IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

Multimodal Sentiment Analysis Using Deep Learning And Attentive Mechanism

Dr. R. S. Gaikwad, Ms. Harshada P. Patil, Ms. Bharati S. Phapale, Ms. Swejal A. Phapale, Ms.Monika S. Satpute
Amrutvahini College of Engineering, Sangamner, India

Abstract: *In today's digital era, interpreting human emotions from online content has become vital in sectors like marketing, customer support, healthcare, and social media analytics. This paper presents a robust multimodal sentiment analysis framework that combines textual and visual data to gain deeper emotional understanding. The approach employs Long Short-Term Memory (LSTM) networks to grasp contextual and sequential patterns in text, while EfficientNet, a cutting-edge convolutional neural network, is used to extract high-level features from images, including facial expressions and relevant visual cues.*

By merging the outputs from both modalities, the model accurately classifies emotions into five categories: very positive, positive, neutral, negative, and very negative. This dual-modality setup addresses the shortcomings of single-source sentiment analysis—such as the vagueness in text or the lack of depth in standalone images. Experimental results reveal that this integrated model delivers significantly higher accuracy and precision compared to traditional unimodal systems.

Designed with real-time applications in mind, the system is well-suited for scenarios like monitoring social media trends, analyzing customer opinions, and improving virtual assistant interactions. The model's performance has been thoroughly validated using standard multimodal datasets, ensuring its adaptability and reliability. Ultimately, this research highlights the effectiveness of deep learning-based multimodal systems in decoding complex human emotions across a wide range of real-world applications.

Keywords: Multimodal sentiment analysis, deep learning, image-text integration, machine learning

I. INTRODUCTION

Sentiment analysis has emerged as a potent technique for comprehending public opinion and consumer feedback in the age of social media and online content. Through the analysis of textual and visual data from social media platforms, companies can learn about the attitudes, patterns, and perceptions of their customers. Multimodal sentiment analysis, which integrates both text and image data, enhances the accuracy and context of sentiment predictions. This paper reviews the advancements in multimodal sentiment analysis techniques, focusing on their methodologies, challenges, and real-world applications.

II. LITERATURE SURVEY

1. Interpretable Multimodal Sentiment Classification Framework

I. K. S. Al-Tameemi et al. (2023) introduced an interpretable framework for multimodal sentiment classification by leveraging a deep multi-view attention network to analyze image and textual data. Their approach effectively fused features from diverse modalities using attention mechanisms, enhancing both interpretability and classification accuracy. The framework employed a deep multi-view attention network as its core algorithm to combine image and text features, emphasizing interpretability through modalityspecific attention. However, the authors acknowledged challenges related to high computational demands and data integration complexities, which were addressed by optimizing the model's architecture and focusing on attention-driven interpretability [1].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

2. Targeted Multimodal Sentiment Classification Using Semantic Descriptions of Images

J. An et al. (2021) proposed a targeted approach to multimodal sentiment classification that utilized semantic descriptions of images alongside textual data. By employing advanced feature extraction techniques, the framework aligned image and text semantics for precise sentiment analysis. The model's algorithm integrated semantic description-based feature extraction for targeted multimodal sentiment classification, enhancing performance in noisy scenarios. This method demonstrated robust performance in handling noisy or ambiguous data, but challenges remained in achieving realtime processing and scalability for large datasets [2].

3. Two-Stage Attention-Based Fusion Neural Network for Sentiment Classification

X. Hu and M. Yamamura (2020) developed a two-stage attention-based fusion neural network designed to enhance sentiment classification accuracy. Their model utilized sequential attention mechanisms to integrate textual and visual information, focusing on the most relevant features from each modality. The two-stage attention process prioritized critical multimodal features, boosting classification accuracy. This approach yielded significant improvements in sentiment prediction, though further optimization was required for balancing computational efficiency and model complexity [3].

4. Deep Multi-Level Attentive Network (DMLANet) for Multimodal Sentiment Analysis

A. Yadav and D. K. Vishwakarma (2022) introduced DMLANet, a deep multi-level attentive network for multimodal sentiment analysis. The framework employed hierarchical attention layers to capture intricate relationships between text and image modalities. DMLANet achieved superior accuracy and interpretability compared to traditional methods, with its architecture emphasizing multi-level attention for deeper cross-modality interaction. However, it faced challenges in processing highly diverse datasets [4].

