Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal Abusive Language Detection in Tamil based on Integrated Approach of Machine Learning and Deep Learning Techniques

Md. Tanvir Rahman, Abu Bakkar Siddique Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804002, u1804004, u1804015, u1704057, u1704039}@student.cuet.ac.bd
{avishek, moshiul_240}@cuet.ac.bd

Abstract

Detecting abusive language on social media is a challenging task that needs to be solved effectively. This research addresses the formidable challenge of detecting abusive language in Tamil through a comprehensive multimodal approach, incorporating textual, acoustic, and visual inputs. This study utilized ConvLSTM, 3D-CNN, and a hybrid 3D-CNN with BiLSTM to extract video features. Several models, such as BiLSTM, LR, and CNN, are explored for processing audio data, whereas for textual content, MNB, LR, and LSTM methods are explored. To further enhance overall performance, this work introduced a weighted late fusion model amalgamating predictions from all modalities. The fusion model was then applied to make predictions on the test dataset. The ConvLSTM+BiLSTM+MNB model yielded the highest macro F1 score of **71.43%**. Our methodology allowed us to achieve $\mathbf{1}^{st}$ rank for multimodal abusive language detection in the shared task.

1 Introduction

Recently, the proliferation of social media platforms has played a pivotal role in facilitating the global exchange of ideas, opinions, and information. Although this interconnectedness has engendered a rich diversity of conversations, it has concurrently presented challenges, as noted by Das et al. (2021). The widespread occurrence of abusive language, hate speech, and offensive content exemplifies these challenges. Among the myriad languages employed on these platforms, one notable example is Tamil, a Dravidian language predominantly spoken in South India. The need for effective detection and mitigation of abusive language in Tamil is imperative to ensure a secure and inclusive online environment. Yasaswini et al. (2021) addressed this pressing concern by exploring a multimodal approach to abusive language detection in Tamil. By incorporating multiple modalities, such as text, video, and audio, we aim to enhance the accuracy and robustness of the detection system in the Tamil-speaking digital landscape.

Although studies have been conducted on these issues for multimodal data in the English language, there needs to be more research explicitly looking at abusive language recognition in the context of Dravidian languages (Barman and Das, 2023; Bala and Krishnamurthy, 2023). Limited research on abusive language identification in these languages presents unique challenges. As part of our effort, we investigated abusive language detection in Tamil. The task entails the development of models capable of analyzing the textual, speech, and visual elements within social media videos to predict their classification as either abusive or non-abusive. The main contributions of this work are:

- Implement a weighted late fusion model that effectively amalgamates predictions from text, video, and audio modalities.
- Investigate various Machine Learning (ML), Deep Learning (DL), and their integrated approaches to find a suitable solution for detecting multimodal abusive language in Tamil.

2 Related Work

Online abuse poses a significant threat, prompting the need for sophisticated measures. Multimodal strategies, integrating text, acoustic, and visual analysis, are at the forefront of enhancing content moderation efficacy. Premjith et al. (2023) provides an overview of the shared task on multimodal abusive language detection and sentiment analysis in Dravidian languages, carried out during the third Workshop on Speech and Language Technologies for Dravidian Languages at RANLP 2023. This study covers the word vector enhancement techniques given by Bojanowski et al. (2017), which may help understand word representations in Dravidian languages. In their survey of recent

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 212–217

advancements in multimodal sentiment analysis (text, audio, and video/image), Chandrasekaran et al. (2021) presented a thorough analysis of sentiment datasets, feature extraction algorithms, data fusion techniques, and the effectiveness of various classification strategies. A more comprehensive presentation of multimodal and multilingual hate speech detection was given by Chhabra and Vishwakarma (2023).

In a multilingual social media setting, Sharon et al. (2022) presented MADA, a multimodal technique for abuse detection in conversational audio. (Mozafari et al., 2020) used a range of methods using various architectures, including multi-modal models, video-based models, text-based models, and image-based models. Poria et al. (2018) developed a Multimodal EmotionLines Dataset (MELD) to improve and expand EmotionLines. A novel approach to multimodal sentiment analysis was presented by Poria et al. (2016), which used textual, visual, and audio modalities to extract sentiments from web videos. They combined adequate data from several sources using feature and decision-level fusion techniques. Several methods are used for categorizing audio, video, and natural language text that identify the emotions conveyed as Positive, Negative, or Neutral by Mahendhiran and Kannimuthu (2018). In their analysis of the growth of multimedia communication apps, Soni and Singh (2018) noted both the dangers of cyberbullying and the potential for improved user engagement and natural communication. Few studies have been conducted on multimodal social media data analysis in Dravidian languages. Barman and Das (2023) proposed various unimodal models and introduced a fusion model. Their investigation highlighted mBERT and MFCC's efficacy in classifying abusive language. Furthermore, the Vision Transformer (ViT) demonstrated notable success in sentiment analysis for Tamil and Malayalam, achieving an F1-score (macro) of 57.86% (Barman and Das, 2023). Bala and Krishnamurthy (2023) conducted a study focused on detecting abusive language that amalgamated visual, auditory, and textual features, culminating in an F1-score (macro) of 33%, indicative of the achieved performance in this multimodal endeavor.

