AbhiPaw@ DravidianLangTech: Multimodal Abusive Language Detection and Sentiment Analysis using Transformer based architecture

Abhinaba Bala
IIIT Hyderabad, India
abhinaba.bala@research.iiit.ac.in

Parameswari Krishnamurthy
IIIT Hyderabad, India
param.krishna@iiit.ac.in

Abstract

Detecting abusive language in multimodal videos has become a pressing need in ensuring a safe and inclusive online environment. This paper focuses on addressing this challenge through the development of a novel approach for multimodal abusive language detection in Tamil videos and sentiment analysis for Tamil/Malayalam videos. By leveraging state-of-the-art models such as Multiscale Vision Transformers (MViT) for video analysis, OpenL3 for audio analysis, and the bert-base-multilingual-cased model for textual analysis, our proposed framework integrates visual, auditory, and textual features. Through extensive experiments and evaluations, we demonstrate the effectiveness of our model in accurately detecting abusive content and predicting sentiment categories. The limited availability of effective tools for performing these tasks in Dravidian Languages has prompted a new avenue of research in these domains.

Keywords: abusive language detection, sentiment analysis, multimodal analysis, video analysis, Dravidian languages.

1 Introduction

Abusive content, including hate speech, offensive language, and personal attacks, has become prevalent on social media platforms, posing significant challenges to maintaining a safe and inclusive online environment. Detecting and mitigating such abusive language has become an urgent need for social media platforms, content moderators, and society at large. While the detection of abusive language in textual form has received considerable attention, the analysis of multimodal content, specifically videos, incorporating visual, auditory, and textual information, remains a challenging and under explored task (Chakravarthi et al., 2021), (Premjith et al., 2022).

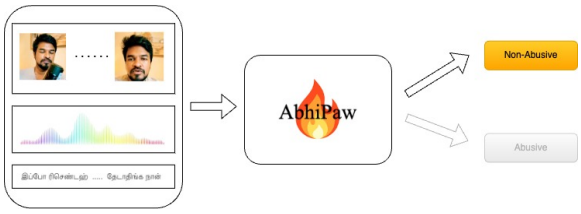

Figure 1: We develop an end-to-end deep network model to learn multimodal representations and perform detection/classification tasks.

The task of multimodal abusive language detection in videos holds great significance due to several reasons. First, the exponential growth of user-generated video content on social media platforms like YouTube demands efficient mechanisms for content moderation and protection against abuse. As videos can convey rich contextual cues through spoken words, facial expressions, and visual content, analyzing multiple modalities becomes crucial to comprehensively understand the abusive intent and impact within such content. By extending abusive language detection to multimodal videos, we can identify and address abusive behaviors more effectively, ensuring a safer and more inclusive digital space.

Second, focusing on videos expands the scope of abusive language detection beyond textual content alone. Abusive language can be embedded in the audio, visual, and textual components of videos, making it essential to develop models that can holistically analyze and interpret these modalities. By leveraging the combined power of visual information, audio cues, and textual context, we can capture nuanced abusive expressions that might be missed by considering only one modality. This multimodal approach enables us to uncover the full spectrum of abusive content, thereby enhancing our ability to combat online abuse.

In this paper, we address the challenge of mul-

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 140–146, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_018

timodal abusive language detection and sentiment analysis in videos, specifically focusing on Tamil (one of the major Dravidian languages spoken in South India and Sri Lanka) and Malayalam(spoken in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry). We propose a novel approach that integrates video, audio, and textual features using state-of-the-art models, including Multiscale Vision Transformers (MViT) for video analysis, OpenL3 for audio analysis, and the bert-base-multilingual-cased model for textual analysis. Through our approach, we aim to advance the field of abusive language detection and/or sentiment classification in videos and contribute to the development of robust models capable of understanding and mitigating online abuse in Dravidian Languages and classify sentiments.

2 Related Work

Multimodal analysis, encompassing tasks such as sentiment analysis, hate speech detection, and humor recognition, has garnered significant attention in recent years. Researchers have explored various fusion methods to effectively combine information from different modalities, leading to improved performance in multimodal analysis tasks. In this section, we review relevant studies and highlight their contributions to the field.

