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 COMPREHENSIVE REVIEW OF HYBRID MULTIMODAL SENTIMENT ANALYSIS IN NLP USING DEEP LEARNING

P Anitha

Assistant Professor, Department of Computer Science, Dr.Umayal Ramanathan College for Women,

Karaikudi, Tamilnadu

Abstract: Sentiment analysis, whether executed within a unimodal or multimodal paradigm, is commonly referred to as opinion mining. It represents a computational methodology used to identify, extract, and quantify subjective information, including perspectives, attitudes, and emotional states. Unlike traditional unimodal sentiment analysis that relies solely on text, multimodal sentiment analysis (MSA) incorporates information from multiple sources—such as speech, tone of voice, facial expressions, and body gestures—to provide a richer and more accurate understanding of emotions. This study presents an extensive survey of research on multimodal fusion techniques and features, with a particular focus on the integration of textual, visual, and audio-visual data. This scholarly investigation explores the historical evolution and theoretical foundations of Multimodal Sentiment Analysis (MSA), examining both current challenges and its benefits. Moreover, this manuscript highlights potential directions for future research, making it a valuable resource for both academic and industry researchers in this domain.

Keywords: Unimodal sentiment analysis, Multimodal sentiment analysis, fusion methodologies, textual, visual, and audio/visual data

1. Introduction

Text sentiment analysis is a type of unimodal sentiment analysis focused on identifying the emotional tone in text data, such as reviews, tweets, or comments. This is relevant to the query, as it deals with sentiment analysis within a single modality—text. Its significance lies in helping businesses and organizations understand public opinion, enabling data-driven decision-making. A foundational perspective on big data visualization, relevant to preprocessing and analysis in sentiment mining was proposed by Angel Preethi and Kumar [1] explore the role of visualization in big data mining, identifying key issues, challenges, and opportunities in creating effective visual representations of large-scale, complex datasets for decision-making and analysis. This paper provides a conceptual overview of visualization techniques, discusses challenges like scalability and data complexity, and highlights opportunities for improving data interpretation in domains such as business analytics and sentiment analysis. In another study by Angel Preethi and Kumar [2] propose a methodological framework for opinion mining (sentiment analysis) targeting text-based data (e.g., social media, reviews), with a pipeline integrating pre-processing, sentiment lexicon scoring, feature extraction, and machine learning classification to extract sentiments (positive, negative, neutral). The authors propose a methodological framework for opinion mining (sentiment analysis) targeting text-based data (e.g., social media, reviews), with a pipeline integrating pre-processing, sentiment lexicon scoring, feature extraction, and machine learning classification to extract sentiments (positive, negative, neutral). While effective for unimodal text analysis, it lacks advanced linguistic nuance handling and multimodal integration.

In a study, Preethi and Kumar [3] propose a lexicon-based approach to enhance sentiment analysis accuracy in large-scale data settings. The study focuses on unimodal (text-only) sentiment analysis, targeting large-scale datasets typically found on social media platforms or in online consumer reviews. The authors proposed a framework for processing textual big data using a dictionary-based sentiment analysis pipeline augmented with machine learning. The dictionary-based approach offers interpretability and scalability for big data, complementing complex deep learning frameworks. This study focuses exclusively on text analysis and depends on the accuracy and comprehensiveness of the sentiment lexicon. However, it may fall short in addressing domain-specific or newly emerging terms and faces challenges in handling sarcasm and negation. In other work, Preethi and Kumar [4] assign frequency-based weights to highlight feature values, integrating them into computer classification to enhance performance. When analyzing common feedback, Dom_classi highlights the significance of words in areas such as film reviews and product feedback. This study struggles with complex linguistic phenomena and does not account for bias in sentiment predictions. In another study, Xu et al. [5] and Adil M et. al. [6] propose a text-based sentiment analysis framework that utilizes a Bidirectional Long Short-Term Memory (BiLSTM) neural network to improve sentiment classification accuracy, especially for applications involving social media data and customer reviews. It employs BiLSTM to capture contextual information from both preceding and succeeding tokens in text sequences, overcoming the limitations of traditional machine learning and unidirectional LSTM models by modeling dependencies in both directions. This paper focuses on unimodal scope, limited contextual nuances, and relies on labeled datasets, which may not generalize to diverse or noisy real-world data. Another study by A. Angel Preethi and Kumar [7] introduces NIC_LBA, a lexicon-based sentiment analysis framework specifically designed for Microblog Data, such as tweets. The study focuses on unimodal (text-only) sentiment analysis, emphasizing the challenges of negations and intensifiers in short, noisy texts. By incorporating rules that account for these linguistic nuances, NIC_LBA improves sentiment classification accuracy, rendering it well-suited for real-time social media monitoring, customer feedback analysis, and opinion mining. Angel Preethi [8] introduced a fuzzy logic framework for sentiment classification that uses linguistic hedges (e.g., “highly,” “moderately”) to enhance the accuracy of sentiment analysis in large text datasets, such as social media posts or product reviews. This study includes fuzzy logic to address the inherent ambiguity in natural language expressions (e.g., “very positive” or “somewhat negative. The framework is designed for text-based datasets, such as social media posts and online reviews, to support informed decision-making in areas like business analytics, customer feedback, and market research. Nonetheless, it faces several challenges, including the time-consuming and domain-dependent task of developing and refining fuzzy rules and hedges, contextual constraints, and a lack of fairness analysis.

