IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY OF TRANSFORMER-DRIVEN HYBRID DEEP LEARNING METHODS FOR MULTIMODAL SENTIMENT ANALYSIS IN NLP

[1]R.Sivagami, [2]P Anitha

[1]Assistant Professor, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi, Tamilnadu

[2] Assistant Professor, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi, Tamilnadu

Abstract: Sentiment analysis (SA) is a computational approach aimed at identifying, extracting, and quantifying subjective information such as emotions, attitudes, and opinions. While traditional unimodal sentiment analysis relies solely on textual data, multimodal sentiment analysis (MSA) leverages diverse information sources, including speech, tone, facial expressions, and body movements, to capture a deeper, more accurate representation of human emotions. This study provides a comprehensive review of existing research on multimodal fusion methods and feature extraction, emphasizing the integration of textual, visual, and audio-visual data through transformer-based deep learning models. It traces the evolution and theoretical underpinnings of MSA, outlining its major advancements, existing challenges, and practical advantages. Furthermore, the paper identifies emerging trends and future research directions, serving as a valuable reference for scholars and practitioners seeking to advance the field of multimodal sentiment analysis.

Keywords: Unimodal sentiment analysis, Multimodal sentiment analysis, fusion methodologies, textual, visual, and audio/visual data, transformer-based deep learning technique

1. Introduction

Text sentiment analysis, a branch of unimodal sentiment analysis, focuses on detecting the emotional tone conveyed in textual data such as reviews, tweets, and comments. This approach is significant as it helps businesses and organizations gauge public opinion and make informed, data-driven decisions. A foundational contribution to this field was made by Angelpreethi and Kumar [1] who examined the role of big data visualization in preprocessing and analysis for sentiment mining on text-based data sources such as social media posts and online reviews. The study explored key issues, challenges, and opportunities in designing effective visual representations for large-scale and complex datasets to enhance decision-making and analytical insights. Although this framework performs well for unimodal text analysis, it does not account for deeper linguistic subtleties or the integration of multiple modalities, which limits its applicability in more complex sentiment understanding tasks.

The paper by Nadeem et al. [2] and Sharmila et al. [3] presents a hybrid deep learning framework that combines Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) for sentiment classification on text data. The CNN component is used for extracting local features and key n-gram patterns from textual input, while the GRU component captures sequential dependencies and contextual relationships between words. In a study, a hybrid opinion mining framework [4] were used for big data applications, combining lexicon-based and machine learning techniques to enhance sentiment classification accuracy. The work focuses on unimodal (text-only) sentiment analysis applied to massive datasets such as social media

posts and online reviews. The framework integrates a dictionary-based sentiment pipeline with machine learning techniques, offering both interpretability and scalability suitable for big data environments.

Another study by Shashank Mishra [5] proposes a deep-learning approach that uses a Long Short-Term Memory (LSTM) network for text-based sentiment analysis. The model takes pre-processed text data (e.g., user posts, reviews) and applies embedding/tokenizing techniques followed by the LSTM architecture to classify sentiment into categories like positive, negative or neutral.

A frequency-weighted feature approach [6]**,** assigning weights to highlight important features and incorporating them into sentiment classification models to improve accuracy. Their model, Dom_classi, introduces a hybrid lexicon-based classifier that refines sentiment analysis by detecting domain-specific words whose sentiment polarity in SentiWordNet (SWN) may not reflect their actual usage in labeled reviews; it computes a frequency-based probability (using term frequency in positive vs. negative reviews) to predict the correct polarity class, and then updates the sentiment score of such words with a TF–IDF-based weighting. The authors test this approach on Twitter data from several product domains (mobile, laptops, kitchen, electronics), achieving an overall accuracy of about 84%, and show that their method improves over using SWN alone.

Another study by Chen, X et al. [7] proposes a deep-learning method that explicitly models the effect of negations (e.g., “not good”) and intensifiers (e.g., “very good”)—two linguistic features that traditional sentiment models often mishandle. The authors introduce a mechanism that generates special “supplement vectors” when negation or intensifier cues appear, allowing CNN and LSTM-based models to adjust sentiment representations using full contextual information. However, the gains for intensifiers are smaller, and the approach still does not address more complex semantic issues such as sarcasm, idioms, and domain-specific language, making the method effective but limited to specific linguistic challenges.

