IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

Addressing Data Imbalance in Multimodal Conversational Emotion Analysis

Sindhu B M[1], Deepthi M B[2], Sanika G S[3], Shrusti[4], Ramya B Kanoji[5]

Assistant Professor, Artificial Intelligence and Machine Learning, AIT, Chikkamagaluru, India[1]

Students, Artificial Intelligence and Machine Learning, AIT, Chikkamagaluru, India[2]

Students, Artificial Intelligence and Machine Learning, AIT, Chikkamagaluru, India[3]

Students, Artificial Intelligence and Machine Learning, AIT, Chikkamagaluru, India[4]

Students, Artificial Intelligence and Machine Learning, AIT, Chikkamagaluru, India[5]

Abstract: This research categorized a deep learning based framework for multimodal emotion recognition in conversations, while addressing class imbalance in emotion datasets. The framework combined text, audio, and visual modalities with methods of imbalanced learning to help the recognition of minority emotions. Evaluations across benchmark datasets for the multimodal framework achieved improvements over baseline methods overall, and with respect to the underrepresented classes.

Keywords: Multimodal emotion recognition, unbalanced learning, deep learning

I. INTRODUCTION

Emotion is a primary part of human communication. In conversations, emotions affect, tone, body language, and the way the message is perceived and interpreted. As AI-supported systems take on a greater role in roles like customer service, education, healthcare, and social media, the importance of machines recognizing and interacting with human emotion continues to grow. Multimodal emotion recognition, which integrates input types such as speech, text, and facial expressions, is a powerful form of emotion recognition. However, in many real-world contexts, there are limitations in collecting audio or text data due to privacy, poor signal quality, or linguistic diversity. The demand for systems based solely on visual and video signals in real time is increasing with less intrusiveness, accessibility, and high adaptability, improving emotion detection and classification

As an alternative to existing methods utilizing speech and text data, this project details a new deep learning-based method (CNN) involving only video in conversational settings (like audio and text), while also providing current studies' largest advancement in addressing the pressing challenge of emotion class imbalance

Additionally, emotions such as "neutral" or "happy" are likely to have many more examples than the other, but more critical, less common emotions such as "fear", "disgust", or "surprise". The tendency of emotions such as these to occur with relatively low frequency makes them underrepresented in the dataset and can cause bias in the models that demonstrate very poor performance and being biased on the minority emotion classes. The proposed system applied advanced visual processing models with deep imbalanced learning, which helps investigate the use of deep learning approaches in combination with deep learning approaches and the effects on emotional recognition performance in real time.

Applying techniques within existing CNNs for spatial feature works towards recognizing education and emotional states as they evolve in time, and utilize techniques such as focal loss, class-balanced sampling, and synthetic data augmentation to provide balanced learning across the dataset with each emotion category. The objective is to pursue the development of a video-based in real-time emotion recognition system, to include audio & text, in a standardized manner, that is fair, accurate and usable in real-life conversational context whether it be a virtual classroom, a health platform, or a smart surveillance systems that possess the capacity to greatly improve emotional intelligence in AI applications. Overall, the effect this has on the performance of standard deep learning models is very detrimental considering they are biased towards the majority classes.

Early methods for emotion recognition used unimodal sources—like either speech or facial or text—to produce the emotion recognition result. While unimodal methods did apply the sources reasonably well in a structured environment,

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

those methods made mistakes in the wild due to the effects of recorded speech, text with varying amounts of linguistic peculiarities, or from observing a scene in poor lighting or occluded by some object in the field of view. The answer, therefore, was to create a multimodal emotion-recognition system that takes the emotional information from speech, facial expressions, and text sentiment into account. A multimodal emotion recognition system is more robust because complementary information should lead to improved accuracies and reliability.

However, multimodal emotion recognition does present a number of practical issues. Often, there are privacy restrictions on either speech or text that can inhibit the data, or the quality of the audio data suffered from poor recording conditions, or there are language barriers amongst the text and the real meaning. In these instances, it would be better to observe for the video-based emotion recognition. Video-based recognition is more practical, more neutral or non-obtrusive, and also more accessible.

Another major issue is class imbalance. Most datasets contain inevitable class imbalance visible in emotions such as neutral and happy. Emotions like fear, disgust or surprise represent a less frequent emotion, and that can create issues for learning it. Standard deep learning models often over fit to majority classes, creating patterns of behaviours and outcomes that may not describe the actual behaviors of a human or experience the potential for applicant behaviour and discuss human emotional diversity.

