Audio-Based Hate Speech Detection in Malayalam Using Machine Learning

R V Gayathri Devi, J K Mahanivetha, P Seetharaman, K Devika and G Jyothish Lal

December 16, 2024

Audio-based hate speech detection in Malayalam using Machine learning

Gayathri Devi R V, Maha Nivetha JK, Seetharaman P, Devika K, Jyothish Lal G

Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India.

Contributing authors: cb.en.u4aie21112@cb.students.amrita.edu;
cb.en.u4aie21132@cb.students.amrita.edu;
cb.en.u4aie21164@cb.students.amrita.edu;
cb.en.u4aie21108@cb.students.amrita.edu;
g_jyothishlal@cb.amrita.edu;

Abstract

Detecting hate speech on social media is challenging, particularly in low-resourced languages like Malayalam, due to the scarcity of annotated data. To address this challenge, we introduce a new multiclass dataset for hate speech in the Malayalam language, sourced from YouTube. The study benchmarks the performance of machine learning classifiers for the classification of hate and non-hate speech, in both binary and multi-class classification tasks, using audio features alone. The Random Forest Classifier model performed exceptionally well in binary classification, achieving a macro accuracy of 0.93 and an F1 score of 0.93. Ablation studies conducted with other classifiers, such as Logistic Regression, Support Vector Machines, and Naive Bayes, registered accuracies around 0.85 and macro F1 scores of 0.85. In multiclass classification, the Random Forest model excelled with an accuracy of 0.8289, a macro accuracy of 0.72, and an F1 score of 0.74, outperforming all other models tested in the ablation study. These results demonstrate the effectiveness of the Random Forest Classifier in contributing to a safer online environment by reliably detecting hate speech in Malayalam.

Keywords: Ablation Analysis, Audio Feature Extraction, Hate speech detection, Multiclass Classification.

1 Introduction

Rise of the abusive language on social media platforms is a critical matter that has set an alarm off regarding the need for such strong mechanisms to detect and analyse the sentiment, especially in those languages like Malayalam that have only limited resources. This is a phenomenon that encompasses a wide range of behaviours, from cyber-bullying to the propagation of racial slurs and profanity, leading to unhealthy and unsafe communication environments [1]. In spite of the fact that there have already been made significant strides in hate speech detection algorithms, the majority of which are focused on text-based content, the domain of audio-based hate speech detection is still relatively unexplored [2]. The current benchmark datasets focus on text-based content. This limits the creation of all-encompassing algorithms that can stop hate speech across different formats such as audio and visuals. Tackling this issue brings both difficulties and chances. It highlights the pressing need to develop good detection methods to create safer and more welcoming social media spaces.

The study looks into analysing different types of social media data, with a focus on figuring out sentiments spotting abusive language, and pinpointing hate speech. The work by Premjith et al. [3] really stands out in this area. Leveraging a dataset derived from YouTube videos with transcripts and audio, methodologies encompassed a range of techniques including LSTM, K-means, KNN, logistic regression, TF-IDF features, Multinomial Naive Bayes, and Random Forest Classifier classifiers. The top performing team, Wit Hub, attained notable macro F1-scores: sentiment analysis (0.2444), abusive language detection (0.7143), and hate speech detection (0.2881), underscoring the efficacy of their approach in tackling challenges within the Dravidian languages' social media context.

Efforts to enhance automatic speech recognition (ASR) accuracy, especially for vulnerable individuals, have been demonstrated by Jairam et al. [4] and Bharathi et al. [5]. The LT-EDI@2024 dataset focused on Tamil conversational speech from vulnerable elderly and transgender individuals. Methodologically, state-of-the-art models such as Whisper and XLS-R were fine-tuned on this dataset, with the fine-tuned Whisper ASR model achieving a notable word error rate (WER) of 24.452, outperforming XLS-R. Such advancements contribute to broader inclusivity in ASR technology and facilitate better communication for diverse speakers, particularly in vulnerable populations. Recent studies have advanced hate speech detection and classification using various methodologies and datasets. Asogwa et al. [6] utilized a social media comment dataset and compared SVM and Naive Bayes classifiers, achieving an F1-score of 0.87 with SVM using unigram features and TF-IDF. Similarly, Abro et al. [7] evaluated SVM, Naive Bayes, and Random Forest Classifier classifiers on Twitter and Facebook comments, finding that SVM with TF-IDF features achieved the highest accuracy of 89%. Kurniawan and Budi [8] did a study on Indonesian tweets. They used BOW and TF-IDF with SVM, Naive Bayes, and Random Forest Classifier classifiers. SVM and TF-IDF gave them the best F1-score of 0.84772. These studies show that SVM together with TF-IDF works well for classifying hate speech.

