A Multimodal Approach for Hate and Offensive Content Detection in Tamil: From Corpus Creation to Model Development

JAYANTH MOHAN, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India
SPANDANA REDDY MEKAPATI, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India
PREMJITH B, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India
JYOTHISH LAL G, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India
BHARATHI RAJA CHAKRAVARTHI, Insight Centre for Data Analytics, National University of Ireland Galway, Galway, Ireland

Detecting hate speech on social media platforms is vital to mitigate technology-facilitated violence. Extensive research has targeted widely spoken languages like English, but there is a notable gap in studying hate speech detection in low-resource languages like Tamil. Additionally, with social media platforms now supporting various modalities, including text, speech, and video, effective techniques for hate speech detection in multimodal formats, especially videos, are crucial. However, detecting hate speech in Tamil presents unique challenges due to its morphology and code-mixing nature. This article presents a comprehensive approach for detecting hate speech in Tamil, with a focus on multimodal data. We introduce a new dataset, the MultimodAl Tamil Hate dataset, comprising videos along with their audio and textual transcripts, annotated with four categories of hate speech: offensive, sexist, racist, and casteist. To classify hate speech categories, we leverage transformer-based models. Through a series of experiments, we evaluated the performance of each modality individually and explored their fusion using a multimodal approach. BERT-based models were used for textual analysis to extract informative features, the TimeSformer model was employed for video modality, and Wav2Vec2 was used for audio modality. Specifically, we attained 81.82% accuracy and a 68.65% F1-score for the text modality, 63.63% accuracy and a 50.60% F1-score for the audio modality, and 45.45% accuracy and a 33.64% F1-score for the video modality. By integrating optimal combinations of models from each modality and employing machine learning classifiers, we achieved an accuracy of 81.82% and an F1-score of 66.67% in our hate speech classification task. Our research findings highlight the effectiveness of employing a

This study was supported through the AMRITA Seed Grant (Proposal ID: ASG2022119). B. R. Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under grant SFI/RC/2289_P2 (Insight_2).
Authors' Contact Information: Jayanth Mohan, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India; e-mail: jay.thinkai@gmail.com; Spandana Reddy Mekapati, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India; e-mail: cb.en.u4aie20038@cb.students.amrita.edu; Premjith B (Corresponding author), Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India; e-mail: prem.jb@gmail.com; Jyothish Lal G, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India; e-mail: g_jyothishlal@cb.amrita.edu; Bharathi Raja Chakravarthi, Insight Centre for Data Analytics, National University of Ireland Galway, Galway, Ireland; e-mail: bharathiraja.akr@gmail.com.

ACM 2375-4699/2025/03-ART22
https://doi.org/10.1145/3712260

multimodal approach for hate speech detection in Tamil, showcasing its efficacy in curbing the dissemination of hateful content on social media platforms.

CCS Concepts: • **Computing methodologies → Neural networks;**

Additional Key Words and Phrases: Multimodal, technology-facilitated violence, hate speech detection, transformer, corpus, low-resource languages, Tamil

ACM Reference Format:
Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A Multimodal Approach for Hate and Offensive Content Detection in Tamil: From Corpus Creation to Model Development. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 24, 3, Article 22 (March 2025), 24 pages. https://doi.org/10.1145/3712260

1 Introduction

Social media platforms have become popular mediums for individuals to share various types of content, including text, images, and videos. Users use these platforms to express their ideas and opinions freely. However, this freedom of expression has also led to the proliferation of harmful and violent content, such as **Hate and Offensive Language (HOL)**. HOL content on social media platforms aims to promote hatred or incite violence against individuals or organizations based on factors such as color, gender, ethnicity, race, sexuality, or political and religious affiliations. The dissemination of HOL content has severe repercussions, impacting the lives of individuals and society as a whole [2, 18, 43]. Detecting HOL content has emerged as a crucial task in mitigating the adverse effects of hate speech online. Researchers and experts are actively developing effective techniques to identify and combat HOL content across various languages. The majority of the existing research focuses on widely spoken languages like English, leaving a significant gap in comprehensive studies and resources dedicated to detecting HOL in low-resource languages such as Tamil. In addition, social media platforms allow users to create content using various modalities like text, speech, and video. Video content, in particular, is considered multimodal as it combines text, speech, and a sequence of images [4]. With the exponential growth of user-generated content, especially in multimodal formats, shared online, it is crucial to prioritize developing effective techniques to detect and prevent the spread of hate content accurately. However, it is challenging for various reasons. First, extracting features from diverse modalities and training machine learning models to learn underlying patterns poses difficulties. Additionally, varying video quality and length across social media platforms impact feature extraction from different modalities, affecting applications differently. Furthermore, discerning various types of hate content is more intricate than binary classification due to nuanced language variations and the use of polysemantic words. This complexity increases in the case of low-resource languages such as Tamil, a Dravidian language, which is rich in morphology [37], intensifying the difficulty in detecting different types of hate speech from code-mixed text. As a result, the availability of gold-standard annotated corpora for multimodal and unimodal hate speech detection is minimal for Tamil and other Dravidian languages.

This article proposes a corpus for detecting different types of HOL content—racist, sexist, offensive, and casteist—from multimodal video data in Tamil. Furthermore, we conducted different experiments using transformer-based feature extraction approaches and machine learning algorithms to classify Tamil video data into four different HOL categories. TimeSformer, Wav2Vec2, and various BERT (Bidirectional Encoder Representations from Transformers)-based pre-trained models were used for extracting features from video, speech and text modes, respectively. In addition, we carried out experiments to identify the modality that contributes more to the detection

of HOL categories from the data by taking each modality separately for the detection task. The experiments show that the text modality is the deciding factor for categorizing video data into different classes. For the speech modality, we considered both noised (or raw) data and denoised data, and the experiments show that classification using denoised speech features exhibits better accuracy. The spectral subtraction algorithm is used for denoising the speech signal, which yielded good results in removing noise. To our knowledge, this is the first reported corpus for multimodal HOL classification in Tamil and the first set of machine learning models for this task. We present the following significant contributions:

— A multimodal corpus of HOL content in Tamil language.
— Classification models for detecting four categories of HOL content by leveraging transformer-based feature extraction.
— Identification of the modality that best classifies the HOL content into different categories.

The rest of the article is organized as follows. Section 2 discusses related articles where similar work has been proposed in the literature. Section 3 provides a brief overview of Dravidian languages, highlighting the unique challenges they pose for hate speech and offensive language (HOL) detection and the limited availability of annotated corpora. Section 4 explains the data collection process used in the experiments, including the sources and criteria for selecting the samples. Section 5 details the annotation process for labeling the data with HOL categories, including the guidelines followed and the involvement of annotators. Section 7 presents the experimental setup and methodology used for HOL detection in Tamil, discussing the results obtained from different models and feature extraction approaches while also highlighting the strengths and limitations of each. Section 8 assesses the proposed methodology's robustness via a comprehensive evaluation strategy, focusing on model generalization and overfitting. Finally, Section 9 highlights the effectiveness of our multimodal approach and suggests future research directions for enhancing hate speech detection in low-resource languages like Tamil.

2 Related Work

Understanding human behavior, emotions, opinions, and habits necessitates applying various techniques, including opinion mining, text analysis, sentiment analysis, and facial expression recognition. In recent years, there has been a growing interest in multimodal sentiment analysis owing to its ability to scrutinize verbal and non-verbal behavior, making it a prominent area of research in natural language processing. Multimodal analysis offers a more precise detection of human emotions and feelings compared to analyzing only text. Consequently, several studies have concentrated on developing fusion networks that integrate multimodal representations in these analyses. Researchers have generated numerous datasets in English to explore multimodal sentiment analysis and emotion identification.

The MOSI dataset, introduced by Zadeh et al. [54], is a comprehensive resource for sentiment analysis in English. With more than 2,000 YouTube videos covering diverse topics like movie reviews and political speeches, it offers annotations for sentiment intensity, valence, and arousal, alongside textual and audio transcriptions and visual features such as facial expressions and body language. This dataset is notably challenging due to its broad spectrum of sentiment intensities and nuanced expressions across text, audio, and visual modalities. MOSI has been utilized for sentiment classification, regression, and ranking tasks. Liang et al. [31] utilized this dataset and proposed an RMFN (Recurrent Multistage Fusion Network) model, achieving an F1-score of 0.70 on this dataset.

