

Human sentiment analytics using multi-model deep learning approach

Anil Kumar Muthevi¹, Maganti Venkatesh², Pallavi Gaurav Adke³, Rajashree Gadhawe⁴,
G. L. Narasamba Vanguri⁵, Thiruveedula Srinivasulu⁵

¹Department of Computer Science and Engineering, Aditya University, Surampalem, India

²Department of Computer Science and Engineering (AIML), Aditya University, Surampalem, India

³Institute of Artificial Intelligence, Dr. Vishwanath Karad MIT-World Peace University, Pune, India

⁴Department of Computer Engineering, Pillai HOC College of Engineering and Technology, University of Mumbai, Mumbai, India

⁵Department of Information Technology, Aditya University, Surampalem, India

Article Info

Article history:

Received Feb 14, 2024

Revised Apr 16, 2025

Accepted Jun 8, 2025

Keywords:

Deep learning
Emotional artificial intelligence
Feature-level fusion
Human emotions
Machine learning
Natural language processing
Neural networks

ABSTRACT

For assessing human beings, the measurement of willpower and human emotions plays an important role because human beings are emotional creatures. Emotional analysis, also known as sentiment analysis, is the procedure of using natural language processing (NLP) and machine learning to determine the emotions expressed in speech, text, or other ways of communication. However, critical emotional analysis is limited to human interactions only. Human emotional artificial intelligence, or human sentimental analytics, a sub domain of NLP seeks to improve this understanding. The present study develops a model using multi-model deep learning (DL) approach which is capable of efficiently understanding human emotions and their intentions, closely mirroring human cognition. By extending emotional analysis beyond the traditional limits, this model will collect broad ranging data to uncover clear and hidden emotional details. The main intention of this paper is to build highly effective model which provides in-depth insights into human emotions, leading to logical conclusions depending on all available factors and reasons. The necessary input data for the current study will be collected from audio-visual media covering a vast range of audio and visual samples.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Anil Kumar Muthevi

Department of Computer Science and Engineering, Aditya University

Aditya Nagar, ADB Road, Surampalem 533437, Kakinada District, Andhra Pradesh, India

Email: lettertoanil@gmail.com

1. INTRODUCTION

Emotions are truly mind-boggling elements that can change the entire meaning of a human conversation. Multiple types of emotions influence how we live and interact with other people. Sometimes, it appears that emotions are the ones in control of us. Our decisions, behaviors, and perceptions are all driven by the feelings people encounter in everyday life. Psychologists have made efforts to recognize the different kinds of sentiments people go through from the vast spectrum of human experience. Several distinct perspectives have emerged in an attempt to classify and represent the feelings that individuals possess. Paul Eckman proposed six basic emotions and stated that these are the most widely seen across all human cultures. These commonly recognized emotions include fear, surprise, disgust, happiness, sadness, and anger. These 6 emotions are considered the fundamental origin of many other emotions, except for those of neutrality and calmness.

Artificial intelligence (AI) is essentially the imitation of human natural cleverness in machines to perform tasks in a human-like manner. It involves designing a machine capable of thinking, handling everything from basic functions to complex processes, with varying levels of cognitive skill. Advancements in brain science have enabled the shift from artificial narrow intelligence (ANI) to artificial general intelligence (AGI), which allows machines to perform, think, and carry out tasks similarly to humans. Although AGI research remains decades away, basic human evaluations aid in refining the techniques used for replicating the human mind. This specific domain within AI, focusing on the analysis of human communication, is loosely termed as natural language understanding.

Human emotional analytics is derived from a specific component of natural language processing (NLP). A wide array of human emotions needs to be categorized to determine the correct polarity, feeling, or intent behind a statement. NLP emphasizes interpreting text in human language to generate insights that assist in simplifying business decision-making. However, the human emotional spectrum is significantly more intricate. It primarily relies on visual cues, tone of voice, or spoken words. With the growing capabilities of AI and machine learning (ML), there is promising potential to develop a machine capable of identifying a user's emotions. Is there any perfect and complete system for emotion detection? The answer for this question is the objective of this research is to build a highly efficient system that provides deep insights into human emotions, leading to logical conclusions based on all available factors and contextual reasoning. The necessary input data for this analysis will be sourced from audio-visual media, incorporating a diverse range of audio and visual samples.

2. LITERATURE SURVEY

Machine intelligence, which has always been considered a daydream since the early 1900s, aimed to enable computers to comprehend natural data. With the rise of many imaginary stories and movies, it appeared to remain a daydream until the early 1950s. That was the period when the foundations of AI were established. Figure 1 illustrates the different approaches to sentiment analysis. Research persisted, experimenting with diverse methods such as supervised ML and unsupervised ML, among others, in efforts to determine the polarity of a face in an image or to detect traces of polarity within a piece of text.