(Zadeh et al., 2017) introduced Tensor Fusion Network to pose the problem of multimodal sentiment analysis as *intra-modality* and *inter-modality* dynamics. (Zadeh et al., 2018) introduced a novel interpretable fusion mechanism called Dynamic Fusion Graph (DFG). (Poria et al., 2016b) described a novel temporal deep convolutional neural network for visual and textual feature extraction and used multiple kernel learning to fuse heterogeneous features extracted from different modalities.

(Majumder et al., 2018) introduced an innovative hierarchical feature fusion strategy that sequentially combines modalities in pairs before fusing all three modalities together. (Poria et al., 2016a) employed a combination of feature-level and decision-level fusion techniques to integrate affective information derived from multiple modalities. (Hazarika et al., 2018) introduced a multimodal emotion detection framework that extracts multimodal features from conversational videos and hierarchically models the self- and inter-speaker emotional influences into global memories. (Liu et al., 2023) propose a cascaded multichannel hierarchical fusion method for multimodal emotion recognition.

(Poria et al., 2017) propose a recurrent model that is able to capture contextual information among utterances. They also introduce attention based networks for improving both context learning and dynamic feature fusion. (Chauhan et al., 2019) introduce a recurrent neural network based approach for the multi-modal sentiment and emotion analysis. The proposed model learns the inter-modal interaction among the participating modalities through an auto-encoder mechanism. They employ a context-aware attention module to exploit the correspondence among the neighboring utterances. (Ghosal et al., 2018) also proposed a recurrent neural network based multi-modal attention framework that leverages the contextual information for utterance-level sentiment prediction. (Chen and Li, 2020) first applies the cross-modal co-attention mechanism to learn the long range of context information and then use a sentimental words classification auxiliary task to guide and learn the sentimental words aware final multimodal fusion representation.

(Han et al., 2021b) propose a framework named MultiModal InfoMax (MMIM), which hierarchically maximizes the Mutual Information (MI) in unimodal input pairs (inter-modality) and between multimodal fusion result and unimodal input in order to maintain task related information through multimodal fusion. (Han et al., 2021a) propose the Bi-Bimodal Fusion Network (BBFN), a end-to-end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations.

(Zadeh et al., 2018) introduce CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), then largest dataset of sentiment analysis and emotion recognition. EmotionLines (Hsu et al., 2018) was introduced as the first dataset with emotions labeling on all utterances in each dialogue only based on their textual content. (Poria et al., 2019) was created by enhancing and extending the EmotionLines Dataset.

A comprehensive understanding of the field can be gained through an examination of the historical context and the establishment of baseline methodologies. Notable works such as (Poria et al., 2018), (Cambria et al., 2017), and (Gandhi et al., 2022) provide valuable insights into the broader perspective of the subject.

3 Method

In this section, we present the methodology employed for the multimodal abusive language detection and sentiment analysis tasks. We first provide an overview of the problem statement and task formulation. Then, we discuss the feature extraction process for video, audio, and text modalities. Next, we describe our approach in detail, including the model architecture and the steps involved. Finally, we outline the training and inference procedures.

3.1 Problem Statement

The aim of this study is to determine whether a particular set of video, audio, and text is abusive or non-abusive. Additionally, we seek to predict the sentiment expressed in the given video clip with audio and text information as a seperate problem. Let (V, A, T) denote the input tuple, respectively for video, audio and text.

3.2 Feature Extraction

For the feature extraction process, we utilize specific techniques for each modality:

Video Features To extract video features, we employ Facebook Research's Multiscale Vision Transformers (MViT). MViT connects the concept of multiscale feature hierarchies with transformer models. The architecture consists of multiple stages that hierarchically expand the channel capacity while reducing the spatial resolution. This creates a multiscale pyramid of features, enabling the modeling of both simple low-level visual information and complex, high-dimensional features. MViT has been shown to outperform other vision transformers in terms of computation and parameter efficiency across various video recognition tasks.

$$F_v = \text{MVIT}(V) \tag{1.a.}$$

Audio Features OpenL3 is specifically designed for audio feature extraction using deep learning models. It provides pre-trained models that can generate high-dimensional embeddings for audio signals. OpenL3 supports different audio feature representations, such as raw embeddings and intermediate representations like log-mel spectrograms. It is particularly useful when you want to leverage the power of deep learning for audio analysis tasks.

$$F_a = \text{OpenL3}(A) \tag{1.b.}$$

Text Features To extract text features, we employ the bert-base-multilingual-cased model. This model is pre-trained on a large corpus of text data and is capable of capturing contextual information across multiple languages, including Tamil and Malayalam.

$$F_t = \text{BERT}(T) \tag{1.c.}$$

3.3 Our Approach

In our approach, which we refer to as AbhiPaw, we leverage the given data comprising three modalities, each ranging from 40 to 80 seconds in length. The AbhiPaw model is built upon a transformer-based architecture, enabling the detection of abusive language content in Tamil videos and separate training for multiclass sentiment analysis on video modalities (Tamil and Malayalam videos).

