Emotion sentiment analysis from text and image using machine learning models

Kanishk[1]*, Akshi Pathania[2]*, Divya Rani[3]*, Akanksha Puri[4], Anamika Rangra[5]

{kanishkpathania5@gmail.com[1], akshipathania001@gmail.com[2], Divyakanungo00@gmail.com[3] }

[1,2,3,5]Himachal Pradesh Technical University Hamirpur, India, [4]National Institute of Technology Hamirpur, India

Abstract. Artificial intelligence is ability of machines to think and act like humans. This research uses AI, NLP to extract emotion and sentiments of humans. Emotion and Sentiment analysis have vital role in development of intelligent systems that can understand human behaviour. Analysis of human behaviour using either text or image, unimodal approach often fails to capture full human emotions. Hence, in this paper we identify the emotion and sentiments of humans by text and image by using multimodal. Multimodal means using more than one type of data-like combining text, image, audio or video, to understand and solve problems. For text-based analysis, use the MELD dataset, and for image- based analysis, use the FER2013 dataset. Using the SBERT, for textual embedding, enabling deeper contextual understanding and Convolution Neural Network for image classification. Experimental result shows that the text-based modal achieves the accuracy of 95% and image- based achieves the 72.41%. Merged the extracted features from both modalities using Random Forest algorithm and by using this accuracy achieved is 79.77%. This study demonstrates the effectiveness of a multimodal approach to emotion and sentiment analysis by combining both text and image information. This significantly improve the classification performance compared to unimodal. Future work may involve integration with speech, emotion tracking over time, language independent emotion classification.

Keywords: Sentiment Analysis, MELD (Multimodal Emotion Lines Dataset), FER2013, SBERT (Sentence Bidirectional Encoder Representation from Transformers), CNN (Convolutional Neural Network).

1 Introduction

In today's digital era, it has become necessary to understand human behavior in order to improve the interaction between humans and machines. Human behavior can only be understood when we know about the sentiment and emotion of humans. Sentiments and emotion play a crucial role in decision-making, perception, and social interaction. With an increase in dependency on technology in daily life, the demand for intelligent systems which can respond according to human emotions and sentiments has grown significantly. Emotion and sentiment analysis, a subfield of artificial intelligence (AI), plays a key role in developing such systems by enabling machines to interpret the emotional tone and sentiment expressed through text, speech, facial

* These authors contributed equally to this work.

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362581

expressions, and other modalities.

Traditionally, sentiment and emotion analysis only used a single modality, like only text or image or voice, but not multimodal. However, analyzing a single modality may not fully capture the complexity of human emotions and sentiments. Humans frequently express their emotions and sentiments through a combination of verbal and non-verbal ways, such as text and facial expressions. Therefore, using multiple modalities, such as text and image, can improve the accuracy of emotion and sentiment understanding.

Machine learning algorithms have been used to capture and recognize the patterns from both textual and visual data. For text inputs, model SBERT is utilized, which is productive in extracting semantic meaning and emotional context. For image-based input, CNNs are used because of their ability to learn and recognize facial features. Datasets like MELD for text and FER-2013 for images are useful for training and evaluating such models. By fusing or merging the outputs of both models, the system attempts to provide a better understanding of the user's emotional state.

The main objective of this study is to increase classification performance over various emotion categories, as well as sentiment categories. The model is evaluated on datasets (MELD and FER2013), achieving great results, with up to 72.41% accuracy in image-based emotion classification and 95% in text-based sentiment detection. By extracting the features from both modalities using the Random Forest algorithm, it has achieved an accuracy of 79.77%. This system has wide applications in fields such as healthcare, customer service platforms, education, and human-computer interaction.

2 Motivation

Significance of Problem: Understanding human emotions and sentiments is a critical task in the modern digital ecosystem, where individuals interact with technology through text, images, voice and other online expressions. The way people interact or express opinions with each other is shifting from simple text-based to other multi-modal forms. The past research is mostly focused on single modalities, such as only textual data, which often fail to capture accurate human emotional expressions. Social media platforms, customer service systems, digital classrooms and e-shopping platforms are relied heavily on user sentiments and emotions to provide meaningful messages, feedback and responses.

