CLASIFICACIÓN DE AUDIO PARA PREDECIR SENTIMIENTOS

Lic. Hugo Arnoldo Oliva Castillo

Maestría en Ciencia de Datos. Procesamiento y Clasificación de Datos

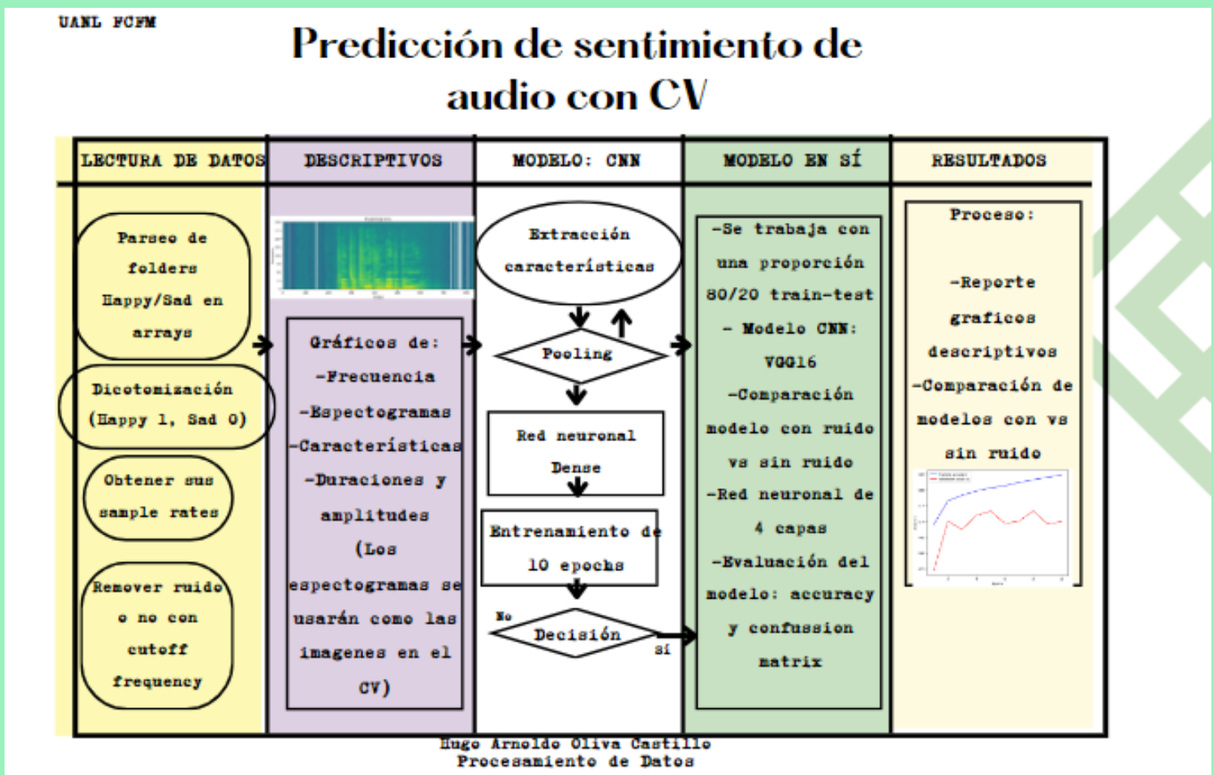
INTRODUCCIÓN

El sentimiento puede estar presente en como una persona pronuncia una oración, por eso en esta ocasión se emplearán técnicas de análisis de audio plasmadas en imágenes para predecir la emoción imperante en un audio (feliz o triste). Las técnicas que se usaron fueron análisis de frecuencias, espectogramas, MFCC (extracción de características), entre otros. Las muestras se componen de frases de no más de 20 segundos segmentados en dos listas de 2167 records de archivos .wav/.mp3

Objetivo principal: Mediante el uso de técnicas de análisis de audio y Machine Learning, predecir la emoción preponderante de dicho audio.

Objetivos secundarios: el procesamiento adecuado de la información para llevar a cabo dichos procesos y la obtención de aprendizajes interesantes.

Figura A) Resumen del trayecto del proyecto



METODOLOGÍA

Primeramente se parsearon los archivos en formato array, obteniendo sus sample rates (número de muestras de audio por segundo, Hz).

Las muestras fueron puestas en tratamiento de remoción de ruido de fondo con un cutoff frequency, para compararlo vs sin remoción de ruido. (*Figura 1*)

Se procedió a realizar análisis descriptivos de los audios en sus características generales, por ejemplo, duración, amplitud de los sonidos, y “energies” (cuantifican la "fuerza" o "amplitud" general de una señal). (*Figura 2).*

Figura 1: Comparativa audio, forma original vs remoción de ruido

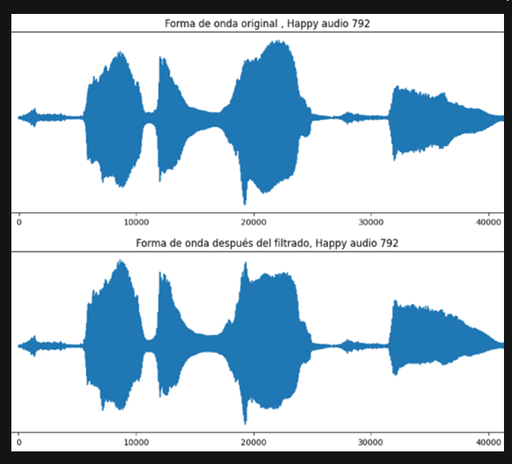
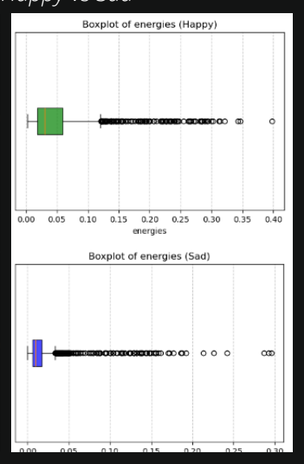


