593 Digital Publisher
CEIT
ISSN 2588-0705

Detección de emociones en discursos utilizando machine learning

Emotion detection in speeches using machine learning

Mercedes Jamileth Miranda-Leon[1]
Universidad Técnica de Manabí - Ecuador
mmiranda9730@utm.edu.ec

Ramón Alfredo Toala-Dueñas[2]
Universidad Técnica de Manabí - Ecuador
ramon.toala@utm.edu.ec

doi.org/10.33386/593dp.2024.4.2367

V9-N4 (jul-ago) 2024, pp 72-101 | Recibido: 11 de febrero del 2024 - Aceptado: 23 de abril del 2024 (2 ronda rev.)

1 ORCID: https://orcid.org/0000-0003-4372-8221
2 ORCID: https://orcid.org/0000-0001-5397-9054

Cómo citar este artículo en norma APA:

Miranda-Leon, M., Toala-Dueñas, R.,(2024). Detección de emociones en discursos utilizando machine learning. 593 Digital Publisher CEIT, 9(4), 72-101, https://doi.org/10.33386/593dp.2024.4.2367

Descargar para Mendeley y Zotero

RESUMEN

En el contexto actual, donde las interacciones humanas se expanden en la era digital, la detección de emociones en discursos se establece como un área de investigación crucial. Este artículo se enfoca en detectar las emociones empleando técnicas de Machine Learning con el procesamiento de audio para discernir emociones en diversos discursos. La investigación subraya la influencia de las emociones en la comunicación y señala la falta de una teoría integral que abarque el espectro emocional completo. Desde la búsqueda en fuentes académicas hasta la implementación en Google Colab con herramientas como Pydub y Librosa, la metodología abarca todas las etapas. Se recopilan discursos de distintas categorías, etiquetados manualmente en emociones positivas, negativas y neutras. El procesamiento de datos implica la conversión a formato WAV, segmentación y etiquetado. Se implementa una Red Neuronal Convolucional (CNN) para la clasificación, con una precisión del 80.00% en el conjunto de prueba, respaldando la eficacia del modelo. El análisis incluye visualizaciones de la matriz de confusión y presentación de informes de clasificación. Las conclusiones destacan la viabilidad del ML y procesamiento de audio en la detección de emociones en discursos en español, resaltando la importancia del procesamiento de datos y sugiriendo mejoras para futuras investigaciones. Este trabajo se presenta como una contribución significativa al análisis emocional proporcionando un sólido marco para investigaciones posteriores.

Palabras claves: procesamiento de emociones por voz, machine learning, deep learning, extracción de características, clasificación de emociones.

ABSTRACT

In the current context, where human interactions expand in the digital age, emotion detection in speeches is established as a crucial research area. This article focuses on detecting emotions using Machine Learning techniques with audio processing to discern emotions in various speeches. The research highlights the influence of emotions on communication and points out the lack of a comprehensive theory that encompasses the full emotional spectrum. From searching in academic sources to implementation in Google Colab with tools like Pydub and Librosa, the methodology covers all stages. Speeches from different categories are collected, manually labeled in positive, negative, and neutral emotions. Data processing involves conversion to WAV format, segmentation, and labeling. A Convolutional Neural Network (CNN) is implemented for classification, with an accuracy of 80.00% on the test set, supporting the effectiveness of the model. The analysis includes visualizations of the confusion matrix and presentation of classification reports. The conclusions highlight the feasibility of ML and audio processing in emotion detection in Spanish speeches, emphasizing the importance of data processing and suggesting improvements for future research. This work is presented as a significant contribution to emotional analysis providing a solid framework for subsequent research.

Keywords: speech emotion processing, machine learning, deep learning, feature extraction, emotion classification.

Introducción

La comunicación humana es un proceso complejo que implica el intercambio de información verbal y no verbal. La comunicación no verbal desempeña un papel crucial en el proceso de comunicación humana, aunque la comunicación verbal transmite información explícita, la comunicación no verbal permite transmitir emociones, actitudes y estados de ánimo de manera más sutil y precisa, las expresiones faciales, el lenguaje corporal y el tono de voz pueden complementar o contradecir el mensaje verbal, influyendo en la interpretación y comprensión del mensaje. Además, la comunicación no verbal es universal y trasciende las barreras del lenguaje, permitiendo la comunicación efectiva incluso entre personas que hablan diferentes idiomas. (Trak, 2023)

En el contexto de la sociedad moderna, donde la interacción humano-computador (HCI) se ha vuelto de vital importancia, la detección de emociones en discursos se ha convertido en un área de investigación crucial, alimentada por la necesidad de comprender la complejidad de las interacciones humanas en línea y fuera de ella(Al-Dujaili & Ebrahimi-Moghadam, 2023).

En los últimos años, ha existido un creciente interés del desarrollo de métodos automáticos para detección de emociones en discursos, ya que el habla constituye una forma de comunicación compleja que transmite información en varios niveles además del contenido verbal(Rovetta, 2020). Este interés surge de la importancia de comprender y analizar las emociones expresadas en los discursos, ya que estas pueden influir en las interacciones humanas y en la toma de decisiones, los avances en el aprendizaje automático han permitido el desarrollo de métodos cada vez más precisos para esta tarea, lo cual ha impulsado la investigación en el campo de la detección automática de emociones en discursos.

Kerkeni et al. (2020), sugieren que el reconocimiento automático de emociones en el habla es desafiante debido a la diversidad de expresiones emocionales y los factores contextuales. Sin embargo, los avances en el aprendizaje automático han permitido métodos más precisos para identificar y categorizar las emociones en las conversaciones, determinando si son positivas, negativas o neutrales.

Esta área del análisis de texto se lleva a cabo mediante diversas herramientas y recursos, así como mediante distintos lenguajes de programación que permiten extraer información y comprender la polaridad de las emociones presentes en el texto(Martín De Diego et al, 1970).

El reconocimiento automático del habla y las emociones (ESR) normalmente consta de dos partes principales: extracción de características y clasificación. Sin embargo, la mayoría de los sistemas ESR se centran únicamente en la parte de extracción de características e ignoraban la parte de clasificación (Albadr et al., 2022).

La literatura acerca de este tema infiere que este campo de estudio especifico tiene gran potencial de avances en áreas como HCI y sistemas, además de mejorar la capacidad de las máquinas para comprender y responder a las emociones humanas tiene aplicaciones prácticas en diversas áreas. El potencial del tema y el avance tecnológico en progreso acerca del mismo fomentan la investigación acerca de la exploración y optimización de modelos específicamente para la detección de emociones en el habla para obtener algoritmos más eficientes y precisos, razón por la cual el objetivo de este trabajo es encontrar un método para la clasificación de emociones en audios que sea preciso y eficiente en base a investigaciones en la prueba de algoritmo para así obtener una propuesta.

Objetivos

General

Detectar las emociones en discursos, aplicando técnicas de Machine Learning.

Específicos

Digital Publisher
CEIT
ISSN 2588-0705

Realizar un estudio del arte respecto a la temática de estudio, de conceptos y modelos de algoritmos en machine learning.

Investigar las técnicas utilizadas para el reconocimiento de emociones utilizando algoritmos de machine learning.

Diseñar el data set a ser utilizado para la detección de emociones.

Realizar las pruebas correspondientes con la red neuronal.

Preguntas de investigación

RQ1: ¿Qué aspectos aborda el estudio del arte en el contexto de la detección de emociones en discursos utilizando machine learning?

RQ2: ¿Qué tipo de técnicas se investigaron para el reconocimiento de emociones en discursos mediante algoritmos de machine learning?

RQ3: ¿Cómo se diseñó el data set para la detección de emociones en discursos en el estudio mencionado?

RQ4: ¿Qué tipo de pruebas se llevaron a cabo con la red neuronal en el proceso de detección de emociones en discursos?

Metodología:

Revisión Bibliográfica

En esta primera parte, se hizo uso de la recopilación de información que se investigó con el tema detección de emociones en discursos, esto ayudo a comprender los conceptos clave y los modelos de algoritmos mayormente utilizados en este campo.

Búsqueda en Línea

Se utilizo bases de datos académicas, bibliotecas digitales y motores de búsqueda para encontrar artículos científicos y recursos relevantes sobre el tema, que fueron los siguientes.

Biblioteca Digital ACM

SpringerLink

Science direct

Recopilación y Preprocesamiento de Datos de Voz

Se recolectaron grabaciones de discursos de voz en diferentes contextos, como discursos políticos, graduaciones de estudiantes de bachillerato, motivación, conferencias de drogadicción, y días de las madres. Los audios se etiquetaron manualmente, el sistema de etiquetado es de tres categorías: positivo, negativo y neutro, se utilizó las siguientes técnicas:

Conversión de formato: Los datos se convirtieron al formato WAV.

Segmentación: Los datos se dividieron en segmentos de un minuto.

Arquitectura

En base a la literatura que se analizó se modeló una arquitectura para llevar a cabo el estudio la cual se plantea en cinco etapas que se visualiza en la Figura 1 además de ser explicadas a continuación:

Figura 1
Arquitectura

Recopilación de datos

El primer paso en la arquitectura fue la recolección de un conjunto de datos desde la plataforma web de videos YouTube donde se seleccionaron diferentes tipos de discursos de audios en diversos ámbitos en el cual se mencionan en recopilación y procesamiento de datos de voz. A su vez los diferentes audios seleccionados fueron puestos en el servicio de alojamiento de archivos de Google drive para

mayor facilidad de ser utilizados y optimizar el etiquetado.

Procesamiento de datos

En el siguiente paso fue realizado el procesamiento de los audios convirtiéndolos en formato WAV para su respectivo análisis, aquí además fueron divididos en segmentos de 1 minuto, y separados por el sexo del hablante, es decir, la voz femenina o masculina, esta clasificación fue alojada dentro de carpetas que a su vez fueron subdivididas en otras, donde se categorizaron por el etiquetado manual, que se planteó en el paso anterior, para así poder procesarlos de manera más óptima con el servicio de la misma empresa de Google Colab para realizar la respectiva extracción de sus características.

Selección de algoritmos de machine learning

El siguiente paso es la clasificación de las emociones en tres categorías: "positivo," "negativo" y "neutro,". A partir de la literatura analizada anteriormente se utilizó el algoritmo de aprendizaje profundo denominada Redes neuronales, ya que este algoritmo se adecua a los objetivos planteados para el presente trabajo sobre procesamiento de características vocales de los discursos.

Clasificación

En este componente se realizó clasificación donde se utilizó un modelo de aprendizaje automático para catalogar las emociones en el audio, específicamente el modelo que fue seleccionado y utilizado en este caso es una red neuronal convolucional (CNN), debido a que según en la literatura, han demostrado ser algoritmos de aprendizaje más destacados para el análisis de contenido demostrando además un rendimiento excepcional en tareas como segmentación, clasificación, detección y recuperación de imágenes.

Documentación y presentación

Finalmente prosiguió el proceso de documentación que precedió la presentación de todo el trabajo, en esta se contemplan la recopilación de todos datos obtenidos y utilizados, hasta la implementación de la red neuronal y los resultados obtenidos a partir de estas.

Además, se incluyeron métricas de evaluación que fueron manejadas, como fueron precisión, recall, F1-score y matriz de confusión, para cada emoción, utilizando gráficos y tablas etc.

Enfoques

La arquitectura planteada tendrá dos enfoques que ayudarán a la comprensión de los resultados obtenidos y permitirán un mejor análisis de los mismo, estos son:

Cualitativo:

El método cualitativo en investigación se enfoca en comprender fenómenos desde la perspectiva de los participantes, recolectando datos no numéricos como entrevistas y observaciones. En este estudio, el método cualitativo fue crucial para entender la complejidad de los datos utilizados en el entrenamiento de algoritmos de machine learning. Se emplearon técnicas cualitativas, como el análisis de documentos, para interpretar datos provenientes de textos, imágenes y audios descargados de YouTube (aproximadamente 631 archivos en formato wav). Estas estrategias no solo proporcionaron significado a los datos, sino que también contribuyeron a la evaluación del rendimiento de los algoritmos.