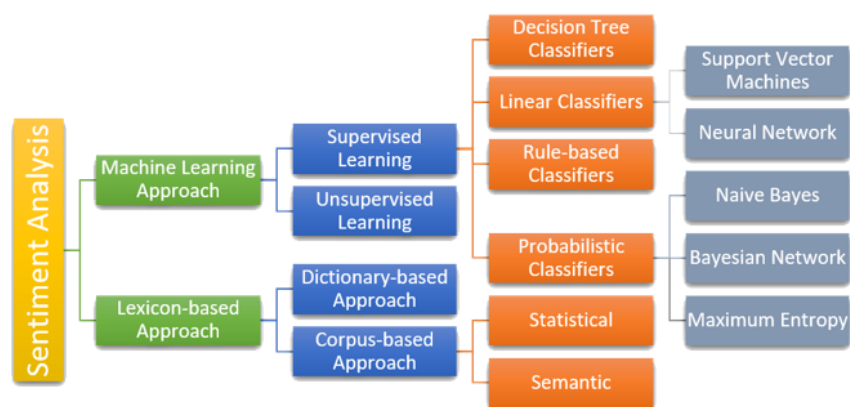


Figure 1. Sentiment analysis and methods

The starter of advanced concepts like artificial neural network (ANN) expanded the scope of emotion detection, enabling machines to work together more effectively with human users. Through the application of feature extraction and deep learning (DL) techniques, machines achieved significantly improved outcomes in analyzing facial expressions and word order polarity. Despite substantial advancements, extracting human-like results from multi-dimensional data remains underexplored. The fusion of visual and auditory content presents the potential to uncover insights that are imperceptible through independent processing. Processing intricate and complex inputs using basic conventional neural networks can be laborious. In such scenarios, deeply-fused networks can play a pivotal role in evaluating the depth of the data within the deep network. While deriving emotions from equivalent expressions is manageable, human emotions are inherently diverse and often defy simplistic classification scales. Mixed emotions and emotional fluctuations are unique to humans, as the most emotionally complex species. Understanding deep

and combined emotions based on their intensity adds another layer of difficulty, and classifying an entity accordingly becomes an even greater challenge.

Over the years, sentiment analysis research has inspired researchers to develop a variety of systems to aid in analysis. Most of these systems are tailored to analyze a single type of content for sentiment classification within their specific domains. Estimation mining involves classifying unstructured data and text into negative, positive, neutral, and categories [1]. Significant advancements have been achieved in the fields of emotion and sentiment analysis through various ML techniques [2], [3]. Traditionally, sentiments are classified into two main categories positive and negative [4], [5]. Numerous ML techniques have been developed for sentiment classification, including stochastic gradient descent (SGD), which enables learning from classifiers based on non-differentiable loss functions, as presented by Bifet and Frank [6]. Another well-known and effective algorithm is naïve Bayes, initially introduced by Thomas Bayes and later elaborated upon by Ciresan *et al.* [7]. Among supervised learning algorithms, support vector machine (SVM) is highly prominent [8]. While many tools and methods are available for sentiment analysis using ML, SVM consistently demonstrates superior accuracy and efficiency compared to other approaches, according to comparative studies in [9]–[11], who thoroughly explored text and audio-visual cues for multimodal emotional analysis. As noted by Zhang *et al.* [12], emotion and sentimentality analysis both pertain to an individual's internal state, and only two notable methodologies for multimodal emotional analysis exist, as proposed by [13], [14]. Prior research in multimodal emotional analysis generally falls into two categories: one focusing on feature extraction from individual modalities, and the other on techniques to fuse features derived from multiple modalities. In 1970, authors in [15] [16] conducted comprehensive studies on facial expressions, concluding that universal facial expressions help in identifying emotions. They recognized anger, surprise, disgust, fear, sadness, and joy as six basic emotional categories. These categories effectively represent most facially expressed emotions. A seventh category, contempt, was later introduced by Ciresan *et al.* [7]. Pak and Paroubek [17] developed the facial action coding system (FACS), which decodes facial language by breaking down expressions into a series of action units (AU).

Recent investigation on speech-based sentiment analysis has focused on recognizing audio features like fundamental frequency (pitch), bandwidth, speech intensity, and duration, as explored by Chen [18]. Speaker-dependent approaches often yield better outcomes compared to speaker-independent ones. This is evident in the impressive results of Navas *et al.* [19], who achieved around 98% accuracy using Gaussian mixture models (GMM) and incorporating prosodic, vocal quality, and mel frequency cepstral coefficients (MFCC) as speech characteristics. However, speaker-dependent methods are impractical for applications involving a large user base. Visual sentiment examination based on text descriptions is effectively described by Ortis *et al.* [20].

Text-based sentiment recognition is a rapidly growing field in NLP, drawing significant attention from both academic and industrial sectors. Traditionally, sentiment and emotion detection in text has relied on rule-based systems, bag-of-words models using expansive emotion or sentiment lexicons, as mentioned by Mishne [21]. Data-driven approaches leveraging large annotated datasets are also used, as described by Muthevi *et al.* [22] and Xia *et al.* [23].

According to Wei *et al.* [24], deep neural networks (DNN) have seen notable improvements in recent years, especially in optimization techniques, activation functions, regularization, pooling, and network design. Multi-column DNN introduced by Ahmad *et al.* [25] explored decision fusion, later expanded to include weighted averaging and adaptive methods based on input conditions by Matsumoto [26]. The current methodology takes a different route by deeply integrating features across multiple intermediate layers, concurrently learning the demonstration of base networks. Wang *et al.* [24] proposed a novel DL method deeply-fused nets centered on deep fusion. Data pre-processing techniques are comprehensively addressed by Ilyas and Chu [27], while Malley *et al.* [28] detail a variety of pre-processing methods. Modern sentiment analysis approaches using DL are described in [29], [30].

