales, no existe mayor comunicación por parte de la participante para inferir que el estrés puede estar influyendo negativamente en su bienestar general. Estrategias de afrontamiento adecuadas al estrés, maneja tiempos y no experimenta necesidad de tener todo el tiempo internet para realizar sus actividades. Reconoce las TICs como una herramienta de ayuda para la agilización de procesos. Menciona satisfacción por el uso de las TICs en su área laboral.

Sujeto de Prueba 12

- Género: Femenino
- Área Laboral: Contabilidad

Figura 124

Video 12

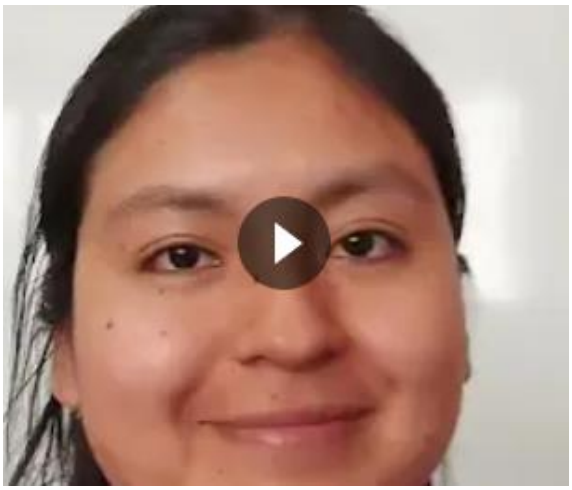
Análisis Psicológico. Facies neutrales, se mantiene tranquila durante toda la entrevista. Estrategias de afrontamiento adecuadas para el estrés (centradas en el problema). De acuerdo con su impresión personal el uso de las tecnologías no causa molestias o preocupaciones exageradas más bien ayudan a agilizar los trabajos y a ser más eficientes. La participante aparentemente experimenta niveles de estrés y preocupación normales para la resolución de conflictos. Menciona satisfacción al tener un logro y al poder usar las TICs en su trabajo y vida diaria.

Sujeto de Prueba 13

- Género: Femenino
- Área Laboral: Contabilidad

Figura 125

Video 13



Análisis Psicológico. Facies neutrales, se mantiene tranquila durante toda la entrevista. Estrategias de afrontamiento adecuadas para el estrés (centradas en el problema). De acuerdo con su impresión personal el uso de las tecnologías no causa molestias o preocupaciones exageradas masa bien ayudan a agilizar los trabajaos y a ser más eficientes. La participante aparentemente experimenta niveles de estrés y preocupación normales para la resolución de conflictos. Menciona satisfacción al tener un logro y al poder usar las TICs en su trabajo y vida diaria.

Sujeto de Prueba 14

- Género: Femenino
- Área Laboral: Contabilidad

Figura 126

Video 14

Análisis Psicológico. La participante tiene dificultades en organizar su discurso, es entendible. Experimenta niveles de estrés normales frente al uso de una tecnología completamente nueva en su trabajo. Reconoce las actividades que puede realizar cuando el uso prolongado de TICs empieza a causar malestar. Tiene adecuadas estrategias de afrontamiento (centradas en el problema). Si experimenta ansiedad cuando no tiene acceso al internet, más cuando tiene que realizar trabajos importantes. Reconoce los medios tecnológicos como herramientas que ayudan, pero también reconoce el mal uso de estas puede ser perjudicial. Logra reconocer sus logros.

Sujeto de Prueba 15

- Género: Femenino
- Área Laboral: Contabilidad

Figura 127

Video 15

Análisis Psicológico. Inquietud psicomotora. Mantiene buena conexión con la entrevista. Reconoce las TICs como herramientas que facilitan en varios sentidos el trabajo. Reconoce adecuadamente las actividades que puede utilizar cuando el uso prolongado de las TICs empieza a causar malestar. Afirma que experimenta ansiedad o estrés cuando tiene trabajo pendiente o está manejando información importante que puede dañarse o perderse. De acuerdo con las respuestas de la participante sí experimenta estrés en niveles normales en las actividades de la vida diaria y no generan malestar más allá del oportuno para dar solución a ciertas situaciones. Logra sentir satisfacción cuando tiene éxito un proyecto o trabajo. Adecuadas estrategias de afrontamiento al estrés.

Sujeto de Prueba 16

- Género: Masculino
- Área Laboral: Sistemas

Figura 128

Video 16

Análisis Psicológico. Inquietud psicomotora, facies de incomodidad y aparente dolor muscular en la zona cervical. Actitud poco profesional y despreocupada. Cuando el uso prolongado de TICs causa malestar, sus facies cambian drásticamente acompañando al sentir de inconformidad, reconoce que en esta situación debe dejar el trabajo, menciona que “hasta que le pase”, no tiene claras que acciones específicas le pueden ayudar a disminuir ese malestar. Estrategias de afrontamiento poco adecuadas ante el estrés. Reconoce que el nivel de estrés que experimenta es considerable en situaciones de daño o perdida de información y en la acumulación de trabajo, además afirma que en su caso personal el uso de las TICs sí le significa varias horas de pérdida de tiempo sin que se dé cuenta. Reconoce que al tener éxito en proyectos o trabajos desarrollados en las TICs es gratificante. Existe duda en su respuesta a la satisfacción que siente al utilizar las TICs; su respuesta es ambigua.

Sujeto de Prueba 17

- Género: Masculino
- Área Laboral: Sistemas

Figura 129

Video 17

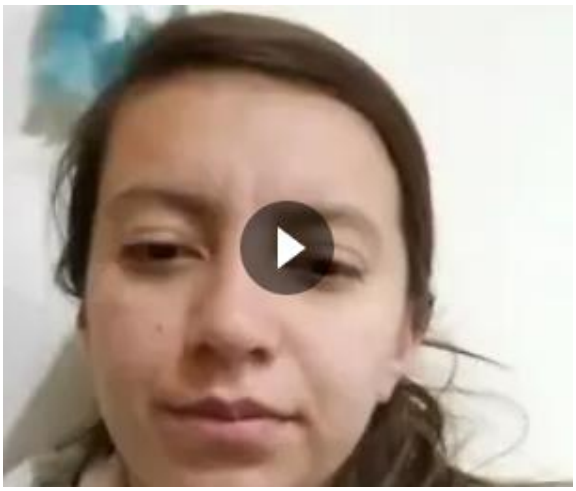
Análisis Psicológico. El participante presenta nerviosismo, evita el contacto visual, usa muletillas constantemente. Entiende y aplica la pausa activa cuando el uso prolongado de TICs empieza a causar molestas, cansancio, agotamiento mental. Siente temor ante la posibilidad de daño o perdida información importante, por lo cual previene con backup. No experimenta ansiedad al no estar en contacto con las TICs o internet. Expresa satisfacción al tener un proyecto exitoso, reconoce las TICs como herramientas que solucionan problemas; no existe rechazo hacia las tecnologías, sino los acoge para su área laboral como personal.

