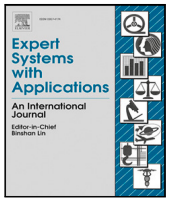
Expert Systems With Applications 240 (2024) 122579

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Deep CNN with late fusion for real time multimodal emotion recognition

Chhavi Dixit [a], Shashank Mouli Satapathy [b,*]

[a] Associate Software Engineer, Shell India Markets Pvt. Ltd., Bengaluru, Karnataka 560103, India
[b] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

ARTICLE INFO

Keywords:
CNN
Cross dataset
Ensemble learning
FastText
Multimodal emotion recognition
Stacking

ABSTRACT

Emotion recognition is a fundamental aspect of human communication and plays a crucial role in various domains. This project aims at developing an efficient model for real-time multimodal emotion recognition in videos of human oration (opinion videos), where the speakers express their opinions about various topics. Four separate datasets contributing 20,000 samples for text, 1,440 for audio, 35,889 for images, and 3,879 videos for multimodal analysis respectively are used. One model is trained for each of the modalities: fastText for text analysis because of its efficiency, robustness to noise, and pre-trained embeddings; customized 1-D CNN for audio analysis using its translation invariance, hierarchical feature extraction, scalability, and generalization; custom 2-D CNN for image analysis because of its ability to capture local features and handle variations in image content. They are tested and combined on the CMU-MOSEI dataset using both bagging and stacking to find the most effective architecture. They are then used for real-time analysis of speeches. Each of the models is trained on 80% of the datasets, the remaining 20% is used for testing individual and combined accuracies in CMU-MOSEI. The emotions finally predicted by the architecture correspond to the six classes in the CMU-MOSEI dataset. This cross-dataset training and testing of the models makes them robust and efficient for general use, removes reliance on a specific domain or dataset, and adds more data points for model training. The proposed architecture was able to achieve an accuracy of 85.85% and an F1-score of 83 on the CMU-MOSEI dataset.

1. Introduction

Human emotions can be understood best when multiple aspects of a person's actions can be analyzed, the content of their speech, their pitch and tone of speaking, and their body language, among others. Hence, recommender systems (Patro, Mishra, Panda, & Hota, 2022), review analyzers, virtual assistants (Singh, Srivastava, Rana, & Kumar, 2021), behavioral analysis, and many such systems, which rely on emotions expressed by humans, can benefit greatly from an advanced multimodal emotion recognition architecture (Agarwal, Yadav, & Vishwakarma, 2019). This project uses text, audio, and video modalities for its purpose. The use of visual and audio cues along with text, where the emotion recognition studies have been quite advanced (Agarwal et al., 2019; Cunningham, Ridley, Weinel, & Picking, 2021), as compared to other modalities, can make the sentiment prediction and analysis more efficient, overcoming the ambiguities that might be faced while using only one modality (Wang, Wan, & Wan, 2020). The opinions described while speaking may also reflect partial or mixed sentiments, for example, a combination of happiness and surprise is present in the statement "Oh! I did not expect to perform so well in these exams". (Chaturvedi, Satapathy, Cavallari, & Cambria, 2019). Systems where multiple emotions can be predicted based on dominance, can be useful for such use cases.

Multimodal analysis models must extract the most appropriate features from the modalities present (Chaturvedi et al., 2019). Extraction and representations are the fundamental steps that can greatly affect the outcome of a model. Another challenge that is most commonly encountered while developing models for multimodal emotion recognition is the method of fusion of the different modalities (Cunningham et al., 2021). Some go for early fusion, leading to a homogeneous input for the prediction models, some do late fusion (Das & Singh, 2022) owing to the differences in the modality representations, or some follow other variations of modality fusions (Gandhi, Adhvaryu, Poria, Cambria, & Hussain, 2022).

The wide range of applications of an efficient multimodal multi-class emotion recognition model makes it an important field to focus research on Yu et al. (2022). This motivated the study for an efficient real-time multimodal analysis model, which is not limited to only binary or ternary, but multi-class classifications. The advancements in individual domains of text, audio, and image emotion recognition

* Corresponding author.
E-mail addresses: chhavi1212@gmail.com (C. Dixit), shashankamouli@gmail.com (S.M. Satapathy).

https://doi.org/10.1016/j.eswa.2023.122579
Received 25 January 2023; Received in revised form 1 November 2023; Accepted 11 November 2023
Available online 14 November 2023

have motivated the use of different kinds of feature extraction, pre-processing, and deep learning techniques combined using ensemble learning for accurate emotion prediction.

This project aims to study the field of multimodal emotion recognition and apply the methodology for accurate and fast predictions using independent emotion recognition in domains of text, audio, and image. The objectives of this study are as follows:

- To explore the field of multimodal emotion recognition with accurate and fast architecture.
- To use cross-dataset training and testing for developing a generalized accurate and robust emotion recognition model.
- To focus on flexible multi-class, multi-label classification and reduce class bias using multiple models.

Since more research has been done in the three fields individually, the datasets available in them are more elaborate and the techniques for emotion recognition also have been much developed, especially in the textual emotion recognition segment (Agarwal et al., 2019). Hence, this has been used to the advantage of this project by training individual models as if they had to predict emotions only in the individual text, audio, and image datasets they are used in. Three benchmark datasets have been used here, for training. It has been noted that each of the datasets has multi-class emotion labels. The datasets used have labels representing the emotions expressed by humans on a usual basis. They include happiness, sadness, anger, surprise, etc. The number of data points in each of the three datasets differs due to the mode of data they represent. Each of the datasets contains raw information, the features are extracted from each according to the needs of the relevant prediction model. Based on the properties of the three modalities used, the prediction models also differ from each other to suit their modalities. Text prediction uses fastText, audio prediction uses 1D CNN and image prediction is done with a 2D CNN model. CNN models have been used instead of high-performing transformers (Ghorbanali, Sohrabi, & Yaghmaee, 2022), to have a light-weight architecture with optimal efficiency. More details about these models and their impact on emotion recognition from the respective datasets have been elaborated further in the paper.

A usual multimodal dataset may be one of the opinion videos. An opinion video can be fragmented and each fragment can be segmented into the three modalities as stated and the models trained on those individual datasets can be used for emotion recognition on the video. This has been followed here. The final prediction for each fragment of the video is done after comparing the performance of bagging with equal weightage and stacking using machine learning (Wen & Hughes, 2020). The overall sentiment expressed comes from the majority percentage of expressed emotions during the whole video.

For this complete project, deep learning models were trained and used due to their advantage of high accuracy. Also, since all the datasets used have a high number of datapoints, they result in good deep-learning models. In cases of emotion recognition as well, they have proven to show good efficiency. Overall, this project aims to explore the field of multimodal emotion recognition and go deeper into the methods used and introduce a novel method for analysis in the field of real-time emotion recognition of human oration.

The structure of the manuscript is as follows:

- Section 2, Background, is exploring the work done in this domain and the research in emotion recognition in the individual modalities.
- Section 3 is Materials and Methods which explains in detail the datasets that were used and the complete flow of the project from data extraction and pre-processing to testing and GUI implementation.
- Section 4, Results and Discussion is showcasing the results obtained with the implemented architecture and analyzing the same.
- Section 5 is Real-Time Applications which is discussing the practical applications of the proposed architecture in different domains of society.
- Section 6, Threats to Validity mentions the factors that have been taken into assumption during the implementation of the said architecture, which might threaten our results when they are not followed.
- The last section, 7 is Conclusion and Future Scope which is summarizing the complete project and mentioning the improvements and experimentation that can be performed in the future based on this project.

2. Background

The field of human emotion recognition as a whole has attracted a lot of research over the past few years, one of the reasons being the wide variety of applications that this domain has. This section elaborates on the background of the topic and the literature survey done while preparing for this project. As mentioned earlier, the field of multimodal emotion recognition has applications in a large number of fields. Many challenges are still present in this field which are slowly being tackled to make the models more efficient and ready for practical applications. This section is divided into multiple subsections. Section 2.1 elaborates on the research done for textual emotion recognition. It is followed by Sections 2.2 and 2.3 for literature survey on image and audio analysis respectively. Section 2.4 explains the research, studied in the field of multimodal emotion recognition. This is followed by Section 2.5, which discusses the limitations observed in the literature survey. The last Section 2.6 mentions the contributions that this study intends to make in the field of multimodal emotion recognition.

2.1. Text emotion recognition

Sailunaz and Alhajj (2019) tried to predict the sentiments in a Twitter dataset by introducing different types of scores while training the models and were able to achieve a high accuracy for the benchmark ISEAR dataset using both Naive Bayes (classification) and KNN (clustering) for the final prediction. In 2020, Liu (2020) explored the combination of using a CBOW-based model along with CNN for predicting the sentiments in the IMDB and COAE2014 datasets and was able to achieve a high accuracy of 87.2% and 90.5% respectively. Meng, Long, Yu, Zhao, and Liu (2019) proposed a transfer learning technique based on a multi-layer CNN model by capturing context from the embedding of text and got improved accuracies on the Amazon reviews and Chinese product reviews dataset. Lee, Kim, and Song (2021) proposed a novel way for sentiment analysis with the hamper of lesser data to train and were able to achieve accuracies comparable with state-of-art machine learning algorithms. They used a semi-supervised approach for the same by creating a base dictionary using a lasso-based ensemble model. Xu, Meng, Qiu, Yu, and Wu (2019) used sentiment information to improve the lightweight TF-IDF word embedding model, and train the BiLSTM model for sentiment prediction. They were able to achieve an F1-score of 92.18%. In 2019, Ahuja, Chug, Kohli, Gupta, and Ahuja (2019) explained the impact of feature extraction on sentiment analysis models including different types of machine learning as well as deep learning models and different embedding schemes for text.

2.2. Audio emotion recognition

For emotion recognition in the audio dataset, a novel approach of using computer vision for the bag of visual words on the spectrogram was used by Pikramenos et al. (2020). The accuracy reaches 68% and is independent of language. Li, Dimitriadis, and Stolcke (2019), used a combination of lexical and acoustic features using late fusion techniques to analyze audio data in their study. The result gave a high

accuracy for ternary classification. Another approach for audio analysis was proposed by Huang and Bao (2019), in which multiple conventional, deep learning and contextual approaches were combined to get an average accuracy of 70.3%. Cunningham et al. (2021) focused on the use of non-musical audio for emotion prediction using two machine learning algorithms, regression, and artificial intelligence models. They were able to achieve good accuracy using both models on the IADS dataset. They suggest the exploration and use of more combinations of audio features for optimal results. Panda, Malheiro, and Paiva (2020) conducted a survey for the use of audio features and methods for emotion recognition from music. They gave a comprehensive analysis of the correlation between audio features and the emotions they represent. They used both computational methods as well as music psychology for the same.

