Hate Speech Detection Using Mel Spectrogram and Vision Transformer on Custom Made Bengali Data

Anushka Singh
CSE(AI) 3rd Year Batch A
Amrita School of Artificial Intelligence
CB.EN.U4AIE22006

Devika S
CSE(AI) 3rd Year Batch A
Amrita School of Artificial Intelligence
CB.EN.U4AIE22014

Mohana Priya M
CSE(AI) 3rd Year Batch A
Amrita School of Artificial Intelligence
CB.EN.U4AIE22035

Vaishnavi Venkat
CSE(AI) 3rd Year Batch B
Amrita School of Artificial Intelligence
CB.EN.U4AIE22158

***Abstract*—The Classification of Hate and Non-Hate Speech in Bengali Language is quite limited, considering the huge number of people speaking this language. This is due to the very limited number of available datasets as it is a *low-resourced* language. Significant research has been conducted in this language and in hate detection through other sources of media such as memes and textual content, but audio content remains unexplored. This gives an opportunity to create our own dataset primarily from YouTube and also analyze the audio content through speech processing applications helping us differentiate between hate and non-hate audio.**

This project aims at generating Mel Spectrograms for all the audio samples and feeding the images of Mel Spectrogram into the Vision Transformer model for classification, with a 5 cross-validation of the dataset, it has achieved an accuracy of 98 percent.

***Index Terms*—Bengali audio data set, Speech processing, Noise reduction, Mel Spectrogram, Vision Transformer.**

I. Introduction

As social networks and online communication tools spread rapidly among a wide range of age groups, maintaining a polite and safe online environment has become crucial. Audio-based communication is becoming more and more popular among on-line content types, especially through voice messages, podcasts, and video platforms. However, verbal hate detection in audio content—particularly in low-resource languages such as Bengali—remains understudied, despite significant advances in the detection of hate speech in text and images.

This study focuses on the important and timely task of identifying and categorizing hate speech in Bengali audio.

Our *motivation* to do this project is that there are methods of hate detection existing only for the significant amount of text and image data available and speech techniques although existing are used less due to unavailability of datasets - especially for low resourced language like Bengali.Creating an efficient classification model to differentiate between hate and non-hate speech in Bengali audio is the main goal of this project.

The following crucial steps are involved in the project:

- Dataset Creation: Gathering audio samples in Bengali from multiple sources and classifying them as hate or non-hate speech.
- Pre-processing and cleaning: Using normalization and noise reduction techniques to improve audio quality.
- Mel-Spectrograms are used in feature extraction to transform audio signals into visual representations that are appropriate for deep learning.
- Model Development: Using a Vision Transformer (ViT) model, the audio samples are categorized according to features that have been extracted.

II. Literature Survey

Numerous research methods have been carried out to identify and detect hate and abuse speech through audio. The few papers mentioned have used ADIMA as their data set and traditional methods of obtaining a mel spectrogram and other pre-trained methods such as VGG,Wav2Vec2 models, XLSR-53, CLSRIL-23, Him-4200. The final deep learning methods used are Mean Pool, Max Pool,LSTM and GRU. This has achieved an overall 79.67% accuracy [1].In the second article [2] the authors have used the ADIMA data set along with generation of the Mel spectrogram, as it clearly aligns with the human pitch and finally used CNN - takes input from the LNN spectrogram - takes data of the extracted texts. This overall methodology achieves an accuracy of 77.47% percent. A vision transformer approach is adopted in paper [3] for Telugu speech, which combined with an MLP classifier and using 2 hours of data, achieved 87.09% accuracy. Paper [4] discusses about Malayalam hate speech detection using a multimodal approach of 1D-CNN and BERT, with traditional features extracted like MFCC and Chromagram, with an overall 98.39% accuracy. Another paper focuses on Emotion recognition, [5] where VGG19 and data augmentation techniques on RAVDESS and EMODB datasetsare performed, with gaining an overall accuracy of 97.71% and 98.79%, respectively. Paper [6] talks about research made on multilingual abuse detection using classical classifiers (LR, RF, MLP) with features like MFCC and Chroma, reporting moderate gains over baseline ADIMA models. Similarly, paper [7] captures acoustic features extracted thoruogh OpenSMILE and evaluates models like RF and SVM, finding RF to get the best results across multilingual and cross-lingual conditions.

