A unified multimodal learning framework for sentiment analysis and mental health indicators from YouTube videos

Discover Mental Health

RESEARCH Open Access

A unified multimodal learning framework for sentiment analysis and mental health indicators from YouTube videos

Priyanshu Satapathy[1], Onushka Chauhan[1], Deepika Kumar[1*], Preeti Sharma[1], Sumit Kumar Banshal[2], Oana Geman[3], Lucia Morosan-Danila[4*] and Jude D. Hemanth[5]

*Correspondence:
Deepika Kumar
deepika.kumar@bharatividyapeeth.edu
Lucia Morosan-Danila
lucia.danila@usm.ro
[1]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India
[2]Department of Computer Science and Engineering, Alliance University, Bangalore, India
[3]Chalmers University of Technology, and Gothenburg University, Gothenburg, Sweden
[4]Stefan cel Mare University of Suceava, Suceava, Romania
[5]Department of Electronics & Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India

Abstract

This study presents a multimodal deep learning framework designed to analyze sentiment patterns in YouTube videos and explore their association with early indicators of mental well-being. The approach integrates textual transcripts, vocal characteristics, and facial expressions into a unified representation to capture the emotional depth that individual modalities often miss. The model was trained and evaluated on a curated dataset of diverse YouTube content, and the results consistently showed that the fused architecture performed better than unimodal baselines. Compared with text-only or audio-only systems, the multimodal model achieved higher accuracy and fewer misclassifications, particularly in cases where speakers displayed subtle or mixed emotions. The integration of vocal cues such as pitch variation, speaking rate, and stress patterns helped clarify emotional ambiguity, while visual features such as micro-expressions, gaze direction, and facial tension added further clarity to the sentiment shifts within the videos. Transformer-based fusion delivered the most stable performance, demonstrating strong generalization across varied communication styles and recording conditions. In addition to reporting classification outcomes, the study examined how specific multimodal patterns correlate with non-clinical markers of mental health. Consistent associations were observed between fluctuating sentiment trajectories and indicators such as emotional instability, sustained negative tone, and reduced expressive variability. Instances where facial expressions contradicted verbal sentiment also showed relevance for identifying mild distress signals. These findings suggest that multimodal emotional cues can offer valuable insights into the affective state of content creators and may support research on digital well-being. The analysis also revealed challenges related to background noise, varying video quality, and inconsistent facial visibility, which influenced the reliability of certain features. Despite these limitations, the study demonstrates that combining audio, visual, and textual information provides a more complete and reliable picture of sentiment expression on social media platforms. The proposed framework offers a foundation for future systems aimed at understanding online emotional behavior and contributes to ongoing discussions on the responsible use of machine learning in mental-health-oriented applications.

Keywords Multimodal sentiment analysis, Multimodal fusion, Natural language processing, Tokenization, Machine learning, Deep learning, Extreme learning machine (ELM)

1 Introduction

Nowadays, people often share their views through social media, and it is not easy to understand what they really mean by it. Sites like YouTube, Facebook and Instagram is a place where millions of videos, posts and comments are published regularly [1]. Most of these posts consists of mix words, different tone, gestures and complex facial expressions. Therefore, to really figure out what are the intentions and feeling behind it, we need to look at all signals together. That's why sentiment analysis is both complex and interesting. Sentiment analysis, also known as opinion mining, aims to identify, understand and classify the sentiment behind videos [2]. It makes use of linguistics, psychology and computation to decide if statement is positive, negative or neutral. Sentiment analysis has gain importance with the increase in user- generated content online such as blogs, tweets, reels, etc. It helps organizations, researchers and even teachers to understand public opinion or reactions towards day-to-day trending topics [3, 4]. Earlier, sentiment analysis was majorly focused on text, researchers used methods like Naïve Bayes, SVM or Maximum Entropy to interpret ordering and linguistic relationship between words to get detailed context, attention to unigrams, bigrams and parts-of-speech [5, 6]. However, with evolution of social media, people started posting short, casual messages, full of slangs, emojis and sarcasm for which text-based approaches didn't work well so improvised ways to clean, process and perceive the text were introduced [7, 8], still these methods were not able to catch sarcasm, subtle emotions or conflicting feelings [9]. This led to need for multimodal sentiment analysis which look at audio and visual cues too along with text. Videos not only consist of words but tone, pauses and facial expressions. Sometimes, focusing only on one signal can be misleading. To overcome this, feature-level fusion [7] is used that combines all the features to get single vector, or decision-level fusion that merges predictions obtained from each modality [10, 11]. These approaches performed better than focusing on single modality and are better at identifying mixed or conflicting signals. Unifying multiple signals is not easy as text, audio and visuals don't always gives a single outcome and sometimes oppose each other [12]. Data gathered from various sources can be messy, with different sizes and structure, which makes integration difficult. It is also difficult to handle languages with low resources such as Hindi or tackle real-time problems. Despite these issues, it is required to look at different modalities in video to interpret sentiments just like humans and gain useful findings.

Various studies show that emotions and reasonings are correlated to each other strongly, establishing link between thoughts and feelings. Sentic blending is one of the methods that merges common reasoning with emotions to extract hidden clues that is not possible with traditional text analysis [13]. Studies also highlighted that combining text, audio and visuals extracted from real-time videos performs better for sentiment classification than using only one modality [14]. In short, sentiment analysis has evolved significantly. It is now more crucial to combine words, tone and visual gestures rather than simply classifying text. This approach will give a far clear understanding of how people feel [11]. Despite all of the developments, models still face problem to deal with

novel subjects, multiple local languages, and changing circumstances. These problems need to be resolved for advancement of sentiment analysis in field of social media, education and other sectors. The research contributions are as follows:-.

1. To bridge the gap between unimodal and multimodal sentiment analysis approaches in video understanding.
2. To enhance understanding of multimodal emotion expression in digital video content.
3. To develop a multimodal sentiment analysis framework integrating text, audio, and visual modalities for YouTube videos.
4. To extract acoustic features (mel spectrograms, MFCC, RMS energy) and visual features (brightness, contrast, emotion distribution) for comprehensive analysis.

The paper follows following structure: Sect. 1 introduces the Sect. 2 explains the literature review. Section 3 shows the architecture and workflow of pipeline developed. Section 4 discusses results and analysis, accompanied by conclusion.

2 Literature review

The primary goal of sentiment analysis, also referred to as opinion mining is to determine emotional tone behind people's words and behaviour. In comparison to traditional text mining, which only focuses on facts and words, sentiment analysis aims to gather individual opinions, feelings, attitude and classify it as positive, negative or neutral. These viewpoints have become immensely valuable for companies and researchers to make decisions as the web is continuing to expand quickly, with reviews, blogs, social media posts, and forums. Bing Liu [3] conducted significant studies in this area, including topics like identifying subjective remarks and identifying spam perspectives, as well as finding comparative reviews. He employed supervised as well as unsupervised machine learning, lexicons and parsing methods to convert unorganized text into form which have real life applications ranging from refining product to perform market trend analysis. With integration of computer science, linguistics and NLP, sentiment analysis has gradually developed and evolved over time. Its applications have reached broader scope rather than just classifying data as positive, negative or neutral. For instance, Maite Taboada [4] focused on enhancing sentiment detection through the use of linguistic cues, context, and sentence structure, including dependency parsing and part-of-speech tagging, the techniques picked up on nuances that would be missed by a simple keyword count. Authors in [5] concentrated on feature engineering in a similar manner. In addition to standard classifiers like Naïve Bayes, Maximum Entropy, and SVM, unigrams, bigrams, and POS tags were employed. Moreover, for analysing posts on social media that involve informal language full of slangs, dictionaries containing acronyms and emojis were used. But challenges still appear while understanding conflicting emotions in such data or to adjust models across different channels. Because of social media, massive amount of short, opinion-focused content is regularly generated by platforms such as Twitter. According to Aliza Sarlan et al. [6], manual analysis of this large volume of data is not possible. They explained the sentiment classification through machine learning in addition with features like emojis, hashtags and token frequency. However, real social-media language is much less predictable. It includes slang, acronyms, and frequent style changes that obscure emotional cues. To stay effective, models need to interpret this rapid and often unstable pattern of communication. To clean up messy

social media posts, text-based techniques have advanced with preprocessing steps like tokenization, stop word removal, and stemming or lemmatization [7]. Traditional models frequently miss the order and context of words, but deep learning—particularly models like Bi-Directional LSTMs—helps capture these aspects. It seems to be even more effective to combine deep learning with traditional methods, which strengthens and expands models for all kinds of online text [10]. Researchers are now investigating multimodal sentiment analysis in addition to text. Combining text, audio, and video features results in more accurate sentiment detection, according to Quentin Portes et al. [12]. It's difficult to analyse you tube video spoken words alone and deriving a meaningful text. Elements such as vocal tone, facial movements, pauses, and other non-verbal cues contribute significantly to the underlying meaning. In multimodal systems, these signals are merged into a unified representation through feature-level fusion, enabling classifiers like SVMs and neural networks to process them collectively. Tools such as OpenFace for visual extraction, OpenSMILE for acoustic features, and various text-embedding methods support this integration. However, deploying such computationally intensive models on devices with limited processing capacity remains a substantial challenge. Soujanya Poria et al. [10] analysed and processed text, audio, and video altogether using CNN-LSTM to get better results. This method collects context and timing for a variety of data types. Harika Abburi et al. [11] included acoustic features like pitch, intensity, mel-spectrograms and MFCC with visual features like facial expressions and gestures. They used deep neural networks, SVMs and Markov models for classification. Results revealed that combining multiple modalities show better performance especially for variety of languages and subjects. However, more research is need to be done for low-resource languages like Hindi, Marathi, Devnagari etc. Recent papers engaged with quick learning algorithms and fusion techniques to merge text, audio and video features for handling multimodal data. Feature-level fusion or deep learning combined with traditional methods is frequently used to enhance performance [14, 15]. Mansoorizadeh et al. [16] introduced asynchronous feature-level fusion that aligns features with varying timestamps to form a single vector for classification. Moreover, studies proved that adding facial expressions and body movements with decision and feature-level enhances emotion recognition [17]. For fast and scalable sentiment analysis, Extreme Learning Machine (ELM) is very useful. Researchers demonstrated that ELM requires less training time compared to basic neural networks as it randomly sets parameters and calculates for output directly [18]. Some variations, like Circular-ELM (C-ELM), are used for visual quality evaluation [19] and speech recognition [20] because of their stability. ELM determine sentiment just like humans in real time by integrating common reasoning with emotions [15]. It is supported by tri-modal sentiment classification that can process text, audio and video quickly and precisely [21]. This technique is further improved by sentic fusion that combines semantic and emotional data together. It allows system to observe changes in sentiment over time consistently, even when using data from multiple sources with varying scales or labels [11]. Fuzzy logic is also used in identification and characterization of sentiments. This is because sentiment always cannot fit into the intervals of 0 or 1. Sometimes, sentiment can be combination of multiple intensities of emotions [22]. Robust and scalable sentiment analysis method can be developed by employing ELM, sentic and multimodal fusion [23, 24]. These mechanisms can be used