In another paper by Angel Preethi and Kumar [9] introduces OPINE_NEG, a method for detecting negations (e.g., “not good”) and intensifiers (e.g., “very good”) in social media text to improve sentiment analysis accuracy, addressing the challenge of linguistic nuances in informal, noisy data. It proposes a hybrid framework combining rule-based techniques for negation/intensifier detection with machine learning for sentiment classification, leveraging social media datasets (e.g., Twitter) and linguistic features like part-of-speech (POS) tagging. Its strengths include improved accuracy and scalability, but it is limited by its unimodal scope and lack of fairness analysis. In another study, Parimala et al. [10] propose a hybrid framework combining Improved Moth Flame Optimization (IMFO) with a Deep Convolutional Neural Network (DCNN) for colorectal cancer classification from biomedical images. IMFO optimizes feature selection and DCNN hyperparameters, enhancing classification accuracy (~85-90%) and robustness to noisy data. While focused on medical imaging, the framework’s optimization techniques could inform feature selection in cross-lingual or multimodal sentiment analysis. The limitations include its unimodal scope and lack of fairness analysis, but it demonstrates machine learning versatility for healthcare applications.

In a survey paper by P.Q. Dao et al. [11], both sentiment analysis and emotion analysis often utilize similar forms of expression. In this context, traditional input devices like the keyboard and mouse are not essential for sentiment analysis to operate. Instead, it relies on innovative modalities such as speech, gestures, messaging, and facial expressions to interpret opinions, emotions, and polarity. These modalities are inherently subjective and can convey a wide range of sentiments, including positive, negative, neutral, joy, and delight. In recent years, sentiment analysis has seen significant research attention, particularly in extracting emotions from voice, text, and facial expressions.

2. Multimodal Sentiment Analysis (MSA)

In today's world, people are increasingly using a combination of text and visual imagery to express their thoughts and emotions. Multimodal Sentiment Analysis (MSA) [11] is an emerging field that aims to analyse and recognize sentiments using data from various modalities. In a study by D. Hazarika et al. [12], the MISA framework was proposed for multimodal sentiment analysis (MSA), utilizing text, audio, and visual data. It effectively captures both shared and modality-specific features, thereby improving sentiment prediction accuracy. The authors utilize modality-specific feature extraction—GloVe or BERT for text, openSMILE for audio, and ResNet or OpenFace for visuals—combined with a two-stream architecture to capture both invariant and modality-specific representations, which are then combined through a fusion mechanism (either concatenation or attention) for sentiment classification. The two-stream representation, which is an invariant stream that captures shared sentiment across modalities using shared subspaces or adversarial training, and the specific stream preserves unique modality features (e.g., text-specific semantics). This study makes an important contribution to Multimodal Sentiment Analysis (MSA) by presenting a new two-stream architecture that enhances performance and enables smoother integration of text, audio, and visual modalities. While the Multimodal Interactive Sentiment Analysis (MISA) addresses the challenges of modality heterogeneity and alignment, it is deficient in bias mitigation, exhibits limited evaluation against adversarial perturbations, and is not explicitly designed to accommodate multilingual contexts.