3. CONTRIBUTED WORK

The previous approaches in this domain involved utilizing a variety of ML algorithms and logical rule removals on single, exact datasets to extract results. However, they presented several limitations that could not be resolved due to a restricted perspective on data features. These limitations include tone and subjectivity, communication context, polarity inference, sarcasm and irony, limited class labels, enslavement on dataset. To address the aforementioned limitations observed in earlier models, the current work aims to construct a new machine that overcomes certain shortcomings of prior failed methodologies. Consequently, we have adopted newly developed DL methods and neural network modules to capture essential data features that are often hidden or difficult to detect in conventional analysis.

3.1. Databases used

To build a machine capable of detecting various emotional aspects from AV data, we are constrained by the availability of suitable datasets. Several major sources have contributed relevant data, including: i) SAVEE: this dataset contains both audio and video recordings from four male actors using phonetically balanced, generic British English sentences to represent various emotions across multiple repetitions; ii) RAVDESS: this dataset includes 24 participants (12 female and 12 male), who articulate lexically-matched phrases in a neutral North American accent; iii) TESS: this comprising recordings from two female actors one younger and one older this dataset portrays a variety of emotions with neutral emotional intensity using generic statements; iv) YouTube: a global video-sharing platform offering thousands of videos from various categories, contributed by diverse users and organizations across the web; v) FER 2013: this is a visual dataset featuring facial expressions of male and female actors, collected from films and other resources, depicting multiple emotional states; and vi) Google News Vectors: this resource is part of Google's code project, containing a vast dictionary of English vocabulary and terms, intended for classifying textual content into precise groups. These datasets vary in content, covering audio, video, facial expressions, and textual vectors, and provide essential resources for detecting multiple emotional states.

3.2. Data pre-processing

The data obtained from these extensive datasets and manually gathered sources is initially unstructured and mixed in content. Therefore, to ensure usability, it is essential to organize and normalize this data. Datasets undergo pre-processing functions to categorize data by emotional type, gender (e.g., male or female voice), and to reformat them using specific identifiers. This facilitates diversity handling and classification efficiency. The preferred dimensional standards are maintained to be lossless, minimizing information loss and maximizing feature extraction. All collections are transformed into structured formats to ensure emotional attributes are retained distinctly. However, the heterogeneity of the data still poses a challenge, necessitating standardized rule sets for smoother neural network training. In the same way, only consistent data visuals that offer rich feature sets are included, while inconsistent or misleading data units are filtered out using several classifiers to ensure uniform correctness.

3.3. Proposed method

Primary aim of this study is to create a structure accomplished of integrating multiple sentiment analysis modalities into a unified outcome using DL neural networks. Multimodal input data is considered in this process, where each modality is individually analyzed and their outcomes are combined to yield a comprehensive sentiment conclusion. This approach enhances sentiment reliability and uncovers additional data characteristics. The CNN performs on par with human experts across tasks, demonstrating the ability to detect sentiment polarity and classify emotions with a competence level comparable to human judgment. Neural network models differ significantly from traditional ML techniques such as SVM, Naïve Bayes, and linear regression, offering improvements in areas where earlier models faltered.

3.3.1. Data discrimination

The audio and visual data is processed through three separate modules focused on tone, video, and text. Data segregation here refers to isolating each modality rather than combining multiple types. We extract audio tones, visual cues from videos, and text transcribed from audio, assigning corresponding tension-weight levels. This segmentation process is depicted in Figure 2.

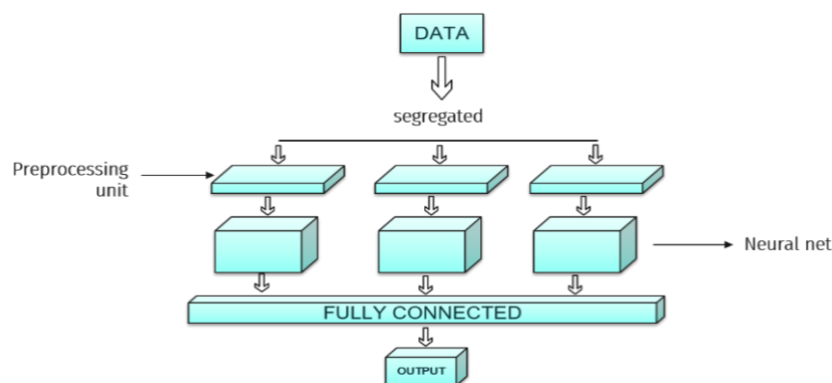


Figure 2. Separated data distributed to individual neural nets

To ensure a structured methodology and adaptability for future research modifications, the algorithmic progression is outlined as follows:

- Step 1: the proposed system extracts multimodal features from videos, each depicting a unique scenario. These features include a sequence of video frames, an audio signal in WAV format, and text obtained through audio transcription.
- Step 2: facial expression and emotion recognition techniques are applied to the video frames to extract visual emotional features.
- Step 3: the audio signal is analyzed to obtain relevant acoustic features, identifying speech instances and extracting components like voicing probability, tonality, and the main frequency of speech variations.
- Step 4: the audio data is further processed to drill-out only the transcribed textual content. The segregated outputs from all three modules are then forwarded to pre-processing units.