Another paper introduced a hybrid sentiment analysis framework, OPINE_NEG [8] that combines rule-based detection of negations and intensifiers with machine-learning classification to handle the linguistic complexity and noise of social media text, using features like POS tagging to improve accuracy and scalability, though it remains limited by its unimodal design and lack of fairness evaluation. Garg and Subrahmanyam [9] focus on improving sentiment analysis by effectively handling negation, which often reverses or alters sentiment polarity in text. The authors propose a super-ensemble deep-learning approach that combines multiple models—such as LSTM, Bi-LSTM, GRU, and CNN—to enhance the detection of negation cues and their impact on nearby sentiment-bearing words.

Angelpreethi and Ramesh Kumar [10] both a lexicon-based method (called Senti_Lexi) and a machine-learning based method (called Senti_Mac) to analyze and classify opinions of students about the COVID-19 pandemic — particularly their views on online education and pandemic-related experiences — into positive, negative, or neutral sentiment. By comparing both approaches, the study aims to examine which method performs better for this domain of student feedback during COVID-19. As a contribution, this comparative analysis helps understand how traditional lexicon-based sentiment analysis stacks up against machine-learning classification for pandemic-era student opinions — potentially aiding educators or institutions to gauge student sentiment.

Neha Punetha et al. [11] proposed a novel unsupervised method — called EOT-NH (Evaluation-Based on Distance from Average Solution Optimization Technique Negation Handling Tagger) — to perform sentiment classification by mathematically modeling negation handling without relying on pre-training or large labeled datasets. The approach combines multi-criteria decision-making (MCDM) with game-theoretic ideas: it computes textual features like context scores and emotion scores for input sentences (whether they contain negations or not), then uses those as criteria in a decision matrix to classify sentiment as positive, negative, or neutral.

Another study [12] provides an overview of sentiment-analysis techniques and argues that hybrid methods—combining lexicon-based and machine-learning approaches—are more effective for handling large, unstructured big-data environments such as social media. Its strength lies in offering a clear, conceptual survey of existing methods and highlighting why hybrid approaches may overcome the weaknesses of individual techniques. The paper by Huan et al. [13] proposes an emotionally charged text-classification model that integrates deep learning with sentiment semantic features by combining a recurrent neural network (RNN) architecture with sentiment-specific vectors derived from lexicons to enrich the input representation. This hybrid design enables the model to capture both contextual dependencies and sentiment-oriented cues,

resulting in improved performance on multiple benchmark datasets compared to baseline deep learning approaches that rely solely on word embeddings.

The paper by Angelpreethi [14] given a fuzzy-logic-driven model for classifying sentiments expressed in textual data, focusing on reducing ambiguity that commonly arises in opinionated language. The study uses fuzzy linguistic hedges to represent varying degrees of sentiment intensity—such as "slightly," "highly," or "extremely"—allowing the system to capture subtle gradations that traditional binary or numerical classifiers often miss. By constructing a fuzzy rule-based framework and applying it to text datasets, the model enhances interpretability and provides smoother, human-like decision boundaries for sentiment categorization. While the approach improves handling of linguistic uncertainty, it remains limited to text-based inputs and depends heavily on well-designed fuzzy rules, which may require domain expertise and may not generalize easily across datasets or languages.

In a survey conducted by P.Q. Dao et al. [15], it is noted that sentiment analysis and emotion analysis often share similar expressive forms. Unlike traditional methods that depend on keyboards and mice as input devices, modern sentiment analysis leverages alternative modalities such as speech, gestures, text messaging, and facial expressions to interpret human opinions, emotions, and polarity. These subjective modalities can communicate a broad spectrum of sentiments, including positive, negative, neutral, joy, and delight**.** In recent years, there has been growing research interest in sentiment analysis, particularly focused on extracting emotional cues from voice, text, and facial expressions**.**

2. Multimodal Sentiment Analysis (MSA)

Another study proposed a lexicon-based sentiment-analysis method [16] tailored for microblog data, with a particular focus on handling negations, intensifiers, slang words, emoticons (emojis as text characters)**,** and emotion terms — linguistic features that ordinary lexicon-based methods often ignore or treat poorly. They argue that correctly detecting negation (e.g. "not good"), intensifiers (e.g. "very good"), slang or emoticons is crucial for correctly assigning sentiment polarity (positive vs negative) to short, informal microblog texts. Their objective is to classify microblog posts into positive or negative sentiment by incorporating these additional linguistic cues instead of relying solely on standard sentiment lexicons.