for smart learning in schools, social media monitoring and human-computer interaction (Table 1).

Existing sentiment analysis studies demonstrate strong performance using hybrid machine learning and deep learning approaches, particularly for text-based social media data. However, these models often struggle with contextual ambiguity, scalability, and adaptation to multilingual or low-resource settings. Recent transformer-based multimodal models address contextual and cross-modal dependencies more effectively, but they introduce high computational complexity and limited deploy ability in resource-constrained environments. Furthermore, most transformer-driven approaches remain data-intensive and underexplored for Indian and other low-resource languages. Motivated by these gaps, this research aims to explore an efficient sentiment analysis framework that balances contextual understanding, computational efficiency, and adaptability across modalities and domains.

3 Proposed methodology

The multimodal sentiment analysis framework built for videos examine three major inter-related data modalities: text, audio and images. Textual information is provided by video transcripts, audio information is provided by speech and acoustic characteristics, and visual cues are provided by facial expressions and frame-level details. An independent analytical technique that detects sentiment and emotion-specific trends over time is applied to each modality. Several sophisticated algorithms are employed for each modality to guarantee robustness, and a comparison is carried out using a 10-fold

Table 1 Literature review

Source	Authors	Data/modality	Methodology	Key contribution	Limitations
[4]	Taboada et al.	Text	Linguistic analysis	Context-driven sentiment interpretation	Weak on informal text
[5]	Sahayak et al.	Twitter (Text)	NB, SVM, MaxEnt	Feature-rich tweet sentiment analysis	Mixed sentiment handling
[7]	Albladi et al.	Twitter (Text)	Review of ML & DL models	Comprehensive comparison of sentiment techniques	Limited multilingual focus
[8]	Jonnala et al.	Twitter (Text)	Bi-LSTM + Logistic Regression	Context-aware hybrid sentiment classification	High computational cost
[12]	Portes et al.	Multimodal	Feature-level fusion DNN	Multimodal sentiment detection for embedded systems	Resource-intensive
[11]	Abburi et al.	Multimodal	DNN, SVM, HMM	Comparison of fusion strategies	Low-resource language gap
[21]	Poria et al.	Multimodal	ELM-based fusion	Improved accuracy over unimodal models	Population variability
[25]	Smith-Mutegi et al.	Twitter (Text)	Bi-LSTM with ML classifier	Analysis of AI perception in STEM education	Social media bias
[26]	Tsai et al.	Text, Audio, Video	Multimodal Transformer (MulT)	Cross-modal attention for sentiment fusion	High computational demand
[27]	Yang et al.	Multimodal	BERT-based fusion	Context-rich multimodal representation	Limited real-time use
[28]	Yu et al.	Text, Visual	Vision-Language Transformer	Improved multimodal alignment	Large data requirement
[29]	Mai et al.	Multimodal	Self-Attention Transformer	Temporal sentiment modeling	Low-resource adaptability
[30]	Wang et al.	Multimodal	Unified Transformer Encoder	Robust sentiment integration	Deployment complexity

cross-validation. The entire process for the suggested multimodal sentiment analysis framework is depicted in Fig. 1.

3.1 Data acquisition and preprocessing

A dataset of 100 YouTube videos which included mix of TED Talks, how-to guides, emotional music, documentaries and educational content has been utilised for this research. yt-dlp videos are saved as MP4 files and maxes out at 720p. The aim is to get a good spread across different types of sentiment for different videos. YouTube transcript API has been used to capture the captions with their start time, end time and how long they playing. First five segments were gathered from each video, where each segment ranges from 3 to 5 s. Each video is assigned a label that describes its sentiment as – positive, neutral or negative. Each segment is passed through individual modality pipelines to get extensive understanding. The newly added videos were carefully selected enhance generalizability and reduce domain bias, cover diverse tones, genres, and topics, including: Motivational and inspirational talks, Educational and scientific lectures, Documentary-style narratives, TED/TEDx talks with varied speaking styles, Emotionally negative content (e.g., sadness, anger, distress). These videos involve different speakers, recording environments, lighting conditions, facial expressions, and camera qualities, ensuring exposure to a broad range of real-world scenarios. The original dataset exhibited a high proportion of neutral segments. To counter this issue, we adopted a category-aware video selection strategy, intentionally incorporating emotionally expressive content (both positive and negative) alongside neutral material. As a result: the sentiment distribution is now significantly more balanced across positive, neutral, and negative classes. This reduces bias toward the majority class and enables fairer model evaluation. In addition, segment-level aggregation across multiple frames and videos further stabilizes class distributions during evaluation.

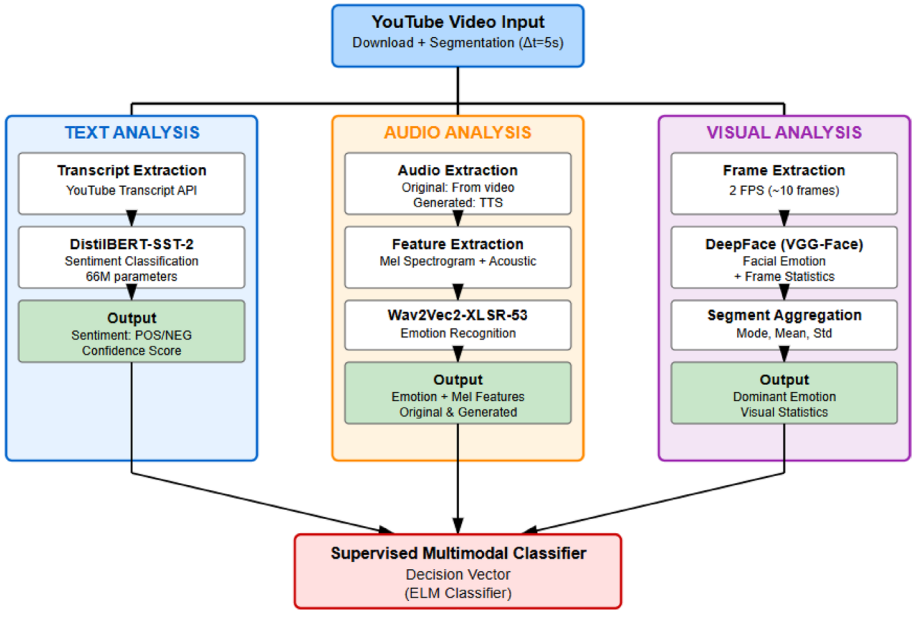


Fig. 1 Multimodal sentiment analysis system

3.2 Text modality analysis

The text modality uses a set of six complementary sentiment analysis models, including transformer-based deep learning approaches and basic rule-based algorithms, to forecast the general sentiment for each time segment based on transcript data. This multi-model architecture combines multiple algorithms to provide reliable sentiment classification using transformer based models, rule based models and ensemble integration of models. The text modality pipeline is shown in Fig. 2:

3.2.1 Transformer based models

- DistilBERT-SST-2. It is a BERT version with 66 million parameters that is fine tuned. Transcripts are tokenized (up to 512 tokens) and classified as binary sentiment.

$$\text{y\_distilbert} \in \{\text{POSITIVE, NEGATIVE}\}$$

- BERT-Multilingual: is a model with 110 million parameters that has been trained for multiple star ratings. The outputs are mapped to three categories of sentiment:

$$\begin{aligned}\text{y\_bert} = \text{POSITIVE, if rating} \in \{\text{4-star}, \text{5-star}\}\\ \text{NEUTRAL, if rating} = \text{3-star}\\ \text{NEGATIVE, if rating } \in \{\text{1-star, 2-star}\}.\end{aligned}$$

- RoBERTa-Twitter: The cardiffnlp/twitter-roberta-base model (125 M parameters) optimizes for social media sentiment and makes three-class predictions.
- FinBERT: A financial domain model (ProsusAI/finbert) is utilized to assess cross-domain generalization for general sentiment tasks.