Zhang et al. [13] present a multimodal sentiment classification through a semi-supervised learning approach. It captures independent knowledge by extracting modality-specific sentiment features (e.g., text semantics, audio tone) through separate processing streams. It also captures interactive knowledge by modeling cross-modal interactions (e.g., text-audio or audio-visual correlations) using techniques such as attention mechanisms or shared subspaces to leverage complementary sentiment cues and improve prediction accuracy. By adopting a semi-supervised learning approach, the framework leverages both labeled and unlabeled data to identify modality-specific (independent) as well as cross-modal (interactive) features. ChatGPT said: It leverages methods such as attention mechanisms or shared subspaces to capture complementary sentiment cues, thereby improving robustness and enhancing generalization capabilities. It employs modality-specific feature extraction, a semi-supervised learning strategy to leverage unlabeled data, addressing the challenge of limited labeled multimodal datasets and a fusion mechanism to model independent and interactive knowledge for sentiment classification. Although this paper successfully addresses data scarcity, it is limited by high computational complexity and insufficient emphasis on fairness and interpretability.

L. Sun et al. [14] proposed a multimodal framework for emotion recognition and sentiment analysis that integrates text, audio, and visual modalities, employing an attention-enhanced recurrent neural network (RNN) to boost performance. It extracts modality-specific features, applies an attention mechanism to highlight important information from each modality, and uses a recurrent model (e.g., LSTM or GRU) to capture temporal patterns for emotion and sentiment prediction. The framework integrates text, audio, and visual modalities to detect sentiments and emotions in multimodal datasets, particularly those involving video-based interactions. By combining attention mechanisms with recurrent neural network architectures, it efficiently models temporal dependencies and highlights the influence of pertinent modalities, overcoming the limitations of static fusion approaches. The paper excels in modality integration and temporal robustness but is limited by high complexity and no fairness analysis.

The paper by K. Vasanth et al., [15] presents a dynamic fusion framework for MSA, effectively integrating text, video, and audio to enhance sentiment prediction on social media. Its adaptive approach outperforms static fusion methods by effectively addressing issues like modality noise and heterogeneity. The framework handles multimodal data through dynamic fusion, where an adaptive mechanism—such as attention or gating—contextually weights and combines modality features based on their relevance (e.g., emphasizing text in text-heavy posts or video in visually rich content). In contrast to static fusion, it adjusts weights according to context, improving robustness. The author introduces a dynamic fusion MSA framework for social media that enhances accuracy through adaptive integration of modalities. Its strengths lie in robustness and broad applicability, though it is limited by high computational complexity.

Garcia et al. [16] proposed a three-level framework for a multimodal emotion recognition approach that fuses textual, acoustic, and visual data using a hierarchical processing approach to predict emotions in multimedia content. The study centers on detecting emotions (such as happiness, sadness, and anger) conveyed through text, audio, and visual modalities—a vital aspect of affective computing due to the complexity of multimedia content on platforms like social media. The hierarchical framework is well-suited for emotion-aware systems, customer feedback analysis, and interactive interfaces, as it functions through three stages: feature extraction, modality-specific modeling, and multimodal fusion. While its major drawback lies in high computational complexity, its structured design, robustness, and broad applicability remain key strengths.

To capture inter-modal interactions with dynamic weighting, M. Jiang et al. [17] proposed a cross-modality gated attention fusion framework for multimodal sentiment analysis (MSA), which leverages text, audio, and visual modalities for sentiment prediction. The approach first extracts modality-specific features and then applies gated attention-based fusion, which dynamically assigns weights to cross-modal relationships, leading to more robust sentiment classification. Its advantages include strong cross-modal modeling, resistance to noise, and suitability for multimedia-rich contexts, though it is hindered by high computational complexity. Using multimodal embeddings from BERT (text), Wav2Vec 2.0 (audio), ELECTRA (text), and Vision Transformer (ViT) (visual), T. Grosz et al. [18] propose a framework designed for recognizing humor and imitated emotions. They use integrated gradients to find relevant subspaces of these embeddings for better performance. It combines pre-trained transformer models for feature extraction with an interpretability technique (integrated gradients) to discover task-relevant embedding sub-spaces, followed by fusion and classification for humor and mimicked emotion tasks in the MuSe 2023 challenge. Despite its complexity, the work stands out for task specificity, interpretability, and use of advanced embeddings.

Sun et al. [19] and Tingting Zhang et. al. [20] proposed a general debiasing framework for multimodal sentiment analysis (MSA), integrating text, audio, and visual modalities to mitigate biases such as gender, cultural, and modality-specific influences, thereby enabling fairer and more accurate sentiment prediction. The approach employs a debiasing strategy—likely through adversarial training—where an adversarial network is trained to strip bias-related features (e.g., gender or cultural cues) from modality representations, ensuring the model prioritizes sentiment-relevant information. Alternatively, reweighting can be applied to adjust the influence of biased samples or features during training, reducing their impact on predictions. This approach is combined with modality-specific feature extraction and fusion to mitigate bias while maintaining sentiment classification performance. The resulting debiased representations are integrated using a fusion strategy, employing attention to dynamically weigh each modality's contribution and concatenation or gating to merge features into a unified representation for sentiment prediction. The strengths lie in its fairness-performance balance, robustness, and applicability to diverse contexts, though computational complexity is a limitation.