These modalities, the audio, visual, and textual modalities (i.e., multiple media sources), are each handled by dedicated pre-processing units tailored to their specific type. These units perform individual operations such as data cleaning and conversion. They are subsequently linked to separate neural networks.

3.3.2. Visual processing

We employ the Haar cascade classifier from Open CV's computer vision modules to detect faces in the visual data. To reduce the dimensional complexity associated with RGB color storage, the extracted frames from input videos are converted to greyscale. Detecting edges and boundaries in colored visuals is notably more complex; hence, greyscale conversion retains intensity levels and enhances classifier performance, while ensuring no bias is introduced, irrespective of the subject's race.

After the transformation, facial key points are detected and the visual is segmented to isolate features specific to the intended individual. The face coordinates obtained are mapped to produce a segmented image, minimizing interference from external elements that could introduce unintended noise during analysis. Upon completion of feature extraction, the segmented visual is transformed into a multidimensional array that preserves pixel-level data, which is subsequently fed into the neural network.

A sequential model is employed, with characteristic values and forms configured to initialize it for visual data processing. The model comprises multiple layers involving pooling and dropout iterations to retain optimal features and eliminate weak connections. Activation functions used in this phase include rectified linear unit (ReLU) and SoftMax. These were selected based on their effectiveness for our specific task. The ReLU is a piecewise linear activation function that returns the input itself when it is positive, and zero when it is not. The resulting vector is an intermediate representation stored for further processing by deeper network layers.

3.3.3. Tonal analysis

Traditional emotion recognition techniques employ NLP to analyze the semantics of words and phrases, then assess sentiment accordingly. However, language is inherently complex, and such conventional analysis often overlooks nuances like regional dialects, tone, pitch and volume. Hence, we propose a system that not only analyzes the content of speech but also its delivery. Audio features are derived from each segmented portion of the videos using a 48 kHz sampling rate and a 100 ms sliding window, allowing for the capture of fine-grained details. To normalize the audio data, Z-standardization is applied, enhancing the visibility of diverse acoustic features for further analysis. The speech waveform is then transformed into a parametric symbolic representation, which reduces the data rate and facilitates efficient downstream processing. The effectiveness of classification relies heavily on the distinctiveness and quality of these extracted features. For this purpose, we utilize MFCC. The formula for calculating the mel frequency for a given input frequency is (1) and MFCCs are computed using the (2).

$$Mel(f) = 2595 \times \log_{10}(1 + f/700) \quad (1)$$

Here, $Mel(f)$ is the frequency in mels and f is the frequency in Hz.

$$\hat{C}_n = \sum k = 1^n (\log \hat{S}_k) \times \cos [n(k - 1/2)\pi/k] \quad (2)$$

Where k is the number of mel cepstrum coefficients, \hat{C}_n is the final MFCC coefficient, and \hat{S}_k is the output of the filter bank.

The MFCC data is compressed into 13 coefficients, representing the frequency spectrum from 20 Hz to 22 kHz. These coefficients correspond to specific frequency regions, with their intensities visualized through varying color depths at mapped coordinate points. The LibROSA library is used to convert stereo audio into mono while maintaining the original sampling rate, ensuring that essential audio characteristics are preserved. The extracted MFCC features are then structured into an n-dimensional array and organized into a

data frame. After completing MFCC segmentation and feature extraction, the dataset is prepared for input into neural network models. Another sequential model is used for tonal information processing. Model values and shapes are appropriately configured. This network undergoes numerous iterations of pooling, convolution, dropout, and flattening to enhance feature recognition in hidden layers and remove inconsistent data links.

Activation functions used again include ReLU and SoftMax, applied across multiple layers to evaluate each node connection. A small learning rate is used as a hyperparameter to ensure the neural network learns gradually, improving its ability to detect tone-dominant features. Optimizers used include root mean square propagation (RMSprop) and Adam, selected through one-vs-one comparison to determine the optimal choice per scenario. The output is an intermediate data vector stored for additional processing. Once the speech segments are identified, the extracted audio is passed through a speech-to-text module to recover the spoken content. To construct a reliable textual analysis model, various grammatical and syntactic rules are applied, including subject noun rule, direct insignificant objects, negation, modifiers (adjectival, adverbial, participial), prepositional phrases, noun compound modifiers. These rules ensure that the resulting model maintains the integrity of the textual information while delivering consistent and meaningful predictions. The model outputs a vector, which forms another intermediate result ready for integration in the final fused neural network.

3.3.4. Synthesis of tri-modal analysis

This module focuses on feature-level fusion, combining information from textual, audio, and visual modalities. Multimodal fusion serves as a core element in any effective emotion detection system, significantly contributing to the improvement of agent–user interaction quality. A primary challenge in this domain lies in devising an effective strategy for integrating cognitive and functional information from diverse sources each characterized by unique temporal scales and data dimensions.