3.2.2 Rule-based models

- VADER (Valence Aware Dictionary and Sentiment Reasoner): A lexicon-based method for social media content analysis. It generates a compound sentiment score, s_compound ∈ [-1, 1], which is subsequently transformed to three-class sentiment using threshold-based methods:

$$\begin{aligned}\text{y\_vader} = \{\text{POSITIVE, if s\_compound} \geq ?0.05.\\ \text{NEGATIVE, if s\_compound} \leq -0.05\\ \text{NEUTRAL, otherwise}\}\end{aligned}$$

- TextBlob: A pattern-based sentiment analyzer that assigns polarity ratings from −1 to 1. Classification employs similar thresholding:

$$\begin{aligned}\text{y\_textblob} = \{\text{POSITIVE, if } p > 0.1\\ \text{NEGATIVE, if } p < -0.1\\ \text{NEUTRAL, otherwise}\}\end{aligned}$$

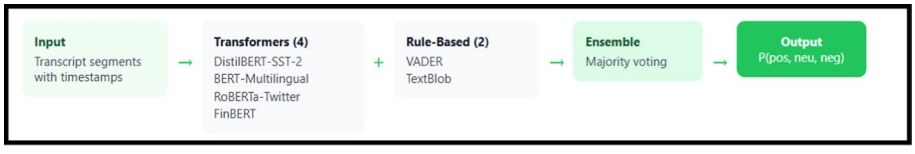


Fig. 2 Text modality pipeline

3.2.3 Ensemble integration

This mechanism encompasses predictions from all the six models applied through majority voting strategy to generate accurate sentiment classification. It merges results from four transformer-based models with rule-based models to decide final sentiment for each text segment. Predictions of individual models was merged using majority voting to create a reliable ensemble prediction:

$$\mathrm{y\_text\_ensemble} = \mathrm{mode}(\{\mathrm{y\_distilbert}, \mathrm{y\_bert}, \mathrm{y\_roberta}, \mathrm{y\_finbert}, \mathrm{y}_{\mathrm{vader}}, \mathrm{y\_textblob}\}),$$

where Mode(·) indicates which of the six models' sentiment classes is most prevalent.

In case of equal votes, transformer-based models are considered rather than rule-based methods because they offer a deeper understanding.

3.3 Audio modality analysis

This pipeline computes acoustic features by extracting five second audio segments in WAV format. Librosa library is employed to compute features like Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) to capture timbral patterns. Other features include Root Mean Square (RMS) energy, spectral centroid, spectral rolloff and zero-crossing rate. Two emotion recognition algorithms are applied: the transformer-based Wav2Vec2-XLSR-53 model that generates probability distribution across seven emotion classes, and HuBERT Large which is also transformer-based model to categorize sentiment as positive, negative or neutral. The audio modality analysis pipeline is shown in Fig. 3:

3.3.1 Audio extraction

The auditory modality examines acoustic and paralinguistic features that transmit emotional meaning other than verbal semantics. Five-second audio segments in WAV format with a 16 kHz sample rate, a mono channel, and the PCM S16LE codec are extracted using FFmpeg. To ensure exact alignment with transcript boundaries, the extraction tool performs temporal segmentation utilizing the -ss argument for start time and the -t parameter for duration. For the purpose of feature extraction and model inference, the gathered audio recordings are temporarily retained.

3.3.2 Feature extraction

Acoustic features were extracted from audio using Librosa library. Mel spectrograms are calculated using n_mels-128 frequency bands that ranges up to f_max = 8000 Hz. Significant statistical measures, like mean spectral energy (μ_mel), standard deviation (σ_mel), that represents variability, and min/max values (S_max, S_min), are obtained by transforming spectrograms into decibel scale.

Fig. 3 Audio modality analysis

The Librosa library is used to extract acoustic features. Mel spectrograms are calculated using n_mels = 128 frequency bands that span up to f_max = 8000 Hz. Major statistical measures, such as mean spectral energy (μ_mel), standard deviation (σ_mel), which indicates variability, and maximum/minimum values (S_max, S_min), are obtained by converting power spectrograms into the decibel scale. In addition, several more acoustic features are calculated: Mel-Frequency Cepstral Coefficients (MFCCs) that capture timbral patterns, Root Mean Square (RMS) Energy that indicates sound intensity. The signal brightness is represented by the spectral centroid. Spectral Rolloff, which describes the form of the frequency energy distribution, and Zero-Crossing Rate, which indicates frequency fluctuation through changes in signal sign. RMS energy is defined as:

$$E_{RMS} = \sqrt{\frac{\sum x^2(n)}{N}} \tag{1}$$

where x(n) is the audio signal and N is the number of samples.

3.3.3 Emotion recognition algorithms

Audio-based emotion classification has been done by using two complementary algorithms. The first algorithms uses Wav2Vec2-XLSR-53. This transformer-based model is fine-tuned for speech emotion recognition after undergoing self-supervised pre-training on 56,000 h of multilingual speech data. The model creates probability distributions for the emotion classes "neutral, happy, sad, angry, fear, disgust, and surprise" based on raw audio waveforms. The transformation maps emotions to three classes of sentiment are represented as:

$$\begin{aligned} \text{y\_wav2vec} \;=\; \{\; & \text{POSITIVE, if e} \in \{\text{happy, joy, excited, calm}\} \\ & \text{NEGATIVE, if e} \in \{\text{sad, angry, fear, disgust}\} \\ & \text{NEUTRAL, otherwise}\} \end{aligned}$$

where e represents the recognized emotion with the highest posterior probability.

Another algorithm employed is HuBERT Large (Hidden-Unit BERT), which is also a transformer- based speech recognition model that undergoes self-supervised pre-training on large -volume unlabelled speech data.

Let: H = {h_1, h_2, ..., h_T} be frame-level embeddings from HuBERT Large.

$\bar{h} = \frac{1}{T}\sum_{t=1}^{T} h_t$ be the pooled utterance-level representation.

f (.) be a trained classifier.

$$\begin{aligned} \text{y\_HuBERT} \;=\; \{\; & \text{POSITIVE, if } P_{\text{pos}}(\bar{h}) > \tau_{pos} \\ & \text{NEGATIVE, if } P_{\text{neg}}(\bar{h}) > \tau_{neg} \\ & \text{NEUTRAL, otherwise}\} \end{aligned}$$

where P_{pos}, P_{neg} are class probabilities from classifier and τ_{pos}, τ_{neg} are learned or empirically chosen thresholds.

3.4 Visual modality analysis

This pipeline examines frame-level data and facial expressions to detect emotions through low-level visual statistics and provide precise emotion recognition. FFmpeg is employed to extract frames for segment with temporal synchronization. OpenCV is used

to compute statistics like brightness, sharpness and contrast to achieve comprehensive analysis. Two algorithms are applied for facial emotion recognition: DeepFace with Haar Cascade classifiers for face detection, and FER that uses Multi-task Cascaded Convolutional Networks for face localization. These algorithms classify frames into seven emotional categories which are then converted to three sentiment classes. Figure 4 shows the visual modality pipeline developed. Visual Modality Analysis depicted in Fig. 4:

3.4.1 Feature extraction and processing

The frames are extracted and processed using FFmpeg library, which generates ten frames for every five-second video segment, that is two frames per second. Each frame is stored in JPEG format with mild compression. For synchronization, it lines up frames with both audio and transcript extracted before. Further, each frame is checked for low-level visual details and also for facial emotions to interpret what people are feeling. This method not only focus on people but also considers nearby environment to give more detailed understanding. Apart from emotion recognition, three image statistics is calculated per frame using OpenCV, brightness (B) represents the average pixel intensity as shown in Eq. (2):

$$B = \frac{W \times H}{\sum I(i,j)} \tag{2}$$

W and H represent the width and height of the image. I(i, j) represent pixel intensity at position (i, j). Contrast represents standard deviation of these pixel intensities. Sharpness, on the other hand, comes from variance of Laplacian operator and represents strongness of edges. For each segment, we aggregate frame-level data of all the 10 frames to get single that represent entire segment. The most frequent prediction across these frames becomes the dominant emotion for the segment.

Emotion consistency is about how consistent a particular emotion is across all these frames:

$$\alpha\_visual = \frac{count(e\_dominant)}{F} \tag{3}$$

Face detection reliability is expressed as:

$$\rho\_face = \frac{frames\_with\_faces}{F} \tag{4}$$

indicating analysis reliability. The average visual statistics (brightness, contrast, and sharpness) are computed as follows:

$$u\_stat = \frac{\sum stat\_f}{F} \tag{5}$$

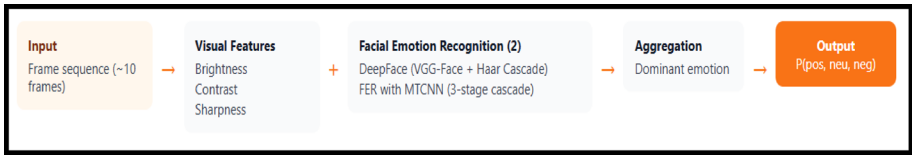


Fig. 4 Visual modality pipeline

3.4.2 Facial emotion recognition algorithms

DeepFace (VGG-Face Backend) and FER (facial expression Recognition) algorithms are used for facial emotion recognition. DeepFace (VGG-Face Backend) consists of 16 convolutional layers that is trained on millions of images. To ensure robustness when faces are not clearly visible or multiple faces appear, OpenCV Haar Cascade classifiers are used for face identification. The algorithm assigns a probability distribution (p_e) to each of the seven categories of emotions—disgust, anger, fear, happiness, sadness, surprise, and neutral. For all emotion classes, e_dominant = argmax of p_e.Emotions are then translated to three types of sentiment:

$$\begin{aligned} \text{y\_deepface} \;=\; & \{\ \text{POSITIVE, if e\_dominant} \in \{\text{happy, surprise}\} \\ & \text{NEGATIVE, if e\_dominant} \in \{\text{sad, angry, fear, disgust}\} \\ & \text{NEUTRAL, if e\_dominant} \;=\; \text{neutral}\} \end{aligned}$$

The FER algorithm uses a CNN-based emotion classifier and Multi-task Cascaded Convolutional Networks (MTCNN) to identify faces. It gradually improves face localization using three-layer architecture (P-Net, R-Net, and O-Net) and offers high precision in face detection. Face confidence score is calculated as:

$$c\_face\,(f) = \frac{w\_f \times\ h\_f}{W \times\ H} \tag{6}$$

3.5 Multimodal supervised fusion classifier

The Supervised Classification Layer generates the final sentiment decision by processing the fused representation derived from the text, audio, and visual streams. To ensure robust prediction, three complementary supervised learning models are employed. The Extreme Learning Machine (ELM) serves as the primary classifier due to its exceptionally fast training capability and its effectiveness in handling linearly separable patterns within high-dimensional multimodal features. In parallel, a Support Vector Machine (SVM) is incorporated as a comparative model, leveraging its margin-maximization strategy to deliver stable classification boundaries and strong generalization across varied sentiment classes. The third classifier, Random Forest, contributes an ensemble-based perspective by aggregating multiple decision trees, enabling the system to capture deeper nonlinear relationships and maintain reliability even when the multimodal inputs contain irregularities or noise. Together, these classifiers form a complementary decision framework that strengthens the overall predictive quality of the system. Given the multimodal vector xfusion, the output prediction is computed using Eq. (7):

$$\hat{y} = sign(H\ \dagger\ T) \tag{7}$$

where

- H is the hidden-layer activation matrix.
- H†T is its Moore–Penrose pseudoinverse.
- T is the target label matrix.