Similarly, Busso et al. [9] introduced the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset to facilitate emotion recognition research in English. This dataset comprises audio and video recordings capturing naturalistic conversations between pairs of speakers. These

interactions are designed to simulate real-life scenarios, which involve 10 actors engaging in conversations with a trained actor assuming the role of a therapist. The dataset includes annotations of emotional states, facial expressions, body movements, and speech features, spanning four sessions, each lasting approximately 1 hour. Consisting of 151 recorded dialogues, with 302 speakers across five sessions, each segment is annotated for the presence of nine emotions (Angry, Excited, Fear, Sad, Surprised, Frustrated, Happy, Disappointed, and Neutral). Tripathi et al. [45] proposed a deep learning based approach for multimodal emotion recognition using this IEMOCAP dataset. Their method combines speech, text, and facial expression features using a **Convolutional Neural Network (CNN)** for visual input and a **Long Short-Term Memory (LSTM)** network for speech and text input separately. By fusing outputs from these networks, they achieve an accuracy of 82.08% on the emotion recognition task. Poria et al. [35] introduced the MELD dataset, specifically curated to advance research in multimodal emotion recognition in English. This dataset includes naturalistic conversations among three speakers with detailed emotional annotations for each throughout the dialogue. This dataset consists of 1,424 utterances from 634 videos recorded in a laboratory setting to simulate real-life interactions. Notably, a significant portion of the dataset comprises dialogues and utterances from the *Friends* TV series, with annotations covering seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. In a subsequent study, Song et al. [42] leveraged this dataset to propose a novel approach called *SPCL* (Supervised Prototypical Contrastive Learning) for conversational emotion recognition. Their method involves learning prototypical representations for each emotion class and using contrastive learning to enhance the discriminative capabilities of these representations across different emotions. Combining classification and contrastive learning objectives achieved state-of-the-art results on emotion recognition tasks, yielding an accuracy of 73.2%. The paper by Zadeh et al. [56] introduces the CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) dataset. This dataset comprises videos featuring individuals in diverse natural settings, totaling more than 23,000 video segments and 40,000 sentences with sentiment and emotion annotations. The authors also propose IDFG, an interpretable dynamic fusion graph model, which is a graph-based approach that dynamically integrates audio, visual, and textual information to predict speaker sentiment and emotion. The IDFG model achieves state-of-the-art results on sentiment analysis and emotion recognition tasks and offers interpretable explanations for its predictions.

Hasan et al. [22] introduced the UR-FUNNY dataset, aimed at enhancing the researcher's understanding of humor and facilitating the development of computational models for humor recognition in English. This dataset comprises short comedic video clips accompanied by transcripts, audio tracks, and visual features like facial expressions and body gestures. Additionally, they propose a multimodal fusion approach for humor recognition by integrating information from text, audio, and visual features. Their approach demonstrates superior performance compared to other methods on the UR-FUNNY dataset, underscoring the importance of incorporating multiple modalities in humor recognition tasks.

2.1 Datasets for Dravidian Languages

Recent studies focusing on Dravidian and low-resource languages (e.g., [10]) concentrate on detecting offensive language in code-mixed comments/posts. Chakravarthi et al. [10] introduced a new dataset for Dravidian languages, specifically Tamil and Malayalam, sourced from social media platforms. The dataset comprises 6,534 Tamil comments, 5,951 Malayalam comments, and 5,941 code-mixed Malayalam-English and Tamil-English comments. Roy et al. [39] tackled this challenge using an ensemble framework integrating CNNs, LSTM networks, and transformer models like BERT and MuRIL (Multilingual Representations for Indian Languages). The effectiveness of this ensemble approach is compared with traditional machine learning algorithms utilizing TF-IDF

for feature vector generation. Chakravarthi et al. [11] also introduced a dataset for identifying offensive language in code-mixed comments/posts from Dravidian languages (Malayalam-English and Tamil-English) sourced from social media. This dataset comprised 4,000 samples categorized into offensive and non-offensive classes. Subramanian et al. [44] address this challenge, focusing specifically on Tamil language comments by employing traditional machine learning models such as Bernoulli Naïve Bayes, **Support Vector Machine (SVM)**, Logistic Regression, and K-Nearest Neighbor, alongside state-of-the-art transformer-based models like mBERT, MuRIL, and XLM-RoBERTa for offensive language detection. Additionally, they leveraged the concept of adapters [32] to fine-tune mBERT, XLM-RoBERTa, and MuRIL on a smaller dataset of offensive language detection, achieving state-of-the-art performance. There are very few works like that of Chakravarthi et al. [12] where a multimodal dataset is explicitly introduced for Tamil and Malayalam languages, focusing on sentiment analysis. This dataset is sourced from YouTube and consists of 134 product or movie review videos, divided into 70 in Malayalam and 64 in Tamil. Each video is annotated with sentiment labels (Highly Positive, Positive, Neutral, Negative, and Highly Negative), making it ideal for classification tasks.

To our knowledge, no multimodal annotated video dataset is available for hate speech analysis in Tamil, even though numerous datasets are available for English and other languages. Therefore, we developed a hate speech dataset exclusively for the Tamil language called the **MultimodAl Tamil Hate (MATH)** dataset; the dataset consists of videos along with audio and text transcripts. For the current study, a total of 52 videos were curated. The approach for hate speech detection followed in this work is similar to the one adopted by Wöllmer et al. [50]. In our study, we have annotated the hate category at the video clip level. Despite the smaller size of our dataset as compared to some of the previous datasets for other languages, we emphasize that it still has significant value on its own, given that volunteer annotators have annotated it. We anticipate that the findings of this study will stimulate further research on multimodal analysis in Tamil and other low-resource languages. Our study's findings are expected to spur additional research on multimodal analysis in Tamil and other low-resource languages. Despite our dataset's modest size, it is our hope that our work will inspire researchers to pursue similar analyses in Tamil and other low-resource languages. Table 1 presents a comparative study of different datasets and the models employed, highlighting areas for potential improvement.

3 Dravidian Languages

Dravidian languages are a group of languages primarily spoken in South India and surrounding regions. The term *Dravidian* was first used by Robert Caldwell [41], and it refers to the prominent languages of South India, including Tamil, Kannada, Telugu, and Malayalam, which are also the official languages of Tamil Nadu, Karnataka, Andhra Pradesh, and Kerala, respectively. Among the Dravidian languages, Tamil holds a special place as the oldest living language in the family. It is spoken in India, Sri Lanka, Singapore, Malaysia, South Africa, and the Tamil diaspora. With more than 89.7 million speakers [49], Tamil is one of the official languages of Tamil Nadu, Puducherry, India, Sri Lanka, and Singapore. Tamil is considered a *diglossic* language, meaning that its written and spoken forms can be different. The government of Tamil Nadu, India, and Sri Lanka standardizes the written form of present-day Tamil. There are various spoken forms of Tamil spoken across different regions, each having its own dialects and accents. These include the Central Tamil dialect, Kongu Tamil, Chennai Tamil, Madurai Tamil, Nellai Tamil, Kumari Tamil, and Palakkad Tamil in India, as well as the Batticaloa Tamil, Jaffna Tamil, and Negombo Tamil in Sri Lanka. Other Dravidian languages, such as Kannada, have also been influenced by Tamil, which has heavily influenced the Sankethi Tamil dialect in Karnataka. Additionally, the people of Lakshadweep speak a different

Table 1. Comparison of Recent Works

Author	Dataset	Models Used	Scope of Improvement
[20]	MMHS150K	Inception V3 for images and LSTM for text are concatenated, followed by fully connected (FC) layers.	- A small set of multimodal examples is used; it can be increased - Authors could use a more diverse set of models for evaluation; the current study only considers three models
[7]	New dataset sourced from YouTube and EMBY	Energy, zero crossing rate, spectral centroid, and cepstral coefficients for audio, TF-IDF for text, and pixel representation for videos followed by machine learning algorithms for classification	- Feature extraction methods can be improved
[34]	Sourced from Twitter	ResNet for images and BERT for text with early-fusion architecture followed by a feedforward network	- Only ResNet and BERT models are used to extract features; other models could have been assessed
[38]	IEMOCAP and HSDVD (Hate Speech Detection Video Dataset)	BERT and ALBERT for text, and energy, spectral, and MFCC for audio followed by fusion and multi-layer perceptron for classification	- The authors have utilized the audio and text modalities for classification, but the inclusion of video modality could have improved performance of the model
[26]	Bengali Hate Speech, sourced from Facebook, newspaper articles, and YouTube	Efficient Net for images and BERT for text followed by fusion and FC layers for classification	- Increase in size of the dataset - Explore other ways to improve performance by utilizing transformer-based image feature extraction
[29]	Facebook Hateful Memes	Vision-language models like ViLBERT and Visual BERT	- Could have improved the annotation to include diverse classes
[40]	MMHS150K	BERT for text, and Inception V3, Inception, and ResNet for image	- Make their code publicly available for reproducibility
[52]	New dataset sourced from Facebook	Convolutional text model and CNN-based model for image followed by different fusion approaches	- Could have made the dataset used open source
[15]	Twitter	Pretrained BERT and Convolutional Graph Neural Networks for text, and VGG for images	- The size of the dataset is small, which can be improved to increase accuracy
[51]	Multimodal Hate Speech and Sarcasm	Vision-language models like ViLBERT and Visual BERT followed by Domain adaptation module	- The number of classes is less, making it less robust
[36]	MMHS150K	BERT for text, and Capsule Network and EfficientNet for text and image	- Other feature extraction models and the use of multi-head attention is not explored
[16]	Amharic Hate Speech Detection	word2vec for text and MFCC for audio followed by LSTM, BI-LSTM, GRU, and BI-GRU for classification	- The dataset is not open sourced for further research
[8]	MMHS150K	Context-Invariant LSTM and RoBERTa-large for text, and Xception, NASNet, and Inception-ResNet V2 for image	- The paper did not discuss potential overfitting concerns, as it does not evaluate the model's performance on external datasets
[28]	Facebook Hateful Memes	LSTM, Bidirectional LSTM for text, and vgg16 for image followed by fusion for classification	- The models are not able to identify all offensive memes, and transformer-based approaches for both modalities are not explored
[5]	Abusive Comment Detection (Tamil-ACL 2022)	(1) n-gram for text followed by multi-layer perceptron for classification (2) 1D Convolutional LSTM for classification	- Only two approaches are explored; other methods involving transformer-based feature extraction can be explored
[17]	FIRE 2020—HASOC	RNN, LSTM, GRU, BI-LSTM, Hierarchical Attention Networks, and BI-GRU	- Transformer-based models are unexplored, and the architecture is only for text modality
[3]	Facebook Hateful Memes	Contrastive language-image pretraining	- The model used (CLIP) can be fine-tuned to the used dataset for improved accuracy
[48]	CrisisHateMM	RoBERTa for text and Swin Transformers for images followed by multi-layer perceptron fusion	- The performance can be compared with other feature extraction methods like ConvNeXts, Dino V2 for images, and XLNet for text

dialect called *Jeseri*. Understanding the various dialects and accents spoken in different regions is crucial for studying the Tamil language.