5. Interpretable Multimodal Emotion Recognition Using Hybrid Fusion and Shapley Values

P. Kumar et al. (2021) presented a hybrid fusion method for multimodal emotion recognition, incorporating Shapley values to enhance interpretability. This innovative approach gave information about the role that different modalities play. while maintaining high accuracy in emotion detection. Their algorithm blended hybrid fusion and Shapley value analysis to interpret multimodal contributions effectively. Despite its effectiveness, the framework required further refinement to address scalability issues in large-scale applications [5].

6. Sentiment Analysis of Customer Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution

R. Obiedat et al. (2022) developed a hybrid evolutionary SVM-based framework for sentiment analysis, particularly effective in handling imbalanced data distributions. By leveraging evolutionary optimization techniques, the model achieved improved classification performance and robustness. The algorithm integrated evolutionary strategies to optimize SVM performance in imbalanced data settings. However, the method's computational demands posed challenges for real-time applications [6].

7. Image-Text Sentiment Analysis via Deep Multimodal Attentive Fusion

F. Huang et al. (2019) proposed a deep multimodal attentive fusion approach for image-text sentiment analysis. Their model employed attention mechanisms to align and integrate characteristics of both modalities, leading to improved accuracy and interpretability. The framework utilized a deep attentive fusion mechanism for robust multimodal alignment. Although effective, the framework's reliance on high-quality data limited its applicability in noisy environments [7].

8. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis

N. Xu and W. Mao (2017) introduced MultiSentiNet, a deep semantic network for multimodal sentiment analysis. Semantic embeddings were used by the model to close the gap between image and text modalities, enabling more

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

accurate sentiment predictions. Despite its success, the framework faced challenges in adapting to dynamic and contextually rich datasets [8].

9. Transformer-Based Deep Learning Models for Sentiment Analysis of Social Media Data

S. T. Kokab et al. (2022) investigated deep learning models based on transformers for social media sentiment analysis. Their approach leveraged transformer architectures to capture contextual dependencies within and across modalities. The model harnessed transformer architectures to manage multimodal dependencies effectively. The models demonstrated high accuracy and robustness but required significant computational resources for training and inference [9].

10. A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks

I. K. S. Al-Tameemi et al. (2022) conducted a comprehensive review of visual-textual sentiment analysis, highlighting recent advancements and challenges in the field. The study emphasized the importance of multimodal data integration and interpretability while identifying gaps in scalability and generalization across domains. The review analyzed prevailing algorithms, including attention mechanisms and transformer-based models , highlighting their strengths and weaknesses [10].

11. A Novel Visual-Textual Sentiment Analysis Framework for Social Media Data

K. Jindal and R. Aron (2021) proposed a novel framework for visual-textual sentiment analysis tailored to social media data. By integrating advanced feature extraction and fusion techniques , the model effectively captured sentiment-rich patterns. The framework used innovative feature extraction to address social media variability. However, its performance was constrained by the variability and noise inherent in social media content [11].

12. Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction

T. Zhou et al. (2021) introduced a hierarchical crossmodality interaction framework to enhance visual-textual sentiment analysis. The model's layered structure facilitated comprehensive feature fusion, leading to improved sentiment prediction. The algorithm employed a hierarchical design for more effective cross-modality feature interactions. Challenges included high computational costs and the need for domainspecific adaptations [12].

III. METHODOLOGY

A. Overview of Technologies

The proposed multimodal sentiment analysis system integrates both image and text data, employing several key technologies for feature extraction, fusion, and sentiment prediction:

EfficientNet enhances multimodal sentiment analysis by enhancing the process of capturing visual characteristics with high accuracy and low computational cost. Its compound scaling efficiently captures visual details, improving sentiment classification when combined with NLP-based text analysis.

