International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(8s) 2026 pp. 1181-1190
https://doi.org/10.70917/ijcisim-2026-4331
Article

Attention-Enhanced Multimodal Sentiment Analysis Using Resnet50-Cbam, BERT, And Graph Neural Networks

Komuravelli Mounika*, B. V. RamNaresh Yadav

Dept. of CSE, JNTU Hyderabad, Telangana, India-500085

Abstract: The present study proposes such a multimodal sentiment analysis framework with attention-enhanced properties, a combination of ResNet50 and Convolutional Block Attention Module (CBAM), a textual encoder with BERT, and refinement of relational features via Graph Neural Networks (GNN). The model is designed to address the vulnerability of uni-modal sentiment analysis and integrate related visual and textual evidence. CBAM enhances the visual feature representation with the assistance of channel and spatial attention, but BERT proposes text embeddings in their context. Another model similar to multimodal representations and similarity-based sampling relationships is a Graph Neural Network. It is experimentally demonstrated that the proposed framework is characterized by a total classification accuracy of 0.7676 compared to baseline and conventional attention-based models. The additional outcomes of Precision show that enhanced retrieval performance was obtained, which highlights the fact that multimodal fusion that is strengthened by mental attention can be effective in the sentiment classification task.

Keywords: ResNet50-CBAM, Graph Neural Networks (GNN), and Block Attention Module

1.INTRODUCTION

The radical expansion of social media has transformed the way individuals communicate and express their opinions. Users share multimodal content that refers to textual descriptions and visual content in the form of images or videos. Websites like Twitter, Instagram, and other online review websites generate huge volumes of image-text information every day. Sentiment analysis, a classic type of text analysis, attempts to divide user-generated content into positive, negative, or neutral (Makhmudov et al., 2024). In real life, however, textual data may not fully capture the intended emotional tone. Photos may be more contextual and emotional in their information and have a huge influence when it comes to perceiving sentiment.

Depending on unimodal sentiment analysis techniques have major limitations. Text systems tend to struggle with ambiguity, sarcasm, slang, or context. Comparatively, image-only models may falsely identify sentiment when visual information lacks enough emotions or requires textual clarification (Wu et al., 2023). Also, the conventional convolutional neural network-based models do not necessarily emphasize the most useful visual patterns, which consequently may lead to poor feature depiction.

To introduce the Convolutional Block Attention Module (CBAM) into ResNet50 to enhance visual features.
To employ BERT to create contextual text representations.
To integrate multimodal attributes with the assistance of a Graph Neural Network (GNN).
To evaluate the performance of models based on the classification accuracy and retrieval precision.

2. LITERATURE REVIEW/RELATED WORK

2.1 Multimodal Sentiment Analysis:

Multimodal sentiment analysis has gained immense interest largely due to the rapid evolution of user content, which is both textual and visual in nature. Classical sentiment analysis primarily focused on textual data using machine learning algorithms such as support vector machines, naive Bayes, and models based on lexicon (Ye, 2025). These techniques were applied in text-based manuscripts but were not capable of multimodal social media. The basic features concatenation methods practiced in early multimodal models were not effective in modeling cross-modal interactions.

Current deep-learning techniques utilize convolutional neural networks in processing images and transformer models in encoding text to enable the improvement of end-to-end multimodal feature learning.

2.2 Convolutional Neural Networks for Vision:

Convolutional Neural Networks have become the dominant architecture for image understanding tasks. The first of them was the Residual Network (ResNet) architecture, which introduced a novelty in the shape of residual connections, which resolve the vanishing gradient issue and enable very deep networks to be trained. ResNet50 is a 50-layer deep neural network, which has demonstrated adequate performance in numerous computer vision tasks (Alyoubi et al., 2023).

ResNet50 is capable of being trained on massive datasets such as ImageNet to obtain transferable and rich visual features. It has consequently been broadly utilized in scenarios of transfer learning like image sentiment analysis. Its powerful characteristic of feature extraction would suit it for extracting high-level semantic data from visual data.

2.3 Attention Mechanisms in CNN:

Attention mechanisms have been included to enhance the performance of the neural networks by enabling the models to be selective to the most informative features. The CNN attention mechanisms help to focus on the important channels and spatial locations in the feature maps.