In the modern digital era, individuals increasingly communicate their thoughts and emotions through a combination of text and visual content**.** This trend has given rise to Multimodal Sentiment Analysis (MSA) [15], an emerging research area that seeks to analyze and interpret sentiments using data from multiple modalities. In a study by D. Hazarika et al. [17], the authors introduced the MISA (Multimodal Interactive Sentiment Analysis) framework, which integrates text, audio, and visual data to improve sentiment prediction accuracy. The framework employs modality-specific feature extraction techniques—such as GloVe or BERT for textual data, openSMILE for audio, and ResNet or OpenFace for visual features—within a two-stream architecture. This design captures both shared (invariant) and unique (modality-specific) representations.

Zhang et al. [18] proposed a multimodal sentiment classification framework based on a semi-supervised learning approach**.** The model extracts modality-specific sentiment features—such as textual semantics and audio tone—through independent processing streams, while also modeling cross-modal interactions (e.g., text–audio or audio–visual relationships) to capture complementary emotional cues. Techniques like attention mechanisms and shared subspaces are employed to integrate these multimodal features, enhancing prediction accuracy**,** robustness, and generalization. L. Sun et al. [19] introduced a multimodal framework for emotion recognition and sentiment analysis that combines text, audio, and visual modalities using an attention-enhanced recurrent neural network (RNN) to improve predictive performance. The framework extracts modality-specific features, applies an attention mechanism to emphasize the most informative signals from each modality, and employs recurrent models (such as LSTM or GRU) to capture temporal dependencies in sequential data. Designed particularly for video-based multimodal datasets, the model effectively detects emotions and sentiments by integrating information across modalities.

Similarly, K. Vasanth et al. [20] proposed a dynamic fusion framework for Multimodal Sentiment Analysis (MSA) that integrates text, audio, and video data to enhance sentiment prediction on social media platforms. Unlike static fusion techniques, this framework employs an adaptive fusion mechanism—such as attention or gating—to contextually weight and combine modality features based on their relevance. Garcia et al. [21] introduced a three-level multimodal emotion recognition framework that integrates textual, acoustic, and visual data using a hierarchical processing approach to predict emotions in multimedia content. The study

focuses on identifying emotions such as happiness, sadness, and anger conveyed through multiple modalities—a crucial component of affective computing, especially given the complexity of social media and multimedia environments. To better capture inter-modal dependencies with dynamic weighting, M. Jiang et al. [22] proposed a cross-modality gated attention fusion framework for Multimodal Sentiment Analysis (MSA). This approach utilizes text, audio, and visual features for sentiment prediction, beginning with modality-specific feature extraction followed by gated attention-based fusion. The fusion mechanism dynamically adjusts weights across modalities, enabling more accurate and noise-resistant sentiment classification. The framework's strengths include effective cross-modal interaction modeling, robustness to noisy inputs, and adaptability to multimedia-rich environments, though it is limited by high computational demands.

Additionally, T. Grosz et al. [23] presented a multimodal framework for humor and imitated emotion recognition, leveraging pre-trained transformer embeddings—including BERT and ELECTRA for text, Wav2Vec 2.0 for audio, and Vision Transformer (ViT) for visuals. The study employs integrated gradients to identify task-relevant subspaces within these embeddings, enhancing both performance and interpretability.

Sun et al. [24], Tingting Zhang et al. [25], and Yanjing Wang et al. [26] introduced a general debiasing framework for Multimodal Sentiment Analysis (MSA) that integrates text, audio, and visual modalities to reduce biases stemming from gender, culture, and modality-specific influences, thereby promoting fairer and more accurate sentiment prediction. The framework adopts a debiasing strategy, often implemented through adversarial training, where an adversarial network is trained to remove bias-related attributes (e.g., gender or cultural cues) from modality representations so that the model focuses solely on sentiment-relevant features. Alternatively, a reweighting mechanism can be employed to adjust the influence of biased samples or features during training, further minimizing their impact on sentiment predictions. This approach combines modality-specific feature extraction with fusion mechanisms that preserve sentiment classification performance while mitigating bias. The debiased representations are fused using strategies such as attention-based weighting, which dynamically evaluates each modality's contribution, and concatenation or gating methods to produce a unified representation for final sentiment prediction.