Necessity of this Research: The motivation for this research is that using a single modality cannot fully understand human emotions. Human feelings are often expressed through multiple ways - such as text, images, facial expressions, and tone – so relying on one type of data often leads to incomplete or incorrect results. Therefore, a multimodal approach that helps integrate both text and image features can significantly improve the accuracy and reliability of emotion and sentiment detection models.

Another motivation comes from the challenges identified in existing literature—such as label inconsistencies across datasets, overfitting in image-based models, limited contextual understanding in text-based models, and the lack of real-time deployable architectures. Our study addresses these gaps by introducing a **joint-labelled multimodal dataset setup**, a late-fusion mechanism using Random Forest, and a productive combination of SBERT and MobileNetV2 to design a scalable, generalizable, and practical solution.

Context Leading to this Research: Platforms like e-shopping apps, social media websites, and customer service systems depend on users' feedback to understand their emotions, feelings and experiences. As people now express feelings through text, images, and other digital forms, single-modality methods often fail because they can capture only a single type of emotional signal, such as text or image alone. This leads to incomplete understanding of a user's true feelings. Therefore, there is a strong need for improved multimodal emotion analysis that can combine information from multiple sources for better accuracy.

3 Related Work

The topic of multimodal emotion and sentiment analysis has attracted much interest over the past few years, particularly due to the rising accessibility of deep learning technology and multimodal data sets. Our application is the contribution to this study since it suggests a hybrid model that consists of SBERT as the textual emotion representation method and CNN as the facial emotion expression method. Another distinctive aspect of our strategy is that we use a multimodal dataset which is jointly labelled so that the emotion tags that are used in text and image modalities are the same. This enhances the alignment of labels, which is one of the demerits of most current multimodal research. Our model uses a hybrid fusion strategy to integrate SBERT embeddings with CNN features, enabling better recognition of nuanced emotions. The system is evaluated on standard datasets like MELD and FER-2013 and shows improved performance over unimodal baselines, particularly in detecting complex or overlapping emotional expressions.

A late fusion is proposed in the present paper where we perform the facial emotion detection through the FER2013 dataset using the MobileNetV2 model and the SBERT embedding + XGBoost on the text emotion recognition on the MELD dataset. We use a Random Forest classifier to mix the predictions of the two models with the expectation that the accuracy of the emotion recognition in general could be enhanced by utilizing both an image and the textual data in a rather effective way.

Several prior studies have explored related directions. Zadeh et al. [1] introduced the Tensor Fusion Network (TFN), which explicitly models intra- and inter modality dynamics across text, audio, and video. Tzirakis et al. [2] proposed an end-to-end CNN and LSTM framework combining audio and visual cues. Poria et al. [3][4] also investigated multi-party conversations using attention-driven LSTM-based models on the IEMOCAP and MELD datasets.

Priyasad et al. [5] implemented cross-modal attention mechanisms to dynamically weight each modality. Zhining He et al. [6] used simple concatenation-based fusion and highlighted the importance of modality specific preprocessing. Kumar et al. [7] presented VISTANet, combining image, speech, and text in a hybrid explainable model. Lai et al. [8] gave an extensive overview of the fusion techniques and difficulties, dividing them into early, late, and hybrid fusion. The study by O Adel et al. [9] was focused on multimodal learning of social media spectrograms. Multimodal emotion recognition in 4 modalities (text, image, speech, and motion) was performed with the use of DeBERTa and 3D CNNs [10], on which complex architectures exhibited significant improvements. Alzamzami and Saddik [11] suggested a transformer-based feature fusion method for recognising multimodal sentiment by effectively merging textual and visual features from actual Twitter data. In contrast, Zhang et al. [12] presented a cross-modal semantic fusion network with alignment mechanisms to minimise the

semantic disparity between text and image modalities. Finally, Sailunaz and Alhajj [13] examined emotion and sentiment analysis of Twitter text, emphasising the significance of incorporating emotional context into sentiment classification. In comparison to these works, our manner also focuses on simplified joint-text-image fusion with readily available architectures (SBERT and CNN) and puts emphasis on label consistency, a factor that is frequently ignored. This renders our model more implementable (as well as deployable) in a real-time or resource - constrained setting and yet it also successfully attains an excellent accuracy with regard to the tasks of emotion and sentiment classification.