Sujeto de Prueba 18

- Género: Femenino
- Área Laboral: Sistemas

Figura 130

Video 18



Análisis Psicológico. La participante tiene gesticulaciones neutrales durante la entrevista. En momentos se puede divisar incertidumbre, esto debido a la recuperación de la información para dar respuesta a las preguntas en la entrevista. Reconoce las TICs como una herramienta que facilita la comunicación y el trabajo. Reconoce las estrategias necesarias para cuando el uso prolongado de TICs genera cansancio, desenfoque y malestar o tensión muscular. El mantener pendientes en el trabajo (p.j. correos electrónicos) le genera estrés por una valoración subjetiva de acumulación de trabajo que puede mantenerse durante el día. En la participante es más notorio los rasgos de personalidad ansiosa, situación por la cual es más propensa a experimentar mayores niveles de estrés que el resto de las personas. Sin embargo, las estrategias de afrontamiento son adecuadas hacia el estrés, aunque podría existir cierta rumiación en el pensamiento por los pendientes o errores que se puedan cometer en el ambiente laboral. Expresa satisfacción al trabajar con las TICs al considerarlas como una herramienta de ayuda en varias áreas del desarrollo personal.

Sujeto de Prueba 19

- Género: Masculino
- Área Laboral: Sistemas

Figura 131

Video 19

Análisis Psicológico. El participante presenta facies de incertidumbre en el ceño, sin embargo, se desenvuelve de manera adecuada al dar respuestas específicas y claras en la entrevista. Reconoce las TICs como un medio de ayuda a resolución de conflictos y trabajos. El participante conoce sobre las estrategias a usar cuando el uso prolongado de las TICs empieza a causar molestias, cansancio, etc. También menciona que el trabajo con uso de TICs llega a tener mayores distracciones lo que puede enlentecer los resultados y rapidez del trabajo. Existe miedo al tener que manejar información o datos que no tengan backup de respaldo. El participante a pesar de que en ciertas circunstancias existe temor y malestar maneja de manera adecuada el estrés y además está interesado en continuar conocer nuevas tecnologías con las que pueda trabajar. Facies (gesticulaciones).

Sujeto de Prueba 20

- Género: Masculino
- Área Laboral: Sistemas

Figura 132

Video 20

Análisis Psicológico. El participante se muestra satisfecho en relación con el uso de las TICs en su experiencia laboral. Muestra facies que denotan seguridad, satisfacción y en ocasiones de duda, esta última mayormente relacionada con el procesamiento de la información al recordar experiencias para brindar una respuesta acorde. La respuesta del participante cuando experimenta cansancio ansiedad, tensión muscular al usar por tiempos prolongados las TICs, es bastante favorable y beneficiosa, tanto para su desempeño laboral como su cuidado personal y emocional. Por las respuestas brindadas en la entrevista se puede entender que el participante tiene buenos estrategias de afrontamiento al tecnoestrés, llegando a experimentarlo en niveles normales de estrés diario.

Resultados de Datos específicos

Para el siguiente análisis se toma de la muestra total 3 individuos específicos considerando características especiales propias de cada sujeto, de esta manera se analizará bajo criterio psicológico el comportamiento emocional a través del video cargado en el sistema, mismo que permite el reconocimiento del rostro, audio y el procesamiento textual determinando el estado del individuo bajo la opinión por parte de la persona experta en salud psicológica y ocupacional.

Sujeto de Prueba 19

- Género: Masculino
- Área Laboral: Sistemas

Condiciones. El individuo analizado se caracteriza por poseer rasgos fisiológicos fuertes, de forma física el usuario lleva accesorios extras como gorra, lentes y en su rostro se puede apreciar presencia de barba, estas características influyen en mayor impacto al momento de realizar el proceso de reconocimiento.

Análisis del Sistema

Reconocimiento Facial. En la Figura 133 se observa el video original cargado al sistema, seguido del video procesado en el cual se manifiestan emociones a lo largo de la respuesta dada por el individuo. La figura procesada marca un cuadro alrededor del rostro del individuo identificando las emociones expresadas a lo largo de la duración del video.

