iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77563

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Multi-Modal Sentiment Analysis Using Text, Audio, And Facial Expressions for Human Emotion Detection- A Survey

Anshika Saxena[1], Dr Shweta Singh[2]
[1] Research Scholar, [2]Assistant Professor, Sagar Institute of Science & Technology Gandhinagar, Bhopal (M.P.)

Abstract: ***Human emotion recognition has become a significant research focus within artificial intelligence due to its growing importance in human–computer interaction, affective computing, and intelligent decision-support systems. Conventional emotion recognition methods have largely relied on unimodal data sources, such as text, speech, or facial expressions. Although effective in controlled settings, unimodal approaches often provide an incomplete and ambiguous understanding of emotional expression, as human emotions are inherently multimodal. This review paper critically examines a dissertation that proposes a deep learning-based multimodal sentiment analysis framework for human emotion detection by integrating textual, acoustic, and facial expression modalities. The reviewed framework employs a Long Short-Term Memory (LSTM)-based architecture to effectively model temporal and contextual dependencies present in multimodal data. Textual information is encoded using embedded word sequences, audio data captures emotional prosody through acoustic features, and visual inputs represent facial expression patterns. These modality-specific features are fused within a unified deep learning framework to perform binary emotion classification. Experimental evaluation using standard performance metrics, including accuracy, precision, recall, F1-score, confusion matrix analysis, and training–validation curves, demonstrates an overall classification accuracy of 82.22 percent, along with balanced precision and recall values. The review highlights the robustness, methodological soundness, and practical relevance of multimodal sentiment analysis, emphasizing its advantages over unimodal approaches and its contribution to the advancement of affective computing research.***

Keywords: ***Multimodal Sentiment Analysis, Human Emotion Recognition, Affective Computing, Deep Learning, LSTM Networks, Text–Audio–Visual Fusion.***

I. INTRODUCTION

Emotion recognition constitutes a foundational challenge in artificial intelligence, as emotions play a decisive role in shaping human perception, cognition, communication, and decision-making processes. Emotional states influence how individuals interpret information, respond to stimuli, and interact with both humans and intelligent systems. As artificial intelligence technologies increasingly permeate everyday life—through virtual assistants, recommender systems, intelligent tutoring platforms, and decision-support tools—the ability of machines to detect, interpret, and respond appropriately to human emotions has become a critical requirement. Emotion-aware systems have the potential to enhance user experience, improve system adaptability, and enable more natural and empathetic interactions between humans and machines. Consequently, emotion recognition has emerged as a central research area within affective computing and human–computer interaction.Early computational approaches to emotion recognition predominantly relied on unimodal data sources, particularly textual sentiment analysis. Text-based methods sought to infer emotional polarity or affective states by analysing word usage, syntactic patterns, and semantic structures within written or transcribed speech. These approaches demonstrated considerable success in structured domains such as product reviews, opinion mining, and social media analysis, where emotions are often explicitly expressed.

However, language alone rarely conveys the full emotional intent underlying human communication. Emotional meaning is frequently influenced by tone of voice, facial expressions, and contextual cues that are not directly captured in textual data. As a result, unimodal text-based systems often struggle with implicit emotions, sarcasm, irony, and emotionally ambiguous statements, leading to misinterpretation in real-world settings.In parallel, unimodal approaches based on speech and facial expressions were explored to capture non-verbal emotional cues. Speech-based emotion recognition systems analyse acoustic and prosodic features such as pitch, energy, speech rate, and rhythm, which vary significantly across emotional states. Similarly, facial expression recognition systems examine facial muscle movements and visual patterns to identify emotions.

While these approaches provide valuable insights into emotional expression, they are also subject to substantial limitations. Audio-based systems are highly sensitive to background noise, recording conditions, and individual speaking styles, whereas facial expression-based systems can be affected by lighting variations, occlusion, head pose changes, and cultural differences in emotional display. These challenges underscore a fundamental limitation of unimodal emotion recognition: reliance on a single modality fails to capture the multifaceted and context-dependent nature of human emotions.The dissertation reviewed in this paper directly addresses these limitations by adopting a multimodal sentiment analysis paradigm. This paradigm is grounded in the understanding that human emotions are inherently multimodal and are typically expressed through a combination of linguistic content, vocal characteristics, and facial expressions. Humans naturally integrate information from multiple sensory channels when interpreting emotions, and replicating this process computationally is essential for developing robust and human-like emotion recognition systems. By integrating text, audio, and facial expression data within a unified deep learning framework, the reviewed study aims to provide a more comprehensive, reliable, and context-aware approach to human emotion detection.A key methodological contribution of the reviewed dissertation lies in its use of deep learning techniques, particularly Long Short-Term Memory networks, to model temporal and contextual dependencies across modalities. Emotional expression unfolds over time, especially in spoken communication, where changes in tone, emphasis, and facial expressions convey evolving emotional states.