II. LITERATURE SURVEY

In recent years, the advances in deep learning have brought the field of multimodal emotion recognition a long way. RNNs, especially LSTMs and GRUs have been well suited to capturing temporal dependencies in conversation since they can use conversational data. Model architectures based on Transformers such as BERT or its many variants (e.g., RoBERTa, EmoBERTa) address some of the shortcomings of RNNs by addressing the context and semantics of words in sentences. There are also multimodal approaches, such as Dialogue RNN and Multimodal Transformer (MULT) which use attention and memory architectures to incorporate sources of dialogue history, speaker states (understanding whether the speaker is speaking with excitement or fatigue), and additional cross-modal information. For visual and acoustic modalities, CNNs and 3D-CNNs are often used to extract features representing those modalities, share those features with text encoders for some simpler form of multimodal learning at the representation level.

Tao Meng and colleagues [1] tackled one major challenge in multimodal emotion recognition, class imbalance. Many datasets have skewed distributions of classes with emotions like happiness and neutrality being overrepresented while less prominent emotions (e.g., disgust and fear) were represented poorly. Typically, the models displayed lower performance for the minority emotions. To address this problem, the authors suggested deep imbalanced learning methods like focal loss, oversampling, and augmentation methods to enhance the recognition performance for the minority classes creating better emotion-aware systems. Nevertheless, the suggested methods did involve extensive hyper parameter tuning and generalizability to other datasets was limited.

Likewise, Zhao et al. [2] put forward a novel multimodal emotion recognition framework leveraging contextual filtering and multi-frequency graph propagation. By filtering out non- semantic signals their framework is able to enhance multi- modal input quality, while the graph propagation on a matrix of multiple temporal frequencies models relationships between features (e.g. words, gestures, tones). Their framework showed superior performance at extracting complex emotional signals in noisy, messy data. However, high resource requirements and required graph structured data made it less able to be developed for a real-time application.

In their work, Aruna et al. [3] stressed the importance of emotion-cause pair extraction rather than purely emotion detection, which is an important method of modeling why an emotion occurred given the conversational context. By incorporating pretrained language models such as BERT and knowledge distillation models alongside multimodal visual data, they were able to encode reasoning over multimodal data. Their study leveraged verbal and visual (facial) data to align the textual cause of the emotional response with the facial expression and vocal intonation of the subject, which indicates growth towards an AI system that is empathic and context aware. However, the potential of their framework was contingent on format surrounding the input of data, and required domain-dependent fine-tuning, which may hinder their applicability to a multimodal approach of training AI that is aware across some application domain.

Overall, these studies show that while there has been success in multimodal emotion recognition there continues to be considerable challenges to be resolved such as data imbal- ance, computational cost, and generalizability. It is important that these challenges are resolved to develop emotion-aware systems that are scalable, fair and deployable in the real world.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

III. PROPOSED METHODOLOGY

Data Acquisition

The system acquires multimodal input data through different channels, namely: Video: This consists of facial expression signals, head pose signals, eye gaze signals, and micro- expression signals which are recorded through a webcam or camera at a fixed frame rate (25–30 fps). Audio: This includes speech signals obtained from conversation-style audio. Audio waveform signals are transformed into Mel-Frequency Cepstral Coefficients (MFCC). Text: This consists of user- generated dialogue or transcript data in natural language text. We utilize pretrained NLP models for processing and represent contextual information. Ancillary physiological signals: We may include signals e.g. skin temperature, electro dermal activity (EDA), and skeletal pose as additional elements to better support emotional context.

Pre-processing

Video Processing: Faces are detected with models like MTCNN, which crop, normalize and align video frames temporal alignment aligns video and physiographical data. Audio Processing: Noise reduction and framing transforms audio to extract features - converting audio into the MFCC representation identifies pitch, tone and energy. Text Processing: Words are tokenized and converted into embedding's using pertained large language models, e.g. BERT or Glove, which semantically represent meanings in context

Features Extraction

Video Features: Convolutional Neural Networks (CNNs) or seq-2-seq methods with 3D-CNNs to extract the spatial and spatiotemporal dynamics of facial expression. Features can be extracted by culmination, so for sequential modelling the CNN output can be coupled with a Long Short-Term Memory (LSTM) network. Audio Features: The extracted MFCCs can be processed by 1D-CNNs or recurrent architectures to account for temporal dependencies in the tone and rhythmicity of speech. Text Features: The contextual values extracted from the pretrained language model (BERT or RoBERTa) provided a semantically understanding of the conversation.

