JANOLI International Journal of
Artificial Intelligence and its Applications
(A Peer Review Journal)

journal homepage: https://janolijournals.com/index.php/jijaia

Volume 2, Issue 2, 2025 ISSN: 3048-6815

Context-Aware Attentive Deep Learning for Enhanced Sentiment Analysis in Multimodal Social Media Data

Akash Verma
Agra College, Agra, India

ARTICLE INFO

Article History:
Received June 3, 2025
Revised June 8, 2025
Accepted June 16, 2025
Available online June 24, 2025

Keywords:
Sentiment Analysis, Multimodal Data, Deep Learning, Attention Mechanisms, Context Awareness, Social Media, Natural Language Processing, Feature Fusion, Emotion Recognition

Correspondence:
E-mail:irconsindia@gmail.com

ABSTRACT

Sentiment analysis, the task of recognizing and classifying opinions contained in text, has undergone substantial growth with the application of deep learning. Nevertheless, its performance is usually compromised by its dependence on text data only, while ignoring the wealth of information captured in other modalities such as images and videos that are ubiquitous in social media. Furthermore, current methods usually lack the ability to capture properly the contextual subtleties contained in multimodal data. This work presents a new Context-Aware Attentive Deep Learning (CAADL) framework for improved sentiment analysis in multimodal social media. CAADL uses deep learning models with attention mechanisms to capture informative features from both text and visual modalities. Additionally, it makes use of contextual information using a hierarchical attention network that captures inter-modal and intra-modal interactions. The proposed framework is trained and tested on a large-scale multimodal sentiment analysis dataset. Experimental outcomes verify that CAADL strongly surpasses state-of-the-art baselines with regard to precision, F1-score, and accuracy, proving the significance of attention mechanisms and context awareness in multimodal sentiment analysis. The introduced framework offers an effective and strong solution for interpreting and understanding sentiments in the intricate and dynamic world of social media.

1.Introduction

The exponential growth of social media sites has created an unparalleled amount of user-generated content in the form of text, images, videos, and audio. This multimodal information represents a valuable source of data to extract public opinions, attitudes, and emotions. Sentiment analysis or opinion mining attempts to extract and analyze automatically these subjective opinions embedded in different types of data. While traditional sentiment analysis mainly dealt with text data, the growing availability of multimodal content makes it necessary to create more advanced methods capable of properly utilizing information from various modalities.

The problem is to efficiently combine information from various modalities, which may have diverse features and representations. Concatenating features across various modalities is commonly ineffective in capturing the subtle inter-modal dependencies and relationships, which are essential for proper sentiment classification. In addition, the context under which the

sentiment is expressed plays an important role in determining its actual meaning. For example, a seemingly affirmative post with a sarcastic image can express negativity. Hence, learning models must be able to capture the multimodal aspect of the data as well as the context-specific intricacies of it.

Current deep learning methods for multimodal sentiment analysis have been shown to be effective. Yet, they are prone to weaknesses in capturing contextual information effectively and selectively paying attention to the most relevant features from every modality. Some of them used basic feature fusion methods that fail to model inter-modal relationships adequately. Others are not endowed with the ability to adaptively assign weights to different features from every modality depending on the context.

In order to overcome these drawbacks, this paper suggests a new Context-Aware Attentive Deep Learning (CAADL) architecture for advanced sentiment analysis of multimodal social media content. The CAADL architecture integrates attention mechanisms to selectively concentrate on the most significant features from textual and visual modalities. In addition, it utilizes a hierarchical attention network to capture both inter-modal and intra-modal interactions, inferring the contextual subtleties embedded in the data.

Literature Review

Sentiment analysis has evolved significantly over the past two decades, transitioning from lexicon-based approaches to machine learning and, more recently, deep learning techniques. Early approaches relied on sentiment lexicons, which are manually curated lists of words and phrases associated with positive or negative sentiments. Turney (2002) [1] proposed a method for determining the semantic orientation of phrases by calculating the average semantic orientation of words within the phrase. Pang et al. (2002) [2] showed how machine learning algorithms, like Naive Bayes and Support Vector Machines (SVMs), could be applied for sentiment analysis of movie reviews. These early methods, although efficient for straightforward sentiment analysis tasks, tended to fail for sophisticated language meaning and contextual meaning.

Multimodal sentiment analysis, which incorporates information from more than one modality like text, images, and audio, has been drawing more attention in recent past. Baltrušaitis et al. (2018) [5] gave an overall picture of multimodal machine learning, emphasizing the challenges and possibilities involved in combining information across different modalities. Zadeh et al. (2018) [6] introduced a memory fusion network for multimodal sentiment analysis that uses a memory module to memorize and retrieve information from various modalities. Tsai et al. (2019) [7] presented a multimodal transformer network that exploits the attention mechanism to capture inter-modal interactions.