2.3. Image emotion recognition

Asaithambi, Venkatraman, and Venkatraman (2021) worked on sentiment analysis on an image dataset and proposed an end-to-end architecture for processing big data and using machine learning for sentiment prediction. They were able to achieve an impressive value of 0.952 for the AUC metric. Babajee, Suddul, Armoogum, and Foogooa (2020) used a combined model of CNN and SVM to predict emotions from human expressions. They were able to successfully predict 7 emotion classes with an accuracy of 79.8% on the KDEF dataset. Pathak, Bhalsing, Desai, Gandhi, and Patwardhan (2020) have also explored the use of deep learning models for facial emotion recognition and got an accuracy of 70% on the validation dataset using the CNN model. Yadav, Kumar, Ranga, and Rawat (2020) proposed another deep learning algorithm based on CNN for emotion recognition and got an accuracy of 57% on the FER-2013 dataset. Yadav and Vishwakarma (2020) proposed a novel way of using residual attention-based deep learning networks (RA-DLNet) with convolutional neural networks for image emotion recognition. They were able to achieve an accuracy of 70% for the same on binary classification. Semi-supervised learning is often used in image classification, Feng et al. (2022) proposed dynamic mutual training between multiple models for correction of the overall model. Though this architecture is yet to see its effectiveness in the image emotion recognition domain, it seems to be a promising technique to overcome the lack of labeled data and generalize the model to extend to unseen data input as well. In 2020, Zheng, Li, and Wang (2020), in their paper, proposed the use of local region features against the global sentiment for effective visual sentiment analysis. However, they stated that the use of local region features for global sentiment analysis can lead to lower confidence scores. This can directly impact the accuracy and efficiency of the model proposed.

2.4. Multimodal emotion recognition

A Multi-Attention RNN network was proposed by Kim and Lee (2020) for emotion recognition on multimodal data. The proposed architecture was able to achieve 84.31% accuracy on the CMU MOSI dataset. Sun, Sarma, Sethares, and Bucy (2019) have also proposed a novel approach for multimodal sentiment analysis using multimodal embeddings from text, audio, and video modes together to improve downstream sentiment classification, resulting in an accuracy of up to 93%. Another contribution relating to multimodal embeddings was explored by Yang, Xu, and Gao (2020) which used the interaction between text and audio to fine-tune the pre-trained BERT embedding algorithm for a Cross-Modal BERT. The authors saw significant improvement in performance on the CMU-MOSI and CMU-MOSEI datasets. Chaturvedi et al. (2019) proposed the use of ECG signals along with text and projected the features onto a four-dimensional emotion space, to get a good and fast approximation of the emotion expressed. They observed an improvement of about 10%–20% in performance while predicting 24 different emotions in the test samples. Yang, Shao, Wu, and Lin (2022) argued that not just modality fusion, but the focus is required on the quality of features extracted from individual modalities as well. For this they proposed the translation of audio and visual features as well to convert to that of BERT, they got an accuracy of 52.9% on the CMU-MOSEI dataset for 7 classes, which was comparable to or better than most of the recent research. The conclusion of their studies can be combined with those of other studies to create an efficient architecture. Shad Akhtar, Singh Chauhan, Ghosal, Poria, Ekbal, and Bhattacharyya (2019) also proposed a multi-tasking multi-attention framework for simultaneous sentiment and emotion analysis. They used modality-specific transformers to individually capture representations of each modality. Then cross-modal fusion module is introduced to combine the individual representations. They observed an improvement in their framework as compared to single-task architectures. Kumar and Vepa (2020) tried to address three major problems faced in multimodal sentiment analysis, out of those they found that learning and cross-modal interactions have the most benefits and they were able to get an accuracy of 81.1% on CMU-MOSEI. This was an absolute increase of 1.34% over state-of-art techniques tested on the dataset. They used Bi-GRUs for individual modalities and then learned cross-modal interactions between them inspired by methods used by other researchers. The issues raised by them are still explored in many recent works and still have scope for improvement. Focus on the improvement of unimodal representations was illustrated by Guo, Kong, Zhou, Wang, and Wang (2021) where they continuously improved the weightage given and representations of individual modalities, later projecting them on a common latent space. First, unimodal LSTM networks were employed to encode the information from each modality separately. Then their cross-modal interactions were captured using attention mechanisms. Both the results were then refined further for final predictions. Another context-aware methodology was explored by Mittal, Bera, and Manocha (2021) which included semantic information and socio-dynamic interactions during two-way communications among others. A quantum-inspired approach towards multimodal sentiment analysis proposed by Li, Gkoumas, Lioma, and Melucci (2021) in 2021, using principled methods for modeling complicated interactions and correlations. Huddar, Sannakki, and Rajpurohit (2021) proposed a pair-wise attention mechanism using RNNs to better understand the relationship between different modalities for improved representation. They got an accuracy of 81.29% for their work. They suggested more focus on feature selection for improved results. To tackle the problem of lack of large-scale, well-labeled datasets for multimodal emotion recognition, Zhang et al. (2021) proposed using semi-supervised learning along with cross-modal knowledge transfer for efficient sentiment analysis. This allowed them to work with a large set of unlabeled data, making their model more generalized as well.

2.5. Research gaps

Certain research gaps were identified in these papers. Many researchers focused more on the binary or ternary classification of sentiments, over emotion identification. This has been noted by Li, Zhang, Wang, and Gao (2021) in their research as well. These have multiple applications but a more detailed classification can be more beneficial in certain other aspects. As a field, text emotion recognition is way more advanced as compared to other modalities. Even in terms of the availability of datasets, text emotion recognition has an abundance of the same, unlike other modalities or multimodal emotion datasets. There is still a lack of large-scale and diverse multimodal emotion datasets. Semi-supervised learning is another interesting path to take for emotion recognition with a lack of fully annotated datasets. However, since the datasets that we have used are annotated well, this is not required. It has been shown by Dashtipour, Gogate, Cambria, and Hussain (2021) and other researchers that multiple modalities tend to give a more efficient performance than unimodal architectures, yet, in multimodal emotion

recognition, only some have included more than two modalities. A major reason for the same is another research gap, since all the modalities have different features and are represented in different manners, their optimal fusion, be it late or early, is still an ongoing research. The segmentation of complete text to make meaningful sentences is another topic of research and hampers text emotion recognition, because of the lack of context for any given input when it is a part of a bigger piece of text. Learning from the research done till now and with the hope of filling a few of the research gaps, this study proposes an efficient and fast model for real-time analysis of oration using the three modes of text, audio, and image on various benchmark datasets.

2.6. Contributions

The contributions that this study intends to make in the field of multimodal emotion recognition can be summarized as follows:

- The use of multiple benchmark datasets for individual modalities as well as complete stacking and testing, which added more data points for model training; good performance of those models can be interpreted as a generalized efficient and robust model as they have performed well in multiple benchmark datasets at various stages.
- Consideration of each of the modes of a multimodal oration input as independent and predicting individually with separate models. Use of stacking using a light machine learning model on the output of individual models which is advantageous for real-time emotion prediction.
- Multi-class, multi-emotion, real-time analysis of opinion videos for practical usage.
- Use of custom CNN models for image and audio emotion recognition to develop a light model from scratch with easy interpretability and visualization and robustness to noise and variations. Spatial hierarchy representations of CNNs are good for handling the multiple modalities used here as well.

3. Materials and methods

Section 3.1 gives an overview about the datasets used. It explains their sources, features, and usage in this study. It is followed by Section 3.2 which discusses the pre-processing done on the data extracted from these datasets for further training and testing of the respective models. The last Section 3.3 gives a detailed explanation of the deep learning methodologies and algorithms adopted for the complete architecture.

3.1. Datasets

Multiple datasets have been used over the course of this project for various steps. Three of the datasets are benchmark datasets to replicate a uniform scale for accuracy for each of the parts. In total four datasets are used, one each for text, audio, and image emotion recognition models, and one for stacking and multimodal testing.

For textual emotion recognition, "Emotions dataset for nlp" has been used with Kaggle as the source. It has separate training, validation, and testing datasets, with 16,000, 2,000, and 2,000 entries respectively. There are six emotion labels for the dataset, sadness, anger, surprise, joy, fear, and love. The dataset is unbalanced, with the highest number of observations for happiness and the lowest for surprise in all three subsets. Table 1 summarizes the class-wise distribution of the dataset. The numbers in each of the cells represent the number of string inputs available in the dataset against each of the emotions for the training, validation, and testing subsets respectively.

For audio model training, "RAVDESS emotion speech audio" (Livingstone & Russo, 2018) dataset has been used from Kaggle. It is a small dataset with 1,440 audio files, yet, it has a variety of audio files to represent each of the eight emotions, neutral, calm, happy, sad, angry, fearful, disgust, and surprise. It is a well-balanced dataset with an equal number of files for each emotion, as can be seen in Table 2. It takes into account gender, the channel of use, etc. There are 24 professional actors, 12 male, 12 female, and each of the actors has contributed to 60 audio entries in the North American Accent in the English language. For visual emotion recognition, the dataset used is "FER-2013" (Goodfellow et al., 2013), with Kaggle as the source. It has a total of 35,889 data points with a total of 7 emotion labels, namely surprise, sad, neutral, happiness, anger, disgust, and fear. The details of the same can be seen in Table 3. Each cell in the table represents the number of images present in the unprocessed dataset for each of the above-stated emotions in the training and testing subsets. It has separate folders for each of the training and testing files and more sub-folders for each of the emotions represented. It is unbalanced with the least amount of data for the emotion of disgust. The images are gray-scale with dimensions of 48*48. Each of the images consists of different human expressions, with mostly the face visible at different angles (Sarangi, Nayak, Panda, & Majhi, 2022).