As illustrated in Figure 3, two main fusion techniques are utilized: i) feature-level combination: this approach merges attributes from each modality into a unified joint vector before any classification step is undertaken and ii) decision-level combination: each modality is modeled and categorized separately. The individual results are then combined using established methods, such as expert rules or simple mathematical operations comprising summation, product, majority voting, and statistical weighting. Multi-model analysis fusion can be observed in Figure 4.

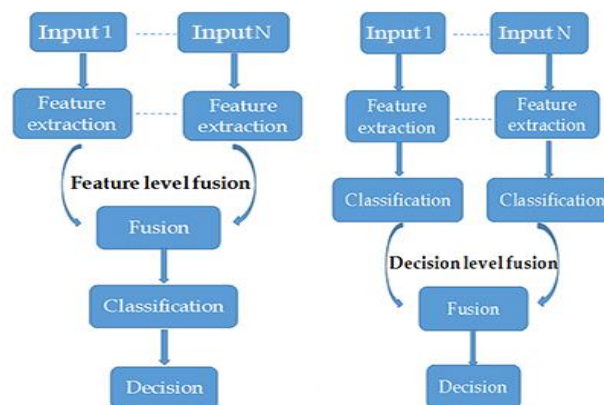


Figure 3. Fusion methods and types

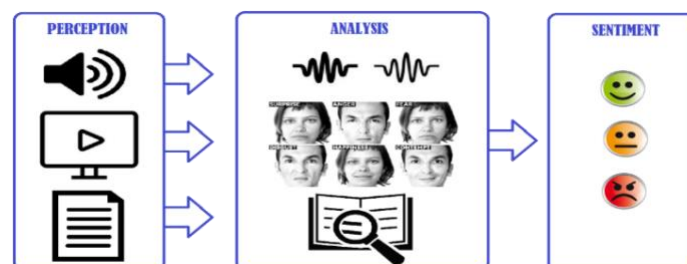


Figure 4. Fusion of multimodal analysis

In the current study, feature-level fusion was applied by combining the feature vectors from all modalities to construct a unified, extended feature vector. Additionally, decision-level fusion was applied across various emotional intents to examine how different analytical approaches perform when processed through the subsequent classifier module to obtain emotional descriptions. Feature vectors from each modality were also merged into a single feature stream in this study. The assumption that a simple fusion generating a resultant vector limited to a basic emotion set would increase machine assurance is a commonly drawn but mistaken conclusion from such analysis. However, the tri-modality approach introduces a broader perspective, enabling the detection of a wide range of emotional intensities expressed by human subjects.

Humans, as inherently complex emotional beings capable of experiencing and expressing mixed emotions, offer a valuable foundation for utilizing this emotional model to explore deeper emotional states. The model's depth depends on the intermediate support and confidence levels derived from the fusion process. Referencing Plutchik's wheel of emotions, as depicted in Figure 5, we can identify diverse emotional spectrums and their influence, facilitating a more detailed emotional recognition process in humans, not merely capturing the emotional sequence but also interpreting the intent behind communicative expressions. Plutchik's model serves as a framework for viewing emotional literacy through a more expansive lens.

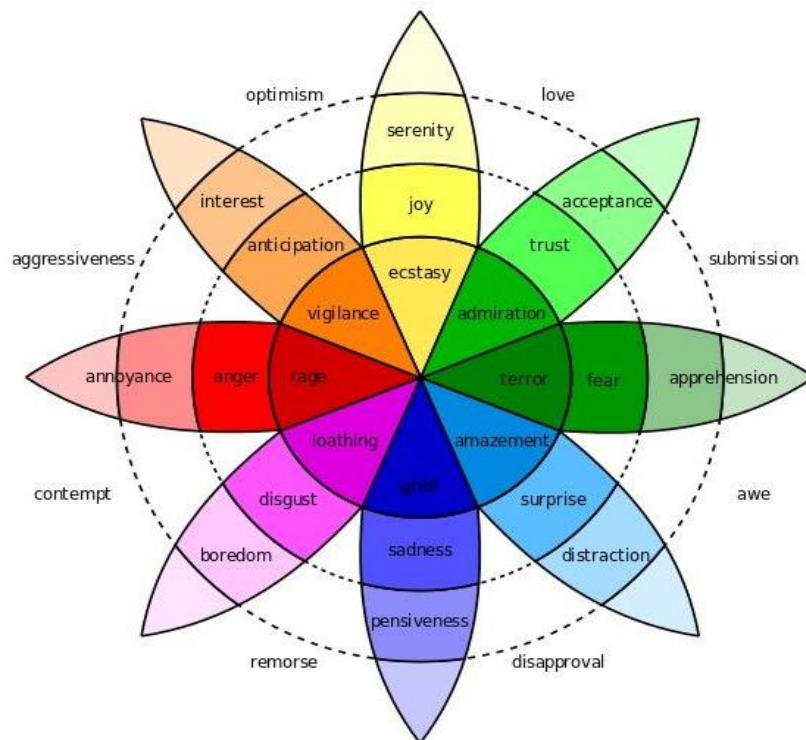


Figure 5. Plutchik's wheel of emotions

Utilizing the same tri-model outcomes, both behavioral irony and communicative irony can be identified to a significant degree. Thus, enhancing emotional literacy involves more than expanding vocabulary for emotions; it encompasses understanding the interrelationships among emotions and recognizing how they evolve over time. Leveraging the tri-model results, this work also demonstrates the potential to detect behavioral and communicative irony with greater precision.