Modality Separation: The input modalities are separately processed, as discussed in Section 3.2.

Neural Layer Fusion: The separated modalities are then passed through a single neural layer to output features with consistent dimensions. This is akin to normalisation. This step is important for fair fusion since different modalities might have different ranges of values.

$$\begin{aligned} F'_v &= \text{Linear}(F_v) \\ F'_a &= \text{Linear}(F_a) \\ F'_t &= \text{Linear}(F_t) \end{aligned} \tag{2}$$

Positional Encoding: We incorporate positional encoding to capture the spatial information of the modalities. This allows the model to understand the relative positions of elements within each modality.

$$F''_v, F''_a, F''_t = PositionalEncoding(F'_v, F'_a, F'_t) \tag{3}$$

Modality Type Embeddings: Type embeddings corresponding to the three modalities are added. These embeddings do not encode any specific meaning or imposed order but serve to distinguish one modality from another.

$$F'''_v, F'''_a, F'''_t = TypeEmbedding(F''_v, F''_a, F''_t) \tag{4}$$

Classifier Tokens: Similar to the classic CLS tokens in Transformer models, we employ learnable classifier tokens to detect abuse in videos. A single token is used to generate the output.

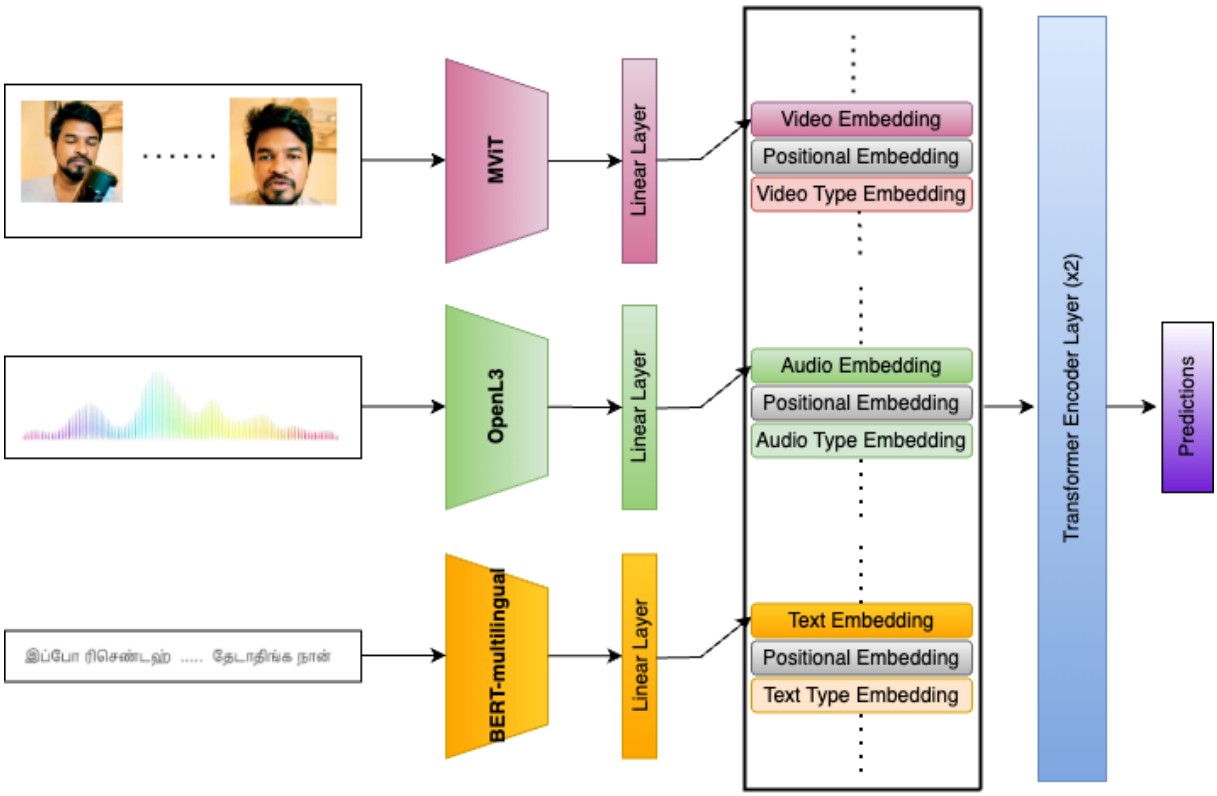

Figure 2: Overview of our work, which is inspired by (Srivastava et al., 2023) and (Hazarika et al., 2020). We utilize separate frozen backbones to get features out of different modalities. We also use linear layers to project them to same dimensionality. To capture positional information between the token we use positional embeddings. For our downstream task of detection, we also add the CLS embeddings before passing to transformer layer followed by linear layer for classification.

Masking: Appropriate masking is applied to prevent self-attention on padded tokens, ensuring accurate attention across the modalities.

Transformer Encoder: The processed features are then passed through a Transformer encoder. Importantly, self-attention across modalities is not applied. For abuse detection, only the outputs corresponding to the classification tokens are considered.

$$z_k = \text{TransformerEncoder}\begin{pmatrix}\text{CLS} + F_v''' \\ + F_a''' + F_t'''\end{pmatrix} \quad (5)$$

