Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multimodal Natural Language Processing: Integrating Text, Vision, And Speech for Enhanced Artificial Intelligence Understanding

Dr Basant Kumar[1*], Praveen Nainar Balasubramanian[2], Dr. Essam Al-Husseini[3]

Asst Professor, Department of Mathematics and Computer Science, Modern College of Business and Science, Muscat, Sultanate of Oman[1*]

University of North Carolina at Charlotte[2]

Assistant Professor of Business Administration, Al-Esraa University, Iraq[3]

*Corresponding Author Email: dr.basantkumar19@outlook.com

ARTICLE INFO

Received: 10 July 2025

Revised: 24 Aug 2025

Accepted: 05 Sept 2025

ABSTRACT

Multimodal sentiment analysis is a developing research area that aims at using multiple and diverse inputs like text, speech and vision to increase the efficiency of emotion identification. In this research, the MELD dataset is used for the classification of sentiment using an integration of Random Forest (text), SVM (speech), and ANN/CNN (vision). The results reveal that the vision models outcompeted ANN with the models attaining 75% accuracy, the Random Forest attained about 56%, while SVM has a seventeen percent tested speech rate and most often misclassified sentiments as the neutrality. The approaches are also in line with the need to perform multimodal fusion in which the talk, text and vision modes are used in a complementary manner in order to minimise the classification error. The areas to be improved in the future are the transformer text and speech models (BERT, Wav2Vec), the attention-based CNNs for facial analysis, and advanced fusion methods, such as early, late fusion, and hybrid fusion. Multimodal sentiment analysis has real-world uses in areas such as human-computer interaction, monitoring the sentiment in customer behaviour with the help of AI systems, tracking mental health conditions, and moderation of content in social media.

Keywords: Multimodal Sentiment Analysis, Machine Learning, Deep Learning, Emotion Recognition, Speech Processing, Computer Vision, Multimodal Fusion, Human-Computer Interaction

Introduction

Artificial Intelligence (AI) has developed great advances in understanding language in terms of processing and analysing human communication, but it remains a challenge to identify sentiment and emotion in conversation. [1]. While the existing natural language processing (NLP) models are mostly based on textual inputs, human communication in general, and emotions in particular, includes text, speech and non-verbal signals. Multimodal NLP is an advanced form of such use of artificial intelligence techniques that combines these various modalities in enhancing understanding and decision-making. [2]. The strengths of learning from textual, audio and visual input are that it results in a system that has context, has common knowledge, or is more human-like or easily able to understand emotions required in an interactive system [1]. The AI techniques need to incorporate information from multiple modalities because the single-modality models are inadequate for detecting human emotions and intents. [3]. Body language and tone of voice are usually involved in communication that cannot be easily captured and relayed through written or oral words. For instance, where the message says, 'That is so great', the same may be perceived as negative if given with a sarcastic look or intonation. This kind of limitation of unimodal NLP underlines

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

how faster and more accurate results can be obtained if multiple data sources are taken into consideration in order to get a better answer to the specific questions of context and sentiment classification [4].

Emotion recognition is perhaps the largest component of the area of application of multimodal learning. [5]. Sentiment analysis is one of the primary elements of AI systems, which can be used in customers' support chatbots, virtual assistants, and mental health monitoring [6] [7] [8]. Such systems are made better by Multimodal AI as it identifies the psychological states such as frustration, happiness or sadness besides other if verbal stimuli are sometimes not easily discernible. Emotions are very vital in areas of application, including automated therapy and human-robot interaction, so as to be in a position to independently respond appropriately [9].

Another area is human-computer interaction (HCI), where multimodal learning increases the interactive nature and reality of information processing by AI systems [10]. Modern voice assistants such as Siri, Alexa or Google Assistant can recognise the emotions of people increasingly better and more accurately [11]. Using tone of voice and facial muscles, there are ways available for multichannel artificial intelligence to alter the responses based on a user's mood and disposition, including conversations. It also becomes applicable in teaching where an artificial tutor can sense that the learner is frustrated and change strategy [12]. As human-computer interactions become more integrated into everyday life, understanding and responding to human emotions is becoming a critical component of creating more engaging, intuitive, and helpful AI systems. Multimodal sentiment analysis provides a solution by enabling AI systems to interpret emotions more accurately by considering text, speech, and vision.

Despite the improvements in traditional approaches to unimodal sentiment analysis, they are still accompanied by some disadvantageous elements. Implicational meanings, sarcasm and contextual sensitivity are difficult for text-based NLP models; hence, emotions are often misclassified [13]. Although speech-based approaches can capture information flow, tone and, pitch and intensity of the conversation, it has drawbacks such as noise, differences in speakers' characteristics and irregular intonation. Similarly, in vision-based sentiment models, facial expressions are used, which may not be accurate due to low expressiveness, bad lighting or occlusion. [14]. These challenges point out the fact that the use of a single model is not able to capture all the dimensions of sentiment within humans.

Multimodal learning is the process of combining written language, speech, and vision, but it faces the challenge of data fusion, which is the incorporation of features from all three without losing valuable information about each modality. [2]. Different data types require different structures in their feature space and processing, and therefore, management into the same model is expensive. When fused improperly, it can create redundancy of information or is more likely to contain inconsistent or lost data, which will affect the sentiment classification rate [15]. Other issues are equally encountered in feature extraction and selection when dealing with multimodal sentiment classification. Each modality is processed with individual pre-processing techniques. The text is processed using TF-IDF and word embedding, speech is used in the form of MFCCs, while images are converted to convolutional feature maps [16]. This is one of the unsolved problems in data pre-processing and feature extraction area. In addition, data used in real life also pose problems of asynchrony in facial expressions and speech tones as compared to the textual content being spoken [17]. Therefore, poor temporal synchronisation when working with multifaceted models may cause misinterpretation of sentiments and, thus, lower performance results.

The existing sentiment analysis techniques tend to be unimodal (normally text-based model), which does not account for the broader range of human emotion, especially in complex situations where sarcasm, tone or facial expression is involved. Speech-based models find it difficult to handle variations in accents of speakers, background noise and lack of wide range of expression, while vision-based systems could have problems with lighting conditions and low expressiveness. The limitations lead to misclassification and poor accuracy of applications to real-world scenarios. This gap is bridged by this research through the proposition of a multimodal sentiment analysis framework to incorporate text, speech and vision to represent complimentary affective signals. The methodology integrates machine learning (Random Forest, SVM) and deep learning (ANN, CNN) techniques in conjunction with modality-specific ones that enhances over all accuracy and robustness. The study, achieving this goal, applies the MELD

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

dataset and numerous fusion strategies to overcome the modality-specific weaknesses and showcase a scalable strategy to construct more emotionally intelligent AI systems for such purposes as virtual assistants and sentiment-aware customer service platforms.

The main goal of this research is to establish a multimodal sentiment classification model that utilises both text, speech, and vision for the purpose of increasing the rates of detecting emotions. To meet this goal, the following specific objectives are developed. The first specific objective is to build a single multi-modal classifier through ANN, CNN, Random Forests, and SVM. The enhancement of the general accuracy of sentiment predictions is also the primary motivation to explore features derived from deep learning and classification through traditional ML models. Such a hybrid approach makes it possible to combine the strengths of each model so as to enhance several modality sentiment classifications. The second objective involves comparing unimodal and multimodal sentiment classification. In order to compare text-only, speech-only, vision-only, and fusion-based models, a detailed evaluation shall be carried out. Hence, this research seeks to quantify the value given to the methods of integrating different modalities in a bid to understand why the process of integration makes the enhancement of sentiment classification possible. The third objective focuses on the assessment of the methods within the framework of the proposed multimodal approach to the MELD dataset. The MELD dataset, which carries out real-world conversation-style text, speech and visual information getting involved, facilitates developing proficiency in learning all these parallel data. Thus, this research will compare early integration, which fuses the raw signals; late integration, which fuses the prediction results; and feature-level integration, which fuses the derived features, in order to identify the best method for processing multimodal information. In the analysis, performance measures, which include accuracy, precision, recall, F1-score, and ROC-AUC, will be employed in making the comparison.