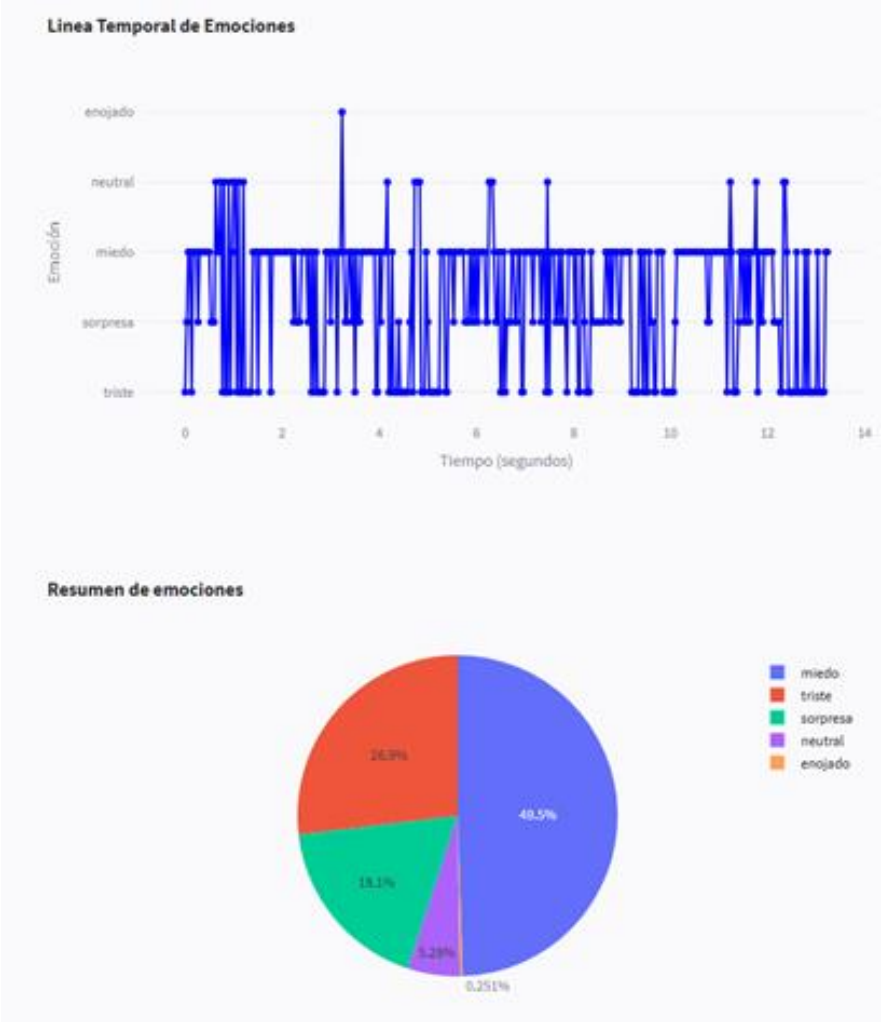
Figura 133

Video Cargado en el Sistema

La Figura 134 permite observar a través de una línea de tiempo los picos de las emociones que el sujeto a experimentando según el modelo predictivo facial, en el gráfico tipo pastel se recepta el conjunto de emociones detectadas, dando como resultado un resumen de emociones basado en valores porcentuales elevados, mismos que predominan en el usuario durante la realización del video. Las emociones predominantes indican que el usuario muestra miedo, tristeza y sorpresa al responder la entrevista realizada.

Figura 134

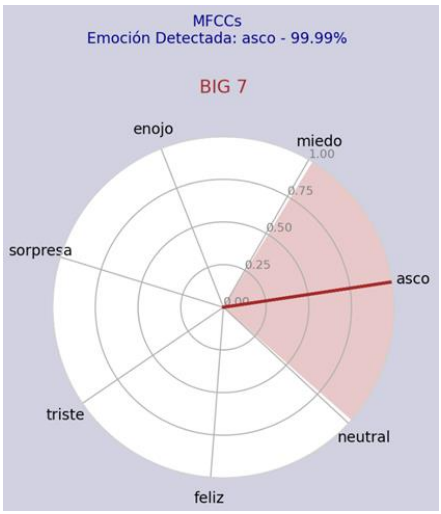
Emociones detectadas en el individuo procesado



Reconocimiento Auditivo: En la Figura 135 se observa el procesamiento del audio extraído a partir del video cargado en el sistema, el sujeto indica como emoción “asco” esto se debe a que la entrevista es grabada, enfocándose en sus facciones, causando incomodidad al momento de dar respuesta al test realizada.

Figura 135

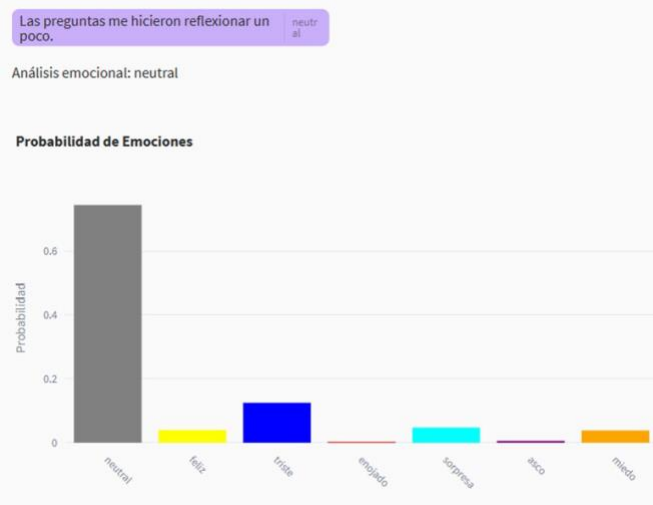
Procesamiento del audio extraído del video del individuo procesado



Reconocimiento Textual: La Figura 136 a través de un gráfico de barras indica el procesamiento textual a partir del video cargado donde el individuo muestra mayor nivel de felicidad complementando con la emoción neutral, esto se debe a momentos de incertidumbre en su respuesta.

Figura 136

Procesamiento textual a partir del video del individuo procesado



Análisis Psicológico. El individuo siente incertidumbre, incomodidad, porque no sabe que responder o está nervioso. En otra sintomatología muestra miedo, nerviosismo, responde al

hecho mismo de la entrevista. Tiene buena respuesta en torno al estrés, es muy flexible a aprender nuevas tecnologías a pesar de sentir temor. Por la voz su respuesta varía esto se debe a nerviosismo por ser una entrevista grabada.

Sujeto de Prueba 14

- Género: Femenino
- Área Laboral: Contabilidad

Condiciones. El usuario analizado se caracteriza por no llevar objetos y/o accesorios adicionales, de manera que permite apreciar su rostro con mayor facilidad, determinando puntos faciales clave para el proceso de reconocimiento.

Análisis del Sistema

Reconocimiento Facial: La Figura 137 muestra el video original cargado al sistema, seguido del video procesado, se observa que durante la reproducción del video se marca un cuadro alrededor del rostro del individuo identificando las emociones expresadas.