Once the models have been trained with appropriate hyper-parameters and deep learning layers to suit the individual datasets, they are combined to predict the sentiments in a multimodal dataset, "CMU MOSEI dataset" (Zadeh & Pu, 2018). This dataset has 3228 opinion videos from 1,000 distinct speakers on 250 distinct topics, amounting to approximately 65 h. This benchmark dataset is the largest available multimodal dataset and represents sentiments, in terms of ternary classification, numerical classification, and label-based emotion classification. It has 6 emotions, anger, surprise, happiness, sadness, disgust, and fear. Since each video has multiple sentiments labeled, hence, the number of sentiment data points exceeds the number of videos by a large number. This can further be observed in Table 4, which shows the rounded-off number of data points for each of the emotions. The numbers under each of the emotions in the table do not add up to the total as they convey the number of data points for those emotions in the 3,879 videos present in the unprocessed dataset.

Multiple datasets have been used for cross-dataset training and testing to make a generalized architecture, not biased towards any particular dataset. It has also been noted that each of the datasets represents emotions commonly expressed by humans. The emotions vary slightly, though, in one or two class labels, which are adjusted in the final multimodal prediction to get the six most common classes of happiness, sadness, anger, surprise, fear, and disgust. These are also the six labels that the videos in CMU-MOSEI dataset have been classified into.

As can be observed from Tables 1, 2, 3 and 4, each of the datasets has large variations in the number of observations and the classes of the classification as well. The imbalance observed in all datasets except RAVDESS reflects the frequency of occurrence of those emotions in the real world, hence making emotions like happiness more probable than those like disgust. It might be argued that the imbalance hampers the training of the respective deep learning models (Obaid & Nassif, 2022), but appropriate steps have been taken to ensure that does not happen. These steps encompass the usage of the fastText model for text classification which uses the Huffman Tree. The tree is formed based on the occurrence frequency of the classes, more frequent ones being closer to the root. Taking into consideration the imbalance, the image model uses an image data generator which applies data augmentation techniques. It increases the data points creating variations and reducing over-fitting of input data. Training with the generator creates a model robust to class imbalance. Data augmentation advantages have been reviewed in Chlap et al. (2021), mentioning its use in handling imbalanced data. These steps and the complete methodology have been explained in detail in Section 3.3. Also, we have used metrics like precision, recall and F1-score. These metrics are essential and helpful in getting a comprehensive analysis of the performance.

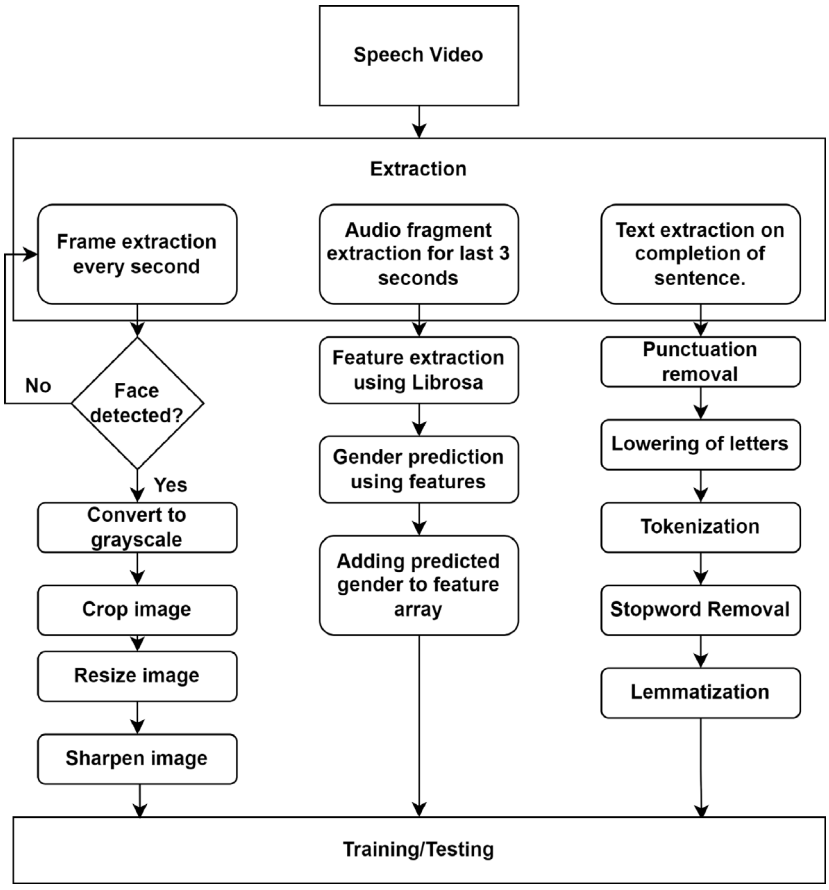


Fig. 1. Pre-processing diagram.

Table 1
Class-wise distribution of text dataset.

Subset	Sadness	Anger	Love	Surprise	Fear	Happiness	Total
Training	4666	2159	1304	572	1937	5362	16 000
Validation	550	275	178	81	212	704	2000
Testing	581	275	159	66	224	695	2000
Total	5797	2709	1641	719	2373	6761	20 000

3.2. Data pre-processing

Each of the three modalities of a video is pre-processed separately for appropriate prediction. The pre-processing is done to match the input data of training datasets and then pre-processed further, to extract relevant features from them and efficiently train and test the models.

For text sentiment classification, each sentence of the complete transcript is treated as a separate observation to predict the relevant sentiment. The division in sentences is done to keep the context of the individual words for better understanding and prediction. Hence, for real-time analysis, the model waits for the completion of a sentence to give the final prediction for the same. Each of the sentences is processed further in the following order: removal of punctuation, lowering of capital letters, tokenization, removal of stop words, and lemmatization. The resultant string with only keywords is ready for input in the text emotion recognition model.

The dataset for audio emotion recognition consists of many details about each of the audio files, which contribute to the input for the emotion recognition model. Out of the information, only the gender of the actor is used. Since gender is already given here, other extracted features from this dataset are used for creating a Random Forest Classifier model that can effectively predict the gender of the speakers where it is not present. Multiple studies have shown a positive impact of gender knowledge on audio emotion recognition (Thakare, Chaurasia, Rathod, Joshi, & Gudadhe, 2021), hence specific model for gender prediction has been created for better prediction on unknown videos. Since the audio files in the training subset are 3 s long, the audio from the complete video in testing is also processed in fragments of 3 s. From each of the fragments, multiple features are extracted, using the Librosa library, in terms of arrays and numerical values and stored in multi-dimensional arrays. The features are absolute short-time Fourier transform, an average of 40 Mel-frequency cepstral coefficients, chromogram, mel-scaled spectrogram, average RMS value, average energy contrast between the highest and lowest energy band, and average tonal centroid features. They result in an array of 195 values. These are used for gender prediction. Predicted gender is added as a feature for sentiment prediction.

For image emotion recognition, each of the images in the dataset is pre-processed. The ImageDataGenerator is used to create more image data with variations in the angle of the image, stretching, zooming, etc. keeping the target size as an image of 48*48 size. The data is ready for input in the 2D CNN model for sentiment prediction. For the multimodal dataset, the frame is extracted at intervals of one second and run through the pre-trained Deepface module for Python with 'opencv' as the backend for face detection. If the face is detected, the frame is converted to a Grayscale and cropped to have only the face in the frame. The image is sharpened by performing convolution of the image with a 3×3 sharpening filter, $K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$. The image is then resized to 48*48 size. This resized image is then fed into the model for prediction. The convolution operation is illustrated in Eq. (1):

$$[I * K](x, y) = \sum_m \sum_n I(x - m, y - n) \cdot K(m, n) \tag{1}$$

Table 2
Class-wise distribution of audio dataset.

Gender	Calmness	Happiness	Sadness	Anger	Fear	Surprise	Disgust	Neutral	Total
Male (12)	96	96	96	96	96	96	96	48	720
Female (12)	96	96	96	96	96	96	96	48	720
Total (24)	192	192	192	192	192	192	192	96	1440

Table 3
Class-wise distribution of image dataset.

Subset	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Total
Train	3995	436	4097	7215	4965	4830	3172	28 710
Test	958	111	1024	1774	1233	1247	832	7179
Total	4953	547	5121	8989	6198	6077	4004	35 889

Table 4
Class-wise distribution of multimodal dataset.

Happiness	Sadness	Anger	Disgust	Surprise	Fear	Total videos
12 400	6000	5000	4100	2300	1900	3879

where the output is the result of convolution, I represents image matrix, K represents the sharpening kernel defined above, (x, y) represents the location in the output image, and m and n are the coordinates within the kernel. It involves sliding a convolution kernel over the image, computing the element-wise product at each position, and summing the results to produce the convolved output, which emphasizes image features and patterns.

Fig. 1 summarizes the extraction of data of individual modes and the steps followed for pre-processing before putting them as input in the respective deep learning models.

3.3. Architecture and methodology

The complete architecture can be divided into five different modules. The first three pertain to the individual modalities of a multimodal input, the fourth module corresponds to ensemble learning, performance optimization, and testing. The last module is reserved for real-time testing using a GUI.