This research also presents a way to further enhance the application of multimodal NLP and sentiment classification as it examines the ways through which the various modalities can contribute to emotion identification. Sentiment analysis typically focuses on text data alone, but this approach often misses emotional cues found in non-verbal forms of communication, such as speech tone and facial expressions. Multimodal sentiment analysis attempts to bridge this gap by integrating these diverse forms of input, offering a more accurate and robust approach to emotion identification. The first significant contribution includes proposing an advanced supervised model, which utilises ML and the DL mechanism to accomplish the task of sentiment classification. The second contribution is a comparative analysis between unimodal and multimodal models, as well as the investigation of benefits linked to the fusion of multiple modes. The third contribution is a study of the various fusion techniques to identify the most effective fusion for text, speech and vision in sentiment classification. At last, the research presents a realistic evaluation of the MELD dataset for the advancement of the domain of multimodal AI and emotion detection.

Literature Review

1.1. Overview of NLP in Sentiment Analysis

Sentiment analysis is common in natural language processing which helps the machine to understand the Human emotions and opinions from text [18]. It is used in the analysis of brand discussions on social media platforms, customer feedback analysis, political opinion tracking, and conversational AI agents. There are two key classification types for sentiment analysis, exhaustive classification and intensive classification, which are positive and negative, as well as anger, joy, and sadness, respectively. Traditional approaches include the semantic rule-based approach and semantic lexicon-based approach, where there is an expansion to the more advanced ML-DL-based text-based sentiment analysis [19] [20].

1.1.1. Traditional NLP Methods for Sentiment Analysis

One of the most common techniques in text categorisation, the Bag-of-Words (BoW) model, transforms documents into multi-set words training words irrespective of localisation within the document.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Since each document is turned into word occurrences, it undergoes the input of classifiers such as Naïve Bayes or SVM classifiers. Even though BoW is computationally efficient, it fails to understand context very well, and the problem of sparsity reduces its performance in the case of sentiment analysis. [21].

TF-IDF (Term Frequency-Inverse Document Frequency) is more effective as compared to BoW because the word importance score is done based on the frequency of the word in a document with reference to its frequency in a collection of documents [22]. This is used to make sure that more emphasis is given to significant words by eliminating the frequent words. However, TF-IDF is not capable of handling polysemy, which is the words that have two or more different meanings, and it does not account for the position of the word in the context as well as the relation between the words. [23].

In order to overcome the limitations of BoW and TF-IDF representation methods, word embedding approaches like Word2Vec, GloVe and FastText were proposed [24]. These techniques encode words into continuous vectors so as to render modelling the semantics of word usage by recognising the contextual relevancies. Unlike BoW, word embeddings are good at understanding synonyms and associated words, enhancing the performance of the sentiment analysis models.

1.1.2. Advanced Approaches in NLP Sentiment Analysis

Deep learning brought more significant changes in sentiment analysis by introducing models that use long-range dependencies and context. Long Short-Term Memory (LSTM) networks, as an improvement over Recurrent Neural Networks (RNNs), are well-suited to model sequential dependencies and flow of data across contexts in sentiments present in most generic sentences [25]. Nevertheless, LSTMs have drawbacks in terms of the extreme length of the texts and slow time in training the models. Currently, Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-Trained Transformer), have become revolutionary both in the field of NLP, particularly for performing sentiment analysis [26]. All the Transformers operate from self-attention mechanisms, allowing for the processing of entire sentences at once to differentiate meanings. [27]. These models surpass NLP techniques, and for these reasons this tool is considered the current best in sentiment analysis.

Apart from deep learning, there remain other machine learning models like Random Forest, SVM, and Naïve Bayes in sentiment classification. These models work fairly well for learning from specific texts, but they cannot identify contextual correlations between words in comparison to deep models [28]. Nevertheless, they are highly efficient from the computational perspective as they can be utilised for small datasets and used in real-time applications where deep learning methods cannot be applied.

1.2.Speech Processing for Sentiment Analysis

Speech-based sentiment analysis involves biological and linguistic characteristics in the voice, such as melody, tone, loudness, and tempo. While performing speech analysis, the classification takes place after extracting features from the audio data [29]. There are different methods of extracting speech characteristics, and the most common one is the Mel-Frequency Cepstral Coefficients (MFCCs).

The human body can show emotions through the voice, stress and variations in the pitch [30]. For instance, the high pitch and the increased speech rate is considered to indicate a state of excitement or anger, while low pitch and the slowed rate of reporting are associated with sadness. These aspects cannot be readily identified by the traditional NLP techniques employed in models, which is why the identification of emotions from speech is vital for real-time use cases like voice-operated devices, mental health assessment and call centre monitoring [31].

MFCCs are one of the most common feature extraction techniques applied in the speech domain, which characterises the short-term power spectrum of sound. They replicate the human ear rhythms or distribution of sensitivity to various frequencies so as to be highly useful in classifying speech-based emotion [32]. To enhance the conventional features of the audio signals, Spectrograms enable the spectrum of frequency energies to be discovered, and the Mel Frequency Cepstral Coefficients are used to identify the variations of the tone of voice and three states of emotion, namely Happy, Angry and Sadness. [33].

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.2.1. Comparison of Speech-Based Classifiers

Support Vector Machines (SVM) and Decision Tree classifiers, which are the parts of the classical Machine Learning approaches have been employed for detecting the emotions from the speech. Such models are often applied to hand-crafted speech features such as MFCCs, pitch and energy and perform poorly in highly variable speech data [34]. With the evolution of deep learning, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) have also been applied to raw spectrograms of speech for end-to-end learning. CNNs extract spatial patterns from spectrograms, while LSTM looks at sequences of spectrograms in the data [35]. These models are better than conventional ML models, but they demand massive amounts of labelled data and computing ability.

1.3.Vision-Based Emotion Recognition

Facial expressions are significant in the classification of sentiment when used in multimodal analysis since the face contains the features of attitude. Vision-based sentiment analysis is implemented through the use of deep learning variants like CNNs as well as ANNs to learn facial features from images and videos [36].

CNNs are well suited for image-based emotion recognition because they can learn the features in a hierarchical structure from the facial expressions. While most of the prior computer vision approaches needed the design of features, CNNs can identify the slight differences in facial muscle movements, eye movement, and mouth movements [37]. Facial emotion can also be detected using ANN, but in most cases, the output is fed to the ANN after feature extraction. ANNs do not discover spatial patterns, as is the case with CNN. Instead, it provides feature-level classification. [38].

1.3.1. Challenges in Image-Based Sentiment Analysis

Despite the success of vision-based sentiment classifiers, it is clear that there are challenges that exist when it comes to image-based sentiment analysis. One of the potential problems that has a great impact on users is image noise and light change, which may cause distortions in facial features and thus result in the wrong classification [39]. Also, emotional recognition in real-time remains an issue since facial data obtained from videos requires models with low time delay that are suitable for real-time analysis of emotions [40]. Besides, there is a shortage of data diversity, and many of the existing facial emotions datasets lack collection quality, which prevents the models from being adaptive to different populations.

1.4.Hybrid Approaches and Multimodal Fusion

Multimodal sentiment analysis involves the use of text, speech, and vision to improve the situation awareness of AI regarding emotions [2]. Different strategies have been tested in order to join the information from different modalities. There are various methods to merge multiple data inputs, which are used to enhance the context in sentiment classification. These models employ a combination of ML techniques such as the Random Forest, Support Vector Machine and deep learning frameworks, including Convolutional Neural Networks, Long Short-Term Memory and Transformers.

1.4.1. Fusion Techniques

Early Fusion

Early fusion is a strategy that combines features from the different modalities at the input layer before passing them through the classification model. [41]. Despite having a lot of information, this approach is useful, although normally, the phenomenon investigated may need further reductions in dimensions.

Late Fusion

Late fusion processes each modality initially and then integrates the results towards the last stage of data processing [41]. While this approach is computationally efficient, it has a main drawback in that it is not able to handle cross-modal interactions optimally.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Feature-Level Fusion

Feature-level fusion is used to extract equivalent features from the different modalities of data and understand the relation between them [41]. This approach is the most effective, but aligning such inputs can be a complex process if a lot of modalities are involved.

Several studies have employed early fusion (integrating features at the input level) and late fusion (combining results at the decision level) for multimodal sentiment analysis. While early fusion allows models to learn relationships between modalities, late fusion provides simplicity and can avoid the complexities of joint representation learning. Hybrid fusion, combining both approaches, has also been explored.