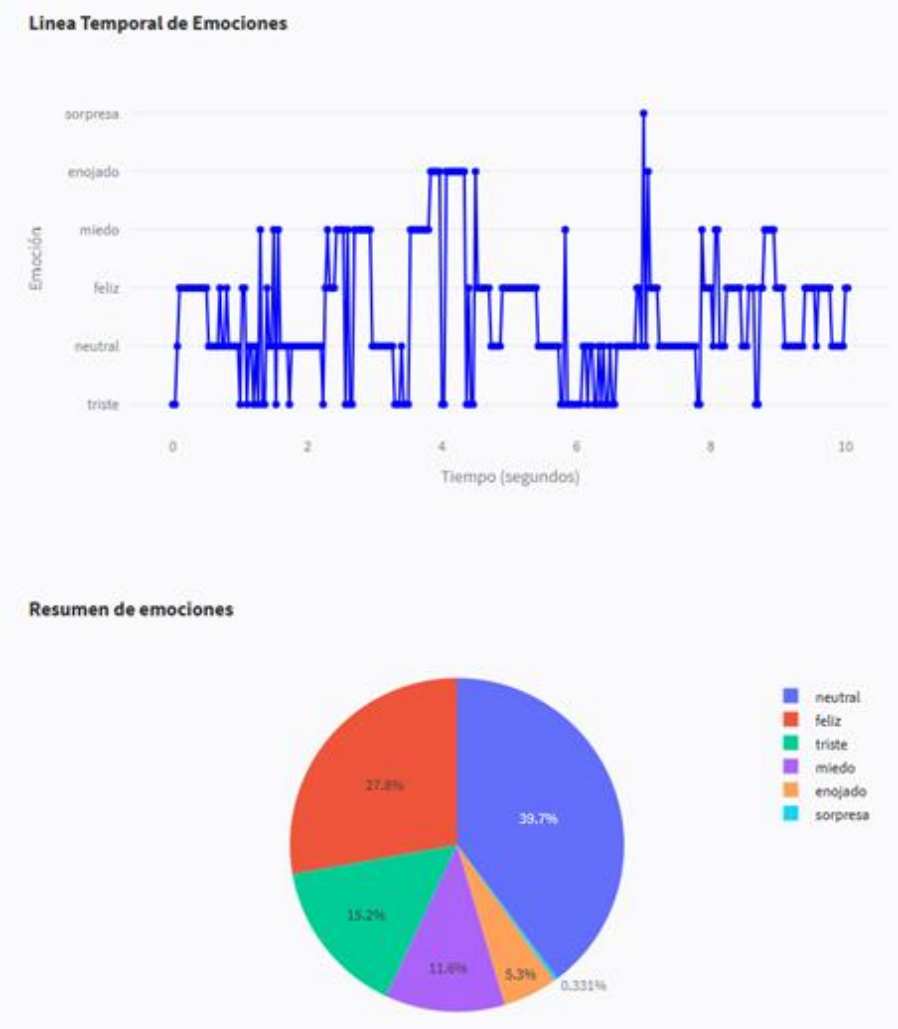
Figura 137

Video Cargado en el Sistema

La Figura 138 permite observar a través de una línea de tiempo los picos de las emociones que el sujeto a experimentando según el modelo predictivo facial, en el gráfico tipo pastel se recepta el conjunto de emociones detectadas, en esta ocasión el sujeto muestra neutralidad, felicidad y tristeza con valores porcentuales más altos a diferencia del resto de emociones.

Figura 138

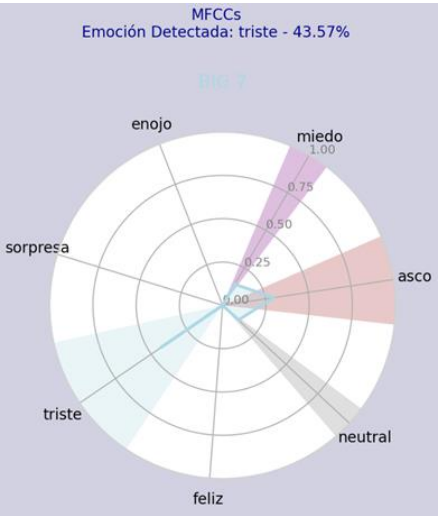
Emociones detectadas en el individuo procesado



Reconocimiento Auditivo. En la Figura 139 se observa el procesamiento del audio extraído a partir del video cargado en el sistema, se puede apreciar de forma polar que el usuario durante la realización del test expresa miedo, neutralidad y en mayor nivel asco.

Figura 139

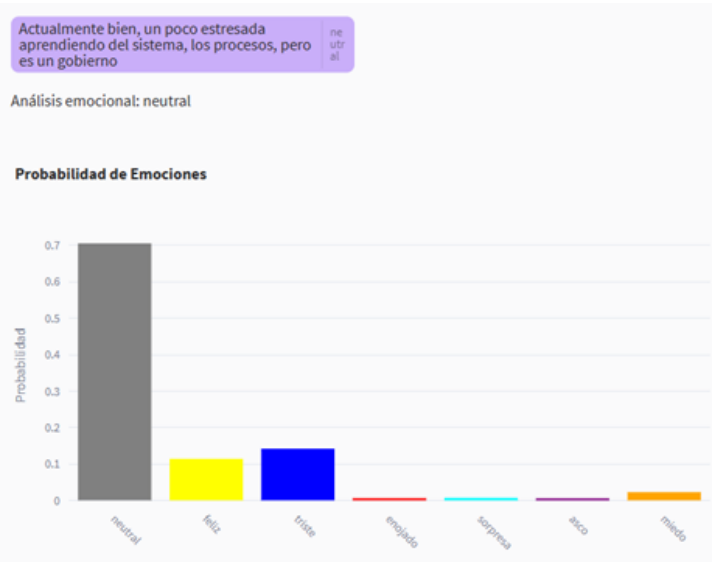
Procesamiento del audio extraído del video del individuo procesado



Reconocimiento Textual. En la Figura 140, a través de un gráfico de barras indica el procesamiento textual a partir del video cargado, se puede apreciar que el sujeto analizado expresa neutralidad, eso se debe a sintomatología de incertidumbre e inseguridad en su respuesta.

Figura 140

Procesamiento textual a partir del video del individuo procesado



Análisis Psicológico. La entrevistada presenta niveles normales de estrés laboral que son necesarios para que ella puede resolver problemas durante sus actividades diarias. Posee

estrategias adecuadas de afrontamiento para sobrellevar la carga laboral, se organiza muy bien y muestra satisfacción cuando tiene éxito en su trabajo.

Sujeto de Prueba 16

- Género: Masculino
- Área Laboral: Sistemas

Condiciones. El individuo analizado se caracteriza por poseer lentes, este accesorio es clave al momento de realizar el proceso de reconocimiento, debido a la fijación de su mirada ante la cámara durante la entrevista aplicada.

Análisis del Sistema

Reconocimiento Facial: La Figura 141 muestra el video original cargado al sistema, seguido del video procesado, se observa que durante la reproducción del video se marca un cuadro alrededor del rostro del individuo identificando las emociones expresadas.