Linear Classification: The feature representation from the Transformer encoder is fed into a linear layer for classification. The output logits are compared with the ground truth labels

3.4 Training and Inference

Training Our model is trained end-to-end with *Cross Entropy* Loss and *Adam* optimizer.

Inference We take in un-seen test data and pass to the model to get output.

4 Experiments

4.1 Evaluation Metrics

The evaluation of the model was done based on their F1 score, which is a common metric used in NLP to measure the performance of classification models.

4.2 Datasets

Two different datasets were provided for the shared task at Multimodal Abusive Language Detection and Sentiment Analysis : Dravidian-LangTech@RANLP 2023.

Task 1 : Multimodal detection of abusive content in Tamil: This sub-task involves developing models that can analyze textual, speech and visual components of videos from social media platforms, such as YouTube, and predict whether they are abusive or non-abusive.

Task 2 : Multimodal sentiment analysis in Dravidian languages: This sub-task involves developing models that can analyze textual, speech and visual components of videos in Tamil and Malayalam

from social media platforms, such as YouTube, and identify the sentiments expressed in them. The videos are labelled into five categories: highly positive, positive, neutral, negative and highly negative There are two substasks corresponding to Tamil and Malayalam languages.

4.2.1 Dataset Analysis

Distribution across categories

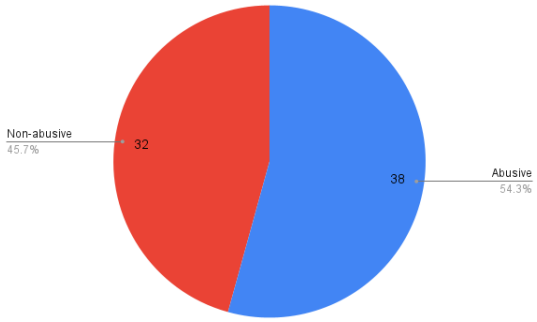

Figure 3: Pie chart showing the number of instances belonging to each category in abusive comment detection task

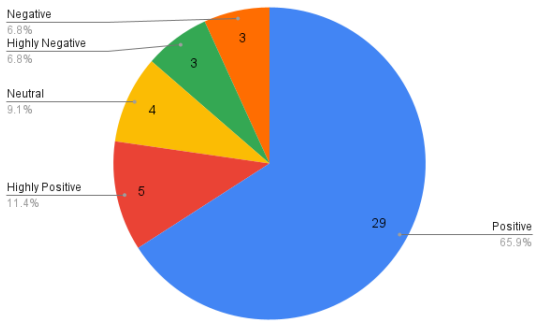

Figure 4: Pie chart showing the number of instances belonging to each category in sentiment analysis task - Tamil

Class Imbalance The training dataset used in our study presents a notable class imbalance issue, with a substantial data points being refered to *Positive*, Figure 4 and 5.

Low sample size The available training data was insufficient in terms of quantity and diversity to fully capture the complexity and variability of the problem domain. We had 70 samples for training for first task, and for second task we had 44 and 50 instances for Tamil and Malayalam sub-tasks respectively.

