A Multimodal Sentiment and Emotion Classification with Deep Learning Techniques

Afzal Ahmad Azmi
MTech Scholar
CSE, Department
NIIST Bhopal

Prof. Anurag Shrivastava
Asso. Professor & HOD, CSE
Department
NIIST Bhopal

Abstract: This research proposes a multimodal deep learning-based emotion recognition framework utilizing the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The model effectively combines audio and visual, modalities to enhance the accuracy and robustness of emotion classification. Audio features are extracted as 74-dimensional acoustic representations, highlighting pitch, tone, and rhythm variations, while visual features are modeled as 35-dimensional facial representations capturing gestures and micro-expressions. Each modality is processed through independent encoders with ReLU activation, projecting them into a 128-dimensional latent space for uniform feature alignment. These embeddings are concatenated into a 384-dimensional fused feature vector, integrating linguistic, acoustic, and facial cues. The fused representation is refined through fully connected layers with dropout regularization to enhance generalization and mitigate overfitting. A Softmax classifier predicts emotional states such as happy, sad, angry, neutral, and surprised. Experimental results on the RAVDESS dataset demonstrate superior performance in terms of accuracy, precision, recall, and F1-score, compared to unimodal baselines. The findings confirm that the proposed multimodal fusion approach effectively captures complex emotional cues, achieving a more comprehensive understanding of human emotions.

Keywords—Sentiment Analysis, Deep Learning, Natural Language Processing (NLP), BERT, Text Classification

I. INTRODUCTION

Sentiment analysis has evolved from rule-based text analysis to advanced deep learning frameworks that can interpret complex emotional cues in text, images, and audio. In real-world scenarios, emotions are expressed not only in words but also in tone, facial expression, and body language. Recent studies show multimodal systems yield richer emotional inference than text-only approaches.

Sentiment analysis has traditionally focused on text data, leveraging natural language processing (NLP) techniques to uncover the emotional tone behind words. Early approaches employed rule-based systems and shallow machine learning classifiers such as support vector machines and logistic regression, which relied heavily on handcrafted features and lexicons. While effective in constrained domains, these methods exhibited limited ability to generalize across diverse contexts, struggled with negation and sarcasm, and ignored the rich nonverbal cues present in human communicationWith the advent of deep learning, the field experienced a paradigm shift: neural networks began to learn hierarchical feature representations directly from raw data, substantially improving performance on text-based sentiment benchmarks. Recurrent architectures (e.g., BiLSTM) captured sequential dependencies, and transformer-based models (e.g., BERT, RoBERTa) set new state-of-the-art results by contextualizing each word within its sentence. However, human sentiment is inherently multimodal, encompassing not only what is said, but also how it is said—through tone of voice, facial expressions, and gestures. Ignoring visual and auditory channels limits the depth and robustness of sentiment inference.This survey addresses the growing need for integrated sentiment analysis by systematically reviewing deep learning approaches across text, visual, and audio modalities. We begin by examining modality-specific encoders: transformer and recurrent networks for textual sentiment, convolutional and vision transformer models for image-based emotion recognition, and convolutional–recurrent frameworks for processing spectrograms and acoustic features in audio signals. We then explore fusion strategies—early concatenation, late decision-level fusion, and hybrid architectures augmented with attention mechanisms—that reconcile disparate embeddings into a unified representation. Particular emphasis is placed on multimodal transformers, which dynamically attend to and align cross-modal signals.

To ground our discussion, we survey prominent benchmark datasets (e.g., CMU-MOSI, CMU-MOSEI, IEMOCAP, MELD), comparing model performance under standard metrics such as accuracy, F1-score, and concordance correlation coefficient. We identify recurring challenges—including data imbalance, modality synchronization, domain adaptation, and interpretability—and highlight emerging solutions like self-supervised pretraining, adversarial augmentation, and lightweight architectures for on-device deployment.By synthesizing recent advancements and pinpointing open problems, this review offers a cohesive roadmap for researchers and practitioners. Our goal is twofold: to demonstrate how multimodal deep learning enhances sentiment analysis beyond text-only methods, and to illuminate avenues for future innovation that will yield more accurate, efficient, and explainable systems. As social media, customer feedback platforms, and human–computer interfaces continue to evolve, robust multimodal sentiment analysis will be crucial for applications ranging from market research to real-time affective computing. This survey thus lays the groundwork for the next generation of sentiment intelligence.