4 Data Collection

The data collection starts by selecting Tamil language videos with a focus on hate speech and offensive content found on YouTube. Our aim is to build a diverse corpus for multimodal HOL detection in Tamil, covering different dialects and topics worldwide. This presents a significant challenge in discerning different manifestations of hate speech, compounded by the inherent complexities of the language. Figures 1 and 2 are examples of the types of content we are considering for inclusion in the dataset. We used a two-step process to create multimodal data. First, we extracted speech from videos and transcribed it into text, generating a text modality that captures the spoken content. To further analyze the text, Figure 3 presents a word cloud of the text modality, offering a visual representation of the most frequent terms. This resulted in a dataset featuring three modalities: text, speech, and video. The next step involves annotating the collected data, with studies [24, 46] focusing on four categories of hate speech: offensive, racist, sexist, and casteist. These labels were assigned using a 4-point scale to evaluate the intensity of hate speech in each video comprehensively. We selected these specific categories to create a focused dataset for HOL detection, as they are prevalent in online platforms and have significant societal impacts [46]. By concentrating on these labels, the dataset offers insights into various dimensions of hate speech, facilitating more targeted analysis and classification of HOL content. Our dataset comprises 52 videos collected from

Fig. 1. First example image.

Fig. 2. Second example image.

Fig. 3. Word cloud of text modality.

YouTube, sourced using various downloading programs to ensure reasonable resolution for analysis. The videos range from 192 to 720 pixels, offering diverse visual content. Table 3 illustrates the distribution of video durations, enhancing the dataset's richness and enabling comprehensive exploration of HOL content across various temporal contexts.

4.1 Acquisition of the Videos for Developing the Dataset

During acquisition, the primary challenge was identifying videos in the four specified categories. We carefully reviewed content shared by various YouTubers, focusing solely on hate speech. After scrutinizing each video, we formed a dataset exclusively containing offensive, racist, sexist, and casteist content. The selection criteria for MATH videos are listed next:

— *Duration of the video*: We fixed the minimum length of the video as 1 minute and the maximum as 3 minutes as shown in Table 2. Therefore, we can ensure that the selected videos clearly explain their HOL characteristics. To avoid a disparity in the data length, we decided to restrict the maximum length of the video to 3 minutes.
— *Presence of hate content*: The other challenge we faced was the presence of multiple instances of hateful content in a single video. To reduce complexity, we listened to the utterances

Table 2. Distribution of Video Durations in the Dataset

Duration of Video	Count of Video
Count ≤ 1	3
Count > 1 and <1.5	45
Count > 1.5 and <2	2
Count < 1	2
Mean	1.207

Table 3. Distribution of Class Labels in the Dataset

Various Class Labels	Count of Each Label
Caste	15
Offensive	17
Racist	16
Sexist	4

corresponding to each video, split them into multiple segments, and selected the most relevant portion containing proper HOL information.

— *Annotation of hate content*: The dataset was prepared by considering only those videos with hate content. Videos were carefully selected to curate the dataset with accurate hate content. It was manually annotated and labeled based on the content of the video.

There is a scarcity of research work examining hate content in Tamil. This scarcity poses challenges in terms of identifying suitable videos and annotating them. Some of the limitations and difficulties encountered during the selection and annotation process are discussed next:

— *Code-mixed Tamil*: Users often code-mixed Tamil with languages like English or other regional languages while speaking or writing, which poses a challenge in hate speech detection algorithms and language processing models. Code-mixed content may contain a blend of different languages, making it difficult to accurately classify hate speech or determine the context solely based on the Tamil component.

— *Quality of audio*: The audio quality in YouTube videos can vary significantly, with background noise, low volume, or distorted audio, making it difficult to extract and transcribe the speech accurately. A low-quality audio can result in poor transcription when automated speech-to-text models are used.

— *Limited caption availability*: The subtitles for Tamil YouTube videos are rare. Besides, transcribing Tamil audio extracted from videos using automated tools does not generally provide accurate, adequate, and fluent text.

— *Ambiguous context*: Capturing the proper context using only one modality is challenging, as it may not convey the intended sense of the input. Therefore, incorporating features extracted from multiple modalities, such as visual cues and gestures from the video data, along with the tone, pitch of the speech, and the word meaning of the tokens in the transcripts, is essential to properly distinguish various HOL video contents.

— *Dialect variations*: YouTube videos in Tamil may contain multiple dialects, making it challenging to accurately capture the diverse linguistic nuances and variations in the dataset.

Ultimately, only videos that were devoid of the aforementioned challenges were included in the dataset. Given the intricacy of the underlying distribution, having diverse training samples is crucial for comprehensive multimodal language analyses.

4.2 How Did We Overcome the Challenges?

The difficulties we encountered in gathering the data at various phases are explained next, along with the steps we took to address those difficulties.

4.2.1 Selection of the Videos. The initial hurdle encountered during dataset preparation was related to video selection. Only a small percentage of Tamil videos posted on YouTube contained hate content. Consequently, we began by searching for specific topics, assuming that videos about those topics might contain hate speech. However, some videos representing the identified hate

categories were relatively short. We carefully viewed all videos and selected only those at least 1 minute long. Another challenge we faced was the quality of the videos, encompassing both visual and auditory aspects. Videos with resolutions below 192 pixels and those with poor perceptual audio quality were excluded from the dataset. Factors such as the quality of camera lenses, lighting, camera placement, and upload quality affected the visual resolution. At the same time, the absence of a high-quality external microphone impacted the sound quality. Background noise present in the videos was also a crucial consideration. Videos with background soundtracks that degraded the audio quality were eliminated from the dataset. Moreover, we ensured that we downloaded the best-quality versions of these videos, meeting the aforementioned criteria. Overall, these challenges required us to carefully select videos of appropriate length, high visual resolution, and good audio quality, thereby ensuring the quality and reliability of the dataset.

4.2.2 Preparation of the Transcripts. The most difficult and time-consuming process in the dataset preparation was the generation of transcripts. This process involved a significant amount of human assistance. Initially, we transcribed the collected videos using IBM speech-to-text [23] and Google speech-to-text [21] models. Even though the preceding models are the benchmark models, the results were unsatisfactory because of the accent, poor pronunciation, lack of speech clarity, and awkward sentence structure. Therefore, we manually compiled the transcripts. The other challenges encountered while creating the transcripts are listed next:

— Users speak quickly, often ending sentences without completing them, which makes it challenging to understand.
— Slip of the tongue, regional slang, and ignorance of the correct pronunciation might lead to unclear utterances.
— Identification of the position of the punctuation was another challenge. The beginning and end of sentences were unclear in some of the cases. Besides, speakers rarely follow formal sentence structures to create a sentence. Punctuation was placed in the transcripts after listening to the audio carefully multiple times.
— The use of different dialects caused another problem. The annotators unfamiliar with the dialects may not understand the uttered words and hence can result in wrong spellings. Therefore, we sought assistance from people familiar with the dialects to comprehend the words and spelling. The dialect issues also had an impact on the annotating process. The majority of words in Tamil have many meanings. They can vary depending on the context, cultural background, and personal perspectives of different individuals and depending on the situation.
— Sometimes the meanings are reversed as well. Neologisms made this situation more complicated. Today, words that are typically used to indicate negative feelings are frequently employed to express good ones. If the annotator is unfamiliar with such usages in a particular dialect, it leads to ambiguity while annotating the video. Therefore, we had a team of annotators to rectify it.

5 Annotation Process

Hate speech annotation is the final step of dataset development. It is also a complicated job, as it requires the opinions of multiple annotators to finalize the labels for each data. In this work, we annotated the data into four categories: offensive, sexist, racist, and caste.