Paper [8], compares CNN-LSTM and Vision Transformers for emotion recognition, finding CNN-LSTM performing with an overall 88.5%accuracy.Paper [9], where machine learning techniques including logistic regression utilize Mel spectrograms and MFCCs for hate speech detection, achieving 87% accuracy. Lastly, paper [10] compares CNN and LSTM on EMODB using MFCCs.LSTMs performs better in handling large data without increasing network size,and an accuracy of 85.5%.

III. DATA SET DESCRIPTION

A. DATASET COLLECTION FROM YOUTUBE

The primary source for the custom made data for hate speech detection is taken from YouTube. Here there are two classes Hate and Non - Hate. Each of the two distinct sets comprise of 1 hour 10 minutes and 1 hour 5 minutes of audio samples respectively.After chunking some of the larger audio samples the total audio samples produced are 124 and 102 for the two classes respectively.

For finding and downloading the Hate audios, the searches were made with regard to political hate history and geopolitical fights , personal fights, and humiliation. The audio is also collected from news channels that reported any defaming fights or verbal accusations. There were additional audios available from adult cartoon comedy which was completely based on using foul language and hate language and individual videos which captured verbal accusations and foul conversations.

Non-hate audios were relatively available with less complex searches from various motivational videos, tutorial sessions and personal, general conversation from a single speaker from radio channels that exist on YouTube and a few children cartoons.

The data set has no repeated speakers to ensure that there are no two Mel Spectrograms that look similar. At least two audio samples with the same speaker exists due to chunking the large audio file, this is managed to a great extent to reduce the audio samples with the same speakers.

B. TABULAR DESCRIPTION

TABLE I
COMPLETE DATASET STATISTICS

Statistics	**HATE**	**NON-HATE**
Minimum Duration	12 sec	6 sec
Maximum Duration	30 sec	30 sec
Total Duration	1 hour 10 min	1 hour 5 min
Total No. of Samples	124	102
Total Female Samples	21	19
Total Male Samples	103	83
Unique Speakers	74	85
Unique Speakers (Female)	8	16
Unique Speakers (Male)	66	69

IV. METHODOLOGY

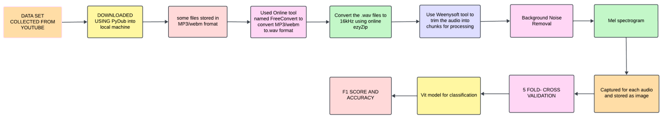

DATA SET COLLECTED FROM YOUTUBE
DOWNLOADED USING PyDub into local machine
some files stored in MP3/webm fromat
Used Online tool named FreeConvert to convert MP3/webm to.wav format
Convert the .wav files to 16kHz using online ezyZip
Use Weenysoft tool to trim the audio into chunks for processing
Background Noise Removal
Mel spectrogram
Captured for each audio and stored as image
5 FOLD- CROSS VALIDATION
Vit model for classification
F1 SCORE AND ACCURACY

Fig. 1. PIPELINE

A. EXTRACTING AUDIOS

The audios are selected from youtube and downloaded into the local machine using the Python tool called Pydub. The audios are downloaded and stored in either MP3 or webm format. A tool named FreeConvert is used to convert all the audio into .wav format. WAV format is required as the files are lossless, meaning they retain all original audio data, ensuring the highest possible sound quality without compression artifact. WAV format is also compatiple with generating the Mel Spectrogram. The WAV format files are converted into 16Khz because it is one of the most common pre-processing steps where most speech models are trained on 16kHz audio which is nothing but the sampling rate.