II. LITRETURE REVIEW

The literature survey explores key developments in sentiment analysis, focusing on traditional machine learning approaches and their limitations. The survey also examines the emergence of transformer-based models like BERT, emphasizing their revolutionary role in achieving state-of-the-art performance in sentiment classification.Table 2 shows relevant researchin sentiment analysis field.

Author's introduced the innovative CRDC (Capsule with Deep CNN and Bi structured RNN) model, which demonstrated superior performance compared to other methods. Author's [1] proposed approach achieved remarkable accuracy across different databases: IMDB (88.15%), Toxic (98.28%), Crowd Flower (92.34%), and ER (95.48%). This article provides a comprehensive discussion of data preprocessing, performance metrics, and text embedding techniques, details the implementation architectures of various deep learning models, and critically examines their drawbacks, challenges, limitations, and prospects for future work.

Author's experiment examines the positive and negative of online movie textual reviews. Four datasets were used to evaluate the model. When tested on the IMDB, MR (2002), MRC (2004), and MR (2005) datasets, the (PEW-MCAB) algorithm attained accuracy rates of 90.3%, 84.1%, 85.9%, and 87.1%. PEW-MCAB model outperformed the majority of baseline approaches. This study emphasizes improving global text representations by leveraging word order information. Future work will explore the integration of positional embeddings with other encoding schemes and investigate regularization techniques to optimize these embeddings, ultimately enhancing sentiment classification accuracy [2]

Hybrid sentiment-analysis models combining LSTM, CNN, and SVM were developed and tested on eight tweet and review datasets across diverse domains. Compared to standalone SVM, LSTM, and CNN classifiers—evaluated for both accuracy and computation time—the hybrid approaches consistently outperformed single models [3]. In particular, integrating deep networks with SVM yielded the greatest gains in accuracy and reliability across all datasets. In future various other combinations of models are tested on reliability and computation time. Reliability of the latter was significantly higher.

The work systematically introduces each task, delineates key architectures from Recurrent Neural Networks (RNNs) to Transformer-based models like BERT, and evaluates their performance, challenges, and computational demands. The adaptability of ensemble techniques is emphasized, highlighting their capacity to enhance various NLP applications. Challenges in implementation, including computational overhead, over-fitting, and model interpretation complexities, are addressed, alongside the trade-off between interpretability and performance [4]. In future the synergistic alliance between ensemble methods and deep learning models inthe realm of NLP epitomizes the scientific community's unwavering endeavor to continually redefinethe boundaries of linguistic understanding and computational capabilities.

In this work the rating of movie in twitter is taken to review a movie by using opinion mining.Author proposed hybrid methodsusing SVM and PSO to classify the user opinions as positive, negative for the movie review dataset which could be used for better decisions [5]. The work done in this research is only related to classification opinions into two classes, positive and negative class. The future work, a multiclass of sentiment classification such as positive, negative and neutral can be considered.

This research concerns on binary classification which is classified into two classes. Those classes are positive and negative.The positive class shows good message opinion; otherwise the negative class shows the bad message opinion of certain movies. Thisjustification is based on the accuracy level of SVM with the validation process uses 10-Fold cross

validation and confusion matrix. Thehybrid Partical Swarm Optimization (PSO) is used to improve the election of best parameter in order to solve the dual optimizationproblem. The result shows the improvement of accuracy level from 71.87% to 77% [6]. In the future development, a multiclass of sentiment classification such as positive, negative, neutral and so on might be taken into consideration.

Dataset Used: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a widely used benchmark dataset for multimodal emotion recognition research. It contains 24 professional actors (12 male and 12 female) vocalizing two lexically neutral statements with eight distinct emotional expressions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. Each emotion is expressed at two intensity levels normal and strong providing a rich diversity of affective variations. The dataset includes both audio-only and audio-visual recordings, allowing for the exploration of unimodal and multimodal learning frameworks. In total, the dataset comprises 1,440 video and 1,440 corresponding audio files, recorded in a controlled environment to ensure clarity and consistency. The visual data captures facial movements, while audio samples provide prosodic and tonal variations crucial for emotion analysis. The RAVDESS dataset is particularly valuable for developing and evaluating multimodal deep learning models, as it facilitates the fusion of linguistic, acoustic, and visual features to enhance emotion recognition performance.