3 Task and Dataset Description

The task aims to develop advanced models for detecting abusive content in Tamil videos on social media, particularly YouTube. These models scrutinize diverse video elements to predict whether the content is abusive or non-abusive. Abusive content includes offensive language or visuals intended to cause harm, distress, or discomfort, while non-abusive content aligns with guidelines promoting respectful and positive engagement. The organizers of the competition, DravidianLangTech@EACL 2024, have released a Tamil language dataset for abusive content detection (Chakravarthi et al., 2021; Premjith et al., 2022; B et al., 2024). Table 1 illustrates the training and test set distribution for all three modalities.

Dataset	Abusive	Non-Abusive	Total
Train	38	32	70
Test	9	9	18

Table 1: Summary of abusive language detection dataset

The provided dataset included components encompassing video, audio, and extracted text. In the training dataset, the duration for both audio and video spans from a minimum of 23.38 seconds to a maximum of 86.36 seconds, with an average time of 47.90 seconds. The extracted text data contains 19,743 words, consisting of 642 unique words. After removing stopwords, the number of unique words reduces to 588. The text samples exhibit a minimum of 65 words, a maximum of 500 words, and an average of 282.04 words.

4 Methodology

The proposed work starts by examining the videos' visual elements and then investigates the audio data before moving on to the textual aspects of the modeling. To combine the textual, audio, and visual elements, we employed weighted late fusion to create a more reliable classification of abusive and non-abusive content. Figure 1 displays the schematic framework for detecting multimodal abusive language in Tamil.

4.1 Visual Approach

This study targets the detection of abusive content in videos. Using OpenCV, we extract 15 sequential frames from each video, each being 128x128 in size. The extracted frames are normalized for consistent pixel value comparison in subsequent analysis. These frames are then input into three distinct models: ConvLSTM, 3D-CNN, and 3D-CNN combined with BiLSTM. To extract spatial and

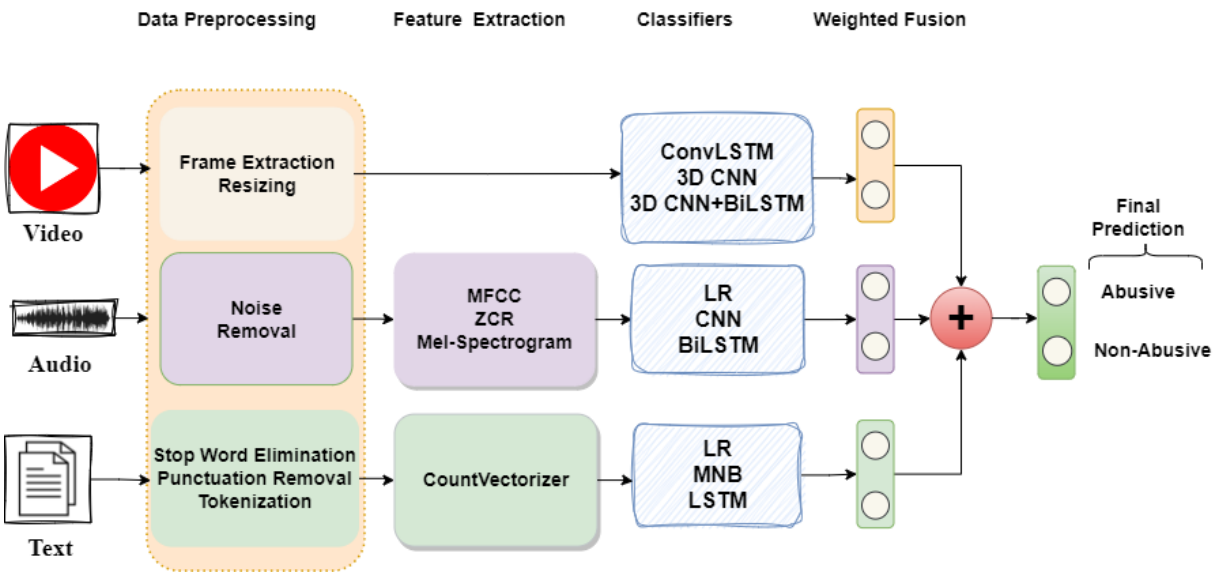

Figure 1: Schematic framework for Tamil abusive language detection

temporal features simultaneously, we employed ConvLSTM, which combines CNNs and LSTMs. We also used a hybrid model combining 3D-CNNs with LSTM to use their complementary abilities for improved abusive content identification. Additionally, standalone 3D-CNNs are used for extracting direct spatiotemporal features. The experimental setup for the deep learning models employed in video is outlined in Table 2.