Besides individual research, work on spotting hate speech across various languages and platforms has become more important. Prasad et al. [9] took on the task of finding hate speech in languages with few resources by using datasets from Twitter, Facebook, and YouTube. They used Model Agnostic Meta-Learning (MAML) along with the Cross lingual Language Model - RoBERTa (XLM-R). Their study got good accuracy scores between 0.80 and 0.88 doing better than usual fine-tuning methods. This showed how well it works when there's not much labelled data to use.

Additionally, Barman and Das [1] present a novel approach to multimodal sentiment analysis and abusive language detection in Dravidian languages using datasets sourced from YouTube videos. Integrating methodologies from computer vision, speech processing, and natural language processing, the study achieves a weighted average F1 score of 0.5786 for abusive language detection and weighted average F1 scores of 0.357 for Tamil and 0.233 for Malayalam for sentiment analysis.

Gupta et al. [10] introduce the ADIMA dataset, a novel multilingual audio dataset for abusive content detection, consisting of 11,775 samples across 10 Indic languages. Methodologically, it utilizes VGG and Wav2Vec2 models for feature extraction and explores various pooling and recurrent network architectures for classification. With accuracy ranging from 76.96% to 79.67% and macro F1 scores from 76.90% to 79.48%, the study demonstrates robust performance in abusive content detection across languages. Moreover, multimodal approaches have been instrumental in enhancing hate speech detection accuracy, as demonstrated in studies like [11], [12], and [13]. These studies leveraged diverse datasets and methodologies, integrating text-based features with audio-based features to achieve impressive accuracy rates and macro F1 scores, surpassing previous state-of-the-art techniques and highlighting the effectiveness of multimodal approaches in hate speech detection.

Bhesra et al. [2] introduce a novel hate speech dataset comprising both audio and text modalities. This dataset addresses a gap in understanding hate speech in audio content and consists of 600 samples, including hate and non-hate classes, covering diverse demographic entities. Machine learning classifiers and text encoders are utilized to evaluate hate speech detection performance on both modalities. The proposed multimodal hate detection algorithm achieves an accuracy of 80.5% ± 4.7%, surpassing the performance of using audio alone (79.0% ± 5.4%), highlighting the effectiveness of combining text and audio modalities for more accurate hate speech detection. Boishakhi et al. [14] use a dataset of 1051 videos from sites like YouTube and EMBY. Their study extracts separate features from image, audio, and text data using methods such as Recursive Feature Selection and Maximum Relevance - Minimum Redundancy. They tested seven traditional machine learning classifiers and used a hard voting ensemble method to make the final prediction. AdaBoost and Naive Bayes classifiers scored the highest accuracy at 87% and 75% beating individual modalities.

This study by Kshirsagar et al. [15] employs three datasets for hate speech detection: the Sexist/Racist (SR) data set, HATE dataset and HAR datasets. It can be noted that the F1 scores increase significantly when they make use of a neural network combined with SWEM architecture as well as pre-trained word embeddings. The F1 scores also move within large ranges: for instance, there was an increase of F1 scores from 0.74 to 0.86 in the case of SR dataset alone, from 0.90 to 0.924 in case of HATE datasets as well as from 0.170 to 0.319 when considering HAR datasets as well. Recent research endeavors have made significant strides in hate speech detection, sentiment analysis, and abusive language identification across diverse languages and modalities. These studies not only showcase the efficacy of their methodologies but also underscore the importance of inclusive datasets and robust methodologies in advancing the field and fostering a safer online environment.