Table 1: Comparison Table of Key Related Works

Study	**Modalities**	**Fusion Technique**	**Dataset(s)**	**Key Contribution**
Ours	Text + Image	Hybrid fusion (SBERT + CNN)	MELD +FER-2013	Joint-labeled dataset, fusion architecture
Zadeh et al. [1]	Text + Audio+ Visual	Tensor Fusion Network (TFN)	MOSI, MOSEI	Cross-modality tensor interactions
Tzirakis et al. [2]	Audio +Visual	CNN + LSTM	RECOLA	End-to-end multimodal fusion
Poria et al. [3][4]	Text + Audio+ Visual	Attention-based LSTMs	IEMOCAP, MELD	Multi-party emotional dialogue modeling
Priyasad et al. [5]	Text + Audio	Cross-modal attention	IEMOCAP	Dynamicattentionfor modalities
Zhining He et al. [6]	All three	Simple concatenation	AVEC, MOSI	Lightweight fusion architecture
Kumar et al.[7]	Image+Speech+ T ext	Hybrid + Explainability	IIT-RMMEmoRec	Explainable multimodal model (VISTANet)
Lai et al.[8]	Multiple modalities	Survey of fusion strategies	Various	Comparative study of early,late, hybrid
O Adel et al. [9]	Image+ Audio	Concatenated CNN	Social Media	Spectrogram CNN for social emotion detection
E. Boitel [10]	Text + Image + Audio	Multi-stream fusion	SAVEE, BAUM-1	DeBERTa + ResNet + 3D-CNN
Alzamzami & Saddik [11]	Image+ Audio	Transformer-based feature fusion	Twitter (in-the-wild tweets)	Uses transformers to fuse visual and textual features for improved sentiment recognition in noisy real-world data

Zhang et al. [12]	Text + Image	Cross-modal semantic fusion with alignment	Multimodal sentiment datasets	Introduces cross-modal alignment to better connect text and image semantics
Sailunaz & Alhajj [13]	Text	No fusion (single modality)	Twitter dataset	Combines emotion + sentiment detection from tweets using NLP techniques

4 Our Work

The key findings of this study can be concluded as follows:

- We suggest an efficient multimodal system of emotion and sentiment analysis that combines textual embeddings of SBERT and visual features of CNN.
- We introduce a joint-labelled approach that aligns emotion categories across MELD (text) and FER-2013 (image) datasets for consistent multimodal learning.
- We implement a late-fusion strategy using Random Forest to combine predictions from both modalities, improving overall accuracy and robustness.
- Our model achieves strong performance with 94.98% accuracy on text, 72.41% accuracy on images, and 79.77% accuracy in multimodal fusion—surpassing several baseline models.
- The proposed system is simple, scalable, and suitable for real-time emotion-aware applications such as social media analysis, customer support, and human–computer interaction.

The study illustrates a multimodal machine learning system, that is, a combination of NLP and visual sentiment/emotion -detecting analysis. It solves text and image data by means of an ordered manner. ML with using BERT/RoBERTa and ResNet. Late Fusion combines the results of the two modalities, enhancing accuracy. The model was reviewed using the traditional metrics and is cross-validated to enable holistic real-world inputs. Such applications span human-computer interaction, mental health, customer testimonials, and social media and provide a sound basis of emotion -conscious intelligent systems.

4.1 Dataset Used

In the present research project that is sentiment analysis and emotion detection based on multimodal data-oriented research, two different datasets have been used, i.e. one using text data and another using image-based data.

In the case of the textual modality, we have used the MELD (Multimodal EmotionLines Dataset), a well-annotated corpus of dialogues of the popular TV series friends. It offers utterance-level remarks in both sentiment and emotion and hence it is suitable in supervised learning tasks in natural language processing. Namely, we have exploited the dev_sent_emo.csv file, which can be found at github [19] and it is a csv file, used on processing model of machine learning.

In the case of the image modality, we used FER-2013 (Facial Expression Recognition 2013) which can be accessed on Kaggle and can be downloaded [20]. This data will have thousands of grey scale facial images that are classified under different emotions namely happiness, sad, anger and neutral. The FER-2013 dataset has been selected because it is of much importance in tasks that revolve around emotion classification. The two datasets were suitably assembled and divided to fit the model training, validation as well as testing processes as per the need of system specification.