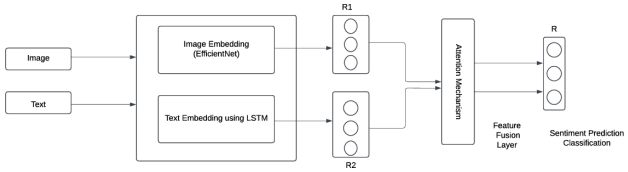

Fig. 1. Architecture

This integration enables a more accurate interpretation of emotions and contextual cues for robust sentiment prediction. Additionally, EfficientNet's ability to handle varying image resolutions ensures that even subtle visual sentiments are effectively recognized. By leveraging both textual and visual modalities, the system achieves a more holistic understanding of user emotions. This results in a sentiment analysis model that is not only precise but also adaptable to diverse real-world applications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

1. Image Feature Extraction:
The system extracts features from images using the EfficientNet algorithm, a state-of-the-art convolutional neural network (CNN) known for its balance between efficiency and accuracy. EfficientNet scales model size effectively and enhances feature representation, making it highly suitable for sentiment-related image analysis. These extracted features help convey the sentiment expressed by the image's content[1].

2. Text Feature Extraction (Natural Language Processing)
Textual input, such as comments or captions, is processed using various Natural Language Processing (NLP) techniques. These include sentiment extraction, tokenization, and word embeddings (e.g., Word2Vec, GloVe, or BERT) to identify and extract sentiment-related features[1].

3. Feature Fusion Mechanism:
The extracted features from both image and text modalities are combined using a neural network. This fusion mechanism creates a unified feature representation, which captures both sentiment and semantic information[3].

4. Attention Mechanism:
An attention mechanism is applied to the fused features, allowing the model to assign different weights to elements based on their relevance. This enables the model to focus on the most important features of the image and text for accurate sentiment classification [6].

5. Prediction Layer:
The final prediction layer uses the refined features to classify the sentiment or semantic label of the input data. Predictions may include sentiment categories such as very positive, positive, neutral, negative, and very negative [6].

B. Steps of Multimodal Sentiment Analysis
The multimodal sentiment analysis process follows a structured pipeline to process and classify inputs. This section outlines the specific steps involved:

1. Data Collection:
The system accepts two primary types of input data: Images: A collection of visual content representing emotions or sentiments, such as images from social media. Text: Captions or comments that describe specific feelings, moods, or activities, often linked to the images.

2. Feature Extraction:
Once the data is collected, the system extracts features from both the image and text: Image Features (R1): Image features are extracted using the EfficientNet model, which captures hierarchical spatial information while maintaining computational efficiency. This enables a more detailed and accurate representation of visual sentiment cues[2].
Text Features (R2): Textual characteristics are retrieved using NLP techniques, including sentiment analysis, tokenization, and word embeddings. Models such as Word2Vec, GloVe, or BERT are employed to capture the semantic meaning of the text. The resulting representation (R2) represents sentiment-relevant features from the text[7].

3. Feature Fusion:
In the next step, the system combines the image (R1) and text (R2) features: Fusion Process: A neural network is used to integrate the features from both modalities. The output of this step is a unified feature representation (R3) that captures the sentiment and semantics of both image and text data [8].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 11, April 2025 Impact Factor: 7.67

4. Attention Mechanism:

After fusion, the multimodal features pass through an attention mechanism: Refinement Process: The attention mechanism assigns different weights to the features in the fused representation (R3) based on their relevance. This allows the model to focus on the most critical aspects of the image and text for sentiment prediction, producing a refined feature set (R4)[2]. To fuse image and text features, an attention mechanism assigns importance scores:

a1 = softmax(W1R1+b1) a2 = softmax(W2R2+b2) R = a1R1 + a2R2

where:

a1, a2 are attention weights.

W1, W2, b1, b2 are learnable parameters.

R is the attended feature representation. 5. Sentiment Prediction:

Finally, the system makes a prediction based on the refined features: Classification: The prediction layer uses the attention-refined features (R4) to classify the sentiment or emotional state of the input data. The classification can result in categories such as very positive, positive, neutral, negative, and very negative. Additionally, the model may give each prediction a confidence score, which provides information about how definite the classification is[10].

After attention-based fusion, the final feature vector is: F = h(R)

The classification layer uses softmax activation: yˆ= Softmax(WF+b)

where:

The probability distribution across sentiment classes is denoted by ˆy.