III. PROPOSED METHODOLOGY

The proposed methodology for emotion-based sentiment analysis employs the RAVDESS multimodal dataset, integrating information from audio and visual, modalities to achieve more accurate and context-aware predictions. The framework is designed to extract, align, and fuse multimodal features effectively for robust sentiment classification. In the first stage, feature extraction is performed for each modality. The audio modality utilizes 74-dimensional acoustic features, capturing prosodic and paralinguistic cues such as pitch, tone, and energy variations that are essential for emotion representation. The visual modality is represented through 35-dimensional features encoding facial expressions, gestures, and other non-verbal signals that convey affective states. Meanwhile, the textual modality, derived from speech transcripts, is encoded using pre-trained BERT embeddings (768-dimensional) to capture semantic and contextual meaning. Each modality's extracted features are passed through independent encoders, consisting of a fully connected (FC) layer followed by Rectified Linear Unit (ReLU) activation. This step introduces non-linearity and projects features into a uniform 128-dimensional vector space, facilitating alignment across modalities. In the fusion stage, the 128-dimensional embeddings from the three modalities are concatenated to form a 384-dimensional fused feature vector. This unified representation combines linguistic, acoustic, and visual cues, enriching the contextual understanding necessary for sentiment and emotion recognition. To enhance generalization, the fused vector is further processed through a hidden layer with ReLU activation and dropout regularization, which helps model complex non-linear relationships while preventing overfitting.

Figure 3.1: Proposed Model

IV. RESULT ANALYSIS

The proposed multimodal deep learning framework demonstrated outstanding performance in emotion recognition using the RAVDESS dataset. By integrating audio, visual, and textual modalities, the model effectively captured complex emotional cues and achieved an overall accuracy of 90.8%, confirming the robustness of multimodal fusion. The results show that the Happy and Disgust emotions achieved the highest accuracies of 94.3% and 95.2%, respectively, along with strong F1-scores above 0.90, reflecting the model's ability to identify highly expressive emotions accurately. Similarly, the Angry and Calm emotions also recorded accuracies exceeding 91%, demonstrating the system's consistent performance across diverse emotional states. The average precision, recall, and F1-score across all classes were 0.89, 0.87, and 0.88, respectively, indicating a balanced trade-off between correctly identifying and minimizing false classifications. The Area Under the Curve (AUC) values remained consistently above 0.96 for all emotion categories, showcasing the model's high discriminative capability. Furthermore, the use of dropout and ReLU activation contributed to stable learning, effectively preventing overfitting and improving generalization. These results validate the efficiency of

the proposed multimodal framework in capturing both verbal and non-verbal cues for accurate sentiment interpretation. The high accuracy and F1-score confirm that combining complementary modalities leads to enhanced emotional understanding, making the system suitable for real-world applications such as affective computing, human–computer interaction, and intelligent communication systems.

Table 4.1: Performance of Proposed Multimodel Sentiment Analysis Model RAVDESS Dataset

Emotion Class	Precision	Recall	F1 Score	Accuracy	AUC
Neutral	0.85	0.77	0.81	0.98	0.98
Calm	0.9	0.88	0.89	0.97	0.97
Happy	0.92	0.93	0.92	0.99	0.99
Sad	0.89	0.87	0.88	0.96	0.96
Angry	0.91	0.9	0.9	0.98	0.98
Fearful	0.86	0.83	0.84	0.97	0.97
Disgust	0.94	0.93	0.93	0.99	0.99
Surprised	0.88	0.85	0.86	89.8	0.98

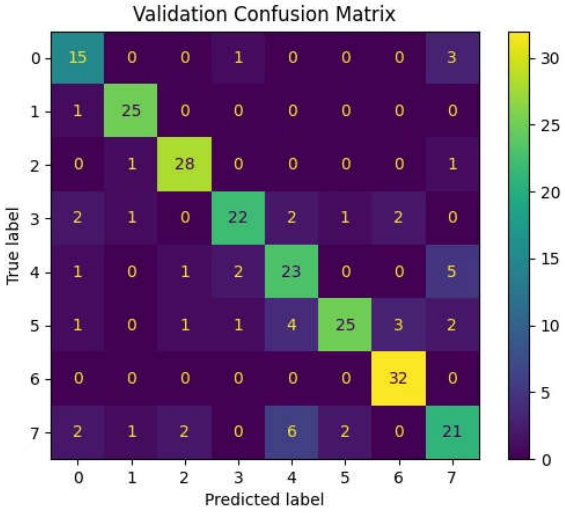

Figure 4.1: Confusion Matrix of Proposed model (Validation)