Cuantitativo:

Se empleó el método cuantitativo para evaluar el rendimiento de algoritmos en la detección de emociones en discursos. Se calcularon métricas como exactitud, precisión, sensibilidad y especificidad, utilizando un conjunto de datos de audios etiquetados manualmente. Estas métricas proporcionaron una evaluación objetiva del desempeño de los algoritmos de Machine Learning entrenados. Además, se aplicó la matriz de confusión para visualizar la distribución de las predicciones en

593 Digital Publisher CEIT ISSN 2588-0705

comparación con los valores reales, brindando una perspectiva detallada del rendimiento de los algoritmos en la clasificación de emociones en discursos. Este enfoque cuantitativo permitió una medición precisa y estadística del éxito de los algoritmos, respaldando así la validez y eficiencia del modelo propuesto en la detección de emociones en discursos.

Herramientas:

En las herramientas que se utilizaron en el trabajo fueron hardware, software, internet:

Software

YouTube

Se utilizo para la recopilación de datos que son los audios de diferentes discursos.

Conversor de video a WAV

Es software se utilizó para el procesamiento de datos, ayudo a convertir los audios a formato WAV y para dividirlos en segmentos.

Kaggle

Este software ayudo en la extracción de características del audio, como la energía, el espectrograma y la frecuencia fundamental, dando facilidad al código que se utilizó en el trabajo, así como para entrenar la red neuronal.

Google Colab

Es un servicio gratuito de nube alojado por Google para fomentar la investigación sobre Machine Learning e inteligencia Artificial, este permite escribir, correr y ejecutar código Python dentro de Google drive, en este será el ejecutado el algoritmo planteado.

Google Drive

Es un servicio de alojamiento y sincronización de archivos desarrollado por Google, se utilizó para ejecutar y almacenar los resultados a través del Google Colab, además de almacenar los audios de prueba que se usaran en el algoritmo.

Hardware

Se utilizo computadora una laptop personal, para ejecutar los softwares de recopilación de datos, procesamiento de datos, aprendizaje automático y evaluación, la misma que tenía la siguiente descripción:

-Procesador: Procesador de 64 bits, 4 núcleos.

-Memoria RAM: 16 GB.

-Almacenamiento: 1 TB.

-Sistema operativo: Windows 10.

Internet

Desde el internet se recopilo audios de discursos de diferentes discursos desde YouTube, y con el procesamiento de datos, el internet se utilizó para acceder a herramientas y bibliotecas de software para procesar los datos de audio.

Ayudo a acceder a herramientas web, que permitieron realizar tareas básicas de procesamiento de audio, como la conversión de formato, la división en segmentos y la extracción de características, así como acceso a librerías de Python que permitieron realizar tareas avanzadas de procesamiento de audio, como el análisis de espectrogramas y la extracción de características.

También se utilizó para la búsqueda de documentos de investigaciones importantes de otros investigadores para avanzar en el trabajo.

Revisión de literatura

Definiciones

Detección de emociones

La detección de emociones es la tarea de identificar las emociones que experimenta una persona, esto puede realizarse a través de una variedad de señales, incluyendo expresiones faciales, lenguaje corporal, habla y escritura.

El acto de hablar implica la comunicación verbal, incluyendo información lingüística para identificar palabras y expresiones. La voz, generada por la vibración de las cuerdas vocales, exhibe características únicas como timbre, intensidad y calidad sonora (Jahangir et al., 2021).

En el contexto de la detección de emociones en discursos, la tarea es identificar las emociones que expresa una persona a través de su voz, esto se realiza analizando las características acústicas del habla, como la prosodia, la frecuencia fundamental y la intensidad. Esta tarea es importante para una variedad de aplicaciones (Kerkeni et al., 2020), incluyendo áreas como la atención al cliente, educación, o salud, también puede ayudar en la construcción de sistemas inclusivos y mejorar la accesibilidad de las interfaces.

Machine Learning

El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender a partir de datos y mejorar su rendimiento sin intervención humana explícita (Kavitha et al., 2022). Los algoritmos de aprendizaje automático se pueden utilizar para realizar una variedad de tareas, que incluyen la detección de emociones, el reconocimiento del habla y la clasificación de imágenes.

En el contexto de la detección de emociones, el aprendizaje automático se puede utilizar para desarrollar modelos que puedan identificar las emociones a partir de las características acústicas del habla, siendo fundamental para el progreso en la detección de emociones en discursos.

Reconocimiento Automático del Habla y las Emociones (ESR)

Las emociones desempeñan un papel esencial en la comunicación humana, manifestándose a través de expresiones faciales, gestos, posturas, tono de voz, selección de palabras, respiración y temperatura corporal. Estos elementos emocionales pueden modificar significativamente el mensaje transmitido, siendo la forma de comunicación a veces más relevante que el contenido. Aunque son fundamentales, aún carecemos de una teoría integral que pueda describir completamente el espectro emocional (Martín De Diego et al., 1970).

El Reconocimiento Automático del habla y las emociones (ESR) se basa en el análisis de la voz y las expresiones faciales para comprender y procesar la información emocional transmitida por los individuos, este campo de investigación se ocupa del desarrollo de sistemas que reconocen el habla y las emociones (Zhang et al., 2018). Este campo ha ganado interés en los últimos años debido a su potencial aplicación en una variedad de áreas.

ESR utilizan algoritmos de reconocimiento de voz para convertir las palabras habladas en texto o comandos entendibles por las máquinas. Lo que implica el uso de software o dispositivos especializados capaces de capturar, procesar y responder a los datos de voz. Estas herramientas son cada vez más utilizadas en diferentes campos, incluyendo la tecnología de asistentes virtuales y la automatización del hogar (Schuller et al., 2019).

Tabla 1

Clasificación de emociones

Clasificación	Descripción
Emociones Positivas	Las emociones positivas se destacan por su valencia emocional positiva, induciendo sensaciones de bienestar, alegría y satisfacción.
Emociones Negativas	Las emociones negativas presentan una valencia emocional negativa, asociada a sensaciones de malestar, angustia o insatisfacción.
Emociones Neutras	Las emociones neutras carecen de una valencia emocional clara y no provocan un impacto emocional significativo.

Según Russell (1980), uno de los modelos dimensionales más reconocidos en dos dimensiones es el modelo circunflejo afectivo, sus seguidores afirman que cada experiencia afectiva resulta de una combinación lineal de dos dimensiones, lo que produce una emoción específica, como se ilustra en la Figura 2.

Digital Publisher
CEIT
ISSN 2588-0705

Figura 2
Modelo circunflejo afectivo (Russell, 1980).

Trabajos Relacionados

En esta investigación se realizó búsquedas de artículos en base al tema de investigación en diversos estudios realizados por instituciones educativas tanto nacionales como internacionales, así como en artículos científicos de revistas especializadas. Estas fuentes nos permitirán realizar una exhaustiva revisión del conocimiento actual, consultar diversas fuentes bibliográficas y recopilar información relacionada con la aplicación de técnicas de aprendizaje automático, conceptos y modelos de machine learning, para identificar los estados emocionales de las personas a través de su voz, este enfoque se aplicará en el ámbito de discursos de la Política, Graduación, Motivación, Drogadicción y Día de las Madres. A continuación, se presentará investigaciones realizadas por diferentes autores sobre plataformas utilizadas para analizar emociones.

En esta investigación de Guerrón Pantoja (2023) destaca el desarrollo de un sistema de reconocimiento de emociones a través de la voz utilizando técnicas de aprendizaje profundo. Este campo de estudio se encuentra en constante evolución y tiene aplicaciones prácticas en diversos sectores, como la tecnología de asistentes virtuales, la atención médica y la psicología.

Este enfoque se centra en analizar las características acústicas de la voz para identificar patrones asociados con diferentes estados emocionales. Para lograrlo, se extraen características de la señal de voz, como el tono, la entonación y el ritmo, y se analizan utilizando algoritmos de aprendizaje automático.

Se ha demostrado que las técnicas de aprendizaje profundo son efectivas en la detección de emociones como la alegría, la tristeza, la ira y el miedo. Estas técnicas permiten que el sistema mejore su precisión a medida que se le proporciona más datos, lo que lo hace particularmente útil en situaciones que requieren una detección precisa y en tiempo real de las emociones, como en asistentes virtuales con capacidad de respuesta emocional o en el análisis de emociones en interacciones sociales mediadas por computadora.

La aplicación de técnicas de aprendizaje profundo ha demostrado ser eficaz en la detección de emociones como la alegría, la tristeza, la ira y el miedo, entre otras. Estas técnicas permiten que el sistema mejore su precisión a medida que se le alimenta con más datos, lo que lo hace particularmente útil en situaciones que requieren una detección precisa y en tiempo real de las emociones, como en asistentes virtuales con capacidad de respuesta emocional o en el análisis de emociones en interacciones sociales mediadas por computadora (Guerrón Pantoja, 2023).

La investigación de Padilla (2022) sobre la voz como reacción emocional y la prosodia, se profundiza en el reconocimiento de emociones a través del análisis de la prosodia vocal. Esta área de investigación se centra en cómo las características vocales, como el tono, la entonación y el ritmo, pueden revelar información sobre el estado emocional de una persona.

El reconocimiento de emociones basado en la prosodia vocal es un campo importante en la inteligencia artificial y la computación afectiva. Se utiliza en aplicaciones como asistentes virtuales con capacidad de respuesta emocional, análisis de emociones en llamadas de servicio al cliente y detección de emociones en interacciones sociales mediadas por computadora.

Las tecnologías utilizadas en el reconocimiento de emociones basado en la prosodia vocal suelen implicar el uso de algoritmos de aprendizaje automático y procesamiento de señales de voz. Estos algoritmos pueden entrenarse para identificar patrones en la prosodia vocal que correspondan a diferentes emociones, lo que permite a las máquinas interpretar y responder de manera más efectiva a las señales emocionales en el habla humana.

En resumen, el reconocimiento de emociones basado en la prosodia vocal es un área de investigación emocionante y en evolución que tiene aplicaciones prácticas en una variedad de campos, desde la tecnología de asistentes virtuales hasta la atención médica y la psicología.

Principio del formulario

El estudio de Xu et al. (2020) se centra en el análisis del estado cognitivo del aprendizaje sensible a las emociones en entornos educativos. Para lograr esto, proponen un enfoque basado en el aprendizaje profundo para analizar expresiones faciales, utilizando información ordinal para mejorar la precisión en la identificación de estados emocionales. Utilizan una red neuronal convolucional profunda para extraer características emocionales de las expresiones faciales de los estudiantes. Los resultados sugieren que este enfoque puede ser una herramienta efectiva para comprender el estado emocional de los estudiantes en contextos educativos, lo que podría tener implicaciones significativas en el diseño de estrategias de enseñanza más personalizadas y efectivas.

Según García-Ancira (2020), los estudiantes con un alto nivel de inteligencia emocional tienden a obtener mejores calificaciones en comparación con aquellos con un nivel más bajo de inteligencia emocional. Además, se observó que estos estudiantes son capaces de manejar de manera más efectiva el estrés y las presiones académicas, lo que les permite mantener un equilibrio saludable entre sus responsabilidades académicas y su bienestar emocional. El análisis también mostró que los estudiantes con habilidades emocionales desarrolladas tienen una mayor motivación intrínseca para el aprendizaje, lo que se traduce en una mayor persistencia y compromiso con sus estudios. Estos estudiantes también tienden a tener una mayor autoestima y confianza en sí mismos, lo que les ayuda a enfrentar los desafíos académicos con una actitud positiva y proactiva.