3.3.1. Text emotion recognition model

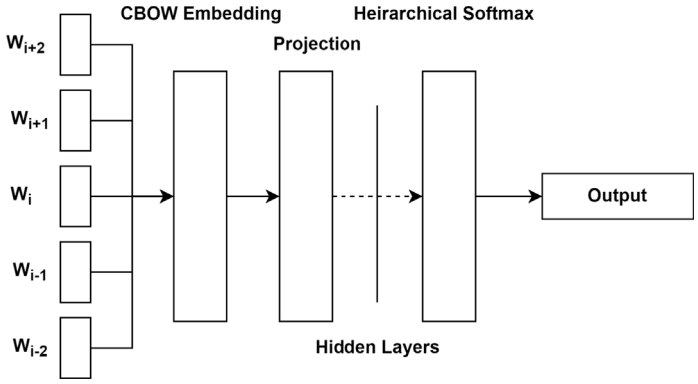
The deep learning model for text emotion recognition was trained based on the "Emotions dataset for nlp" taken from Kaggle. The details of the dataset have been mentioned in Section 3.1. After the pre-processing, as stated in the previous section, a fastText model (Bojanowski, Grave, Joulin, & Mikolov, 2017) was trained on the input from the dataset for sentiment classification. The fastText model was introduced by Facebook Research as a text embedding and classification model which uses a continuous bag of words, a shallow neural network, for category prediction. It uses character n-grams, meaning that words that are present rarely in the training dataset are also represented well enough. This makes it a suitable option for cross-corpus usage. The most advantageous characteristic of this model is that it gives results comparable to other models but has a very high speed. Hence, it is suitable for real-time analysis. It gives the flexibility to change the number of epochs, type of loss function, learning rate, etc. as is seen fit based on the training dataset and target of classification. In this study, the pre-trained fastText model was trained for 8 epochs on the 80% training data with a learning rate of 0.7 and the number of word grams as 4. The loss function used was one versus all, for efficient multi-class classification. This loss function creates 6 different binary classification models for each of the 6 classes and trains each with respect to all others. Despite the availability of other efficient embedding models like variations of BERT, GloVe, etc. Kumar et al. (2021), Vijayvargiya, Kumar, Malapati, Murthy, and Krishna (2022), fastText was also preferred due to its ability for good performance when unknown words are present because of character n-gram division. Since the architecture used multiple datasets for training and testing, this feature proved to be of great advantage. Fig. 2(a) shows the basic layers of a fastText model. The words are taken based on the n-gram specified and each word gets a unique embedding vector based on the order of words around it using the CBOW model (Mikolov, Chen, Corrado, & Dean, 2013). The large embedding vectors are then projected to retain only the important features, 300 by default. After going through more hidden layers, finally, the hierarchical softmax function is used as the last layer. Fig. 2(b) shows the resultant binary tree structure from the hierarchical softmax function. This tree uses the Huffman algorithm resulting in the occurrence of more frequent classes near the root and the rare ones further away. The leaves in the figure represent the classes present in the training dataset for text classification as explained in Section 3.1. The six classes in the figure occur based on the number of observations in them in the training dataset. This structure has an edge over other similar models because it is able to handle imbalanced classes well without external manipulations and is also efficient for multi-class classifications.

For sentiment prediction in the transcript of the multimodal dataset, first, they were pre-processed to only extract the sentences said by the speaker. Next, the prediction confidence scores for each of the sentences were added. The minimum threshold for the sentiments to be listed was kept to 0 to include sentiments with even the lowest possibilities. Based on aggregated confidence scores, the sentiments were ranked and the top 'n', based on the number of sentiments for the video, was concluded as the predicted sentiments for the same.

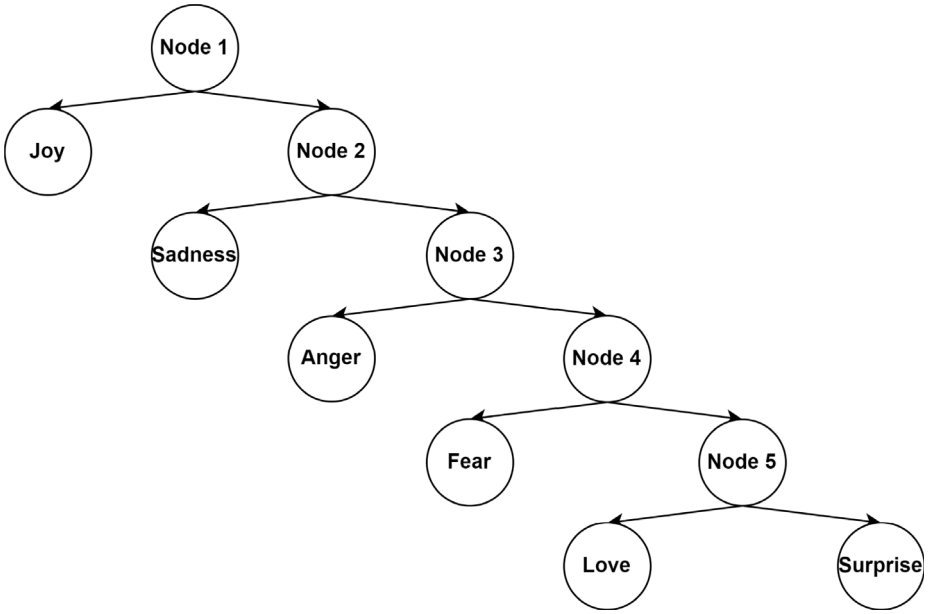
Fig. 3 shows the complete flow for the training of the text analysis model on its training dataset. The pre-processing shown is elaborated well in Fig. 1 as well.

3.3.2. Audio emotion recognition

The second module can be considered as the audio emotion recognition module, which uses the 'RAVDESS emotion speech audio' dataset from Kaggle. The dataset comes with much more information, other than the audio files, that can be extracted from the names of the files. These include the intensity of speech, the statement spoken, the identity of the actor, the vocal channel, etc. However, since such information is not available in other datasets, only the information regarding the gender of the speaker is used. It was noticed during this study that correct identification of the gender of the speaker can increase the accuracy of sentiment prediction by a great margin. The features, as stated in the pre-processing section, are first extracted using the Librosa library and those along with the gender information are used to train a Random Forest model with 100 trees. In the sentiment prediction for the multimodal dataset, this model was used to predict the gender and the extracted features and the predicted gender were then used for sentiment prediction. The model used for sentiment prediction was a one-dimensional CNN model. Convolutional neural networks were first introduced for image analysis (Fukushima & Miyake, 1982). Hence accordingly, they were in two dimensions considering the pixels along both height and width of the image. Here, the CNN model used is one-dimensional, hence, the kernel moves along only one direction. This



(a) Important layers



(b) Binary Tree for Class Prediction

Fig. 2. FastText Model.

variation was introduced keeping in mind the multiple applications of the same in various domains (Das, Dutta, Sharma, & Godboley, 2020). As can be seen in Fig. 4, the model consisted of a total of 11 layers. The input vector was given to a convolution layer with 64 filters. This was followed by the next convolution layer of 128 filters. A max pooling layer with pool size 8 was used for reducing the size of the vector, followed by a dropout layer with 0.4 as a probability. The same block of 1D convolution layer, max pooling, and dropout layers was repeated. This was followed by a flatten layer, a dense layer, another dropout with 0.4 probability, and the final output dense layer with activation function softmax. The loss is used for categorical cross-entropy. The model was trained on the input for 81 epochs. The resultant prediction confidence scores for each of the audio fragments of 3 s were aggregated and averaged to get the final confidence scores. The sentiments are then ranked in decreasing order of confidence scores and picked as predictions based on the number of emotions labeled for the video.

Fig. 5 summarizes the steps followed for training of the audio analysis model on the RAVDESS dataset (Livingstone & Russo, 2018). The detailed features extracted are explained in Fig. 1. The CNN model was developed making sure that it was efficient while being lightweight.

3.3.3. Image emotion recognition

For image emotion recognition, the benchmark 'FER-2013' (Goodfellow et al., 2013) dataset was used from Kaggle. Each of the images in the dataset consists of a face with the relevant expression. The images are in grayscale format. The ImageDataGenerator module from Tensorflow is used to create variations for the given images for a better input dataset. Once the pre-processing is complete, the data is given as input in a two-dimensional CNN model (Fukushima & Miyake, 1982). This deep learning model is most commonly used for image classification, object detection, etc. The model used here is trained with 4 blocks, each consisting of 6 layers including two convolution layers, two batch normalization layers, a dropout layer, and a max pooling layer. This is followed by a flattened layer. The next two blocks, each contain a dense layer, one batch normalization layer, and a dropout layer. The last layer of the model is a dense layer that gives the final output. This is also summarized in Fig. 6. The model is run on 133 epochs and gives the result in terms of a vector of probabilities of occurrence of each of the sentiments. While using the trained model on opinion videos, the pre-processing includes an extra step of face detection and image manipulation to keep only the face in the image. If the image has a face, the sentiment is predicted. The confidence score for each of the sentiments is added and averaged to get the

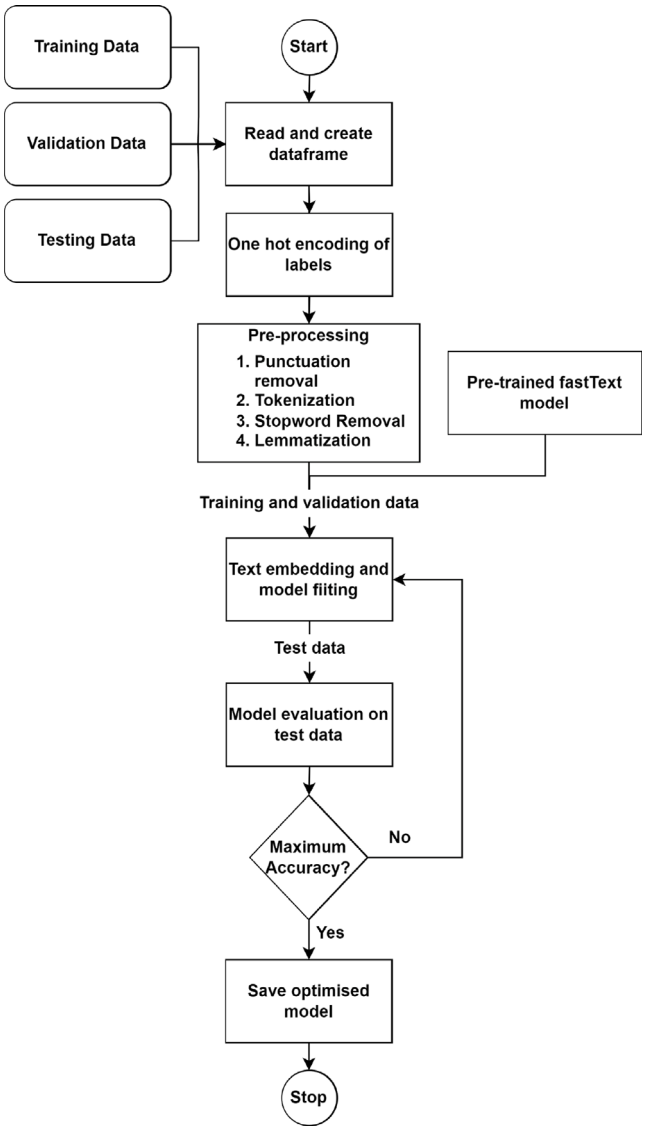


Fig. 3. Flow for text analysis model.

final sentiment confidence score for the complete video. The sorting in descending order of confidence scores gives the top emotions depicted in the video.