LSTM-based architectures are well suited for capturing such temporal dynamics, enabling the system to retain relevant contextual information while processing sequential multimodal data. Through modality-specific feature representation and effective fusion strategies, the proposed framework seeks to leverage complementary emotional cues while mitigating the weaknesses associated with individual modalities.This review paper reorganizes and synthesizes the contributions of the dissertation into a coherent review-oriented narrative. Rather than presenting the work as an experimental study alone, the review emphasizes the broader conceptual foundations of multimodal emotion recognition, the methodological design choices underlying the proposed framework, and the empirical findings obtained through systematic evaluation. Additionally, the review situates the dissertation within the wider evolution of sentiment analysis and affective computing research, highlighting its relevance in addressing long-standing challenges associated with unimodal approaches. By critically examining the reviewed work, this paper aims to elucidate the significance of multimodal sentiment analysis as a practical and scalable solution for emotion recognition and to underscore its potential impact on the development of emotion-aware intelligent systems across diverse application domains.

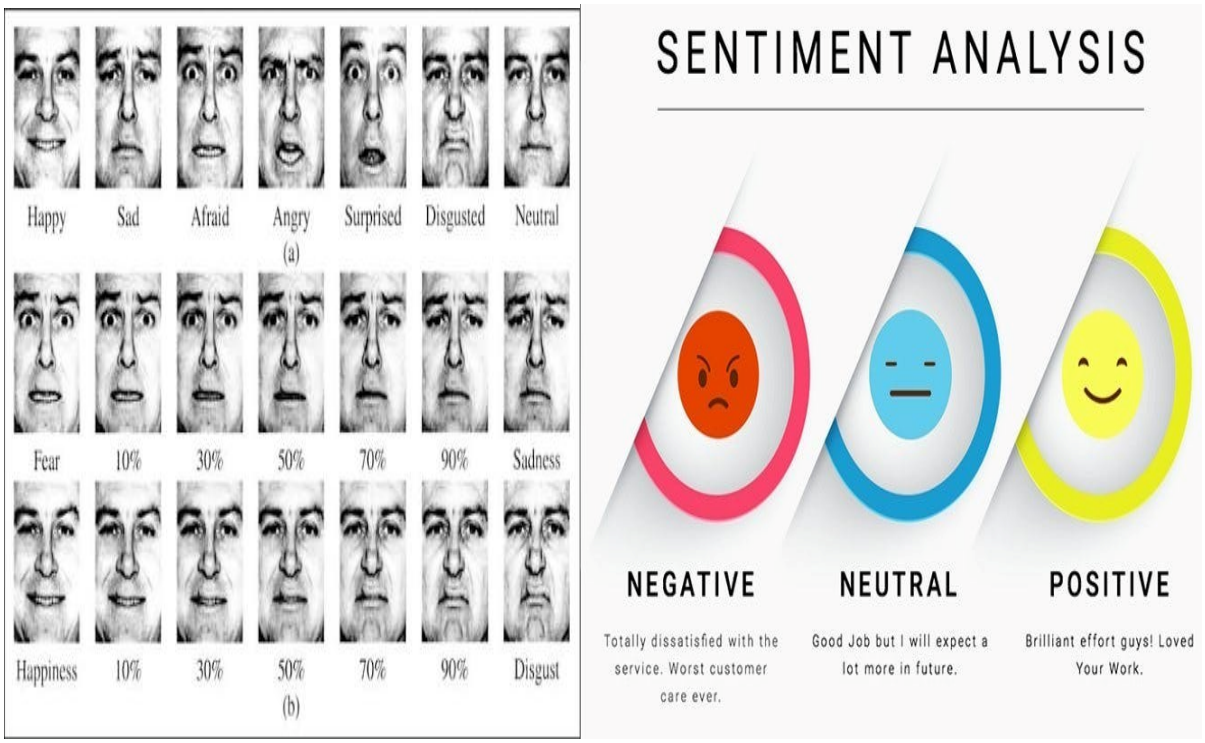


Figure 1.1: Illustration of primary human emotion modalities, including facial expressions, speech-based emotional cues, and textual sentiment representation.