En su estudio sobre el reconocimiento de emociones en el habla, Zhang et al. (2021) emplearon métodos de aprendizaje profundo, incluyendo redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN), redes neuronales convolucionales 1D (CNN1D) y redes neuronales recurrentes bidireccionales (Bi-RNN). Estos métodos fueron utilizados para analizar características del habla que reflejan diferentes emociones, como enojo, alegría, tristeza, entre otras.

Para la detección de emociones, los investigadores aplicaron estas técnicas a conjuntos de datos de voz que contenían expresiones emocionales. A través del entrenamiento de los modelos con estos datos, lograron detectar y clasificar las emociones presentes en el habla con un alto grado de precisión. Este enfoque permitió identificar patrones en el habla asociados con diferentes emociones, lo que contribuye al desarrollo de sistemas más efectivos para el reconocimiento y la interpretación de las emociones humanas en contextos de interacción humano-máquina.

En la investigación de Hernández Tamayo et al. (2020), se utilizaron varios métodos de clasificación, como Naïve Bayes, Perceptrón Multicapa, Máquinas de Vectores de Soporte y

Bosque Aleatorio, para identificar emociones a través de la voz en el contexto del español hablado en México. Para este estudio, se emplearon dos conjuntos de datos, conocidos como corpus, que contenían expresiones emocionales: Emo_voz.mx1 y EmoWisconsin. Estos conjuntos representaban emociones inducidas y actuadas, respectivamente.

593 Digital Publisher CEIT ISSN 2588-0705

En el caso de Emo_voz.mx1, constaba de tres conjuntos de datos de voz, cada uno conteniendo 40 palabras seleccionadas de la lista Swadesh en español, 40 oraciones, párrafos con un promedio de 450 palabras y un poema de 94 palabras. Las emociones capturadas fueron enojo, disgusto, miedo, alegría, tristeza, sorpresa y neutral. Por su parte, EmoWisconsin trabajó con siete emociones: molesto, seguro, inseguro, motivado, nervioso, neutral e indeterminado, utilizando un grupo de 28 niñas y 17 niños con edades entre 7 y 13 años. Las características acústicas empleadas incluyeron MFCCs, Tasa de Cruce 0 (ZCR), energía, centroide espectral, dispersión espectral, entre otras. En los resultados, se destacó que el mejor rendimiento se logró mediante un algoritmo basado en Máquinas de Vectores de Soporte en comparación con los otros métodos de clasificación evaluados.

En otro estudio realizado por Basmmi et al. (2020), se señala la dificultad de determinar y comparar la eficacia de distintos servicios web de acceso gratuito para el análisis de sentimientos, dada la amplia gama de opciones disponibles. Para recolectar datos, emplearon RStudio, un software de análisis de datos de código abierto, y extrajeron información a través de la API de Twitter. Después de comparar cuatro servicios web, entre los cuales se incluía MonkeyLearn, destacan que esta última plataforma obtuvo los resultados más sobresalientes en comparación con las demás, debido a su alto nivel de calidad en los procesos que ofrece. Concluyen mencionando que esta plataforma resulta apropiada para otros investigadores, ya que puede ser beneficiosa en el análisis de temas relacionados con el análisis de sentimientos.

En la investigación de Kurniawan et al. (2020), se destaca la amplia variedad de programas disponibles para llevar a cabo la minería de texto, con especial énfasis en la popularidad de RapidMiner como una destacada solución de minería de datos de código abierto en todo el mundo. Su estudio se centró en analizar y determinar la polaridad de un conjunto específico de datos. En su enfoque metodológico, emplearon un proceso de cinco etapas que abarcaba desde el procesamiento del texto hasta la extracción de la polaridad de las opiniones, categorizándolas como positivas o negativas. Los resultados de la investigación demostraron una confiabilidad del 82% en la aplicación durante el proceso de análisis de sentimientos, lo que, según ellos, indica un alto nivel de efectividad en la minería de texto llevada a cabo en la aplicación.

El análisis de sentimientos se ha convertido en un área de investigación fundamental en el contexto de la creciente cantidad de contenido generado por usuarios en la web. En este sentido, el estudio de Ligthart et al. (2021) proporciona una visión detallada sobre los avances en este campo, destacando la importancia de los algoritmos de aprendizaje profundo en el análisis de sentimientos.

Ligthart et al. (2021) observaron un notable incremento de contenido generado por usuarios en la web, impulsado por la avanzada digitalización, que proporciona una amplia gama de opiniones sobre diversos temas, el propósito central de su investigación fue realizar una revisión sistemática de la literatura y un estudio de mapeo en el ámbito del análisis de sentimientos. Su enfoque se centró en identificar aspectos generales, tipos de conjuntos de datos y los algoritmos empleados en el análisis de sentimientos. Los resultados del estudio resaltan la relevancia y la utilidad del análisis de sentimientos como una herramienta significativa y oportuna para la sociedad en general. Según su análisis, se constató que los algoritmos de aprendizaje profundo son los más ampliamente utilizados en el proceso de análisis de sentimientos.

Por último, Senthilkumar, N. (2022) presenta un nuevo algoritmo para extraer características de archivos de audio utilizando una Red Neuronal Convolucional (CNN) Resnet101. Este enfoque involucra tres etapas: En el primer bloque, el audio se divide en segmentos y se identifican las diferencias entre ellos. Se emplea el algoritmo de clustering k-means para agrupar los segmentos de manera eficiente. La estrategia Radial Base Function (RBF) se utiliza para seleccionar los segmentos vecinos más similares en la secuencia de clusters. En el segundo

bloque, la secuencia de espectrogramas de audio se introduce en la red Resnet101 para aprender características distintivas. Las salidas de la red se normalizan después. En el tercer bloque, los datos pasan por una red neuronal bidireccional de tipo LSTM (BiLTSM).

El algoritmo se probó en varios conjuntos de datos, incluidos EMO-DB, Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS) e Interactive Emotional Dyadic Motion Capture (IEMOCAP), obteniendo tasas de precisión del 85.57%, 77.02% y 72.25%, respectivamente. Se observa un aumento del 0.92% en la tasa de acierto en comparación con investigaciones previas en el conjunto de datos compartido, Emo-DB. En resumen, el método propuesto se considera innovador, logrando mejoras en la precisión y reduciendo la complejidad computacional.

A continuación, se muestra el cuadro comparativo de las investigaciones mencionadas en función de sus enfoques, métodos y resultados:

Ver Tabla 2.

Este cuadro resalta los enfoques principales de cada investigación, los métodos destacados utilizados y los resultados notables logrados en cada caso, estos estudios abordan la detección de emociones en discursos desde diversas perspectivas y utilizando una variedad de técnicas, lo que demuestra la diversidad de enfoques en este campo de investigación.

Investigación

Técnicas De Algoritmos De Machine Learning

En el mundo actual, los algoritmos de machine learning desempeñan un papel fundamental al enfrentarse a grandes y complejos conjuntos de datos, su eficiencia permite a las máquinas procesar información y extraer conocimientos valiosos. Lo que otorga a las organizaciones una ventaja competitiva al identificar patrones, realizar predicciones precisas y aplicarse en diversas áreas como el análisis de datos, la toma de decisiones y la personalización de recomendaciones, su existencia es crucial para los avances actuales. (Guzmán Moyano, 2023).

El aprendizaje automático es un campo amplio con una gran variedad de algoritmos, además las técnicas de algoritmos de aprendizaje automático pueden clasificarse en dos tipos principales:

Técnicas de aprendizaje supervisado

El concepto de aprendizaje supervisado se refiere a una técnica de la inteligencia artificial en la cual se utiliza un conjunto de datos etiquetados para entrenar un modelo. Este conjunto de datos etiquetados consiste en ejemplos de datos de entrada y sus correspondientes datos de salida deseados. El aprendizaje supervisado usa datos etiquetados para entrenar modelos, buscando una función que relacione la entrada con la salida, se basa en la existencia de una relación aprendida a partir de ejemplos etiquetados (Carvajal Jaramillo, 2022).

Las técnicas de aprendizaje supervisado se pueden clasificar en dos categorías principales, clasificación y regresión, la clasificación se utiliza para asignar una categoría a una entrada, mientras que la regresión se utiliza para predecir un valor continuo a partir de una entrada.

Técnicas de aprendizaje no supervisado

Estas técnicas se utilizan cuando no se dispone de un conjunto de datos etiquetados. El algoritmo de aprendizaje no supervisado utiliza los datos de entrada para aprender una estructura o patrones en los datos.

Las técnicas de aprendizaje no supervisado se pueden dividir en tres categorías principales: agrupación, reducción de dimensionalidad y detección de anomalías.

- Agrupación: Esta categoría se utiliza para agrupar datos similares.

- Reducción de dimensionalidad: Esta categoría se utiliza para reducir el número de dimensiones de un conjunto de datos.

593 Digital Publisher CEIT ISSN 2588-0705

Tabla 2

Cuadro comparativo de las investigaciones

Investigación	Enfoque y Objetivo	Métodos Destacados	Resultados Notables
Guerrón Pantoja (2023)	Desarrollo de un sistema de reconocimiento de emociones a través de la voz utilizando técnicas de aprendizaje profundo.	Análisis de características acústicas de la voz para identificar patrones emocionales.	Eficacia demostrada en la detección de emociones como alegría, tristeza, ira y miedo. Mejora de la precisión con más datos. Aplicaciones prácticas en tecnología de asistentes virtuales, atención médica y psicología.
(Padilla (2022)	La voz como reacción emocional y la prosodia	Análisis de características vocales como tono, entonación y ritmo para inferir el estado emocional. Profundizar en el reconocimiento de emociones mediante el análisis de la prosodia vocal.	Utilización de algoritmos de aprendizaje automático y procesamiento de señales de voz para identificar patrones asociados con diferentes emociones y mejorar la respuesta a las señales emocionales en el habla humana.
Xu et al. (2020)	Análisis de las expresiones faciales en entornos educativos utilizando técnicas de aprendizaje profundo para comprender el estado emocional de los estudiantes. Principio del formulario	Utilización de una red neuronal convolucional profunda para extraer características emocionales de las expresiones faciales de los estudiantes.	Los resultados sugieren que este enfoque puede ser efectivo para comprender el estado emocional de los estudiantes en contextos educativos, lo que podría tener implicaciones significativas en el diseño de estrategias de enseñanza más personalizadas y efectivas.
García-Ancira (2020)	Evaluar la relación entre la inteligencia emocional y el rendimiento académico, así como su impacto en el manejo del estrés y las presiones académicas.	Análisis de correlación entre inteligencia emocional y calificaciones, evaluación de la capacidad de manejo del estrés mediante cuestionarios y entrevistas.	Los estudiantes con un alto nivel de inteligencia emocional tienden a obtener mejores calificaciones y a manejar el estrés académico de manera más efectiva. Además, tienen una mayor motivación intrínseca para el aprendizaje, mayor persistencia y autoestima. Estos hallazgos resaltan la importancia de la inteligencia emocional en el éxito académico y el bienestar emocional de los estudiantes universitarios.
Zhang et al. (2021)	Reconocimiento de Emociones en el Habla utilizando aprendizaje profundo	Análisis de redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN), redes neuronales convolucionales 1D (CNN1D), y redes neuronales recurrentes bidireccionales (Bi-RNN)	Resumen de avances recientes y desafíos en el campo, identificación de técnicas prometedoras y áreas para futuras investigaciones
Hernandez et al. (2020)	Identificación de Emociones en el Habla	Naïve Bayes, Perceptrón Multicapa, Máquinas de	SVM superó a otros métodos de clasificación.
Basmmi et al. (2020)	Evaluación de Servicios Web de Análisis de Sentimientos	Comparación de servicios web de análisis de sentimientos	MonkeyLearn obtuvo los mejores resultados.
Kurniawan et al. (2020)	Minería de Texto y Análisis de Sentimientos	Uso de RapidMiner y análisis de polaridad de opiniones	Confianza del 82% en la aplicación de RapidMiner.
Ligthart et al. (2021)	Análisis de Sentimientos	Revisión sistemática de la literatura y mapeo en el	Uso generalizado de algoritmos de aprendizaje profundo.
Senthilkumar et al. (2022)	Extracción de Características de Audio	Red Neuronal Convolucional (CNN) Resnet101, clustering	Mejora en la precisión y reducción de complejidad.

