VNU Journal of Science: Com. Science & Com. Eng., Vol. 42, No. 1 (2026) 12–22

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

ViMACSA-GAT: A Graph Attention Network-Based Vietnamese Multimodal Sentiment Analysis Model with Cross-Modal Fusion

Hoang Nam Do, Dinh Tai Pham*

Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam

Received 04th February 2026
Revised 10th March 2026; Accepted 26th March 2026

Abstract: Multimodal sentiment analysis (MSA) models often use simple integration methods that overlook diverse intramodal relationships. Vietnamese, with its rich and complex grammatical and lexical features, presents significant challenges. This paper proposes a new ViMACSA-GAT model using a graph attention network with two parallel branches and a multimodal integration converter. The model has the following main steps: The text modality branch employs the PhoBERT encoder to process Vietnamese text, and the graph attention networks (GAT) model contextual dependencies. For the image modality, the Image branch uses a vision transformer (ViT) to extract features, with nodes representing both the global image and specific regions of interest (ROIs). Then, GAT is used to capture the relationships between the extracted image elements. Next, using a cross-modal transformer, deep merging is performed by simultaneously processing node-level representations from both graphs, allowing for detailed intermodal alignment. The model is optimized with focal loss to handle class imbalance and evaluated on the Vietnamese dataset called ViMACSA, achieving advanced performance in Accuracy (Acc = 85.50%), Precision = 74.01%, Recall = 72.04%, F1-score = 72.63%. The results evaluate the generality of the proposed model for Vietnamese MSA.

Keywords: Multimodal sentiment analysis, graph attention networks, multimodal fusion, vision transformers, Vietnamese language processing, region of interest.

1. Introduction

Sentiment analysis is a key task in natural language processing, aiming to identify and categorize the sentiment or opinion expressed in text, such as positive, negative, or neutral sentiment. With the development of social media platforms, the field of sentiment analysis has become important for applications such as customer feedback analysis, market research, and public opinion monitoring. Traditional sentiment analysis is primarily based on text data, but real-world communication often

*Corresponding author.
E-mail address: namdh@ntt.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6927

involves multiple modes, such as text combined with images, videos, or audio. This has spurred MSA, integrating information from different modes to capture a more comprehensive understanding of user intent and sentiment [1]. Currently, multimodal approaches have improved performance by leveraging complementary signals between modes, such as an image accompanying a text post that can provide visual context to clarify ambiguous sentiments found only in text. GATs [2–6] address key problems of MSA, such as efficiently modeling complex relationships within methods and between methods combined together. The hardest problem in MSA is understanding interactions. For example, a phrase "Great trip!" accompanied by a picture of a stormy day can carry ironic connotations. GATs help capture these nuances by constructing and analyzing graphs representing multimodal data. However, most advances in multimodal sentiment analysis have focused on resource-rich languages like English and Chinese, leaving resource-poor languages like Vietnamese undervalued and less popular. Vietnamese is particularly noteworthy due to its tonal nature, syllabic structure, and the frequent use of informal slang, abbreviations, and spelling errors in the context of social media [7]. These linguistic features complicate NLP tasks, and the limited number of annotated multimodal datasets restricts the development of deep learning models for the Vietnamese language. Addressing this gap through pre-trained models on standard datasets such as PhoBERT, a monolingual BERT-based model pre-trained on a large Vietnamese corpus, has improved performance for tasks such as part-of-speech tagging and named entity recognition [8]. Similarly, the ViSoBERT model [7], specifically designed for Vietnamese social media texts, has improved the ability to handle informal language and achieved good results in sentiment-related tasks. However, multimodal sentiment analysis for Vietnamese remains limited. Datasets such as UIT-VSFC [9], containing over 16,000 student responses annotated by sentiment and topic, provide a foundation for text analysis but lack multimodal elements. The UIT-ViQuAD Vietnamese machine-reading comprehension dataset [10], with 23,000 question-answer pairs. The ViMACSA multimodal aspect-category sentiment analysis dataset [11], with text-image pairs in the hotel sector, highlights the increasing availability of resources. However, there is a need for models that effectively combine multimodal data while addressing the specific challenges of the Vietnamese language.

This paper introduces a MSA model based on graph convolutional neural networks, designed to process Vietnamese data by modeling intermodal interactions. We conduct experiments on a multimodal Vietnamese dataset, evaluating the model's ability to integrate textual and visual features to improve sentiment classification. The main contributions of this paper include:

- Proposed multimodal processing model based on a graph attention network.
- Deep merging mechanism using intermodal transformers for multimodal Vietnamese data, including images and text.
- Model optimized using focal loss to handle class imbalance in the dataset.
- Model achieved experimental results on the advanced ViMACSA dataset with an accuracy of 85.50% and an F1-score of 72.63%.