3.6 Evaluation metrics

Cross validation strategy has been used for evaluating the performance of various algorithms. Dataset of segments is divided into 10 folds with stratified sampling, so every

fold has the uniform balance of sentiment classes. To avoid overfitting and comparing the performance of all algorithm 10-fold cross-validation is used. This method generates ten different train-test splits, with each segment acting as a test sample just once. For each round, k ∈ {1, 2,…, 10}:

- The training set contains 90% of the data.
- Tested 10% of data across five segments.
- To assure reproducibility, generate a random state with a seed of 42.
- Shuffle to eliminate temporal bias.

Accuracy, Precision, Recall and F1 score has been used for the evaluation of classification algorithm. The number of samples with true class I predicted as class J, aggregated across all 10 folds, is represented by the element CM[i, j] in a C × C matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times \ Precision \times \ Recall}{Precision + Recall} \tag{11}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

4 Results and analysis

The detailed analysis of proposed multi-video text sentiment analysis pipeline is discussed in this section, which is constructed by integrating six already trained sentiment analysis models: DistilBERT, BERT Multilingual, RoBERTa-Twitter, FinBERT, VADER, and TextBlob. These transcripts were picked based on 100 different YouTube videos on the various topics and emotional expressions. The hundred videos were overall sampled into text segments that were analysed. A 10-Fold cross-validation was performed to make sure that the evaluation was comprehensive and reliable, as it gave each model a better insight into its performance. The label of the end segment was decided in the form of a majority-vote mechanism of each segment of text across the various models utilized. It is identified that each of the models worked in aligned rather differently with each other, presumably due to each of them having strengths and weaknesses regarding language understanding, processing messy or short transcripts, and different data training.

- **Cross-validation evaluation metrics**.

10-Fold Cross-Validation was used to evaluate each model's ability to generalise across domains. For Accuracy, Precision, Recall, and F1-Score, the mean and standard deviation were calculated across ten folds.

4.1 Text model evaluation

In the comparison of models, DistilBERT turned out to be the weakest performer. Since it was originally trained on SST-2, which consists of neat and concise movie-review style sentences, it could not cope well with the messy, unstructured nature of YouTube transcripts. YouTube transcripts frequently include hesitations, filler expressions, informal wording, and abrupt conversational shifts, which collectively make sentiment interpretation considerably more challenging. Among the transformer models evaluated, BERT-Multilingual showed comparatively stronger performance, largely due to its exposure to diverse linguistic structures during pre-training, which helped it manage grammatical variability. However, its broad lexical coverage occasionally reduced its sensitivity to subtle sentiment cues that appear in everyday conversational English. RoBERTa-Twitter demonstrated the highest effectiveness within the transformer group; its adaptation to social media discourse enabled it to process short, informal utterances with ease and to recognize elements such as emojis and casual phrasing that commonly appear in spoken YouTube content. In contrast, FinBERT produced only moderate results, reflecting the mismatch between its domain-specific training on financial documents and the general emotional expressions found in user-generated speech. Notably, VADER, despite being a simple lexicon-driven model, surpassed majority of transformer-based approaches by accurately capturing conversational tone, punctuation-driven emphasis, and sentiment intensity markers typical of online interactions. TextBlob and RoBERTa-Twitter achieved the strongest overall performance, recording an accuracy of 0.8100 along with balanced precision and recall, supported by its stable polarity scoring mechanism and suitability for natural, dialogue-like text. A comparative summary of the performance of the text-based models is presented in Table 2.

The comparative performance table highlights notable differences in how each text-based sentiment model generalizes across the 10-fold cross-validation. TextBlob and RoBERTa emerged as the strongest performers, achieving accuracies of 81% for both, with RoBERTa additionally obtaining the highest overall precision of 84.5%. Lexicon-driven approaches showed particular strength when handling short, conversational transcripts and segments with frequent topic shifts, which is reflected in their low standard deviation values and consistent behaviour across folds. RoBERTa-Twitter also delivered competitive results, underscoring its suitability for informal, dialogue-oriented content. In contrast, domain-specific transformer models such as FinBERT and DistilBERT recorded lower accuracies of 60% and 39%, indicating a clear mismatch between their pre-training objectives and the diverse linguistic patterns characteristic of YouTube speech. BERT-Multilingual reached an accuracy of 50%, suggesting challenges in maintaining stable performance across folds, likely due to the heterogeneous nature of the dataset. The relatively higher standard deviations observed in the transformer-based models indicate their sensitivity to limited training samples and potential class

Table 2 Comparative analysis of performance of various text based models

Model	Accuracy	Precision	Recall	F1-Score
TextBlob	0.8100 ± 0.1136	0.8260 ± 0.1237	0.8100 ± 0.1136	0.8042 ± 0.1171
RoBERTa	0.8100 ± 0.1446	0.8455 ± 0.1015	0.8100 ± 0.1446	0.7941 ± 0.1495
VADER	0.7500 ± 0.1628	0.7848 ± 0.1701	0.7500 ± 0.1628	0.7424 ± 0.1705
FinBERT	0.6000 ± 0.0894	0.4184 ± 0.1062	0.6000 ± 0.0894	0.4758 ± 0.0979
BERT-Multilingual	0.5000 ± 0.1673	0.5602 ± 0.2906	0.5000 ± 0.1673	0.4654 ± 0.2187
DistilBERT	0.3900 ± 0.1578	0.2145 ± 0.1325	0.3900 ± 0.1578	0.2673 ± 0.1504

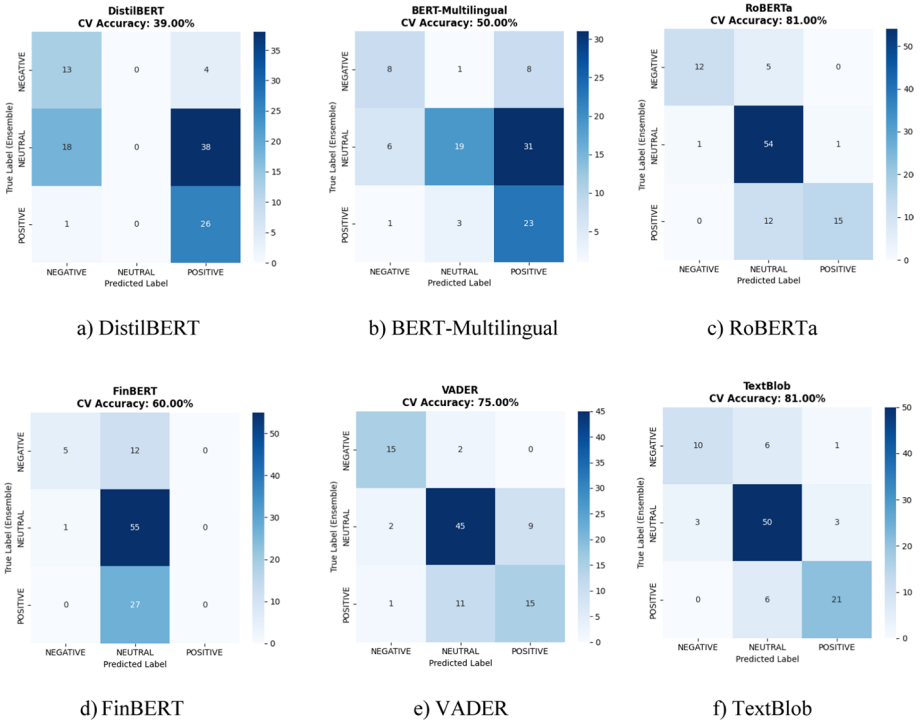
imbalance. Overall, the findings show that lexicon-based models offer the most reliable and stable performance for sentiment analysis of short, informal, and multi-topic transcripts, making them particularly suitable for real-world YouTube data.

Lexicon-based models consistently outperformed transformer models in this dataset, with TextBlob (81%) achieving the highest accuracy and showing strong stability across all 10 folds. These models were less affected by shifts in video domain or topic variability, whereas transformer-based models struggled with domain mismatch—particularly when trained on narrow or specialised corpora—which led to high variance and inconsistent performance across folds. BERT-Multilingual and DistilBERT showed notable instability, and transformers generally found it difficult to handle poetic, scientific, or abstract transcripts, resulting in misclassifications. Among them, RoBERTa-Twitter performed the best with 81% accuracy, likely due to its alignment with conversational and social-media-like language, while FinBERT and DistilBERT underperformed due to poor generalisation on diverse YouTube transcripts. High standard deviations across folds further indicate that all transformer models were sensitive to class imbalance and the limited dataset size. Figure 5a–f shows the confusion matrices of different text models. Figures 6 and 7 depicts model performance comparison of 10-cross validation and accuracy of 10 folds respectively.