The steps followed for the training of the image analysis model is illustrated in Fig. 7. Image sharpening and resizing are not necessary for the training dataset. This is done to maintain uniformity for both the training and the testing datasets. The steps followed for training are shown in the flow diagram.

3.3.4. Model testing, aggregation and GUI

The fourth module comes into play once the training for each of the three emotion recognition models is complete. The models are run on the CMU-MOSEI (Zadeh & Pu, 2018) dataset. The transcript for the videos was provided in the dataset itself. All non-zero emotions were taken as valid emotions for the videos. The individual modality extraction, pre-processing, and predictions were done as mentioned in the previous sections. Each of the models was treated independently. Once the results were obtained, they were put into a Random Forest model as input for late fusion. Techniques for choosing the ensemble model to be used have been explored in Injadat, Moubayed, Nassif, and Shami (2020) for various tasks. In our proposed model, two models, that is, bagging and stacking using the Random Forest model were considered. Owing to the better performance of stacking, it was chosen and refined for the final output. The model was created with input being an 18 elements long vector with values 1 if the sentiment exists according to the model, and 0 otherwise. The output consisted of a one-dimensional vector of length 6. Each of the elements had a value of 1 if the corresponding sentiment was represented in the video, else, 0. The elements were in the sequence: anger, disgust, fear, happiness, sadness, and surprise. For this, the MultiOutputClassifier class of sklearn library was used. Fig. 8 shows an overview of the architecture, with the pre-processing elaborated in Fig. 1. Since steps like importing of libraries are redundant and obvious, they have not been added in the individual flow charts for the modules. The fourth module is the final stage in model training and optimization.

Once the models were trained and the highest accuracy was obtained on the benchmark multimodal dataset, a simple GUI was created using the Tkinter library in Python for real-time analysis of videos, with the least amount of lag, the only lag would be while the text model waits for completion of sentence, audio model waits for 3 s for an audio segment, and the image model for 1 s while waiting to extract the next frame. The same procedure as followed for the CMU-MOSEI dataset is followed here. The individual results of each of the models are continuously given as input in the stacking model for regular updation of the overall sentiments expressed.

4. Results and discussion

All of the models are able to get accuracies comparable with those of state-of-art models. Even on the CMU-MOSEI dataset, the architecture gives good results, despite the models being trained on other datasets. Fig. 9(a), 9(b), and 9(c) show the epoch vs accuracy graphs for training and validation for each of the three models. The orange line in each of the graphs corresponds to testing accuracy and the blue line corresponds to training accuracy. The number of epochs finally used is mentioned in Section 3.3, along with the details of layers used in deep learning models. The resultant accuracy can be found in Table 6.

As can be observed in Fig. 9(a), there is a steep increase in accuracy till 2 epochs, followed by a gradual increase in both, training and testing accuracies. The difference between the training and testing accuracies was reduced for other loss functions, but, to facilitate multi-class classifications, an appropriate loss function was chosen, without sticking to the individual dataset parameters and generalizing the results. The smoothness in the graph can be attributed to the high sophistication of the fastText model, whereas other models have been self-developed. Fig. 9(b) shows a gradual increase in the accuracy of the model from epoch 30. After that, the training accuracy starts reaching stability while training accuracy keeps increasing, signifying possible over-fitting if epochs continue to be trained. Fig. 9(c) shows the data for the image model. It differs from usual training and testing accuracy graphs in the sense that it continues to have higher testing accuracy from approximately 15 epochs onward. This trend was observed for all testing subsets on the FER-2013 dataset. It should also be noted that though the testing accuracy peaks and troughs throughout, the mean of the oscillations were gradually increasing along with the smooth increment of the training accuracy, indicating improvement of the model.

Each of the datasets for individual models was split in an 80:20 ratio for training and testing. Hence, all the models were trained only on 80% of the datasets. Once optimized, they were individually tested on the CMU-MOSEI dataset, using the complete multimodal dataset for testing. The accuracies obtained are shown in Table 6. The fine-tuned fastText model was able to give a higher accuracy on the training dataset with other parameters. However, the loss function had to be chosen to suit the requirements of both training and testing datasets, hence, the trade-off between the lower accuracy and generalization of the model, led to an accuracy of 89.52%. The image model was chosen

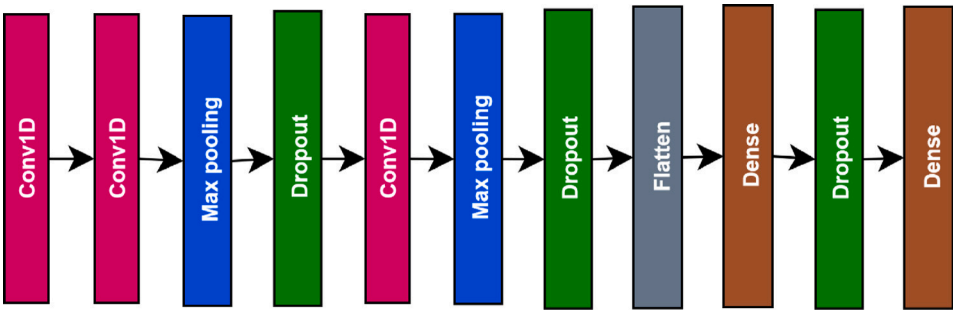


Fig. 4. 1D CNN model for audio emotion recognition.

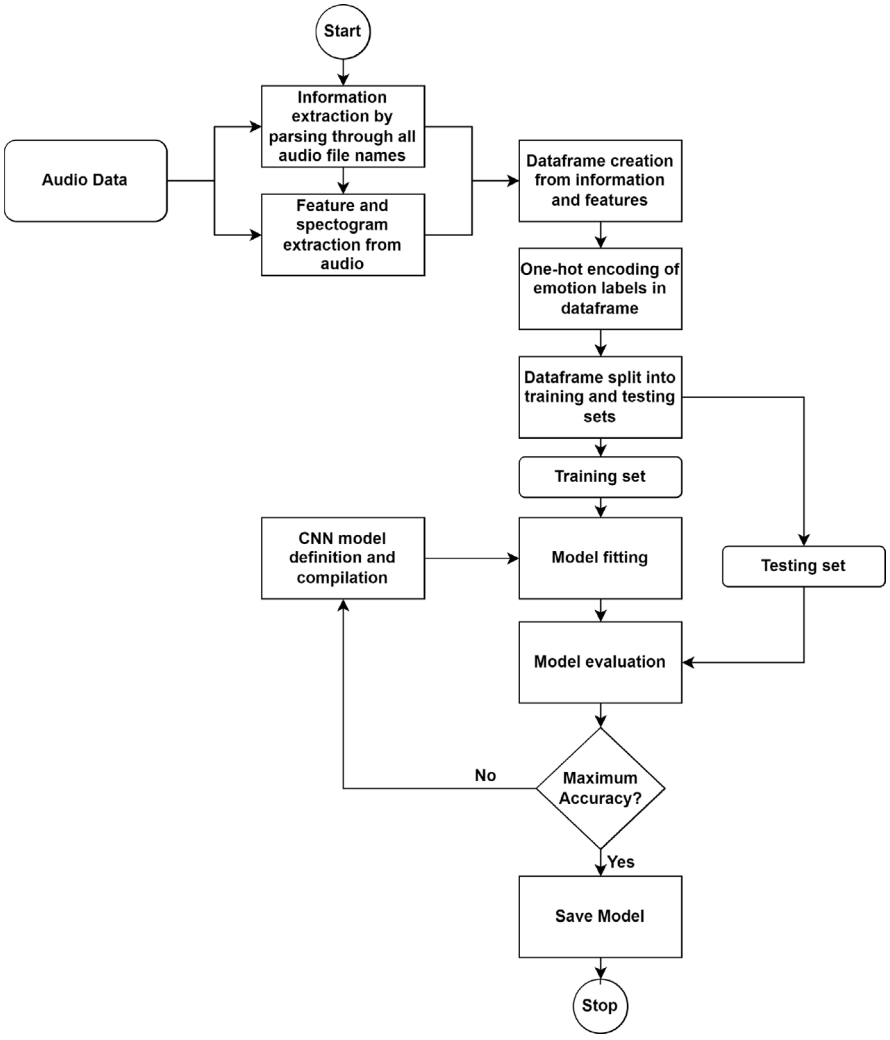


Fig. 5. Flow for audio analysis model.

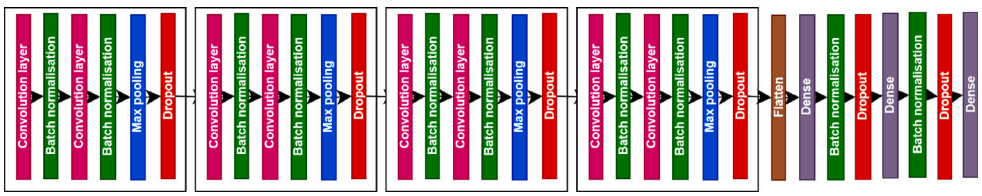


Fig. 6. 2D CNN model for image emotion recognition.

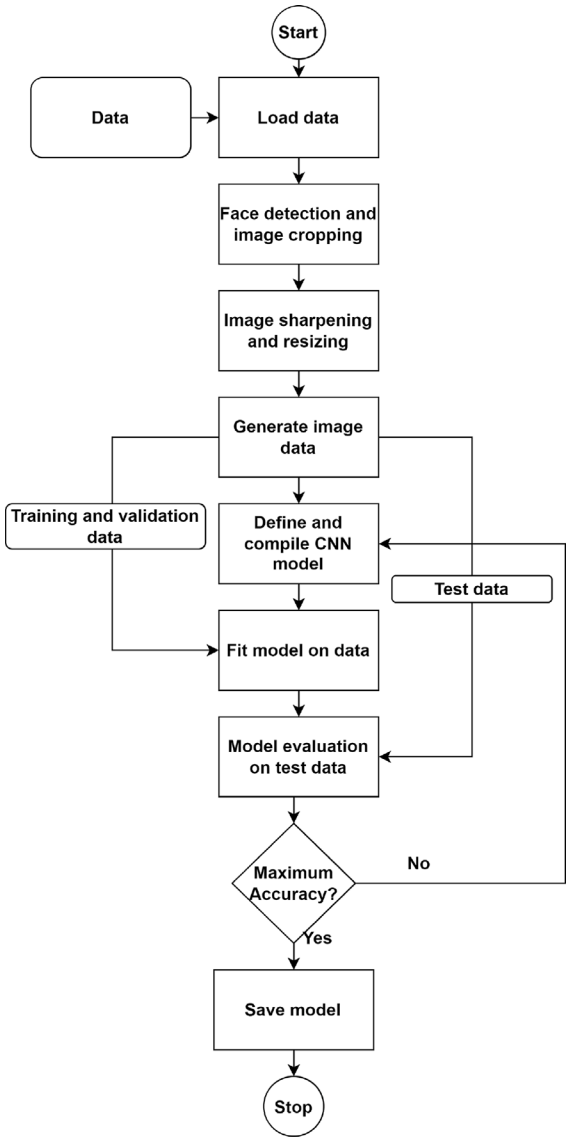


Fig. 7. Flow for image analysis model.