B. REMOVAL OF NOISE

Removal of noise is a crucial part especially in the Non-Hate audios as it contains a lot of background music. When generating Mel Spectrograms for the Non-Hate audios it generates Mel Spectrogram similar to Hate audios, thus noise removal is a key part in the pre-processing.This function is called using the noisereduce.

C. GENERATING THE MEL SPECTROGRAM

Mel Spectrogram is the main objective of this work where it conveys the audio representation of a speaker. This is generated for all samples of the Hate and Non-Hate classes and stored for feeding it into the deep learning model.Mel spectrogram usually depicts the high energy levels and silenced areas which are one of the important features for differentiating the types of audios

D. MODEL BUILDING

The goal of the model is to classify whether an audio clip belong to the hate ot the non hate class. But instead of feeding the audio directly into the model the images which are the mel spectrograms are fed in to the ViT (Vision Transformer Model) to do the classification.

The mel spectrogram pictures the sound, where it shows how the frequency content of the audio changes over time. ViT model is originally mage for image classification, it works like a trasformer for images, where it breaks the images into small patchers and process them. It learns which part (the patches) of the image are important to make the decision

The 5- fold cross validation is done as the images captured totally are only 226. Here it train on 4 parts, and test on the remaining 1 part thus it makes sure the model performs well on different parts of data

The ViT takes an image and breaks them into pieces which are called patches then uses a transformer to understand the image. The image is split into small square chunks, where the image is reduce to a 16x16 size. Each patch is flattened which is turned into a 1D vectore so that it can be processed in the model easily. At each image when the image is processed ViT adds positional embeddings so that the models knows where each patch from the original image has occured from. The transformer learns relationships between different patches focusing on important parts of the images