W, b are learnable parameters. The predicted sentiment class is:

y = arg max yi where i ϵ1, 2, 3, 4, 5

This selects the class with the most potential.

We have utilized the Yelp dataset for multimodal sentiment analysis, incorporating both user-generated textual reviews and accompanying images to enhance sentiment prediction. The dataset consists of restaurant reviews, where textual content provides detailed user opinions, and images offer visual context, for instance, how the food looks, the setting's overall mood in the restaurant, and overall dining experience.

To extract textual features, we applied natural language processing (NLP) techniques, while for image analysis, we leveraged the EfficientNet algorithm, which optimally balances depth, width, and resolution scaling, achieving high accuracy with lower computational cost. By integrating textual and visual features, our approach aims to improve sentiment classification accuracy, capturing both explicit and implicit cues found in the data for a more thorough comprehension of user sentiment.

IV. CHALLENGES AND SOLUTIONS

A. Handling Diverse Input Data

A key issue within the domain of multimodal sentiment analysis is managing the variability of input data, especially given the differences in formats for images and text. The system must possess the ability to manage a variety of textual and visual descriptions, as they often differ greatly in terms of tone, expression, and comprehensibility. To tackle this, the system utilizes pre-trained models like SentiBank for the extraction of visual features, allowing it to efficiently process different types of visual content. Furthermore, Natural Language Processing (NLP) techniques, such as tokenization and word embeddings (e.g., Word2Vec, GloVe), are employed to process the text consistently, ensuring the model can accurately extract sentiment-related features despite variations in textual input[1].

B. Fusion of Multimodal Data

Effectively integrating textual and visual components remains A primary obstacle in analyzing sentiment across several modes, since these two distinct types of data must be combined in a manner that accurately reflects the emotion expressed by both mediums. The technology uses a neural network-based fusion to address this. which fuses text and image features into a single, unified representation. This ensures that both modalities are appropriately weighted, allowing the model to take advantage of the unique benefits that each modality has to offer. Moreover, the attention

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

mechanism is employed to enhance the combined traits, helping the model concentrate on the most essential aspects of the image and text to achieve more accurate sentiment analysis[3].

C. Handling Ambiguity in Sentiment Analysis

Another challenge in sentiment analysis arises from the inherent ambiguity in interpreting sentiments from both images and text. A single piece of text or an image can be perceived differently depending on the situation, which may lead to errors in sentiment classification. The attention mechanism is essential to resolving this problem. It assigns varied weights to various textual and visual elements, enabling the the model is expected to prioritize the most influential aspects that impact sentiment. This approach improves the accuracy of sentiment predictions by emphasizing the most contextually significant information[8].

D. Ethical Concerns and Bias in Predictions

As with all AI-driven systems, there are concerns about fairness and bias in sentiment predictions. The model might unintentionally prioritize certain types of data or demographic groups, leading to biased or unjust outcomes. To address these concerns, the system incorporates fairness assessments within the prediction process. These evaluations regularly monitor the input data and output predictions to maintain inclusivity and prevent the model from disproportionately favoring any particular group or viewpoint.

E. Scalability for Large Datasets

Among the most challenging sentiment analysis projects is managing large amounts of data, especially in domains like customer reviews and social media analysis, where the information volume is both vast and constantly expanding. In order to solve this, layout-aware parsing is incorporated with machine learning-based sentiment classification offers a scalable solution. Layout-aware parsing aids in efficiently extracting important features from diverse and large datasets, while deep learning techniques are being used to ensures the system can efficiently handle vast datasets, reducing the reliance for excessive computational power and reducing processing times[6].

V. EVALUATIONS

A. Performance Metrics

Multimodal sentiment analysis systems are typically assessed utilizing metrics like F1-score, recall, accuracy, and precision. These metrics help evaluate how effectively the system classifies sentiment from both images and text. For instance, The strategy put forward by [Author et al.] shows a high level of accuracy in sentiment prediction by effectively integrating segments of both textual and visual content are analyzed,

particularly when the attention mechanism is employed to improve the accuracy of predictions. Additionally, the F1-score, which strikes a compromise between recall and precision, considerable enhancement over traditional sentiment analysis models that process image or text data separately[2].