Table 1: Comparison with existing approaches

Models	**Metrics**
mBERT, ViT, MFCC [1]	Weighted Avg F1: 0.5786 (abusive language), 0.357 (sentiment - Tamil), 0.233 (sentiment - Malayalam)
Machine learning classifiers, text encoders [2]	Accuracy: 80.5% ± 4.7%
LSTM, K-means, KNN, logistic regression, TF-IDF features, Multinomial NB, Random Forest Classifier classifiers [3]	Macro F1: 0.2444 (sentiment), 0.7143 (abusive language), 0.2881 (hate speech)
Fine-tuned Whisper ASR model, XLS-R [4]	Word Error Rate: 24.452
Pre-trained models (Whisper, transformer-based architectures) [5]	Word Error Rate: 24.452 (Team 1), 29.297 (Team 2), 37.7333 (Team 3)
SVM, NB [6]	Accuracy: 95.8%, 94.3%
SVM, NB, RF, LR [7]	Accuracy: 93%, 89%, 91%, 92%
SVM with TF-IDF unigram features [8]	F1-Score: 0.84772
Model Agnostic Meta-Learning (MAML), Cross-lingual Language Model - RoBERTa (XLM-R) [9]	Accuracy: 0.80 - 0.88
VGG, Wav2Vec2, pooling, recurrent network architectures [10]	Accuracy: 76.96% - 79.67%, Macro F1: 76.90% - 79.48%
LSTM, BILSTM, GRU, BIGRUt [11]	Accuracy: 88.15%
Multimodal learning, IEMOCAP dataset [12]	Precision: 93.00%
Transformer framework, "Attentive Fusion" layer [13]	Macro F1: 0.927
Recursive Feature Selection, Maximum Relevance - Minimum Redundancy, classical machine learning classifiers, hard voting ensemble method [14]	Accuracy: 87% (AdaBoost), 75% (Naive Bayes)
Pre-trained word embeddings, SWEM architecture [15]	F1 Score: 0.74 - 0.924
Random Forest Classifier (our model)	Macro Accuracy is 0.722 and the Macro F1 Score is 0.74

Most existing studies focus on widely spoken languages or languages with more resources. Our research focuses on languages with limited resources like Malayalam, filling a crucial gap in hate speech detection research. This novel focus highlights the importance of inclusivity in social media analysis and contributes to the understanding of hate speech dynamics in linguistically diverse contexts. With a large dataset comprising audio samples from social media platforms in Malayalam, our research contributes significantly to addressing the scarcity of annotated data, particularly in less-resourced languages.
A summary of the existing works and their metrics are mentioned in Table 1.

2 Methodology

As illustrated in Figure 1, this study's methodology includes data preparation, feature extraction, classifier training, and testing. The speech data sourced from YouTube is divided into training and testing sets. Specifically, 80% of the data is used for training, while the remaining 20% is reserved for testing. labeled as either hate or non-hate speech, with hate speech further categorized into four classes.Mel-frequency cepstral coefficients (MFCC) are extracted from the audio data to capture essential sound features. These MFCC features are then used to train a Random Forest Classifier to distinguish between hate and non-hate speech and classify specific hate categories. During testing, the model applies these learned patterns to new data.

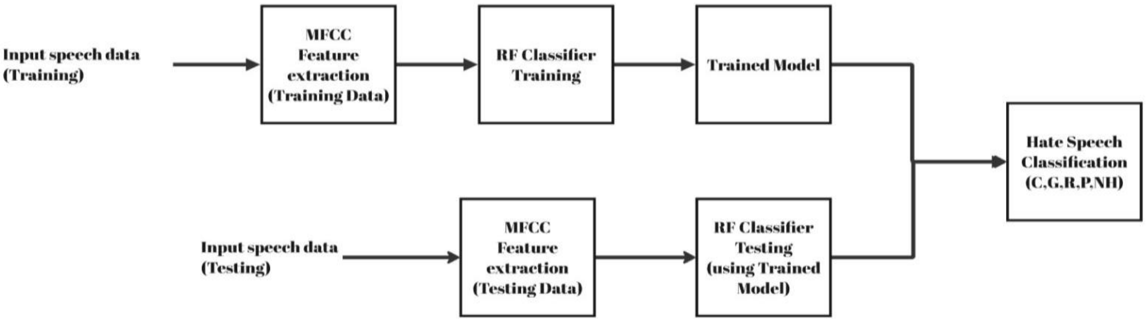
Input speech data (Training)
MFCC Feature extraction (Training Data)
RF Classifier Training
Trained Model
Hate Speech Classification (C,G,R,P,NH)
Input speech data (Testing)
MFCC Feature extraction (Testing Data)
RF Classifier Testing (using Trained Model)

Figure 1: Block Diagram - G: Gender P: Politics C: Personal Defamation R: Religion NH: Non-Hate