The remainder of the paper is organized as follows: Section 2 evaluates related works, Section 3 describes the proposed model, Section 4 details the experimental setup and results, and Section 5 concludes.

2. Related Work

MSA is a dynamic field of research that aims to understand and categorize sentiment from a

variety of data modes. Work in this field can be categorized into several main approaches.

2.1. Approaches in Multimodal Sentiment Analysis

Early studies on SA [12, 13] often focused on unimodal approaches. Text-based sentiment prediction is the most traditional and common approach, exploiting linguistic content. Image-based methods analyze visual elements such as facial expressions, color, and context, while sound-based methods use features such as tone and pitch to identify sentiment. However, relying solely on unimodal approaches is often insufficient to predict the complexity of human sentiment. Therefore, a multimodal integration approach is needed for new research. This approach integrates information from two or more modes to provide more accurate predictions [1]. Integration techniques vary widely, from joining feature vectors (early integration) to combining decisions from individual models (late integration), and more complex architectures such as tensor-based and attention-based integration to model intermodal interactions [14].

2.2. Applications of Graph Neural Networks in Multimodal Learning

Graph Neural Networks (GNNs) [15] have addressed many challenges in the field of multimodal learning. GNNs allow modeling multimodal data as graph structures. In this, nodes can represent words, objects in images, or different modes. The edges of the graph represent the relationships between these nodes. Variants of GNNs [16], such as graph convolutional neural networks and graph attentional networks, can learn context-dependent representations, effectively capturing complex interactions within the same mode and between different modes. This approach is suitable in MSA because it can model the interaction between a word in a comment and a specific object in the accompanying image.

In summary, advanced techniques such as GNNs have been successfully applied in MSA on resource-rich languages like English, but there is still a research gap on Vietnamese data. The Vietnamese language has unique linguistic characteristics (isolating language, tones) and informal expressions on social networks, creating new challenges with unique characteristics [17]. Currently, to our knowledge, the number of research works applying GCN or GAT to Vietnamese multimodal data is still very limited. The FCMF method [11] has pioneered the proposal of an efficient model for merging two methods, creating a standard for future research on the ViMACSA dataset. However, building a graph-based architecture to explicitly model intramodal relationships for Vietnamese remains a challenge. This paper aims to directly address that problem, proposing a GAT-based architecture to further exploit the multimodal data structures of the Vietnamese language.

3. Proposed Method

3.1. Problem Statement

Consider a multimodal data sample from the dataset $\mathcal{D}$ denoted as $S = [\mathbf{X}_T, \mathbf{X}_I]$, where: $\mathbf{X}_T \in \mathbb{R}^{T \times d_T}$ represents the text modality, T is the number of tokens in the sentence or comment, and d_T is the hidden representation size of each token; $\mathbf{X}_I \in \mathbb{R}^{N_{\text{img}} \times d_I}$ represents the image modality, N_{img} is the number of images in sample S, and d_I is the size of the image feature vector. The notation I (short for *Image*) represents the entire image component associated with the post or comment. Each data sample $S^{(i)} = [\mathbf{X}_T^{(i)}, \mathbf{X}_I^{(i)}]$ is associated with a corresponding sentiment label $y^{(i)} \in \mathcal{Y} = \{\text{negative}, \text{neutral}, \text{positive}\}$.

The goal of the MSA problem is to determine a parameterization mapping:

$$f_\theta : (\mathbf{X}_T, \mathbf{X}_I) \longrightarrow \hat{y} \tag{1}$$

where $\hat{y}$ is the sentiment label predicted by the model and θ is the entire set of parameters to be trained.

The problem is modeled as experimental risk optimization on the dataset $\mathcal{D} = \{(S^{(i)}, y^{(i)})\}_{i=1}^{N}$, where N is the total number of training samples, by minimizing the average loss function:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta\left(\mathbf{X}_T^{(i)}, \mathbf{X}_I^{(i)}\right), y^{(i)}\right) \quad (2)$$

where, $\mathcal{L}(\cdot)$ is the loss function *Focal Loss*, which helps reduce the influence of easy patterns and focus on difficult patterns to address class imbalance in sentiment data. In the proposed model, the text component $\mathbf{X}_T$ is represented as a token-level graph to model contextual relationships between words, while the image component $\mathbf{X}_I$ is represented as a region-level graph to learn spatial relationships between regions of interest (ROI) in the image. The ultimate goal is to learn a unified multimodal representation:

$$\mathbf{h} = F_\theta(\mathbf{X}_T, \mathbf{X}_I) \quad (3)$$

where $\mathbf{h}$ is a composite feature vector containing both linguistic and visual information, enabling the model to accurately identify sentiment in multimodal posts, even when the data is noisy or partially missing information.