5.1 Annotation Setup

Zadeh et al. [55] and Mohammad [33] proposed annotation procedures for video clips. Using this procedure, three annotators independently annotated each video clip, considering the multimodal

nature of the data. The annotation process equally addressed the video, audio, and text modalities, as comprehensive analysis required considering all three. Facial expressions of individuals in the videos played a crucial role in determining the type of hate content; thus, each unique facial expression was taken into account during annotation. Vocal modulations and the transcript of the vlogger's speech were also utilized to provide additional clues for annotation. Notably, facial expressions and tone of speech were particularly useful in distinguishing between different types of hate content in the video clips, especially when sarcastic comments were employed by the speaker. The annotation schema used for this task is provided next:

- *Offensive*: A video clip can be annotated as offensive if the person uses Offensive language or offensive remarks with a Disrespectful or Insensitive facial expression. Additionally, if the person exhibits a Mocking or Provocative tone along with a Smirking or Disapproving facial expression, it may indicate the presence of offensive content. These indicators suggest content that is likely to cause distress, discomfort, or harm to individuals or groups.
- *Sexist*: A video clip can be annotated as sexist if the person uses Sexist language or derogatory remarks with a Condescending or Dismissive facial expression. Additionally, if the person exhibits a Patronizing or Objectifying tone along with a Smirking or Mocking facial expression, it may indicate sexist attitudes or behaviors. These indicators suggest a biased or discriminatory view toward individuals or groups based on their gender.
- *Racist*: A video clip can be annotated as racist if the person uses Racial slurs or derogatory language with a Contemptuous or Disparaging facial expression. Furthermore, if the person exhibits a superior or hateful tone along with a Hostile or Angry facial expression, it may indicate racist attitudes or beliefs. These indicators suggest a biased or discriminatory perspective toward individuals or groups based on their race.
- *Caste*: A video clip can be annotated as casteist if the person uses Caste-based slurs or derogatory language with a Dismissive facial expression. Additionally, if the person exhibits a superior or Condescending tone along with a Contemptuous or Scornful facial expression, it may also indicate caste-based discrimination. These indicators suggest a bias or discriminatory attitude toward individuals or groups based on their caste.

5.2 Annotators

The annotation team consists of three people. All three annotators are postgraduates and have full native language proficiency in Tamil. In addition, they undertook Tamil as their second language in their schooling. The annotators had good fluency in Tamil, combined with their postgraduate qualifications, enabling them to comprehend the nuances and subtleties of the language. They can effectively discern the type of hate content expressed in the multimodal (MATH) dataset, considering the cultural and contextual elements unique to the Tamil language. By leveraging their linguistic expertise and educational background, these annotators contribute to accurately analyzing and interpreting the type of hate content. Separate Google Sheets were provided to each annotator fostering a conducive work environment. This separation helps minimize the risk of unintentional bias or undue influence on the annotation process. It allows each annotator to independently evaluate the dataset, ensuring a more objective and unbiased annotation process. Labeling was done manually by considering the following aspects after observing the video clips.

5.3 Words and Sentences

The choice of words in a phrase is pivotal in discerning the existence of offensive and discriminatory content. Specific words and phrases can serve as indicators of hate speech, and their presence within a text can aid in identifying the underlying sentiment. Here are a few illustrations:

— *Sexist*: Phrases like "She's such a tomboy," "Sow your wild oats," "She looks curvy," and "That's so gay" objectifying men/women and discriminating against genders are signals of sexist hate speech.
— *Racist*: Expressions such as racial slurs and derogatory remarks targeting specific ethnicities or races like "nigger," "negro," "black ass," and "nigga" signify racist hate speech.
— *Casteist*: Phrases like "untouchables should be kept separate," "Chappri," "We don't associate with people from that caste," "Inter-caste marriages are unacceptable," and "Are you gonna marry an SC girl?" indicate casteist hate speech.
— *Offensive*: Words or phrases that are profane, vulgar, derogatory, or target a person or group based on their identity, such as "hate speech," "bigoted remarks," "homophobic slurs," "religious insults," and "disability mockery," reflect offensive hate speech.

It is quintessential to analyze the context, intent, and impact of the words used to detect hate speech. Here are some instances of terms in Tamil that are sexist, racist, offensive, and caste:

— *Sexist*: "Poda Potta" (a derogatory term stating men as women), "ponnunga kuninja thala nimirama nadakanum" (gender-based term stating that walking without lifting their heads is important for a girl), and "Dupatta podunga doli" (policing of women's clothing choices).
— *Racist*: "Kakka mari dhane iruka" (a term discriminating based on color by mentioning a crow's color) and "poda karuva payale" (a term teasing a pale-skinned person).
— *Offensive*: "Yelavu edutha Sirukki"(a curse word used on a girl) and "otha kiruku koothi" (a curse word mentioning one's private parts).
— *Casteist*: "Jathi veri" (a term referring to caste), "namba jathi ponna dhan nee katikinu" (a term promoting same-caste marriage), and "avangalam palu para kitaye pogatha" (a phrase referring to lower-caste people as untouchables).

5.4 Attention to Detail

Sometimes videos may not contain explicit hate words or phrases, making it necessary for the annotator to assess the content manually. In such cases, the annotator should consider factors like tone, emotional expression, and the sequence of words used to identify the underlying sentiment of the video. By carefully examining the tone of voice, facial expressions, and body language of the individuals in the video, the annotator can gain insights into the conveyed intention. They should pay attention to the emotional cues exhibited by the speakers and analyze the overall context in which the words are spoken. Additionally, the annotator should consider the sequence of words used and their implications. Even if no explicit hate speech is present, certain combinations of words or phrases can be offensive. Analyzing the language and implied meaning behind the words is crucial to label the video accurately. In addition, the video's overall message should be considered by the annotator, as the context, theme, and purpose of the content can also provide clues about the type of hate content present. Considering all of these factors, the annotator can effectively identify the type of hate in the video, even if it does not contain direct hate words or phrases.

5.5 Usage of Words in the Speech

The ambiguity of words is a universal problem in language. People may use words with opposite meanings in different contexts in regional languages such as Tamil. Vernacularism or dialects in these languages need to be noted to avoid such ambiguity. Apart from dialects, words used among a particular age group also create ambiguity. Some examples include 'Bhayangaram' (fearful and awesome), and 'malai' (evening and garland) has different meanings based on the tone in which it is conveyed.

Table 4. Detailed Explanation of the Different Ranges of Kappa Scores

Kappa Score	Definition
<0.00	Poor agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect or perfect agreement

5.6 Inter-Annotator Agreement

The inter-annotator agreement is used to compute the agreement between the annotators for labeling data based on the annotation guidelines. In this work, we used the kappa score proposed by Landis and Koch [30] for computing the agreement score. The equation is given next:

$$k = \frac{P(A) - P(E))}{1 - P(E)}. \tag{1}$$

$P(A)$ and $P(E)$ are the probability of all three annotators agreeing on the same score and the likelihood of the predicted agreement to occur, respectively. Table 4 defines the agreement score value. In this work, we obtained a kappa score of 0.7182 for annotating the videos in Tamil, which shows a substantial agreement between the annotators.

6 Ablation Study

To thoroughly evaluate the dataset's performance, we conducted separate experiments for each modality (text, audio, video) and also with the proposed multimodal approach by fusing the modalities. We employed various pretrained models for the text modality, such as Tamil-BERT, MuRIL, LaBSE (Language-agnostic BERT Sentence Embedding), and MuRIL-Large-Cased, to analyze the textual content and detect hate speech patterns. The audio modality involved utilizing the Wav2Vec2 model to extract relevant audio features and identify hate speech based on audio cues. The video modality utilized the TimeSformer model to capture spatiotemporal relationships and visual cues within videos for hate speech detection. We then applied machine learning models, including Logistic Regression, Decision Trees, Random Forests, and SVM, to comprehensively evaluate the dataset's performance in hate speech classification. This comprehensive approach allowed us to understand each modality's strengths and explore their effects on classification.

6.1 Text Modality

In our analysis of the text modality, we utilized four transformer-based models: Tamil-BERT [25], LaBSE [19], Hate-MuRIL [13], and MuRIL-Large-Cased [27]. To have a baseline comparison between other similar methods used, we utilized BERT as our baseline comparison model [38, 39]. Our main goal was to extract feature vectors from the text data, particularly the video transcripts, to facilitate effective training of machine learning models for classification. We began by extracting text from the dataset stored in docx files. Employing a systematic approach as shown in Figure 4, we initialized the tokenizer for each model and tokenized the input text. Leveraging the capabilities of the respective transformer-based model, we generated highly informative and contextually rich feature vectors. These feature vectors encapsulate the semantic essence and deep understanding of the text. Subsequently, we employed these feature vectors for classification using four machine learning models: Logistic Regression, Decision Tree, Random Forest, and SVM. After thorough comparison and analysis, our findings confirmed Tamil-BERT's superiority among

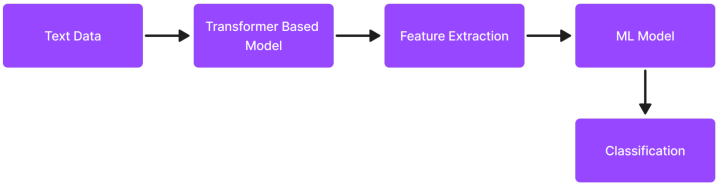

Fig. 4. Flow diagram of text modality.

the transformer-based models for HOL classification in Tamil. Remarkably, by combining Logistic Regression with Tamil-BERT for feature extraction, we attained an impressive accuracy of 81.82%. These findings strongly showcase Tamil-BERT's exceptional capacity to extract relevant and discriminative features from Tamil text. This notably boosts classification performance, highlighting its effectiveness and value in hate speech classification. We conducted many experiments with each model, and their performance and results are shown in Table 5.