1.4.2. Benefits of Combining Machine Learning with Deep Learning (RF, SVM + ANN, CNN)

Combining two or more classifiers, including random forest and support vector machine for machine learning with artificial neural networks and convolutional neural networks with deep learning, increases the amount of accuracy and generalisation of sentiment classification [42] [43]. In terms of structural textural characteristics, it augments features from audio revolutionary changes effectively and also works well for facial expression, which makes it very applicable for multimodal AI models. [44].

Method

1.5. Dataset Description: MELD

The Multimodal EmotionLines Dataset (MELD) can be categorised as one of the widely used benchmark datasets intended for sentiment as well as emotion classification. MELD augments the original EmotionLines database by including not just textual data but also modalities of both audio and visual data that was from the Friends TV series [45]. This is done by using a multimodal rather than a text-based dataset, which offers a more holistic approach to identifying the emotions and sentiments of people in conversations, which in turn makes it suitable for testing the effectiveness of multimodal NLP models.

The sentiment and emotion of the dataset are divided into recognised dialogue-utterance formats, which means that the dataset is further sorted by dividing dialogues into turns. A sample contains text, a corresponding speaker label, a corresponding emotion label, a sentiment label and acoustic audio-visual modalities. There are positive, neutral and negative sentiments, as well as anger, disgust, fear, joy, neutral-emotion, sadness, and surprise as the categorisation of emotions.

Descriptive to the dataset, MELD has a total number amounting to 13,708 utterances spoken in 1,433 dialogues. About 21% of the samples are positive, while negative samples rank third, even though there are relatively few of them, 14% only. This is a problem that increases the difficulty of classification models, and to overcome this, data augmentation or recollection methods and re-sampling should be applied. Every message is also linked with the audio that has information about the speaker's voice as well as face landmarks extracted from the video. Hence, MELD can be well used for training and testing multimodal models that simultaneously analyse textual, spoken, and visual information.

1.6. Data Pre-processing Techniques

In order to pre-process text, speech audio, and image, data pre-processing will be employed to facilitate training and model evaluation of different modalities. Before the classification task can be performed, different modality needs to be pre-processed using certain feature extraction techniques.

1.6.1. Text Processing

In text pre-processing, some of the tasks carried out included converting the raw data collected in the form of utterances into a format appropriate for machine learning algorithms. In order to pre-process the text data, they were tokenised, stop words were removed, and the words were lemmatised. Feature

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

extraction will be done using Term Frequency-Inverse Document Frequency (TF-IDF), a method that refers to the frequency of the terms within the documents. As compared to the relative frequency approach, TF-IDF makes low and high values of the correlated terms but emphasises the importance of more unique terms. For the overall sentiment classification, an RF classifier is used as the main classifier. Random forest, a type of ensemble learning analysis, uses a number of decision trees that are created from samples of the data set and integrates their results. The use of RF, as opposed to a single decision tree, is the generalisation of decision and minimisation of the overfitting, which is beneficial when it comes to text classification.

1.6.2. Speech Processing

The speech data was pre-processed using Wav2Vec, which provides a robust feature extraction from raw audio, converting it into a suitable format for classification. To perform discrete speech-based sentiment classification speech features need to be extracted from raw audio files. To characterise speech, raw waveforms contain much redundancy and noise, and therefore, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted. It should be noted that the properties of MFCCs mimic the structure of the human auditory system, as they focus on changes in the frequency of change, which is important for recognising emotions.

Using Librosa, the 13 MFCC features per frame were extracted with delta and delta-delta features to augment the dynamics in the sound. These extracted features were then averaged and varied, which produced a numerical matrix representing the structure of the respective utterance's audio part.

The support vector machine (SVM) with an RBF kernel will be used for classification purposes. SVMs apply well to speech classification because of their capability to work in a high-dimensional space and ability to track intricate patterns of speech. The RBF kernel is used for this purpose because it is capable of working on non-linear structures in the features of the speech, leading to better classification.

1.6.3. Image Processing

In order to perform facial emotion recognition, the frames extracted from MELD clips had to be pre-processed. The steps described above for each utterance come with visual data; therefore, the pre-processing involved the use of OpenCV's Haar Cascades for face detection and the *Dlib* library for facial alignments. This helped in keeping the model abstracted to the regions of the face which were important for a particular gender.

The vision data was pre-processed by detecting faces in the video frames and extracting facial landmarks, which were then passed to the CNN for feature learning. For emotion classification from facial images, two deep learning algorithms, namely Artificial Neural Network (ANN) and Convolutional Neural Network (CNN), were applied. Before inputting the data into these models, the image pre-processing techniques, namely, Grayscale conversion, Histogram equalisation, and Image normalisation, were used to improve contrast and remove noise. Other Augmentation methods used included flipping, rotation, and brightness adjustment in order to enhance the model generalisation.

1.7. Fusion Approach

The current research uses late fusion to combine the outputs from the text, speech, and vision classifiers. In particular, results of classification of Random Forest (text), SVM (speech), and CNN/ANN (vision) models' fusion were implemented by a majority voting technique. This approach was selected to retain modality-specific features integrity with minimizing the noise or the non-informative data impact from any single source. Contrary to early fusion, which may lead to redundancy or incompatibility of features, the late fusion will enable independent contributions of each modality to the final sentiment prediction. This approach enhanced the reliability and comprehensibility of the multimodal system, and particularly in diverse data settings, such as MELD.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.8. Model Architectures

1.8.1. Text Model: TF-IDF + Random Forest

In the case of text-based sentiment classification, the feature vector based on TF-IDF is used to train or classify a Random Forest classifier. Random Forest was selected for text sentiment classification due to its resilience to overfitting and effectiveness in handling high-dimensional, sparse data such as TF-IDF vectors. RF model consisted of 100 decision trees that were built using a randomly selected part of the dataset. The final prediction is made by majority rule of the trees constructed. In an attempt to optimise the performance of Boosting, hyper-parameter tuning is done using Cross validation utilising such factors as tree depth and number of estimators with GridSearchCV. It is important to note that in the final model, accuracy, precision, recall, and F1-score were used for assessing the performance.

1.8.2. Speech Model: MFCC + SVM

In the segmentation and classification of speech, the extracted MFCC features were applied in an SVM classifier with an RBF kernel. Support Vector Machine (SVM) was used for speech classification, leveraging its strength in modelling non-linear decision boundaries based on MFCC features. To achieve a good balance of training samples, both for classification the data is balanced between the two classes. Subsequently, GridSearchCV is employed to fine-tune for a proper combination of C, the regularisation parameter and gamma, which is the RBF kernel parameter, to classify the given data set as accurately as possible. To overcome the issue of overfitting, cross-validation of the models is conducted so that they provide good performance on unknown speech data.

1.8.3. Vision-based Models: ANN and CNN

In this study, ANN and CNN architectures were employed for facial emotion recognition. Convolutional Neural Networks (CNN) were chosen for their superior ability to capture hierarchical spatial features in facial expressions, while Artificial Neural Networks (ANN) provided a baseline for comparison. The ANN model has three layers, and all three layers are fully connected layers. It receives the flattened pixels of the input images, then proceeds to the hidden layers with 128 neurons and another with 64 neurons using ReLU activation for emotion categorisation, which is made into four different categories; a softmax layer will be employed. This is done to avoid overfitting and dropout regularisation of the type L2, with a strength of 0.3 is used. The CNN model involved two Conv2D layers and two max-pooling layers down-sampling the data fed into the convolutional layers. The first convolutional point applied 32 filters of size 3x3, and the second one applied 64 filters of the same size. A dense layer of 128 neurons is incorporated before the Softmax output layer in order to make the network fully connected. This CNN architecture made it possible to extract high and low-level features from the facial images which are related to the different emotions.

1.9.Model Training and Hyper-Parameter Tuning

For all the models, 80% of the given dataset is used for training and 20% for testing. Stratified sampling is followed as a method to partition the datasets into the training and testing sets with equal distribution of sentiment classes. In Random Forest (text-based classification), a grid search is conducted on the number of trees, maximum depth of the trees, minimum number of samples required to split the tree and the number of trees to build the forest. It has been found that an optimal configuration of the features results in a higher classification accuracy. In the case of SVM for speech classification, tuning is centred sharply around two parameters, namely C (regularisation) and gamma (kernel width), to discern the adequate decision boundary for sentiment classification. This improved the generalisation between the data from both sets, thus even decreasing the classification errors for speech emotion recognition.