3.2. Overall Model Architecture

To address the challenges in MSA on Vietnamese data, we propose a ViMACSA-GAT architecture as shown in Figure 1, comprising three main components: two parallel branches for encoding features for text and images, a transformer-based method merging module, and a classification head for predicting sentiment. The text processing branch is designed to capture sequential contextual relationships between words. Using the PhoBERT method, a transformer model pre-trained for Vietnamese, we extract text features to generate feature vectors for each token. Instead of using a dependency tree-based syntactic graph, which often has high error rates on informally structured social network evaluation data, we construct a custom sequence graph with a neighborhood window of $k = 2$. This design is calculated based on the monosyllabic nature of Vietnamese, where compound words are often composed of two syllables. Connecting neighboring nodes within a narrow range allows for this. Through its attention mechanism, GAT enables the model to automatically learn and assign weights to important tokens, thereby effectively capturing local semantic dependencies and the lexical structures specific to Vietnamese. As a result, GAT extracts context-rich node representations before performing intermodal interactions. These representations are then aggregated to create a global representation vector for the text. Simultaneously, the image branch extracts visual features and models the spatial relationships between objects. ViT is used to extract features. The ROI features, together with the global image features, are used to construct a fully connected graph, which is then fed into another GAT. This mechanism allows the model to identify and focus on image regions that have the most influence on overall sentiment. Next, instead of using crude merging methods (such as vector joining or feature addition), we propose a deep cross-modal alignment mechanism via a cross-modal transformer. Here, the model doesn't simply mix information but maps node-level representations from text and image graphs into a shared attention space. The self-attention mechanism in the transformer acts as a multi-dimensional correlation filter. It allows each text token (e.g., "dirty room") to directly query and match its weights against ROIs. This detailed alignment process helps the model detect subtle inconsistencies (such as irony) or complementarities between text and image-features that traditional merging methods often miss due to the loss of data locality.

Finally, this representation is passed through a classification head to predict logits for the three sentiment classes. The entire model is optimized using Focal Loss, which addresses the problem of data imbalance by focusing on learning difficult patterns.

To learn the mapping f_θ, we propose the architecture *ViMACSA-GAT*. The f_θ model is decomposed into three main components: (i) the text processing branch E_T with the representation $Z_T = E_T(X_T)$; (ii) the image processing branch E_I with the representation $Z_I = E_I(X_I)$; and (iii) the fusion and classification module consisting of F_{fusion} and C_{head}, where $h_{\text{fused}} = F_{\text{fusion}}(Z_T, Z_I)$ and Logits $= C_{\text{head}}(h_{\text{fused}})$. In the above notation, X_T and X_I are the input data (section 3.1); Z_T and Z_I are the node-level feature representations from each branch (details in sections 3.3 and 3.4); h_{fused} is the final fusion vector representation (section 3.5); and *Logits* is the output of the classification layer (section 3.6).

3.3. Text Processing Branch

Assume the text input is a string of T tokens $W = \{w_1, w_2, \ldots, w_T\}$ This token string is fed into PhoBERT to extract the hidden representations as follows:

$$H^t = \text{PhoBERT}(W) \in \mathbb{R}^{B\times T\times d_{\text{bert}}} \tag{4}$$

where $H^t = \{h_1^t, h_2^t, \ldots, h_T^t\}$ is the string of feature vectors; B is the batch size; and d_{bert} is the hidden dimension of PhoBERT (768).

Next, the chain graph $G_t = (V_t, E_t)$ is constructed with tokens as nodes ($V_t = H^t$). The adjacency matrix A_t is generated based on the neighborhood window. The node representations are updated via GATEncoder as follows:

$$Z^t = \text{GATEncoder}(H^t, A_t) \in \mathbb{R}^{B\times T\times d_{\text{model}}} \tag{5}$$

where $Z^t = \{z_1^t, z_2^t, \ldots, z_T^t\}$ are the node features after GAT, and d_{model} is the hidden dimension of model (256).