4.2 Visual model evaluation

The visual sentiment analysis pipeline, which analyses face expressions and frame-level visual characteristics taken from 100 legitimate video clips, is thoroughly evaluated in this part. We evaluated two algorithms:

- A deep CNN-based facial expression classifier called DeepFace (VGG-Face backend).



a) DistilBERT b) BERT-Multilingual c) RoBERTa

d) FinBERT e) VADER f) TextBlob

Fig. 5 **a**–**f** Confusion matrix of text models used for analysis

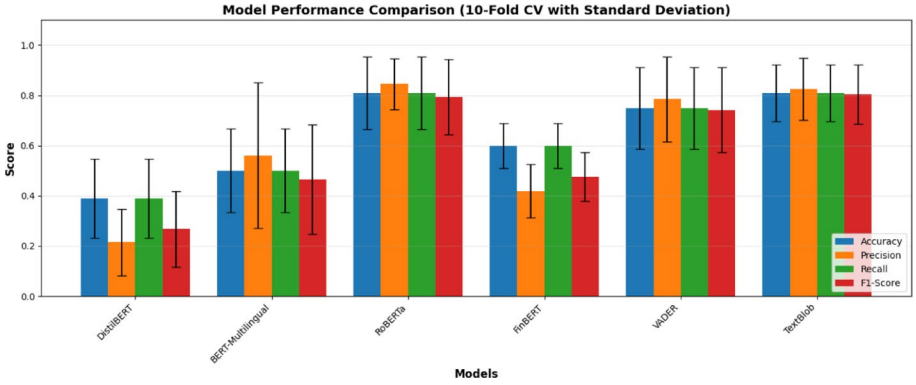


Fig. 6 Model performance comparison (10-fold cross-validation)

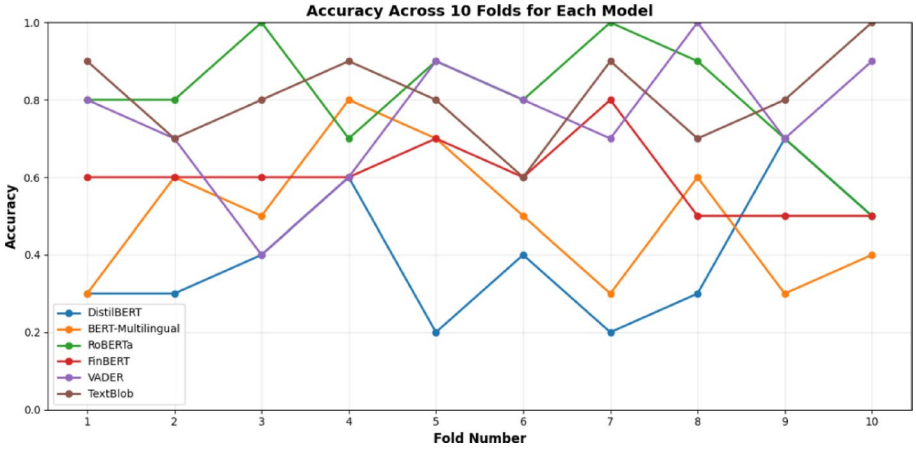


Fig. 7 Accuracy across 10 folds for each model

- FER is a lightweight model for facial expression recognition (OpenCV Haar Cascade is used as a fallback when built-in FER is not available).

There were ten frames that were always sampled in each segment. In order to facilitate the qualitative analysis, a set of low-level visual features were created in individual frames, including brightness, contrast, sharpness, hue, saturation, edge density. Sentiment predictions were generated by DeepFace and FER on each of the segments. A 10-Fold Cross-Validation method was used in order to ensure reliability in this study in spite of the small sample size. The end results are provided as means of performance of all folds, and the corresponding standard deviations.

The visual features are extracted using sharpness, Hue and saturation, edge density, Face detection rate. Sharpness is calculated through Laplacian variance to get the edge detail and the level of focus. Hue and Saturation is derived from the HSV colour space to analyse colour composition and richness.

The edge density is identified by proportionating edges using Canny edge detection which indicates frame complexity. The face detection rate is measured as the percentage of frames in which a face was successfully identified whereas face confidence score represents the confidence assigned by the algorithm, indicating the reliability of detection.

DeepFace was found to perform the best, largely due to its deep CNN architecture trained on large and diverse facial expression datasets, along with its highly reliable face

Table 3 Comparative analysis of performance of various visual based models

Algorithm	Accuracy	Precision	Recall	F1-Score
DeepFace	0.8525 ± 0.0888	0.8874 ± 0.0622	0.8525 ± 0.0888	0.8529 ± 0.0873
FER	0.4025 ± 0.1345	0.5255 ± 0.1899	0.4025 ± 0.1345	0.3644 ± 0.1350

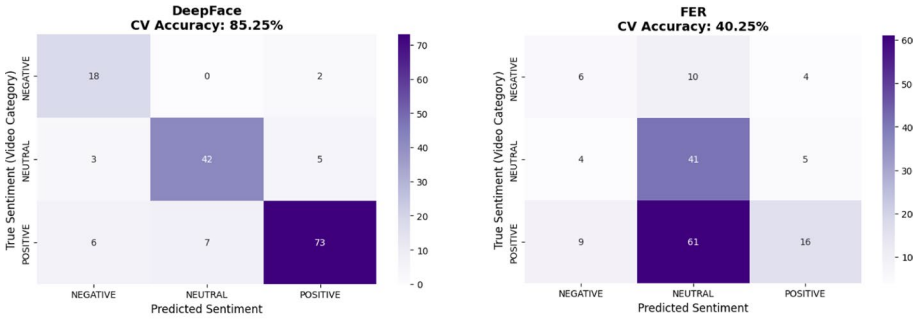


Fig. 8 Confusion matrix of different visual models

detection that consistently achieved close to 100% accuracy across all video segments. It also demonstrated stable, high-confidence predictions even under varying lighting, contrast, and frame-quality conditions. In contrast, the FER algorithm performed significantly worse because of its lower model complexity, which limits its ability to capture subtle expressions, and its reliance on Haar Cascade fallback for face detection, resulting in inconsistent and less accurate detections. FER was also highly sensitive to challenging visual conditions such as profile faces, low-light environments, and motion blur, and it produced lower, unstable confidence scores that reduced the overall reliability of its sentiment predictions. The performance comparison of various visual based models is depicted in Table 3:

FER accuracy drops significantly in segments with low brightness or high sharpness variability, showing its sensitivity to visual noise, while the dataset's high proportion of neutral expressions tends to bias shallow or transformer-based models toward predicting the neutral class. DeepFace, however, consistently identifies subtle facial cues such as calmness, focus, and mild smiles, demonstrating strong feature representation, whereas FER often misclassifies neutral expressions as positive or negative due to its limited ability to detect fine-grained emotional details. Overall, DeepFace achieved the highest accuracy at 85.25%, with stable performance and highly reliable face detection, while FER showed considerable variation across folds and a much lower accuracy of 40.25%. DeepFace also exhibited smaller standard deviations, indicating stronger resilience to dataset fluctuations and class imbalance, and it performed better under varying illumination, sharpness, and face-visibility conditions—all crucial factors in visual sentiment classification shown in Figs. 8, 9 and 10.

4.3 Audio model evaluation

In this section, the audio-based sentiment analysis component is evaluated in detail. Two algorithms were used for this purpose:

- Transformer-based Speech Emotion Recogniser, or Wav2Vec2-XLSR.
- HuBERT Large, which is also a transformer- based speech recognition model.

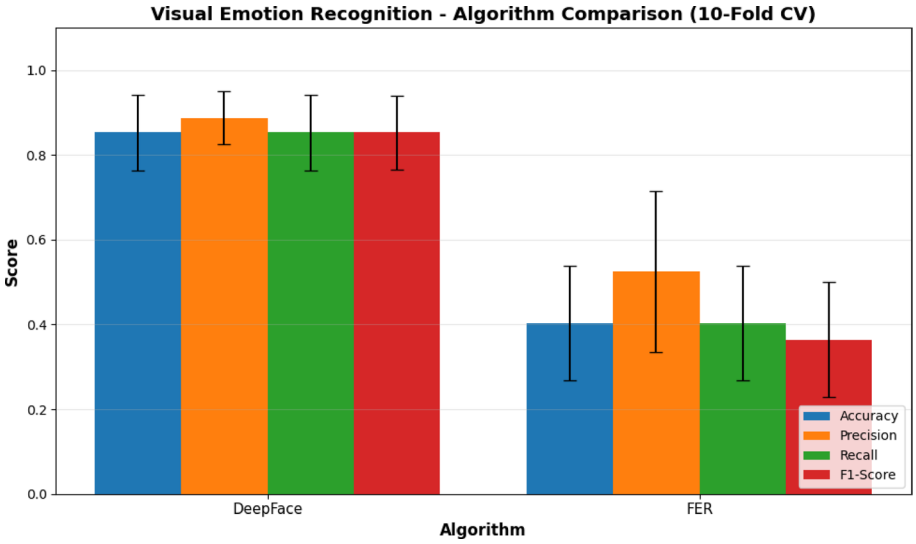


Fig. 9 Model performance comparison (10-fold cross-validation)