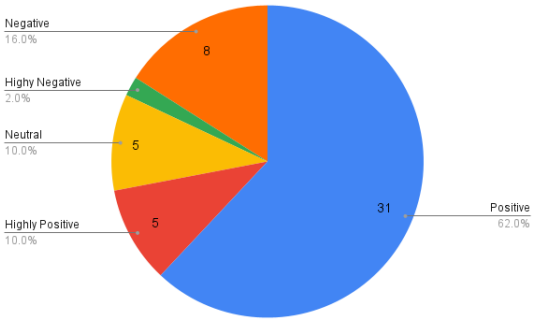

Figure 5: Pie chart showing the number of instances belonging to each category in sentiment analysis task - Malayalam

Table 1: Abusive Language Detection - Tamil

Team	F1-score (macro)	Rank
hate-alert	0.5786	1
AbhiPaw	0.3333	2

4.3 Implementation Details

Our model is trained on a single NVIDIA T4 GPU. We trained our PyTorch model using specific hyperparameters to ensure optimal performance and effective training. The maximum number of epochs was set to 150, allowing the model to undergo multiple iterations through the training dataset. To efficiently process the data, we employed a batch size of 4, which divided the dataset into smaller subsets for parallel computation. Adam optimizer was utilized to optimize the model's weights. Additionally, we set the initial learning rate to 10^{-3}, which determined the step size for adjusting the model's parameters during training. These carefully chosen hyperparameters played a crucial role in achieving the desired results and advancing the effectiveness of our model.

5 Results

We obtain an F1 score of 0.3333 in Abusive Language Detection - Tamil, *Table* 1

For multi-modal sentiment analysis we got an F1 score of 0.1333 for Tamil, *Table* 2 and a score of 0.0923 for Malayalam *Table* 3

Table 2: Sentiment Analysis - Tamil

Team	F1-score (macro)	Rank
hate-alert	0.1429	1
AbhiPaw	0.1333	2

Table 3: Sentiment Analysis - Malayalam

Team	F1-score (macro)	Rank
hate-alert	0.1889	1
AbhiPaw	0.0923	2

6 Conclusion

We present a novel approach for detecting abusive language in low-resource language videos by integrating visual, auditory, and textual features. Our framework demonstrates promising results in accurately identifying abusive content and predicting sentiment categories.

To advance the field, future work should focus on expanding the dataset to address resource scarcity, exploring advanced fusion techniques for multimodal integration, incorporating contextual information and temporal dependencies, and tackling class imbalance challenges. By refining these techniques and considering the linguistic and cultural nuances of Dravidian Languages, we can make significant strides towards ensuring a safer and more inclusive online environment.

7 Acknowledgements

We thank our anonymous reviewers for their invaluable insights and feedback, which have greatly enriched this work. It is important to note that the opinions, conclusions, and findings presented in this material solely represent the views of the authors and do not necessarily reflect the perspectives of their respective graduate institutions or affiliations. We would like to thank Dhruv Srivastava, Aditya Kumar Singh, Prasha Srivastava and Sagar Joshi for their generous assistance and contributions throughout the course of this research. Their support has been instrumental in shaping the development and outcomes of this study. This work was submitted as a part of the DravidianLangTech workshop, 2023 (B et al., 2023).

References

Premjith B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Rajeswari Natarajan, Nandhini K, Abirami Murugappan, Bharathi B, Kaushik M, Prasanth SN, Aswin Raj R, and Vijai Simmon S. 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*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

E. Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R. B. V. Subramanyam. 2017. Benchmarking multimodal sentiment analysis. *ArXiv*, abs/1707.09538.

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*.

Dushyant Singh Chauhan, Md. Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Minping Chen and Xia Li. 2020. SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1067–1077, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2022. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. *Proceedings of the 2021 International Conference on Multimodal Interaction*.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *ArXiv*, abs/2109.00412.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. *arXiv preprint arXiv:2005.03545*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xia Liu, Zhijing Xu, and Huang Kan. 2023. Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion. Florence, Italy. Computational Intelligence and Neuroscience.

N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133.

Soujanya Poria, E. Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038.

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

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016b. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, E. Cambria, Amir Hussain, and Alexander Gelbukh. 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33:17–25.

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.

Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. 2023. How you feelin'? learning emotions and mental states in movie scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amir Zadeh, Minghai Chen, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics*.