- Detección de anomalías: Esta categoría se utiliza para identificar datos que son inusuales o atípicos.

En el ámbito del Procesamiento del Lenguaje Natural (NLP), el Análisis de Sentimiento se destaca al emplear máquinas para procesar textos y asignarles clasificaciones comprensibles. Este enfoque implica el uso de algoritmos de procesamiento de lenguaje que extraen características, como la frecuencia de las palabras, y se apoya en algoritmos de aprendizaje automático supervisado. Estos últimos son entrenados inicialmente con datos clasificados por humanos, permitiendo a las máquinas aprender patrones y contextos para posteriormente realizar análisis de sentimiento de manera automatizada (S. & K.V., 2020).

Los algoritmos de Machine Learning están en constante evolución, adaptándose y mejorando su rendimiento a medida que adquieren más información. Se emplean en diversas tareas como clasificación, regresión y clustering. Las técnicas de aprendizaje profundo, especialmente las redes neuronales artificiales, han destacado en los últimos años al aprender patrones complejos en los datos, la combinación de características de voz y lenguaje ha demostrado un rendimiento superior en comparación con el uso de una sola fuente de información. Estos algoritmos pueden aplicarse al reconocimiento de emociones al extraer características de expresiones faciales, tono de voz o patrones de escritura, entrenando modelos para asignar emociones a conjuntos específicos de características.

Las técnicas de reconocimiento de emociones utilizando algoritmos de machine learning se pueden clasificar en dos categorías principales:

Técnicas basadas en características

Estas técnicas extraen características de los datos de entrada, como la expresión facial, el tono de voz o los patrones de escritura, y luego utilizan un algoritmo de clasificación para asignar una emoción a cada conjunto de características.

Técnicas de aprendizaje profundo

Estas técnicas utilizan redes neuronales artificiales para aprender directamente la relación entre los datos de entrada y las emociones.

Técnicas basadas en características

Las técnicas basadas en características son las más utilizadas para el reconocimiento de emociones, extraen características de los datos de entrada como la expresión facial, el tono de voz o los patrones de escritura, y luego utilizan un algoritmo de clasificación para asignar una emoción a cada conjunto de características.

Algunos de los algoritmos de clasificación más utilizados para el reconocimiento de emociones incluyen:

Clasificadores lineales:

Estos clasificadores utilizan funciones lineales para asignar una emoción a cada conjunto de características, los clasificadores lineales se utilizan en una amplia gama de aplicaciones, como:

Clasificación de imágenes

Los clasificadores lineales se utilizan para clasificar imágenes en diferentes categorías, como gatos, perros, personas, etc.

Clasificación de texto

Los clasificadores lineales se utilizan para clasificar texto en diferentes categorías, como noticias, publicidad, etc.

Clasificación de audio

Los clasificadores lineales se utilizan para clasificar audio en diferentes categorías, como música, voz, etc.

Clasificadores no lineales

Estos clasificadores utilizan funciones no lineales para asignar una emoción a cada conjunto de características, existen otros tipos de clasificadores no lineales como:

Digital Publisher CEIT ISSN 2588-0705

Redes neuronales:

Las redes neuronales son un tipo de clasificador no lineal que se utiliza para aprender patrones complejos en datos.

Árboles de decisión:

Los árboles de decisión son un tipo de clasificador no lineal que se utiliza para tomar decisiones en función de un conjunto de reglas.

Métodos de agrupación:

Los métodos de agrupación se utilizan para agrupar datos en función de su similitud.

Clasificadores basados en árboles:

Estos clasificadores construyen un árbol de decisión para asignar una emoción a cada conjunto de características, los clasificadores basados en árboles son fáciles de entender y de interpretar (Alourani et al., 2019). También son muy eficientes en términos de tiempo de cálculo, sin embargo, los clasificadores basados en árboles pueden ser sensibles al ruido en los datos.

Existen otros tipos de clasificadores basados en árboles:

Árboles de decisión CART:

Los árboles de decisión CART son una técnica de aprendizaje supervisado utilizada para clasificar datos. Se dividen en subconjuntos, creando nodos terminales, mediante la subdivisión sucesiva del conjunto de datos de entrenamiento. Este proceso se repite hasta alcanzar un número predeterminado de nodos terminales, proporcionando un modelo efectivo para clasificación.

Random Forest:

Random Forest es un método de aprendizaje automático que utiliza un conjunto de árboles de decisión para clasificar datos, funciona construyendo un conjunto de árboles de decisión aleatorios, cada árbol de decisión se construye utilizando una muestra aleatoria de los datos de entrenamiento y un subconjunto aleatorio de las características.

Gradient Boosted Trees:

Gradient Boosted Trees es un método de aprendizaje automático que utiliza un conjunto de árboles de decisión para clasificar datos.

Para entrenar un GBT, se sigue el siguiente procedimiento:

(1) Inicializar el modelo: Se inicializa el modelo con un árbol de decisión débil.

(2) Calcular el gradiente: Se calcula el gradiente de la función de pérdida con respecto a las predicciones del modelo actual.

(3) Ajustar un árbol de decisión débil al gradiente: Se ajusta un árbol de decisión débil al gradiente.

(4) Actualizar el modelo: Se actualiza el modelo sumando el árbol de decisión débil ajustado al modelo actual.

(5) Repetir los pasos 2-4: Se repiten los pasos 2-4 hasta que se alcance un número predefinido de árboles o hasta que el modelo converja.

Una vez que el modelo está entrenado, se puede utilizar para predecir o clasificar nuevos datos.

Técnicas De Aprendizaje Profundo:

Las técnicas de aprendizaje profundo son un tipo de aprendizaje automático que utiliza redes neuronales artificiales para aprender patrones complejos en los datos, y han demostrado ser muy efectivas para el reconocimiento de emociones, ya que son capaces de aprender relaciones no lineales entre los datos de entrada y las emociones.

Algunos de los tipos de redes neuronales artificiales más utilizadas para el reconocimiento de emociones incluyen:

Redes neuronales convolucionales:

Están diseñadas para aprender patrones en datos de imagen, como las expresiones faciales, una red neuronal convolucional es un tipo de red neuronal artificial se componen de capas de convolución, que son filtros que se aplican a la imagen para identificar características importantes.

Estas son muy efectivas para el reconocimiento de objetos en imágenes ayudan en una amplia gama de aplicaciones, como el reconocimiento facial, el reconocimiento de imágenes médicas y el procesamiento de imágenes de satélite.

Redes neuronales recurrentes:

El aprendizaje de patrones en datos secuenciales es fundamental en campos como el análisis de discurso y el reconocimiento de voz. Las redes neuronales recurrentes (RNN) permiten analizar el contexto y las relaciones entre los elementos de una secuencia, mejorando la comprensión y el rendimiento en diversas aplicaciones (Figueroa Sacoto, 2021).

Las redes neuronales recurrentes son muy efectivas para el reconocimiento de patrones en datos secuenciales, como el discurso, el texto y el código. Se utilizan en una amplia gama de aplicaciones, como el reconocimiento de voz, la traducción automática y el procesamiento del lenguaje natural.

Redes neuronales híbridas:

Pueden combinar diferentes tipos de redes neuronales para mejorar el rendimiento como las:

-Redes neuronales convolucionales-recurrentes:

Combinan redes neuronales convolucionales y las redes neuronales recurrentes, pueden utilizarse para tareas como el reconocimiento de objetos en imágenes y el reconocimiento de voz.

-Redes neuronales recurrentes-transformer:

Pueden utilizarse para tareas como el procesamiento del lenguaje natural y la traducción automática.

-Redes neuronales convolucionales-transformer:

Se utilizan para tareas como el reconocimiento de imágenes y el reconocimiento de objetos.

Aplicaciones de las redes neuronales híbridas

Las redes neuronales híbridas son versátiles y se aplican en diversas áreas, como el reconocimiento de imágenes para objetos, caras y escenas, estos también se utilizan para el reconocimiento y procesamiento de voz, incluyendo traducción automática y transcripción de audio. En el procesamiento del lenguaje natural, contribuyen a tareas como traducción automática, resumen de texto y clasificación. Estas redes representan una herramienta poderosa para mejorar el rendimiento en diversas aplicaciones.

Modelos De Algoritmos De Machine Learning:

El tema de Tipos de Aprendizaje en machine learning se centra en la comprensión de las distintas formas en que las máquinas pueden aprender de los datos, los principales tipos de aprendizaje son:

Regresión logística:

La regresión logística se emplea para la catalogación binaria de los datos, siendo una técnica esencial en el aprendizaje automático (Panesar et al., 2019). En el contexto del machine learning, las técnicas de clasificación son fundamentales, la regresión logística es una de ellas. Esta técnica, es una de las más simples y ampliamente utilizadas para la clasificar variables (Fernandes et al., 2020) the most appropriate statistical technique to deal with dichotomous dependent variables. Materials and Methods: we estimate the effect of corruption scandals on the chance of reelection of candidates running for the Brazilian Chamber of Deputies using data from Castro and Nunes

Digital Publisher CEIT ISSN 2588-0705

(2014, se emplea para predecir el resultado de una variable definida, al proporcionar una estimación de la probabilidad de un resultado específico, la regresión logística permite tomar decisiones informadas en diversos campos como la medicina, la economía y la investigación de mercado. Además, al comprender la importancia de la regresión logística, los profesionales del aprendizaje automático pueden utilizarla como base para desarrollar técnicas más complejas y avanzadas en el campo de la clasificación de datos (Cordero et al., 2023).

Redes Neuronales Recurrentes (RNN):

Las Redes Neuronales Recurrentes (RNN) son modelos eficaces para datos secuenciales, destacándose en tareas como reconocimiento de voz, clasificación de texto, predicción de palabras, análisis de sentimientos, correlación frase-imagen y modelado de lenguaje (Manchev & Spratling, 2020). Al retener información sobre eventos anteriores, se destacan en el procesamiento secuencial, convirtiéndose en modelos potentes y versátiles para abordar problemas de aprendizaje que involucran datos en secuencia (Nasir et al., 2021).

Redes neuronales de Memoria a Corto y Largo Plazo (LSTM):

Las redes neuronales de Memoria a Corto y Largo Plazo (LSTM) han ganado considerable atención debido a su eficacia en diversas aplicaciones prácticas, dentro de las Redes Neuronales Recurrentes (RNN), las LSTM destacan por su capacidad para capturar relaciones temporales de largo plazo sin enfrentar desafíos de optimización. Han demostrado éxito en una amplia gama de problemas, como reconocimiento de escritura, modelado de lenguaje, traducción, análisis de video y procesamiento de señales de audio, entre otros. Su versatilidad las posiciona como una herramienta fundamental en múltiples contextos (Sherstinsky, 2020).