The global vector representing the text, denoted t_{graph}, is obtained by masked mean pooling on the nodes:

$$t_{\text{graph}} = \text{MaskedMeanPooling}(Z^t) \in \mathbb{R}^{B\times d_{\text{model}}} \tag{6}$$

3.4. Image Processing Branch

Assume the input is a set of N_{img} images, each image containing R regions of interest (ROI). Image features are extracted using ViT, where each image I_k is fed into ViT to obtain a patch feature grid $P_k \in \mathbb{R}^{(G\times G)\times d_{\text{vit}}}$, with d_{vit} being the hidden dimension of ViT (768). For the jth ROI region, the feature r_{kj} is calculated by averaging the patches belonging to that region. In addition, a global node g_k is created by averaging all the patch features of the image. The node representation for image I_k is defined as follows:

$$H^{i_k} = \{\, g_k,\, r_{k1},\, \ldots,\, r_{kR} \,\} \in \mathbb{R}^{(1+R)\times d_{\text{vit}}} \tag{7}$$

Next, the nodes in H^{i_k} are encoded using GAT. A fully connected graph G_{i_k} is constructed, where A_{full} is the corresponding adjacency matrix. The node representations are updated by GATEncoder:

$$Z^{i_k} = \text{GATEncoder}(H^{i_k}, A_{\text{full}}) \in \mathbb{R}^{(1+R)\times d_{\text{model}}} \tag{8}$$

where d_{model} is the hidden dimension of the model (256). The global representation for each image is obtained by mean pooling on the nodes. Then, the final representation for all images in a sample is calculated by averaging over N_{img} images:

$$i_{\text{graph}} = \frac{1}{N_{\text{img}}} \sum_{k=1}^{N_{\text{img}}} \text{MeanPooling}\left(Z^{i_k}\right) \in \mathbb{R}^{B\times d_{\text{model}}} \tag{9}$$

where B is the batch size.

Additionally, ROI-level node representations from the entire image are combined into a tensor $Z^i \in \mathbb{R}^{B\times \text{maxR}\times d_{\text{model}}}$ to serve the subsequent merging steps. This study uses available ROI labels from the ViMACSA dataset to ensure

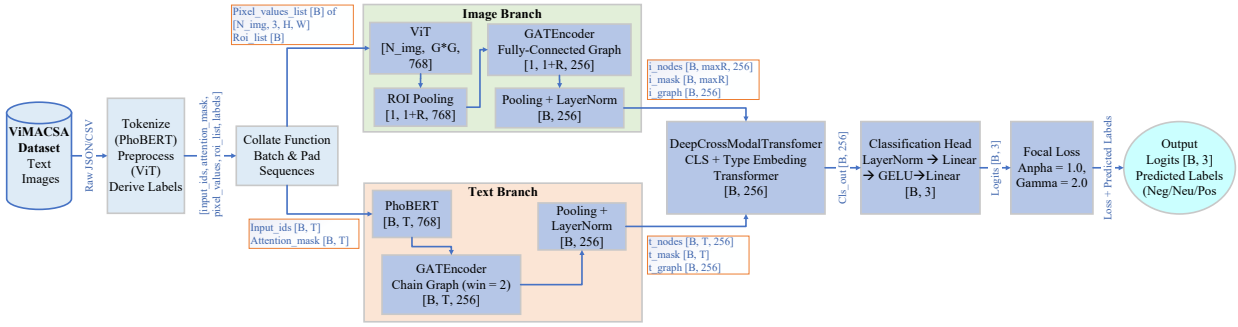

Figure 1. Overall architecture of the ViMACSA-GAT model.

accuracy in training and alignment; the graph architecture of the Image branch is compatible with automated extraction methods. In practical applications, manual ROI labels can be replaced with proposed regions from object detection models such as Faster R-CNN or YOLO. Additionally, the model can also be adapted to use virtual nodes extracted from the regions with the highest attention weight in the ViT attention map, completely eliminating the reliance on manual labeling while maintaining the ability to focus on important local features.

3.5. Bimodal Fusion

The representations from the two branches are combined with a special token `[CLS]` and type embeddings E_{type} to form the merged input string:

$$S = [\,\mathbf{x}_{\text{cls}};\, Z^t;\, Z^i\,] + E_{\text{type}} \in \mathbb{R}^{B\times(1+T+\text{maxR})\times d_{\text{model}}} \tag{10}$$

where B is the batch size, T is the number of text tokens, and maxR is the maximum ROI. The string S is then fed into the Transformer encoder to model the cross-interaction between modalitys:

$$H^{\text{cross}} = \text{TransformerEncoder}(S) \in \mathbb{R}^{B\times(1+T+\text{maxR})\times d_{\text{model}}} \tag{11}$$

In the Transformer output, the representation of the token `[CLS]` (first position) is used as the final merged vector representation:

$$\mathbf{cls}_{\text{out}} = H^{\text{cross}}[:, 0, :] \in \mathbb{R}^{B\times d_{\text{model}}} \tag{12}$$