4.1.1 Text Dataset: MELD (Multimodal Emotion Lines Dataset):

MELD has good quality labels at utterance level and most appropriate in multi- turn tasks of classification of dialogue emotions. All the utterances of the dataset are tagged with one of the seven distinct categories of emotion and one of three classes of universal sentiments, which are used as basis to test and train models using natural language..

The table below provides a conclusion to the utterance's distribution in our chosen MELD CSV file (dev_sent_emo.csv):

Table 2: Emotion and Sentiment Distribution in MELD Dataset

Emotion	Utterances	Sentiment	Utterances
Anger	153	Negative	408
Disgust	22	Positive	233
Fear	40	Neutral	940
Joy	163		
Neutral	470		
Sadness	111		
Surprise	150		

4.1.2 Image Dataset: FER-2013 (Facial Expression Recognition):

This dataset has grayscale face individual photos, each given different emotional labels, and is very popular due to emotion classification tasks. When carrying out this project, we particularly targeted four major classes of emotions extracted in the dataset:

1.)Happy **2.) Sad**

3.)Angry **4.) Neutral**

The selection of these classes was done on the basis of relevance and frequency on analysis of expression of emotion as occurs in the real world. The FER-2013 dataset is provided into two major directories namely train and test which could be handled in an organized way which means models can be developed, trained and their performance evaluated.

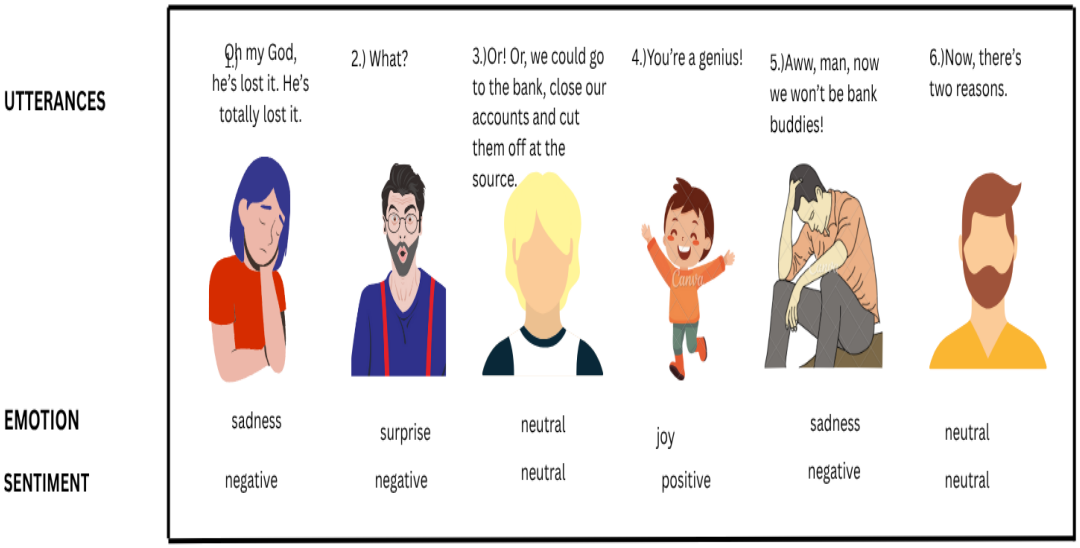


Figure 1: Emotion and sentiment shift of speakers in a dialogue

5 Methodology

The studies used in this research involve a two-modality approach, which assesses both textual and image data to determine mood and emotion. The methodology will be structured into three significant phases: Pre-processing, Feature Extraction and Classification and finally Fusion Strategy in order to integrate two modality results..