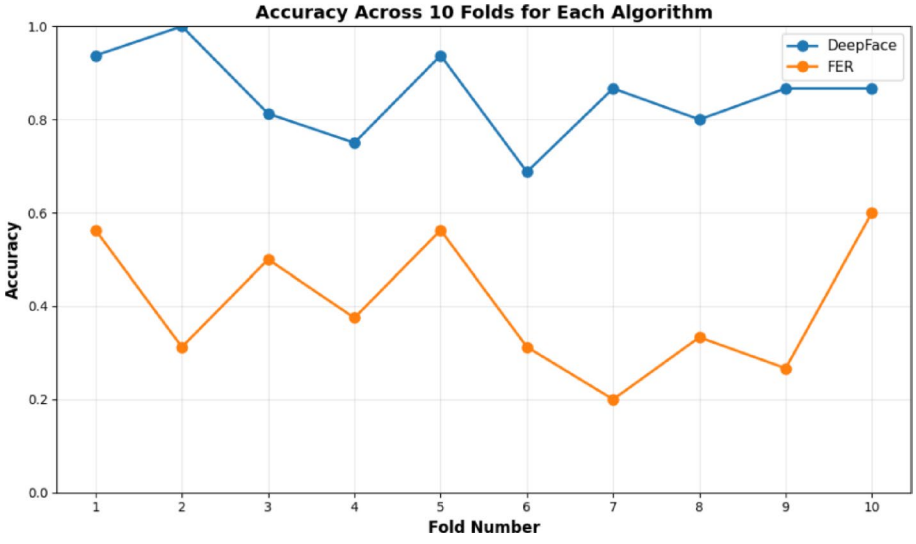


Fig. 10 Accuracy across 10 folds for each model

Forty good audio fragments were collected out of the ten videos. In every segment, Mel spectrograms,2 MFCCs, RMS energy, spectral centroid, spectral rolloff and zero-crossing rate were calculated. The emotional classification of every audio piece was identified on the basis of the genre of the beginning video.

A standard collection of spectral and low-level audio descriptors was applied to the analysis of all audio segments. The MFCCs are utilized to trace the phonological and articulatory deflections of the speech, The Spectral centroid provides information on the brightness of audio, whereas Spectral rolloff is used to measure the distribution of frequency bandwidth. Two algorithm viz. Transformer-based deep model (Wav2Vec2-XLSR) and HuBERT Large has been used and their performance is evaluated using 10-Fold Cross-Validation. The performance analysis of audio-based models is shown in Table 4:

Table 4 Comparative analysis of performance of various audio based models

Algorithm	Accuracy	Precision	Recall	F1-Score
Wav2Vec2-XLSR	0.8400 ± 0.1236	0.8761 ± 0.1278	0.8400 ± 0.1236	0.8490 ± 0.1277
HuBERT Large	0.2733 ± 0.1133	0.3376 ± 0.2773	0.2733 ± 0.1133	0.2031 ± 0.0902

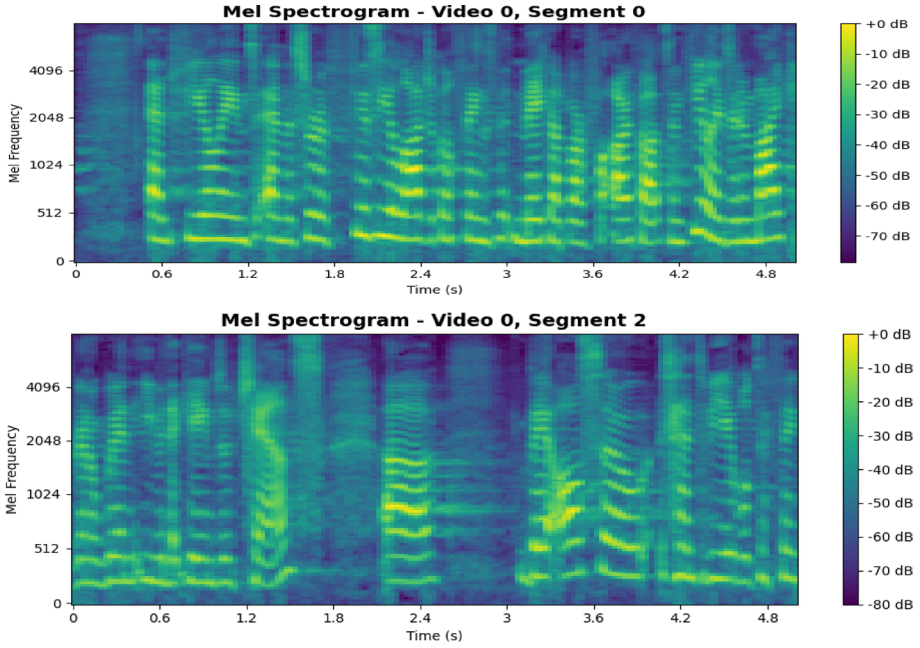


Fig. 11 Mel spectrogram for different video segments

With an accuracy of 84%, the Wav2Vec2-XLSR model outperformed HuBERT Large, which only obtained 27.33% accuracy. With the best accuracy of 84% on Wav2Vec2-XLSR, it can be seen that it was quite consistent and yielded a relatively small number of false positives on this task to classify the segments as positive, negative, or neutral. By contrast, the HuBERT Large model had an accuracy of 27.33% which indicates that it was not consistent in its performance and it had difficulties in picking up emotional cues using features.

Wav2Vec2-XLSR reliably detects positive and negative prosody even in short phrases and under varying microphone conditions, outperforming handmade feature models that struggle with narration, background music, or monotone speech. The sample of mel frequency spectrograms are shown in Fig. 11.

Consequently, Wav2Vec2-XLSR is recommended as the primary audio sentiment model for the multimodal system. Figure 12 shows the confusion matrices of the two audio models. Figures 13 and 14 depicts model performance comparison of 10-cross validation and accuracy of 10 folds respectively.

4.4 Multimodal decision-level fusion evaluation

The proposed Multimodal Sentiment Analysis Framework, which combines the predictions from three separate modalities—Text, Audio, and Visual—into a single supervised classifier, is thoroughly evaluated in this part. The multimodal approach uses complementing cues to increase resilience and classification accuracy in contrast to single-modality systems that only use language, prosody, or facial expressions. In the final stage,

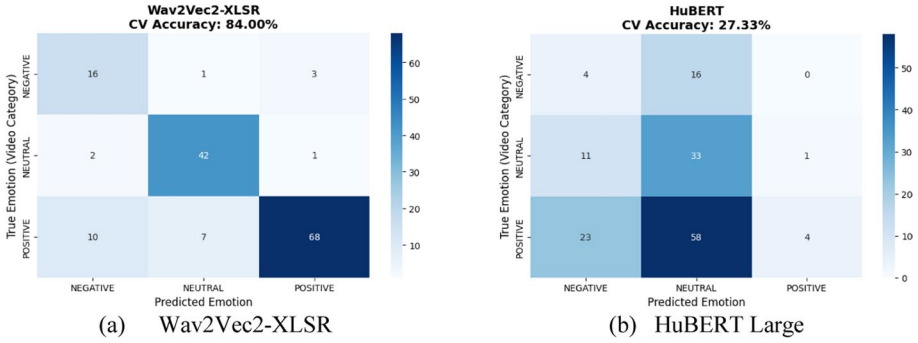


Fig. 12 Confusion matrix of different audio models

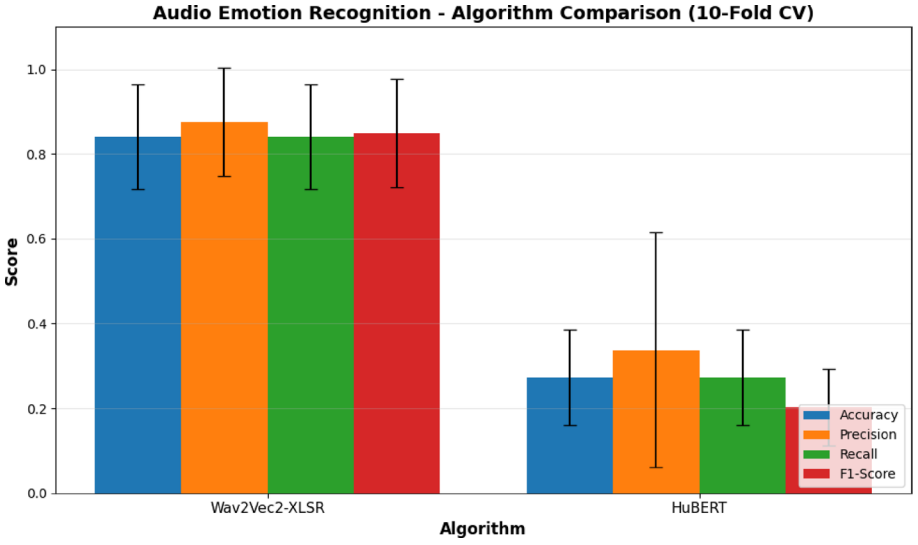


Fig. 13 Model performance comparison (10-fold cross-validation)

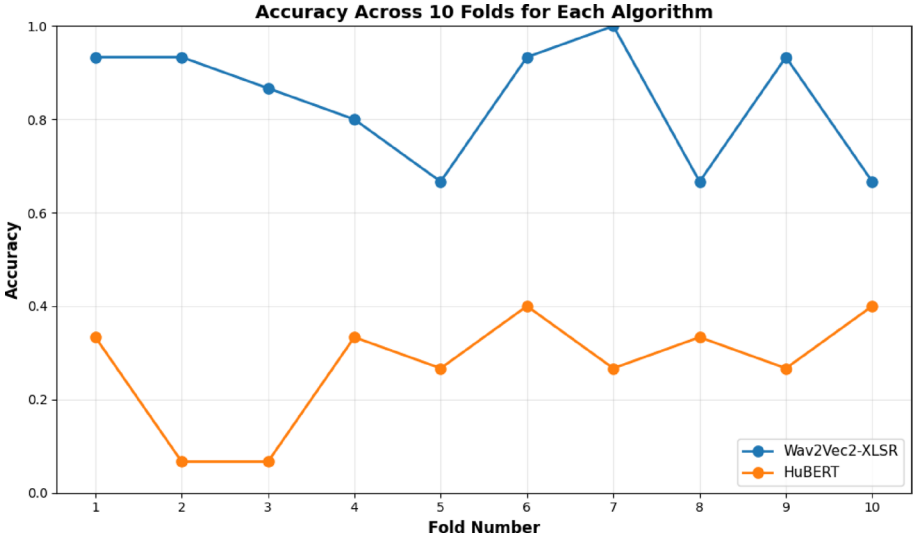


Fig. 14 Accuracy across 10 folds for each model

each modality establishes probabilistic sentiment scores (positive, neutral, and negative) as a result of the last stage, which is a Decision-Level Fusion. Four classifiers were compared for fusion system:

- Extreme Learning Machine (ELM): This is used as a convergence classifier.
- Support Vector Machine (SVM): This is used as a Strong traditional baseline.
- Random Forest: Nonlinear ensemble classifier.
- Equal-Weighting Voting Model: Simple probability averaging baseline.