Bidirectional Long Short-Term Memory (BI LSTM):

Las redes neuronales recurrentes (RNN) convencionales utilizan solo el contexto previo, mientras que las Long Short-Term Memory (LSTM) y otras RNN solo aprenden de información pasada en secuencias temporales. Las Bidirectional LSTMs (BI LSTM) superan esta limitación al constar de dos partes: una red LSTM estándar y otra en sentido inverso, capturando información contextual de secuencias temporales de manera más completa. Este enfoque ha demostrado eficacia en diversas aplicaciones, como procesamiento de lenguaje natural, reconocimiento de voz y análisis de audio, al capturar patrones y relaciones temporales de manera más exhaustiva que las RNN convencionales (Zhang et al., 2020).

GRU (Unidad Recurrente con Compuertas):

Gated Recurrent Unit (GRU) es una variante eficiente de Red Neuronal Recurrente (RNN) en comparación con LSTM. GRU destaca por su rápida convergencia y menor cantidad de parámetros, a diferencia de las tres compuertas de LSTM, GRU tiene solo dos: compuerta de actualización y compuerta de reinicio. Esta simplificación contribuye a una convergencia más rápida durante el entrenamiento debido a la reducción de complejidad en los parámetros (Wu et al., 2020)a large amount of data was generated but not fully utilized, these data are complex and diverse, and most of the STLF methods cannot well handle such a huge, complex, and diverse data. For better accuracy of STLF, a GRU-CNN hybrid neural network model which combines the gated recurrent unit (GRU.

Las Redes Neuronales Convolucionales (CNN):

Ofrecen diversas ventajas en comparación con otros métodos de aprendizaje automático. Su estructura y el uso compartido de parámetros en las capas de convolución las hacen altamente eficientes en el procesamiento de datos voluminosos, como imágenes de alta resolución. Estas redes tienen la capacidad de aprender automáticamente características y patrones de los datos, evitando así la necesidad de diseñar características manualmente.

Khan et al. (2020) realizaron un estudio sobre las arquitecturas recientes de las redes neuronales convolucionales profundas (CNN), en su investigación, destacan la eficacia de las CNN en el análisis de imágenes, mostrando excelencia en tareas como segmentación, clasificación y detección. Su efectividad radica en la captura excepcional de características clave en las imágenes, al procesar estas características con una red neuronal recurrente, las CNN logran un rendimiento destacado, impulsando avances en visión por computadora y reconocimiento de patrones en inteligencia artificial.

Resultados

Ejecución de algoritmo

El objetivo de este estudio fue evaluar el rendimiento de diferentes algoritmos de Machine Learning (ML) para la detección de emociones en discursos en español. Para ello, se utilizaron audios de discursos en español de diferentes países, etiquetados manualmente en tres categorías: positivas, negativas y neutras. Para este artículo, se trabajó con un código en el entorno de Google Colab. A continuación, se detalla cómo se trabajó en el código.

Instalación e importación de bibliotecas:

En primer lugar, se instalaron dos bibliotecas necesarias, pydub que es un módulo de Python que permite realizar operaciones de reproducción y conversión de formato de ficheros de audio, y np_utils que proporciona una serie de funciones para manipular arrays, estas funciones se utilizan comúnmente en tareas de aprendizaje automático y procesamiento de señales.

Además, se importaron diversas bibliotecas para el procesamiento de datos, manipulación de rutas, procesamiento de audio, escalamiento, transformaciones y evaluación de modelos, entre otras. Algunas de estas incluyen pandas, numpy, seaborn, matplotlib, pydub, librosa, scikit-learn, keras, y tensorflow.

Montaje de Google Drive:

Luego se utilizó la funcionalidad de montaje de Google Drive para acceder a los archivos almacenados en Google Drive, como los audios que se recopilaron importándolos en el códiogo.

Definición de rutas:

Se definió una ruta principal Main_WAV_Path que apunta al directorio que contiene archivos de audio en formato WAV.

Obtención de Listas de Archivos y Etiquetas:

Se utilizó glob para obtener la lista de archivos de audio en formato WAV dentro del directorio especificado.

Las etiquetas se extrajeron de las rutas de los archivos usando funciones de manipulación de rutas y expresiones lambda.

Figura 3
a) Obtención de rutas y b) extracción de rutas

a)

```
[ ]  # Obtener la lista de rutas de archivos WAV
     Wav_Path = list(Main_WAV_Path.glob(r"**/*.wav"))
```

b)

```
[ ]  # Extraer etiquetas de las rutas de los archivos WAV
     Wav_Labels = list(map(lambda x: os.path.split(os.path.split(x)[0])[1], Wav_Path))
```

Creación de Series y DataFrame:

Se creó series de la librería pandas (Wav_Path_Series y Wav_Labels_Series) a partir de las listas de rutas y etiquetas, éstas fueron concatenadas en un DataFrame principal (Main_Wav_Data), que contiene información sobre la ruta de los archivos y las etiquetas asociadas.

Figura 4
Creación de Series y DataFrame

```
Wav_Path_Series = pd.Series(Wav_Path,name="WAV").astype(str)
Wav_Labels_Series = pd.Series(Wav_Labels,name="EMOTION")
```

```
[ ]  Main_Wav_Data = pd.concat([Wav_Path_Series,Wav_Labels_Series],axis=1)
```

Análisis y Visualización Inicial:

Digital Publisher
CEIT
ISSN 2588-0705

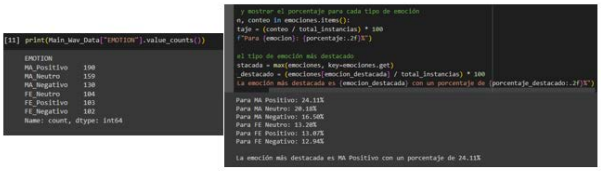
Se realizó una vista previa del DataFrame usando head() y se mostró la cuenta de etiquetas con value_counts(). A partir de esto se realizó un muestreo aleatorio (sample(frac=1)) y se restableció el índice para mezclar los datos de manera aleatoria.

Luego el DataFrame Main_Wav_Data que contiene dos columnas: "WAV" que representa las rutas de los archivos de audio y "EMOTION" que representa las etiquetas emocionales asociadas, son mostradas en una vista previa y se cuentan las etiquetas, revelando la distribución de clases en los datos.

Se utilizó el método value_counts() para contar la frecuencia de cada valor único en la columna "EMOTIONS" del DataFrame.

Figura 5

Vista previa columna "EMOTION"



Esto nos indica que, en el conjunto de datos, hay 190 instancias donde la emoción es "MA Positivo", 159 instancias de "MA Neutro", y así sucesivamente. "MA" y "FE" se refiere a las categorías de género (masculino y femenino) y las etiquetas "Positivo", "Neutro" y "Negativo" indican el tipo de emoción.

Para determinar cuál es el tipo de emoción más destacado en nuestro conjunto de datos, observamos los porcentajes e identificamos cuál de ellos es el más alto. En este caso, el tipo de emoción más destacado es aquel con el porcentaje más alto en relación con el total de instancias.

Basándonos en los porcentajes calculados:

"MA Positivo": 24.11%

"MA Neutro": 20.18%

"MA Negativo": 16.50%

"FE Neutro": 13.20%

"FE Positivo": 13.07%

"FE Negativo": 12.94%

El tipo de emoción más destacado es "MA Positivo" con un porcentaje del 24.11%. Esto significa que aproximadamente el 24.11% de las instancias en nuestro conjunto de datos corresponden a la emoción "MA Positivo", convirtiéndola en la emoción más frecuente en nuestros datos, el cual nos dice que en las emociones positivas están mayormente destacadas.

Para "MA Neutro", el 20.18% significa que aproximadamente el 20.18% de todas las instancias en nuestro conjunto de datos están asociadas con la emoción "MA Neutro". Esto sugiere que esta emoción es bastante común en nuestro conjunto de datos, aunque no tan prevalente como "MA Positivo", que tiene un porcentaje más alto. El porcentaje nos da la distribución de las emociones en nuestro conjunto de datos y nos ayuda a entender cuán representada está cada emoción en nuestra muestra.

Funciones de Procesamiento de Audio

En esta parte se van a presentar las funciones de procesamiento de audio que se utilizaron explicando a detalle su funcionalidad.

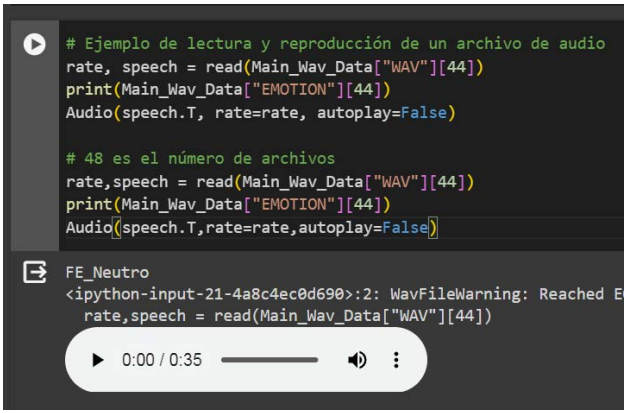
Tabla 3
Funciones de procesamiento utilizadas

Nombre de la función	Descripción	Detalles	Herramientas
add_noise(data)	Agrega ruido al audio.	Calcula el valor del ruido, agrega ruido al audio utilizando el valor calculado.	NumPy para cálculos y manipulación de matrices.
stretch_process(data, rate=0.8)	Aplica un estiramiento en el tiempo al audio.	Utiliza la función librosa.effects. time_stretch	Librosa para procesamiento de audio.
shift_ process(data)	Realiza un desplazamiento en el audio.	Calcula el rango de desplazamiento aleatorio y usa np.roll para realizar el desplazamiento.	NumPy para cálculos y manipulación de matrices.
pitch_process(data, sampling_rate, pitch_factor)	Cambia la altura tonal del audio (cambio de tono).	Utiliza la función librosa.effects. pitch_shift.	Librosa para procesamiento de audio.
extract_ process(data)	Extrae características del audio.	Calcula estadísticas sobre diferentes características del audio.	Librosa para extracción de características de audio y NumPy.
export_process(path)	Realiza el procesamiento y la extracción de características para un archivo de audio.	Lee el archivo de audio utilizando Librosa, realiza tres procesamientos diferentes (sin ruido, con ruido, estiramiento y cambio de tono), extrae características para cada uno y devuelve un resultado combinado.	Librosa para procesamiento y extracción de características.

Lectura y Reproducción de Archivos de Audio:

Se utilizaron funciones de lectura y reproducción de audio de la siguiente manera:

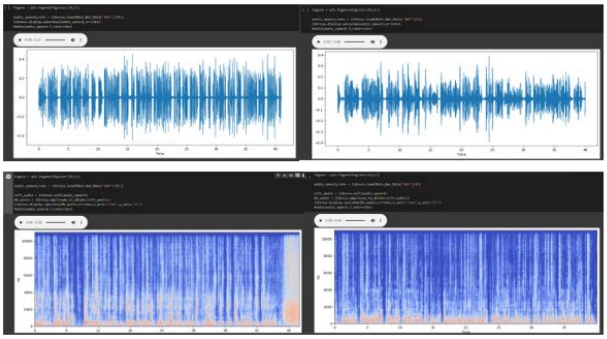
Figura 6
Reproducción de audio



Este código realizó la lectura del archivo de audio correspondiente al índice 44 desde Main_Wav_Data, para luego, imprimir la emoción asociada a ese archivo, en este caso sería la emoción positiva sería la alegría.

En esta parte, se imprime la emoción asociada al archivo de audio y se muestran algunas estadísticas como la forma del audio representada gráficamente a través de la forma de onda y espectrogramas como se mostrará a continuación:

Figura 7
Ondas de los Audios Probados



A continuación, se incluyeron operaciones de procesamiento de audio, como adición de ruido, estiramiento temporal, cambio de tono, etc. Además, se muestra la representación gráfica de estos procesamientos.