5.1 Text Modality

Intensive pre-processing of text data is performed including:

- **Lowercasing:** Converting all text to lowercase
- **Stop Word Removal:** Eliminating non- informative words (e.g., "is", "the")
- **Tokenization:** Segmenting text into meaningful units
- **Lemmatization**: Reducing words to their base form

After pre-processing, textual features are retrieved using Sentence-BERT (SBERT), which produces dense sentence-level embeddings. These embeddings are classified using XGBoost, an efficient gradient boosting classifier. The complete text processing workflow is shown in **Figure 2.**

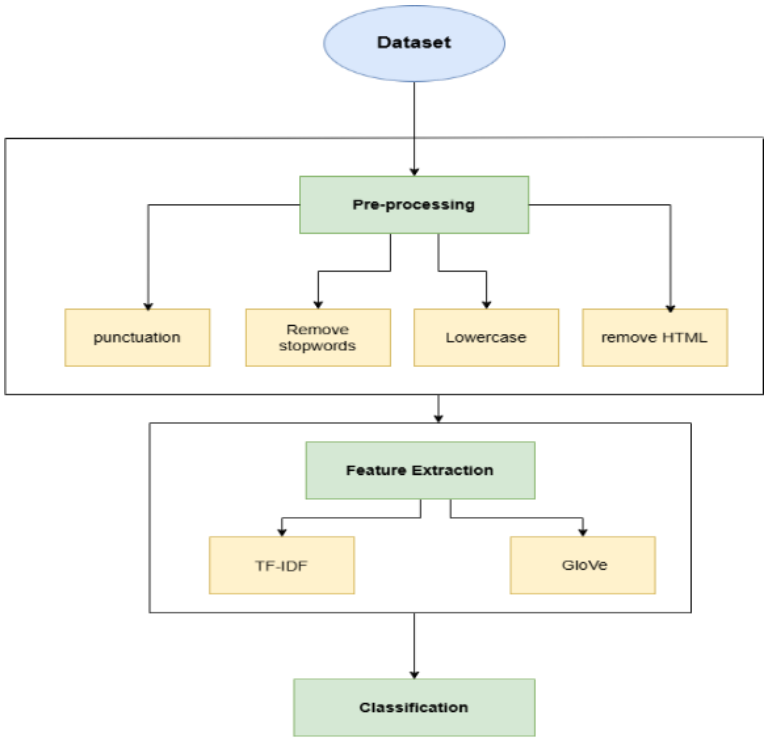


Figure 2: Text Modality Pipeline

5.2 Image Modality

For image data, pre-processing includes:

- Resizing and normalization
- Noise reduction
- Data augmentation techniques (e.g., flipping, zooming)

Feature extraction is carried out using CNNs, where convolutional layers act as automatic feature extractors. The output is flattened and passed through fully connected layers to classify the image into the appropriate emotional category. Image pipeline is presented in **Figure 3**.