The comparative analysis of Multimodal Fusion Classifiers shown in Table 5:

ELM classifier exhibited good and stable performance, obtaining good values of all the evaluation measures, as the Accuracy, Precision, Recall, and F1-score fall within the range of about 95–96. These findings suggest that the suggested fusion strategy provided a good separation of class representations so that the multimodal feature space could be efficiently discriminated in the compressed version. The small scale of the fused features was especially very apt to ELM where there was fast acquisition of the decision boundaries and the impractical generalization capabilities. On the same note, the SVM classifier was performing at a competitive level in classification as the test results continuously gave evaluation scores near 95 indicating that the multimodal feature fusion was working effectively to improve class separability. Due to its margin based learning process, SVM was effective to classify the joint probability vectors and exhibited a predictive stability. In spite of the fact that the inference time with SVM was a little bit more than with ELM, SVM was still more capable of participating in such training tasks as it was faster and did not require as many computations as ELM. Random Forest performs well but does not achieve perfect accuracy, which is 98%. The naive equal-weighting method reaches 92% accuracy. This shows that the slight differences between modalities cannot be fully captured by just averaging probabilities. It highlights the advantages of a supervised fusion strategy. Figure 15 shows the confusion matrices of the multimodal fusion classifier. Figures 16 and 17 depicts model performance comparison of 10-cross validation and accuracy of 10 folds respectively.

5 Multimodal sentiment analysis as a foundation for mental-health monitoring

Although the primary objective of this work is sentiment classification in YouTube videos, the extracted multimodal features—textual sentiment polarity, speech prosodic characteristics, and facial expressiveness—closely correspond to behavioural cues widely discussed in psychological and affective computing literature. Previous studies have reported that reduced pitch variability, low vocal energy, negative or neutral linguistic patterns, and diminished facial expressiveness are commonly observed in individuals experiencing prolonged psychological strain. The proposed multimodal framework

Table 5 Comparative analysis of performance of multimodal fusion classifier

Classifier	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Test Time (s)
ELM	0.9524 ± 0.0121	0.9582 ± 0.0108	0.9549 ± 0.0133	0.9565 ± 0.0115	0.0527	0.0001
SVM	0.9478 ± 0.0146	0.9491 ± 0.0139	0.9463 ± 0.0152	0.9477 ± 0.0144	0.0041	0.0004
Random Forest	0.9800 ± 0.0600	0.9900 ± 0.0300	0.9800 ± 0.0600	0.9800 ± 0.0600	0.3534	0.0163
Equal Weighting	0.9200 ± 0.0980	0.9500 ± 0.0671	0.9200 ± 0.0980	0.9227 ± 0.0948	0	0