II. EVOLUTION OF SENTIMENT ANALYSIS AND EMOTION RECOGNITION

The evolution of sentiment analysis and emotion recognition reflects a gradual progression from simple rule-based methods to sophisticated deep learning-driven multimodal frameworks. The reviewed dissertation positions its contribution within this broader historical trajectory, highlighting how methodological advancements have been shaped by an increasing understanding of the complexity and multimodal nature of human emotional expression.Early research in sentiment analysis was dominated by lexicon-based approaches, which relied on predefined dictionaries of words annotated with positive, negative, or neutral sentiment values. These methods operated on the assumption that emotional meaning could be inferred directly from the presence or frequency of sentiment-bearing words in a text. Lexicon-based techniques were attractive due to their simplicity, transparency, and low computational requirements, making them suitable for early opinion mining tasks such as product review analysis and social media monitoring. However, despite their interpretability, these approaches exhibited substantial limitations. They were unable to effectively handle contextual ambiguity, linguistic nuances, negation structures, sarcasm, or domain-specific language usage. For instance, the emotional meaning of a sentence often depends on surrounding words or situational context, which lexicon-based methods were ill-equipped to capture.

As a result, these techniques frequently produced inaccurate or overly simplistic sentiment classifications in complex or conversational scenarios.The introduction of machine learning-based approaches marked a significant advancement in sentiment analysis and emotion recognition. Supervised learning algorithms such as Naïve Bayes classifiers, Support Vector Machines, and logistic regression models enabled data-driven sentiment classification by learning statistical patterns from annotated datasets. These methods incorporated feature extraction techniques such as n-grams, term frequency–inverse document frequency representations, and syntactic features to improve classification performance. Compared to lexicon-based approaches, machine learning models demonstrated improved adaptability and accuracy, particularly when trained on domain-specific data. Nevertheless, these approaches remained constrained by their reliance on manual feature engineering, which required significant domain expertise and experimentation. Moreover, traditional machine learning models struggled to capture sequential relationships and long-range contextual dependencies inherent in natural language, limiting their effectiveness in emotion recognition tasks where emotional meaning evolves across sentences or conversational turns.A major paradigm shift occurred with the emergence of deep learning techniques, which fundamentally transformed sentiment analysis and emotion recognition. Deep learning models enabled end-to-end learning from raw or minimally processed data, reducing the dependence on handcrafted features. Recurrent Neural Networks and, in particular, Long Short-Term Memory architectures demonstrated strong capabilities in modeling sequential and temporal dependencies in textual and acoustic data. LSTM networks addressed the vanishing gradient problem and allowed models to retain contextual information over extended sequences, thereby improving the detection of nuanced emotional expressions in language and speech. In parallel, Convolutional Neural Networks achieved remarkable success in facial expression recognition by automatically learning hierarchical spatial features from images and video frames. These advances significantly improved the robustness and accuracy of visual emotion recognition systems.Collectively, the maturation of deep learning techniques across text, audio, and visual domains laid the foundation for multimodal sentiment analysis. By integrating multiple sources of emotional information, multimodal frameworks aim to overcome the inherent limitations of unimodal approaches. The reviewed dissertation builds upon this evolution by adopting a multimodal deep learning architecture that leverages complementary emotional cues from linguistic content, vocal characteristics, and facial expressions. This historical progression underscores the necessity of multimodal sentiment analysis as a natural and essential advancement in the pursuit of accurate, reliable, and human-like emotion recognition systems.

III. LIMITATIONS OF UNIMODAL EMOTION RECOGNITION

A central argument emphasized in the reviewed dissertation is the inherent inadequacy of unimodal emotion recognition systems for accurately interpreting human emotional states in real-world environments. Although unimodal approaches have demonstrated reasonable performance under controlled experimental conditions, their reliance on a single source of information significantly restricts their ability to capture the multifaceted and context-dependent nature of emotional expression. Human emotions are rarely conveyed through one channel alone; instead, they emerge through a complex interaction of language, vocal tone, facial expressions, and situational context. Consequently, systems that analyse emotions using only one modality often produce incomplete or ambiguous interpretations.Text-only emotion recognition models illustrate this limitation clearly. While linguistic content provides valuable semantic and contextual information, it frequently fails to capture emotional intensity, intent, or nuance.

