Deep Learning Approaches to Sentiment Analysis on Text, Visual, and Audio Modalities: A Review

Afzal Ahmad Azmi[1], Prof. Anurag Srivastava[2]
[1]*MTech Scholar, CSE, Department, NIIST Bhopal*
[2]*HOD, CSE, Department, NIIST Bhopal*

Abstract- **Sentiment analysis has rapidly evolved beyond text-only approaches to embrace the rich information contained in images and audio, giving rise to multimodal sentiment analysis. This survey provides a comprehensive review of deep learning methods applied to sentiment classification across text, visual, and audio modalities. We first examine modality-specific encoders: transformer-based and recurrent networks (e.g., BERT, BiLSTM) for textual sentiment, convolutional and vision transformer models (e.g., ResNet, ViT) for image-based emotion recognition, and convolutional–recurrent architectures for audio signals using spectrogram and MFCC features. Next, we analyze fusion strategies—early, late, and hybrid fusion—that integrate modality representations, highlighting the role of attention mechanisms and multimodal transformers (e.g., MMT) in dynamically weighting cross-modal interactions. Benchmark datasets such as CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD are surveyed, along with evaluation metrics including accuracy, F1-score, and concordance correlation coefficient. We discuss practical applications in social media monitoring, customer feedback analysis, and human–computer interaction. Key challenges such as data imbalance, modality synchronization, domain adaptation, and model interpretability are addressed, alongside proposed solutions like data augmentation, adversarial training, and self-supervised pretraining. Finally, we outline future research directions, including lightweight architectures for edge deployment, advanced fusion techniques, and explainable multimodal sentiment frameworks. By synthesizing recent advances, this survey serves as a roadmap for researchers developing robust and scalable multimodal sentiment analysis systems.**

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

I. INTRODUCTION

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 communication with 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 research in sentiment analysis field.

Table 1 Review of Related Work

S.No.	Title	Description	Future work
1	Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach	Author's introduced the innovative CRDC (Capsule with Deep CNN and Bi structured RNN) model, which demonstrated superior performance compared to other methods. Our proposed approach achieved remarkable accuracy across different databases: IMDB (88.15%), Toxic (98.28%), CrowdFlower (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.
2	Scalable deep learning framework for sentiment analysis prediction for online movie reviews	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.
3	Hybrid Deep learning models for sentiment Analysis	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	In future various other combinations of models are tested on reliability and computation time. Reliability of the latter was significantly higher.

		accuracy and computation time—the hybrid approaches consistently outperformed single models. In particular, integrating deep networks with SVM yielded the greatest gains in accuracy and reliability across all datasets	
4	AReview of Hybrid and Ensemble in Deep Learning for Natural Language Processing	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.	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.
5	Opinion Mining using Hybrid Methods	In this work the rating of movie in twitter is taken to review a movie by using opinion mining. Author proposed hybrid methods using SVM and PSO to classify the user opinions as positive, negative for the movie review dataset which could be used for better decisions.	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.
6	Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization	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. This justification is based on the accuracy level of SVM with the validation process uses 10-Fold cross validation and confusion matrix. The hybrid Partical Swarm Optimization (PSO) is used to improve the election of best parameter in order to solve the dual optimization problem. The result shows the improvement of accuracy level from 71.87% to 77%.	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 hybrid models were evaluated on eight publicly available text-based sentiment datasets spanning both tweets and longer-form reviews:

Sentiment140 (Twitter)

Sentiment140 is a large-scale Twitter dataset containing 1.6 million tweets labeled for binary sentiment (positive or negative) based on emoticon heuristics. Tweets were collected via Twitter's API and automatically annotated using smiley and frowny emoticons as proxies for sentiment. The dataset captures informal language, abbreviations, and hashtags typical of social media, making it valuable for training and evaluating models on real-world, short-text sentiment classification tasks.

Twitter US Airline Sentiment

The Twitter US Airline Sentiment dataset comprises approximately 14 000 tweets directed at major U.S. airlines, each manually labeled as "positive," "negative," or "neutral." Collected between February and April 2015, it reflects customer opinions on airline service, delays, and staff interactions. The dataset's moderate size and balanced classes make it ideal for evaluating sentiment analysis models on domain-specific, customer-feedback data in the transportation sector.

SemEval-2017 Task 4 (Twitter)

SemEval-2017 Task 4 is a benchmark Twitter sentiment dataset introduced for the International Workshop on Semantic Evaluation. It includes hundreds of thousands of tweets annotated for binary and fine-grained sentiment categories: "very negative," "negative," "neutral," "positive," and "very positive." Expert annotators provide high-quality labels, enabling detailed model assessment. The dataset's diverse topics and timely social media context offer robust evaluation for advanced sentiment classification techniques.

Stanford Sentiment Treebank (SST-2)

The Stanford Sentiment Treebank 2 (SST-2) features over 67 000 movie-review sentences labeled as "positive" or "negative." Derived from film reviews on Rotten Tomatoes, it provides phrase-level annotations organized in a parse tree structure. Its fine-grained sentiment labels and syntactic information facilitate the development of models that understand compositional sentiment, making SST-2 a cornerstone dataset for evaluating deep learning architectures on sentence-level sentiment tasks.