Training of ANN will be done by mini-batch gradient descent while for CNN (facial emotion recognition), training is optimised using Adam optimiser. Batch normalisation is used in the process of normalising the data to enhance the training process, while early stopping is applied to avoid overfitting. Dropout layers were added to both ANN and CNN models equal to 0.3-0.4 in order to reduce overfitting. The learning rate is predefined and started at 0.001 in order to get the best out of the neural network model.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The performance of each of the models is measured using accuracy, precision, recall, and F1-score. To expand on the evaluation of the classification performance, confusion matrices were obtained for each of the modalities, which would illustrate the areas of confusion. The proposed multimodal fusion technique incorporated textual, speech, and vision features in order to enhance the general sentiment analysis classification. Applying these two features of machine learning, the methodology guarantees accurate sentiment classification models that are also scalable and reliable in responding to real-life complex data inputs. The proposed approach of TF-IDF + RF for the text MFCC + SVM for speech and ANN/CNN for visions shows that multimodal is more efficient in educating AI to recognise emotions compared to using single data-only classification.

Future versions of this study will include transformer-based models such as BERT [text] and Wav2Vec [speech] for contextual understanding and nuance of feelings. These models provide better explainability with the use of attention mechanisms that bring into focus significant input features that make predictions. In addition, XAI methods including SHAP (SHapley Additive exPlanations) and Grad-CAM will be discussed to obtain transparent explanations regarding how each of the modalities plays its role in classifying sentiment, with ethical AI deployment in sensitive uses such as those of monitoring mental health and tracking customer emotion.

Results

Exploratory Data Analysis

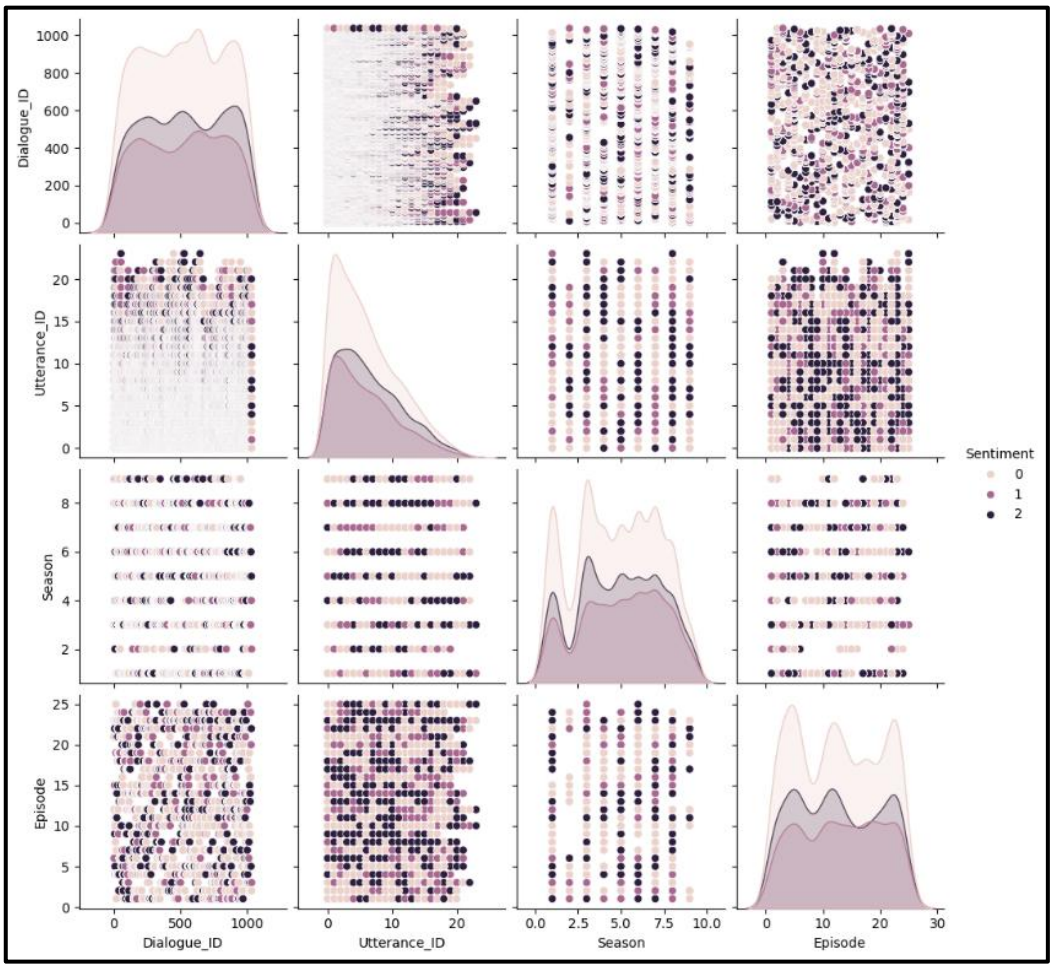

Figure 1. Pair plots – Relationships between numerical features in the MELD dataset

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 1 is a pair plot that shows the correlation of numerical inputs in the MELD dataset by sentiment classes, where $0 = Neutral$, $1 = Positive$, $2 = Negative$. The Kernel density estimates on the diagonal show higher density for the neutral sentiments. Scatter plots indicate that there is a variation in sentiment based on dialogues and episodes, thus the need for multimodal learning.

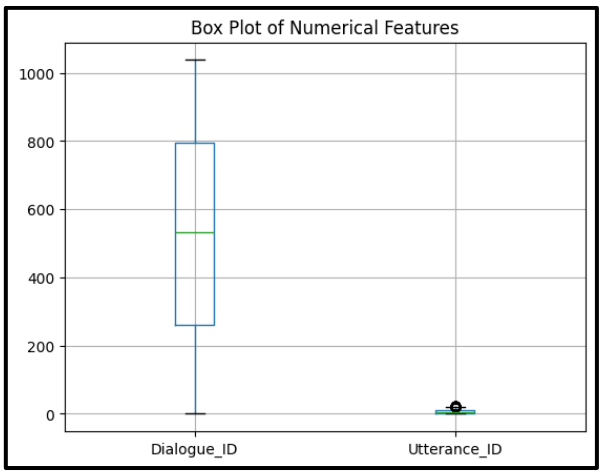

Figure 2. Box Plot – Anomalies detection in numerical features

Figure 2 represents the box plot of the entire *Dialogue_ID* and *Utterance_ID*. *Dialogue_ID* Varies which means there are many different dialogues while *Utterance_ID* is more limited, which signifies that there are fewer utterances. Variations in the *Utterance_ID* values indicate that some of the dialogues consist of way more utterances than others. This supports the proposal of treating dialogues as hierarchical in sentiment classification to suit the different structures of utterances.

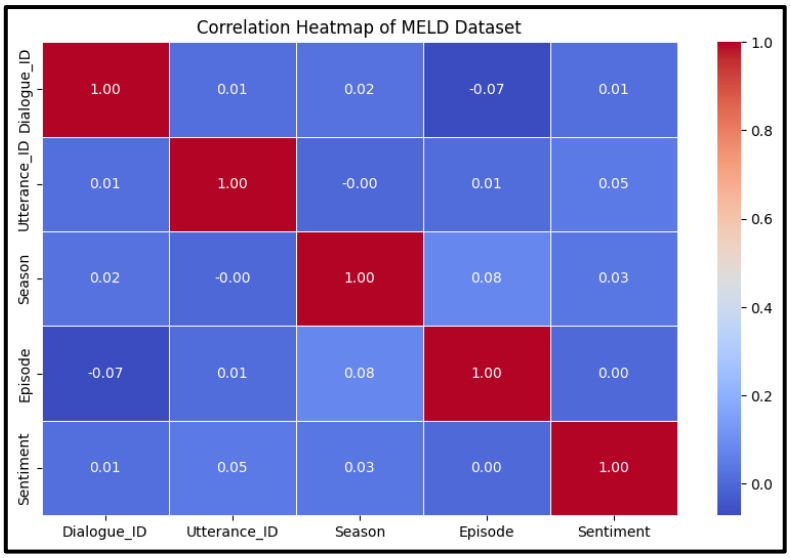

Figure 3. Correlation Heat map – Relationships among MELD's numerical features

Figure 3 presents a correlation heat map to signify the interconnection of numerical attributes of the MELD. *Dialogue_ID* is related to the Episode, which means that this field can be useful for assigning a dialogue to a certain episode. These weak associations of the sentiment polarity with numeric features mean that sentiment classification needs to utilise inputs from multiple modalities since the metadata can only provide limited emotional information based on the given numerical values.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