3.6. Classification Head

The vector $\mathbf{cls}_{\text{out}}$ is passed through a feed-forward neural network (FFN) to compute logits for $C = 3$ of the sentiment class:

$$\text{Logits} = \text{FFN}(\text{LayerNorm}(\mathbf{cls}_{\text{out}})) \in \mathbb{R}^{B\times C} \tag{13}$$

3.7. Focal Loss Function

The Focal Loss function is defined as follows:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{14}$$

where:

- p_t is the model's prediction probability for the correct class;
- γ is the focusing parameter, with $\gamma \geq 0$;
- α_t is the balancing coefficient to handle imbalances between classes.

3.8. ViMACSA-GAT Algorithm

The following algorithm summarizes the complete processing pipeline of the proposed model, clearly illustrating the input–output relationships among its components.

Algorithm 1 ViMACSA-GAT Model

Require: D_{text}: Text comment dataset; D_{img}: Image list dataset; D_{roi}: ROI dataset; θ: Model parameters
Ensure: $\hat{Y}$: Predicted label vector
Main-Procedure: MSA_Pipeline($D_{\text{text}}, D_{\text{img}}, D_{\text{roi}}, \theta$)
// Stage 1: Data Preparation
dataset ← ViMACSADataset($D_{\text{text}}, D_{\text{img}}, D_{\text{roi}}$)
dataloader ← DataLoader(dataset, collate_fn = CollateFn)
// Stage 2: Batch Processing
for each batch in dataloader **do**
  input_ids, **att_mask** ← $batch['input\_ids'], batch['attention\_mask']$
  px_values_list, **roi_lists** ← $batch['pixel\_values\_list'], batch['roi\_lists']$
  // Step 2.1: Intra-modal Encoding
  t_nodes, **t_mask** ← TextBranch(**input_ids**, **att_mask**)
  i_nodes, **i_mask** ← ImageBranch(**px_values_list**, **roi_lists**)
  // Step 2.2: Cross-modal Fusion
  $\mathbf{cls}_{out}$ ← CrossModalFusion(**t_nodes**, **t_mask**, **i_nodes**, **i_mask**)
  // Step 2.3: Classification
  logits ← ClassificationHead($\mathbf{cls}_{out}$)
  $\hat{\mathbf{y}}_{\text{batch}}$ ← arg max(softmax(**logits**))
  Append $\hat{\mathbf{y}}_{\text{batch}}$ to $\hat{Y}$
end for
return $\hat{Y}$

Algorithm 2 Text Branch

Require: **input_ids**, **att_mask**
Ensure: $\mathbf{T}_{nodes}$, **att_mask**
$\mathbf{H}^t$ ← PhoBERT(**input_ids**, **att_mask**)
$\mathbf{A}_t$ ← ConstructChainGraph($\mathbf{H}^t$)
$\mathbf{T}_{nodes}$ ← GATEncoder($\mathbf{H}^t, \mathbf{A}_t$)
return $\mathbf{T}_{nodes}$, **att_mask**

Algorithm 3 Image Branch

Require: **px_values_list**, **roi_lists**
Ensure: $\mathbf{I}_{nodes\_batch}$, corresponding_mask
$\mathbf{I}_{nodes\_batch}$ ← [ ]
for each (**sample_imgs**, **sample_rois**) **do**
  $\mathbf{H}^i_{sample}$ ← [ ]
  for each (**img**, **rois**) **do**
    P ← ViT(**img**)
    $\mathbf{H}^i$ ← ROIPooling(**P**, **rois**)
    $\mathbf{A}_i$ ← ConstructFullyConnectedGraph($\mathbf{H}^i$)
    $\mathbf{Z}^i$ ← GATEncoder($\mathbf{H}^i, \mathbf{A}_i$)
    Append $\mathbf{Z}^i$ to $\mathbf{H}^i_{sample}$
  end for
  $\mathbf{I}_{nodes\_padded}$ ← PadOrTruncate($\mathbf{H}^i_{sample}$)
  Append $\mathbf{I}_{nodes\_padded}$ to $\mathbf{I}_{nodes\_batch}$
end for
return $\mathbf{I}_{nodes\_batch}$, corresponding_mask