A paper authored by R. Jain et. al., [21] focuses on processing text, audio, and visual modalities to detect sentiments in real-time, a critical task in affective computing given the growing volume of multimedia content on platforms like social media. The suggested architecture uses a pipeline that combines feature extraction, deep learning models, and fusion techniques to analyze multimodal data in real-time. In the feature extraction, the textual features are extracted using transformer-based model like BERT, the acoustic features are extracted using wav2vec and the visual features using CNN or pre-trained models like ResNet. The framework is designed for real-time processing without sacrificing accuracy, making it ideal for applications such as live customer feedback analysis, public opinion monitoring, and interactive virtual assistants. Although dataset dependency and computational trade-offs may be drawbacks, its low-latency performance, resilience, and suitability for dynamic settings are its main advantages.

Y. Zheng et al. [22] proposed the Discriminative Joint Multi-Task Framework (DJMF), a novel multimodal sentiment analysis (MSA) approach that integrates text, audio, and visual modalities within a multi-task learning paradigm. The framework models both within-task and cross-task relationships, enabling sentiment classification together with related tasks such as emotion recognition. It employs modality-specific feature extraction, a unified multi-task learning architecture with discriminative modeling, and dynamic task interaction mechanisms. A weakness in the DJMF framework is its computational complexity; however, it is notable for its joint multi-task learning, which uses intra- and inter-task dynamics to enhance sentiment analysis and related tasks.

Ayetrian et al. [23] introduced an inter-modal attention-based deep learning architecture that combines text, visual, and potentially other modalities (e.g., audio or metadata) to detect hate speech, fake news, and abusive language. It delivers strong performance in classification tasks by leveraging an attention mechanism to model interactions across modalities as a unified representation. A major contribution is the inter-modal attention mechanism, which dynamically captures cross-modal relationships, crucial for detecting subtle harmful content. Although computational complexity and dataset dependency may be drawbacks, its dynamic attention mechanism, multitasking abilities, and suitability for social media moderation are its main advantages. Compared to related studies, it provides a flexible solution for harmful content detection while complementing sentiment- and emotion-oriented frameworks.

In another study, Q. Lu et al. [24] introduced the Coordinated-Joint Translation Fusion (CJTF) framework with Sentiment-Interactive Graph Convolutional Networks (SI-GCNs) for multimodal sentiment analysis (MSA). This framework integrates text, audio, and visual inputs for sentiment prediction, combining a translation-based fusion strategy with graph convolutional networks (GCNs) to model sentiment interactions across modalities, thereby achieving improved classification performance. The CJTF framework consists of feature extraction, Coordinated-Joint Translation Fusion, SI-GCNs, and sentiment classification, efficiently processing and fusing multimodal data while emphasizing inter-modal sentiment interactions. The framework's key strengths include its innovative fusion strategy and practical applicability, while its main limitations are high computational complexity and dependence on specific datasets.

Table 1 [11], provides a summary of the datasets, multimodal features (MF), fusion methods (MFM), models, and accuracy for the multimodal research papers reviewed.

Table 1. [11] Datasets, Multimodal Features (MF), Multimodal Fusion Methods (MFM), Models, and Accuracy

S.No.	Study & Year	Datasets	MF	MFM	Models & Accuracy
1	Hazarika et. al. [12], 2020	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	MISA(Acc:83.4,85.5,70.61)
2	Dong Zhang et. al. [13], 2020	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	Bi-modal (Acc: 70.9, F1:70.9), Tri modal(acc:71.2,F1:71.2)
3	Sun et. al. [14], 2021	3rd+OMuSe-CaR	Text+ Visual +Audio	Late Fusion	Temporal model (Acc:0.5549)
4	Vasanth et. al. [15], 2022	Self+NO	Text+ Visual +Audio	Early Fusion	N/A
5	Garcia et. al. [16], 2022	Self+NO	Text+ Visual +Audio	Late Fusion	HERA framework
6	Jiang et. al. [17], 2022	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	CMGA (Acc:53.03)
7	Grosz et. al. [18], 2023	3rd+NO	Text+ Visual +Audio	Late Fusion	Auc:0.8420
8	Sun et. al. [19], 2023	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	GEAR (Acc: 84.39%)
9	Jain et. al. [21], 2023	Self+NO	Text+ Visual +Audio	Late Fusion	MTCNN Model, NLP model, SVM model, Google API