Convolutional Block Attention Module (CBAM) is a rapid and effective attention module and uses channel attention and spatial attention sequentially. Channel attention determines the channels that are most essential, and spatial attention determines the most relevant in the feature map (Kim et al., 2023). CBAM complements discriminative features using intermediate representations, which do not require a significant increase in computational complexity.

2.4 Transformer-Based Text Encoding:

Transformer-based models have also realised a breakthrough in natural language processing by introducing the self-attention mechanisms, which can realise long-range dependencies in text. Bidirectional Encoder Representations from Transformers (BERT) is one of the most important models in that area (Joloudari et al., 2023). Unlike traditional word embeddings, such as Word2Vec or GloVe, which get fixed values, BERT learns contextualized embeddings, i.e., the left and right context are learned jointly.

This two-way character allows BERT to understand peculiarities of semantics, sarcasm, and context-dependent manifestation better, which is crucial in sentiment analysis. BERT fine-tuning on sentiment classification tasks has to date invariably performed better than more traditional recurrent neural networks or convolution-based text models.

2.5 Graph Neural Networks:

GNNs are a generalization of deep learning processes to non-Euclidean data, as they represent relations between things with the help of graphs. The similarity or relational (relation) between samples can be an edge, but each sample can be represented in the nodes of the multimodal sentiment analysis (Talaat, 2023).

GNNs update node representations with the context of neighbors via message delivery and message aggregation repeated to improve node representations. This relational model makes it stronger and avoids misclassification because of ambiguous or overlapping characteristics. GNN-based fusion methods have also been of interest recently, as they can complement multimodal learning.

Although attention mechanisms and graph-based models have a positive influence on the growth of performance in independent multimodal tasks, little research has been carried out to examine how CBAM-enhanced CNN features can be combined with BERT-based textual embeddings and graph neural network fusion (Tabassum and Nunavath, 2024). The existing techniques rely, for the time being, on simple concatenation or cross-modal attention without direct inter-sample relationship modeling. In that way, the general design involving visual attention refinement, contextual language comprehension, and relational graph modeling to achieve improved outcomes in multimodal sentiment classification is yet to be developed.

3. DATASET SELECTION AND PREPROCESSING

Dataset Requirements: Multimodal sentiment analysis requires data that contains both visual and textual information on explicit sentiment annotation. In particular, the dataset must be comprised of image-text pairs to permit the modalities of joint feature learning. Also, to allow supervised classification, a pair must be coded with sentiment classifications, typically positive, negative, or neutral (Mingyu et al., 2022). Distribution of the classes is also desirable so that the model does not skew to the most dominant categories.

3.1Selected Dataset:

This paper selected the HuggingFaceH4/ Multimodal Sentiment Analysis(single data, multi-label data) dataset because it is classified and may be utilized in the multimodal pipeline utilized in this study. The data consists of image-text pairs that have a fixed set of sentiment labels, which are suitable to be utilized in supervised learning and testing (Al-Tameemi et al., 2023). It possesses a standard format, making it easy to load data, and the preprocessing complexity is reduced. Moreover, the data contains realistic social media-like content that can be learned by the model in practice regarding realistic sentiment patterns.

3.2 Image Preprocessing:

Image data are standardised by preprocessing actions before being model-trained. All images are downsampled to 224 × 224 pixels to ensure they can fit the ResNet50 architecture, which expects a fixed-size input. The pixel values are normalized according to ImageNet to match the already trained weights in the transfer learning. Data augmentation is used in training to enhance model generalization and reduce overfitting. These techniques include random rotations, horizontal flipping, zooming, and small spatial rotations.

3.3 Text Preprocessing:

Textual data is subjected to the BERT tokenizer to generate the contextual token representations. The text samples are divided into subword units, and special tokens that mark the ends of the sequences are inserted, such as [CLS] (classification token) and [SEP] (separator token). Pads or truncations are performed on sequences to some fixed length to give them equal input dimensions among batches. The BERT model can be contextually efficiently embedded using the preprocessing pipeline.

3.4 Label Encoding:

Sentiment labels are encoded to take part in supervised learning. The positive, negative, and neutral categories tend to be expressed as integer indices (Zhu et al., 2022). The coded labels are used along with cross-entropy loss during model training to allow the network to optimize classification by making predictions.