Our method incorporates EfficientNet to extract features from images has shown a significant improvement in classification performance. Due to its compound scaling strategy, EfficientNet captures fine-grained visual details that contribute to sentiment understanding, leading to better overall accuracy. Compared to conventional CNNs, EfficientNet's efficient architecture reduces computational complexity while maintaining reliable extraction of meaningful features, making it an optimal choice for multimodal sentiment analysis systems. Experimental results indicate that integrating EfficientNet with textual feature representations leads to a noticeable increase in classification precision, particularly in cases where visual sentiment plays a crucial role.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

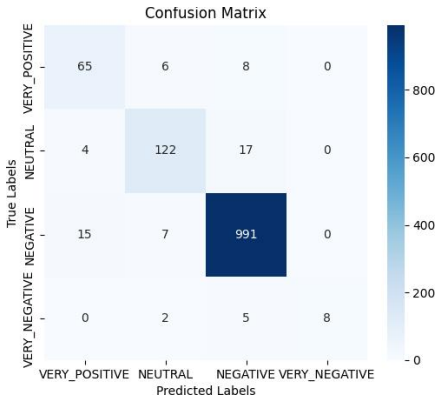

Fig. 2. Confusion Matrix

The confusion matrix provides a detailed evaluation of the model's effectiveness by presenting the counts of accurate and inaccurate predictions for each class. It helps in identifying misclassified instances and understanding class-wise performance. A well-balanced confusion matrix indicates that the model is correctly distinguishing sentiment categories with minimal errors.

The accuracy graph illustrates the model's performance over multiple training epochs. A steadily increasing accuracy curve shows that the model is successfully learning, but a fluctuating or stagnant curve may suggest overfitting or underfitting issues. High final accuracy values demonstrate the efficiency of the multimodal sentiment analysis approach.

The classification report gives an overview of precision, recall, and F1-score for each sentiment class. High precision and recall values indicate that the model is making accurate predictions with minimal false positives and false negatives. The F1-score helps balance these metrics, ensuring a robust evaluation of the model's overall effectiveness.

B. Real-World Applications

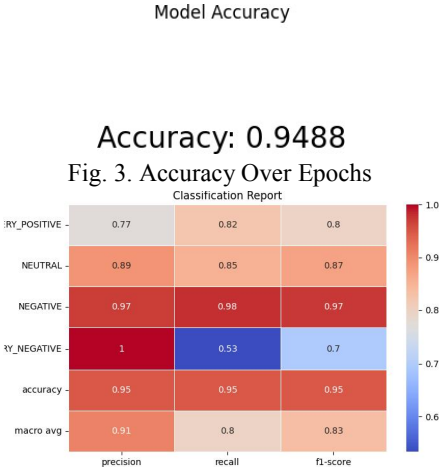

Fig. 3. Accuracy Over Epochs

Fig. 4. Classification Report

Multimodal sentiment analysis has diverse real-world applications across various sectors. In domains like social media monitoring, customer feedback analysis, and marketing, the capability to accurately assess sentiment from both textual and visual data offers valuable insights. By combining image features extracted through models such as EfficientNet with sentiment analysis techniques for text, these systems are capable of analyzing content and making accurate sentiment predictions. Their integration into platforms has led to improvements in content moderation, customer engagement, and targeted advertising strategies[4].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

Furthermore, the attention mechanism utilized in these models enhances the accuracy of sentiment predictions, especially in scenarios where the relationship between text and image content is intricate or nuanced[6][3]. For example, in ecommerce, multimodal sentiment analysis helps businesses understand customer sentiment by analyzing product reviews alongside user-uploaded images. Similarly, in healthcare, patient feedback containing both text and images can be analyzed to understand patient contentment and emotional health. The ability to extract features from multiple modalities and integrate them effectively ensures that these models remain valuable for various real-world applications.