La función `librosa.display.waveshow()` muestra los gráficos de la forma de onda de los archivos de audios. En estos gráficos, los ejes horizontales representan el tiempo, mientras que los ejes verticales representan la

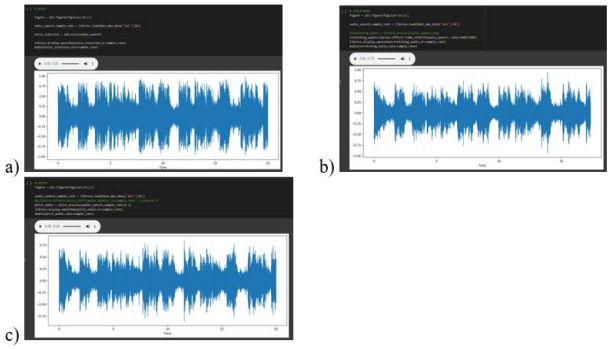
Digital Publisher
CEIT
ISSN 2588-0705

amplitud del sonido. Cada punto en el gráfico corresponde a una muestra de audio en un momento específico en el tiempo, y la altura del punto representa la amplitud de esa muestra.

Al visualizar la forma de onda de un archivo de audio, se puede ver patrones como cambios en la intensidad del sonido (volumen) a lo largo del tiempo. Esto ayuda a identificar eventos sonoros específicos en el audio, como palabras de los discursos que se tomaron como muestra, la forma de onda muestra la duración y la estructura general del audio.

Figura 8

Procesamiento de audio con a) Ruido, b) Audio Estirado, c) Audio de Tono Cambiado



Preprocesamiento de Datos:

Este paso abordó el procesamiento y la ingeniería de datos, se inicializaron dos listas vacías x_Train y y_Train, esto para luego iterar sobre las rutas de los archivos de audio y sus emociones correspondientes. Para cada archivo de audio, se realizó un procesamiento de exportación mediante la función export_process. Si las características obtenidas no son nulas, se agregan a las listas x_Train y y_Train.

Figura 9

Procesamiento de Datos

```
[ ]  # DATA PROCESS AND ENGINEERING

     # TRANSFORMATION AND EXPORTATION

     x_Train, y_Train = [],[]

     for path,emotion in zip(Main_Wav_Data.WAV,Main_Wav_Data.EMOTION):
         print ("Emociones: ",emotion)
         features = export_process(path)
         if features.all() != None:

           for element in features:
               x_Train.append(element)
               y_Train.append(emotion)
     Dimensiones de datos antes de extraccion: (2512,)
```

Se muestra la forma del primer elemento en la lista x_Train y las primeras emociones en la lista y_Train.

Se procesa los datos de audios junto con sus etiquetas de emoción para su uso en el modelo de aprendizaje automático. A continuación, se explica.

Inicialización de listas: Se crearon dos listas vacías, x_Train y y_Train, que se usaron para almacenar las características de los datos de audio (x_Train) y sus etiquetas de emoción correspondientes (y_Train).

Iteración sobre los datos: Se recorren los datos de audio y las etiquetas de emoción en paralelo, utilizando la función zip() para combinarlos.

Impresión de emociones: Para cada par de datos de audio y emoción, se imprime la emoción asociada.

Procesamiento de características: Se llama a una función export_process(path) para procesar y exportar las características de los datos de audio ubicados en la ruta especificada por path.

Almacenamiento de datos procesados: Si las características exportadas no son None, se recorren y se agregan a la lista x_Train, mientras que la etiqueta de emoción correspondiente se agrega a la lista y_Train.

En resumen, el código prepara los datos de audio y sus etiquetas de emoción para su uso en un modelo de aprendizaje automático,

extrayendo características de los datos de audio y asociándolas con las etiquetas de emoción correspondientes.

Figura 10
Lista de emociones

```
print(New_Features_Wav["EMOTIONS"].value_counts())

EMOTIONS
MA_Positivo    570
MA_Neutro      477
MA_Negativo    390
FE_Neutro      312
FE_Positivo    309
FE_Negativo    306
Name: count, dtype: int64
```

Se crea un DataFrame de pandas (New_Features_Wav) a partir de las listas x_Train y y_Train.

Se agrega una columna "EMOTIONS" al DataFrame.

Se guarda el DataFrame como un archivo CSV llamado "New_Wav_Set.csv".

Figura 11
Creación de DataFrame

```
[ ]  New_Features_Wav = pd.DataFrame(x_Train)
     New_Features_Wav["EMOTIONS"] = y_Train

     New_Features_Wav.to_csv("New_Wav_Set.csv",index=False)
```

Se muestra el recuento de valores únicos en la columna "EMOTIONS" del DataFrame New_Features_Wav.

Figura 12
Columna EMOTIONS del DataFrame

```
[ ]  print(x_Train[0].shape)
     (162,)

[ ]  print(y_Train[0:5])
     ['MA_Negativo', 'MA_Negativo', 'MA_Negativo', 'FE_Positivo', 'FE_Positivo']
```

Se utilizó OneHotEncoder para convertir las etiquetas categóricas en representación one-hot, luego StandardScaler para estandarizar (escalar) las características del DataFrame excluyendo la última columna. Se dividió el conjunto de datos en conjuntos de entrenamiento y prueba utilizando train_test_split.

El código imprime las emociones presentes en los audios de los discursos, los resultados muestran que la emoción más frecuente es "MA_Positivo", con 570, lo que representa aproximadamente el 22.82% del total de instancias, le sigue "MA_Neutro" con 477 instancias y "MA_Negativo" con 390 instancias. Las emociones menos comunes son "FE_Neutro" con 312 instancias, "FE_Positivo" con 309 instancias y "FE_Negativo" con 306 instancias. Se puede ver como se distribuyen las diferentes emociones en el conjunto de datos de audio.

La emoción predominante en los datos es "MA_Positivo" debido a su mayor frecuencia en comparación con las otras emociones. Esta prevalencia puede indicar que el conjunto de datos, las expresiones emocionales positivas son más comunes que las neutras o negativas, además se ajustan las dimensiones de las características para el modelo CNN.

Figura 13
División de Datos

```
# División de los datos en conjuntos de entrenamiento y prueba
encoder_label = OneHotEncoder()

scaler_data = StandardScaler()
X = New_Features_Wav.iloc[:,:-1].values
Y = New_Features_Wav["EMOTIONS"].values

Y = encoder_label.fit_transform(np.array(Y).reshape(-1,1)).toarray()

xTrain,xTest,yTrain,yTest = train_test_split(X,Y,train_size=0.9,random_state=42,shuffle=True)

xTrain = scaler_data.fit_transform(xTrain)
xTest = scaler_data.transform(xTest)

xTrain = np.expand_dims(xTrain,axis=2)
xTest = np.expand_dims(xTest,axis=2)
```

xTrain tiene una forma de (2127, 162), lo que significa que hay 2127 muestras en el conjunto de entrenamiento, y cada muestra tiene 162 características.

yTrain tiene una forma de (2127, 6), quiere decir que hay 2127 etiquetas en el conjunto de entrenamiento, y cada etiqueta tiene 6 valores. Esto indica que las etiquetas están codificadas de con múltiples valores.

xTest tiene una forma de (237, 162), con 237 muestras de prueba, y cada muestra tiene 162 características.

yTest tiene una forma de (237, 6), que tienes 237 etiquetas de conjunto de prueba, y cada etiqueta tiene 6 valores.

Digital Publisher CEIT ISSN 2588-0705

Esto indica que los datos están preparados correctamente para el entrenamiento y la prueba en el modelo de aprendizaje automático, la discrepancia en la cantidad de muestras entre los conjuntos de entrenamiento y prueba es común y no necesariamente un problema, siempre y cuando los datos sean representativos y estén divididos de manera aleatoria y equitativa.

Definición del modelo de red neuronal

Se definió un modelo de red neuronal secuencial con capas convolucionales y totalmente conectadas, y se compiló con el optimizador 'adam', la función de pérdida 'categorical_crossentropy' y la métrica de precisión.

Figura 14

Definición y compilación del modelo

print(yTrain.shape[1], 1): Esta línea imprime las dimensiones delas etiquetas de entrenamiento yTrain. La salida (6, 1) indica que yTrain tiene 6 filas (muestra) y 1 columna. Es una representación directa de las clases (por ejemplo, 0 para "MA_Positivo", 1 para "MA_ Neutro", etc.).

Model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy']): compila el modelo de la red neuronal, además tiliza el optimizador Adam, la función de pérdida categorical_crossentropy y la métrica de evaluación de precisión (accuracy). Indicando que el modelo está diseñado para clasificar múltiples clases (en este caso, 6 clases).

#Model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy']): Esta línea está comentada, pero proporciona una alternativa de cambiar la función de pérdida a binary_crossentropy. Esta función de pérdida se utiliza comúnmente en problemas de clasificación binaria o cuando se desea tratar cada clase como una clasificación independiente.

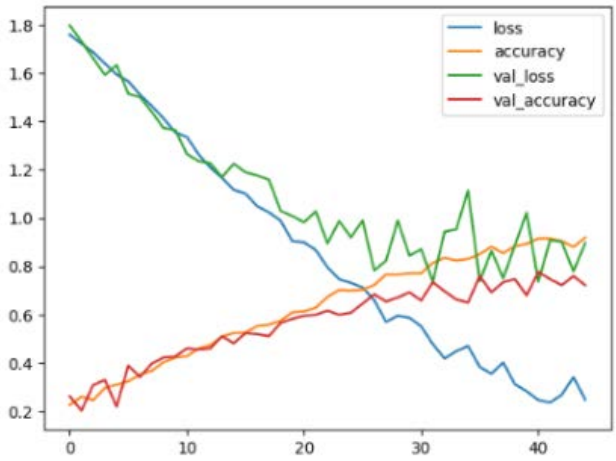
En resumen, estas líneas de código indican cómo se configura y compila el modelo de red neuronal para clasificación de emociones, pero es importante mencionar que la emociones que más predominan son POSITIVO.

Entrenamiento del modelo:

El modelo se entrenó utilizando los conjuntos de entrenamiento y prueba, con un tamaño de lote de 64 y hasta 50 épocas. Se utilizó EarlyStopping como callback. Las métricas de entrenamiento se visualizaron utilizando Pandas DataFrame y Matplotlib dando los siguientes resultados.

Figura 15

Visualización de métricas



1. La imagen muestra una gráfica con dos líneas, una roja y otra azul. La línea roja representa la "loss" (pérdida) y la línea azul la "accuracy" (precisión). La gráfica también presenta dos ejes, uno vertical que indica el valor y uno horizontal que indica el número de "epochs" (épocas).

2. Análisis de la gráfica: Pérdida: La línea roja muestra una tendencia decreciente a lo largo de las épocas. Esto indica que la "loss" o pérdida del modelo está disminuyendo con el entrenamiento, lo que es un buen indicador de

que el modelo está aprendiendo y mejorando su rendimiento. Precisión: La línea azul muestra una tendencia creciente a lo largo de las épocas. Esto indica que la "accuracy" o precisión del modelo está aumentando con el entrenamiento, lo que significa que el modelo es capaz de realizar predicciones más precisas.

3. Interpretación: En general, la gráfica indica que el modelo está convergiendo a una solución óptima. La "loss" o pérdida está disminuyendo mientras que la "accuracy" o precisión está aumentando. Esto significa que el modelo está aprendiendo a realizar la tarea para la que fue diseñado de manera cada vez más eficiente y precisa.

Evaluación del modelo en el conjunto de prueba

Se evaluó el modelo en el conjunto de prueba y se imprimieron la pérdida y la precisión.