In another study, R. Jain et al. [27] proposed a real-time multimodal sentiment detection framework that processes text, audio, and visual modalities—a key requirement in affective computing given the surge of multimedia content on social media platforms. The architecture follows a pipeline-based design encompassing feature extraction, deep learning-based analysis, and fusion techniques to process multimodal data efficiently in real-time. Specifically, BERT is used for extracting textual features, Wav2Vec for acoustic features, and CNN or ResNet for visual features. The framework is optimized for low-latency performance while maintaining high accuracy, making it well-suited for applications such as live customer feedback monitoring, public opinion analysis, and interactive virtual assistants. Y. Zheng et al. [28] introduced the Discriminative Joint Multi-Task Framework (DJMF), an innovative approach to Multimodal Sentiment Analysis (MSA) that integrates text, audio, and visual modalities within a multi-task learning framework. This model captures both within-task and cross-task relationships, allowing it to perform sentiment classification alongside related tasks such as emotion recognition. The framework incorporates modality-specific feature extraction, a unified multi-task architecture with discriminative modeling, and dynamic task interaction mechanisms to enhance learning efficiency.

Similarly, Ayetrian et al. [29] and Md. Mithun Hossain et al. [30] proposed an inter-modal attention-based deep learning architecture that integrates text, visual, and optionally other modalities (such as audio or metadata) for detecting hate speech, fake news, and abusive language. The framework achieves strong classification performance by employing an inter-modal attention mechanism that dynamically models cross-modal interactions to form a unified representation. In another study, Q. Lu et al. [31] proposed the Coordinated-Joint Translation Fusion (CJTF) framework integrated with Sentiment-Interactive Graph Convolutional Networks (SI-GCNs) for Multimodal Sentiment Analysis (MSA). This approach combines text, audio, and visual inputs to enhance sentiment prediction, employing a translation-based fusion strategy alongside graph convolutional networks (GCNs) to model sentiment interactions across modalities, resulting in improved classification accuracy.

Table 1 [15], provides a summary of the datasets, multimodal features (MF) and fusion methods (MFM) for the multimodal research papers reviewed.

Table 1. [15] Datasets, Multimodal Features (MF), Multimodal Fusion Methods (MFM)

S.No.	Study & Year	Datasets	MF	MFM
1	Hazarika et. al. [17], 2020	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion
2	Dong Zhang et. al. [18], 2020	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion
3	Sun et. al. [19], 2021	3rd+OMuSe-CaR	Text+ Visual +Audio	Late Fusion
4	Vasanth et. al. [20], 2022	Self+NO	Text+ Visual +Audio	Early Fusion
5	Garcia et. al. [21], 2022	Self+NO	Text+ Visual +Audio	Late Fusion
6	Jiang et. al. [22], 2022	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion
7	Grosz et. al. [23], 2023	3rd+NO	Text+ Visual +Audio	Late Fusion
8	Sun et. al. [24], 2023	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion
9	Jain et. al. [27], 2023	Self+NO	Text+ Visual +Audio	Late Fusion
10	Zheng et. al. [28], 2024	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion
11	Ayetrian et. al. [29], 2024	Self+NO	Text+ Visual +Audio	Early Fusion
12	Lu et. al. [31], 2024	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion

3. Classification and Analysis

3.1 Datasets

The authors utilize various datasets, discussed in this article, with their features summarized in Table 2. [15].

Table 2 [15] Datasets used by the researchers

Dataset	Description	Size / Stats
CMU-MOSI	The dataset consists of 93 selected YouTube videos featuring a single speaker facing the camera and expressing opinions in English. It includes 89 speakers (48 male and 41 female), with no restrictions on recording setting, distance, or camera model.	The dataset contains 2,199 opinion segments, each annotated with sentiment intensity scores ranging from -3 to +3
CMU-MOSEI	Large-scale dataset for sentiment and emotion detection, from diverse YouTube speakers. The dataset ensures balanced gender representation and includes randomly selected phrases from both thematic and monologue videos.	~23,500 videos, 1,000 speakers
T4SA	Twitter dataset combining textual and visual content (tweets + photos). Neutral class somewhat ambiguous due to annotation quality.	1M tweets, 1.5M photos
DFMSD	A domain-free multimedia sentiment dataset collected from uncontrolled social media and outdoor environments using the Twitter Stream API, with unbiased annotations provided by psychologists.	The dataset comprises 14,488 tweets, including 10,244 photos
Fakeddit	Multimodal dataset of fake news from Reddit (text, images, metadata, comments). Offers multiple labeling schemes.	1M records (Mar 2008–Oct 2019)
MuSe-CaR	Dataset for MuSe 2021 challenges (text, audio, visual). YouTube automotive reviews with challenging in-the-wild conditions.	291 videos, 70 speakers
MVSA	Twitter-based multimodal dataset (text + images) with sentiment labels. Two subsets: MVSA-Single and MVSA-Multiple.	MVSA-S: 4,869 pairs; MVSA-M: 19,598 pairs
ReactionGIF	Dataset of tweets with GIF replies for two-turn conversation. GIFs mapped to emotions and sentiment.	30,000 tweet-GIF pairs