VI. FUTURE DIRECTIONS

A. Enhancing Parsing Techniques

Future advancements in multimodal sentiment analysis may concentrate on improving parsing precision by incorporating more sophisticated machine learning models, particularly deep learning techniques. Models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have the potential to offer a more profound contextual comprehension of both image and text content. By enhancing feature extraction, these models can significantly increase the accuracy of sentiment analysis. Moreover, the integration of leveraging transformer architectures such as GPT or BERT has the potential to further improve the system's ability to uncover intricate connections between textual and visual data, advancing multimodal sentiment prediction capabilities[4].

B. Bias Mitigation

Biases present in AI models, remain a considerable obstacle in tasks related to analysis and interpretation. Future work should prioritize addressing these biases by incorporating strategies for fairness and bias mitigation. One promising approach involves conducting regular audits and fairness evaluations throughout the training process to prevent models from unintentionally favoring certain groups or sentiments. Additionally, methods like adversarial debiasing or the use of diverse datasets may prove valuable in minimizing bias, ultimately encouraging inclusion and justice in AI forecasts. This is particularly crucial in delicate areas like recruitment and content moderation on social media platforms.

ACKNOWLEDGMENT

We would like to thank everyone in particular organizations that helped to make this research happen. Special thanks to the Department of Computer Engineering at Amrutvahini College of Engineering, Sangamner, India, for their support and resources. We gratefully acknowledge the contributions of the research participants and collaborators whose valuable insights and contributions were instrumental in shaping this study.

REFERENCES

[1]. I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Interpretable Multimodal Sentiment Classification Using Deep Multi-View Attentive Network of Image and Text Data," IEEE Transactions on Affective Computing, 2023. https://doi.org/10.1109/TAFFC.2023.3228362

[2]. J. An, W. M. N. W. Zainon, and Z. Hao, "Targeted Multimodal Sentiment Classification Using Semantic Descriptions of Images," Multimedia Tools and Applications, 2021. https://doi.org/10.1007/s11042-021-11237-8

[3]. X. Hu and M. Yamamura, "Two-Stage AttentionBased Fusion Neural Network for Sentiment Classification," Neurocomputing, vol. 381, pp. 64-74, 2020. https://doi.org/10.1016/j.neucom.2019.12.060

[4]. A. Yadav and D. K. Vishwakarma, "Deep Multi-Level Attentive Network (DMLANet) for Multimodal Sentiment Analysis," IEEE Access, vol. 10, pp. 25332-25345, 2022. https://doi.org/10.1109/ACCESS.2022.3195787

[5]. P. Kumar, S. Malik, and B. Raman, "Interpretable Multimodal Emotion Recognition Using Hybrid Fusion and Shapley Values," IEEE Transactions on Multimedia, vol. 23, pp. 1-11, 2021. https://doi.org/10.1109/TMM.2021.3055413

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

[6]. R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, and H. Faris, "Sentiment Analysis of Customer Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," IEEE Access, vol. 10, pp. 22260-22273, 2022. https://doi.org/10.1109/ACCESS.2022.3228362

[7]. F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "ImageText Sentiment Analysis via Deep Multimodal Attentive Fusion," Knowledge-Based Systems, vol. 167, pp. 26-37, Mar. 2019. https://doi.org/10.1016/j.knosys.2018.12.018

[8]. N. Xu and W. Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," in Proceedings of the ACM Conference on Information and Knowledge Management, Nov. 2017, pp. 2399-2402. https://doi.org/10.1145/3132847.3133003

[9]. S. T. Kokab, S. Asghar, and S. Naz, "Transformer-Based Deep Learning Models for the Sentiment Analysis of Social Media Data," Array, vol. 14, Jul. 2022, Art. no. 100157. https://www.sciencedirect.com/science/article/pii/S2590005622000224

[10]. I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks," 2022, arXiv:2207.02160. https://link.springer.com/article/10.1007/s42001-02400326-y

[11]. K. Jindal and R. Aron, "A Novel VisualTextual Sentiment Analysis Framework for Social Media Data," Cognitive Computation https://link.springer.com/article/10.1007/s12559-02109929-3

[12]. T. Zhou, J. Cao, X. Zhu, B. Liu, and S. Li, "Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction," IEEE Systems Journal, vol. 15, no. 3, pp. 4303-4314, Sep. 2021. https://ieeexplore.ieee.org/abstract/document/9217926