V. RESULTS

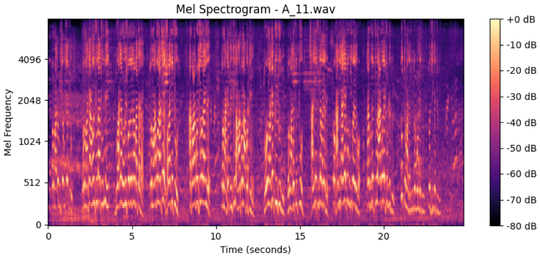


Fig. 2. HATE SPEECH DETECTION IMAGE 1

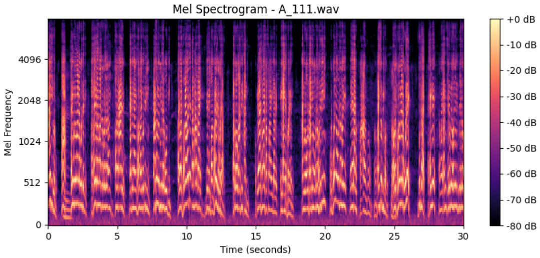


Fig. 3. HATE SPEECH MEL SPECTOGRAM IMAGE 2

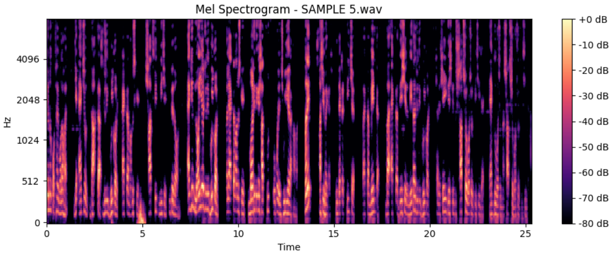


Fig. 4. NON-HATE SPEECH MEL SPECTROGRAM IMAGE 1

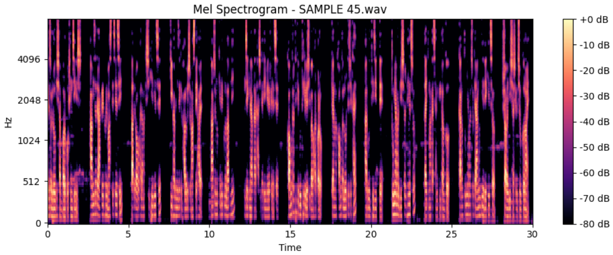
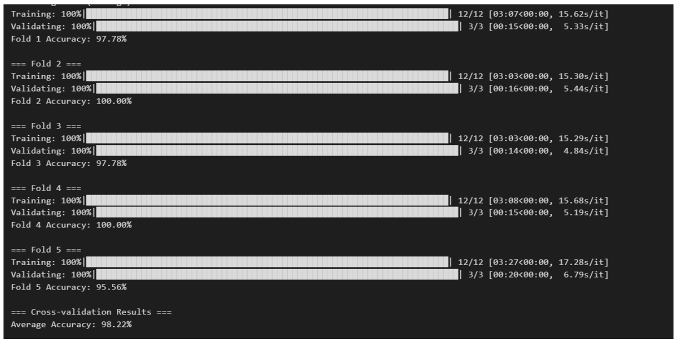


Fig. 5. NON-HATE SPEECH MEL SPECTROGRAM IMAGE 2



```
Training: 100%|████████████████████████████████████████| 12/12 [03:07<00:00, 15.62s/it]
Validating: 100%|████████████████████████████████████████| 3/3 [00:15<00:00,  5.33s/it]
Fold 1 Accuracy: 97.78%

=== Fold 2 ===
Training: 100%|████████████████████████████████████████| 12/12 [03:03<00:00, 15.30s/it]
Validating: 100%|████████████████████████████████████████| 3/3 [00:16<00:00,  5.44s/it]
Fold 2 Accuracy: 100.00%

=== Fold 3 ===
Training: 100%|████████████████████████████████████████| 12/12 [03:03<00:00, 15.29s/it]
Validating: 100%|████████████████████████████████████████| 3/3 [00:14<00:00,  4.84s/it]
Fold 3 Accuracy: 97.78%

=== Fold 4 ===
Training: 100%|████████████████████████████████████████| 12/12 [03:08<00:00, 15.68s/it]
Validating: 100%|████████████████████████████████████████| 3/3 [00:15<00:00,  5.19s/it]
Fold 4 Accuracy: 100.00%

=== Fold 5 ===
Training: 100%|████████████████████████████████████████| 12/12 [03:27<00:00, 17.28s/it]
Validating: 100%|████████████████████████████████████████| 3/3 [00:20<00:00,  6.79s/it]
Fold 5 Accuracy: 95.56%

=== Cross-validation Results ===
Average Accuracy: 98.22%
```