6.1.1 BERT-Base-Uncased. BERT is a self-supervised model trained in English using a masked language modeling objective. It is uncased, meaning that it treats "english" and "English" the same. It is designed for fine-tuning on tasks like sequence classification, token classification, or question answering, where the entire sentence (possibly masked) is utilized for decision making. It is pretrained on English text with two objectives: masked language modeling and next sentence prediction. We used BERT as our baseline for comparison with other transformer-based models fine-tuned on Tamil or hate data. However, our results indicate that despite being a transformer model, BERT was unable to classify language-agnostic content into the four categories effectively. Table 5(a) shows the results obtained using this baseline model.

6.1.2 Tamil-BERT. Among various text models utilized in our study, one notable model is Tamil-BERT. Developed by Google AI, Tamil-BERT is a pretrained transformer-based language model tailored for natural language processing tasks in Tamil. Built upon the BERT model, Tamil-BERT offers impressive capabilities for fine-tuning it to various downstream tasks, including hate speech classification. Its exceptional performance across tasks such as sentiment analysis and text classification underscores its value. With its deep linguistic understanding and proficiency in capturing contextual nuances, Tamil-BERT significantly enhances our hate speech classification efforts, as demonstrated in Table 5(b).

6.1.3 MuRIL-Large-Cased. In our examination of various text modalities, we included MuRIL (Multilingual Representations for Indian Languages) as one of our selected models. Developed by Google, MuRIL is a pre-trained transformer-based language model capable of handling natural language processing tasks across multiple languages. MuRIL-Large-Cased, the largest version of the model, boasts 680 million parameters and has been trained on extensive text data from 101 languages, making it comprehensive and versatile. Leveraging the transformer architecture, MuRIL consistently achieves state-of-the-art results in various tasks, including text classification and named entity recognition. Notably, MuRIL excels in cross-lingual settings, leveraging its multilingual training data to handle diverse languages and nuances effectively. While MuRIL demonstrates exceptional performance and versatility across languages, it did not surpass the language-specific capabilities of Tamil-BERT in our hate speech classification tasks shown in Table 5(b).

6.1.4 LaBSE. We integrated LaBSE as one of our four selected models due to its remarkable ability to generate high-quality sentence embeddings across 109 languages. Its dual-encoder architecture effectively captures cross-lingual contexts and nuances, yet it did not outperform in

Table 5. Performance Comparison between Different Text Models

Maching Learning Method	Accuracy	Precision	Recall	F1-Score
SVM	45.45	60.42	41.67	36.90
LOGISTIC REGRESSION	54.54	70.83	50.00	43.33
RANDOM FOREST	54.54	65.00	50.00	43.04
DECISION TREE	45.45	62.50	47.92	37.50

(a) BERT-Base-Uncased (Baseline)

Machine Learning Method	Tamil-BERT				MuRIL-Large-Cased			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	72.77	79.17	68.75	60.18	72.72	87.50	68.75	63.10
LOGISTIC REGRESSION	**81.82**	**88.75**	**75.00**	**68.65**	72.72	87.50	68.75	63.10
RANDOM FOREST	72.72	83.75	68.75	62.50	72.72	87.50	62.50	58.33
DECISION TREE	45.455	60.00	41.67	36.79	63.64	50.00	58.33	53.33

(b) Tamil-BERT vs MuRIL-Large-Cased

Machine Learning Method	LaBSE				Hate-MuRIL			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	63.64	77.78	61.11	46.97	63.30	72.92	60.42	53.04
LOGISTIC REGRESSION	18.18	72.73	33.33	10.26	54.55	77.08	52.08	43.75
RANDOM FOREST	27.27	42.86	38.89	24.34	63.64	72.92	60.42	53.04
DECISION TREE	18.18	31.25	50.00	10.00	63.64	72.92	60.42	53.04

(c) LaBSE vs Hate-MuRIL

our hate speech classification tasks, as illustrated in Table 5(c). LaBSE excels in cross-lingual scenarios, addressing challenges posed by diverse linguistic sources. However, its lack of specific training in Tamil, unlike Tamil-BERT, affects its performance. This underscores the importance of language-specific models like Tamil-BERT, which excel in tasks targeting Tamil content. While LaBSE contributes to cross-lingual tasks, specialized training of language-specific models enhances their effectiveness in language-specific tasks.

6.1.5 Hate-MuRIL. We chose Hate-MuRIL as the final model for HOL classification. Developed by Google, Hate-MuRIL is a robust transformer-based language model tailored for detecting abusive speech in multiple languages. While it boasts a strong track record, including state-of-the-art results in various hate speech classification benchmarks, it did not match the performance of Tamil-BERT in our evaluation, as shown in Table 5(c). This discrepancy can be attributed to Hate-MuRIL's lack of explicit training in Tamil. Although Hate-MuRIL can classify hate content in several languages, including English and various Indian languages, language-specific models like Tamil-BERT excel due to their finely tuned capabilities for capturing language-specific nuances. Nonetheless, including Hate-MuRIL in our evaluation underscores its significance as a valuable resource for detecting hate speech across multiple languages. Our results demonstrate the effectiveness of using task-specific transformer-based models for extracting feature vectors from text data for hate speech classification. The Tamil-BERT model showed promising results and could potentially be used for other similar classification tasks in Tamil, as it is already trained on the Tamil dataset.

6.2 Audio Modality

The audio modality plays a crucial role in hate speech classification, and in this study, we conducted experiments on the audio modality using our dataset. To begin with, we loaded the audio files into a list and sorted them based on their file names, ensuring that they matched their corresponding labels. Next, we performed a necessary preprocessing step by resampling the audio files to a standardized sampling rate of 16 kHz. This step ensures consistency in the audio data and prepares it for further analysis. Following resampling, we utilized the Wav2Vec2 model [53] as

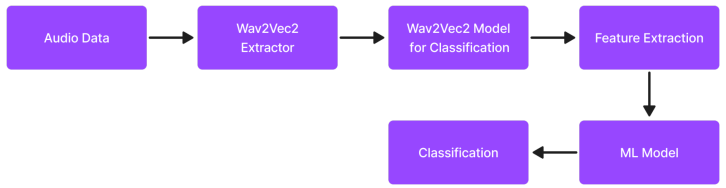

Fig. 5. Flow diagram of audio modality.

Table 6. Performance Comparison of Audio Modality

Machine Learning Method	Accuracy	Precision	Recall	F1-Score
SVM	54.54	55.00	50.00	48.81
LOGISTIC REGRESSION	63.64	78.75	62.50	52.38
RANDOM FOREST	54.54	71.67	47.92	45.24
DECISION TREE	36.36	37.50	29.17	30.95

(a) Classification of audio using time and frequency domain features (baseline)

Machine Learning Method	Wav2Vec2 for Raw Audio				Wav2Vec2 for Noise Removed Audio			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	63.63	58.75	56.25	54.76	63.60	54.17	62.50	57.50
LOGISTIC REGRESSION	63.63	75.00	56.25	50.60	72.72	81.25	68.75	59.52
RANDOM FOREST	63.63	70.83	56.25	50.18	45.50	63.33	41.67	38.10
DECISION TREE	54.54	73.21	50.00	44.17	45.45	58.75	43.75	37.50

(b) Performance comparison of Wav2Vec2 for Raw Audio and Noise Removed Audio

shown in Figure 5 for feature extraction to extract relevant features from the audio files. We then obtained logits, which are the outputs of the model representing the feature vector for each class. To organize the logits obtained from the Wav2Vec2 model, we stored them in a list and converted them into a numpy array. The resulting array had dimensions of (52, 12), where 52 represents the number of audio files in our dataset, and 12 represents the feature dimension associated with each audio file. This facilitated the subsequent machine learning modeling process. With the extracted features, we applied various machine learning models for hate speech classification. The models we employed included Logistic Regression, Decision Tree, Random Forest, and SVM. These models are widely used for classification tasks and offer different strengths in handling diverse datasets. We trained each model using the extracted features and their corresponding labels. Finally, we compared the accuracy of each machine learning model on our dataset, as shown in Table 6(b). This comparison provided insights into their effectiveness in classifying hate speech only using the audio modality.

6.2.1 Using Time and Frequency Domain Features. We conducted feature extraction from the audio data to establish a baseline comparison [1] with our transformer-based model. This process involved extracting four key features: energy, zero crossing rate, spectral centroid, and **Mel-Frequency Cepstral Coefficients (MFCCs)**, where we computed 13 MFCCs using the Mel scale. Energy signifies the magnitude of the sound, calculated by summing squared amplitude values over time in a given audio frame. Zero crossing rate measures the rate at which the audio signal changes its sign. The spectral centroid indicates the "average" frequency of a sound. MFCCs, computed using the Mel scale, represent the short-term power spectrum of a sound. They capture essential characteristics of an audio signal for speech and audio processing tasks. The feature extraction process was applied to each audio file individually, resulting in a feature vector for each file. Subsequently, these feature vectors were concatenated across all audio files, yielding a consolidated feature matrix with dimensions (52,16). Here, the first dimension (52) represents the total number of audio files processed, while the second dimension (16) signifies the length of the feature

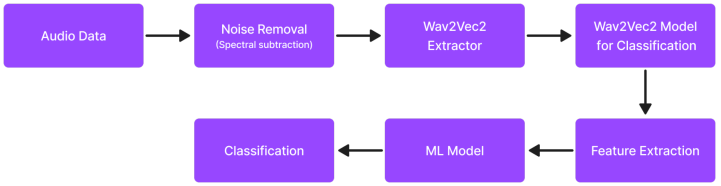

Fig. 6. Flow diagram of the noise-reduced audio modality.

vector for each audio file. This is then classified using machine learning models shown in Table 6(a). This approach allowed us to capture essential characteristics of the audio data in a compact and analyzable format, facilitating subsequent comparison and analysis with our transformer-based models.