4. Experimental Setup

4.1. Dataset Characteristics

In this study, we employ the ViMACSA [11] which is a novel multimodal dataset for sentiment analysis in Vietnamese. ViMACSA is built in the context of the hotel industry, where customer experiences and sentiments are often expressed simultaneously through text and images. The dataset includes 4,876 text-image pairs, collected from online review and experience-sharing platforms. A special feature of ViMACSA is that in addition to the overall sentiment label, the data also comes with 14,618 detailed annotations assigned specifically to both text and images. This allows the model to learn not only general sentiment but also to delve deeper into the relationship between language and visuals. The training dataset consisted of 2,876 reviews, with an average length of approximately 42.42 words, each containing an average of 3.01 sentimental aspects, 6,421 positive labels, 1,402 neutral labels, and 830 negative labels, accompanied by 5,428 images and 8,656 ROI. The validation dataset consisted of 1,000 reviews, with an average length of 39.36 words, an average of 2.98 sentimental aspects per review, 2,230 positive labels, 463 neutral labels, and 291 negative labels, along with 1,789 images and 2,880 ROI. The test set also included 1,000 reviews, with an average length of 42.17 words, an average of 2.98 aspects per review, 2,178 positive labels, 485 neutral labels, 318 negative labels, along with 1,841 images and 3,097 ROI.

4.2. Implementation Details

This study employs the following training configurations. The model is trained for 20 epochs, representing the maximum number of full passes over the training dataset. Early stopping with a patience of 8 is applied; training terminates if the F1-score on the validation set does not improve for 8 consecutive epochs, which helps prevent overfitting and reduces unnecessary computational cost.

A batch size of 16 is used, indicating the number of samples processed in each parameter update step. The initial learning rate is set to 1×10^{-5} using the AdamW optimizer, a commonly adopted small learning rate for fine-tuning transformer-based models. The learning rate is decayed by 20% after each epoch, with a decay factor $\gamma = 0.8$, to facilitate better convergence in later training stages.

To address class imbalance, we adopt the focal loss function. The focusing parameter γ is set to 2.0 to emphasize hard-to-classify samples, while the balancing parameter α is fixed at 1.0, implying no additional class-wise weighting.

For the GAT and cross-modal transformer layers, the hidden size is set to 256, with 2 layers and 4 attention heads. The maximum text length is limited to 128 tokens, and longer comments are truncated accordingly. To ensure reproducibility, the random seed is fixed at 42. The model is implemented in PyTorch and trained on an NVIDIA GeForce RTX 4070 Ti SUPER GPU.

4.3. Evaluation Metrics

Accuracy (Acc): is defined as the ratio of correctly predicted samples to the total number of samples in the dataset:

$$\text{Acc} = \frac{\text{Number of correctly predicted samples}}{\text{Total number of samples}} \tag{1}$$

Precision, Recall, and F1-score: These metrics provide a comprehensive evaluation of the model's performance across classes. In this multi-class classification task (Negative, Neutral, Positive), we adopt the macro-average scheme, where the metric is computed independently for each class and then averaged without weighting. This ensures that all classes contribute equally, regardless of their frequency.

- **Precision:** Measures the reliability of positive predictions. High precision for a class indicates that when the model predicts a sample as belonging to that class, it is likely to be correct.
- **Recall:** Measures the model's ability to identify all true positive samples. High recall indicates that most actual samples of a class are correctly detected.
- **F1-score:** The harmonic mean of Precision and Recall, providing a balanced evaluation metric. It is the primary metric used in this study to compare overall model performance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

4.4. Experimental Results and Discussion

After the training process, the ViMACSA-GAT model performed best on the test set (selected based on the highest F1-score on the test set). Table 1 summarizes the model's final performance.

The model achieved an overall F1-score of 72.63%, demonstrating balanced classification across sentiment classes. The 85.50% accuracy also confirms the superior performance of the proposed model. Although the results were recorded in the hotel sector, the model's high performance is partly due to PhoBERT's ability to extract generalized features for text and ViT for images. This suggests that the model has the potential to maintain stable performance when applied to other multimodal data domains with similar linguistic characteristics.

Table 1. Performance comparison of ViMACSA-GAT with baseline models on the ViMACSA dataset

Modality	Model	Precision	Recall	F1-score (%)	Accuracy (%)
Text	MemNet*	57.71	51.53	51.88	–
	GCAE*	61.92	54.06	55.39	–
	IAN*	65.02	59.05	60.28	–
Text + Visual	ESAFN*	64.70	59.79	60.32	–
	MIMN*	67.01	62.74	63.73	–
	MACSA-LSTM*	72.53	69.04	69.54	–
Text	RoBERTa*	66.76	61.69	62.53	–
	LCF-RoBERTa*	73.63	69.69	70.96	–
Text + Visual	FCMF*	**81.43**	**78.80**	**79.73**	–
	ViMACSA-GAT (ours)	**74.01**	**72.04**	**72.63**	**85.50**

*Results are reported from FCMF [11].