Figura 2: Comparación “energies” de todos los audios Happy vs Sad



Se visualizaron las Ondas en manera visual con gráficos de frecuencias y espectogramas y extracciones de características con MFCC. (*Figura 3).*

Figura 3: Ejemplo de MFCC y Espectograma

Los espectogramas en sus dos tratamientos fueron exportados a imagen. Esto con el fin de ser trabajados en una red convolucional, (bajo la idea de que el espectograma es una representación visual fidedigna del audio, *Fig 4.).*

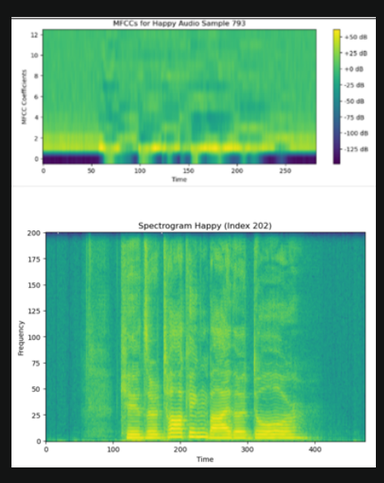
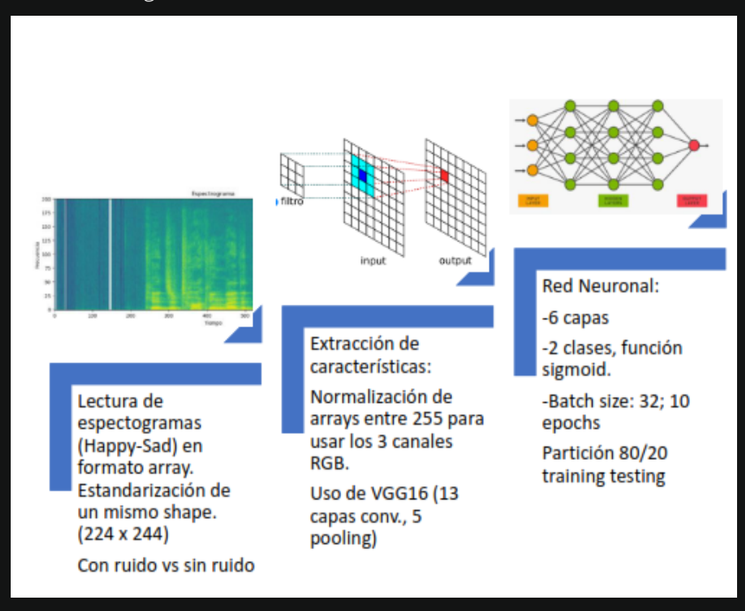


Figura 4: Modelo Red Neuronal Convolucional

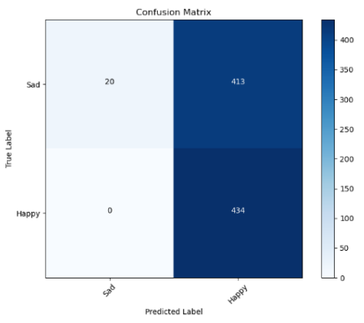


COMPLEMENTO/TRABAJO A FUTURO

Elementos que no se abordaron en esta metodología y que pudieran ser trabajo futuro: trabajar usando los gráficos MFCC; explorar con otro cutoff rate en el ruido de fondo; implementar más categorías de emociones; y técnicas de análisis de sentimiento en las frases, modificar la red neuronal o usar otra red convolucional, comparativa de todos los modelos posibles para hacerlo determinístico, entre otros.

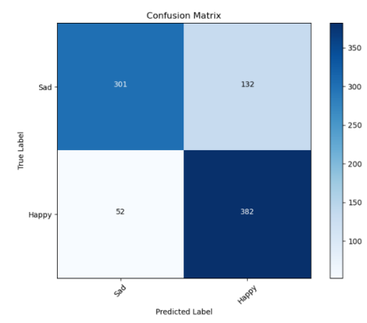
RESULTADOS/CONCLUSIONES

Fig. 5: Resultados CM modelo con ruido



Accuracy: 52.36%

Fig. 6: Resultados CM modelo sin ruido



Accuracy: 78.78%

*E*n las matrices de confusión, se destaca que el modelo con ruido mostró confusiones significativas, principalmente etiquetando la mayoría de las muestras como "Happy". En contraste, el modelo sin ruido, a pesar de tener menor precisión en la categoría "Happy", logró una clasificación más equilibrada para ambas categorías, mejorando así la exactitud general. Ambos modelos comenzaron con una precisión de entrenamiento alrededor del 78%, pero la diferencia clave surgió durante las pruebas, donde el modelo sin ruido demostró ser más capaz de generalizar los resultados del aprendizaje.

Esta discrepancia sugiere que tratar las muestras al eliminar el ruido de fondo si es conveniente para el accuracy del modelo, debido a que los modelos sin ruido pudieran ser más limpios y por lo tanto más aptos para generalizar.

REFERENCIAS

Referencias en el código QR.

Muchas gracias. Contacto:

arnoldo.oliva12@gmail.com