3.2 Multimodal Fusion Methods

Multimodal data, which capture information from multiple perspectives, are more informative than single-modal data, as different modalities can effectively complement each other. Major challenges in multimodal sentiment analysis involve maintaining the semantic integrity of each modality, ensuring effective cross-modal fusion, and integrating features from different modalities. The fusion methods used by various researchers are summarized in Table 3. Depending on the different types of modal fusion [15], it can be summed up as feature-based multimodal fusion in the early stages and decision-based multimodal fusion in the latter stages.

Table 3 Multimodal Fusion Methods utilized by the researchers

Method	Description	Advantages	Limitations
Early Fusion (Feature-based / Shallow Fusion)	Combines features from multiple modalities right after the first feature extraction stage. Features from different modalities are integrated into the same parameter space.	- Enables robust and accurate multimodal sentiment analysis .- Captures cross-modal relationships early.	- Parameter space mismatch between modalities can reduce effectiveness .- Requires large training datasets. – High model complexity → longer training times.
Late Fusion (Decision-level Fusion)	Combines outputs from independently trained models on different modalities at the decision stage.	- Flexible and lightweight .- Can handle missing modalities during inference.	- May lose cross-modal interactions present in early stages - Relies on each single-modal model's quality.

MSA models can be extended to develop new modalities based on complex temporal models and fusion procedures, or they may require further optimization to improve accuracy and other performance metrics. Recently, transfer learning has gained prominence as a key technique. Furthermore, multimodal sentiment analysis (MSA) algorithms can evaluate user credibility by integrating metadata and comments with user-related information.

4. Transfer Learning Techniques – A Rise in Multimodal Approaches

The goal was to advance the earlier approaches by developing more sophisticated multimodal fusion techniques that could fully leverage the complementary strengths of textual and visual data**.** Consequently, new strategies emerged that enhanced multimodal integration via attention mechanisms**,** contrastive learning, and cross-attention networks.

Xu et al. [32] and Adil M. et al. [33] developed a text-based sentiment analysis framework leveraging a Bidirectional Long Short-Term Memory (BiLSTM) neural network to enhance classification accuracy in social media and customer review data. The BiLSTM model captures contextual information from both preceding and succeeding tokens, overcoming the limitations of traditional machine learning and unidirectional LSTM models. However, their study remains restricted to unimodal text analysis, is limited in its adaptability to diverse or noisy real-world data, and relies on labeled datasets for training.

Al-Alshaqi, Rawat, and Liu et al. [34] proposed a BERT-based multimodal framework for fake news detection that integrates textual and visual data fusion to improve classification accuracy. The model leverages BERT for text feature extraction and a convolutional neural network (CNN) for image feature extraction, combining them through a fusion mechanism to capture the complementary information present in both modalities. This multimodal approach enhances the detection of misleading content by analyzing the semantic alignment between text and images, outperforming traditional text-only methods. The framework is computationally intensive, relies on high-quality multimodal datasets**,** and may face challenges in generalizing to diverse real-world news sources or handling manipulated media.

Xue et al. [35] and Deng et al. [36] proposed a multimodal fake news detection framework that emphasizes exploring the semantic consistency between textual and visual data to identify misinformation. By analyzing the degree of alignment between image content and accompanying text, the model effectively detects inconsistencies that often indicate falsified or misleading information. This approach enhances fake news detection performance by leveraging both modalities, outperforming traditional text-only methods on benchmark datasets. However, the study has certain limitations. Its effectiveness depends heavily on the quality and proper alignment of text–image pairs, which can be challenging in real-world data. Additionally, it may fail to detect fake news where both the text and image are fabricated yet internally consistent.