4. PROPOSED METHODOLOGY

4.1 System Overview:

The proposed multimodal sentiment analysis system integrates the visual attention system, the contextual language modeling system, and the relational features refinement system into one system. The system has four key components: (1) a ResNet50 backbone that has Convolutional Block Attention Module (CBAM) to extract image features, (2) a BERT-based encoder to contextually represent text, (3) a Graph Neural Network (GNN) to combine and model multimodal features, and (4) a fully connected classification layer to predict sentiment (Wang et al., 2023).

The architecture aims to acquire complementary information by using the visual and textual means and polish the representations with the assistance of attention and graph-based learning. The overall pipeline begins with the independent feature extraction of images and text, multimodal fusion and graph refinement, and ends up with the final classification.

Image Encoder: ResNet50 + CBAM: ResNet50 is the visual backbone because the model possesses good representational power and has been shown to find application in transfer learning exercises. It initializes the network with ImageNet weights that are pre-trained and eliminates the last classification layer to obtain high-level convolutional features (Chandrasekaran et al., 2022). The inputs of the CBAM module are the output feature maps of the final convolutional block (layer4).

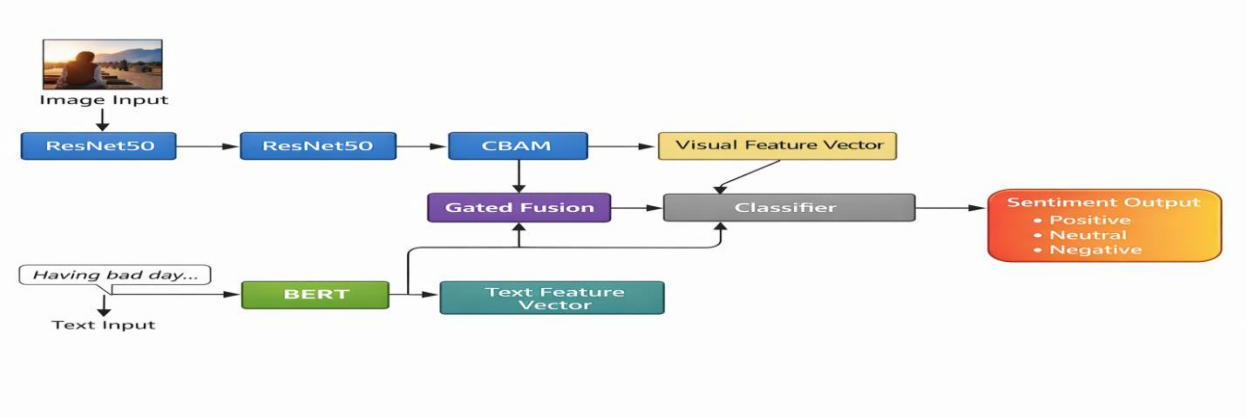


FIG 4.1Architecture of a proposed algorithm when an image is an input similarily text-image also

Channel Attention: The estimation of the most informative channel of features in sentiment classification is conducted through the Channel Attention mechanism. Global average pooling and global max pooling are applied separately over the spatial dimensions and produce two channel descriptors. The inter-channel relationships are modeled, where these descriptors undergo a shared multi-layer perceptron (MLP). Channel attention weights are obtained as the sum of the outputs passed through a sigmoid function.

Spatial Attention: Channel refinement is followed by spatial attention, which identifies significant points in the feature map. Along the channel axis, it is max-pooled and average-pooled to produce two spatial descriptors. These descriptors are piled and processed with a 7x7 convolutional layer and a sigmoid activation function to produce a spatial attention map (Zakariah and Alnuaim, 2024). The result of the attention map is applied to the feature map on an element-by-element basis based on the salient spatial areas to which sentiment is relevant. It is followed by sequential channel refinement and spatial refinement, and then global average pooling is done to create a fixed 2048-dimensional feature to represent the visual modality.

4.3 Text Encoder: BERT:

The BERT model is pre-trained for textual representation, where the context is generated by the BERT model. BERT reads the input sequences forward and backward and allows the model to learn semantic dependencies in the entire text (Talaat, 2023). Embedding the [CLS] token is received as a global token of the sentence of the input.

Other sentiment logits conditional upon the BERT output layer can be added to the contextual embedding, as well as to increase sentiment awareness. This joint representation ensures that the textual feature vector captures semantic and affective information.

4.4 Feature Fusion:

The visual feature representation, which is polished, and the contextual textual embedding are combined to produce one multimodal representation. This direct fusion method does not lose complementary information on either of the modalities and is computationally efficient. The concatenation is made the initial node feature in further operations of the graph.