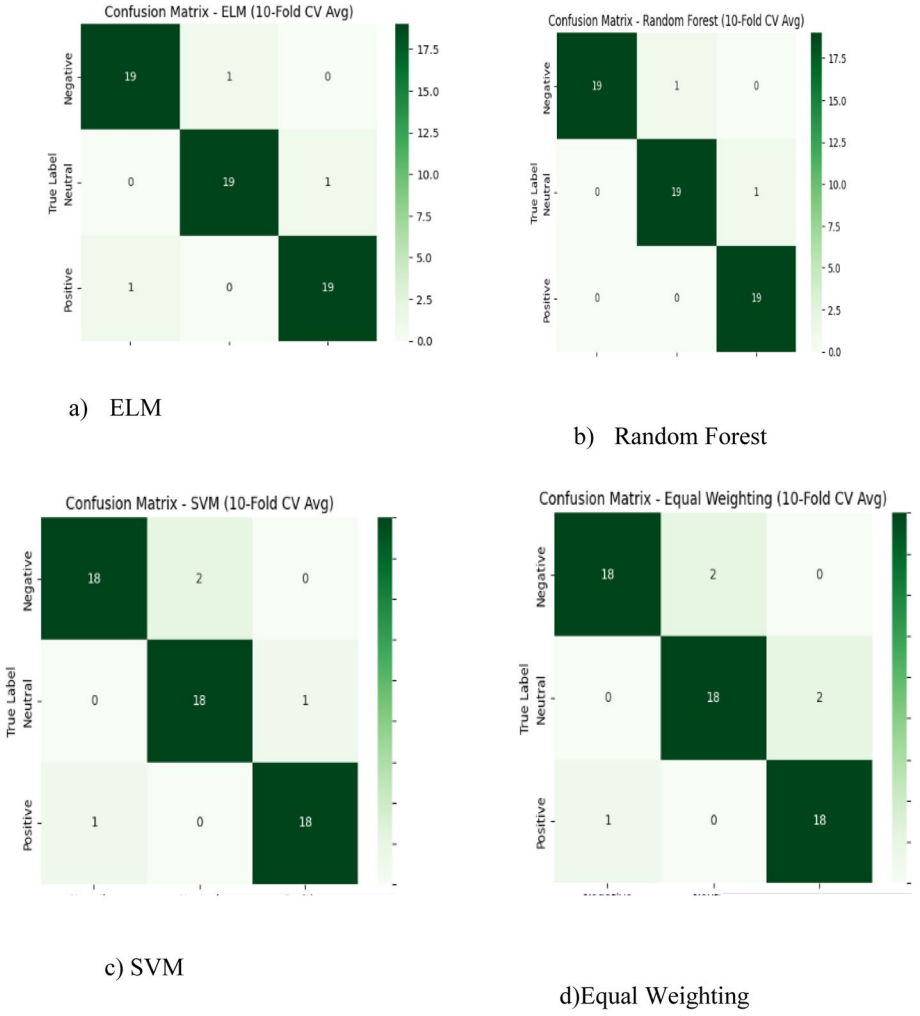


Fig. 15 Confusion matrix of different multimodal fusion models

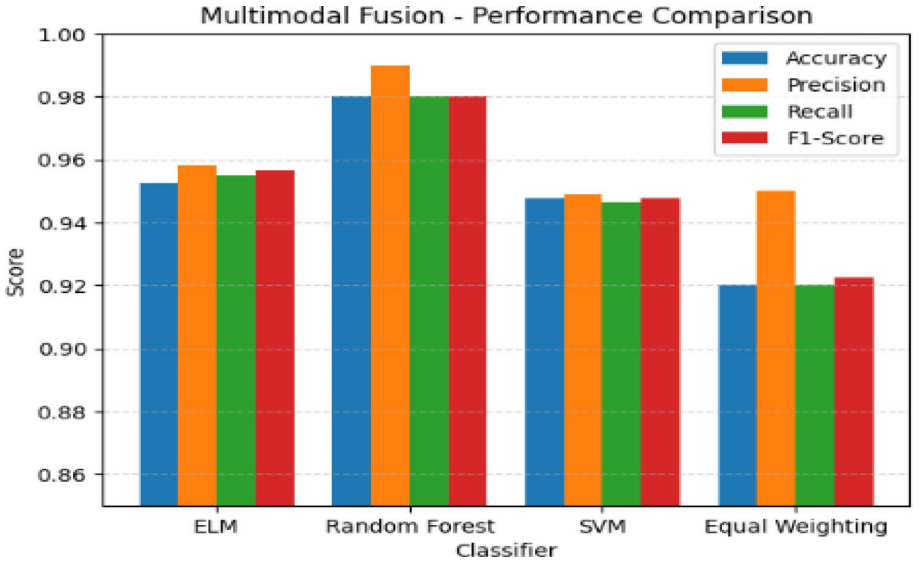


Fig. 16 Multimodal performance comparison

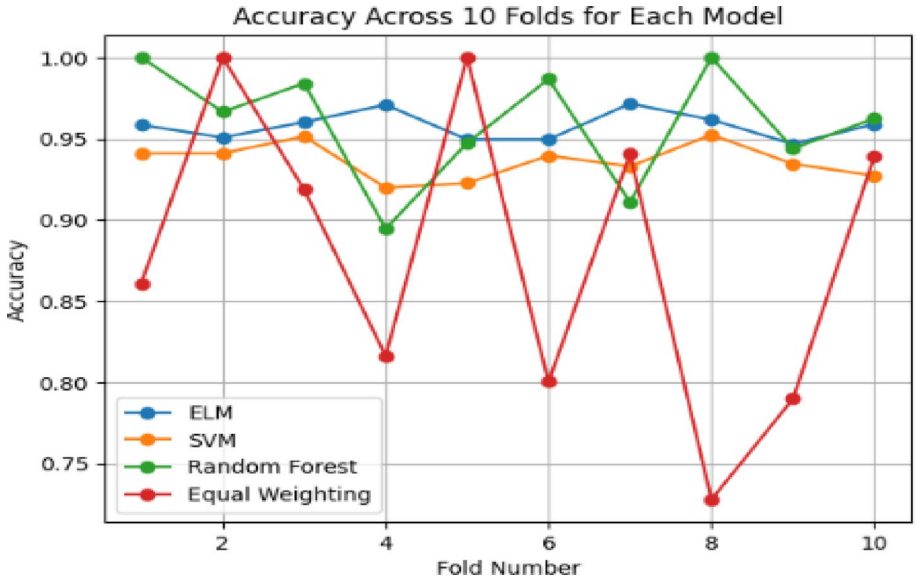


Fig. 17 Accuracy across 10 folds for each model

captures these behavioural attributes through sentiment distributions, acoustic energy measures, and facial emotion probabilities, enabling an observational and analytical interpretation of affective behaviour without reliance on clinical assessment. By aggregating system outputs over time, such as average sentiment polarity, facial expressiveness ratios, and audio-derived emotional indicators, longitudinal trends in user behaviour can be examined and qualitatively related to established behavioural observations reported in prior research. In particular, content exhibiting subdued prosody, reduced articulation energy, and limited facial expressiveness reflects patterns consistently associated with lower affective engagement. These observations suggest that behavioural signals derived from user-generated content, including YouTube videos and other self-expressive media, can provide meaningful insights into mood-related variations through non-intrusive, data-driven analysis.

6 Conclusion

This research demonstrated that a reliable multimodal sentiment system may be developed by integrating text, audio, and visual data with a decision-level fusion approach. DeepFace was discovered to be more efficient in visual sentiment issues since it recognized faces dependably and gathered expressive details excellently. Text analyses were observed to be performed well using lexicon-based models such as TextBlob and VADER and performed even better than some transformer models in a number of instances. In the audio section, it was demonstrated that the Wav2Vec2-XLSR was better at the task of detecting emotional miracles in speech compared to the HuBERT Large. On the condition of combining three modalities together, the Extreme Learning Machine (ELM) based classifier outperformed SVM, Random Forest and equal-weighting baselines, and it got the highest values of accuracy with consistency, that sentiment could not be judged on the basis of one modality. Instead, combining modalities creates a richer and more discriminative feature space, which allows for highly accurate sentiment prediction. Overall, the project produces a comprehensive multimodal sentiment analysis pipeline that is methodologically scalable, computationally efficient, and empirically stable.

Even if the suggested approach works incredibly well on the carefully chosen dataset, there are a number of intriguing directions for further development. Applications in social media analytics, human-robot interaction, smart classrooms, and mental health monitoring are made possible by the model's ability to recognise multimodal sentiment in real-time. Instead of depending only on decision-level fusion, the system may be able to learn richer cross-modal associations by using transformer-based multimodal encoders (such as VideoBERT, CLIP, or multimodal LLaMA models) or attention-based fusion methods. Furthermore, broadening the dataset to include a wider range of emotional expressions, languages, and contextual circumstances might enhance generalisation. By including SHAP-based interpretability for multimodal predictions and implementing the system as a scalable web service or API, future research may also concentrate on explainability.

Author contributions
Conceptualisation, methodology, and validation, P.S., O.C. and D.K.; formal analysis, D.K., J.D.H and S.K.B; resources, J.D.H., O.G. and L.M.D.; data curation, D.K., P. and O.C.; review and editing, D.K., and P.S.; visualisation, D.K, P. and S.K.B.; project administration, J.D.H., O.G. and L.M.D.; funding acquisition, O.G. and L.M.D. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Data availability
[https://github.com/Preeti061204/Multimodal-Sentiment-Analysis-Pipeline-for-YouTube-Videos](https:/github.com/Preeti061204/Multimodal-Sentiment-Analysis-Pipeline-for-YouTube-Videos).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Ethical standards
The authors declare that they do not have any conflict of interest. The research was carried out bearing in mind the ethical and privacy-related issues. The videos accessed to obtain all data in the experiments were all publicly available and accessed using official APIs without circumventing restrictions or privacy settings. No user- authenticated, restricted or private content was gathered or examined. In a further attempt to reduce the privacy risks, a filtering system was implemented which blocked videos with sensitive personal data, blatant personal disclosures or content that might cause privacy issues. The analysis was conducted at the aggregated level of segment and no personally identifiable data were saved, deduced or made with reference to individual identities. Also, the suggested framework is not expected to give clinical, diagnostic, or individual-level measures; instead, it targets high-level sentiment and affective pattern analysis. These will guarantee the use of the data in a responsible manner and will bring the study to the standards of the ethical research when analysing publicly available user-generated content.

Clinical trial number
Not applicable.

Competing interests
All authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript. All authors declare that they have no conflict of interest.

Received: 3 December 2025 / Accepted: 5 February 2026
Published online: 18 February 2026

References

1. Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. Comput Sci Rev. 2018;27:16–32.
2. Chaudhary L, Girdhar N, Sharma D, Andreu-Perez J, Doucet A, Renz M. A review of deep learning models for Twitter sentiment analysis: challenges and opportunities. IEEE Trans Comput Social Syst. 2023;11(3):3550–79.
3. Liu B. Sentiment analysis and subjectivity. In: Handbook of natural language processing 2.2010. 2010. pp. 627–666.
4. Taboada M. Sentiment analysis: an overview from linguistics. Ann Rev Linguist. 2016;2(1):325–47.
5. Sahayak V, Shete V, Pathan A. Sentiment analysis on Twitter data. Int J Innov Res Adv Eng (IJIRAE). 2015;2(1):178–83.

6. Sarlan A, Nadam C, Basri S. Twitter sentiment analysis. In: Proceedings of the 6th International conference on Information TechnologyMultimedia. IEEE; 2014.
7. Albladi A, Islam M, Seals C. Sentiment analysis of Twitter data using NLP models: a comprehensive review. IEEE Access. 2025.
8. Jonnala N, Surekha, et al. Leveraging hybrid model for accurate sentiment analysis of Twitter data. Sci Rep. 2025;15(1):24438.
9. Kafi A, Abdullah SK, Banshal N, Sultana, Gupta V. Source recommendation system using context-based classification: empirical study on multi-level ensemble methods. J Scientometr Res. 2024;13(2):475–84.
10. Poria S et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing. 2016;174: 50–59.
11. Abburi H et al. Multimodal sentiment analysis using deep neural networks. In: International conference on mining intelligence and knowledge exploration. Springer, Cham; 2016.
12. Portes Q et al. Multimodal neural network for sentiment analysis in embedded systems. In: VISIGRAPP (5: VISAPP). 2021.
13. Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing. 2006;70(1–3):489–501.
14. Subramanian M, Veerappampalayam Easwaramoorthy S, Subbarayan N, Moorthy U. Empowering sentiment analysis in social media: a comprehensive approach to enhance the classification of abusive Tamil comments using transformer models. J Big Data. 2025;12(1):208.
15. Principi E, et al. Acoustic template-matching for automatic emergency state detection: an ELM based algorithm. Neurocomputing. 2015;149:426–34.
16. Mansoorizadeh M, Charkari NM. Multimodal information fusion application to human emotion recognition from face and speech. Multimedia Tools Appl. 2010;49(2):277–97.
17. Gunes H. Bi-modal emotion recognition from expressive face and body gestures. J Netw Comput Appl. 2007;30(4):1334–45.
18. Huang G, et al. Trends in extreme learning machines: a review. Neural Netw. 2015;61:32–48.
19. Decherchi S, et al. Circular-ELM for the reduced-reference assessment of perceived image quality. Neurocomputing. 2013;102:78–89.
20. Cambria E, et al. An ELM-based model for affective analogical reasoning. Neurocomputing. 2015;149:443–55.
21. Poria S, et al. Towards an intelligent framework for multimodal affective data analysis. Neural Netw. 2015;63:104–16.
22. Kindra M, Dixit V, Gupta V. A fuzzy-based approach for characterization and identification of sentiments. In: Computational intelligence for information retrieval. CRC; 2021. pp. 219–36.
23. Yagnik M, Hashmi M, Kumar D, Jain K, Grover E, Hemanth JD. A multi-output BERT framework for abusive comment detection and sentiment analysis on Low-Resource Language. ACM Trans Asian Low-Resource Lang Inform Process. 2025;24(11):1–25.
24. Cambria E et al. Sentic blending: scalable multimodal fusion for the continuous interpretation of semantics and sentics. In: 2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI). IEEE; 2013.
25. Smith-Mutegi D, Mamo Y, Kim J, Crompton H, McConnell M. Perceptions of STEM education and artificial intelligence: a Twitter sentiment analysis. Educ Inform Technol. 2025;30:1–18.
26. Tsai Y-HH, Bai S, Yamada M, Morency L-P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th annual meeting of the association for computational linguistics; 2019. pp. 6558–6569.
27. Yang J, Sun Y, Chen J, Li Y. Multimodal sentiment analysis using BERT-based cross-modal fusion. IEEE Access. 2020;8:154478–89.
28. Yu W, Xu H, Wang Z, Shen HT. Vision-language transformer for multimodal sentiment analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2021. pp. 251–260.
29. Mai S, Hu H, Xing S, Zong Y, Zhang B. Multimodal sentiment analysis with Temporal self-attention transformer. IEEE Trans Affect Comput. 2022;13(3):1214–25.
30. Wang Y, Guo D, Li J, Li B. Unified transformer framework for multimodal sentiment analysis. Inform Fus. 2023;89:1–12.

Publisher's note