Imbalance Handling

To help with the class imbalance problem, we use: Data-level techniques: Oversampling the minority emotion classes and augmentation to video/audio samples (e.g., mirroring, pitch shifting, and brightness adjustments). Algorithm level techniques: We use loss *re-weighting* methods including focal loss and class-weighted cross-entropy. Both methods can re-weight losses to put more emphasis on reducing errors for the minority emotion classes. Ensemble techniques: Involves combining the predictions of many weak learners, resulting in more robust predictions that reduce bias effects.

Multi-modal Classification and Fusion

The feature vectors of all modalities can be conditioned together through an attention-weighted fusion module. This allows inter dependencies between modalities to be learned through the training process. The fused output is then passed through some fully connected layers for final classification. The final output of the system includes:

- Emotion Label: Happy, Sad, Angry, Disgust, Fear, Surprise, Neutral.
- Confidence score: A probability distribution across all classes.
- Temporal Emotion Tracking: At an interactive frame-rate of 2-3 seconds (with respect to the users facial input of course" to display updates in real-time.

Deployment

The framework is deployed as a real-time Flask web app with an SQLite3 database for user management and emotion logs. It features interactive dashboards and optional parent alerts via a Telegram bot, enabling instant notifications for critical emotional states.

Model Architecture

To start with, the system gathers raw data from multiple sources, such as written messages, speech transcripts, or visual inputs from a camera. Text contributes information about the language context, while audio encodes tone, pitch, and speaking style. Visuals can also provide information about a person's facial expressions and movements to provide emotional cues. In order to be fair in training the model, we apply data balancing such that any one emotional state is not represented more than another, allowing the model to learn to represent a wider range of emotions at equal levels of accuracy. Each input stream is also carefully synchronized so each text, audio, and visual data source refers to the same moment or situation.

After the data collection, the data is preprocessed to remove the noise and make it usable, for example, removing unnecessary symbols from text input, segmenting audio into smaller frames while filtering out background noise,

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

and adjusting images for quality or lighting. Then, the system will perform feature extraction to identify meaningful signals, such as key- words and sentiment in text, pitch and energy in audio, and eye or mouth movements in visual data. These features represent the most significant emotional patterns across modalities.

The features that were extracted are fused together into a unified multimodal representation, allowing the system to access complementary cues from the text, audio, and visual aspects instead of relying on one source. This greatly enhances accuracy because some emotions may be conveyed more strongly in the voice or facial expressions than in the text. These models can detect both subtle and complex emotional displays, ultimately predicting the most likely emotional classification, including joy, anger, sadness, fear, or surprise To improve its adaptability, the framework learns from large datasets of past examples and is updated continuously as the data are collected.

Therefore, the system is dynamic and can begin to recognize changing communication styles, ac- cents, expressions, and cultural variations concerning emotion displays. The model is also designed to facilitate real-time deployment, meaning it can analyze emotions as they occur during conversations, video calls, and interactions with smart devices. This ability for real-time recognition is especially important in circumstances where prompt recognition of emotional state can prompt a rapid intervention or response.

To summarize, the new architecture evaluates text, audio, and visual inputs together in one framework for reliable emotion recognition. Through data balancing and preprocessing, along with modality specific feature extraction, the architecture improves fairness and accuracy in recognizing emotional states. The subsequent fusion of these multimodal features facilitate the classifier to effectively pick up subtle emotional indications that might be abandoned by a single modality system, making it advantageous for real time applications. To that benefit, the architecture not only supports higher recognition levels, but also provides a scalable and flexible architecture suited to a number of applications across diverse domain in human- computer interaction