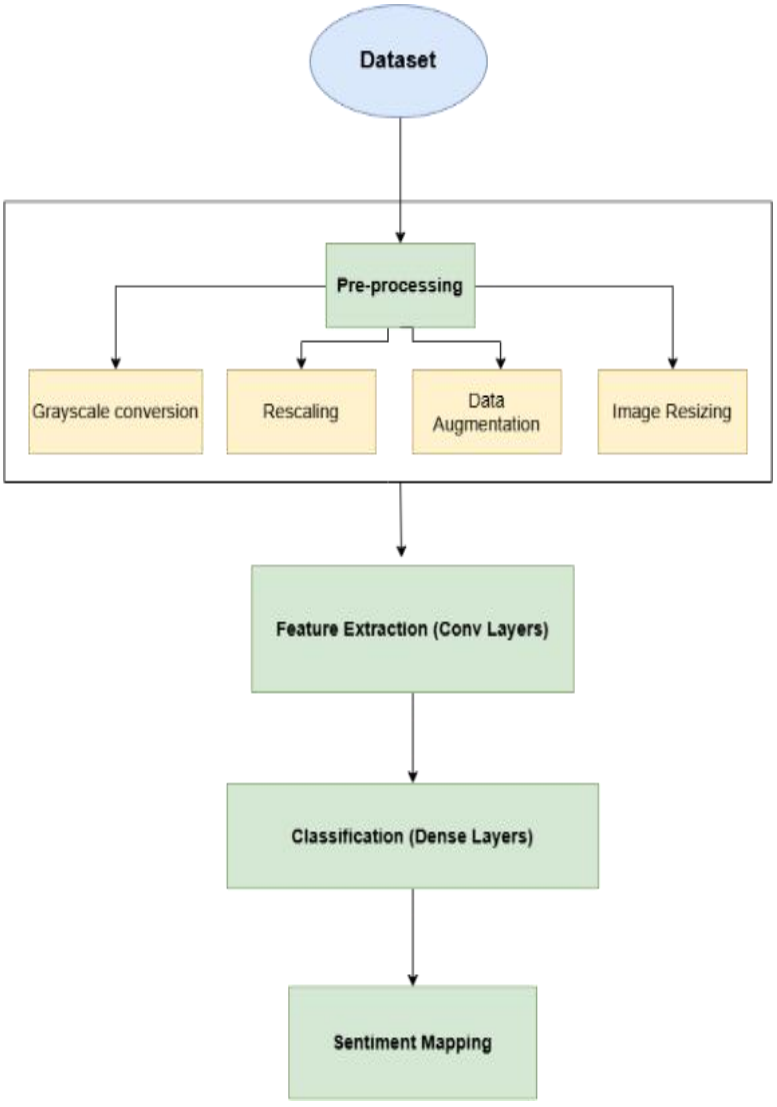


Figure 3: Image Modality Pipeline

5.3 Fusion Strategy: Late Fusion

To combine insights from both modalities, a late fusion strategy is applied where individual predictions from the text-based (SBERT + XGBoost) and image-based (CNN) models are merged at the decision level by using Random Forest algorithm of machine learning. Each model outputs a probability distribution over class. These outputs are then combined using weighted averaging, depending on the performance of each model during validation.

Late fusion enhances the overall accuracy of the system by leveraging complementary strengths of each modality, thus producing a unified prediction, such example shows as in **figure 4**.

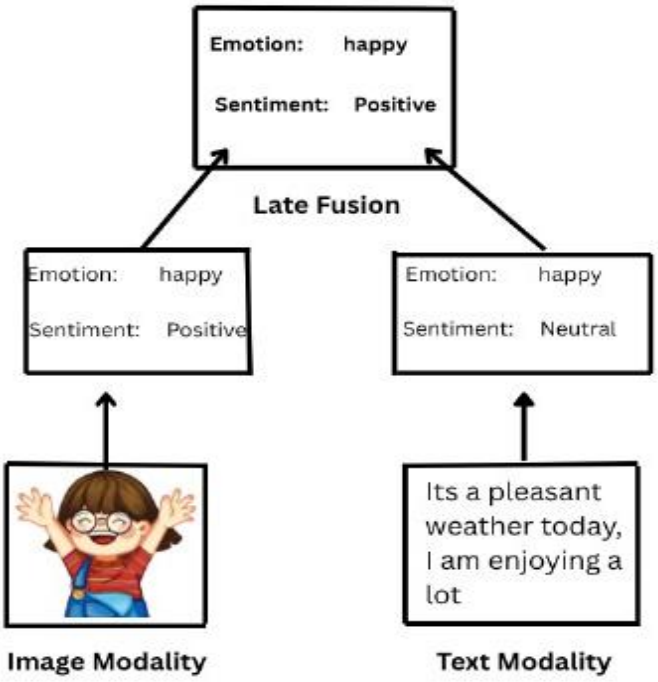


Figure 4: Late Fusion

6 Experiments

6.1 Datasets

We tested our suggested multimodal emotion and sentiment detection model on two publicly reachable datasets, FER2013 and MELD. There are more than 35,000 greyscale images of faces in FER2013 divided into 7 categories of emotions. In this study, we were interested in four categories, namely angry, happy, sad and neutral. MELD has more than 13,000 text utterances in multi-party conversations tagged with the seven categories of emotions: anger, disgust, fear, joy, neutral, sadness and surprise.

To harmonize emotion categories across modalities, we projected the seven divisions of MELD to four common divisions corresponding to FER2013. We divided both datasets to have training (80%), validation (10%), and testing (10%) ones. Preprocessing of the data involved cleaning of text data and normalization of image data as well as balancing classes to make the model training and evaluation robust.

We also implemented uniform preprocessing techniques based on the past works on multimodal sentiment and emotion analysis but adapted them to our particular model (MobileNetV2 and SBERT based feature extraction pipelines).