2.1 Dataset Description and annotation

The DravLangGuard dataset, a newly developed as part of the study, was curated to specifically address the challenges of detecting hate speech in Malayalam, a Dravidian language, on social media. This multimodal dataset includes both speech and corresponding text data sourced from YouTube videos [16]. In this study, we used only the speech dataset. Hate speech in the dataset is categorized into four main classes as defined by YouTube's hate speech policy: gender-based (based on sexual orientation), religion-based (based on religious comments), political/nationality-based (statements that degrade individuals or groups based on their nationality or political affiliations), and personal defamation (dehumanizing content that compares individuals or groups to animals, diseases, or pests), with an additional non-hate category for neutrality. Non-hate category is collected from generic videos and motivational talks.

Furthermore, speech samples were collected from YouTube channels with more than 50,000 subscribers. The videos were downloaded and converted to WAV format, with each sample ranging in duration from 2 to 49 seconds. Table 2 shows the class distribution for the collected dataset, with a breakdown of the number of samples for each hate category and the non-hate class. The dataset is almost balanced (44%) with 416 non-hate and 517 hate speech samples. Three native Malayalam-speaking annotators (2 male, 1 female) reviewed all the speech samples, following YouTube's hate speech policy guidelines. The inter-annotator agreement, measured by Cohen's Kappa measure, was 0.84, with final labels decided by majority vote in case of disagreements. Structure of File Names: LanguageCode: A two-letter code representing the language of the speech sample.ML for Malayalam

SpeechType: A two-letter code indicating whether the speech is hate speech or non-hate speech.

- HS for Hate Speech
- NH for Non-Hate Speech

HateCategory: A one-letter code representing the category of hate speech. This component is only present in hate speech files (HS).

- G for Gender-based hate speech
- R for Religion-based hate speech
- P for Nationality/Political hate speech
- C for personal defamation

SequenceNumber: A three-digit sequence number that uniquely identifies the file within its category.

Table 2: Class Distribution by Language

Language	**C**	**G**	**NH**	**P**	**R**
Malayalam	196	92	416	128	101

2.2 Feature Extraction

In this study, we employed Mel-frequency cepstral coefficients (MFCC) as the main feature extraction technique for processing speech data in the task of hate speech detection. MFCC is

a widely used feature in speech and audio processing because it effectively captures the spectral properties of audio signals, closely mimicking the human auditory system's response to sound. The process of MFCC feature extraction involves converting the time-domain audio signal into a frequency-domain representation using a Fourier transform. This is followed by mapping the resulting frequencies onto the Mel scale, which emphasizes the perceptually relevant aspects of sound—particularly those frequencies to which the human ear is most sensitive. The MFCCs are derived by taking the logarithm of the power spectrum on the Mel scale and then applying a discrete cosine transform (DCT) to decorrelate the coefficients. This process results in a compact representation of the spectral envelope, capturing the most important features that contribute to the perception of the speech content.

2.3 Classification and Performance evaluation

For each audio file in our dataset, we extracted and averaged MFCC features over a three-second duration, providing a concise yet informative feature set that encapsulates the phonetic and tonal characteristics of the speech. These extracted features were subsequently used to train the machine learning model, Random Forest classifier to accurately classify speech into hate and non-hate categories, as well as distinguish between multiple classes within hate-speech. The use of MFCC ensured that the models were fed with rich, discriminative features, which are crucial for achieving high classification accuracy in the context of speech-based hate speech detection. The model's performance is evaluated with accuracy and F1 score, ensuring effective hate speech detection in Malayalam social media.

3 Results and Discussion

We evaluated the performance of the proposed approach using standard measures such as accuracy and F1-score. The proposed method achieved macro accuracy and macro F1 scores equally around 0.93. As shown in Figure 2, the proposed approach is very good in differentiating the speech between "Hate" and "Non-Hate," having a great level of precision and recall for both the given classes. For multi-class classification the classifier showed high competency in detecting "Personal Defamation" and "Non-Hate" speech, however, performing average on other classes. Nevertheless, the macro-accuracy of 0.80, and a macro F1 score of 0.798 showed that the proposed approach sets a benchmark result on the dataset developed. This is significant in the sense that no other existing method perform a multi class classification of the hate-speech data.

Table *3* shows the comparison of accuracy and F1-scores for the binary and multiclass classification tasks in Malayalam using Random Forest Classifier.