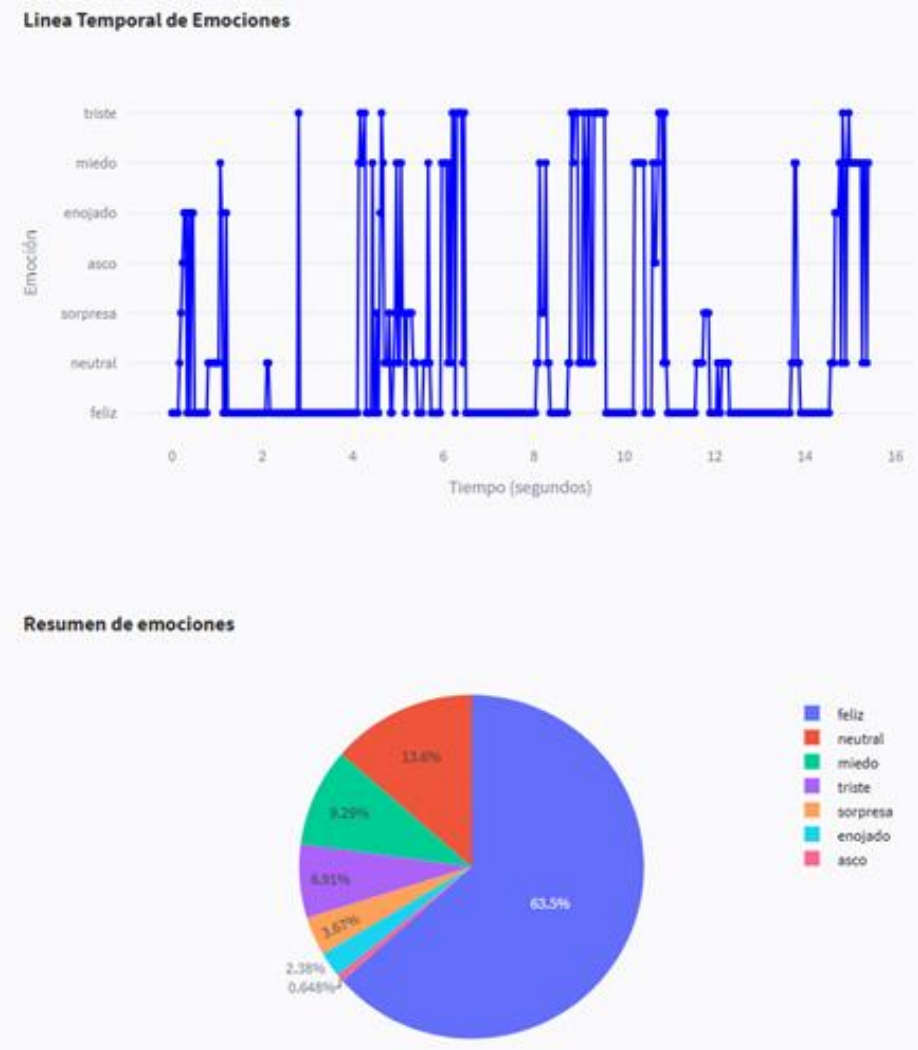
Figura 141

Video Cargado en el Sistema

La Figura 142 permite observar a través de una línea de tiempo los picos de las emociones que el sujeto a experimentando según el modelo predictivo facial, en el gráfico tipo pastel se recepta el conjunto de emociones detectadas, en esta ocasión el sujeto muestra felicidad en mayor porcentaje seguido de neutralidad y miedo con valores porcentuales menores.

Figura 142

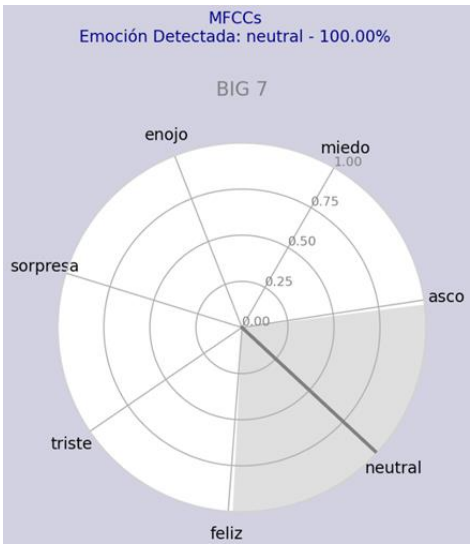
Emociones detectadas en el individuo procesado



Reconocimiento Auditivo. En la Figura 143 se observa el procesamiento del audio extraído a partir del video cargado en el sistema, se puede apreciar de forma polar que el sujeto durante la realización del test expresa neutralidad constante al momento de dar su respuesta.

Figura 143

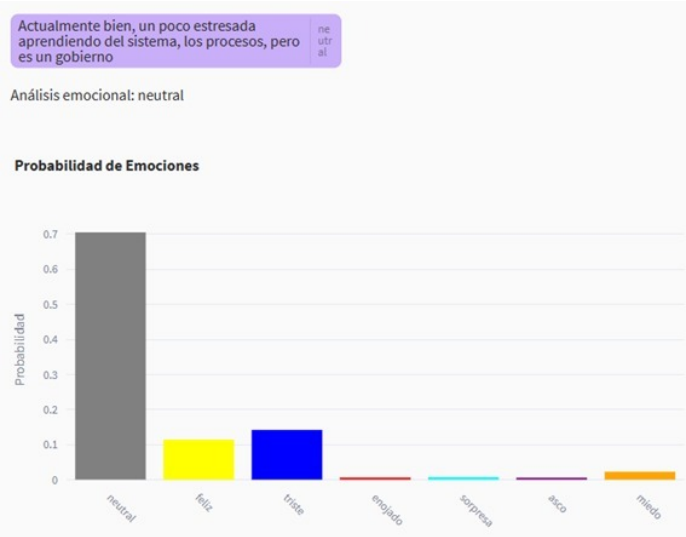
Procesamiento del audio extraído del video del individuo procesado



Reconocimiento Textual. La Figura 144 a través de un gráfico de barras indica el procesamiento textual a partir del video cargado, el sujeto analizado expresa neutralidad a través del texto, siendo su principal emoción durante el proceso de entrenamiento. La sintomatología neutral particularmente oculta expresiones emocionales específicas.