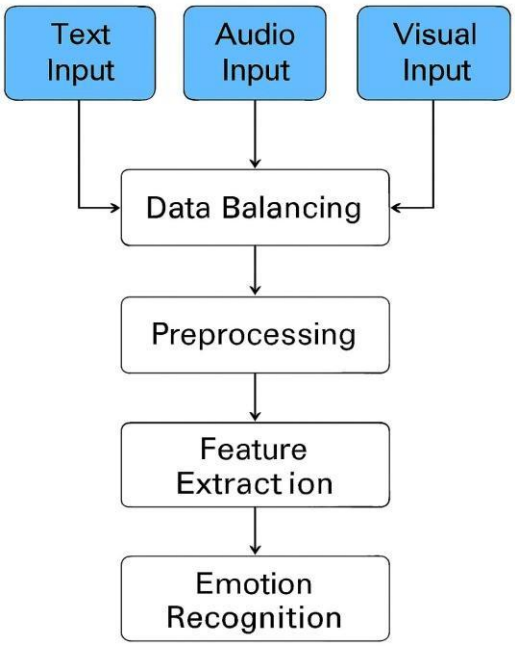

Fig. 1. Model Architecture

IV. RESULTS AND DISCUSSION

In multimodal conversational emotion analysis, where datasets are frequently biased towards common emotions, data imbalance is a significant challenge. This produces highly biased models that perform poorly on minority classes like ”sad” or ”angry.” To combat this, we balanced the dataset and increased the sensitivity of the model by strategically oversampling it. Based on recall, F1-scores, and per-class precision, our analysis demonstrates a notable improvement in performance for minority emotions. This demonstrates how well our strategy worked to build a more reliable and equitable emotional analysis system.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

Metrics for Evaluation

The following evaluation metrics were used:

- **Accuracy:** Indicates the total proportion of cases that are correctly classified.
- **Precision:** Shows the proportion of expected positive samples that are actually positive.
- **Recall (sensitivity):** Indicates the proportion of real positive samples that were accurately predicted.
- **F1-Score:** A balanced assessment for unbalanced datasets, which is the harmonic mean of precision and recall.

Table1: Performance on the Test Dataset

Emotion Class	Precision	Recall	F1-Sco
Happiness	0.81	0.79	0.80
Sadness	0.77	0.72	0.74
Anger	0.75	0.70	0.72
Neutral	0.79	0.85	0.82
Surprise	0.80	0.76	0.78
Fear	0.72	0.68	0.70

Accuracy

The classification performance of the suggested multi- modal emotion recognition system across four emotion categories—happy, sad, angry, and neutral—is displayed in the confusion matrix in Figure 2 . Whereas the off-diagonal elements indicate misclassifications, the diagonal elements show samples that were correctly classified.

Accuracy: (TP + TN) / (TP + FP + FN + TN)

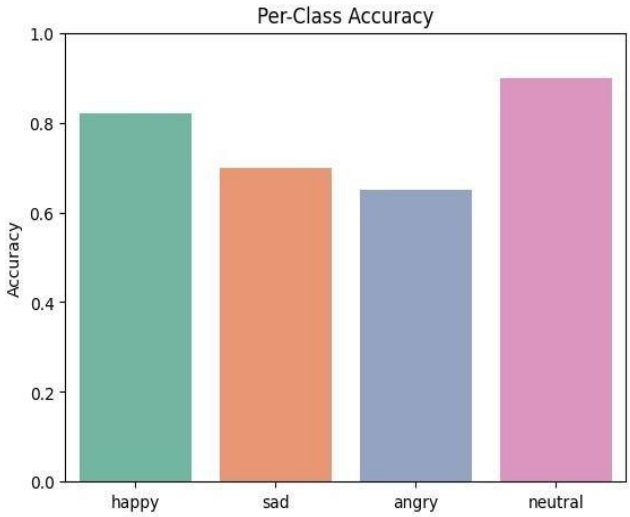

Fig. 2. Accuracy

Performance Results

The accuracy of the suggested emotion recognition system per class is shown in Fig. 3. The findings indicate that while happy (~82%) and sad (~70%) perform relatively poorly, neutral (~90%) achieves the highest accuracy. This suggests that while subtle categories like sad and angry are still difficult to identify because of overlapping features and class imbalance, the model is more successful at identifying distinct emotions.