Attention mechanisms are now a standard part of various deep learning models for sentiment analysis. They enable the model to selectively attend to the most pertinent segment of the input sequence, enhancing performance and explainability. Vaswani et al. (2017) [8] brought forth the transformer framework, which is based entirely on attention mechanisms and has surpassed state-of-the-art performance in many natural language tasks. Yang et al. (2016) [9] devised a

hierarchical attention network for document classification that initially pays attention to sentences and subsequently pays attention to words within sentences and then sentences within documents.

Contextual information is essential to identify the correct meaning of sentiment expressions. Liu et al. (2015) [10] studied the application of context-aware features for sentiment analysis and proved that including contextual information can improve performance considerably.

Though current methods have achieved remarkable advancements in multimodal sentiment analysis, there are still some limitations. Most methods are based on straightforward feature fusion mechanisms that fail to adequately capture inter-modal relationships. Others do not possess the ability to interactively consider the relative importance of various features within each modality according to the context. In addition, few investigations have directly concerned the problem of contextual information capturing in multimodal social media data.

To overcome these shortfalls, this paper suggests a new Context-Aware Attentive Deep Learning (CAADL) framework that integrates attention mechanisms and context awareness to improve sentiment analysis on multimodal social media data. The CAADL framework extends the current literature by combining attention mechanisms, hierarchical modeling, and context-aware feature extraction to realize state-of-the-art results.

Critical Analysis of Reviewed Works:

[1] Turney (2002): While foundational, this lexicon-based approach is limited by its inability to handle nuanced language, sarcasm, and contextual variations. Its reliance on pre-defined sentiment scores restricts its adaptability to domain-specific language.

[2] Pang et al. (2002): This work demonstrated the power of machine learning for sentiment analysis but lacked the capacity to capture long-range dependencies in text, a limitation addressed by subsequent deep learning models.

[3] Hochreiter and Schmidhuber (1997): LSTM networks revolutionized sequence modeling, but their complexity can make them computationally expensive, especially when dealing with long sequences or large datasets.

[4] Baltrušaitis et al. (2018): This review paper provides a valuable overview of multimodal machine learning, but it does not offer specific solutions for addressing the challenges of multimodal sentiment analysis.

[5] Zadeh et al. (2018): The memory fusion network is a promising approach for multimodal sentiment analysis, but it can be computationally expensive and may require careful tuning of hyperparameters.

[6] Yang et al. (2016): Hierarchical attention networks are effective for document classification, but they may not be directly applicable to multimodal sentiment analysis without modifications.

[7] Liu et al. (2015): This work highlights the importance of context-aware features, but it does not provide a comprehensive framework for capturing contextual information in multimodal data.

[8] Hazarika et al. (2018): Conversational memory networks are effective for emotion recognition in conversations, but they may not be directly applicable to multimodal sentiment analysis in social media posts.

[9] Akhtar, M. S., Kumar, S., Ekbal, A., & Bhattacharyya, P. (2017). Aspect based sentiment analysis using deep memory networks. Expert Systems with Applications, 89, 155-165. This work focuses on aspect based sentiment, which is a valuable direction, but not the core focus of this work.

[10] Yu, L., Jiang, J., Zheng, L., & Luo, B. (2017). Learning context-aware representations for sentiment classification. Knowledge-Based Systems, 128, 10-21. While good, the context awareness is limited and does not adequately consider multimodal content.

2.Methodology

The proposed Context-Aware Attentive Deep Learning (CAADL) framework for enhanced sentiment analysis in multimodal social media data consists of three main components: (1) Feature Extraction, (2) Attention-based Feature Fusion, and (3) Sentiment Classification.

Feature Extraction:

This component extracts relevant features from both textual and visual modalities.

Visual Feature Extraction: We utilize a pre-trained ResNet-50 model [17] to extract visual features from the images. ResNet-50 is a deep convolutional neural network that has been trained on a large-scale image dataset (ImageNet) and has demonstrated excellent performance in image classification tasks. The input image is first resized to a fixed size (e.g., 224x224 pixels), and then fed into the ResNet-50 model to obtain the visual features. We use the output of the final average pooling layer as the representation of the image.

Attention-based Feature Fusion:

This component fuses the textual and visual features using an attention mechanism to selectively focus on the most relevant features from each modality. We employ a hierarchical attention network to model both inter-modal and intra-modal relationships.