6.2.2 Wav2Vec2. We chose the Wav2Vec2 model for audio feature extraction due to its exceptional capabilities. Specifically designed for speech recognition, Wav2Vec2 is pretrained on a vast audio corpus, enabling efficient feature extraction. Its self-supervised learning approach allows it to learn from audio data without manual annotations, achieving impressive word error rates on the LibriSpeech dataset with limited labeled data. By pretraining on a large unlabeled dataset, Wav2Vec2 outperforms previous state-of-the-art models while using significantly less labeled data, showcasing its potential for speech recognition tasks with limited labeled data. Hence, we utilized this model for extracting features from our hate audio data.

6.2.3 Noise-Reduced Wav2Vec2. Upon thorough analysis of the audio recordings, we identified unwanted noise, posing a challenge to our hate speech classification task. To mitigate this issue, we conducted experiments to remove noise and evaluate subsequent improvements in a noise-free environment. Employing the Spectral Subtraction technique [47], we harnessed its ability to distinguish noise from desired audio signals like speech. This technique estimates noise floor through power spectrum analysis and subtracts it from the signal, effectively attenuating noise components. The resulting modified power spectrum reconstructs the enhanced quality filtered audio signal as illustrated in the pipeline process in Figure 6. To assess the impact of noise removal on hate speech classification, we applied the spectral subtraction algorithm to our dataset. We denoised the audio files using a noise threshold parameter of 0.05 and an alpha parameter of 2.0, and standardized their frequency to 16 kHz. Subsequently, we utilized the Wav2Vec2 model to extract relevant features from the audio, organizing them into a numpy array with dimensions (52,12), representing 52 audio files and 12 feature representations per audio. These features served as input for subsequent machine learning models for classification. Our results shown in Table 6 revealed the efficacy of spectral subtraction in noise removal, significantly improving hate speech classification. In particular, Logistic Regression emerged as the most successful among employed techniques, achieving an impressive accuracy rate of 72.72%.

6.3 Video Modality

When we tried to classify hate speech using the video modality, we encountered challenges in capturing and analyzing relevant information from the videos. One of the complexities we encountered was the difficulty in visually capturing emotions expressed by the YouTubers, making hate speech classification through videos a demanding task. To address these challenges, we devised a comprehensive approach. We initiated the video modality classification by setting the path to the video file and specifying the desired frame size for the transformer model to extract important visual cues and temporal information. Leveraging the OpenCV library, we loaded the video file,

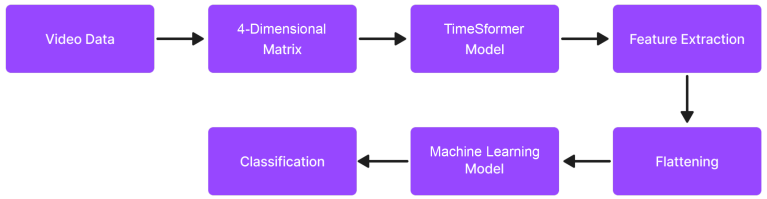

Fig. 7. Flow diagram of video modality.

Table 7. Performance Comparison of Video Modality

Machine Learning Method	TimeSformer				Inception Net (Baseline)			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	45.45	71.88	43.75	38.64	35.63	35.17	32.17	34.12
LOGISTIC REGRESSION	45.45	76.04	50.00	33.64	35.63	32.24	32.17	30.95
RANDOM FOREST	45.45	73.21	50.00	31.67	45.45	65.00	47.92	43.33
DECISION TREE	27.00	16.67	22.92	18.25	45.45	62.50	35.42	35.00

enabling us to extract essential properties such as the frame rate, width, height, and frame count. We then converted the MP4 to array representation, which involved converting the videos into a 4-dimensional array that accounts for the color channels, frame count, and height and width of the video frames. This transformation allowed us to analyze and extract relevant features from the video data. Furthermore, we padded the frames array, adding zeros along the second axis (frame count) to ensure consistent dimensions across all videos, which varied in length. This step is crucial for an accurate classification task, requiring fixed input dimensions for all videos. We passed the arrays of all videos through the TimeSformer model [6] as shown in Figure 7, generating a sequence of feature vectors that represent the model's understanding. These vectors were stored in a list for further processing. To classify videos effectively, we transformed and appended the collected feature vectors from 52 video files into a numpy array, aligning them with corresponding labels. This numpy array served as the primary input for subsequent classification tasks. We utilized machine learning models for classification, including SVM, Decision Tree, Random Forest, and Logistic Regression, as shown in Table 7.

We employed an 80%-20% train-test split for this process. Training the models on the feature vectors enabled them to learn distinct hate speech patterns, thus improving classification accuracy. We chose Inception Net to classify HOL content for a baseline comparison with the proposed approach [14]. Similar to TimeSformer, we passed video arrays through Inception Net, extracting feature vectors from the last layer before the softmax layer. We utilized machine learning classification algorithms to employ these feature vectors for all 52 videos in our dataset. Our results shown in Table 7 indicate that the TimeSformer model outperformed Inception Net in classifying HOL content.

6.3.1 TimeSformer. TimeSformer [6] is a transformer-based model tailored for video analysis tasks, adept at capturing long-range temporal dependencies and understanding video temporal dynamics. Leveraging TimeSformer, we harnessed its capability to process and extract meaningful features from video data. Its architecture facilitates learning complex patterns and temporal relationships within videos, making it ideal for hate speech classification. With TimeSformer, we could capture crucial contextual information and temporal nuances necessary for identifying hate speech within the video modality.

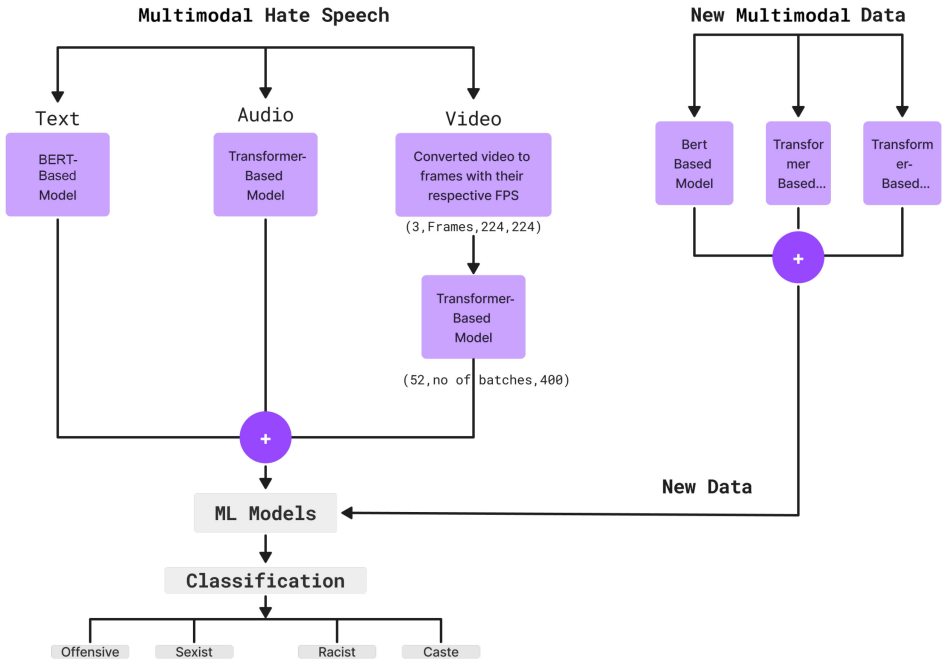

Fig. 8. Proposed methodology.