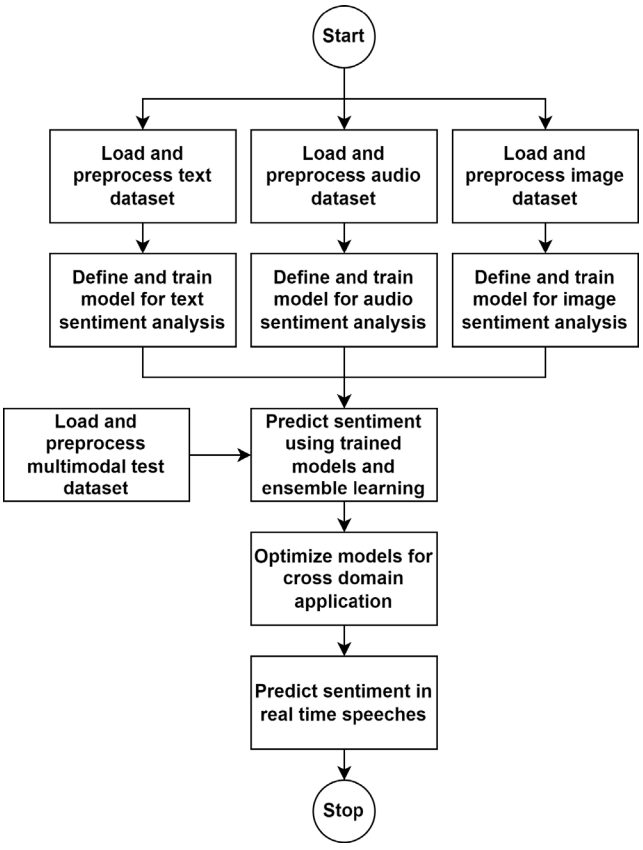


Fig. 8. Overview of the architecture.

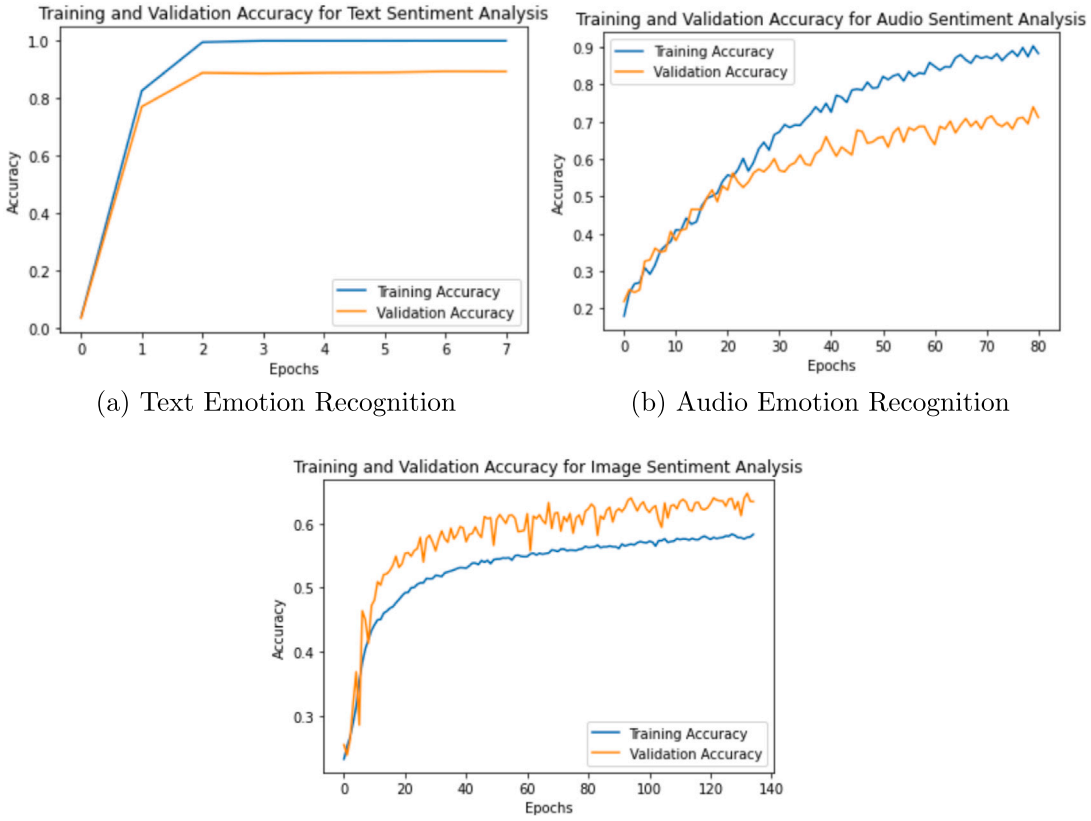
for efficient performance, both in terms of accuracy and time taken, since the analysis has to be real-time. The reason for the significant difference between the accuracies of the audio model on the training dataset and the testing dataset is the difference in the volume range of the training and testing dataset which had to be explicitly manipulated to match. A small Random Forest model was also used along with the audio model for accurate analysis, the model was able to predict the gender based on audio features in the RAVDESS dataset with an accuracy of 97%.

Ablation studies were carried out informally where we experimented with various loss functions and the number of n-grams for fastText. A high learning rate required a lower number of epochs and vice versa. The best results were obtained by reducing the learning rate to 0.7 and increasing the epochs to 8 from the default. The highest accuracy could be achieved for 8 epochs, after which the model started over-fitting. The test accuracy fell from 89.52% for 8 to 87.3% for 9 epochs. For audio emotion recognition, the model was trained on various combinations of MFC and other spectral features, and algorithms for noise reduction. Features like zero crossing rate were also considered, however, they seemed to be extra features, which did not have much impact on the performance of the model, hence, they were removed. Finally, multiple layer combinations of a 1D CNN were experimented on to arrive at the suggested model. For image recognition, we first chose among the various kernels that can be used for face detection, then the use of image generators. Since the dataset was large, the model was first trained without the image generator, however, the class imbalance was creating a bias in the model, so we experimented with various options for image generators. Finally, multiple models were trained on different layer combinations of the 2D CNN model and the number of block repetitions was decided on. Training on multiple iterations gave the optimal number of epochs for a good performance, which was 133 on both the FER-2013 and the CMU-MOSEI datasets. After that, the test accuracy started falling while the training accuracy kept on increasing. For combining the results of models for individual modalities, two different ensembles were tested and the results were compared. The first technique was bagging, which considered the equal weight of all three models, and used the average confidence score from each of the models as the basis for ranking the sentiments and concluding the most probable ones. The second technique used was stacking, which took the final results of the individual models as input and put them into a Random Forest model, giving a vector with length 6 as output, 0 in place of sentiments that were not expressed, and 1 for the ones that were present in the video. Both techniques were first trained multiple times on various parameters to get their optimum models. Then they were checked on two metrics, F1-score, and accuracy. Accuracy was calculated as shown in Eq. (2)

$$Accuracy = \frac{\sum_{i=0}^{n}\sum_{j=0}^{6} f(i,j)}{n \times 6} \times 100 \tag{2}$$

where,

$$f(i,j) = \begin{cases} 1, & \text{if } y_{pred_{i,j}} = y_{test_{i,j}} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$



(a) Text Emotion Recognition (b) Audio Emotion Recognition

(c) Image Emotion Recognition

Fig. 9. Epoch vs accuracy for each model.

Here, y_{pred} and y_{test} represent binary matrices of size n × 6 each. Each element in y_{pred} has the value 0 if the sentiment i is not predicted for observation j and 1 otherwise. Similarly, 1 in y_{test} represents presence of emotion i in observation j. n represents the number of observations in the test subset and 6 stands for the classes present. Eq. (4) shows the formula used for the calculation of another metric, the F1-score to check the efficiency of the architecture proposed.

$$F = \frac{2 \times \sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j)}{2 \times \sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j) + \sum_{i=0}^{n} \sum_{j=0}^{6} x(i,j) + \sum_{i=0}^{n} \sum_{j=0}^{6} y(i,j)} \tag{4}$$

where,

$$w(i,j) = \begin{cases} 1, & \text{if } y_{pred_{i,j}} = y_{testi,j} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$x(i,j) = \begin{cases} 1, & \text{if } y_{pred_{i,j}} = 0 \text{ and } y_{testi,j} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

and

$$y(i,j) = \begin{cases} 1, & \text{if } y_{pred_{i,j}} = 1 \text{ and } y_{testi,j} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

As stated earlier, y_{pred} and y_{test} represent binary matrices of size $n \times 6$ with 1 marking the presence of an emotion and 0 showing its absence in prediction and labeling respectively. $w(i,j), x(i,j)$ and $y(i,j)$ are also used to calculate the precision and recall for each of the classes. This is shown in Eqs. (8) and (9).

$$Precision = \frac{\sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j)}{\sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j) + \sum_{i=0}^{n} \sum_{j=0}^{6} y(i,j)} \tag{8}$$

$$Recall = \frac{\sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j)}{\sum_{i=0}^{n} \sum_{j=0}^{6} w(i,j) + \sum_{i=0}^{n} \sum_{j=0}^{6} x(i,j)} \tag{9}$$

The accuracy for the bagging technique for testing the subset of the videos in the dataset was 80%, the F1-score for the same being 76, whereas, the stacking technique was able to give an average test accuracy of 85.85% and an F1-score of 83 for the same testing subset of the videos. The training accuracy for the stacking was 88.47%. Hence, stacking was chosen as the technique for combining models in the final architecture. Boosting was not considered here as it required sequential execution of the models whereas the input data for all three was heterogeneous. Table 7 shows the class-specific performance of the complete architecture, calculated using Eqs. (8) and (9). In the table, the precision for 'disgust' can be seen as very low, which gives it a low F1-score. This can be attributed to the low number of observations for the same in both the training and testing datasets. The values are obtained by using 80% of the CMU-MOSEI dataset for training the stacking algorithm and testing on the rest 20% of the dataset.