Table 3: Comparison of Accuracy and F1-Score for Malayalam

Task	**Accuracy**	**F1-Score**
Binary class		
Hate	0.9854	0.93
Non hate	0.8724	0.92
Multi-class		
Personal defamation	0.8365	0.90
Gender	0.8178	0.80
Politics	0.7435	0.63
Religion	0.7847	0.73
Non hate	0.8543	0.93

Table 4: Model Performance Comparison for Tamil and Malayalam using MFCC

	Non-Hate	**Politics**	**Religion**	**Personal defamation**	**Gender**
Malayalam					
SVM	0.8377	0.6438	0.6435	0.7346	0.5761
Random Forest Classifier	0.8543	0.7435	0.7847	0.8365	0.8178
Logistic Regression	0.8232	0.70	0.6212	0.7523	0.5645
GaussianNB	0.80	0.6467	0.4458	0.7814	0.6455
Tamil					
SVM	0.6607	0.8751	0.6923	0.5263	0.9012
Random Forest Classifier	0.6607	0.6251	0.7692	0.3684	0.5031
Logistic Regression	0.5089	0.7501	0.5384	0.1578	0.5032
GaussianNB	0.4642	0.5001	0.6923	0.3157	0.1521

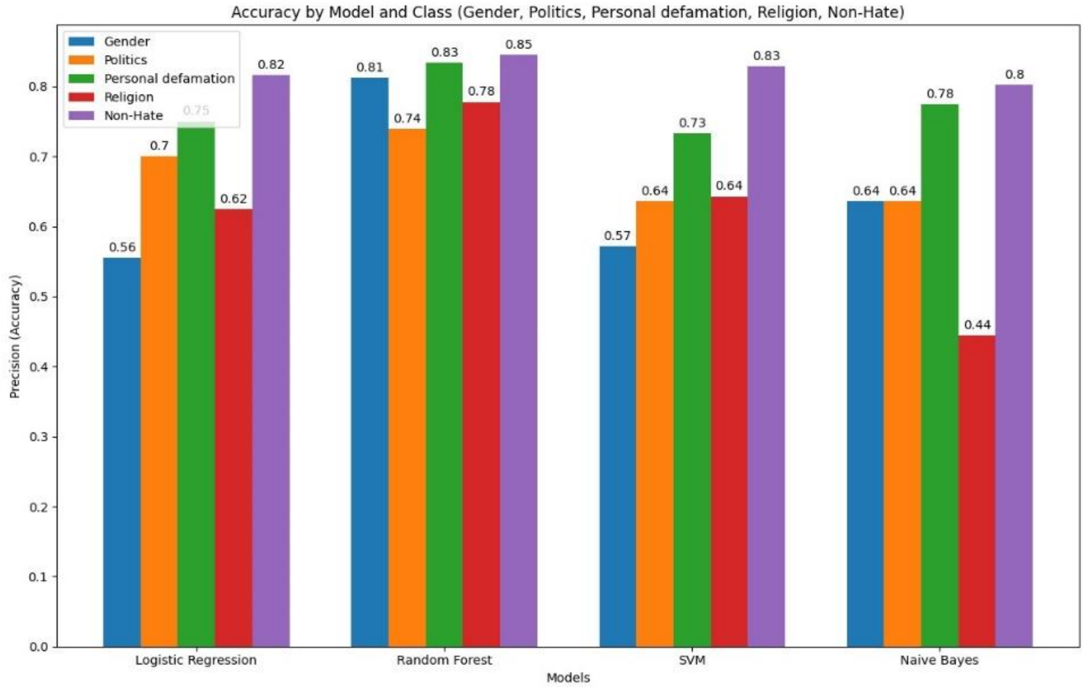

Figure 2: Multi-class results for Malayalam using mfcc feature

4 Ablation Study

Here, we examine the effects of various factors on the performance of our hate speech detection model via ablation studies. In particular, the focus is on three factors: (1) feature selection, (2) model (classifier) selection, and (3) training and testing language.

4.1 Impact of Feature Selection

Firstly, in the ablation study, we focused on more comprehensive feature representation. Precisely, besides MFCCs, we incorporated features from Mel Spectrogram and Chromograms.