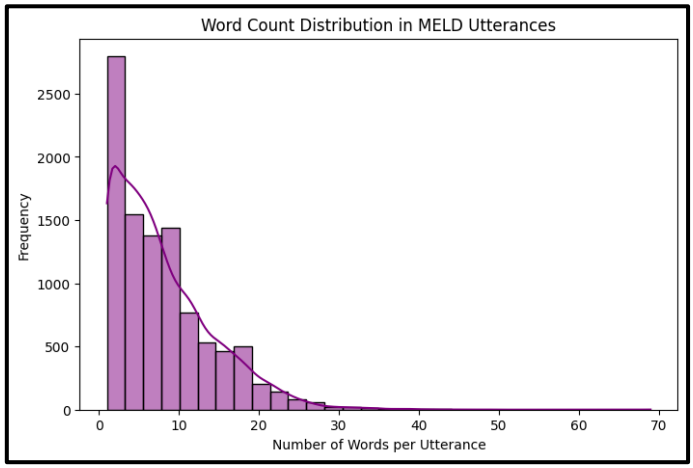

Figure 4. Histogram – Word count Distribution

Figure 4 is the word count distribution of MELD utterances in the form of a histogram. This means most of the utterances are less than ten words or phrases long, with longer ones being the exception rather than the norm. This would mean that textual information may not have sufficient context and that speech and even visuals information is need in sentiment analysis.

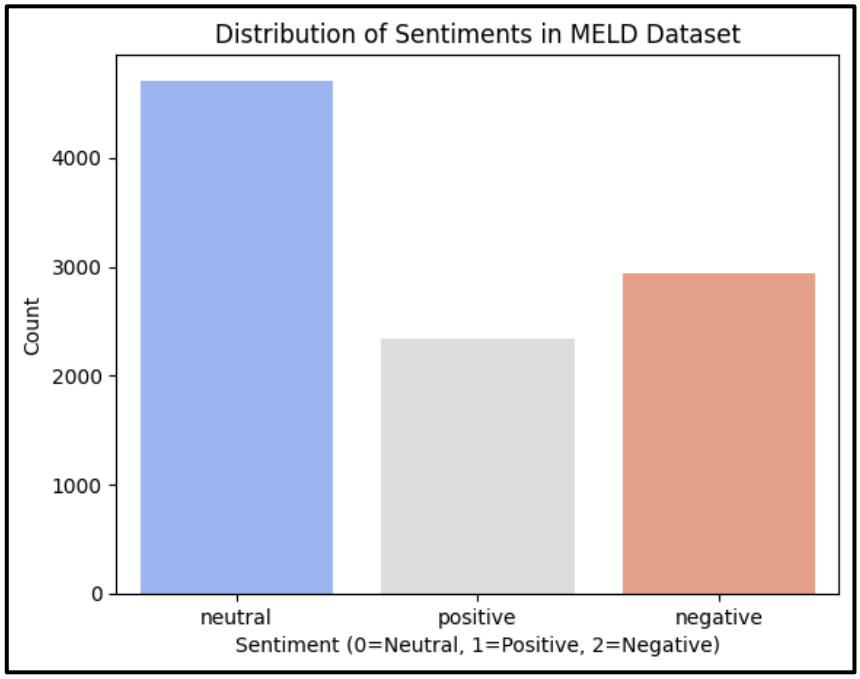

Figure 5. Bar Plot – Distribution of Sentiments

Figure 5 presents a bar plot which describes sentiments in the MELD dataset. The overall sentiments are almost equally divided among the neutral, negative, and positive sentiments. This imbalance may lead to a situation where certain models may be biased, and techniques such as resampling or class weighting may have to be employed. Such a significant prevalence of neutral words indicates the relevance of using non-textual variables for emotion identification.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.10. Model Evaluation

Table 1. Random Forest Classification Report

Random Forest Report (Text):				
	Precision	Recall	F1-score	Support
0	0.58	0.83	0.68	985
1	0.54	0.27	0.36	449
2	0.51	0.33	0.40	564
Accuracy			0.56	1998
Macro-Avg	0.54	0.48	0.48	1998
Weighted-Avg	0.55	0.56	0.53	1998

Table 1 shows the classification report of applying the Random Forest model to textual sentiment classification. Under the evaluation criteria, the proposed model reached an accuracy level of 56% alongside superior recognition performance for neutral sentiments (0.83) while demonstrating lower recognition precision for positive categories (0.27) and negative categories (0.33). The model shows fewer capabilities to identify non-neutral sentiments than neutral ones when measuring model effectiveness.

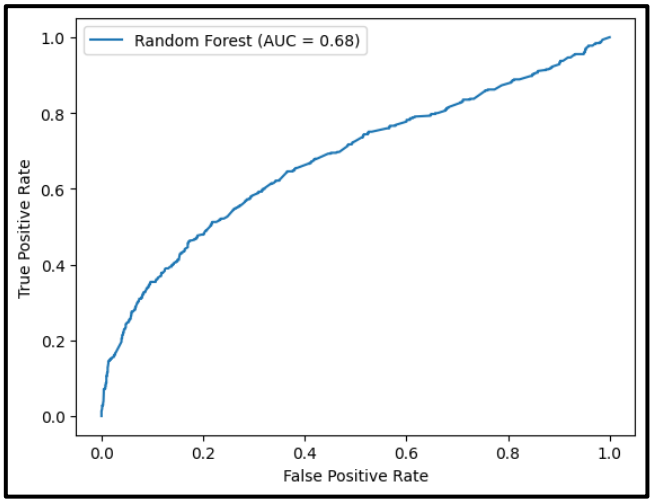

Figure 6. ROC Curve – Random Forest

The Random Forest model acquired a 0.68 AUC score, as shown by its ROC graph in Figure 6. Although the discrimination capabilities are considered moderate, the current model performance requires additional improvements. The curve demonstrates moderate performance in class distinction beyond random chances but highlights that refinement work should be done on feature extraction and model tuning.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

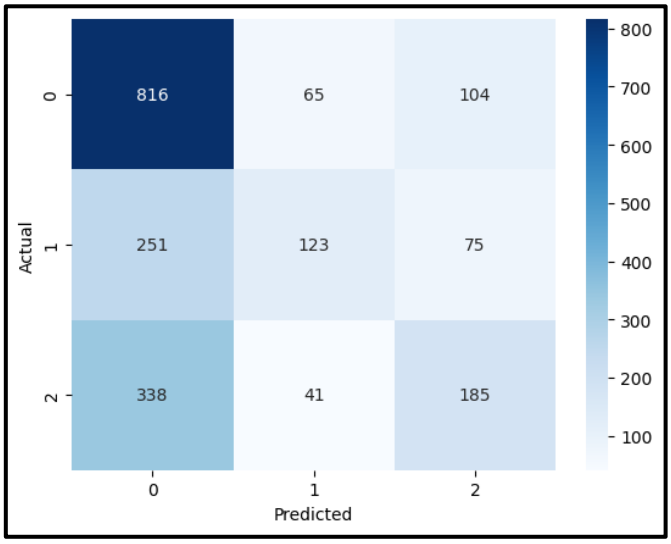

Figure 7. Confusion Matrix – Random Forest

Figure 7 shows a confusion matrix containing sentiment classification model results. The system demonstrates strong performance in identifying neutral statements (816 correct) and shows inadequate abilities when detecting both positive (123 correct) and negative (185 correct) sentiments. Class 2 misclassification stands out as a concern since better feature extraction approaches or class balancing strategies are needed.