7 Results and Discussions

7.1 Multimodal Approach

Based on our ablation study, we discovered that the text modality and audio modality constitute more toward classifying the type of hate content. To assess the performance of the multimodal approach, we further conducted experiments to analyze this. Using the feature embeddings from various models across different modalities, we now fuse and integrate these features to construct an effective multimodal hate speech classification framework as shown in Figure 8. Through a series of experiments, we aim to find the optimal combination of models from each modality that achieves superior performance. To establish a baseline comparison with existing multimodal architectures [14], we concatenate the feature vectors from baseline models for each modality—text, audio and video—for further classification. Specifically, this multimodal baseline comparison utilizes BERT for text, time and frequency domain features for audio, and Inception Net for video. By conducting extensive experiments and exploring various combinations of models from each modality, we aim to identify the configuration yielding the best hate speech detection performance. We systematically analyzed the impact of various model components and hyperparameters on the hate speech classification task. Employing a grid search with cross validation, we explored a range of hyperparameters for our four machine learning classifiers used in this study. By evaluating precision, recall, F1-score, and accuracy metrics, we identified the optimal configurations for each classifier. After extensive experimentation and analysis, we have determined that a specific combination of models yields highly effective results. Through this process, we gained insights into the relative importance of different modalities and their respective models, leading us to select the HateMuRIL model for text, the Noise Removed Wav2Vec2 model for audio, and the TimeSformer model for video. This combined approach achieved superior hate speech classification performance, as

Table 8. Combined Performance Comparison

Machine Learning Method	Accuracy	Precision	Recall	F1-Score
SVM	72.72	81.67	68.75	61.67
RANDOM FOREST	54.54	65.00	47.92	43.44
LOGISTIC REGRESSION	72.72	81.25	68.75	60.42
DECISION TREE	45.45	62.50	35.42	35.00

(a) Performance comparison of the baseline multimodal approach

Machine Learning Method	LaBSE + Audio + Video				Hate-MuRIL + Audio + Video			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	66.70	90.00	50.00	43.75	72.72	77.50	62.50	56.25
LOGISTIC REGRESSION	66.70	90.00	50.00	43.75	72.72	77.08	62.50	56.43
RANDOM FOREST	33.33	45.83	50.00	29.17	54.54	73.21	56.25	48.33
DECISION TREE	16.70	37.50	25.00	16.67	36.36	45.83	37.50	26.67

(b) Performance comparison of the multimodal approach for Raw Audio

Machine Learning Method	Tamil-BERT + Audio + Video				MuRIL-Large + Audio + Video			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	72.72	78.33	62.50	57.22	72.72	77.50	62.50	56.25
LOGISTIC REGRESSION	72.72	76.25	62.50	56.15	63.63	75.00	56.25	50.60
RANDOM FOREST	54.54	75.00	56.25	43.04	45.55	73.21	50.00	31.67
DECISION TREE	63.63	75.00	56.25	50.60	54.54	60.42	58.33	58.93

(c) Performance comparison of the multimodal approach for Raw Audio

Machine Learning Method	LaBSE + Noise Removed Audio + Video				Hate-MuRIL + Noise Removed Audio + Video			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	45.45	35.42	43.42	38.57	72.72	85.00	62.50	57.25
LOGISTIC REGRESSION	45.45	42.50	43.75	40.48	**81.82**	**87.50**	**75.00**	**66.67**
RANDOM FOREST	54.54	75.00	56.25	43.04	54.54	73.21	56.25	48.33
DECISION TREE	54.54	65.00	50.00	43.04	36.36	45.83	37.50	26.67

(d) Performance comparison of the multimodal approach for Noise Removed Audio

Machine Learning Method	Tamil-BERT + Noise Removed Audio + Video				MuRIL-Large + Noise Removed Audio + Video			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	45.45	39.58	43.75	40.48	54.54	48.33	50.00	45.42
LOGISTIC REGRESSION	45.45	42.50	43.75	40.48	63.63	56.25	62.50	54.76
RANDOM FOREST	54.54	75.00	56.25	43.04	45.45	73.21	50.00	31.67
DECISION TREE	63.63	75.00	56.25	50.60	54.54	60.42	58.33	58.93

(e) Performance comparison of the multimodal approach for Noise Removed Audio

demonstrated in Figure 10 and detailed in Table 8. Additionally, various hyperparameters used for this work are presented in Figure 11. Based on our experiments, we found that contrary to traditional expectations, Logistic Regression surprisingly outperforms SVM in our dataset. Despite the conventional wisdom that SVM performs better on nonlinearly separable data due to its kernel trick, our results show that Logistic Regression achieved an accuracy of 81.82% compared to SVM's 72.72%. This unexpected outcome suggests that our data may have characteristics that favor Logistic Regression over SVM, even in scenarios where the data is not linearly separable, as demonstrated in the t-SNE plot in Figure 9. These results can be attributed to several characteristics of our data. Despite the data being nonlinearly separable in the original input space, it is possible that after feature transformation and concatenation, the data becomes more linearly separable, which benefits Logistic Regression. Additionally, Logistic Regression is effective when dealing with a large number of features, especially if some features are more important than others. If certain features have more relevance to the classification task, Logistic Regression may exploit them more effectively, therefore excelling in our classification task.

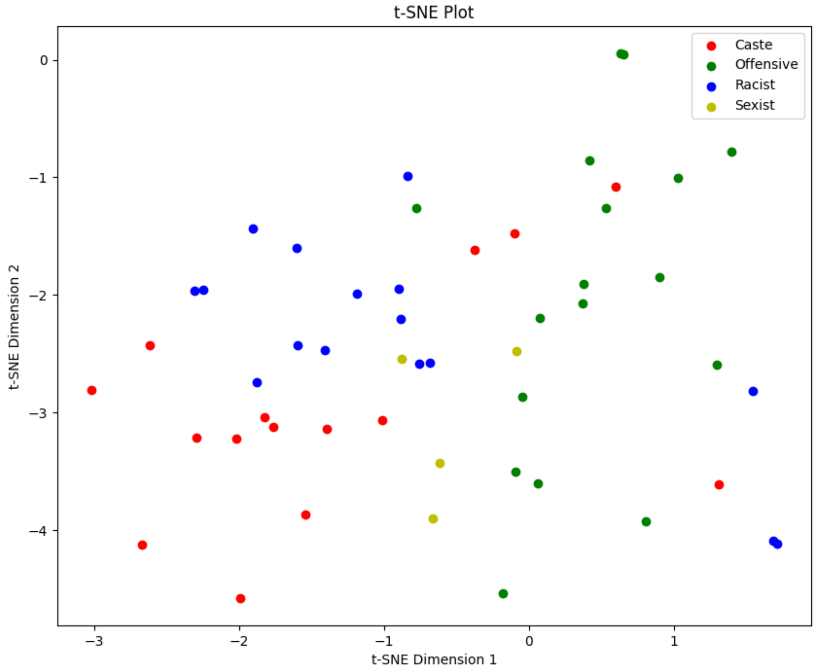

Fig. 9. t-SNE plot of the dataset.

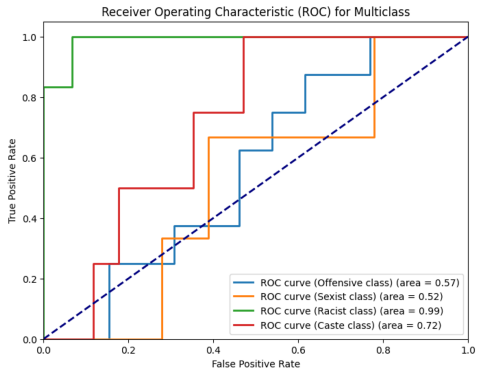

Fig. 10. ROC AUC graph for the best-performing model (Hate-MuRIL + Noise Removed Audio + Video).

Machine Learning Model	Parameter	Value
LOGISTIC REGRESSION	max_iter	10,000
	solver	lbfgs
	multi_class	multinomial
	random_state	42
DECISION TREE	criterion	gini
	random_state	42
SVM	C	0.1
	class_weight	None
	gamma	scale
	kernel	linear
	max_iter	10,000
RANDOM FOREST	max_depth	5
	min_samples_split	2
	n_estimators	200

Fig. 11. Hyperparameters for various classification models.

8 Assessing Model Generalization and Overfitting

To ensure the reliability and robustness of our Logistic Regression model, we employed a comprehensive evaluation strategy focusing on model generalization and potential overfitting. We utilized two key visualizations—loss curves and accuracy plots—which provide insights into the model's performance across varying levels of regularization strength.

8.1 Loss Curves

The loss curves shown in Figure 12 plot the negative log-likelihood loss against the logarithm of the regularization strength. This plot illustrates the model's behavior concerning its capacity to

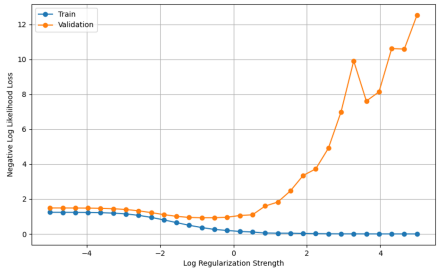

Fig. 12. Train vs validation loss for Logistic Regression.

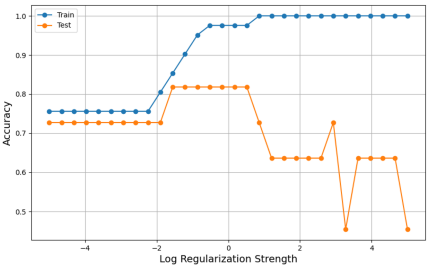

Fig. 13. Train vs test accuracy for Logistic Regression.

fit the training data while maintaining generalization to unseen data. Initially, both training and validation losses decrease consistently as the regularization strength decreases, indicative of the model learning meaningful patterns. However, beyond a certain point, a divergence occurs where the validation loss begins to increase, signaling potential overfitting. Conversely, the training loss continues to decrease, suggesting that the model may be fitting noise in the training data.

8.2 Accuracy Plot

The accuracy plot shown in Figure 13 showcases the performance of the model in terms of classification accuracy on both the training and test sets across different regularization strengths. Similar to the loss curves, the accuracy initially increases as the regularization strength decreases, indicating the model's ability to learn from the data. However, the test accuracy peaks at a certain point, after which it either plateaus or decreases, whereas the training accuracy continues to rise. This divergence signifies a transition from generalization to overfitting, as the model starts to memorize the training data rather than capturing underlying patterns.