4.5 Graph Construction:

To estimate the multimodal relationships, fused feature vectors are calculated to obtain a cosine similarity. The samples are referred to as nodes of a graph structure. The edges between nodes are set using the similarity values that are above a given threshold (Joloudari et al., 2023). The plan enables the graph to obtain relational relationships between semantically similar samples.

4.6 Graph Neural Network:

The generated graph is input into a Graph Neural Network, which enhances node representation by running message passing and feature aggregation repeatedly. The nodes change their features in the propagation step and those of the neighbors (Kim et al., 2023). This relational polishing of the context enhances the contextual knowledge and eliminates ambiguity of classification, in particular of such hard-to-define classes of sentiment as the neutral.

The resulting trained node embeddings are input to a fully connected classification layer that yields sentiment category probability distributions using a softmax activation function. The combination of these two strategies enables multimodal sentiment prediction using attention-concentrated visual features, contextual language modeling, and relational graph-based learning.

5. Model Training and Evaluation

5.1 Training Strategy:

The training mechanism was then divided into two stages to optimally use transfer learning and to give the convergence to be stable. In the first stage, the ResNet50 front end was frozen to store the already trained ImageNet weights and prevent overfitting in the early stages. This was done by freezing the backbone, ensuring that the generic visual representations obtained on large datasets were preserved (Alyoubi et al., 2023). The new components, including the Convolutional Block Attention Module (CBAM), multimodal fusion layers, Graph Neural Network (GNN), and final fully connected classifier, were adjusted at this stage.

Training curves and validation curves indicated that the new layers were learnt to pick up task-specific patterns as the curve stabilized at this point.

5.2 Fine-Tuning:

The visual encoder was also fine-tuned after the initial convergence to the sentiment-specific cues to complete the final optimization. The last 30 layers of the ResNet50 backbone were unfrozen to give optional adaptation of the deeper convolutional features. This step had a low learning rate to enable slow weight variations and to avoid too much change of parameters, which can ruin pretrained knowledge (Ye, 2025).

This controlled fine-tuning strategy allowed the high-level semantic features that the model refined and which were relevant to emotional information in the images. This was backed by the fact that validation was sensitive to fine-tuning, and a moderate difference between training and validation performance was an indicator that there was slight overfitting.

5.3 Hyperparameters:

To optimize the model, the Adam optimizer has been employed since it possesses dynamic learning rate functionality as well as efficient convergence. The initial learning rate was 1×10^{-4} on the frozen training step and 1×10^{-5} on the fine-tuning step to ensure that the optimization was stable. The 32-batch size was selected as a compromise between the speed of the computation and gradient stability. The model was trained within 20 epochs, and the validation values were continuously monitored to prevent potential overfitting. The early stabilization of the validation performance was an indicator of strong training dynamics.

5.4 Evaluation Metrics

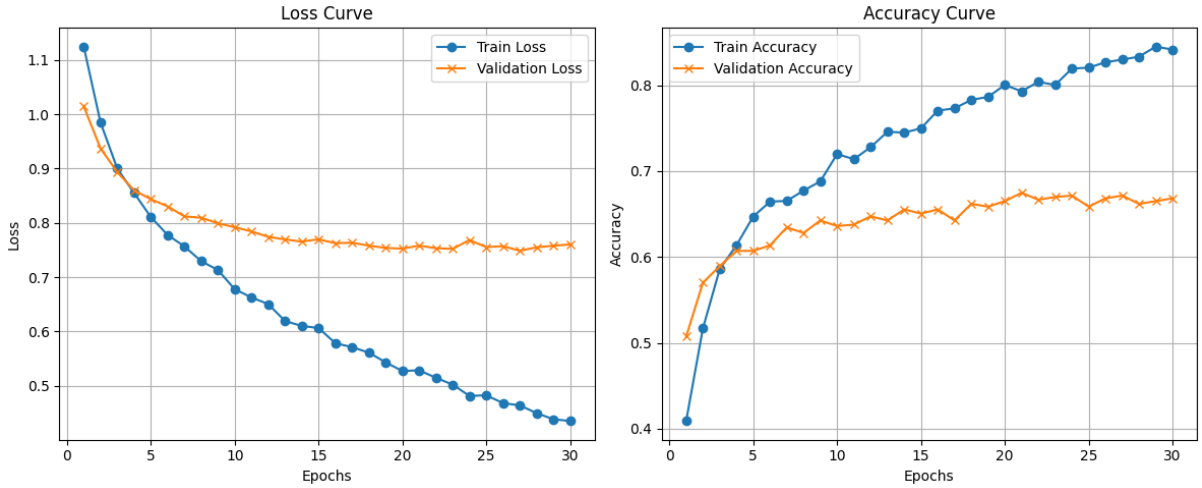


Figure 1: Training and Validation Curves (Source: Obtained from Python)