Implicit affect, sarcasm, irony, and figurative language are particularly challenging for text-based systems, as the emotional meaning of a sentence often depends on tone of voice or facial cues that are absent in textual representations. As a result, statements that appear neutral or positive in text may actually convey frustration, dissatisfaction, or anger when expressed verbally or visually.

Audio-based emotion recognition systems address some of these shortcomings by analysing prosodic features such as pitch, energy, and speech rate. However, these systems are highly sensitive to external factors, including background noise, microphone quality, recording environments, and speaker-specific characteristics such as accent and speaking style. Emotional expression through speech also varies widely across individuals and cultures, making it difficult for audio-only models to generalize effectively across diverse populations.Facial expression-based emotion recognition systems face additional challenges related to visual variability. Factors such as lighting conditions, occlusion caused by glasses or facial accessories, head pose variation, and camera quality can significantly degrade recognition performance. Moreover, cultural norms and individual differences influence how emotions are expressed facially, leading to further ambiguity in visual-only analysis.The reviewed dissertation underscores that these modality-specific limitations are unavoidable in real-world settings, where emotional cues are often subtle, incomplete, or even contradictory across modalities. This recognition provides a compelling motivation for adopting a multimodal sentiment analysis framework, which integrates complementary emotional information from multiple channels to achieve more robust and reliable emotion recognition.

IV. CONCEPTUAL FRAMEWORK OF MULTIMODAL SENTIMENT ANALYSIS

Multimodal sentiment analysis is conceptualized in the reviewed dissertation as an advanced approach to emotion recognition that integrates complementary emotional cues derived from multiple communication channels. This framework is founded on the understanding that human emotions are inherently multimodal and are typically expressed through a combination of linguistic, acoustic, and visual signals. Each modality captures a distinct dimension of emotional expression and contributes unique information that cannot be fully represented by any single channel alone. By jointly analysing these modalities, multimodal sentiment analysis aims to achieve a more holistic, accurate, and human-like interpretation of emotional states.Within this framework, textual data serves as a primary source of semantic and contextual information. Word choice, sentence structure, and linguistic patterns often reflect emotional intent, attitude, and polarity. However, textual information alone may fail to convey emotional intensity or implicit affect. To address this limitation, the audio modality provides critical prosodic and temporal cues, such as pitch, energy, speech rate, and rhythm, which are strongly influenced by emotional state.

Variations in these acoustic features over time offer valuable insight into emotional dynamics that may not be evident from textual content. Facial expressions further complement textual and acoustic information by providing immediate and often spontaneous visual indicators of emotion through facial muscle movements, expressions, and micro-expressions.The integration, or fusion, of these modalities lies at the core of the multimodal sentiment analysis framework. By combining semantic, acoustic, and visual cues, multimodal systems are better equipped to resolve ambiguities that arise in unimodal analysis, such as sarcasm or emotionally contradictory signals. Additionally, multimodal fusion enhances robustness by compensating for noise, missing data, or uncertainty in any individual modality, thereby improving reliability in real-world environments.The reviewed dissertation emphasizes that deep learning architectures, particularly Long Short-Term Memory–based models, are especially well suited for multimodal emotion recognition. LSTM networks are capable of modelling temporal dependencies and contextual relationships across heterogeneous data streams, making them effective for capturing the dynamic nature of emotional expression. By leveraging these capabilities, the proposed multimodal framework provides a structured and scalable approach to emotion recognition that addresses key challenges associated with unimodal systems.

V. METHODOLOGICAL DESIGN OF THE REVIEWED FRAMEWORK

A. Dataset and Problem Formulation

The reviewed work employs a synchronized multimodal dataset consisting of textual transcripts, audio speech signals, and facial expression data. The dataset is annotated for binary emotion classification, enabling the study to focus on emotional presence or polarity while reducing classification complexity. A stratified data splitting strategy ensures proportional class representation across training and testing subsets.

B. Data Preprocessing and Feature Representation

Preprocessing is performed independently for each modality while maintaining temporal alignment. Textual data undergoes cleaning, tokenization, and padding.