Jaiswal, Singh et al. [37] proposed a BERT-VGG19 Multimodal Variational Autoencoder (MVAE) framework for fake news detection that integrates textual and visual features to enhance classification accuracy. The model employs BERT for extracting deep contextual text embeddings and VGG19 for capturing rich visual representations, which are then fused using a variational autoencoder to learn a joint latent representation of multimodal data. This hybrid approach improves the model's ability to detect deceptive content across diverse news types by leveraging complementary cues from both modalities. Ying et al. [38] proposed a Multi-Level Multi-Modal Cross-Attention Network (MMCAN) for fake news detection that effectively integrates textual and visual modalities to improve detection accuracy. The framework employs a cross-attention mechanism to capture fine-grained interactions between text and image features at multiple levels, enabling the model to understand semantic consistency and contextual relationships across modalities.

The summary of the above said papers of transformer based techniques are given in the following Table 4.

Table 4 – Summary of Transfer based techniques

Author(s) & Year	Model / Framework	Key Features / Contributions	Limitations
Xu et al. [32]; Adil M. et al. [33]	BiLSTM-based Sentiment Analysis Framework	Uses Bidirectional LSTM to capture contextual info from both directions; improves classification accuracy in text-based sentiment analysis for social media and reviews.	Limited to text (unimodal) data; requires large labeled datasets; weak adaptability to noisy or diverse real-world data.
Al-Alshaqi, Rawat, & Liu et al. [34]	BERT + CNN Multimodal Framework for Fake News Detection	Integrates BERT for text and CNN for images; fusion captures complementary semantic alignment across modalities; improves fake news detection accuracy.	High computational cost; depends on high-quality multimodal datasets; lacks analysis on bias and explainability.
Xue et al. [35] & Deng et al. [36]	Semantic Consistency Multimodal Framework	Detects misinformation by measuring semantic alignment between text and image content; effective in identifying inconsistencies.	Relies on well-aligned text–image pairs; struggles with fabricated but consistent pairs; high computational requirements; limited generalization.
Jaiswal & Singh et al. [37]	BERT-VGG19 Multimodal Variational Autoencoder (MVAE)	Fuses deep text (BERT) and image (VGG19) embeddings through VAE to learn joint multimodal latent representations for fake news detection.	High computational complexity; limited interpretability; performance drops with noisy or imbalanced datasets; requires aligned text–image pairs.
Ying et al. [38]	Multi-Level Multi-Modal Cross-Attention Network (MMCAN)	Employs hierarchical cross-attention to capture fine-grained interactions between text and image; improves multimodal feature representation.	High computational and memory demands; sensitive to noisy data; poor scalability for real-time use; limited generalization.

5. Conclusion

Researchers across multiple fields have recognized the importance of multimodal sentiment analysis, making it a key focus in feature extraction and fusion research. This review discusses the challenges facing the area of multimodal sentiment analysis and provides advancements in using transfer learning techniques to improve specific model metrics. The transformer-based deep learning techniques have significantly advanced the field of multimodal sentiment analysis by enabling models to effectively capture complex relationships across textual, visual, and auditory modalities. Unlike traditional fusion or unimodal approaches, attention mechanisms selectively focus on the most informative features within and between modalities, improving both interpretability and sentiment classification accuracy. Transformer architectures such as BERT, cross-modal multi-head attention, and hierarchical interactive attention models have further enhanced contextual understanding and semantic alignment among modalities. These approaches achieve superior performance on benchmark datasets, demonstrating their ability to model intricate emotional cues and contextual dependencies. However, challenges such as high computational demands, dependence on large-scale labeled multimodal datasets, and limited robustness to noisy or missing modalities, and lack of interpretability remain open issues. Future research should focus on lightweight and explainable transformer frameworks, adaptive attention mechanisms, and data-efficient multimodal learning strategies to ensure scalable, fair, and generalizable sentiment analysis systems for real-world applications. Nonetheless, MSA approaches show promise in addressing these challenges.