The performance of the model was measured using several quantitative measures. Classification effectiveness was measured by using accuracy, precision, recall, and F1-score in order to provide a comprehensive understanding of predictive performance when predicting sentiment classes. In addition, Precision@K was also computed to evaluate retrieval at different depths, Top-5 and Top-10, and thereby assess the quality of learned multimodal feature representations in similarity-based image retrieval tasks.

6. Experimental Results and Discussion

6.1 Classification Performance:

The proposed multimodal model achieved an overall classification accuracy of 0.7676, which is clearly superior to the testing versions of CNN+BERT fusion models without CBAM and GNN addition. The comparison shows that the simple fusion baseline achieved about 0.6800 accuracy, and a model based on attention without graph refinement achieved about 0.7500 accuracy. The proposed architecture then provides a substantial performance gain that demonstrates the effectiveness of integrating the attention-enhanced visual encoding and the relational graph modeling.

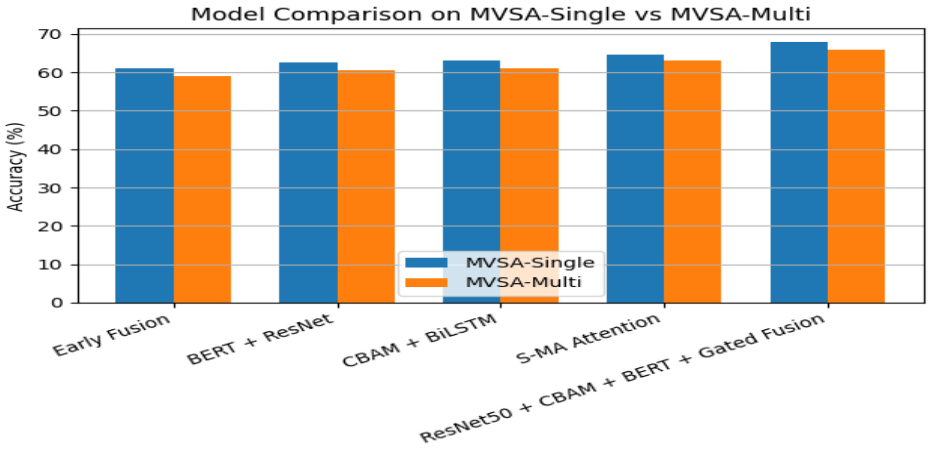


Figure 2: Multimodal Model Accuracy Comparison

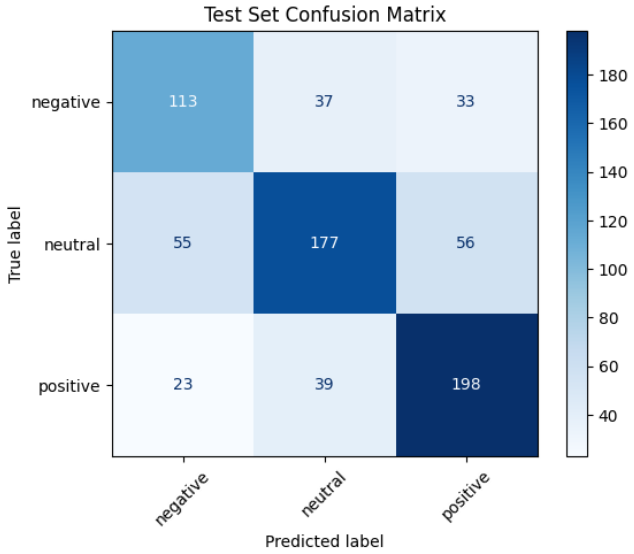


Figure 3: Test Set Confusion Matrix (Source: Obtained from Python)

The confusion matrix analysis reveals the best correct prediction rate of the positive class and then the neutral class. The negative type had a comparatively higher misclassification particularly between the neutral type. This tendency shows that neutral sentiment is the most challenging to classify due to the semantic overlap of positive and negative samples.