4.5. Ablation Study

To verify the contribution of each component in the proposed ViMACSA-GAT architecture, we conduct an ablation study in which key modules are systematically removed or replaced, and the model performance is re-evaluated accordingly. The results are reported in Table 2

Analysis of the results in Table 2 and Figure 2 shows:

- Text branch ablation: When using only image information, performance decreased most significantly (Acc = -7.1% and F1 = -25.76%). This confirms that text information plays a crucial role and provides indispensable supporting emotional signals.
- Image branch ablation: When using only text information, performance decreased significantly (Acc = -1.1% and F1 = -3.01%). This confirms that image information and the objects within it play a crucial role and provide indispensable supplementary emotional signals.
- Cross-modal transformer ablation: When replacing the Transformer merging module with a simpler method of concatenating the two image and text feature vectors and passing them through a feedforward network (FFN), performance decreased (Acc = -1.1% and F1 = -4.47%). This result demonstrates that the cross-attention mechanism in the transformer is effective at capturing complex interactions between the two modalities, outperforming simple fusion methods.
- GAT ablation in both branches: When replacing the GAT encoders in both branches with simple mean pooling, performance decreased (Acc = -1.1% and F1 = -2.54%). This demonstrates the importance of GAT in modeling intramodal relationships between words in text and between regions in images to learn matching feature representations efficiently.

In summary, the excisional experimental portion of this study confirms that all proposed components, including both processing branches, the graph attention mechanism, and the transformer unification module, contribute positively to the model's final performance.

5. Conclusions and Future Work

This paper introduces the ViMACSA-GAT architecture, which effectively combines graph attention networks and transformers to address the challenge of Vietnamese text and

Table 2. Results of the ablation study

Model	Acc (%)	Prec.	Rec.	F1 (%)
Full model (ViMACSA-GAT)	**85.50**	**74.01**	**72.04**	**72.63**
Without Text Branch	78.40	57.65	45.50	46.87
Without Image Branch	84.40	69.29	70.04	69.62
Without Cross-Modal Transformer	84.40	68.09	68.24	68.16
Without GAT (Mean Pooling)	84.40	71.85	69.00	70.09

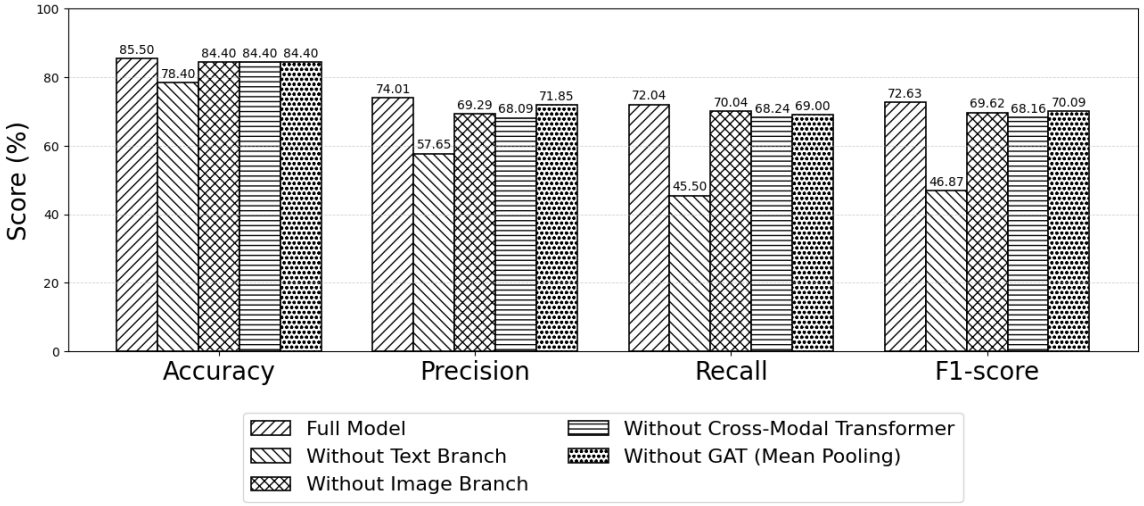

Figure 2. Comparison of model performance.

image datasets the first of its kind in the field of MSA, to our knowledge. Experimental results show that the model achieves good performance with an accuracy of Acc = 85.50% and F1-score = 72.63%. Component removal studies convincingly demonstrate that the integration of image information, text, relational modeling using GAT, and the transformer merging mechanism are all core components that directly contribute to the model's success. This paper affirms the potential of combining graph architectures and attention mechanisms for complex MSA problems. In the future, research will extend the evaluation of ViMACSA-GAT to Vietnamese multimodal datasets across various fields to verify its generalizability. Simultaneously, domain-adaptive approaches and less pattern learning will be studied to reduce reliance on large annotation datasets. Furthermore, integrating an automatic object detection module and further exploiting the cross-attention mechanism between text and images is expected to help build a multimodal end-to-end sentiment analysis system, enhancing its applicability in real-world scenarios.