Fig. 10 shows a simple GUI made using the Tkinter library of Python for implementation of the proposed model for real-time usage on an opinion video. This GUI can download YouTube videos and perform analysis on them, it can also perform the analysis on downloaded videos with the condition that the respective transcript is available.

The snapshot in Fig. 10 of the GUI shows its performance on an opinion video taken from the CMU-MOSEI dataset. The video has labels: happiness, sadness, surprise, and fear. The audio emotion model gets updated every 3 s based on the length of the audio fragments, the image analysis graph is updated every second if the face is detected in the frame, and the text analysis model updates based on the length of the sentence being spoken. If the transcript is well punctuated, the sentences are clearly detected, though, if the transcripts do not have sentence segregation, the model calculates the speed of the talking and uses that with an average of 19.2 words per sentence to detect the sentence as shown in Eq. (10) and (11).

$$Words_{PerSecond} = \frac{Duration_{Speech}}{Words_{Total}} \tag{10}$$

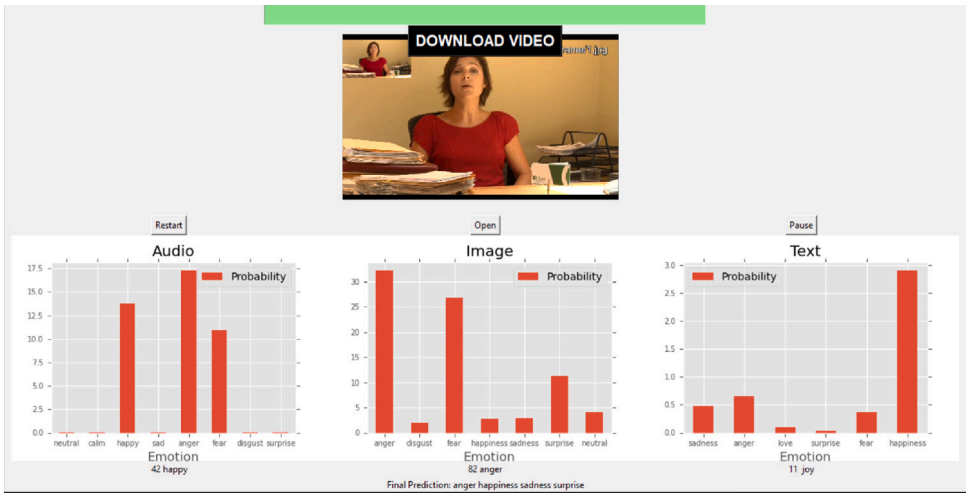


Fig. 10. GUI for real-time opinion video analysis using the proposed model.

Fig. 11. Model overcoming modality bias in GUI.

$$Duration_{Sentence} = Words_{PerSecond} \times 19.2 \tag{11}$$

Here, 19.2 is chosen as average by deriving from table 2 of Zadeh and Pu (2018). Giving the statistics of the CMU-MOSEI dataset, we obtain total number of sentences and total number of words, which are divided to derive the average number of words per sentence used in the equation. Each of the bar graphs in Fig. 10 is updated in real-time and shows the aggregated confidence scores for the respective emotions from the start of the video till the current instance. The number and emotions written under them show the segment number and the emotion depicted in the latest segment respectively. As can be observed, since the fragments of image analysis are the smallest with a duration of one second, the segment number is highest there. The bottom line of "Final Prediction" makes the use of aggregate confidence scores of each of the modes. The total confidence scores till then are then used to determine the top three most dominant sentiments reflected. The top three from each of the modes are put as input into the stacking algorithm to determine the most dominant emotions from the start of the video till the current point. Hence, at the end of the video, the final prediction shows the topmost dominant sentiment of the entire talk. It has the provision to show multiple emotions considering the fact that opinions may not necessarily be reflecting only a single sentiment and emotion perception is subjective.

As can be seen in Fig. 11, the individual prediction models for the modalities do not always work correctly. The audio emotion recognition module is incorrectly predicting disgust as the dominant emotion. However, the stacking model overcomes the bias using the emotions predicted by other modality recognition modules and correctly predicts the overall emotion, as shown in the last line of the GUI.

We have also checked the performance of all the contributing modules in terms of their inference speed. It can be seen in Table 5. The inference speed taken for each module records the time taken right from modality extraction till the final prediction of the modality from the model. It is calculated as shown in Eq. (12).

$$InferenceSpeed = \frac{Time_{Total}}{Samples_{Total}} \tag{12}$$

The total elapsed time was recorded and divided by the number of samples that were predicted to give the inference speed of each module per sample. Out of all the individual modules, audio analysis takes the highest time. This can be attributed to multiple factors like the high number of features that have to be extracted individually during pre-processing and the use of another smaller model for gender prediction. Though the total time seems to be high, as can be seen in the GUI above as well, each of the three modules of text, audio, and image emotion prediction would run in parallel threads. Hence, the maximum time for prediction would be that of audio emotion prediction which is

Table 5
Inference speed of contributing modules.

Module	Inference speed (in sec.)
Text Analysis	0.5209
Audio Analysis	0.9183
Image Analysis	0.7204
Stacking	0.5088

Table 6
Accuracies of individual models.

Model	Accuracy on individual dataset	Accuracy on CMU-MOSEI
Text Model	89.52%	70.41%
Audio Model	73.61%	68.81%
Image Model	66.29%	66.24%

Table 7
Class specific performance of proposed architecture.

Sentiment	Precision	Recall	F1-Score
Anger	0.77	0.78	0.78
Disgust	0.17	1.00	0.29
Fear	0.53	0.70	0.60
Happiness	1.00	0.91	0.95
Sadness	0.89	0.82	0.85
Surprise	0.68	0.79	0.74

0.9 s. Since the text model waits for the completion of a sentence, audio waits for 3 s to take the audio fragment, and the image model takes 1 screenshot each second, the time taken by them for modality extraction, pre-processing, and prediction is negligible. The stacking model waits for a change in any of the three modality predictions and is able to give the final prediction within half a second.

The performance metrics of the proposed architecture with those of the state-of-art models are illustrated in Table 8. The considered metrics are accuracy and F1-score since these were the most commonly available measurements in other researches, which helped to put his project into perspective with respect to their performance. The baseline model introduced by Zadeh and Pu (2018) was the first to introduce the CMU-MOSEI dataset. The authors have used a hierarchical method for studying the edge weights of graphs in MFN, which is a recurrent neural modal, for modality fusion. Their main focus is to study the interaction between the modalities, whereas the focus of this study is to explore an accurate as well as fast architecture for late fusion. Many of the approaches done till now have mostly focused on the interaction between modalities and attention mechanisms to get the best out of the multiple modalities considered. Implementation and key conclusions from the papers mentioned in Table 8 have also been explained in Section 2.4. It can be observed from the table that the proposed model is performing better than the recent state-of-art models, by more than 2%. It must also be noted that all except the proposed model have been trained and tested on the same dataset, that is, only the proposed model is cross-corpus, using multiple benchmark datasets.

5. Real time applications

As for multimodal emotion recognition, the proposed architecture has scope for usage in various domains (Das & Singh, 2023). The real-time analysis gives it an edge over other models for broader usage in practical scenarios. The following section elucidates a few of the many applications of the proposed architecture. In the medical field, it can be used to analyze the behavior of the patient while examining them for mental illnesses. A real-time analysis helps in constant observation of the patient and can help the doctors in determining the correct treatment for the same (Haider, Pollak, Albert, & Luz, 2021).

Similar real-time observations are also helpful in the analysis of behavioral patterns and actions of criminals in order to determine the possibility of repetition of their actions and the circumstances under which the crime was initially committed (Liu & Liu, 2021). Multimodal emotion recognition is also useful in the analysis of political and other similar mass speeches and observing the public reactions to the same. When used prior to elections, emotion recognition can help to assess the emotion most dominated by the speaker, how much it influences the public, and how other speeches must be written and delivered (Zhao et al., 2019). This emotion recognition can also help alert authorities in case inciting emotions are observed to be dominant in such addresses. It can also help curb the spread of misinformation through social media and other means (Ogundokun, Arowolo, Misra, & Oladipo, 2022). Another application of the same can be to take more comprehensive reviews from customers by businesses and analyze those for product and service improvement (Yu et al., 2022). Providing customers with a platform for multimodal expressions would give them the freedom to express themselves in a more natural manner.

The model has been trained and tested specifically on emotion recognition datasets, which has potential for application in multiple domains, as elaborated above as well. Though this is unique to emotion recognition, similar architecture using cross-dataset training and testing can be used for other wide-domain prediction models as well, where the individual predictions are interrelated (Zhou, Liu, Qiao, Xiang, & Loy, 2022).

6. Threats to validity

- The image emotion recognition model is created with the assumption that the speaker's face is visible at all times during the video. The face detector is included in the architecture to discard frames where faces are not detected, but it has not been tested particularly for scenarios of the presence of multiple faces.
- The audio emotion recognition model has been developed for the scenario of a single speaker with a North American accent. Two-way conversations and videos with other noises like music have not been included in the test dataset.
- The textual emotion recognition model is unilingual, trained, and tested only for the English language. Also, the transcript needs to be available with the video for the text model to work.

7. Conclusion and future scope

This study proposes a cross-corpus architecture using various benchmark models for an efficient real-time analysis of opinion videos. It uses the cues of text, audio, and video modalities, hence developing a robust model. The individual models are trained and optimized on independent benchmark text, audio, and image datasets. The optimized models are then tested on a fourth benchmark dataset for individual results and determining the ensemble technique to be used for combining the results. Taking into account the biases posed by the individual models, and their learnings, the stacking technique is determined to be the most suitable. It uses a Random Forest model for stacking the individual inputs. Many smaller models have also been used within these for efficient performance.