Figura 144

Procesamiento textual a partir del video del individuo procesado



Capítulo VI

Conclusiones y Recomendaciones

Conclusiones

Se logró la creación de un prototipo con la capacidad de reconocer y evaluar las emociones de un individuo en el contexto de la salud ocupacional. Esto se alcanzó mediante la combinación de un enfoque multimodal, que utiliza señales provenientes de expresiones faciales, gestos, comunicación no verbal y tono de voz. Para llevar a cabo este análisis integral, se emplearon diversas técnicas avanzadas de Aprendizaje Profundo, como el uso de arquitecturas de redes neuronales convolucionales.

Se realizó la revisión literaria sobre el reconocimiento multimodal de emociones, mediante esto se identificó estudios correspondientes a las arquitecturas y conjuntos de datos más utilizados dentro del campo del Deep Learning, de igual forma se identificaron estudios correspondientes a las emociones, identificando factores de riesgo que contribuyan al tecnoestrés.

Para el desarrollo y entrenamiento del sistema se realizó la implementación de un modelo de red neuronal convolucional de dos capas para el reconocimiento de emociones, la aplicación de técnicas de Hold-out en un conjunto de datos unificado, que abarca los corpus RAVDESS, TESS, MESD y SAVEE, junto con el uso de métricas de evaluación como la matriz de confusión, ha validado la eficacia y la robustez del sistema propuesto.

La arquitectura propuesta de red neuronal convolucional de dos capas logró una precisión general del 93% aplicando técnicas de “Hold-out”, el promedio ponderado sobre las siete clases de emociones en el conjunto de datos de pruebas, logrando específicamente: precisión del 97% para la emoción de miedo, 97% para disgusto, 93% para neutral, 96% para felicidad, 86% para tristeza, 89% para sorpresa y 92% para enojo.

La red neuronal convolucional entrenada logro tener un rendimiento sólido en la mayoría de las categorías emocionales, especialmente en "miedo", "disgusto" y "felicidad". Sin embargo,

hay cierta variación en la precisión para cada emoción, siendo "tristeza" la que tiene la precisión más baja. Estos resultados son valiosos para evaluar la efectividad de la arquitectura en la clasificación de diferentes emociones y pueden ser utilizados para ajustar y mejorar el modelo si es necesario.

Se logró determinar correctamente varias arquitecturas de reconocimiento de aprendizaje profundo y certeras en el reconocimiento de emociones: una red pre-entrenada BlazeFace para el reconocimiento de caras, una red neuronal convolucional de dos capas para el reconocimiento vocal y una red neuronal convolucional de tres capas pre-entrenada de tres capas para el reconocimiento facial.

Se construyó un framework que permite el reconocimiento emocional bajo un método multimodal, usando técnicas de Deep Learning permitiendo detectar emociones en un individuo dentro de un entorno laboral no controlado, de esta manera, bajo el criterio de un experto en salud ocupacional y psicológica se interpretan e identifican patrones relacionados al tecnoestrés, este proceso busca agilizar y fortalecer el conocimiento psicológico entorno a un diagnóstico.

La aplicación de SCRUM como metodología ágil permitió que el proceso se desarrolle bajo comunicación de los desarrolladores y sus colaboradores al momento de receptar los videos, considerando a individuos enfocados en áreas administrativas y de sistemas quienes manejan y conocen sobre herramientas tecnológicas.

El proyecto fue desarrollado en un lenguaje de programación Python, siendo uno de los lenguajes de categoría alta para desarrolladores expertos, de esta manera se supo aprovechar la amplia gama de librerías al momento de realizar la programación enfocada en reconocimiento emocional.

Durante la aplicación del test a la muestra identificada se pudo observar niveles de satisfacción, incomodidad, curiosidad y sintomatología de estrés de esta manera, se logró

identificar emociones predominantes por cada uno de los individuos ayudando al profesional en salud ocupacional y psicológica a determinar de manera más eficiente el estado emocional en un área determinada de trabajo.

La muestra receptada demuestra que posee buenas estrategias de afrontamiento entorno al estrés laboral centradas o enfocadas en la resolución de conflictos a lo largo de la realización de sus actividades diarias.

Los niveles de tecnoestrés son evidentes en la muestra tomada, se puede identificar la presencia de sintomatología, generando un diagnóstico bajo niveles de estrés esperables acorde al trabajo que desarrollan diariamente.

Recomendaciones

Al usar SCRUM como metodología ágil es recomendable mantener reuniones de forma constante entre desarrolladores y colaboradores dando cumplimiento al cronograma de tareas establecido y seguimiento a las soluciones dadas dentro del entorno colaborativo.

Al momento de realizar el entrenamiento de los modelos predictivos de reconocimiento emocional, se recomienda validar continuamente su rendimiento con métricas de evaluación estandarizadas y validadas, mediante esto podemos lograr un porcentaje de precisión más elevado.

Se recomienda que para procesos de reconocimiento emocional se empleen otro tipo de emociones basadas en la rueda de emociones de Plutchik de esta manera se podrá contemplar patrones más precisos entorno al comportamiento humano durante sus actividades diarias.

Para futuras investigaciones se recomienda realizar este tipo de reconocimiento emocional enfocadas en un entorno laboral controlado, empleando buena iluminación y enfoque de manera frontal, de esta manera la captura de gesticulación sea favorable para el receptor de la muestra y para futuros diagnósticos en el área psicológica.

Se recomienda que durante el proceso de extracción de datos el moderador no sólo enfoque la parte fisiológica de un individuo ya que para determinar el comportamiento y

lenguaje no verbal requiere de mayor parte del cuerpo, esto se denomina inquietud psicomotora.

Es recomendable que el sistema permita realizar el reconocimiento emocional bajo la combinación de audio y video con una prueba validado, de esta manera se busca garantizar la fiabilidad de los resultados arrojados en torno a criterios de tecnoestrés.

Es recomendable que este tipo de proyectos sea dirigido a una población más extensa, ya que al estar dirigida en una muestra pequeña la generalidad de los datos influye en los resultados obtenidos bajando el nivel de confiabilidad de estos.

Bibliografía

Abadi, M., Isard, M., & Murray, D. G. (2017). A computational model for TensorFlow: An introduction. *Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, 1–7. https://doi.org/10.1145/3088525.3088527

Abdul, Z. Kh., & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, *10*, 122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444

Abdullah, S. M. S. A., Ameen, S. Y. A., M. Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal Emotion Recognition using Deep Learning. *Journal of Applied Science and Technology Trends*, *2*(02), 52–58. https://doi.org/10.38094/jastt20291

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). *Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild*. https://doi.org/10.48550/ARXIV.1906.02569

Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. En *Advances in Computers* (Vol. 117, pp. 339–368). Elsevier. https://doi.org/10.1016/bs.adcom.2019.09.007

Agarap, A. F. (2018). *Deep Learning using Rectified Linear Units (ReLU)*. https://doi.org/10.48550/ARXIV.1803.08375

Aguilar, A. R., Moreno, R. S., Miranda, L., & Ojeda, W. (2021). *Algoritmo Adam en la Inteligencia Artificial*.

Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. *Chaos, Solitons & Fractals*, *140*, 110120. https://doi.org/10.1016/j.chaos.2020.110120

Almabdy, S., & Elrefaei, L. (2019). Deep Convolutional Neural Network-Based Approaches for Face Recognition. *Applied Sciences*, *9*(20), 4397. https://doi.org/10.3390/app9204397

Ashmore, A., Deen, R., He, Y.-H., & Ovrut, B. A. (2022). Machine learning line bundle connections. *Physics Letters B*, *827*, 136972. https://doi.org/10.1016/j.physletb.2022.136972

Auquilla Vicuña, J. F., & Mora Alvarez, J. C. (2022). *Diseño de un sistema prototipo de diálogo persona-máquina basado en la arquitectura BERT*.

Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, *30*(5), 975–985. https://doi.org/10.1007/s00138-018-0960-9

Bahrami, M., Park, J., Liu, L., & Chen, W.-P. (2018). API Learning: Applying Machine Learning to Manage the Rise of API Economy. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 151–154. https://doi.org/10.1145/3184558.3186966

Baldassarri Santalucía, S. (2016). Computación afectiva: Tecnología y emociones para mejorar la experiencia del usuario. *Bit & Byte*, *2*.

Barrionuevo, C., Ierache, J. S., & Sattolo, I. I. (2020). *Reconocimiento de emociones a través de expresiones faciales con el empleo de aprendizaje supervisado aplicando regresión logística*. XXVI Congreso Argentino de Ciencias de la Computación (CACIC)(Modalidad virtual, 5 al 9 de octubre de 2020).

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2019). *BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs*. https://doi.org/10.48550/ARXIV.1907.05047

Cabada, R. Z., Estrada, M. L. B., & López, H. M. C. (2019). Reconocimiento multimodal de emociones orientadas al aprendizaje. *Research in Computing Science*, *148*(7), 153–165.

Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., & Sobieranski, A. C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, *582*, 593–617.

Carbajal Bacilio, K. A., & Suarez Mariscal, C. F. (2023). *Implementación de un dataset para la evaluación de modelos de análisis de sentimientos en la clasificación de tweets*.

Carneiro, T., Medeiros Da Nobrega, R. V., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767

Chamba, D. A. J., & Jiménez, L. C. (2022). Estudio comprensivo de la Transformada de Fourier Discreta para el análisis de señales digitales. *Revista Científica y Tecnológica UPSE*, *9*(1), 75–84.

Chanchí-Golondrino, G.-E., Hernández-Londoño, C.-E., & Ospina-Alarcón, M.-A. (2022). Aplicación de la computación afectiva en el análisis de la percepción de los asistentes a una feria de emprendimiento del SENA. *Revista Científica*, *44*(2), 215–227. https://doi.org/10.14483/23448350.18971

Chollet, F. (2021). *Deep learning with Python* (Second edition). Manning Publications.

Duville, M. M., Alonso-Valerdi, L. M., & Ibarra-Zarate, D. I. (2022). Neuronal and behavioral affective perceptions of human and naturalness-reduced emotional prosodies. *Frontiers in Computational Neuroscience*, *16*, 1022787. https://doi.org/10.3389/fncom.2022.1022787

Flórez, R., & Fernández, J. (2008). Las redes neuronales artificiales, fundamentos teoricos y aplicaciones practicas. *España, Netbiblo*.

Ghillaume, K. (2023). *Faster Whisper transcription with CTranslate2* [Software]. https://github.com/guillaumekln/faster-whisper/blob/master/README.md

Goldsborough, P. (2016). *A Tour of TensorFlow*. https://doi.org/10.48550/ARXIV.1610.01178

González, Y., Juárez, H., Rocha, O., Hernández, R., & Bermúdez, A. (2019). Evaluación comparativa de sistemas de reconocimiento de locutor basados en los algoritmos LPC,

CC y MFCC. *Memoria. Investigaciones en Ingenieria*, *17*. https://doi.org/10.36561/ING.17.6

Griera i Jiménez, O. (2022). *Multimodal emotion recognition via face and voice*.

Guarino, L. R., Feldman, L., & Roger, D. (2005). La diferencia de la sensibilidad emocional entre británicos y venezolanos. *Psicothema*, 639–644.

Helene Bischel, S. (2015). *El método de la entropía cruzada. Algunas aplicaciones*.

Hewage, N., & Meedeniya, D. (2022). *Machine Learning Operations: A Survey on MLOps Tool Support*. https://doi.org/10.48550/ARXIV.2202.10169

Hoang Thuan, N., Drechsler, A., & Antunes, P. (2019). Construction of Design Science Research Questions. *Communications of the Association for Information Systems*, 332–363. https://doi.org/10.17705/1CAIS.04420

Hron, M., & Obwegeser, N. (2022). Why and how is Scrum being adapted in practice: A systematic review. *Journal of Systems and Software*, *183*, 111110. https://doi.org/10.1016/j.jss.2021.111110

Ierache, J., Sattolo, I., & Chapperón, G. (2020). Framework multimodal emocional en el contexto de ambientes dinámicos. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *40*, 45–59.

Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209–221. https://doi.org/10.1016/j.inffus.2019.06.019

Kamencay, P., Benco, M., Mizdos, T., & Radil, R. (2017). A New Method for Face Recognition Using Convolutional Neural Network. *Advances in Electrical and Electronic Engineering*, *15*(4), 663–672. https://doi.org/10.15598/aeee.v15i4.2389

Khaireddin, Y., & Chen, Z. (2021). *Facial Emotion Recognition: State of the Art Performance on FER2013*. https://doi.org/10.48550/ARXIV.2105.03588

Kim, M., Lee, D., & Kim, K.-Y. (2015). System Architecture for Real-Time Face Detection on Analog Video Camera. *International Journal of Distributed Sensor Networks*, *11*(5), 251386. https://doi.org/10.1155/2015/251386

Ko, B. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, *18*(2), 401. https://doi.org/10.3390/s18020401

Labach, A., Salehinejad, H., & Valaee, S. (2019). *Survey of Dropout Methods for Deep Neural Networks*. https://doi.org/10.48550/ARXIV.1904.13310

Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Frontiers in Computer Science*, *2*, 14. https://doi.org/10.3389/fcomp.2020.00014

Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (2019). A fully trainable network with RNN-based pooling. *Neurocomputing*, *338*, 72–82. https://doi.org/10.1016/j.neucom.2019.02.004

Li, Y., Zhao, Z., Klejch, O., Bell, P., & Lai, C. (2023). *ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition*. https://doi.org/10.48550/ARXIV.2305.16065

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). *MediaPipe: A Framework for Building Perception Pipelines*. https://doi.org/10.48550/ARXIV.1906.08172

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., & Fernández-Martínez, F. (2021). Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors*, *21*(22), 7665. https://doi.org/10.3390/s21227665

Luo, S., Shi, Y., Chin, L. K., Hutchinson, P. E., Zhang, Y., Chierchia, G., Talbot, H., Jiang, X., Bourouina, T., & Liu, A.-Q. (2021). Machine-Learning-Assisted Intelligent Imaging Flow Cytometry: A Review. *Advanced Intelligent Systems*, *3*(11), 2100073. https://doi.org/10.1002/aisy.202100073

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: Review and insights. *Procedia Computer Science*, *175*, 689–694.

Mobile Computing, W. C. A. (2023). Retracted: Multimodal Emotion Recognition Algorithm for Artificial Intelligence Information System. *Wireless Communications and Mobile Computing*, *2023*, 1–1. https://doi.org/10.1155/2023/9761435

Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Computer Science*, *170*, 1168–1173. https://doi.org/10.1016/j.procs.2020.03.049

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of Artificial Neural Networks and Deep Learning. En O. A. Montesinos López, A. Montesinos López, & J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 379–425). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_10

Morcillo Vizuete, F. (2020). *Desarrollo de un sistema de reconocimiento facial utilizando Deep Learning con OpenCV*.

Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A., & Castillo, C. (2021). Análisis del Estado Actual de Procesamiento de Lenguaje Natural. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *E42*, 126–136.

Muhuri, P. S., Chatterjee, P., Yuan, X., Roy, K., & Esterline, A. (2020). Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks. *Information*, *11*(5), 243. https://doi.org/10.3390/info11050243

Musgrave, K., Belongie, S., & Lim, S.-N. (2020). *PyTorch Metric Learning*. https://doi.org/10.48550/ARXIV.2008.09164

Namatēvs, I. (2017). Deep Convolutional Neural Networks: Structure, Feature Extraction and Training. *Information Technology and Management Science*, *20*(1). https://doi.org/10.1515/itms-2017-0007

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Oostwal, E., Straat, M., & Biehl, M. (2021). Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. *Physica A: Statistical Mechanics and Its Applications*, *564*, 125517. https://doi.org/10.1016/j.physa.2020.125517

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). *Automatic differentiation in pytorch*.

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. https://doi.org/10.48550/ARXIV.2106.09462

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (TESS). *Scholars Portal Dataverse*, *1*, 2020.

Pineda Pertuz, C. M. (2022). *Aprendizaje automático y profundo en python*. Ra-Ma Editorial.

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, *7*, 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Quinaluiza Arias, A. I. (2018). *Interfaz de programación de aplicaciones para la generación automática de procedimientos almacenados en Mysql*.

Quiroz Martínez, M. Á., Granda Villon, G. A., Maldonado Cevallos, D. I., & Leyva Vázquez, M. Y. (2020). Análisis comparativo para seleccionar una herramienta de reconocimiento de emociones aplicando mapas de decisión difusos y TOPSIS. *Dilemas contemporáneos: Educación, Política y Valores*. https://doi.org/10.46377/dilemas.v8i1.2441

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision*. 28492–28518.

Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing Ltd.

Rouhiainen, L. (2020). Inteligencia Artificial 101 Cosas Que Debes Saber Hoy Sobre Nuestro Futuro Inteligencia Artificial. *Madrid: Alienta Editorial*.

Sachdeva, S. (2016). Scrum Methodology. *Int. J. Eng. Comput. Sci*, *5*(16792), 16792–16800.

Saks, E. (2019). *JavaScript Frameworks: Angular vs React vs Vue.*

Salanova, M., Cifre, E., & Martín, P. (1999). El proceso de 'Tecnoestrés'y estrategias para su prevención. *Prevención, Trabajo y Salud*, *1*, 18–28.

Sánchez, G. S., & Ramírez, M. E. D. (2013). Optimización de la utilidad esperada de un portafolio a partir del método de entropía cruzada. *Revista de Administración, Finanzas y Economía (Journal of Management, Finance and Economics)*, *7*(2), 83–100.

Saunshi, N., Gupta, A., & Hu, W. (2021). *A representation learning perspective on the importance of train-validation splitting in meta-learning*. 9333–9343.

Serengil, S. I., & Ozpinar, A. (2021). HyperExtended LightFace: A Facial Attribute Analysis Framework. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. *Multimodal Technologies and Interaction*, *6*(6), 47. https://doi.org/10.3390/mti6060047

Southwest Jiaotong University, China, Muhammad, I., Yan, Z., & Southwest Jiaotong University, China. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Universidad Nacional de Córdoba, Narambuena, L., Vaiman, M., & Pereno, G. L. (2016). Reconocimiento de Emociones Faciales en Adultos Mayores de la Ciudad de Córdoba. *Psykhe (Santiago)*, *25*(1), 1–13. https://doi.org/10.7764/psykhe.25.1.791

Velasquez, A. J. C., & Paz, D. C. B. (2020). El Tecnoestrés: Una consecuencia de la inclusión de las TIC en el trabajo. *Cienciamatria*, *6*(1), 295–314.

Villada, F., Muñoz, N., & García-Quintero, E. (2016). Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro. *Información Tecnológica*, *27*(5), 143–150. https://doi.org/10.4067/S0718-07642016000500016

Wang, W. (Ed.). (2011). *Machine Audition: Principles, Algorithms and Systems*. IGI Global. https://doi.org/10.4018/978-1-61520-919-4

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, *9*, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

Apéndices

Apéndice A Certificado – Psicóloga Clínica Betsy Morales

Apéndice B Test de Evaluación de Tecnoestrés