6.2 Feature Extraction

To extract image features, MobileNetV2, pretrained on ImageNet, was fine-tuned using transfer learning with data augmentation on FER2013 images, resized to 96x96. For text, SBERT generated 768-dimensional embeddings from MELD utterances, later classified using XGBoost. Both SBERT and CNN features were saved as CSV files for analysis and training.

6.3 Implementation Details

In our tests we separated the FER2013 as well as MELD into training and test sets with 80 per cent training and 20 per cent testing. The training was conducted on Adam optimizer with the initial learning rate of 0.0001, which was halved when the validation performance stopped improving. The size of the batch was chosen to be 32 in the image model (MobileNetV2) and 16 in the text model (SBERT + XGBoost) because of the limitations in the amount of GPU memory in Google Colab. To prevent overfitting, we used both early stopping (patience=5) and model checkpointing. Random Forest: Roughly 100 estimators were used in the fusion model. We measured the models based on accuracy, precision, recall, and F1-score that were calculated as follows:

$$Accuracy = TP/(TP + FP). \quad (1)$$

$$Recall = TP/(TP + FN). \quad (2)$$

$$F1 - score = 2 \times Precision \times Recall/(Precision + Recall) \quad (3)$$

Where TP is true positives, FP is false positives and FN is false negatives. We fully installed our pipeline TensorFlow, PyTorch, Scikit -learn and Sentence Transformers.

7 Baselines and Comparison

We compare our multimodal emotion-sentiment system with baselines in text, image, and fusion tasks using optimal hyperparameters across models like Text-CNN, ResNet50, and Random Forest variants. Reported metrics are accuracy, precision, recall and F1-score. In Tables 3, 4, and 5, comparative results are given, indicative of the improvement of our methods over those that are based.

Table 3: Text Classification Baselines

Model	Accuracy	F1-score
text-CNN	81.2%	0.79
bcLSTM	84.5%	0.82
DialogueRNN	88.7%	0.87
SBERT+XGBoost (Ours)	94.98%	0.95

Table 4: Image Classification Baselines

Model	Accuracy	F1-score
CNN	65.4%	0.62
ResNet50	70.2%	0.68
MobileNetV2 (Ours)	72.41%	0.71

Table 5: Fusion Baselines

Method	Accuracy	F1-score
Majority Voting	73.4%	0.72
Early Concatenation + RF	75.6%	0.75
Weighted Fusion RF (Ours)	79.77%	0.79

8 Analyses of Results

We give sentiment and emotion classification results with text, image and fusion models. The SBERT+XGBoost found the best accuracy of 95% using a text-based representation of data which showed how many language cues are important in the emotion and sentiment prediction. The accuracy rate of the model of image based on MobileNetV2 was 72.41%, which demonstrates its inefficiency when used independently but its applicability in the detection of visual emotion indications. The hybrid model that merges the image and text predictions based on the late fusion approach of Random Forest scored 79.77 per cent.

This shows that the combination of modalities enhances the performance in comparison to images only, but text holds the greatest predictive power. On balance, the fusion model makes more balanced predictions, which makes it possible to increase its effectiveness in the future, due to more advanced mechanisms of fusion.

9 Conclusions

This research paper includes a multimodal approach for sentiment and emotion analysis by combining textual and visual data using SBERT and CNN, respectively. Our result shows the compatibility of datasets with each other despite the different data in them. Experiments on the MELD and FER-2013 datasets confirmed the effectiveness of this fusion based architecture in recognizing a diverse range of emotions. It is used to show the emotion and sentiment of both text and image in one output by using the fusion technique. Hence, this project is overall interactive and effective to analyze the emotions and sentiments of humans interacting with machines and make the interaction effective.

10 Challenges

Labelling text and image data for sentiment and emotion analysis was challenging due to differing emotion definitions in FER-2013 and MELD. Subtle or mixed emotions, label misalignment, data imbalance, and similar emotional states often led to classification ambiguity and reduced model accuracy, especially on real-world and unseen inputs.

11 Future Scope

The present project was able to apply the labelled multimodal dataset, in which the emotion and sentiment labels were applied to the text and image input. Nevertheless, some extensions can be made to enhance the system by even more people to produce a scalable, generalizable, and workable solution as follows:

Advanced Fusion Techniques: Future research can be able to use more advanced fusion techniques like early fusion, cross-modal attention, transformer-based fusion, or graph neural networks to promote the interaction of text and image features.