IMDB Movie Reviews

The IMDB Movie Reviews dataset comprises 50 000 highly polarized film reviews, evenly split into 25 000 training and 25 000 test samples, with binary labels for positive or negative sentiment. Reviews are longer and more descriptive than tweets, often exceeding 200 words. This dataset challenges models to capture long-range dependencies and nuanced opinions, serving as a standard benchmark for text classification and natural language understanding.

Amazon Fine Food Reviews

Amazon Fine Food Reviews contains approximately 500 000 product reviews from Amazon's food and beverage category, each annotated with a 1–5 star rating. Reviews are binarized into "positive" (4–5 stars) and "negative" (1–2 stars). The dataset includes both text and metadata such as product ID, user ID, and timestamp, enabling analysis of sentiment in the context of product characteristics, user behavior, and temporal trends.

Yelp Review Dataset

The Yelp Review Dataset includes over 650 000 business reviews with corresponding star ratings and business metadata. Reviews are labeled as "positive" for 4–5 stars and "negative" for 1–2 stars, offering a broad cross-section of service-industry feedback. The dataset's scale, varied business categories, and user-provided text enable training and evaluating sentiment models on multi-domain, user-generated content, reflecting real-world applications in recommendation and reputation systems.

III. FINDINGS OF THE SURVEY

The comprehensive review of deep learning approaches for multimodal sentiment analysis across text, visual, and audio modalities yields several key findings:

Modality-Specific Encoders Exhibit Complementary Strengths

Transformer-based and recurrent models (e.g., BERT, BiLSTM) excel at capturing semantic and contextual information from text, while convolutional architectures (e.g., ResNet) and vision transformers (ViT) effectively extract spatial features from images. For audio, convolutional–recurrent hybrids using spectrogram or MFCC inputs accurately model prosodic and acoustic cues. The interplay of these specialized encoders provides a richer representation of sentiment than any single modality alone.

Fusion Strategies Critically Impact Performance

Among the fusion paradigms surveyed, hybrid fusion which blends both early (featurelevel) and late (decisionlevel) fusion consistently outperforms purely early or purely late methods. Attention mechanisms and multimodal transformer layers that dynamically weigh each modality's contribution have proven particularly effective in aligning heterogeneous feature spaces.

Benchmark Datasets Highlight Diversity and Limitations

Experiments on CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD demonstrate that while deep multimodal models achieve high accuracy and F1-scores, performance varies considerably across datasets due to differences in domain (e.g., conversational vs. scripted), class balance, and annotation granularity. The lack of large-scale, well-balanced multimodal corpora remains a bottleneck.

Data Imbalance and Synchronization Challenges Remain

Imbalanced emotional categories such as rare expressions of disgust or surprise—lead to skewed model learning. Moreover, aligning modalities with disparate temporal resolutions (e.g., framelevel audio vs. wordlevel text) introduces synchronization difficulties that can degrade fusion efficacy.

Emerging Directions Show Promise

Self-supervised pretraining on unlabeled multimodal streams, adversarial data augmentation, and lightweight model architectures for edge deployment have demonstrated potential to address data scarcity, improve robustness, and enable realtime inference. Similarly, initial efforts in explainable AI have started to open model "black boxes," enhancing interpretability for end users.

These findings underscore that while deep learning has markedly advanced multimodal sentiment analysis, continued progress hinges on resolving data, synchronization, and interpretability challenges—paving the way for more robust, scalable, and transparent sentiment systems.

CONCLUSION

In this review, we have systematically examined deep learning methodologies for multimodal sentiment analysis across text, visual, and audio data sources. By surveying transformer-based and recurrent architectures (e.g., BERT, BiLSTM) for textual inputs, convolutional and vision transformer models (e.g., ResNet, ViT) for visual sentiment signals, and convolutional–recurrent networks for audio representations, we have highlighted the distinct strengths and limitations of each encoder. Our analysis underscores how these specialized models extract complementary features—semantic, spatial, and acoustic—that are essential for capturing nuanced emotional cues. We then explored fusion strategies that combine modality-specific embeddings, including early fusion, late fusion, and hybrid approaches. Attention mechanisms and multimodal transformers have emerged as particularly effective, dynamically weighting cross-modal interactions to bolster performance. Through comparison on benchmark datasets such as CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD, we demonstrated that hybrid fusion architectures consistently outperform single-modality baselines, as measured by accuracy, F1-score, and concordance correlation coefficient. Despite these advances, several challenges persist. Data imbalance and limited labeled examples in certain emotional categories hinder robust model training. Synchronizing modalities with differing temporal resolutions remains an open problem. Additionally, domain adaptation across diverse contexts (e.g., conversational versus broadcast media) and the interpretability of deep multimodal models require further investigation. Looking ahead, future research should prioritize lightweight, energy-efficient architectures suitable for edge deployment, alongside advanced fusion techniques that can learn optimal cross-modal alignments with minimal supervision. Incorporating explainable AI frameworks will enhance trust and transparency, while self-supervised and adversarial training paradigms can mitigate data scarcity and improve generalization. By synthesizing current developments and identifying key obstacles, this survey provides a comprehensive roadmap for advancing multimodal sentiment analysis. We anticipate that continued innovation in model design, training strategies, and interpretability

will unlock more accurate and adaptable sentiment inference systems, ultimately enriching applications in social media monitoring, human–computer interaction, and beyond.

REFERENCE

[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