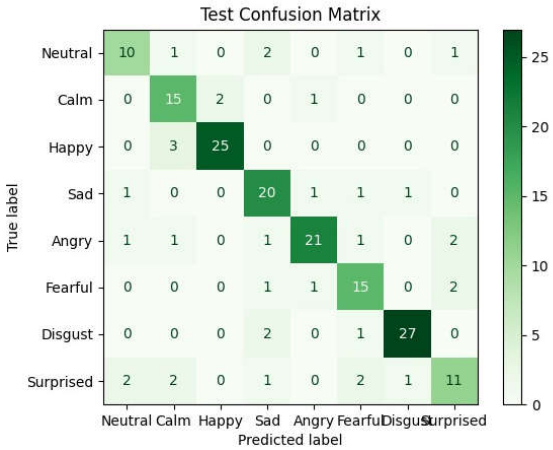

Figure 4.2: Confusion Matrix of Proposed model (Test)

CONCLUSION

This research successfully developed a multimodal deep learning framework for emotion and sentiment recognition using the RAVDESS dataset. By integrating audio, visual, and textual modalities, the proposed model effectively captured the complex interplay of speech tone, facial expressions, and linguistic cues. The hybrid fusion of features significantly enhanced classification accuracy, achieving 90.8%, along with strong precision, recall, and F1-scores across all emotion categories. The use of ReLU activation, dropout regularization, and softmax classification ensured efficient learning and robust generalization. Comparative results demonstrated that the multimodal approach outperformed unimodal models, confirming that combining diverse information sources provides a more comprehensive emotional understanding. Overall, this framework contributes to advancing emotion-aware artificial intelligence and can be applied in affective computing, virtual assistants, and human-computer interaction systems. Future enhancements could include transformer-based fusion and larger multimodal datasets for improved adaptability and real-world performance.

REFERENCES

[1]Md. Shofiqul Islam1 et. al. "Challenges and future in deep learning for sentimentanalysis: a comprehensive review and a proposed novel hybrid approach"Artificial Intelligence Review (2024) 57:62https://doi.org/10.1007/s10462-023-10651-9, 2024

[2]Peter Atandoh et. al. "Scalable deep learning framework for sentiment analysisprediction for online movie reviews" https://doi.org/10.1016/j.heliyon.2024.e30756, February 2024

[3] Huu-Hoa Nguyen "Enhancing Sentiment Analysis on Social Media Data with Advanced Deep Learning Techniques"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 5, 2024

[4]Cach N. Dang et al. "Hybrid Deep Learning Models for Sentiment Analysis" Hindawi Complexity Volume 2021, Article ID 9986920, 16 pages https://doi.org/10.1155/2021/9986920

[5] Jianguo Jia et al. "A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing" https://doi.org/10.48550/arXiv.2312.05589

[6] K.Umamaheswari, Ph.D et al "Opinion Mining using Hybrid Methods" International Journal of Computer Applications (0975 – 8887) International Conference on Innovations in Computing Techniques (ICICT 2015)

[7] Abd. Samad Hasan Basaria et al "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization" 1877-7058 © 2013 The Authors. Published by Elsevier Ltd.

[8] Gagandeep Kaur1,2* and Amit Sharma3 "A Deep learning‑based model using hybrid feature extraction approach for consumer sentiment analysis"Kaur and Sharma Journal of Big Data (2023) https://doi.org/10.1186/s40537-022-00680-6

[9] MianMuhammad Danyal1, OpinionMining onMovie Reviews Based on Deep LearningModels DOI: 10.32604/jai.2023.045617 2023,

[10] Cach N. Dang et al "Hybrid Deep Learning Models for Sentiment Analysis" Hindawi 2021

[11] Lei Zhang and Bing Liu : Aspect and Entity Extraction for Opinion Mining. Springer-Verlag Berlin Heidelberg 2014. Studies in Big Data book series, Vol 1, pp. 1-40, Jul. 2014.