Fig. 6. 5 CROSSFOLD VALIDATION FOR TRAINING THE DATASET

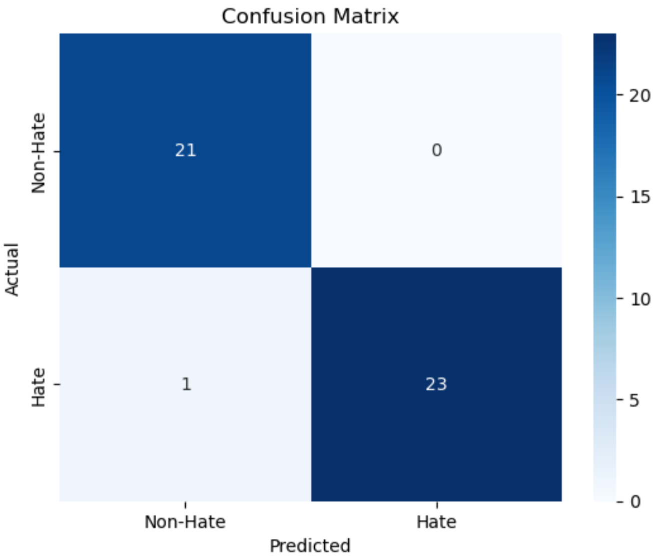


Fig. 7. 5 CROSSFOLD VALIDATION FOR TRAINING THE DATASET

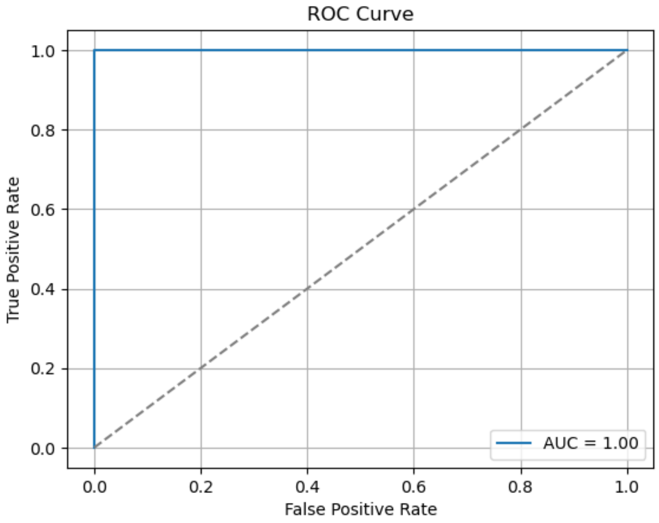


Fig. 8. ROC CURVE

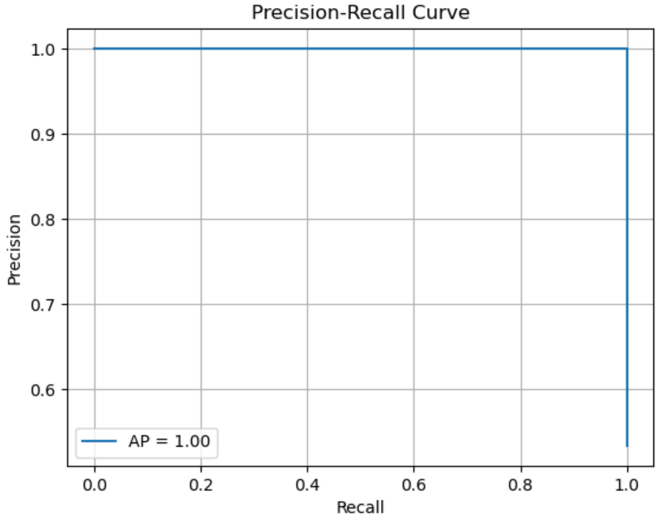


Fig. 9. RECALL CURVE

REFERENCES

[1] V. Gupta, R. Sharon, R. Sawhney and D. Mukherjee, ”ADIMA: Abuse Detection In Multilingual Audio,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6172-6176, doi: 10.1109/ICASSP43922.2022.9746718.

[2] K. Paval, V. Radhakrishnan, K. Krishnan, G. J. Lal and B. Premjith, ”Multimodal Fusion for Abusive Speech Detection Using Liquid Neural Networks and Convolution Neural Network,” 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10724438.

[3] Abusive Speech Detection in Telugu Using Vision Transformers

[4] Audio-Based Hate Speech Detection in Malayalam Using Machine Learning

[5] N. S. S. Reddy, V. V. A. Rohith, V. P. M. S. Reddy, Y. S. Reddy and J. L. G, ”Transfer Learning-Based Emotion Recognition Using Augmented Speech Data,” 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725674.

[6] Multilingual and Multimodal Abuse Detection Rini Sharon, Heet Shah, Debdoot Mukherjee, Vikram Gupta

[7] Abusive Speech Detection in Indic Languages Using Acoustic Feature

[8] Speech Emotion Recognition Using CNN-LSTM and Vision Transformer

[9] ”Multi-modal Hate Speech Detection using Machine Learning,”

[10] Speech Emotion Recognition Using CNN and LSTM