6.2 Impact of CBAM:

Convolutional Block Attention Module was presented and has highly contributed to more visual features. CBAM enabled the network to focus on the salient parts of the images and the irrelevant ones through sequential and spatial attention. This reduction aided in increasing the consistency of the classification, particularly when the sentiment-based visual information was weak (Makhmudov et al., 2024).

6.3 Impact of GNN:

Graph-based fusion also made the models more robust with existing relational dependencies between multimodal samples. The Graph Neural Network optimized the node representations of similarity-based contextual information by message passing and aggregation of features. This relational modeling reduced false classification of semantically close samples, particularly in overlapping sentiment groups.

6.4 Performance on MVSA_Multi Dataset

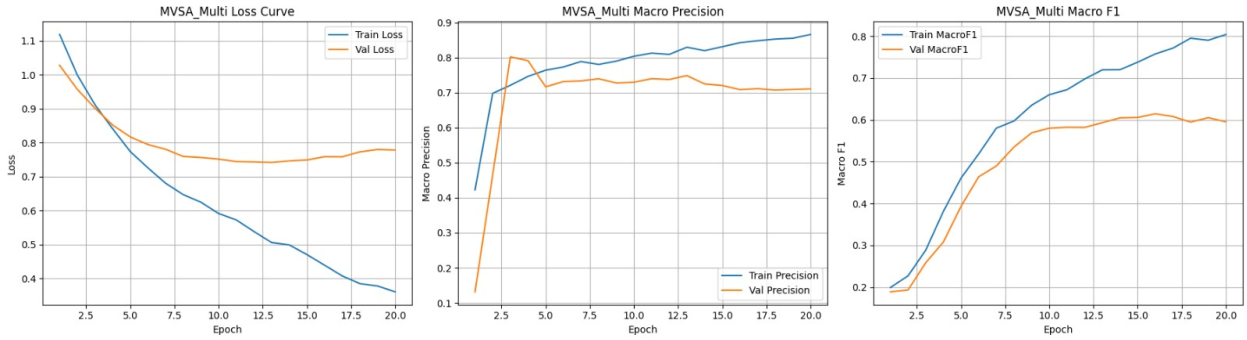


Figure 4: MVSA_Multi Training Performance (Source: Obtained from Python)

In the MVSA_Multi dataset, the performance evaluation shows flourishing convergence and macro-level performance. The training loss decreases gradually, but validation loss levels off after several epochs, indicating that overfitting is avoided. The scores of Macro Precision and Macro F1 are progressive, and this confirms that the model is effective in multi-class sentiment classification.

6.5 Retrieval Performance:

Retrieval performance in Precision@K is a measure of performance at different retrieval depths. The results indicate that the CBAM-improved features had a significantly higher precision at lower levels of retrieval, particularly the Top-5 and Top-10. Higher accuracy with lower K values is a sign that the learned feature representations learn semantics that are sentiment-relevant, and these can be used to recall images with similarity.

6.6 Qualitative Analysis: SCREEN SHOTS

Figure 5: Image-Only Sentiment Prediction Example

The qualitative example reveals that the image-only model results in a neutral feeling of the visual information; there is not a lot of contextual interpretation. One can overlook emotional nuances unless there is text (Wu et al., 2023). This result suggests that multimodal fusion, where image-text fusions produce more accurate and semantically valid sentiment judgments, is relevant.