Parameter	Value
Optimizer	adam
Loss function	categorical_crossentropy
Activation_function (hidden layer)	tanh
Activation_function (output layer)	softmax
Batch size	4
Learning rate	0.001
Epochs	100
Drop-out	0.3

Table 2: Experimental setup for the DL models for video

4.2 Acoustic Approach

This research prioritized noise removal as a preliminary step to refine the audio data. Employing effective noise reduction methods enhances the clarity of the audio signals. Subsequently, we extracted features from each audio sample to discern abusive content. The features include Mel-frequency spectrogram (Mel), Zero Crossing Rate (ZCR), Spectral Contrast, and Mel-frequency Cepstral Coefficients (MFCC). Each feature serves a specific purpose—Spectral Contrast captures spectral texture changes, Mel reflects frequency distribution, ZCR denotes the rate of signal crossings, and MFCC records detailed spectral features. Following extracting these discriminative features, we employed a diverse set of models, namely LR, CNN, and BiLSTM, for a nuanced and thorough analysis leading to the classification of abusive content.

4.3 Textual Approach

In this work, we preprocessed the text using punctuation and Tamil stop-word removal to identify abusive content in text data. We extracted features using CountVectorizer to convert text data into vectorized tokens. The study examined the influence of various preprocessing methods on the performance of conventional machine learning models, such as Logistic Regression (LR) and Multinomial Naive Bayes (MNB). Concurrently, the deep learning model LSTM was also employed. The preprocessed data was utilized to train these models, renowned for their sequential processing capabilities, to discern intricate dependencies and temporal nuances within the text sequences.

4.4 Weighted Late Fusion Approach

This research adopted a weighted late fusion methodology (Pasqualino et al., 2020), wherein distinct models are trained independently for each modality (text, audio, and video). The predictions generated by these individual models are subsequently combined later, employing uniform weights for each modality. This late fusion ap-

Approach	Models	P	R	F1
Visual	ConvLSTM	0.56	0.56	0.56
	3D CNN	0.56	0.56	0.56
	3D CNN+BiLSTM	0.50	0.50	0.50
Acoustic	BiLSTM	0.75	0.72	0.71
	CNN	0.80	0.67	0.62
	LR	0.62	0.61	0.60
Textual	LSTM	0.25	0.50	0.33
	MNB	0.66	0.61	0.58
	LR	0.66	0.61	0.58
Multimodal	**ConvLSTM+BiLSTM+MNB**	0.75	0.72	**0.71**
	ConvLSTM+CNN+LR	0.71	0.67	0.65
	3DCNN+CNN+LSTM	0.62	0.61	0.60
	(3DCNN+BiLSTM)+BiLSTM+MNB	0.71	0.67	0.65

Table 3: Performance of different unimodal and multimodal approaches on the test set, where P, R, and F1 denotes precision, recall and macro F1-score respectively

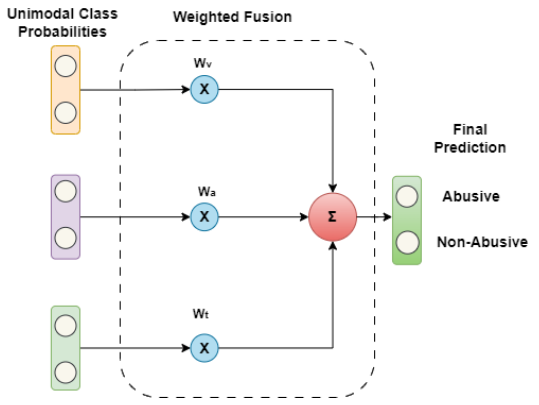

Figure 2: Schematic framework for weighted late fusion

proach allows us to leverage each modality's unique features and effectively amalgamate their contributions. Figure 2 depicts the schematic framework for weighted late fusion for multimodal abusive language detection in Tamil. Mathematically, the late fusion process's final prediction (P) can be expressed 1.

$$P = w_t \cdot P_{text} + w_a \cdot P_{audio} + w_v \cdot P_{video} \quad (1)$$

Here, P is the ultimate output of the late fusion model. The equitable allocation of uniform weights (w_t, w_a, w_v), each assigned a weight of 0.333, guarantees an even contribution from every modality. This approach facilitates a harmonized integration of information originating from text, audio, and video sources within the framework of our weighted late fusion methodology.