Audio data is processed to extract acoustic and temporal features associated with emotional prosody. Facial expression data is normalized to reduce variability caused by lighting and pose. Feature representations are modality-specific, with embedded word sequences for text, sequential acoustic descriptors for audio, and spatial visual representations for facial expressions.

C. LSTM-Based Model Architecture

The proposed architecture centers on stacked LSTM layers designed to capture temporal emotional dynamics. An embedding layer encodes textual input, followed by LSTM layers for sequential modeling. Fully connected dense layers perform feature abstraction, with dropout applied for regularization. The output layer uses a sigmoid activation function to perform binary emotion classification.

VI. EXPERIMENTAL EVALUATION AND RESULTS

The reviewed dissertation conducts a comprehensive experimental evaluation to assess the effectiveness and reliability of the proposed multimodal sentiment analysis framework for human emotion detection. Model performance is evaluated using a combination of standard and widely accepted classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide complementary perspectives on classification behaviour, enabling a balanced assessment of overall correctness, reliability of predictions, and sensitivity to emotional instances. In addition, confusion matrix analysis and training–validation performance curves are employed to gain deeper insights into class-wise prediction patterns and learning dynamics.The experimental results demonstrate that the proposed multimodal model achieves an overall classification accuracy of 82.22 percent, indicating that a substantial majority of emotional instances are correctly classified. Precision, recall, and F1-score values of approximately 0.82, 0.81, and 0.82, respectively, further confirm the balanced nature of the model's performance. This balance is particularly important in emotion recognition tasks, where skewed predictions may lead to systematic misinterpretation of emotional states.

Confusion matrix analysis reveals that most misclassifications occur near emotional boundaries, where textual sentiment, vocal tone, and facial expressions convey subtle, weak, or conflicting emotional cues. Such cases reflect the inherent ambiguity of emotional expression rather than fundamental shortcomings of the model. Notably, the number of false positives remains relatively low, indicating that the model adopts a conservative and reliable approach when predicting emotional states, which is desirable in practical applications.The analysis of training and validation performance curves further demonstrates stable learning behaviour. Both accuracy and loss curves show consistent convergence without significant divergence, suggesting minimal overfitting and effective generalization to unseen data. Collectively, these results validate the robustness and practical viability of the proposed multimodal emotion recognition framework.

VII. DISCUSSION AND SIGNIFICANCE

The findings reviewed in this paper provide strong evidence that multimodal sentiment analysis offers substantial advantages over traditional unimodal emotion recognition approaches. By integrating textual, acoustic, and facial expression modalities, the reviewed framework achieves a more comprehensive and reliable understanding of human emotional expression. The balanced precision and recall values reported in the experimental results indicate consistent and dependable classification behaviour, which is particularly important in emotion-aware applications where both false positives and false negatives can lead to inappropriate system responses or reduced user trust.A key aspect of the reviewed work is the effective use of a Long Short-Term Memory–based deep learning architecture to capture temporal dependencies in emotional expression. Emotions often evolve over time, especially in spoken communication, and the ability to model such temporal dynamics is essential for accurate emotion recognition. The LSTM-based design successfully captures these sequential patterns while avoiding excessive architectural complexity. This balance between representational power and computational efficiency enhances the practical deployability of the framework, making it suitable for real-time or near-real-time applications.

From an academic perspective, the reviewed dissertation makes a meaningful contribution to the field of affective computing by demonstrating the practical effectiveness of tri-modal integration using deep learning techniques. It reinforces the theoretical premise that emotions are inherently multimodal and that their reliable recognition requires the fusion of complementary information sources. The study also provides empirical validation for the use of LSTM-based models in multimodal emotion recognition, offering a structured approach that can be extended or adapted in future research.From an application standpoint, the significance of the reviewed framework extends across multiple domains. In human–computer interaction, accurate emotion recognition can enable more natural and empathetic interactions between users and intelligent systems.

In mental health monitoring and intelligent tutoring systems, emotion-aware capabilities can support personalized interventions and adaptive learning experiences. Similarly, emotion-aware virtual agents can benefit from improved emotional understanding, leading to more responsive and human-centric system behaviour.