References

[1] K. Zhao, M. Zheng, Q. Li, J. Liu, Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective, IEEE Access 13 (2025) 64556–64583, https://doi.org/10.1109/ACCESS.2025.3554665.

[2] X. Jia, M. Jiang, Y. Dong, F. Zhu, H. Lin, Y. Xin, H. Chen, Multimodal Heterogeneous Graph Attention Network, Neural Computing and Applications 35 (4) (2023) 3357–3372, https://doi.org/10.1007/s00521-022-07862-6.

[3] P. Gong, J. Liu, X. Zhang, X. Li, A Multi-Stage Hierarchical Relational Graph Neural Network for Multimodal Sentiment Analysis, in: ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096644.

[4] X. Yang, S. Feng, Y. Zhang, D. Wang, Multimodal Sentiment Detection Based on Multi-Channel Graph Neural Networks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021, pp. 328–339, https://doi.org/10.18653/v1/2021.acl-long.28.

[5] G. Yang, D. Han, Sentiment Analysis for Tourism Reviews Based on Dual-Stream Graph Attention Fusion Network, International Journal of Information and Communication Technology 26 (6) (2025) 117–134, https://doi.org/10.1504/IJICT.2025.145406.

[6] X. Shi, W. Ding, M. Hu, X. Kang, F. Ren, Triple Dimensional Psychology Knowledge Encouraging Graph Attention Networks to Exploit Aspect-Based Sentiment Analysis, Scientific Reports 15 (1) (2025) 27109, https://doi.org/10.1038/s41598-025-08914-2.

[7] N. Nguyen, T. Phan, D.-V. Nguyen, K. Nguyen, ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5191–5207, https://doi.org/10.18653/v1/2023.emnlp-main.315.

[8] D. Q. Nguyen, A.-T. Nguyen, PhoBERT: Pre-Trained Language Models for Vietnamese, in: Findings of EMNLP 2020, 2020, pp. 1037–1042, https://doi.org/10.18653/v1/2020.findings-emnlp.92.

[9] K. Van Nguyen, V. D. Nguyen, P. X. Nguyen, T. T. Truong, N. L.-T. Nguyen, UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis, in: 2018 International Conference on Knowledge and Systems Engineering (KSE), 2018, pp. 19–24, https://doi.org/10.1109/KSE.2018.8573337.

[10] K. Van Nguyen, D.-V. Nguyen, A. G.-T. Nguyen, N. L.-T. Nguyen, A Vietnamese Dataset for Evaluating Machine Reading Comprehension, arXiv preprint arXiv:2009.14725 (2020).

[11] Q. H. Nguyen, M.-V. T. Nguyen, K. Van Nguyen, New Benchmark Dataset and Fine-Grained Cross-Modal Fusion Framework for Vietnamese Multimodal Aspect-Category Sentiment Analysis, Multimedia Systems 31 (1) (2025) 4.

[12] M. Hankar, T. Mzili, M. Kasri, A. Beni-Hssane, Sentiment Analysis Survey: Datasets, Techniques, Applications, Tools, and Challenges, Knowledge and Information Systems (2025).

[13] J. Cui, Z. Wang, S.-B. Ho, E. Cambria, Survey on Sentiment Analysis: Evolution of Research Methods and Topics, Artificial Intelligence Review 56 (8) (2023) 8469–8510.

[14] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal Sentiment Analysis Based on Fusion Methods: A Survey, Information Fusion 95 (2023) 306–325.

[15] F. Liang, C. Qian, W. Yu, D. Griffith, N. Golmie, Survey of Graph Neural Networks and Applications, Wireless Communications and Mobile Computing 2022 (2022) 9261537, https://doi.org/10.1155/2022/9261537.

[16] C. Peng, J. He, F. Xia, Learning on Multimodal Graphs: A Survey, arXiv preprint arXiv:2402.05322 (2024).

[17] K. T. Tran, M. H. Dinh, T. N. N. Tran, V. X. Nguyen, T. T. Bui, Exploring the Potential of Graph Neural Networks for Vietnamese Sentiment Analysis, HUFLIT Journal of Science 9 (2) (2025) 11–11.