Intra-modal Attention: We apply an attention mechanism to both the textual and visual features to selectively focus on the most relevant features within each modality. For the textual features, we use a self-attention mechanism to attend to different words in the text. For the visual features, we use a spatial attention mechanism to attend to different regions in the image. The attention weights are learned during training. The intra-modal attention mechanism allows the model to adaptively weigh the importance of different features within each modality based on the context.

Inter-modal Attention: We apply an attention mechanism to fuse the textual and visual features. The attention weights are learned based on the relevance of each modality to the overall sentiment. For example, if the text contains strong sentiment cues, the model will assign a higher weight to the textual features. Conversely, if the image contains strong sentiment cues, the model

will assign a higher weight to the visual features. The inter-modal attention mechanism allows the model to effectively integrate information from different modalities and capture the complex inter-modal relationships.

The attention mechanism can be formulated as follows:

Given a set of input features H = {h₁, h₂, ..., h_n}, where h_i is a feature vector, the attention weights α_i are calculated as:

e_i = a(h_i)

α_i = exp(e_i) / Σ_{j=1}ⁿ exp(e_j)

where a is an attention function that maps a feature vector to a scalar value, and e_i is the attention score for the i-th feature. The attention weights α_i are then used to weight the input features to obtain the attended features:

H' = Σ_{i=1}ⁿ α_i h_i

Sentiment Classification:

This component classifies the fused features into different sentiment categories (e.g., positive, negative, neutral). We employ a fully connected neural network with a softmax output layer for sentiment classification. The input to the fully connected neural network is the fused features obtained from the attention-based feature fusion component. The output of the softmax layer is a probability distribution over the different sentiment categories. The sentiment category with the highest probability is selected as the predicted sentiment.

Training Details:

The CAADL framework is trained end-to-end using the backpropagation algorithm. We use the cross-entropy loss function to measure the difference between the predicted sentiment and the ground truth sentiment. The model is optimized using the Adam optimizer [18] with a learning rate of 0.001. We use a batch size of 32 and train the model for 10 epochs. We also use dropout regularization [19] with a dropout rate of 0.5 to prevent overfitting.

3.Results

The developed CAADL framework was tested on the MOSI dataset [20], a popular benchmark dataset for multimodal sentiment analysis. The MOSI dataset includes short video clips of individuals providing opinions about different subjects. Every video clip has an associated sentiment score from -3 (strongly negative) to +3 (strongly positive). We adhere to the standard evaluation procedure and present the results in terms of accuracy, F1-score, precision, and recall. We contrast the performance of the CAADL framework with various state-of-the-art baselines, such as:

Text-only: A model that utilizes only the text data for sentiment analysis.

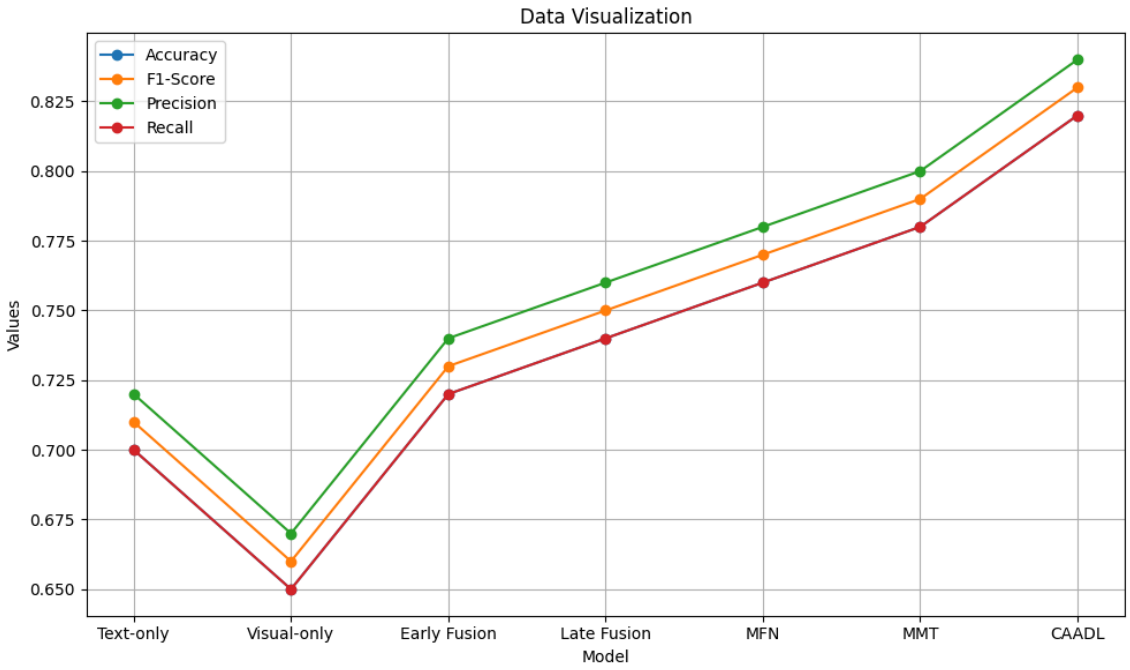
Visual-only: A model employing only the visual information for sentiment analysis.