10	Zheng et. al. [22], 2024	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	DJMF framework
11	Ayetrian et. al. [23], 2024	Self+NO	Text+ Visual +Audio	Early Fusion	Acc:0.94
12	Lu et. al. [24], 2024	3rd+O CMU-MOSI,CMU MOSEI	Text+ Visual +Audio	Late Fusion	Sentiment-interactive graph (Acc: 86.5%, 86.1%)

3. Classification and Analysis

3.1 Datasets

The authors utilize various datasets, discussed in this article, with their features summarized in Table 2. [11].

Table 2 Datasets used by the researchers

Dataset	Description	Size / Stats	Labels / Classes
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	Sentiment intensity (-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	Sentiment and emotion intensity
T4SA	Twitter dataset combining textual and visual content (tweets + photos). Neutral class somewhat ambiguous due to annotation quality.	1M tweets, 1.5M photos	Positive, Negative, Neutral
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	Sentiment distribution — Tweets: 46% Positive, 33% Negative, 21% Neutral; Images: 47% Positive, 10% Negative, 43% Neutral.
Fakeddit	Multimodal dataset of fake news from Reddit (text, images, metadata, comments). Offers multiple labeling schemes.	1M records (Mar 2008–Oct 2019)	The dataset supports multiple classification settings: 2-way (Real/Fake), 3-way (Real/Fake/Mixed), and 6-way

Dataset	Description	Size / Stats	Labels / Classes
			(Satire, True, Fake, Misleading, Manipulated, Imposter)
MuSe-CaR	Dataset for MuSe 2021 challenges (text, audio, visual). YouTube automotive reviews with challenging in-the-wild conditions.	291 videos, 70 speakers	N/A
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	Positive, Neutral, Negative
ReactionGIF	Dataset of tweets with GIF replies for two-turn conversations. GIFs mapped to emotions and sentiment.	30,000 tweet-GIF pairs	Reaction category + sentiment & emotion labels

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, 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	Process / Characteristics	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.	- Merge characteristics of all modalities at shallow model layers . - Often requires **dimensionality reduction** to remove redundant data . - Reduced features are fed into the model for further extraction and prediction. - Uses attention mechanisms to give higher weight to important features (e.g., semantic & physical properties of image/text).	- 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.

Method	Description	Process / Characteristics	Advantages	Limitations
Late Fusion (Decision-level Fusion)	Combines outputs from independently trained models on different modalities at the decision stage.	- Each modality is processed by its own model . - Final decision made by combining predictions via **majority voting**, **averaging**, **weighing**, or other decision rules. - Works even if some modalities are missing.	- 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.

3.3 Future Development

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. Models can account for temporal feature interactions and utilize social context factors, such as propagation dynamics and user profiles. They can also use invariant feature learning techniques to better detect biased features and facilitate bias assessment. 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. They can also utilize adversarial learning and knowledge graphs to strengthen unified inter-modal attention mechanisms. These models can also capture complex interrelationships and assess the relative importance of different modalities. The interpretability of emotion recognition across these modalities can be further enhanced through advanced methodologies, cross-modal connections, and filtering mechanisms.

4. 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 some predictions regarding possible developments in the future, such as the application of transfer learning techniques to improve specific model metrics. Furthermore, the fusion process is still quite difficult because of duplicated information across modalities. While many frameworks with optimal classifiers have been proposed, no single model works for all features—their effectiveness depends on the specific context. 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] Angelpreethi, A., Kiruthika, P., & Ramesh Kumar, S. B. (2018). A methodological framework for opinion mining. *International Journal of Computer Sciences and Engineering*, *6*(Special Issue 2), 6–9. [Online]. Available: https://www.ijcseonline.org/pdf_spl_paper_view.php?paper_id=194&NCTT-2018-02.pdf

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

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

[9] 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

[10] Parimala, S., Kumutha, R., Iram, F., Sunena Rose, M. V., N. R., & Angelpreethi, A. (2024, July). Improved moth flame optimization with deep convolutional neural network for colorectal cancer classification using biomedical images. In *Proceedings of [Conference Name]* (pp. 1–6). IEEE. .[Online].Available: https://doi.org/10.1109/icait61638.2024.10690631

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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).

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.