4. RESULTS AND DISCUSSIONS

A Tkinter GUI application has been used to organize the implemented prototype and assess the performance of the deeply-fused neural network under different conditions. The introductory screen of the application allows users to easily locate and load the data intended for analysis, along with a report button to initiate the evaluation process. A screenshot of the report window is shown in Figure 6.

Human Sentiment Analytics

Main Report OGV SGV OGA SGA

Tri-Modal Result: ecstasy

Facial Dominance: happy Facial Recessive:

Tonal Conclusion: happy

Speech Polarity: Neutral

Vocal Range: male Irony Percentage: 0

Detailed Report:

The basic hypothesis from the analysis 'ecstasy' sentiment with a confidence of 100.
 The Visual analysis reports happy emotion dominantly.
 The Tone analysis reported the overall emotion as happy.
 This shows the user is balanced on projecting fair emotional balance in their facial expressions and voice modulation.
 The intent of communication chosen is observed to be Masculine.
 The content delivered seemed to be neutral.
 There are zero traces of Irony.

Save Report

Figure 6. Performance of the deeply-fused neural network detailed report

The machine demonstrated commendable performance when tested with a generic simulated dataset entity. However, this does not imply that real-time outcomes would necessarily yield the same level of accuracy. Therefore, to thoroughly evaluate the system, it was essential to collect complex data capable of producing critical analytical results. As a result, we selected cinemas and TV shows, where actors portray challenging emotional expressions on monitor. The machine successfully detected the anxiety experienced by the girl, who was simultaneously exhibiting signs of fear and sadness while reflecting on life without a lost loved one. It also accurately identified the individual as female, and determined that the word order used in her speech indicated no specific polarity dominance. Figure 7 presents samples of the enhanced visual and audio segmented data employed by the system.

Human Sentiment Analytics

Main Report OGV SGV OGA SGA

Tri-Modal Result: betrayal

Facial Dominance: sad Facial Recessive: angry fearful

Tonal Conclusion: angry

Speech Polarity: Neutral

Vocal Range: female Irony Percentage: 20

Detailed Report:

The basic hypothesis from the analysis 'betrayal' sentiment with a confidence of 39.
 The Visual analysis reports sad emotion dominantly. The suppressed or minute analysis include angry, fearful emotions.
 The Tone analysis reported the overall emotion as angry.
 The intent of communication chosen is observed to be Feminine.
 The content delivered seemed to be neutral.
 There are slight traces of Irony.

Save Report

Figure 7. Samples of the enhanced visual and audio segmented data

This demonstrates how the machine was capable of handling challenging data and providing deeper insights into the overall emotional state of the individual. Advanced elements such as emotional wellness, behavioral irony, and wordplay detection were fine-tuned to extract uncommon and often unnoticed features. In the detailed report, every conclusion derived from the emotions of different segregated inputs was clearly documented. The report provides a precise evaluation of a person's sentiment, successfully delivering a confidence factor for the resulting interpretation. It validates emotional feeling, polarity, tonal information, and intent of emotion, and irony all simultaneously. Table 1 shows types of emotional outputs.

Table 1. Types of emotional outputs

Entry	Description
Tri-modal result	Presents the conclusion derived from analyzing all data modalities, including visual, audio, and spoken speech.
Facial dominance	Displays the predominant emotion expressed on the individual's face in the visual.
Facial recessive	Displays a list of emotions expressed by the individual, excluding the most apparent ones, if any
Tonal conclusion	Provides the emotional inference derived from the speaker's voice and tone.
Speech polarity	Provides the polarity-based conclusion of the speech sentence spoken by the speaker.
Vocal range	Indicates whether the speaker's voice resembles that of a male or female.
Irony percentage	Indicates the percentage of ironic content detected by the system throughout the analysis.
Detailed report	Provides a comprehensive summary of the analysis and the various factors derived from it.

As a result, the model moves beyond traditional emotion classification, embracing emotion detection to more effectively interpret the rich and overlapping emotional patterns typically found in human behavior. Furthermore, the model is capable of assessing whether an individual in the input data is demonstrating emotional balance. Emotional balance is a vital aspect of mental health, and individuals with frequent fluctuations can be accurately detected by the machine.

5. CONCLUSION

The proposed system addresses several limitations encountered by prior singular models. One suggested improvement for achieving more precise results is to segment the media clips at natural pauses, separators, or sequence endings approximately every ~6 seconds and then process each segment in a sequential iteration, which enhances outcome accuracy. While the current model is still in its research phase, it is already capable of detecting a range of 40-50 different emotional states. The implementation of multimodal sentiment analysis to perform multidimensional emotion analysis could be revolutionary. However, this is not the upper limit of its potential. There are numerous applications for emotional state identification in humans. In today's digital era where communication and reviews are shifting from purely textual content to rich media the proposed model can be used to analyze media content, influencing decisions in business, healthcare, and other sectors, driving progress at an entirely new level. Additionally, it can be employed for fraud detection, where individuals attempt to impersonate others, as even minor discrepancies in their emotional states can be detected by the enhanced system. It may also be integrated into the future of AI, contributing to the creation of intelligent personal assistants like the conceptual "Jarvis" tailored to individual users, recognizing their emotional states and interacting in a more human-like manner by understanding the subtleties of conversation. At present, the system's functionality is restricted to the English language, since speech extraction and emotion interpretation are primarily executed in English. Furthermore, irony and wordplay differ significantly across languages and cultures. One of the future goals is to expand the system to support multilingual and multicultural detection. Currently, tonal data is based mainly on British and North American accents. Expanding the dataset to include other accents such as American, Indian, Australian, and German could significantly enhance the model's adaptability and performance across diverse global populations.