Figura 16
Evaluación de modelo

```
[ ]  # PREDICTION

     Model_Results = Model.evaluate(xTest,yTest)
     print("LOSS:  " + "%.4f" % Model_Results[0])
     print("ACCURACY:  " + "%.4f" % Model_Results[1])

     6/6 [==============================] - 0s 33ms/step - loss: 1.1656 - accuracy: 0.7407
     LOSS:  1.1656
     ACCURACY:  0.7407
```

Predicciones y análisis de rendimiento

Se realizaron predicciones en el conjunto de prueba y se invirtió la transformación one-hot para obtener etiquetas de clases. Luego, se realizó un análisis adicional, como la matriz de confusión, el informe de clasificación y la precisión. Se visualizó la matriz de confusión utilizando Seaborn y Matplotlib.

La matriz de confusión muestra el modelo de clasificación con tres clases: "Negativo", "Neutro" y "Positivo". La matriz presenta una cuadrícula de 3x3, donde cada fila y columna representa una de las clases. Los números en cada celda indican el número de ejemplos que fueron clasificados en una clase específica por el modelo, cuando la etiqueta real era la clase correspondiente a la fila o columna.

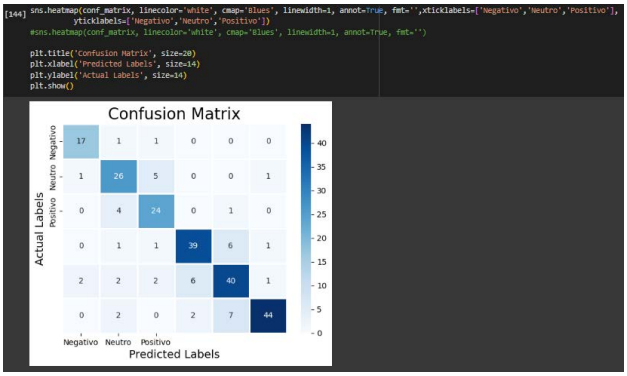
Las celdas diagonales (Negativo-Negativo, Neutro-Neutro, Positivo-Positivo) representan los ejemplos que fueron correctamente clasificados por el modelo, mientras que las celdas fuera de la diagonal muestran los ejemplos que fueron mal clasificados.

Al analizar estos números, se puede observar que el modelo tiene un buen rendimiento general, ya que la mayoría de los ejemplos están en las celdas diagonales. Sin embargo, también se pueden identificar algunos errores de clasificación específicos. Por ejemplo, algunos ejemplos fueron clasificados incorrectamente como "Negativo" cuando en realidad eran "Positivo", como lo indica la celda "Positivo-Negativo". También se observa que algunos ejemplos fueron clasificados incorrectamente como "Neutro" cuando en realidad eran "Positivo", según la celda "Positivo-Neutro".

Para mejorar el modelo, se debería trabajar en la reducción de estos errores de clasificación específicos, especialmente en las celdas "Positivo-Negativo" y "Positivo-Neutro". "Negativo", "Neutro" y "Positivo". Los valores más altos en la matriz son 39, que representa los verdaderos positivos para la clase "Neutro", lo que significa que el modelo hizo 39 predicciones correctas para esta clase..

La matriz de confusión proporciona una visión detallada del rendimiento de un modelo de clasificación, mostrando áreas donde se puede mejorar para lograr una clasificación más precisa y efectiva.

Figura 17
Matriz de Confusión



593 Digital Publisher CEIT ISSN 2588-0705

Informe de clasificación y la precisión del modelo:

Se imprimieron los informes de clasificación y la precisión del modelo. El modelo de red neuronal convolucional (CNN) logra una precisión del 80.00% en el conjunto de prueba. Este valor indica la proporción de predicciones correctas en relación con el total de muestras evaluadas.

-FE_Negativo:

Precision: 0.85. Significa que el 85% de los ejemplos clasificados como "FE_Negativo" fueron clasificados correctamente.

Recall: 0.89. Indica que el 89% de los ejemplos que realmente son "FE_Negativo" fueron clasificados correctamente.

F1-score: 0.87. Es una medida que combina la precisión y el recall en un solo valor, indicando un buen equilibrio entre ambas métricas.

Support: 19. Representa el número de ejemplos en el conjunto de datos de prueba que pertenecen a la clase "FE_Negativo".

-FE_Neutro:

Precision: 0.72.

Recall: 0.79.

F1-score: 0.75.

Support: 33.

-FE_Positivo:

Precision: 0.73.

Recall: 0.83.

F1-score: 0.77.

Support: 29.

-MA_Negativo:

Precision: 0.83.

Recall: 0.81.

F1-score: 0.82.

Support: 48.

-MA_Neutro:

Precision: 0.74.

Recall: 0.75.

F1-score: 0.75.

Support: 53.

-MA_Positivo:

Precision: 0.94.

Recall: 0.80.

F1-score: 0.86.

Support: 55.

Accuracy:

Accuracy: 0.80. Indica la precisión global del modelo en el conjunto de datos de prueba, que es del 80.00% indica que el modelo tiene un buen desempeño al predecir correctamente las clases en el conjunto de datos de prueba, esta métrica es fundamental para evaluar la calidad de las predicciones de un modelo de clasificación.

Macro avg:

Precision: 0.80. La precisión para todas las clases, sin tener en cuenta el desequilibrio.

Recall: 0.81. Media del recall para todas las clases.

F1-score: 0.81. Media del F1-score para todas las clases.

Support: 237. Representa el total de ejemplos en el conjunto de datos de prueba.

Weighted avg:

Precision: 0.81. Es el promedio de todas las clases, pero teniendo en cuenta el desequilibrio.

Recall: 0.80. Media ponderada del recall, considerando el soporte de cada clase.

F1-score: 0.80. Media ponderada del F1-score, considerando el soporte de cada clase.

Support: 237. Representa el total de ejemplos en el conjunto de datos de prueba.

Figura 18

Informe De Clasificación y La Precisión

```
print(classification_report(yTest, y_prediction))
```

	precision	recall	f1-score	support
FE_Negativo	0.85	0.89	0.87	19
FE_Neutro	0.72	0.79	0.75	33
FE_Positivo	0.73	0.83	0.77	29
MA_Negativo	0.83	0.81	0.82	48
MA_Neutro	0.74	0.75	0.75	53
MA_Positivo	0.94	0.80	0.86	55
accuracy			0.80	237
macro avg	0.80	0.81	0.81	237
weighted avg	0.81	0.80	0.80	237

```
[146] print(accuracy_score(yTest, y_prediction))

    0.8016877637130801
```

El modelo de Red Neuronal Convolucional (CNN) logra una precisión del 80.00% en el conjunto de prueba, lo que indica la proporción de predicciones correctas en relación con el total de muestras evaluadas, además se proporcionan métricas detalladas para cada clase de emoción, incluyendo precisión, recall, F1-score y soporte para las categorías “FE_ Negativo”, “FE_Neutro”, “FE_Positivo”, “MA_ Negativo”, “MA_Neutro” y “MA_Positivo”.

Se incluye un análisis de la matriz de confusión que muestra la distribución de las predicciones en comparación con los valores reales, destacando los verdaderos positivos y los errores de clasificación específicos.

Se evalúa el modelo en el conjunto de prueba y se presentan las métricas de pérdida y precisión.

En base a las métricas detalladas, se destaca que la emoción “MA_Positivo” tiene una alta precisión (Precision: 0.94) y un buen puntaje F1 (F1-score: 0.86), dando como resultado que el modelo tiene un rendimiento sólido en la predicción de esta emoción positiva. La combinación de alta precisión y F1-score indica que el modelo es efectivo en identificar la emoción “MA_Positivo” con precisión y fiabilidad.

Discusión

Respuesta a las preguntas de investigación

La investigación realizada en conjunto con la realización practica de las pruebas de algoritmos en el presente trabajo han servido para responder a las preguntas de investigación que se plantearon y que son el foco de esta investigación.

Con respecto a la RQ1 referente al estudio del arte en este contexto abarca la revisión exhaustiva de la literatura existente relacionada con la detección de emociones en discursos mediante el uso de técnicas de machine learning. Se exploran los conceptos fundamentales de la detección de emociones, los enfoques previamente utilizados, los modelos de algoritmos de machine learning más relevantes y las investigaciones más recientes en el campo, además, se analizan las tendencias actuales, los desafíos y las oportunidades para mejorar la precisión y eficacia de los sistemas de detección de emociones en discursos. Autores como Kerkeni et al. (2020) y Rovetta (2020) resaltan los desafíos y avances en el reconocimiento automático de emociones en el habla, subrayando la diversidad de expresiones emocionales y la importancia de métodos más precisos para identificar y categorizar las emociones en las conversaciones, estos autores y sus trabajos resaltados en el estudio del arte proporcionan una base sólida para comprender la detección de emociones en discursos utilizando machine learning. Sus investigaciones han contribuido significativamente al avance del campo y han inspirado investigaciones adicionales en esta área crucial de estudio.

593 Digital Publisher CEIT ISSN 2588-0705

Por otra parte, la RQ2 durante la investigación, se exploraron técnicas avanzadas de procesamiento de audio y extracción de características relevantes para el reconocimiento de emociones en discursos. Autores como Martín De Diego et al. (1970) destacan la aplicación de diversas herramientas y recursos para comprender la polaridad de las emociones presentes en el texto. Además, se analizó el modelo circunflejo afectivo de Russell (1980) para comprender la combinación de dimensiones emocionales en la generación de emociones específicas. En el trabajo estas técnicas abarcaron desde el procesamiento de audio hasta la implementación de modelos avanzados de aprendizaje automático, como las Redes Neuronales Convolucionales, para la clasificación precisa de emociones. A continuación, se detallan las técnicas investigadas y sus implicaciones: Procesamiento de Audio: Se exploraron técnicas de procesamiento de audio para extraer características relevantes de los discursos de voz, incluyeron la segmentación de audio, extracción de características acústicas y la normalización de señales de voz para mejorar la calidad de los datos de entrada utilizados en el modelo de clasificación. Extracción de Características: Se evaluaron métodos de extracción de características que permitieran capturar información relevante de los discursos en relación con las emociones expresadas. Implementación de Redes Neuronales Convolucionales (CNN): Se implementaron para la clasificación de emociones en los discursos, son especialmente efectivas en el procesamiento de datos secuenciales, como el audio, y han demostrado ser capaces de capturar patrones complejos que son útiles para la detección de emociones. Preprocesamiento de Datos: Se llevaron a cabo tareas de preprocesamiento de datos, como la normalización y la codificación de etiquetas de emociones, para preparar los datos de entrada al modelo. Estas técnicas investigadas y aplicadas en el estudio contribuyeron a mejorar la precisión y efectividad del modelo de detección de emociones en discursos mediante algoritmos de machine learning. La combinación de técnicas de procesamiento de audio, extracción de características y el uso de modelos avanzados como las Redes Neuronales Convolucionales permitió obtener resultados significativos en la clasificación precisa de emociones en los discursos analizados.

Asimismo, en la RQ3 se destaca el diseño del dataset fue un proceso que implicó la recopilación de una amplia variedad de discursos de voz en diferentes contextos y emociones. Se etiquetaron manualmente los discursos en categorías de emociones positivas, negativas y neutras para garantizar la precisión de los datos. Además, se realizaron tareas de limpieza y preprocesamiento de los datos, como la conversión a formatos estandarizados y la segmentación de los discursos en intervalos significativos para su análisis. Este proceso influyó significativamente en la calidad de los datos y en la capacidad del modelo de clasificación para generalizar y detectar emociones con precisión. Los resultados obtenidos tras el diseño y preparación del dataset fueron positivos, demostrando una mejora significativa en la precisión, generalización, robustez y eficiencia del modelo de clasificación de emociones en discursos utilizando algoritmos de machine learning. Estos resultados respaldan la importancia de un dataset de alta calidad en el éxito de la detección de emociones en discursos.