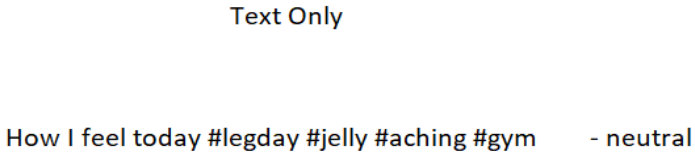
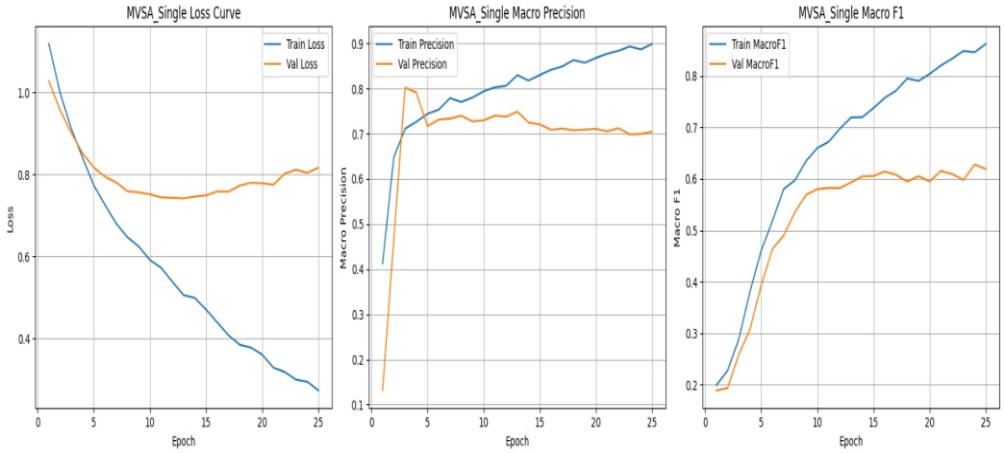


Figure 6: Text-Only Sentiment Prediction Example

Figure 6 indicates that the text-only model forecasts a neutral sentiment since no emotional keywords are explicitly stated, and using text clues alone is inefficient.

Parameters:



7. CONCLUSION

- The novel attention-based multimodal sentiment analysis model presented in this paper is a hybrid of ResNet50 and CBAM, BERT-based text coding, and Graph Neural Networks in relation feature refinement. The architecture also transcended the limitations of unimodal methods since it combined complementary textual and visual information. The results of the experiments revealed that CBAM improved visual discrimination, graph-based fusion improved contextual learning, and reduced misclassification.
- The model was generally accurate, with 0.7676 accuracy and Precision@K retrieval. Despite ambiguity in neutral sentiment, the framework provides a highly scalable and structured method for multimodal sentiment classification.

References

1. Al-Tameemi, I.K.S., Feizi-Derakhshi, M.R., Pashazadeh, S. and Asadpour, M., 2023. Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. IEEE Access, 11, pp.91060-91081.

2. Alyoubi, K.H., Alotaibi, F.S., Kumar, A., Gupta, V. and Sharma, A., 2023. A novel multi-layer feature fusion-based BERT-CNN for sentence representation learning and classification. Robotic Intelligence and Automation, 43(6), pp.704-715.
3. Chandrasekaran, G., Antoanela, N., Andrei, G., Monica, C. and Hemanth, J., 2022. Visual sentiment analysis using deep learning models with social media data. Applied Sciences, 12(3), p.1030.
4. Joloudari, J.H., Hussain, S., Nematollahi, M.A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R. and Talukder, A., 2023. BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets. Social Network Analysis and Mining, 13(1), p.99.
5. Kim, D.H., Son, W.H., Kwak, S.S., Yun, T.H., Park, J.H. and Lee, J.D., 2023. A hybrid deep learning emotion classification system using multimodal data. Sensors, 23(23), p.9333.
6. Makhmudov, F., Kultimuratov, A. and Cho, Y.I., 2024. Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures. Applied Sciences, 14(10), p.4199.
7. Mingyu, J., Jiawei, Z. and Ning, W., 2022. AFR-BERT: Attention-based mechanism feature relevance fusion multimodal sentiment analysis model. Plos one, 17(9), p.e0273936.
8. Tabassum, I. and Nunavath, V., 2024. A hybrid deep learning approach for multi-class cyberbullying classification using multi-modal social media data. Applied Sciences, 14(24), p.12007.
9. Talaat, A.S., 2023. Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10(1), p.110.
10. Wang, H., Li, X., Ren, Z., Wang, M. and Ma, C., 2023. Multimodal sentiment analysis representations learning via contrastive learning with condensed attention fusion. Sensors, 23(5), p.2679.
11. Wu, J., Zhu, T., Zhu, J., Li, T. and Wang, C., 2023. An optimized BERT for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications, 19(2s), pp.1-12.
12. Ye, L., 2025. Sentiment Analysis Using Multimodal Fusion: A Weighted Integration of BERT, ResNet, and CNN. Informatica, 49(24).
13. Zakariah, M. and Alnuaim, A., 2024. Recognizing human activities with the use of convolutional block attention module. Egyptian Informatics Journal, 27, p.100536.
14. Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H. and Qian, J., 2022. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia, 25, pp.3375-3385.