Table 2. SVM Classification Report (Speech)

Support Vector Machine Report (Speech):				
	Precision	Recall	F1-score	Support
0	0.49	1.00	0.66	985
1	0.00	0.00	0.00	449
2	0.00	0.00	0.00	564
Accuracy			0.49	1998
Macro-Avg	0.16	0.33	0.22	1998
Weighted-Avg	0.24	0.49	0.33	1998

The SVM model for speech-based sentiment classification generated a classification report, which is presented in Table 2. The accuracy rate reached only 49% for this model in its operational stage. The model only identified neutral sentiments (0) in responses while missing positive and negative sentiment types, as shown through F1-scores of 0 for these categories.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

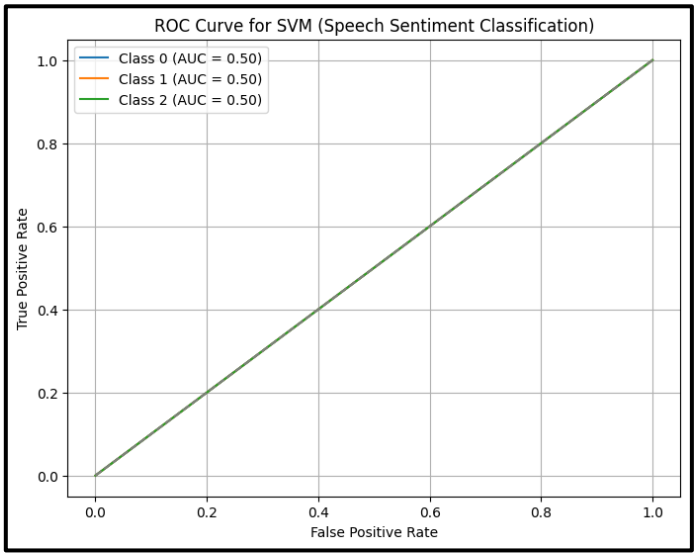

Figure 8. ROC Curve – SVM

Figure 8 shows the ROC curve generated from SVM speech-based sentiment classification model operations. The AUC score of 0.50 demonstrates that the model has no capability beyond basic random chance for making predictions. The model fails to effectively extract sentiment data from speech characteristics, indicating that future processing methods should be improved.

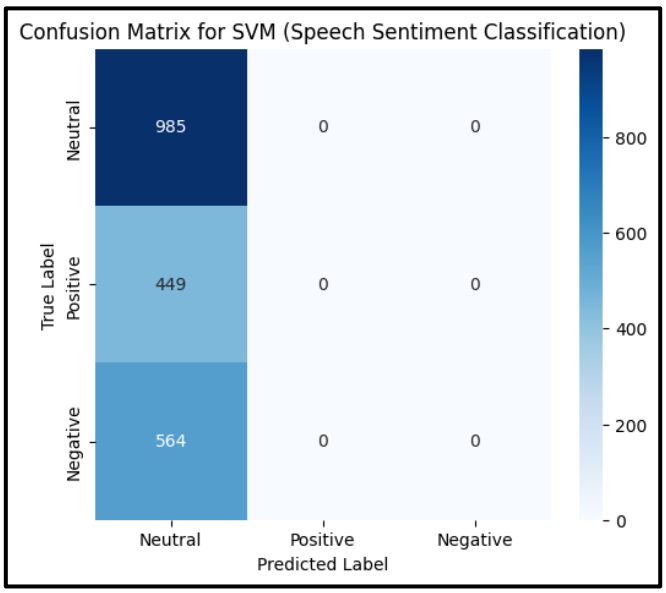

Figure 9. Confusion Matrix – SVM

Figure 9 shows the confusion matrix for the speech sentiment classification using the SVM approach. It indicates that all the test samples had a neutral context where the log of positive as well as negative sentiments was totally misclassified. This suggests a great imbalance in the model that probably favours the majority class; therefore, a need to balance the data and revise other features to increase the level of correct identification.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

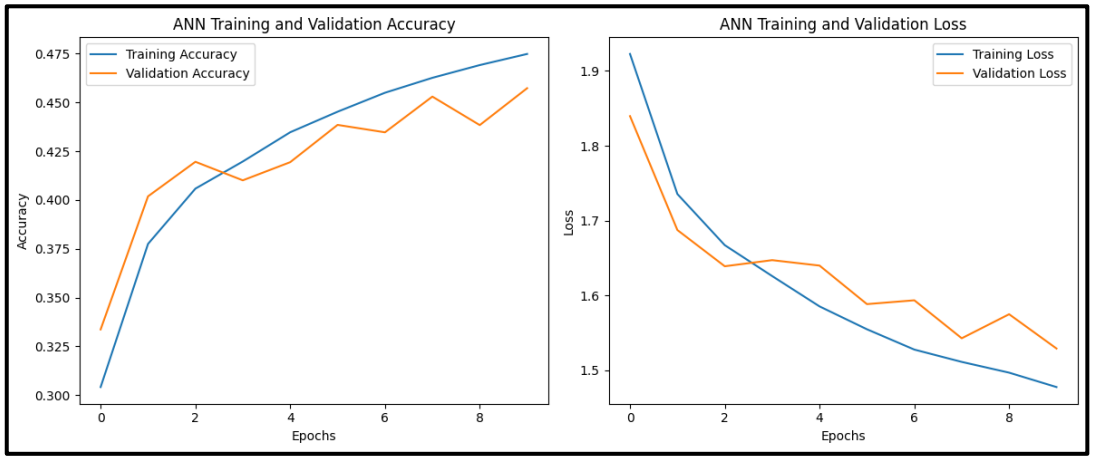

Figure 10. ANN – training and validation accuracy/loss curves

Figure 10 depicts the training and validation accuracy/loss curves for the ANN model. Accuracy rises gradually across epochs and gets to about 47.5%, and herein, validation is slightly below the training. Based on this loss graph, learning takes place, and there is a decrease in the losses. Nonetheless, the difference between the training and validation accuracies indicates signs of overfitting, which must be rectified using regularisation methods.

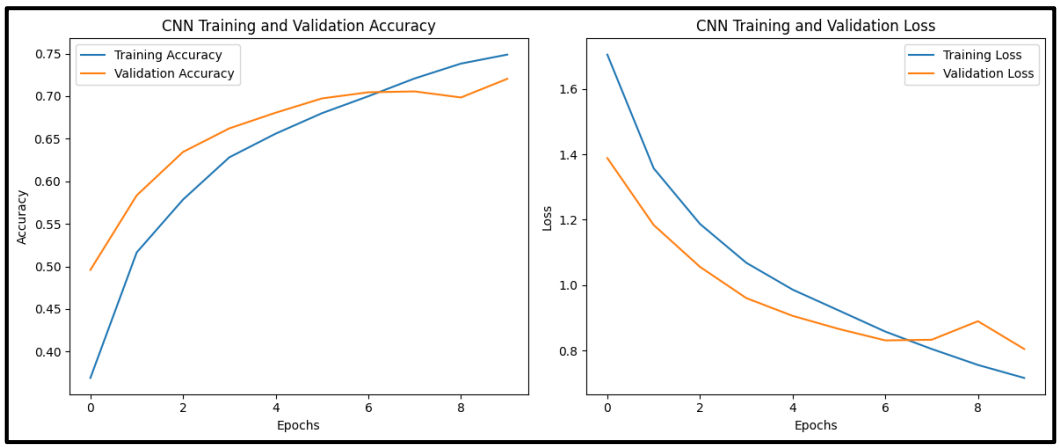

Figure 11. CNN – training and validation accuracy/loss curves

Figure 11 depicts the training and validation accuracy and loss curves for CNN. The CNN has a higher accuracy of about 75%, and the difference between training and validation accuracies is smaller, suggesting better model generalisation. The loss graph shows that it is decreasing progressively, which proves that there is an efficient optimisation. Moreover, CNN is stable in terms of learning as compared to ANN, which makes it suitable for vision-related tasks.

In order to get better insights into the model performance, further set of statistical analysis was carried out. The CNN based vision model gave an accuracy of 75% having a Macro-average F1-score of 0.72, while Random Forest (accuracy: 56% – (macro F1: 0.48) for AAE and SVM for a speech: (accuracy: 49%, macro F1: 0.22). Further, when a one-way ANOVA with a Tukey HSD test was used, it was confirmed that the performance difference between CNN and other models is statistically significant ($p < 0.05$). The advance may be explained by the superior performance of CNN at scanning spatial patterns for facial expressions, which were less blurred than indications of text or speech.