8.3 Model Generalization and Optimal Point

Upon analyzing both plots, we identify a critical point where the test accuracy reaches its highest value, approximately 81.82%. The model balances complexity and generalization at this optimal regularization strength, resulting in robust performance on unseen data. This convergence of evidence from the loss curves and accuracy plot supports the assertion that our Logistic Regression model is well generalized and not overfitted. The identified optimal point, where the test accuracy peaks, signifies the model's ability to generalize effectively to new data while avoiding excessive complexity. This comprehensive evaluation assures the reliability and effectiveness of our model in real-world applications.

9 Conclusion and Future Works

This article presented a thorough investigation into hate speech detection in Tamil, focusing on developing a multimodal dataset and evaluating various machine learning models for classification. This research fills a crucial gap in comprehensive studies and resources dedicated to hate speech detection in low-resource languages like Tamil. The creation of the MATH dataset provides a significant resource for hate speech research in Tamil. This dataset includes videos with audio and text modalities, annotated with four distinct categories of hate speech: offensive, sexist, racist, and caste. Involving multiple annotators with diverse backgrounds ensures the accuracy and reliability of the dataset annotations. Our experiments explored the effectiveness of different modalities and machine learning models for hate speech classification. Transformer-based

models, particularly Tamil-BERT, demonstrate superior performance in the text modality. Employing the Wav2Vec2 model with noise removal techniques in the audio modality significantly improves hate speech classification accuracy. While video modality experiments faced challenges in capturing emotional cues, they still provide valuable insights into the classification of hate speech. The proposed multimodal approach excels in hate speech classification by combining text, audio, and video features, boosting accuracy and depth beyond single-modal methods. These findings contribute significantly to hate speech detection research by addressing the lack of comprehensive studies and resources in low-resource languages like Tamil. The MATH dataset and evaluated multimodal framework offer valuable insights and tools for researchers and practitioners working on hate speech detection in Tamil and similar languages.

Future research in hate speech detection in Tamil can be enhanced by exploring advanced techniques, particularly in video modality, and developing sophisticated multimodal fusion approaches to improve classification accuracy. Continuous updates and expansions to datasets and models are also essential to adapt to the evolving nature of hate speech, ensuring effectiveness in real-world scenarios and staying updated on emerging trends.

References

[1] Sarah A. Abdu, Ahmed H. Yousef, and Ashraf Salem. 2021. Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion* 76 (2021), 204–226.

[2] Natalie Alkiviadou. 2019. Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law* 28, 1 (2019), 19–35.

[3] Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M. Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access* 12 (2024), 22359–22375.

[4] Vidhya Balasubramanian, Sooryanarayan Gobu Doraisamy, and Navaneeth Kumar Kanakarajan. 2016. A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems* 46 (2016), 121–145.

[5] Fazlourrahman Balouchzahi, Anusha Gowda, Hosahalli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2-22: Abusive comment detection in Tamil language using 1D Conv-LSTM. In *Proceedings of the 2nd Workshop on Speech and Language Technologies for Dravidian Languages*. 64–69.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*. 813–824.

[7] Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data '21)*. IEEE, 4496–4499.

[8] Austin Botelho, Bertie Vidgen, and Scott A. Hale. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. *arXiv preprint arXiv:2106.05903* (2021).

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42 (2008), 335–359.

[10] Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Anand Kumar Madasamy, Sajeetha Thavareesan, B. Premjith, K. Sreelakshmi, Subalalitha Chinnaudayar Navaneethakrishnan, John P. McCrae, and Thomas Mandl. 2021. Overview of the HASOC-DravidianCodeMix shared task on offensive language detection in Tamil and Malayalam. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE '21) (Working Notes)*. 589–602.

[11] Bharathi Raja Chakravarthi, Anand Kumar M, John P. McCrae, B. Premjith, K. P. Soman, and Thomas Mandl. 2020. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE '20) (Working Notes)*. 112–120.

[12] Bharathi Raja Chakravarthi, Jishnu Parameswaran P.K., Premjith B, K. P. Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. 2021. DravidianMultiModality: A dataset for multi-modal sentiment analysis in Tamil and Malayalam. *arXiv preprint arXiv:2106.04853* (2021).

[13] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for Indic languages. *arXiv preprint arXiv:2204.12543* (2022).

[14] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. HateMM: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 1014–1023.

[15] Gretel Liz De La Peña Sarracén. 2021. Multilingual and multimodal hate speech analysis in Twitter. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1109–1110.

[16] Abreham Gebremedin Debele and Michael Melese Woldeyohannis. 2022. Multimodal Amharic hate speech detection using deep learning. In *Proceedings of the 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA '22)*. 102–107. https://doi.org/10.1109/ICT4DA56482.2022.9971436

[17] V. Sharmila Devi, S. Kannimuthu, and Anand Kumar Madasamy. 2024. The effect of phrase vector embedding in explainable hierarchical attention-based Tamil code-mixed hate speech and intent detection. *IEEE Access* 12 (2024), 11316–11329.

[18] Medicharla Dinesh Surya Sai Eswar, Nandhini Balaji, Vedula Sudhanva Sarma, Yarlagadda Chamanth Krishna, and S. Thara. 2022. Hope speech detection in Tamil and English language. In *Proceedings of the 2022 International Conference on Inventive Computation Technologies (ICICT '22)*. IEEE, 51–56.

[19] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852* (2020).

[20] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1470–1478.

[21] Google. n.d.. Turn Speech into Text Using Google AI. Retrieved January 19, 2025 from https://cloud.google.com/speech-to-text

[22] Md. Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).

[23] IBM. n.d.. IBM Watson Speech to Text. Retrieved January 19, 2025 from https://www.ibm.com/cloud/watson-speech-to-text

[24] Othman Istaiteh, Razan Al-Omoush, and Sara Tedmori. 2020. Racist and sexist hate speech detection: Literature review. In *Proceedings of the 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA '20)*. 95–99. https://doi.org/10.1109/IDSTA50958.2020.9264052

[25] Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT transformer models for Devanagari based Hindi and Marathi languages. *arXiv preprint arXiv:2211.11418* (2022).

[26] Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md. Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from Bengali memes and texts. In *Proceedings of the International Conference on Speech and Language Technologies for Low-Resource Languages*. 293–308.

[27] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual representations for Indian languages. arXiv:2103.10730 [cs.CL].

[28] Sujata Khedkar, Priya Karsi, Devansh Ahuja, and Anshul Bahrani. 2022. Hateful memes, offensive or non-offensive! In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*. Vol. 2. Springer, 609–621.

[29] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33 (2020), 2611–2624.

[30] J. Richard Landis and Gary G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33, 2 (1977), 363–374.

[31] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920* (2018).

[32] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489* (2021).

[33] Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*. 174–179. https://doi.org/10.18653/v1/W16-0429

[34] Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction* 5, 7 (2021), 34.

[35] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[36] Nishchal Prasad, Sriparna Saha, and Pushpak Bhattacharyya. 2021. A multimodal classification of noisy hate speech using character level embedding and attention. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN '21)*. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533371

[37] B. Premjith and K. P. Soman. 2021. Deep learning approach for the morphological synthesis in Malayalam and Tamil at the character level. *Transactions on Asian and Low-Resource Language Information Processing* 20, 6 (2021), 1–17.

[38] Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218* (2022).

[39] Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language* 75 (2022), 101386.

[40] Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science* 3, 2 (2022), 122.

[41] John Solomon. 2022. Caldwell's Dravidians: Knowledge production and the representational strategies of missionary scholars in colonial South India. *Modern Asian Studies* 56, 6 (2022), 1741–1773. https://doi.org/10.1017/S0026749X21000524

[42] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713* (2022).

[43] K. Sreelakshmi, B. Premjith, and K. P. Soman. 2020. Detection of hate speech text in Hindi-English code-mixed data. *Procedia Computer Science* 171 (2020), 737–744.

[44] Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language* 76 (2022), 101404.

[45] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *arXiv preprint arXiv:1804.05788* (2018).

[46] United Nations. n.d. Understanding Hate Speech. Retrieved January 19, 2025 from https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[47] Saeed V. Vaseghi. 1996. Spectral subtraction. In *Advanced Signal Processing and Digital Noise Reduction*. Springer, 242–260.

[48] Yeshan Wang and Ilia Markov. 2024. CLTL@Multimodal Hate Speech Event Detection 2024: The winning approach to detecting multimodal hate speech and its targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE '24)*. 73–78.

[49] WorldData.info. 2023. Tamil—Worldwide Distribution. Retrieved January 19, 2025 from https://www.worlddata.info/languages/tamil.php

[50] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. YouTube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28, 3 (2013), 46–53. https://doi.org/10.1109/MIS.2013.34

[51] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4505–4514.

[52] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the 3rd Workshop on Abusive Language Online*. 11–18.

[53] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. SUPERB: Speech processing Universal PERformance Benchmark. *arXiv preprint arXiv:2105.01051* (2021).

[54] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[55] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[56] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

Received 9 August 2023; revised 10 May 2024; accepted 23 December 2024