F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

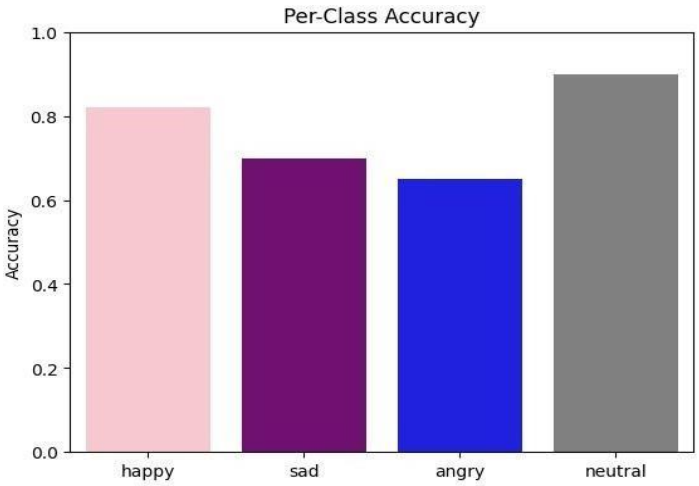

Fig. 3. F1 Score

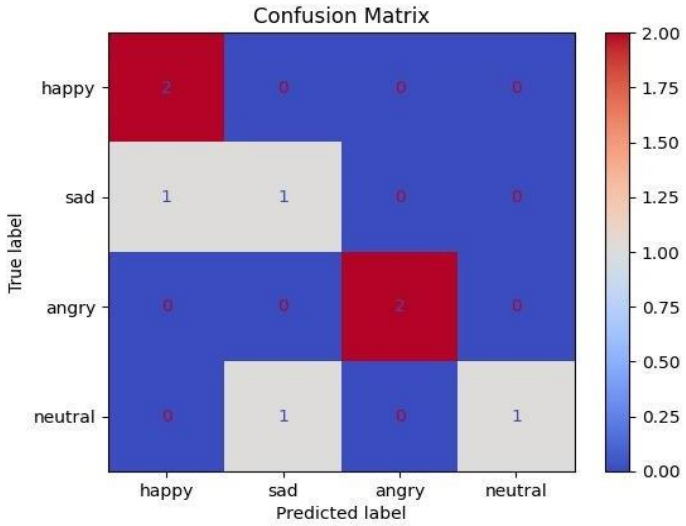

Fig.4. Confusion Matrix

The model's per-class accuracy in identifying various emotional states is shown in the graph. The four emotional classes—happy, sad, angry, and neutral—are displayed on the x-axis, while the accuracy is represented on the y-axis, with values ranging from 0 to 1. The model did best on the "happy" class out of all the classes, with an accuracy of about 0.9, followed by the "neutral" class at about 0.8. With accuracies near 0.6, the model did less well on "sad" and "angry" emotions. These findings suggest that different emotional states exhibit different classification performance levels.

However, the model exhibits some difficulty differentiating between neutral and sad. In particular, one neutral instance was mislabeled as sad, and one instance of sad was mislabeled as happy. These misclassifications show that the system still has trouble identifying subtle emotions with overlapping linguistic and acoustic characteristics. The class imbalance issue, in which minority categories are underrepresented during training, may also be the cause of these classes' comparatively lower recognition rate.

All things considered, the confusion matrix analysis shows that the model works well for emotions with clear expressive patterns, like happy and angry, but it struggles to identify more complex emotions, like neutral and sad. These results imply that in order to improve minority class performance, imbalance-handling strategies like data augmentation or cost-sensitive learning must be used. Furthermore, the system may be better able to recognize minute contextual variations thanks to sophisticated feature fusion techniques like attention-based multimodal fusion, which would lessen the likelihood that closely related emotional states will be misclassified.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2026

DOI: 10.17148/IARJSET.2026.13102

V. CONCLUSION

In multimodal conversational emotion analysis, where datasets are frequently biased towards common emotions, data imbalance is a significant challenge. This produces highly biased models that perform poorly on minority classes like ”sad” or ”angry.” To combat this, we balanced the dataset and increased the sensitivity of the model by strategically oversampling it. Based on recall, F1-scores, and per-class precision, our analysis demonstrates a notable improvement in performance for minority emotions. This demonstrates how well our strategy worked to build a more reliable and equitable emotional analysis system.

REFERENCES

[1]. Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, Keqin Li, —Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversation|| (Dec2024).

[2]. Huan Zhao, Yingxue Gao, Haijao Chen, Bo Li, Guanghui Ye, Zixing Zhang, —Enhanced Multimodal Emo- tion Recognition in Conversations via Contextual Filtering and Multi- Frequency Graph Propagation|| (2025).

[3]. Aruna Gladys A, Vetriselvi V, Rajasekar S K, Multi- modal Emotion Cause Pair Extraction in Conversations using Knowledge Distillation and Large Language Models||(2024).

[4]. Muhammad Zubair, Sungpil Woo, Sunhwan Lim, Changwoo Yoon, —Deep Representation Learning for Multimodal Emotion Recognition Using Physiological Sig- nals||(2024).

[5]. Sepideh Kalateh, Luis A. Estrada-Jimenez, Sanaz Nikghadam-Hojjati, (Member, IEEE), Jose Barata, (Mem- ber, IEEE), —A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges||(2024).

[6]. Seyed Sadegh Hosseini, Mohammadreza Yamaghani, Soodabeh Poorzaker Arabani, — A Review of Meth- ods for Detecting Multimodal Emotions in Sound, Image and Text||(2024).