Other than accuracy, all models were also tested on precision, recall, F1-score and confusion matrix for the imbalanced sentiment classes. For example, the CNN model had a recall of 0.79 for neutral expressions and an F1-score greater than 0.70 in both cases of positive and negative classes, which performed better than text and speech-based models that exhibited biased prediction for neutral sentiment.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The SVM-based speech model significantly misclassified almost all of the positive and negative emotions as neutral, demonstrated in its F1-scores of 0 for classes 1 and 2. Perhaps, this can be attributed to an inadequate representation of subtle prosodic cues of speech by the MFCCs. Future research will involve Wav2Vec or spectrogram CNNs for extracting deeper emotional features and better sensitiveness to the vocal intonations.

Analysis and Discussions

The results obtained from the multi-modal sentiment classification models would like to report the following important findings related to the opinion of text, speech, and vision characteristics in emotion. In the case of the classification of the texts with the Random Forest model, 56% accuracy was obtained, and the best recall of the neutral sentiment was equal to 0.83, while positive sentiment and negative sentiment were equal to 0.27 and 0.33, respectively. This is to suggest that though the model is most efficient in identifying neutral expressions, it falters when it comes to positives and negatives, a shortcoming inherent to the decision taken to use only TF-IDF. The non-neutral, non-zero sentiments in testing are generally less accurately classified as per F1-scores, and this may require pre-processing techniques like word2vec or transformer models for further enhancement.

The SVM model that was applied for speech classification yielded a rather poor result of accuracy of 49%, which suggests that it was not effective in generalising across the sentiment classes. [46]. Therefore, the model classified most of the samples as neutral, as evidenced by the confusion matrix and the model's failure to differentiate between positive and negative samples. This is also upheld in the ROC curve, with the area under it being equal to 0.50, suggesting that the model has a random accuracy. As a result of the high variability in speech patterns, background noise and adequately described the speaker's physical characteristics, which are not by the MFCCs. Further, more elaborate methods like CNNs or LSTMs on spectrograms or other approaches might be required to extract better features from the audio data. [35] [37].

The two types of architectures of the vision-based sentiment classification models showed quite a variation in their performance in the best and worst cases. The feedback from the ANN model was low in terms of training accuracy at 47.5% and almost an equal value for validation accuracy, showing steady though restricted learning potential. The loss curve indicates that the performance improved over time, but the model's accuracy was not very high. Totally, CNN has shown a higher capability to capture the spatial features from the facial expressions and achieved a higher accuracy of 0.75 with a relatively better validation curve than ANN. The CNNs used in this study were able to generalise better to unseen images through their well-structured feature extraction of the convolution layers, proving that CNNs are effective in vision tasks [37].

The strong performance of the vision-based CNN model demonstrates the trustworthiness of facial expressions as an indication of sentiment and outperformed both the forms of texts and speech. This corresponds to the findings of affective computing that visual cues provide rich, direct emotional state indicators that tend to be more reliable than the vocal tone or written expression. CNN's capability of detecting spatial features from facial landmarks enabled it to detect mildly complex emotional patterns with 75% accuracy and over 0.70 F1 score. Conversely, the speech model based on MFCC features had difficulties with the nuanced prosodic cues, which commonly resulted in neutral classifications, which points to the weakness of the handcrafted features in discriminating the emotional variation in audio signals.

Based on the confusion matrix for vision-based classification, it is evident that neutral sentiments were mostly accurately classified, while positive and negative sentiments were mostly misclassified. This indicates that 'happy' and 'angry' faces may actually appear 'similar' or that there is not enough variation between the two classes in the dataset. Based on these challenges, the incorporation of data augmentation techniques, incorporating attention mechanisms or using transfer learning from other facial emotion recognition CNNs can further improve performance. [47]. In order to obtain high accuracy in the overall aspect of multimodal sentiment analysis, features derived from traditional machine learning algorithms

like Random Forest and Support Vector Machine need to be incorporated with features resulting from deep learning algorithms such as Artificial Neural Network and Convolutional Neural Network.

The scope of the implications of this multimodal framework covers several domains. In human-computer interaction emotion-aware systems could dynamically modify dialog and tone in relation to user-emotion, improving the responsiveness of digital assistants and chatbots. In mental health surveillance, the fusion of visual and vocal signals would aid the identification of the early manifestations of depression or distress especially in cases whereby verbal communication is restricted. Emotionally-charged media moderation can be of help for social media content moderation through the use of multimodal sentiment indicators, enhancing digital safety and well-being.

The limitations of this research are defined by the MELD dataset scope, which has a lack of cultural and linguistic diversity despite being extensive. The unimodal models result in performance limits as a result of sparsity of data, class imbalance and basic feature extraction. Future studies need to incorporate transformer-based architectures (BERT, Wav2Vec etc.), to improve the context sensitivity, feature-level fusion and attention mechanisms that will increase interpretability and alignment of the modalities. Extending into physiological signals, which include heart rate or conductance of the skin, which could further boost the emotion detection accuracy in an AI system operating in real-time and in multi-modal surroundings.

Conclusion

This study reveals the efficacy of combining text, speech, and vision in multimodal sentiment analysis. There were vision-based models that utilized Convolutional Neural Networks (CNN) and had the best accuracy of 75%, this being better compared to text-based Random Forest classifiers (56%), and those that were speech-based in nature and performed with limited success because of class imbalance and low feature variability. These results emphasize the power of visual clues in emotion recognition and further emphasize the inability of unimodal models to represent a diverse sentiment gamut of people.

The main contribution of this study is demonstrating the way, a conjugation of machine learning (Random Forest, SVM) and deep learning models (ANN, CNN) can be used. can be very helpful in improving the accuracy of the classification of sentiment. This study also highlights the importance of MELD dataset in training models that could handle real-world dialogue-based multimodal inputs. This framework has immediate application in the enhancement of emotion-aware AI systems like virtual assistants, mental health monitoring tools, intelligent tutoring systems, among others.

For further work, transformer-based networks like BERT for text and Wav2Vec for speech should be investigated to extract the deeper contextual and acoustic features. Further improvements are achieved by the improvement of the fusion strategies, especially at the feature level and the hybrid fusion level. to integrate temporal and semantic cues more between modalities. Besides, the addition of attention mechanisms and video-based facial expression recognition could potentially provide better performance in dynamic, in-the-moment settings.

References

[1] P. Chakriswaran, D. R. Vincent, K. Srinivasan, V. Sharma, C.-Y. Chang, and D. G. Reina, "Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues," *Applied Sciences,* vol. 9, no. 24, p. 5462, 2019.

[2] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion,* vol. 91, pp. 424-444, 2023, doi: https://doi.org/10.1016/j.inffus.2022.09.025.

[3] J. Chen, K. P. Seng, J. Smith, and L. M. Ang, "Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications," *IEEE Access,* 2024, doi: https://doi.org/10.1109/ACCESS.2024.3416370.

[4] M. Binte Rashid, M. S. Rahaman, and P. Rivas, "Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures," *Machine Learning and Knowledge Extraction,* vol. 6, no. 3, pp. 1545-1563, 2024, doi: https://doi.org/10.3390/make6030074.

[5] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces,* vol. 10, pp. 99-111, 2016, doi: https://doi.org/10.1007/s12193-015-0195-2.

[6] M. B. MOHAMED, "Deep learning and Sentiment analysis techniques to ensure customer satisfaction through AI bots," 2023.

[7] A. El-Ansari and A. Beni-Hssane, "Sentiment analysis for personalized chatbots in e-commerce applications," *Wireless Personal Communications,* vol. 129, no. 3, pp. 1623-1644, 2023, doi: https://doi.org/10.1007/s11277-023-10199-5.

[8] B. Omarov, S. Narynov, and Z. Zhumanov, "Artificial intelligence-enabled chatbots in mental health: A systematic review," *Computers, Materials & Continua,* vol. 74, no. 3, 2023, doi: https://doi.org/10.32604/cmc.2023.034655.

[9] M. Szabóová, M. Sarnovský, V. Maslej Krešňáková, and K. Machová, "Emotion analysis in human–robot interaction," *Electronics,* vol. 9, no. 11, p. 1761, 2020, doi: https://doi.org/10.3390/electronics9111761.