The use of multiple models overcomes the possible bias in individual models towards particular emotions due to the number of observations, or other aspects. For example, the image model gets confused between fear and sadness since, generally, the expressions of both classes tend to be similar. The architecture was able to achieve an accuracy of 85.85% and an overall F1-score of 83 on the CMU-MOSEI dataset. This proved to be better than some, and comparable to other recent state-of-art models in this domain for the same dataset. This architecture posed a uniqueness of using multiple benchmark datasets for training and testing respectively. All the models used, and the complete architecture, have been developed while keeping in mind the performance speed of the model and hence, it does not pose any lagging. This makes it ideal for real-time analysis of opinion videos where multiple threads are used for processing each of the modes simultaneously.

Table 8
Comparison of proposed architecture with other state-of-art recent models.

Model	Techniques used	Accuracy	F1-Score
Zadeh and Pu (2018)	MFN		76.3
Shad Akhtar et al. (2019)	Transformers	62.8%	78.6
Kumar and Vepa (2020)	Bi-GRU	81.1%	78.53
Huddar et al. (2021)	RNN	81.29%	73.12
Guo et al. (2021)	LSTM	83.25%	
Proposed Model	**FastText and CNN with late fusion**	**85.85%**	**83**

While the results are promising, work can still be done to make the architecture more efficient. The individual accuracy of the models are not very high on the multimodal dataset. Methodologies can be explored to improve their performance. A possibility might be to use pre-trained models and transfer learning, like VGG-19 for image analysis, etc. This can also increase the overall accuracy of the architecture. Disgust as an emotion has fewer occurrences in the datasets used, though that also reflects the rarity of occurrence of the emotion as compared to others. This lack of observations results in the inability of the model to train well in this class. More diverse datasets can be included while training to overcome this imbalance. This architecture can also be expanded to multiple ethnicities and languages across the World while also making the models dynamic to continue learning from multiple datasets, hence, making them more robust.

CRediT authorship contribution statement

Chhavi Dixit: Data curation, Writing – original draft, Editing, Methodology, Software, Visualization. **Shashank Mouli Satapathy:** Conceptualization, Investigation, Supervision, Validation, Writing – reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Agarwal, A., Yadav, A., & Vishwakarma, D. K. (2019). Multimodal sentiment analysis via RNN variants. In *2019 IEEE international conference on big data, cloud computing, data science & engineering* (pp. 19–23). IEEE.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341–348.

Asaithambi, S. P. R., Venkatraman, S., & Venkatraman, R. (2021). Proposed big data architecture for facial recognition using machine learning. *AIMS Electronics and Electrical Engineering*, *5*(1), 68–92.

Babajee, P., Suddul, G., Armoogum, S., & Foogooa, R. (2020). Identifying human emotions from facial expressions with deep learning. In *2020 zooming innovation in consumer technologies conference* (pp. 36–39). IEEE.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Chaturvedi, I., Satapathy, R., Cavallari, S., & Cambria, E. (2019). Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, *125*, 264–270.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563.

Cunningham, S., Ridley, H., Weinel, J., & Picking, R. (2021). Supervised machine learning for audio emotion recognition. *Personal and Ubiquitous Computing*, *25*(4), 637–650.

Das, S., Dutta, A., Sharma, S., & Godboley, S. (2020). A comparative analysis of a novel anomaly detection algorithm with neural networks. In *Deep learning and neural networks: concepts, methodologies, tools, and applications* (pp. 52–68). IGI Global.

Das, R., & Singh, T. D. (2022). A multi-stage multimodal framework for sentiment analysis of assamese in low resource setting. *Expert Systems with Applications*, *204*, Article 117575.

Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Computing Surveys*, *55*(13s), 38.

Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2021). A novel context-aware multimodal framework for persian sentiment analysis. *Neurocomputing*, *457*, 377–388.

Feng, Z., Zhou, Q., Gu, Q., Tan, X., Cheng, G., Lu, X., et al. (2022). Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, *130*, Article 108777.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Springer.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2022). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*.

Ghorbanali, A., Sohrabi, M. K., & Yaghmaee, F. (2022). Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Information Processing & Management*, *59*(3), Article 102929.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117–124). Springer.

Guo, X., Kong, A., Zhou, H., Wang, X., & Wang, M. (2021). Unimodal and crossmodal refinement network for multimodal sequence fusion. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9143–9153).

Haider, F., Pollak, S., Albert, P., & Luz, S. (2021). Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech and Language*, *65*, Article 101119.

Huang, A., & Bao, P. (2019). Human vocal sentiment analysis. arXiv preprint arXiv:1905.08632.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN. *International Journal of Interactive Multimedia & Artificial Intelligence*, *6*(6).

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, *200*, Article 105992.

Kim, T., & Lee, B. (2020). Multi-attention multimodal sentiment analysis. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 436–441).

Kumar, L., Kumar, M., Murthy, L. B., Misra, S., Kocher, V., & Padmanabhuni, S. (2021). An empirical study on application of word embedding techniques for prediction of software defect severity level. In *2021 16th conference on computer science and intelligence systems* (pp. 477–484). IEEE.

Kumar, A., & Vepa, J. (2020). Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 4477–4481). IEEE.

Lee, G. T., Kim, C. O., & Song, M. (2021). Semisupervised sentiment analysis method for online text reviews. *Journal of Information Science*, *47*(3), 387–403.

Li, B., Dimitriadis, D., & Stolcke, A. (2019). Acoustic and lexical sentiment analysis for customer service calls. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5876–5880). IEEE.

Li, Q., Gkoumas, D., Lioma, C., & Melucci, M. (2021). Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, *65*, 58–71.

Li, Y., Zhang, K., Wang, J., & Gao, X. (2021). A cognitive brain model for multimodal sentiment analysis based on attention neural networks. *Neurocomputing*, *430*, 159–173.

Liu, B. (2020). Text sentiment analysis based on CBOW model and deep learning in big data environment. *Journal of Ambient Intelligence and Humanized Computing*, *11*(2), 451–458.

Liu, Q., & Liu, H. (2021). Criminal psychological emotion recognition based on deep learning and EEG signals. *Neural Computing and Applications*, *33*, 433–447.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One*, *13*(5), Article e0196391.

Meng, J., Long, Y., Yu, Y., Zhao, D., & Liu, S. (2019). Cross-domain text sentiment analysis based on CNN_FT method. *Information*, *10*(5), 162.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mittal, T., Bera, A., & Manocha, D. (2021). Multimodal and context-aware emotion perception model with multiplicative fusion. *IEEE MultiMedia*, *28*(2), 67–75.

Obaid, W., & Nassif, A. B. (2022). The effects of resampling on classifying imbalanced datasets. In *2022 advances in science and engineering technology international conferences* (pp. 1–6). IEEE.

Ogundokun, R. O., Arowolo, M. O., Misra, S., & Oladipo, I. D. (2022). Early detection of fake news from social media networks using computational intelligence approaches. In *Combating fake news with computational intelligence techniques* (pp. 71–89). Springer.

Panda, R., Malheiro, R. M., & Paiva, R. P. (2020). Audio features for music emotion recognition: A survey. *IEEE Transactions on Affective Computing*.

Pathak, A. R., Bhalsing, S., Desai, S., Gandhi, M., & Patwardhan, P. (2020). Deep learning model for facial emotion recognition. In *Proceedings of ICETIT 2019* (pp. 543–558). Springer.

Patro, S. G. K., Mishra, B. K., Panda, S. K., & Hota, A. (2022). Hybrid action-allied recommender mechanism: An unhackneyed attribute for E-commerce. *ECS Transactions*, *107*(1), 4537.

Pikramenos, G., Smyrnis, G., Vernikos, I., Konidaris, T., Spyrou, E., & Perantonis, S. J. (2020). Sentiment analysis from sound spectrograms via soft BoVW and temporal structure modelling. In *ICPRAM* (pp. 361–369).

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computer Science*, *36*, Article 101003.

Sarangi, P. P., Nayak, D. R., Panda, M., & Majhi, B. (2022). A feature-level fusion based improved multimodal biometric recognition system using ear and profile face. *Journal of Ambient Intelligence and Humanized Computing*, *13*(4), 1867–1898.

Shad Akhtar, M., Singh Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. arXiv e-prints, arXiv–1905.

Singh, P., Srivastava, R., Rana, K., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, *229*, Article 107316.

Sun, Z., Sarma, P. K., Sethares, W., & Bucy, E. P. (2019). Multi-modal sentiment analysis using deep canonical correlation analysis. arXiv preprint arXiv:1907.08696.

Thakare, C., Chaurasia, N. K., Rathod, D., Joshi, G., & Gudadhe, S. (2021). Gender aware CNN for speech emotion recognition. In *Health informatics: a computational perspective in healthcare* (pp. 367–377). Springer.

Vijayvargiya, S., Kumar, L., Malapati, A., Murthy, L. B., & Krishna, A. (2022). COVID-19 article classification using word-embedding and extreme learning machine with various kernels. In *International conference on advanced information networking and applications* (pp. 69–81). Springer.

Wang, Z., Wan, Z., & Wan, X. (2020). Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of the web conference 2020* (pp. 2514–2520).

Wen, L., & Hughes, M. (2020). Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, *12*(10), 1683.

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on bilstm. *Ieee Access*, *7*, 51522–51532.

Yadav, Y., Kumar, V., Ranga, V., & Rawat, R. M. (2020). Analysis of facial sentiments: A deep-learning way. In *2020 international conference on electronics and sustainable communication systems* (pp. 541–545). IEEE.

Yadav, A., & Vishwakarma, D. K. (2020). A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimedia Systems*, *26*(4), 431–451.

Yang, B., Shao, B., Wu, L., & Lin, X. (2022). Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing*, *467*, 130–137.

Yang, K., Xu, H., & Gao, K. (2020). Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 521–528).

Yu, B., Wei, J., Yu, B., Cai, X., Wang, K., Sun, H., et al. (2022). Feature-guided multimodal sentiment analysis towards Industry 4.0. *Computers & Electrical Engineering*, *100*, Article 107961.

Zadeh, A., & Pu, P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Long Papers)*.

Zhang, S., Chen, M., Chen, J., Li, Y.-F., Wu, Y., Li, M., et al. (2021). Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. *Knowledge-Based Systems*, *229*, Article 107340.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., et al. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management*, *56*(6), Article 102097.

Zheng, R., Li, W., & Wang, Y. (2020). Visual sentiment analysis by leveraging local regions and human faces. In *International conference on multimedia modeling* (pp. 303–314). Springer.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.