ACKNOWLEDGMENTS

The authors thank the Management for their support and resources throughout this research. Special appreciation is extended to the faculty and staff for their guidance and encouragement.

FUNDING INFORMATION

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Anil Kumar Muthevi	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Maganti Venkatesh		✓				✓		✓	✓	✓		✓		
Pallavi Gaurav Adke	✓		✓	✓			✓			✓	✓		✓	✓
Rajashree Gadhawe	✓		✓	✓			✓			✓	✓		✓	✓
G. L. Narasamba					✓		✓			✓		✓		✓
Vanguri														
Thiruveedula Srinivasulu	✓		✓			✓		✓		✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has complied with all relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [MAK], upon reasonable request.




REFERENCES

- [1] F. Agostinelli, M. R. Anderson, and H. Lee, "Robust image denoising with multi-column deep neural networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2013, pp. 1493–1501.
- [2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
- [3] D. I. H. Farias and P. Rosso, "Irony, sarcasm, and sentiment analysis," in *Sentiment Analysis in Social Networks*, Elsevier, 2017, pp. 113–128, doi: 10.1016/B978-0-12-804412-4.00007-3.
- [4] K. N. Devi and V. M. Bhaskarn, "Online forums hotspot prediction based on sentiment analysis," *Journal of Computer Science*, vol. 8, no. 8, pp. 1219–1224, Aug. 2012, doi: 10.3844/jcssp.2012.1219.1224.
- [5] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "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, Apr. 2013, pp. 108–117, doi: 10.1109/CIHLI.2013.6613272.
- [6] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *Discovery Science*, Berlin, Heidelberg: Springer, 2010, pp. 1–15, doi: 10.1007/978-3-642-16184-1_1.
- [7] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2012, pp. 3642–3649, doi: 10.1109/CVPR.2012.6248110.
- [8] J. Khairnar and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," *International Journal of Scientific and Research Publications*, vol. 3, no. 6, 2013.
- [9] A. Kumar and T. M. Sebastian, "Sentiment analysis on Twitter," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 372–378, 2012.
- [10] P. Ekman and W. V. Friesen, "Facial action coding system," *PsycTESTS Dataset*. Consulting Psychologists Press, Jan. 14, 2019, doi: 10.1037/t27734-000.
- [11] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016, doi: 10.1016/j.neucom.2015.01.095.




- [12] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: a survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1253.
- [13] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis," in *Proceedings of the 13th international conference on multimodal interfaces*, New York, USA: ACM, Nov. 2011, pp. 169–176, doi: 10.1145/2070481.2070509.
- [14] P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, Jan. 2012, doi: 10.4401/ag-5364.
- [15] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–717, 1987, doi: 10.1037/0022-3514.53.4.712.
- [16] D. L. S. Jalligampala, R. V. S. Lalitha, M. Anil Kumar, N. Akhila, S. Challapalli, and P. N. S. Lakshmi, "Boosting accuracy of machine learning classifiers for heart disease forecasting," in *Intelligent Data Engineering and Analytics*, Springer, Singapore, 2022, pp. 109–121, doi: 10.1007/978-981-16-6624-7_12.
- [17] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), 2010, pp. 1320–1326.
- [18] L. S. Chen, "Joint processing of audio - visual information for the recognition of emotional expressions in human - computer interaction," *Ph.D. dissertation*, Department of Electrical Engineering, University of Illinois Urbana-Champaign, USA, 2000.
- [19] E. Navas, I. Hernaez, and I. Luengo, "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1117–1127, Jul. 2006, doi: 10.1109/TASL.2006.876121.
- [20] A. Ortis, G. M. Farinella, G. Torrisi, and S. Battiato, "Visual sentiment analysis based on on objective text description of images," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, Sep. 2018, pp. 1–6, doi: 10.1109/CBMI.2018.8516481.
- [21] G. Mishne, "Experiments with mood classification in blog posts," *CiteSeerX*, pp. 1-8, 2005.
- [22] A. K. Muthevi, K. Srikanth, B. R. Reddy, P. Neelima, M. S. Kumar, and D. Ganesh, "A deep learning approach to Alzheimer's diagnosis: a highlighting the potential impact," *Frontiers in Health Informatics*, vol. 13, no. 3, pp. 8573–8584, 2024.
- [23] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: domain adaptation for sentiment classification (extended abstract)," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, Buenos Aires, 2015, pp. 4229–4233.
- [24] J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," *arXiv-Computer Science*, pp. 1-16, May 2016.
- [25] M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad, "Machine learning techniques for sentiment analysis: a review," *International Journal of Multidisciplinary Sciences and Engineering*, vol. 8, no. 3, pp. 27–32, 2017.
- [26] D. Matsumoto, "More evidence for the universality of a contempt expression," *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, Dec. 1992, doi: 10.1007/BF00992972.
- [27] I. F. Ilyas and X. Chu, *Data cleaning*. New York, USA: Association for Computing Machinery, 2019, doi: 10.1145/3310205.
- [28] B. Malley, D. Ramazzotti, and J. T. Wu, "Data pre-processing," in *Secondary Analysis of Electronic Health Records*, Cham: Springer International Publishing, 2016, pp. 115–141, doi: 10.1007/978-3-319-43742-2_12.
- [29] P. C. Shilpa, R. Shereen, S. Jacob, and P. Vinod, "Sentiment analysis using deep learning," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, Feb. 2021, pp. 930–937, doi: 10.1109/ICICV50876.2021.9388382.
- [30] B. S. Panigrahi *et al.*, "Novel nature-inspired optimization approach-based svm for identifying the android malicious data," *Multimedia Tools and Applications*, vol. 83, no. 28, pp. 71579–71597, Feb. 2024, doi: 10.1007/s11042-023-18097-5.

BIOGRAPHIES OF AUTHORS






Anil Kumar Muthevi    currently working as Professor in Department of Computer Science and Engineering at Aditya University, Surampalem, Andhra Pradesh. He completed M.Tech. Computer Science and Engineering at JNTUK, Kakinada, and Ph.D. Computer Science and Engineering in 2020 from Acharya Nagarjuna University. He has total 18 years of experience in teaching and research. He has authored 21 International Journal Publications and Conferences. His major focused research area is computer vision, image processing, machine learning, deep learning, computer networks and cryptography, and cyber security. He can be contacted at email: lettertoanil@gmail.com.






Maganti Venkatesh    currently working as Associate Professor & HoD – AI&ML, Aditya University, Surampalem, India. He earned his B.Tech. degree in Computer Science & Information Technology from Kakinada Institute of Engineering and Technology in 2005, affiliated with JNTUK, Kakinada, Andhra Pradesh. He completed his M.Tech. in Computer Science and Engineering from Sasi Institute of Technology & Engineering in 2011. He was awarded a Ph.D. by Hindustan Institute of Technology & Science, a Deemed-to-be University, Chennai. Currently serving as an Associate Professor at Aditya University, Surampalem, Andhra Pradesh. He has 19 years of teaching experience. He published in Scopus and SCI-indexed Journals, presented at national and international conferences, life-time member of CSI and ISTE. His research interests include educational data mining, artificial intelligence, machine learning, data science, and optimization algorithms. He can be contacted at email: magantivenkatesh16jan1984@gmail.com.






Pallavi Gaurav Adke    holds a Ph.D. in the Biomedical Image Processing from the Faculty of Electronics and Communication Engineering at Vellore Institute of Technology, Vellore. She received her Master of Engineering in Digital System from Pune University and her Bachelor of Engineering in Electronics and Communication Engineering from Shivaji University. Currently, she is working as a Research Assistant Professor at the Institute of Artificial Intelligence, MIT World Peace University, starting in January 2025. Her research interests include image processing, machine learning, deep learning, signal processing, java programming, and cloud computing. She has received a grant of 24.06 lakhs from Science and Engineering Research Board (SERB-DST) for her research project. She has published 2 patents and 5 copyrights. She can be contacted at email: pallavi.adke@mitwpu.edu.in.






Rajashree Gadhav    holds a Ph.D. in Computer Engineering from University of Mumbai, Maharashtra and currently serving as an Associate Professor and Head of Computer Engineering Department at Pillai HOC College of Engineering and Technology, affiliated to University of Mumbai, Maharashtra. She completed her M.E. and B.E. degree in Computer Engineering from University of Mumbai, Maharashtra. She is having total 15 years of teaching experience. She has published more than 15 research papers in international and national journals like IEEE, WOS, and Springer and presented in conferences. She is life-time member of ISTE. Her research interests include image processing, artificial intelligence, machine learning, and optimization algorithms. She can be contacted at email: rgadhav@mes.ac.in.



G. L. Narasamba Vanguri    has an impressive academic and professional background. She has completed M.Tech. in 2013, she possesses a wealth of knowledge. She has an extensive teaching experience spanning 11 years. Currently serving as an Assistant Professor in the Department of Information Technology at Aditya University, she is also pursuing her Ph.D. at Centurion University of Technology and Management. Her research interests lie in the fields of machine learning and deep learning, showcasing her commitment to advancing knowledge in these areas. Additionally, her membership in the ISTE reflects her dedication to professional development and engagement within her field. She can be contacted at email: gayatrijeedigunta05@gmail.com.



Thiruveedula Srinivasulu    has an impressive academic and professional background. He has completed M.Tech., in 2014, he possesses a wealth of knowledge. He has an extensive teaching experience spanning 10 years. Currently serving as an Assistant Professor in the Department of Information Technology at Aditya University, he is also pursuing Ph.D. at JNTUA. His research interests lie in the fields of big data, data analytics, and deep learning, showcasing his commitment to advancing knowledge in these areas. Additionally, his membership in the ISTE reflects his dedication to professional development and engagement within his field. He can be contacted at email: tstrinu531@gmail.com.