[10] A. Karpov and R. Yusupov, "Multimodal interfaces of human–computer interaction," *Herald of the Russian Academy of Sciences,* vol. 88, pp. 67-74, 2018, doi: https://doi.org/10.1134/S1019331618010094.

[11] T. M. Brill, L. Munoz, and R. J. Miller, "Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications," in *The role of smart technologies in decision making*: Routledge, 2022, pp. 35-70.

[12] J. W. Schofield, D. Evans-Rhodes, and B. R. Huber, "Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students," *Social Science Computer Review,* vol. 8, no. 1, pp. 24-41, 1990, doi: https://doi.org/10.1177/089443939000800104.

[13] L. Weitzel, R. C. Prati, and R. F. Aguiar, "The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques?," *Sentiment analysis and ontology engineering: An environment of computational intelligence,* pp. 49-74, 2016, doi: https://doi.org/10.1007/978-3-319-30319-2_3.

[14] D. Canedo and A. J. Neves, "Facial expression recognition using computer vision: A systematic review," *Applied Sciences,* vol. 9, no. 21, p. 4678, 2019, doi: https://doi.org/10.3390/app9214678.

[15] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion,* vol. 95, pp. 306-325, 2023, doi: https://doi.org/10.1016/j.inffus.2023.02.028.

[16] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," *arXiv preprint arXiv:1906.05681,* 2019, doi: https://doi.org/10.48550/arXiv.1906.05681.

[17] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy,* vol. 25, no. 10, p. 1440, 2023, doi: https://doi.org/10.3390/e25101440.

[18] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social network analysis and mining,* vol. 11, no. 1, p. 81, 2021, doi: https://doi.org/10.1007/s13278-021-00776-6.

[19] P. Berka, "Sentiment analysis using rule-based and case-based reasoning," *Journal of intelligent information systems,* vol. 55, no. 1, pp. 51-66, 2020, doi: https://doi.org/10.1007/s10844-019-00591-8.

[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics,* vol. 37, no. 2, pp. 267-307, 2011, doi: https://doi.org/10.1162/COLI_a_00049.

[21] D. M. El-Din, "Enhancement bag-of-words model for solving the challenges of sentiment analysis," *International Journal of Advanced Computer Science and Applications,* vol. 7, no. 1, p. 9, 2016, doi: https://dx.doi.org/10.14569/IJACSA.2016.070134.

[22] A. P. Pimpalkar and R. J. R. Raj, "Influence of pre-processing strategies on the performance of ML classifiers exploiting TF-IDF and BOW features," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal,* vol. 9, no. 2, p. 49, 2020, doi: http://dx.doi.org/10.14201/ADCAIJ2020924968.

[23] J. A. Nasir, A. Karim, G. Tsatsaronis, and I. Varlamis, "A knowledge-based semantic kernel for text classification," in *String Processing and Information Retrieval: 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011. Proceedings 18*, 2011: Springer, pp. 261-266, doi: https://doi.org/10.1007/978-3-642-24583-1_25.

[24] C. N. Tulu, "Experimental comparison of pre-trained word embedding vectors of Word2Vec, Glove, FastText for word level semantic text similarity measurement in Turkish," *Advances in Science and Technology. Research Journal,* vol. 16, no. 4, pp. 147-156, 2022, doi: https://doi.org/10.12913/22998624/152453.

[25] I. O. William and E. M. Altamimi, "Hierarchical Long Short-Term Memory (LSTM) Model for News Sentiment Analysis," 2024. [Online]. Available: https://www.researchgate.net/publication/381545632_Hierarchical_Long_Short-Term_Memory_LSTM_Model_for_News_Sentiment_Analysis.

[26] N. Passi, M. Raj, and N. A. Shelke, "A Review on Transformer Models: Applications, Taxonomies, Open Issues and Challenges," in *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, 2024: IEEE, pp. 1-6, doi: https://doi.org/10.1109/ASIANCON62057.2024.10838047.

[27] K. Cheng, Y. Yue, and Z. Song, "Sentiment classification based on part-of-speech and self-attention mechanism," *IEEE Access,* vol. 8, pp. 16387-16396, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2967103.

[28] M. Guia, R. R. Silva, and J. Bernardino, "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis," *KDIR,* vol. 1, pp. 525-531, 2019, doi: https://www.doi.org/10.5220/0008364105250531.

[29] K. M. Rezaul *et al.*, "Enhancing audio classification through MFCC feature extraction and data augmentation with CNN and RNN models," *International Journal of Advanced Computer Science and Applications,* vol. 15, no. 7, pp. 37-53, 2024, doi: https://doi.org/10.14569/ijacsa.2024.0150704.

[30] M. Kazemitabar, S. P. Lajoie, and T. Doleck, "Analysis of emotion regulation using posture, voice, and attention: A qualitative case study," *Computers and Education Open,* vol. 2, p. 100030, 2021, doi: https://doi.org/10.1016/j.caeo.2021.100030.

[31] M. Płaza, S. Trusz, J. Kęczkowska, E. Boksa, S. Sadowski, and Z. Koruba, "Machine learning algorithms for detection and classifications of emotions in contact center applications," *Sensors,* vol. 22, no. 14, p. 5311, 2022, doi: https://doi.org/10.3390/s22145311.

[32] F. Zhu-Zhou, R. Gil-Pita, J. García-Gómez, and M. Rosa-Zurera, "Robust multi-scenario speech-based emotion recognition system," *Sensors,* vol. 22, no. 6, p. 2343, 2022, doi: https://doi.org/10.3390/s22062343.

[33] M. D. Pawar and R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients," *Multimedia Tools and Applications,* vol. 80, pp. 15563-15587, 2021, doi: https://doi.org/10.1007/s11042-020-10329-2.

[34] M. I. Alam, F. I. Laiba, T. Nazi, and S. Choudhury, "A comprehensive hybrid framework for Parkinson's disease detection: integrating handcraft features along with deep learning-based feature extraction with variational autoencoder and traditional machine learning techniques for classification," Brac University, 2024. [Online]. Available: https://dspace.bracu.ac.bd/xmlui/handle/10361/25292?show=full

[35] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion,* vol. 99, p. 101869, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101869.

[36] F. Julin, "Vision based facial emotion detection using deep convolutional neural networks," ed, 2019.

[37] M. Shin, M. Kim, and D.-S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 2016: IEEE, pp. 724-729, doi: https://doi.org/10.1109/ROMAN.2016.7745199.

[38] T. H. Le, "Applying artificial neural networks for face recognition," *Advances in Artificial Neural Systems,* vol. 2011, no. 1, p. 673016, 2011, doi: https://doi.org/10.1155/2011/673016.

[39] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing,* vol. 19, no. 6, pp. 1635-1650, 2010, doi: https://doi.org/10.1109/TIP.2010.2042645.

[40] A. Fida *et al.*, "Real time emotions recognition through facial expressions," *Multimedia Tools and Applications,* pp. 1-28, 2023, doi: https://doi.org/10.1007/s11042-023-16722-x.

[41] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications,* vol. 32, no. 6, p. 121, 2021, doi: https://doi.org/10.1007/s00138-021-01249-8.

[42] T. A. Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. B. M. Su'ud, "Sentiment analysis using support vector machine and random forest," *Journal of Informatics and Web Engineering,* vol. 3, no. 1, pp. 67-75, 2024, doi: https://doi.org/10.33093/jiwe.2024.3.1.5.

[43] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment analysis with ensemble hybrid deep learning model," *IEEE Access,* vol. 10, pp. 103694-103704, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3210182.

[44] Z. Chen *et al.*, "Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models," *Computers, Materials & Continua,* vol. 80, no. 2, 2024, doi: https://doi.org/10.32604/cmc.2024.052618.

[45] A. Aguilera, D. Mellado, and F. Rojas, "An assessment of in-the-wild datasets for multimodal emotion recognition," *Sensors,* vol. 23, no. 11, p. 5184, 2023, doi: https://doi.org/10.3390/s23115184.

[46] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications,* vol. 80, no. 16, pp. 23745-23812, 2021, doi: https://doi.org/10.1007/s11042-020-09874-7.

[47] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics,* vol. 10, no. 9, p. 1036, 2021, doi: https://doi.org/10.3390/electronics10091036.