Dynamic and Real-Time Inputs: The existing model is applicable to the case of working with static images and real -time texts. Live video processing and audio emotion recognition as well as real -time streams can be added to it in the future.

Personalized Emotion Recognition: The system of recognizing emotions of a particular user can be customized to fit the individual variation of expressions of faces and language styles.

Context-Aware Analysis: It is possible to introduce the context knowledge (user history, background information, or situation-stipulated hints) to improve the accuracy of perception of emotions.

Emotion-Driven Recommendations: It offers more interactivity and support to the users, as the system can accept emotion-based music recommendations, activity recommendations, or personalized recommendations.

References

[1] A. Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," EMNLP, (2018). 10.18653/v1/D17-1115

[2] P. Tzirakis , G. Trigeorgis et al., "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," IEEE Transactions on Affective Computing, (2017). 10.1109/JSTSP.2017.2764438

[3] S. Poria et al., "Context-Dependent Sentiment Analysis in User-Generated Videos," ACL, (2017). 10.18653/v1/P17-1081

[4] S. Poria et al., "Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines," IEEE Intelligent Systems, (2018). 10.1109/MIS.2018.2882362

[5] D. Priyasad et al., "Attention Driven Fusion for Multi-Modal Emotion Recognition," arXiv:2009.10991, (2020). 10.1109/ICASSP40776.2020.9054441

[6] He, Z., Xiao, Y. LPGNet: A Lightweight Network with Parallel Attention and Gated Fusion for Multimodal Emotion Recognition. *Circuits Syst Signal Process* (2026). 10.1007/s00034-026-03547-w

[7] P. Kumar et al., "VISTANet: Explainable Fusion for Emotion Recognition," arXiv:2208.11450, (2022). 10.1109/TAFFC.2025.3573180

[8] S. Lai et al., "Multimodal Sentiment Analysis: ASurvey," ACM Computing Surveys, (2023). 10.1016/j.displa.2023.102563

[9] O Adel et al., " MM-EMOR: Multi-modal emotion recognition of social media using concatenated deep learning networks," Sensors, MDPI, (2024). **10.3390/bdcc7040164**

[10] Enguerrand Boitel, Alaa Mohasseb, Ella Haig,"MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis," Expert Systems with Applications, (2025). 10.1016/j.eswa.2024.126236

[11] F. Alzamzami and A. E. Saddik, "Transformer-Based Feature Fusion Approach for Multimodal Visual Sentiment Recognition Using Tweets in the Wild," in *IEEE Access*, vol. 11, pp. 47070-47079, 2023, doi: 10.1109/ACCESS.2023.3274744

[12] Zhang, Shunxiang and Liu et al.,"A Multimodal Semantic Fusion Network with Cross-Modal Alignment for Multimodal Sentiment Analysis" 10.1145/3744648

[13] Kashfia Sailunaz a b, Reda Alhajj, " Emotion and sentiment analysis from Twitter text", 10.1016/j.jocs.2019.05.009

[14] Abayomi Bello,Sin-Chun NgORCID andMan-Fai Leung . "A BERT Framework to Sentiment Analysis of Tweets", 10.3390/s23010506

[15] Arghya Sahoo; Ritaban Chanda; Nabanita Das; Bikash Sadhukhan, "Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data" ,10.1109/ICSCC59169.2023.10335061

[16] Manju Priya Arthanarisamy Ramaswamy, Suja Palaniswamy "Multimodal emotion recognition: A comprehensive review, trends, and challenges", **10.1002/widm.1563Digital Object Identifier (DOI)**

[17] Nabeel Al-Milli et al, "Emotion detection and sentiment analysis of textual data", 10.1016/j.eij.2026.100976

[18] Charalampos M. Liapis et al, "Enhancing sentiment analysis with distributional emotion embeddings", 10.1016/j.neucom.2025.129822

[19] https://github.com/declare-lab/MELD/blob/master/data/MELD/dev_sent_emo.csv,

[20] https://www.kaggle.com/datasets/msambare/fer2013.