References

[1] Angelpreethi, A., & Ramesh Kumar, S. B. (2016). Visualizing big data mining: Issues, challenges and opportunities. *International Journal of Control Theory and Applications, 9*(27), 455–460. [Online]. Available: https://serialsjournals.com/abstract/86429_61-182.pdf

[2] Aslam, N., Nadeem, A., Abid, M. K., & Fuzail, M. (2023). Text-based sentiment analysis using CNN-GRU deep learning model. *Journal of Information and Communication Technology Research and Application (JICTRA), 14*(1), 16–28. https://doi.org/10.51239/jictra.v14i1.318

[3] Sharmila, V., Kannadhasan, S., Kannan, A.R., Sivakumar, P., & Vennila, V. (Eds.). (2024*). Challenges in Information, Communication and Computing Technology*: Proceedings of the 2nd International Conference on Challenges in Information, Communication, and Computing Technology (ICCICCT 2024), April 26th & 27th, 2024, Namakkal, Tamil Nadu, India (1st ed.). CRC Press. https://doi.org/10.1201/9781003559092

[4] Angelpreethi, A., & Britto Ramesh Kumar, S. (2018). Opinion mining using hybrid approach on big data: A perspective. *International Journal of Scientific Research in Computer Science and Management Studies, 7*(5). https://doi.org/10.5281/zenodo.17399553

[5] Mishra, S., Aggarwal, M., Yadav, S., & Sharma, Y. (2023). An automated model for sentimental analysis using long short-term memory-based deep learning model. *International Journal of Engineering and Manufacturing (IJEM), 13*(5), 11–20. https://doi.org/10.5815/ijem.2023.05.02

[6] Angelpreethi, A., & Ramesh Kumar, S. B. (2019). Dom_Classi: An enhanced weighting mechanism for domain specific words using frequency-based probability. *International Journal of Applied Engineering Research, 14*(1), 140–148, [Online]. Available: ijaerv14n1_21.pdf

[7] Chen, X., Rao, Y., Xie, H., Li, X., Wang, F. L., Wang, S., & Kwong, C. K. (2019). Sentiment classification using negative and intensive sentiment supplement information. Data Science and Engineering, 4, 109–118. https://doi.org/10.1007/s41019-019-0094-8

[8] Angelpreethi, A. (2025). OPINE_NEG: An approach to detecting negations and intensifiers using social media data. *International Multidisciplinary Research Journal Reviews,* 2(8), 13–17. .[Online].Available: IMRJR.2025.020803.pdf

[9] Garg, S. B., & Subrahmanyam, V. V. (2023). A Research Paper on Negation Handling: Sentiment Analysis Using Super Ensemble Method in Deep Learning. *Indian Journal of Computer Science*, 8(3), 8–16. https://doi.org/10.17010/ijcs/2023/v8/i3/172862

[10] Angelpreethi A. Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic. IJIEMR Transactions. 2023; 12(2):380–387. Available: file:///C:/Users/USER/Downloads/LexiconandMachineLearningBasedComparativeAnalysis.pdf

[11] Punetha, N., Jain, G. Advancing sentiment analysis by addressing negation handling challenge via unsupervised mathematical approach. Soc. Netw. Anal. Min. 15, 20 (2025). https://doi.org/10.1007/s13278-025-01416-z

[12] Angelpreethi, A., & Ramesh Kumar, S. B. (2018). A dictionary-based approach to enhance the accuracy of opinion mining on big data. *International Journal of Research and Analytical Reviews, 5*(4), 1836–1844. [Online]. Available: https://ijrar.com/upload_issue/ijrar_issue_20542657.pdf

[13] Huan, J. L., Sekh, A. A., Quek, C., & Hou, Y. (2022). Emotionally charged text classification with deep learning and sentiment semantic. *Neural Computing and Applications, 34*(3), 2341–2351. https://doi.org/10.1007/s00521-021-06542-1

[14] Angelpreethi, A. (2023). Fuzzy based sentiment classification using fuzzy linguistic hedges for decision making. *Mapana Journal of Sciences, 22*(Special Issue 2), 63–79. https://doi.org/10.12723/mjs.sp2.4

[15] Dao, P. Q., Roantree, M., Nguyen-Tat, T. B., & Ngo, V. M. (2024). Exploring multimodal sentiment analysis models: A comprehensive survey. *Preprints*, 2024080127. https://doi.org/10.20944/preprints202408.0127.v1

[16] Angelpreethi, A., & Ramesh Kumar, S. B. (2019). NIC_LBA: Negations and intensifier classification of microblog data using lexicon based approach. *Journal of Emerging Technologies and Innovative Research, 6*(6), 753–759. [Online].Available: https://www.jetir.org/papers/JETIR1906R05.pdf

[17] Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1122–1131). ACM.