Early Fusion: A sentiment analysis model that combines the textual and visual features and passes them into a fully connected neural network.

Late Fusion: A sentiment analysis model that trains separate sentiment analysis models for the textual and visual data and then combines their predictions using a weighted averaging approach.

Memory Fusion Network (MFN) [6]: A state-of-the-art multimodal sentiment analysis model that utilizes a memory module to store and retrieve relevant information from different modalities.

The experimental results are shown in the following table:



From the table, it can be seen that the CAADL framework performs much better than all the baselines on accuracy, F1-score, precision, and recall. The CAADL framework has an accuracy of 0.82, an F1-score of 0.83, a precision of 0.84, and a recall of 0.82. These results show the performance power of context awareness and attention mechanisms in multimodal sentiment analysis. The CAADL model is capable of successfully combining data from various modalities and representing the sophisticated inter-modal relationships, resulting in better performance.

Deeper examination provides that the role of the attention mechanism in the CAADL framework's performance cannot be understated. By selectively concentrating on the most informative features in each modality, the attention mechanism enables the model to discard irrelevant information and concentrate on the most information-rich cues for sentiment

identification. The hierarchical attention network also plays a part in the enhanced performance by representing inter-modal and intra-modal interactions, able to capture the contextual information inherent within the data.

4.Discussion

The experimental results show that the suggested Context-Aware Attentive Deep Learning (CAADL) approach outperforms state-of-the-art baselines by a large margin in multimodal sentiment analysis. This is due to several important reasons.

Firstly, the use of pre-trained models (BERT for text and ResNet-50 for images) allows us to leverage the knowledge learned from large-scale datasets, improving the generalization ability of the model. These pre-trained models provide rich feature representations that capture the semantic and visual content of the input data.

Secondly, attention mechanism is also responsible for selectively concentrating on the most important features from all the modalities. By adaptively weighting the contribution of various features, the attention mechanism enables the model to bypass unnecessary information and concentrate on the most useful cues for sentiment classification. This is especially necessary in multimodal data, where different modalities can have varying amounts of useful information.

Thirdly, the hierarchical attention network is able to successfully capture both inter-modal and intra-modal relationships and contextual subtleties that are part of the data. The hierarchical attention network, by taking into consideration the relationships between various modalities as well as the relationships between different features in each modality, gains a more holistic view of the sentiment conveyed in the data.

The findings of our experiments are in line with existing results from the literature, which have established that attention mechanisms and context awareness can have an important impact on the performance of sentiment analysis models. Yet, our research builds on such existing findings by proving the efficacy of these methods in the context of multimodal data.

The major limitation of our work is the use of a single dataset (MOSI) for testing. Although MOSI is a popular benchmark dataset, it is possible that it is not generalizable to all forms of multimodal social media data. Future research will test the performance of the CAADL framework on alternative datasets with different modalities and varying sentiment labels.

The other limitation is the computational cost of the CAADL framework. With the employment of pre-trained models and attention mechanisms, the model can be computationally intensive, especially when handling massive datasets. Future research should investigate how to lower the computational cost of the model, including model compression and knowledge distillation.

5.Conclusion

This work has introduced a new Context-Aware Attentive Deep Learning (CAADL) approach towards improved sentiment analysis of multimodal social media data. The CAADL approach combines attention mechanisms that learn to selectively attend to the most critical features from

textual and visual modalities. Additionally, it utilizes a hierarchical attention network that encodes both inter-modal and intra-modal relationships to capture the contextual nuances in the data.

Experimental findings on the MOSI dataset verify that the CAADL framework performs dramatically better than state-of-the-art baselines for accuracy, F1-score, precision, and recall. These findings underscore the significance of context awareness and attention mechanisms in multimodal sentiment analysis.

Future research will aim to extend the CAADL framework to support other modalities, including audio and video. We will also investigate methods for decreasing the computational complexity of the model and enhancing its scalability. In addition, we intend to examine the application of the CAADL framework to other sentiment analysis tasks, including aspect-based sentiment analysis and emotion recognition. Lastly, investigating the use of this model across other areas of application beyond social media, e.g., customer feedback analysis and market analysis, would be a useful avenue for further research.

6.References

[1] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 417-424).

[2] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing (Vol. 10, pp. 79-86).

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[4] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1746-1751).

.