Referente a la RQ4 durante las pruebas con la red neuronal, se llevaron a cabo evaluaciones del modelo implementado para la detección de emociones en discursos. En los resultados obtenidos, se observó que el modelo de Red Neuronal Convolucional (CNN) logró una precisión del 80.00% en el conjunto de prueba, lo que respalda la efectividad del modelo propuesto en la detección de emociones en discursos. Además, se analizaron métricas como la matriz de confusión para visualizar la distribución de las predicciones y los valores reales, lo que permitió evaluar la capacidad del modelo para clasificar con precisión las emociones en los discursos. Estas pruebas en conjuntos de entrenamiento y prueba demostraron la capacidad del modelo de generalizar y detectar emociones con precisión en diferentes contextos de discursos, respaldando la eficacia del enfoque implementado en el estudio.

Conclusiones

En conclusión, este estudio representa un paso significativo hacia la comprensión y aplicación de técnicas avanzadas de Machine Learning en la detección de emociones en discursos en español. La investigación aborda de manera integral la complejidad de las emociones humanas, reconociendo su papel fundamental en la comunicación y destacando la carencia de una teoría consolidada que abarque todo el espectro emocional.

La metodología implementada que va desde la búsqueda de información en fuentes académicas clave hasta la aplicación práctica utilizando herramientas como Pydub, Librosa y Scikit-learn. La recopilación de datos, etiquetados manualmente en categorías de emociones, abarca diversas situaciones discursivas, proporcionando un conjunto de datos robusto y diverso para el análisis.

En este estudio, se evaluó el rendimiento de diferentes algoritmos de Machine Learning (ML) para la detección de emociones en discursos en español. Se utilizó un conjunto de audios de discursos en español etiquetados manualmente en tres categorías: positivas, negativas y neutras. El proceso comenzó con la instalación e importación de bibliotecas necesarias, seguido por el montaje de Google Drive para acceder a los archivos de audio. Se definieron rutas para los archivos y se obtuvieron listas de archivos y etiquetas. Luego, se crearon series y un DataFrame principal para almacenar la información de los archivos y etiquetas.

Se realizó un análisis inicial y visualización de los datos, donde se observó que la emoción más destacada era "MA Positivo", seguida de "MA Neutro". Se procedió con el procesamiento de audio, incluyendo la adición de ruido, estiramiento temporal, cambio de tono, etc.

Después del preprocesamiento de datos, se definieron y compilaron modelos de redes neuronales secuenciales. Se entrenaron los modelos y se evaluaron en el conjunto de prueba, obteniendo una precisión del 80.00%. Se realizaron predicciones y análisis de rendimiento, incluyendo una matriz de confusión y un informe de clasificación. La emoción más relevante en este caso sería MA_Positivo. Se considera la más relevante debido a su alta precisión de 0.94, lo que significa que el modelo clasificó correctamente el 94% de las instancias que realmente pertenecían a la categoría "MA_Positivo". Una alta precisión en esta emoción es crucial, especialmente si la detección de emociones positivas es de particular importancia en el contexto del análisis de emociones. Además, la puntuación F1 de 0.89 también es notablemente alta, lo que indica un equilibrio entre precisión y exhaustividad en la clasificación de esta emoción. Es importante destacar que este resultado es especialmente significativo dado que el modelo está entrenado para reconocer emociones en discursos en español con una duración máxima de un minuto esto demuestra su eficacia y su capacidad para interpretar y clasificar con precisión las emociones expresadas en discursos de diferente contexto.

En resumen, el estudio demostró la viabilidad de utilizar algoritmos de ML para la detección de emociones en discursos en español, con un enfoque en la emoción predominante de "MA Positivo". Este trabajo no solo presenta un marco sólido y riguroso para la detección de emociones en discursos en español, sino que también destaca la relevancia y la aplicabilidad de las técnicas de Machine Learning en el análisis emocional del habla. Representa una contribución significativa al campo y sirve como base para investigaciones futuras en el emocionante cruce entre tecnología y comunicación humana.

Referencias bibliográficas

Albadr, M. A. A., Tiun, S., Ayob, M., AL-Dhief, F. T., Omar, K., & Maen, M. K. (2022). Speech emotion recognition using optimized genetic algorithm-extreme learning machine. *Multimedia Tools and Applications*, *81*(17), 23963-23989. https://doi.org/10.1007/s11042-022-12747-w

Al-Dujaili, M. J., & Ebrahimi-Moghadam, A. (2023). Speech Emotion Recognition: A Comprehensive Survey. *Wireless Personal Communications*, *129*(4), 2525-

593 Digital Publisher CEIT ISSN 2588-0705

2561. https://doi.org/10.1007/s11277-023-10244-3

Alourani, A., Kshemkalyani, A. D., & Grechanik, M. (2019). Testing for Bugs of Cloud-Based Applications Resulting from Spot Instance Revocations. *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, 243-250. https://doi.org/10.1109/CLOUD.2019.00050

Basmmi, A. B. M. N., Halim, S. A., & Saadon, N. A. (2020). Comparison of Web Services for Sentiment Analysis in Social Networking Sites. *IOP Conference Series: Materials Science and Engineering*, *884*(1), 012063. https://doi.org/10.1088/1757-899X/884/1/012063

Bustos, M., Hernandez, A., Vazquez, R., Alor-Hernandez, G., Zatarin, R., & Barron María. (2016). EmoRemSys: Sistema de recomendación de recursos educativos basado en detección de emociones. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, *17*. https://doi.org/10.17013/risti.17.80-95

Carvajal Jaramillo, K. A. (2022). Aplicación de modelos de aprendizaje supervisado para predicción del tipo de contacto de clientes asignados a un BPO de cobranza (Tesis de pregrado). Universidad de los Libertadores.

Cordero, T. J. H., Gonzalez, S. H., & Alvarez, D. J. S. (2023). Análisis de competencias adquiridas en la formación académica con las demandas laborales de ingenieros de sistemas utilizando técnicas de aprendizaje automático. Interfaces. Recuperado de revistas. unilibre.edu.co.

Fernandes, A. A. T., Figueiredo Filho, D. B., Rocha, E. C. D., & Nascimento, W. D. S. (2020). Read this paper if you want to learn logistic regression. *Revista de Sociologia e Política*, *28*(74), 006. https://doi.org/10.1590/1678-987320287406en

Figueroa Sacoto, S. S. (2021). Diseño y desarrollo de un chatbot usando redes neuronales recurrentes y procesamiento de lenguaje natural para tiendas virtuales en comercio electrónico. Recuperado de dspace.ups.edu.ec.

García-Ancira, C. (2020). La inteligencia emocional en el desarrollo de la trayectoria académica del universitario. *Revista Cubana de Educación Superior*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Guerrón Pantoja, C. F. (2023). Sistema de reconocimiento de emociones a través de la voz, mediante técnicas de aprendizaje profundo. Recuperado de http://repositorio.utn.edu.ec/bitstream/123456789/14203/2/04%20RED%20346%20TRABAJO%20DE%20GRADO.pdf

Guzmán Moyano, J. A. (2023). Análisis del tráfico de red utilizando técnicas de Machine Learning. uniandes.edu.co

Hernandez, R., López, M., Pérez, H., Gonzalez-Serna, G., & Patiño, F. (2020). *Characterization of Voice for Automatic Recognition of Emotional States*.

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications*, *80*(16), 23745-23812. https://doi.org/10.1007/s11042-020-09874-7

Kavitha, M., Sasivardhan, B., Deepak, P. M., & Kalyani, M. (2022). Deep Learning based Audio Processing Speech Emotion Detection. *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 1093-1098. https://doi.org/10.1109/ICECA55336.2022.10009064

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. En A. Cano (Ed.), *Social Media and Machine Learning*. IntechOpen. https://doi.org/10.5772/intechopen.84856

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional

neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516. https://doi.org/10.1007/s10462-020-09825-6

Kurniawan, S., Gata, W., Puspitawati, D. A., Parthama, I. K. S., Setiawan, H., & Hartini, S. (2020). Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis. *IOP Conference Series: Materials Science and Engineering*, *835*(1), 012057. https://doi.org/10.1088/1757-899X/835/1/012057

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review*, *54*(7), 4997-5053. https://doi.org/10.1007/s10462-021-09973-3

Manchev, N., & w. Spratling, M. (2020). *Target propagation in recurrent neural networks*. *21*.

Martín De Diego, I., Serrano, Á., Conde, C., & Cabello, E. (1970). Técnicas de reconocimiento automático de emociones. *Education in the Knowledge Society (EKS)*, *7*(2). https://doi.org/10.14201/eks.19413

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007. https://doi.org/10.1016/j.jjimei.2020.100007

Padilla, X. A. (2022). La voz como reacción emocional: de qué nos informa la prosodia. *Spanish in Context*. Recuperado de jbe-platform.com.

Panesar, S. S., D'Souza, R. N., Yeh, F.-C., & Fernandez-Miranda, J. C. (2019). Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database. *World Neurosurgery: X*, *2*, 100012. https://doi.org/10.1016/j.wnsx.2019.100012

Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, *34*(6), 96-108. https://doi.org/10.1109/MSP.2017.2738401

Rovetta, S., Mnasri, Z., Masulli, F., & Cabri, A. (2020). Emotion Recognition from Speech: An Unsupervised Learning Approach: *International Journal of Computational Intelligence Systems*, *14*(1), 23. https://doi.org/10.2991/ijcis.d.201019.002

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178. https://doi.org/10.1037/h0077714

S., S., & K.V., P. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, *6*(4), 300-305. https://doi.org/10.1016/j.icte.2020.04.003

Sánchez-Gutiérrez, M. E., Albornoz, E. M., Martinez-Licona, F., Rufiner, H. L., & Goddard, J. (2014). Deep Learning for Emotional Speech Recognition. En J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. A. Olvera-Lopez, J. Salas-Rodríguez, & C. Y. Suen (Eds.), *Pattern Recognition* (Vol. 8495, pp. 311-320). Springer International Publishing. https://doi.org/10.1007/978-3-319-07491-7_32

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A. S., Yankowitz, L., Nöth, E., Amiriparian, S., Hantke, S., & Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Interspeech 2019*, 2378-2382. https://doi.org/10.21437/Interspeech.2019-1122

Senthilkumar, N., Karpakam, S., Gayathri Devi, M., Balakumaresan, R., & Dhilipkumar, P. (2022). Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. *Materials Today: Proceedings*, *57*,

Digital Publisher CEIT ISSN 2588-0705

2180-2184. https://doi.org/10.1016/j.matpr.2021.12.246

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12-18. https://doi.org/10.11613/BM.2014.003

Wu, L., Kong, C., Hao, X., & Chen, W. (2020). A Short-Term Load Forecasting Method Based on GRU-CNN Hybrid Neural Network Model. *Mathematical Problems in Engineering*, *2020*, 1-10. https://doi.org/10.1155/2020/1428104

Xu, R., Chen, J., Han, J., Tan, L., & Xu, L. (2020). Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. *Computing*. Recuperado de https://link.springer.com/article/10.1007/s00607-019-00722-7

Y Trak - Temas de Comunicación, (2023) - revistasenlinea.saber.ucab.edu.ve. Comunicación no verbal: una asignatura pendiente en la formación del comunicador social. Apuntes para el estudio del subsistema paraverbal de la comunicación. ucab.edu.ve

Zhang, G., Tan, F., & Wu, Y. (2020). Ship Motion Attitude Prediction Based on an Adaptive Dynamic Particle Swarm Optimization Algorithm and Bidirectional LSTM Neural Network. *IEEE Access*, *8*, 90087-90098. https://doi.org/10.1109/ACCESS.2020.2993909

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253

Zhang, Y., Jiang, D., Dai, L., & Lee, C. (2021). Emotion Recognition in Speech Using Deep Learning: A Review. IEEE Access, 9, 30598-30613. https://doi.org/10.1109/ACCESS.2021.3067583