[18] Zhang, D., Dai, H., Wang, L., & Chen, L. (2020). Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. *IEEE Access, 8*, 22945–22954. https://doi.org/10.1109/ACCESS.2020.2970431.

[19] Sun, L., Yu, J., Zhang, R., & He, L. (2021). Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge* (pp. 15–20). ACM.

[20] Vasanth, K., Deepa, N., & Karthikeyan, N. (2022). Dynamic fusion of text, video and audio models for sentiment analysis. *Procedia Computer Science, 215*, 211–219.

[21] Garcia-Garcia, J. M., Cernadas, E., & Luaces, O. (2022). Building a three-level multimodal emotion recognition framework. *Multimedia Tools and Applications, 82*(1), 239–269.

[22] Jiang, M., & Ji, S. (2022). Cross-modality gated attention fusion for multimodal sentiment analysis. *arXiv*. https://doi.org/10.48550/arXiv.2208.11893.

[23] Grósz, T., Stappen, L., Baird, A., Schuller, B., & Cummins, N. (2023). Discovering relevant sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients. In *Proceedings of the 4th Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation (MuSe 2023)* (pp. 27–34).

[24] Sun, T., Li, Y., & Zhang, C. (2023). General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 5861–5869). ACM.

[25] Zhang, T., Zhu, Y., Wu, B., Zheng, C., Tan, J., & Xiong, Z. (2025). A general debiasing framework with counterfactual reasoning for multimodal public speaking anxiety detection. *Neural Networks, 187*, 107314. https://doi.org/10.1016/j.neunet.2025.107314.

[26] Yanjing Wang, Kai Sun, Bin Shi, Hao Wu, Kaihao Zhang, Bo Dong, (2026), A guard against ambiguous sentiment for multimodal aspect-level sentiment classification, Information Processing & Management, Volume 63, Issue 2, Part A, 104375, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2025.104375.

[27] Jain, R., Sharma, P., & Kumar, A. (2023). Real-time sentiment analysis of natural language using multimedia input. *Multimedia Tools and Applications, 82*(26), 41021–41036.

[28] Zheng, Y., Zhang, J., Xu, H., & Li, Y. (2024). Djmf: A discriminative joint multi-task framework for multimodal sentiment analysis based on intra- and inter-task dynamics. *Expert Systems with Applications, 242*, 122728.

[29] Ayetiran, E. F., & Özgöbek, Ö. (2024). An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Information Systems, 123*, 102378.

[30] Md. Mithun Hossain, Md. Shakil Hossain, M.F. Mridha, Nilanjan Dey, (2026),A vision-language model for multitask classification of memes, *Neural Networks*, Volume 194, 108089,ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2025.108089.

[31] Lu, Q., Zhang, Z., Li, J., & Xu, K. (2024). Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. *Information Processing & Management, 61*(1), 103538.

[32] Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). A text sentiment analysis method based on a bidirectional long-short term memory neural network. In *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (pp. 1710–1714). IEEE. https://doi.org/10.1109/IMCEC46724.2019.8984028

[33] Adil, M., Wu, J. Z., Chakrabortty, R. K., Alahmadi, A., Ansari, M. F., & Ryan, M. J. (2021). Attention-based STL-BiLSTM network to forecast tourist arrival. *Processes, 9*(10), 1759. https://doi.org/10.3390/pr9101759

[34] Al-alshaqi, M., Rawat, D. B., & Liu, C. (2025). A BERT-Based Multimodal Framework for Enhanced Fake News Detection Using Text and Image Data Fusion. *Computers, 14*(6), 237. https://doi.org/10.3390/computers14060237

[35] Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management, 58*, 102610.

[36] Deng, Y., Li, Y., Xian, S. *et al.* (2024) Mual: enhancing multimodal sentiment analysis with cross-modal attention and difference loss. *Int J Multimed Info Retr* **13**, 31 . https://doi.org/10.1007/s13735-024-00340-w

[37] Jaiswal, R., Singh, U. P., & Singh, K. P. (2021). Fake news detection using BERT-VGG19 multimodal variational autoencoder. In *Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (pp. 1–5). Dehradun, India.

[38] Ying, L., Yu, H., Wang, J., Ji, Y., & Qian, S. (2021). Multi-level multi-modal cross-attention network for fake news detection. *IEEE Access, 9*, 132363–132373.