- **Mel Spectrogram:** The Mel spectrogram represents how the energy in a speech signal is distributed over time into frequency bands. The mapping of these frequency bands onto the Mel scale—which approximates human auditory perception—becomes beneficial when applied to speech and music analysis, since it highlights the most important frequencies for perception.
- **Chroma Features:** Chromogram features complement this by focusing on the harmonic content of the audio, representing the intensity of pitch classes (such as the 12 semitones in a musical scale). This feature is valuable for identifying tonal and harmonic patterns, which are crucial in distinguishing different types of audio content.

This combined feature vectors were then used in the proposed method and as well as various machine learning models.

4.2 Effect of Classifier Choice

Our baseline model utilized a Random Forest Classifier. To determine the impact of classifier selection, we tested the model using alternative classifiers:

- **Logistic Regression:** Achieved 0.6232 macro accuracy, performing well in non-hate detection but showing lower precision for Religion-related content.
- **Random Forest Classifier:** Delivered the best macro accuracy at 0.72, particularly excelling in personal defamation and non-hate classification. The classification results for hate and non-hate speech in Malayalam, using the Random Forest Classifier model and others, are shown in Figure 3.
- **Support Vector Machine (SVM):** Achieved 0.6126 macro accuracy, with strong performance in non-hate detection but weaker results in Politics-related classification.
- **GausianNB:** Had the lowest macro accuracy at 0.61, performing better in Politics but struggling significantly with Religion-related hate detection.

These results demonstrate that Random Forest Classifier emerged as the most effective model. Table 4 compares the model performances across Tamil and Malayalam languages.

4.3 Variation in Language

To examine the proposed method's adaptability on a different language, we conducted experiments by training and testing the model on different language, Tamil. The dataset for the same is collected following a similar procedure as that of Malayalam. A total of 262 Hate and 297 Non-Hate speech was collected for Tamil language, containing 73, 43, 71 and 75 utterances in G, P, R and C categories as defined in dataset description.

Training and Testing in Tamil: When trained and tested on Tamil data using the Random Forest Classifier model, it achieved an accuracy of 0.5982. The model demonstrated strong performance in the "non-hate" class, with a precision of 0.66 and recall of 0.85, showcasing its effectiveness in recognizing non-hate content in Tamil. Additionally, the Random Forest Classifier model showed promising results across various other categories, reflecting its ability to adapt to the linguistic nuances within the Tamil language. This result highlights the model's effectiveness in processing and classifying Tamil language data, especially in identifying non-hate content accurately. Table 5 summarizes the model performances for Tamil and Malayalam languages, showing the accuracies and F1-scores achieved by different classifiers for both hate and non-hate speech detection tasks.

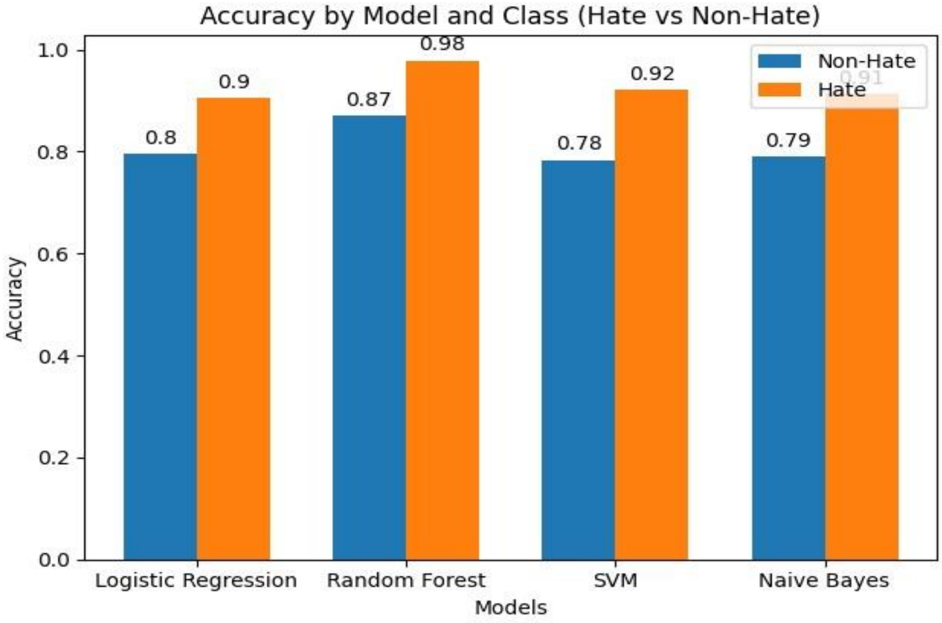

Figure 3: Binary classification for Malayalam

Table 5: Model Performance for Tamil and Malayalam Languages for Binary Classification using mfcc feature