Table 1: Overview of the Reviewed Multimodal Emotion Recognition Framework

Item	Summary
Modalities	Text, Audio, Facial Expressions
Model	LSTM-based deep learning architecture
Task	Binary emotion classification
Fusion	Unified multimodal feature integration
Accuracy	82.22%
Precision / Recall / F1	~0.82 / ~0.81 / ~0.82
Key Advantage	Robust multimodal emotion detection
Limitations	Dataset diversity, binary labels

VIII. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

While the reviewed dissertation demonstrates the effectiveness of multimodal sentiment analysis for human emotion detection, it also acknowledges several limitations that provide important directions for future research. One primary limitation relates to dataset diversity and annotation subjectivity. Emotion recognition datasets often reflect limited demographic, cultural, or contextual variation, which may affect the generalizability of trained models when applied to real-world environments. Additionally, emotional labels are inherently subjective, and differences in annotator perception can introduce noise and inconsistency into the training data, influencing model performance.Another limitation arises from the use of binary emotion classification, which simplifies emotional representation but may obscure finer-grained emotional distinctions. Although binary classification is suitable for certain applications, it does not capture the full spectrum of human emotional states. Future research could extend the framework to multi-class or dimensional emotion recognition, enabling more detailed and expressive emotional analysis. Incorporating advanced mechanisms such as attention layers may further improve performance by allowing the model to focus on the most emotionally salient features within and across modalities. Exploring transformer-based architectures could also enhance the modelling of long-range dependencies and cross-modal interactions.From a practical perspective, optimizing the framework for real-time deployment and resource-constrained environments remains an important research direction. Finally, ethical considerations—including data privacy, informed consent, transparency, and responsible use—must remain central to the development and deployment of emotion recognition systems, particularly in sensitive application domains.

IX. CONCLUSION

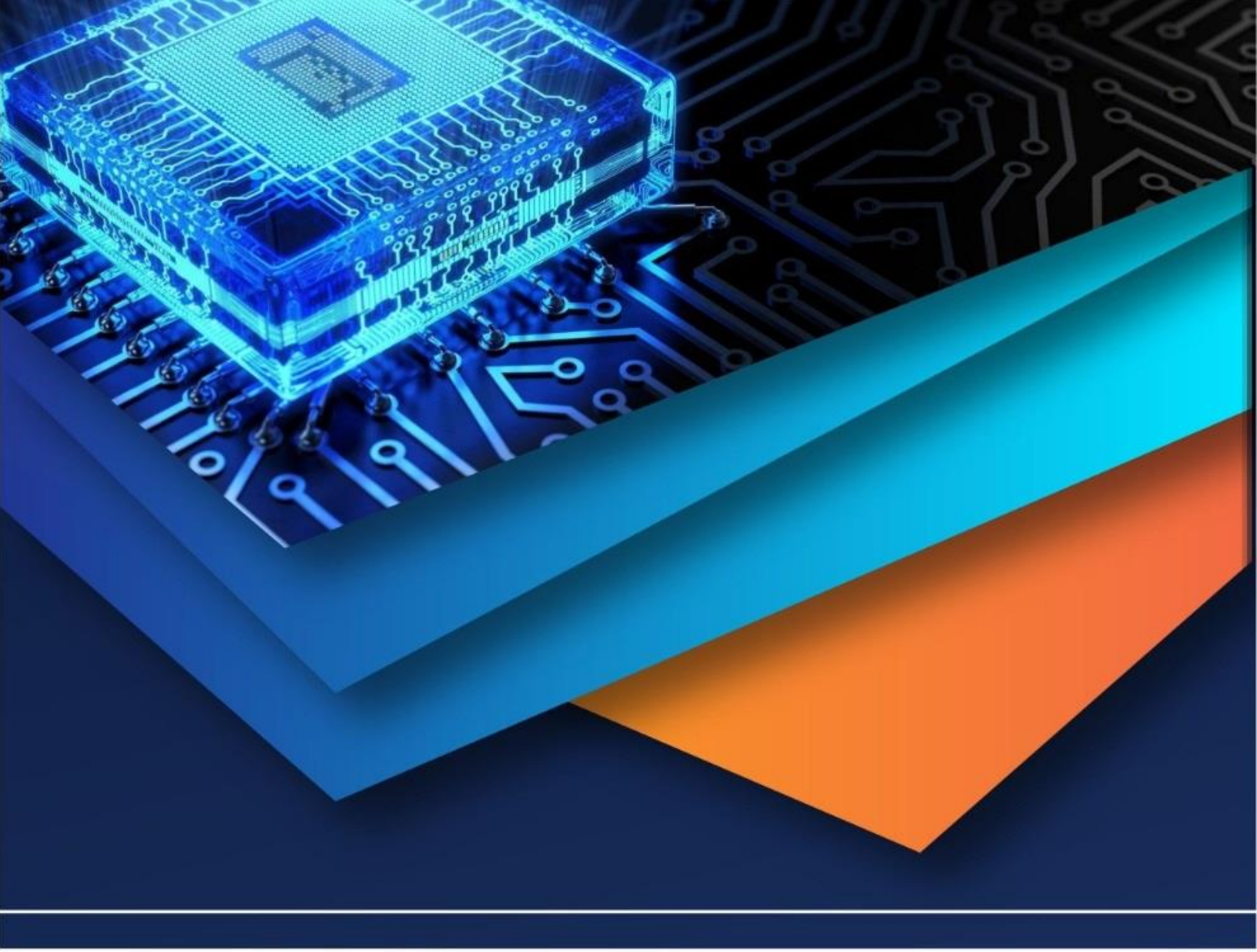
This review paper has synthesized and critically examined a dissertation that proposes a deep learning-based multimodal sentiment analysis framework for human emotion detection. The reviewed work is grounded in the recognition that human emotions are inherently complex, dynamic, and multimodal, and therefore cannot be reliably captured through single-channel analysis alone. By integrating textual, acoustic, and facial expression modalities within a unified Long Short-Term Memory–based deep learning architecture, the dissertation addresses fundamental limitations associated with traditional unimodal emotion recognition approaches. The review highlights how the combined use of multiple modalities enables a more comprehensive and human-like interpretation of emotional states by leveraging complementary emotional cues that are otherwise overlooked in unimodal systems.