5 Results

Precision (P), recall (R), and macro F1-score (F1) are used to assess the model's performance. Results on the test dataset demonstrate that the combined model (ConvLSTM+BiLSTM+MNB) achieved preeminence, securing the topmost performance with a notable F1 (macro) score of 0.7143, as detailed in Table 3.

Table 4 compares the other team's performance with the rank participating in the shared task, where it is evident that our proposed method has achieved the highest F1-score among all participating teams.

Team	F1-(macro	Rank
Binary_Beasts	**0.7143**	**1**
Wit Hub	0.4156	2

Table 4: Competition rank list for Tamil abusive language detection

5.1 Error Analysis

A thorough evaluation of the model's performance on the test data is provided through the presentation of the confusion matrix in Figure 3. The examination of the confusion matrix reveals nearly flawless detection accuracy for abusive content, whereas the performance for non-abusive content is comparatively lower. This discrepancy may be attributed to the model incorrectly predicting non-abusive instances as abusive, potentially due to shared features between the two classes, or it could be a consequence of inadequate training data leading to

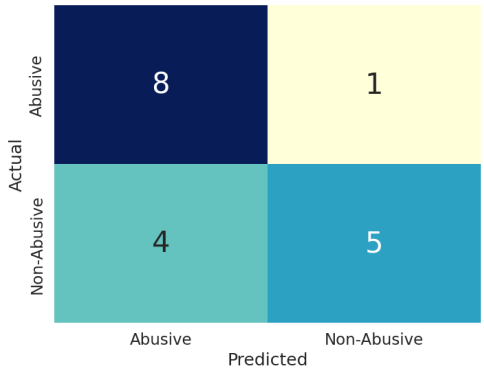

Figure 3: Confusion matrix of the proposed model (**ConvLSTM+BiLSTM+MNB**)

Test sample	Actual	Predicted
test1.txt test1.mp3 test1.mp4	Abusive	Abusive
test2.txt test2.mp3 test2.mp4	Abusive	Non-Abusive
test10.txt test10.mp3 test10.mp4	Non-Abusive	Non-Abusive
test15.txt test15.mp3 test15.mp4	Non-Abusive	Non-Abusive
test16.txt test16.mp3 test16.mp4	Non-Abusive	Abusive

Table 5: Few examples of predicted outputs by the proposed model (**ConvLSTM+BiLSTM+MNB**)

challenges in generalizing to diverse contexts.

Table 5 illustrates some correct and incorrect predicted outcomes by the best-performed model (**ConvLSTM+BiLSTM+MNB**).

Limitations

The current work encountered several hurdles, including:

- Concerns exist about the efficacy of the uniform weighting strategy, potentially compromising the model's responsiveness to specific features.
- The study concentrated on content detection in Tamil, prompting the need to explore the generalizability of the approach to other linguistic domains.
- The dataset's composition and size could influence the model's robustness, necessitating exploration across diverse datasets to validate adaptability in varied contextual settings.

6 Conclusion

This paper endeavors to advance the field of abusive content classification in the Tamil language by employing a multimodal approach that integrates textual, auditory, and visual information. Implementing a weighted late fusion strategy (with an integrated approach of ConvLSTM, BiLSTM, and MNB models), where each modality is assigned a uniform weight, has demonstrated a noteworthy improvement in F1-score compared to unimodal techniques. This outcome underscores the synergistic benefits achieved through the comprehensive analysis of multiple modalities, enhancing the robustness and discriminatory power of our abusive content detection model. Future research endeavors may encompass the refinement of weighting mechanisms and alternative fusion techniques, including the adoption of transformer-based approaches, to enhance modality integration.

References

Premjth B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Nandhini K, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Spandana Reddy Mekapati. 2024. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146.

Shubhankar Barman and Mithun Das. 2023. 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*, pages 217–224.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bharathi Raja Chakravarthi, KP Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, John P McCrae, et al. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1415.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

PD Mahendhiran and SJIJoIT Kannimuthu. 2018. Deep learning techniques for polarity classification in multimodal sentiment analysis. *International Journal of Information Technology & Decision Making*, 17(03):883–910.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Giovanni Pasqualino, Stefano Scafiti, Antonino Furnari, and Giovanni Maria Farinella. 2020. Localizing visitors in natural sites exploiting modality attention on egocentric images and gps data. In *VISIGRAPP (5: VISAPP)*, pages 609–617.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

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*.

B Premjith, Bharathi Raja Chakravarthi, Malliga Subramanian, B Bharathi, Soman Kp, V Dhanalakshmi, K Sreelakshmi, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

Rini Sharon, Heet Shah, Debdoot Mukherjee, and Vikram Gupta. 2022. Multilingual and multimodal abuse detection. *arXiv preprint arXiv:2204.02263*.

Devin Soni and Vivek K Singh. 2018. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194.