Model	**Tamil**		**Malayalam**	
	Hate	**Non-Hate**	**Hate**	**Non-Hate**
	Accuracy			
SVM	0.7112	0.7835	0.9336	0.8663
Random Forest Classifier	0.7695	0.8557	0.9854	0.8724
Logistic Regression	0.7524	0.7565	0.9335	0.8662
GaussianNB	0.7313	0.7832	0.8675	0.8452
	F1-Score			
SVM	0.75	0.75	0.93	0.92
Random Forest Classifier	0.81	0.81	0.93	0.92
Logistic Regression	0.75	0.75	0.90	0.90
GaussianNB	0.76	0.76	0.85	0.85

5 Conclusion

This study introduces a novel approach to hate speech detection in Malayalam social media, addressing challenges in a low-resource language setting. Using the DravLangGuard dataset, this research evaluates multiple ML classifiers, with the Random Forest Classifier demonstrating superior performance in both binary and multiclass tasks. The use of audio-based features like MFCC proves effective in identifying hate speech categories, enhancing detection capabilities in linguistically diverse contexts. These findings underscore the potential of machine learning techniques to improve online safety for less-resourced languages. Future work should explore deeper integration of multimodal data and advanced models to further enhance detection accuracy across digital platforms.

References

[1] Barman, S., & Das, M. (2023, September). hate-alert@ dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (pp. 217-224).

[2] Bhesra, K., Shukla, S. A., & Agarwal, A. Audio vs. Text: Identify a Powerful Modality for Effective Hate Speech Detection. In The Second Tiny Papers Track at ICLR 2024.

[3] Premjith, B., Jyothish, G., Sowmya, V., Chakravarthi, B. R., Nandhini, K., Natarajan, R., ... & Reddy, M. (2024, March). Findings of the Shared Task on Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) @ DravidianLangTech 2024. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 56-61).

[4] Jairam, R., Jyothish, G., Premjith, B., & Viswa, M. (2024, March). CEN Amrita@ LT-EDI 2024: A Transformer based Speech Recognition System for Vulnerable Individuals in Tamil. In Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (pp. 190-195).

[5] Bharathi, B., Chakravarthi, B. R., Sripriya, N., Natarajan, R., & Suhasini, S. (2024, March). Overview of the third shared task on speech recognition for vulnerable individuals in tamil. In Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (pp. 133-138).

[6] Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2022). Hate speech classification using SVM and naive BAYES. arXiv preprint arXiv:2204.07057.

[7] Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications, 11(8).

[8] Kurniawan, S., & Budi, I. (2020, November). Indonesian tweets hate speech target classification using machine learning. In 2020 Fifth International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

[9] Prasad, D., Kadambari, K. V., Mukati, R., & Singariya, S. (2023, October). Real-Time Multi-Lingual Hate and Offensive Speech Detection in Social Networks Using Meta-Learning. In TENCON 2023-2023 IEEE Region 10 Conference (TENCON) (pp. 31-35). IEEE.

[10] Gupta, V., Sharon, R., Sawhney, R., & Mukherjee, D. (2022, May). Adima: Abuse detection in multilingual audio. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6172-6176). IEEE.

[11] Debele, A. G., & Woldeyohannis, M. M. (2022, November). Multimodal Amharic hate speech detection using deep learning. In 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 102-107). IEEE.

[12] Rana, A., & Jha, S. (2022). Emotion based hate speech detection using multimodal learning. arXiv preprint arXiv:2202.06218.

[13] Mandal, A., Roy, G., Barman, A., Dutta, I., & Naskar, S. K. (2024). Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection. arXiv preprint arXiv:2401.10653.

[14] Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2021, December). Multimodal hate speech detection using machine learning. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4496-4499). IEEE.

[15] Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018). Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644.

[16] Abhishek Anilkumar, Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, DravLangGuard: A Multimodal Approach for Hate Speech Detection in Dravidian Social Media, In: Speech and Language Technologies for Low-Resource Languages(SPELLL), Communications in Computer and Information Science (2024)