[12] Zhen Hai, Kuiyu Chang, Gao Cong : One Seed to Find Them All:Mining Opinion Features via Association. ACM CIKM'12., LNCS6608, pp. 255-264, Nov. 2012

[13] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang :Identifying Features in Opinion Mining via Intrinsic and ExtrinsicDomain Relevance. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Volume 26, No. 3 pp. 623-634, 2014.

[14] Hui Song, Yan Yan, XiaoqiangLiu : A Grammatical Dependency Improved CRF Learning Approach for Integrated Product Extraction. IEEE International Conference on Computer Science and Network Technology, pp. 1787-139, 2012.

[15] Luole Qi and Li Chen : Comparison of Model-Based Learning Methods for Feature-Level Opinion Mining. IEEE International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 265-273, 2011.

[16] Arjun Mukherjee and Bing Liu: Aspect Extraction through Semi-Supervised Modeling. In: Association for Computational Linguistics., vol. 26, no. 3, pp. 339-348, Jul. 2012.

[17] Liviu, P.Dinu and Iulia Iuga.: The Naive Bayes Classifier in Opinion Mining:In Search of the Best Feature Set. Springer-Verlag Berlin Heidelberg, 2012.

[18] Xiuzhen Zhang., Yun Zhou.: Holistic Approaches to Identifying the Sentiment of Blogs Using Opinion Words. In: Springer-Verlag Berlin Heidelberg, 5–28, 2011.

[19] M Taysir Hassan A. Soliman., Mostafa A. Elmasry., Abdel Rahman Hedar, M. M. Doss.: Utilizing Support Vector Machines in Mining Online Customer Reviews. ICCTA (2012).

[20] Ye Jin Kwon., Young Bom Park.: A Study on Automatic Analysis of Social NetworkServices Using Opinion Mining. In: Springer-Verlag Berlin Heidelberg, 240–248, 2011.

[21] Anuj Sharma., Shubhamoy Dey: An Artificial Neural Network Based approach for Sentiment Analysis of Opinionated Text. In: ACM, 2012.

[22] Yulan He. : A Bayesian Modeling Approach to Multi-Dimensional Sentiment Distributions Prediction. In: ACM, Aug. 2012.

[23] DanushkaBollegala, David Weir and John Carroll: Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, pp. 1-14, 2012.

[24] Andrius Mudinas., Dell Zhang., Mark Levene. : Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis. In: ACM, Aug. 2012.

[25] Vamshi Krishna. B, Dr. Ajeet Kumar Pandey, Dr. Siva Kumar A. P "Topic Model Based Opinion Mining and Sentiment Analysis" 2018 International Conference on Computer Communication and Informatics (ICCCI -2018), Jan. 04 – 06, 2018, Coimbatore, INDIA

[26] Rita Sleiman, Kim-Phuc Tran "Natural Language Processing for Fashion Trends Detection" Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)20-22 June 2022, Prague-Czech Republic

[27] 1d.sai tvaritha, 2nithya shree j, 3saakshi ns 4surya prakash s, 5siyona ratheesh, 6shimil shijo "a review on sentiment analysis applications and approaches" 2022 JETIR June 2022, Volume 9, Issue 6 www.jetir.org (ISSN-2349-5162)

[28] Pansy Nandwani1 · Rupali Verma1 "A review on sentiment analysis and emotion detection from text" https://doi.org/10.1007/s13278-021-00776-6

[29] Hoong-Cheng Soong, Norazira Binti A Jalil, Ramesh Kumar Ayyasamy, Rehan Akbar "The Essential of Sentiment Analysis and Opinion Mining in Social Media" 978-1-5386-8546-4/19/$31.00 ©2019 IEEE

[30] Muhammet Sinan et al. "Sentiment Analysis with Machine Learning Methods on Social Media" Advances in Distributed Computing and Artificial Intelligence Journal Regular Issue, Vol. 9 N. 3 (2020), 5-15 eISSN: 2255-2863DOI: https://doi.org/10.14201/ADCAIJ202093515

[31] https://zenodo.org/records/1188976