A key contribution emphasized in this review is the effective use of LSTM networks to model temporal and contextual dependencies across multimodal data.

Emotional expression often evolves over time, particularly in spoken communication, and the ability to capture such temporal dynamics is essential for accurate emotion detection. The reviewed framework demonstrates stable learning behaviour, as evidenced by consistent training and validation performance, indicating strong generalization capability and limited overfitting. The reported overall classification accuracy of 82.22 percent, along with balanced precision, recall, and F1-score values, underscores the robustness and practical viability of the proposed approach in handling emotionally ambiguous and noisy real-world data.The review further illustrates that the observed misclassifications primarily occur near emotional boundaries, where textual, vocal, and facial cues may convey mixed or subtle emotional signals. Such cases reflect the inherent subjectivity of emotion perception rather than deficiencies in the model itself. Importantly, the multimodal integration strategy adopted in the reviewed dissertation helps mitigate these ambiguities by compensating for weaknesses in individual modalities, thereby enhancing classification reliability.Overall, the reviewed dissertation represents a meaningful and timely contribution to the field of affective computing and multimodal emotion recognition. It provides a structured and scalable framework that balances accuracy, computational efficiency, and methodological clarity. By demonstrating the tangible benefits of multimodal sentiment analysis over unimodal approaches, the work establishes a strong foundation for future research aimed at developing more adaptive, context-aware, and human-centric emotion recognition systems.

REFERENCES

[1] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
[2] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157–166. https://doi.org/10.1109/72.279181
[3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
[4] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15–21. https://doi.org/10.1109/MIS.2013.30
[5] Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2018). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
[7] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068
[8] Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Springer.
[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[10] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538–541.
[11] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.
[12] Morency, L.-P., Mihalcea, R., & Doshi, P. (2011). Toward multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th International Conference on Multimodal Interfaces, 169–176. https://doi.org/10.1145/2070481.2070516
[13] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011
[14] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
[15] Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2018). Multimodal sentiment analysis: Addressing key issues and setting up baselines. IEEE Intelligent Systems, 33(6), 17–25. https://doi.org/10.1109/MIS.2018.023001545
[16] Schuller, B., Steidl, S., Batliner, A., et al. (2018). A survey on automatic speech emotion recognition: Features, classification, and data sets. Speech Communication, 66, 79–97. https://doi.org/10.1016/j.specom.2014.12.004
[17] Socher, R., Perelygin, A., Wu, J., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1631–1642.
[18] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301–1311. https://doi.org/10.1109/JSTSP.2017.2764438
[19] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52
[20] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. https://doi.org/10.1002/widm.1253



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY