International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 108-121
https://doi.org/10.70917/ijcisim-2026-3571
Article

An Uncertainty-Aware Cross-Modal Graph Attention Model for Improving Aspect-Based Sentiment Analysis

Amitha Joseph[1]*, R Priya [2]

[1]Department of Computer Science, Sree Narayana Guru College, K.G. Chavadi, Coimbatore, Tamilnadu.
* Corresponding author's Email: amithajoseph2016phd@gmail.com

Abstract: The significance of the user-generated content is crucial in influencing business strategies by capturing public sentiment across various platforms. In this context, Aspect-Based Sentiment Analysis (ABSA) is now considered a main technique to determine sentiment polarities related to definite product attributes or aspects. While early ABSA approaches relied solely on textual input, recent advancements have highlighted the benefits of incorporating multimodal data, particularly text and images to improve sentiment interpretation. However, existing models face challenges such as ineffective fusion strategies, loss of informative features, and increased computational cost due to complex architectures. To address these limitations, a novel framework Attention aware Graph Convolutional Multimodal Aspect based Sentiment analysis Network (AGCMASN) is proposed. This model introduces an Uncertainty-Aware Cross-Modal Graph Attention (UACMGA) module for enhanced multimodal aspect-based sentiment classification. Initially, the text features are extracted through Bidirectional Encoder Representations from Transformers (BERT) and the visual features are obtained through the Residual Network (ResNet-50) encoder. The UACMGA module then fuses these features while explicitly modelling uncertainty, which suppress noise and helps retain useful information during cross-modal interactions. A Graph Convolutional Network (GCN) is subsequently employed to predict sentiment polarities at the aspect level. The proposed model provides a strong answer to practical multimodal sentiment analysis problems, as shown by the experimental findings, which show that it achieves better performance in recall, accuracy, precision, and F1-score.
Keywords: Aspect-Based Sentiment Analysis, Multimodality, BERT, Feature Extraction, ResNet-50.

1. Introduction

Extracting valuable insights from online user-generated content plays a key role in numerous real-world applications. For example, studying customer sentiments in e-commerce reviews helps businesses improve their offerings and optimize their marketing approaches. Hence, developing an automated computational framework to interpret opinions embedded in unstructured text has become essential, giving rise to the research area known as Sentiment Analysis (SA) [1]. The primary function of SA in Natural Language Processing (NLP) is emotion detection in text data. Conventional SA methods usually attempt to ascertain the overarching tone of the text by making sentiment predictions at the level of individual sentences or documents [2][3]. These approaches frequently presume, without providing evidence, that the text expresses a univalent opinion on a certain subject. To address this limitation, there has been growing interest in more detailed, aspect-level sentiment detection known as ABSA it has become increasingly popular in the last ten years [4]. Unlike traditional sentiment analysis, which assesses a document's tone in its entirety, ABSA delves deeper towards understand the sentiment expressed to individual aspects. Fig. 1 shows an example of ABSA with its key terms.

Early research in ABSA primarily focused on identifying each sentiment component independently. To illustrate, the Aspect Term Extraction (ATE) algorithm [5] involves extracting

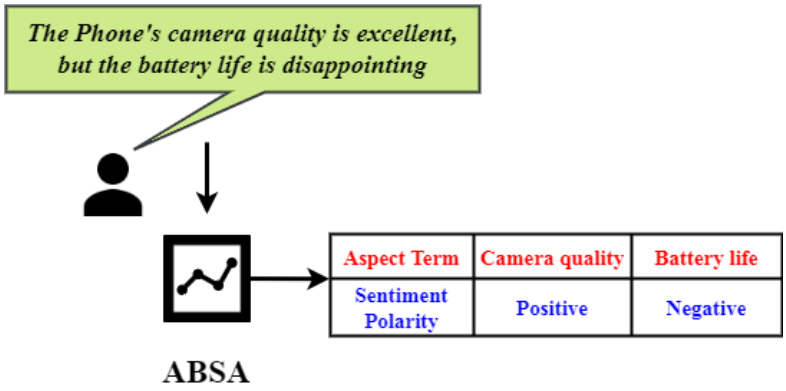


Figure. 1 Key elements of ABSA

all components of a text that are used, while the Aspect Sentiment Classification (ASC) [6] finds out

the emotional tone of a particular phrase clause. Together, these are referred to as single-modal ABSA model.

Yet, identifying an individual sentiment factor remains insufficient for fully understanding aspect-level opinions, especially when opinions are expressed through multiple modalities such as text and images [7]. Fully capturing these opinions requires not only extracting multiple sentiment elements but also recognizing the relationships and dependencies between them across different data types [8]. To address this, a novel multi-modal ABSA (ABMSA) recently implemented to make it easier for prediction of sentiment elements from different modalities. In online reviews and social media posts, people often express their opinions using a combination of images, text, audio, and video [9], as it allows users to convey sentiments more vividly and contextually through written comments, spoken feedback, facial expressions, tone of voice, and even visual content such as photos or short videos [10].

Among these modalities, text and images have proven to be particularly effective and widely used in conveying sentiment, especially on platforms like Instagram, Twitter, and product review sites [11]. But, effectually combining features of different modalities is still a significant challenge. Previous research used pre-trained methods like BERT for text and ResNet-50 for images to extract appropriate features besides employed attention mechanisms to combine the features of both text and images [12]. However, traditional fusion strategies such as simple feature concatenation or basic attention-based fusion often fail to capture complex cross-modal relationships [13]. However, these techniques either select only the top related cross-modal features or apply multilevel fusions, which lead to loss of some informative feature nodes or increasing computational complexity.

To solve these challenges, this study proposes Attention-aware Graph Convolutional Multimodal Aspect-based Sentiment Analysis Network (AGCMASN) for multimodal sentiment classification. The text and image features acquired by BERT and ResNet-50, respectively, the proposed Uncertainty-Aware Cross-Modal Graph Attention (UACMGA) model successfully merges the two modalities by taking uncertainty of cross-modal interactions into consideration. Then, aspect-level sentiment polarities are predicted based on the fused representation by using a Graph Convolutional Network (GCN). The proposed AGCMASN aims at eliminating these redundant and noisy features and retaining the informative ones for better sentiment analysis across multimode representation.

2. Literature survey

The current literature of ABMSA comprises two key subtasks: Multimodal ATE (MATE) and Multimodal ASC (MASC).

A two-step coarse-to-fine Image-Target Matching (ITM) system was suggested in [14] where a rough selection of the relevant image parts is first performed and subsequently the finer structure, the visual objects associated with the target, is classified. The accuracy and F1-score are not very high because of its complex nature of structure.

In [15], Global–Local Features Fusion with Co-Attention (GLFFCA) technique was introduced. In this method, the global module was employed to capture aspect-based global correlations, and the local module was employed to seize the inter-region local semantic alignment. Finally, a feature fusion module integrated both features for final sentiment classification. Nonetheless the accuracy and macro-F1 is rather low.

A variety of multimodal few-shot datasets and a Generative Multimodal Prompt (GMP) [16] was created, which consists of a Multimodal Encoder and N-Stream Decoders. There is also a minor-activity that predicts the

aspects per instance to inform prompt construction. Nonetheless, F1-score is not high because the model is not trained on the sub task.

In [17] a framework for MASC called Chimera was presented that integrates patch-word alignment, fine- and coarse-grained visual features, and textual descriptions to enhance sentiment understanding. It leverages a Large Language Model to infer sentiment causes and affective-cognitive resonance from semantic and visual content. However, its performance requires rationales to be produced of high quality and might therefore lack robustness in various multimodal contexts.

For Joint Multimodal Aspect Based Sentiment Analysis, Aesthetic-oriented Multiple Granularities Fusion Network (Atlantis) [21], was introduced focusing on integrating textual-visual alignment and image aesthetic features on aspect-sentiment extraction. But it complicates the model because of having multiple branches and integrating aesthetic features into the model.

AESAL [22] was implemented for Joint Multimodal Aspect-Based Sentiment Analysis via Aspect Enhancement and Syntactic Adaptive Learning. But it is very much syntactically dependent parser which might affect the performance on noisy social media text. To improve aspect-based sentiment analysis of multimodal inputs, a Text-Dominant Speech-Enhanced Network (TDSEN) [23] was proposed which enhanced the features of speech with text-directed multimodal fusion. However, it has to depend on generated speech features that could result in noise in sentiment prediction.

To improve the performance of multimodal sentiment analysis, an aspect enhancement and text simplification (AETS) [24] model is presented in which the representation of aspects is enhanced, and the multi-aspect text is simplified. However, this depends heavily on syntactic dependency parsing that could decrease performance for more complex and noisy texts.

DaNet [25] was suggested as Dual-Aware Enhanced Alignment Network for Multi modal Aspect-based Sentiment Analysis by fine-grained Text-image Alignment and Multi modal Noise Reduction. However, it still needs extra alignment, denoising modules, thus making the overall model complex.

3. Proposed methodology

The working principle of the proposed AGCMASN model is presented in this section. Fig. 2 portrays the whole architecture of the research.

3.1 Problem definition

This research addresses the challenges of ABMSA, where the predicting the polarity of sentiment of a certain aspect term using both the textual and the visual predictors. Each data instance comprises a text segment T containing the aspect term a, and an associated image I. The sentiment label$y \in \{positive, negative, neutral\}$ corresponds specifically to the aspect a. The goal is to learn a function that maps the triplet (T, I, a) to the sentiment label y. To achieve this, textual features are extracted using the BERT, while image features are obtained via a ResNet-50 encoder. These modality-specific representations are fused using an UACMGA module, which accounts for uncertainty in cross-modal interactions to improve the relevance and robustness of the combined features. After the fused representation is sent via a GCN, each aspect undergoes fine-grained emotion classification.

3.2 Extraction of Multimodal Features

3.2.1 Textual Feature Extraction with BERT

To perform aspect-aware textual feature extraction, the BERT model is employed. BERT is a pre-trained language model containing L transformer layers that generates rich contextualized embeddings by processing the input through multiple transformer layers [18]. Fig. 3 illustrates structure of BERT. Given an input sentence S, which includes a series of tokens S = $\{w_1, w_2, \dots, w_n\}$, BERT produces a collection of hidden states $H = \{h_1, h_2, \dots, h_n\}$, where h_i will be the contextual representation of the i^{th} token in the final transformer layer:

$$BERT(S)_i = h_i^L \quad (1)$$

The output from the BERT encoder is the sequence of text features which is shown below. $h^t = \{h_1^t, h_2^t, \dots, h_n^t\}$ (2)

3.2.2 Image Feature Extraction Using ResNet-50

To extract visual features from images, the ResNet-50 model is utilized due to its robust feature representation capabilities and high generalization performance in image sentiment analysis. Fig. 4 shows the structure of ResNet-50. It uses residual learning to fix deep networks degradation issue, employing residual blocks

that directly learn the residuals between inputs and outputs [19]. For the l^{th} residual block, the input x_l and output x_{l+1} are

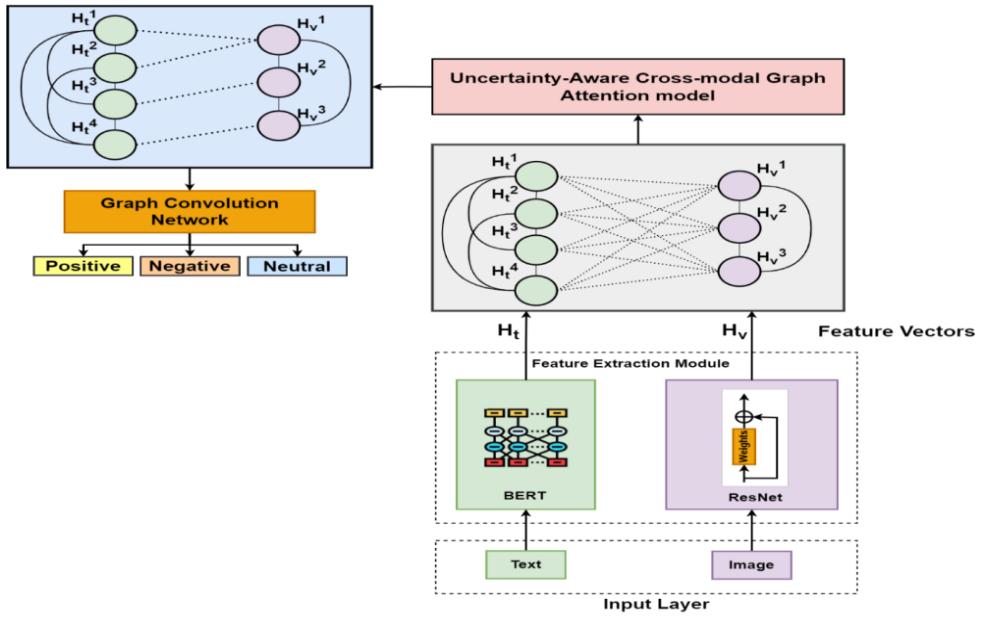


Figure. 2 Overall Architecture of the research

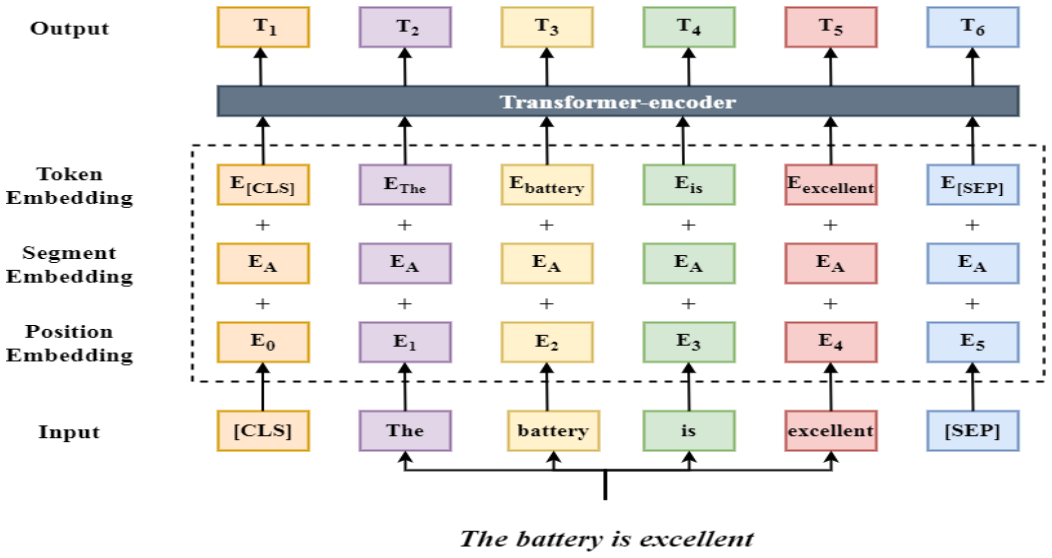


Figure. 3 Structure of BERT

connected by the residual function $F(x_l, W_l)$, computed by:

$$x_{l+1} = F(x_l, W_l) + x_l \quad (3)$$

In Eq. (3), $F(x_l, W_l)$ typically includes a convolutional layer, batch normalization, and ReLU activation, and W_l denotes the learnable weights of the residual function. By stacking multiple such residual blocks, ResNet-50 can construct a deep network that effectively captures complex image features. The output from the ResNet-50 encoder is the sequence of text features which is shown below.

$$h^v = \{h_1^v, h_2^v, \ldots, h_n^v\} \quad (4)$$

3.3 Graph construction

To represent an aspect-level sample with *L* texts and *N* associated images, a heterogeneous graph is constructed where nodes represent both text and image features. The edges capture intra and inter-modal

relationships that contribute to sentiment classification. Node construction involves encoding each text h_t^a using BERT and each image h_i^b using ResNet-50, resulting in $L + N$ total nodes per input instance.

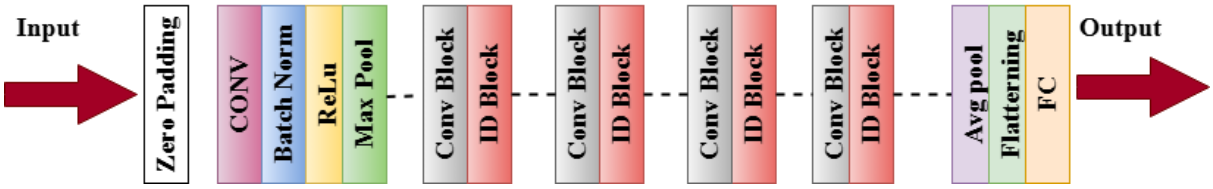


Figure. 4 Structure of ResNet-50 Encoder

Edge construction is guided by attention mechanisms that model dependencies and relevance among nodes. For text-text and image-image relations, attention weights are computed to capture semantic and visual coherence. This is achieved by UACMGA module.

3.4 Uncertainty-Aware Cross-Modal Graph Attention (UACMGA) module

In this module, an enhanced graph attention mechanism is designed to effectively capture both intra- and inter-modal relationships while incorporating uncertainty into the attention computation. In the heterogeneous graph that is generated from Eq. (2) - Eq. (4), a collection of multimodal node features is fed into the UACMGA layer. In this $\overrightarrow{h_i^t} \in \mathbb{R}^F$ and $\overrightarrow{h_i^v} \in \mathbb{R}^F$ represents either text or an image, and the input feature dimension is denoted by F. Fig. 5 shows the working of the module.

The graph contains three distinct types of edges:

1. Text nodes are connected by unimodal edges,
2. Images are connected by unimodal edges., and
3. Linking nodes that include text and images through cross-modal edges.

An edge-weighting is applied to distinguish between the three relations symbolized by the edges, as:

$$A_{(i,j)} = \begin{cases} \mu \times A^t & if\ a<L, b<L \\ \gamma \times A^v & if\ a \geq L, b \geq L \\ A^{tv} & if\ a<L, b \geq L \\ A^{vt} & if\ a \geq L, b<L \end{cases} \quad (5)$$

An adjacency matrix A connects the nodes, and hyper parameters μ and γ are used in this context.

Each node's feature is first linearly transformed via a learnable projection matrix $W \in \mathbb{R}^{F' \times F}$, generating projected node features $W\vec{h}_i^t$ for text and $W\vec{h}_j^v$ for image. For each node of text and image, attention coefficients e_{ij} are computed between a node i and j using a shared attentional mechanism a:

$$e_{ij}^t = a(W\vec{h}_i^t, W\vec{h}_j^t) \quad (6)$$

$$e_{ij}^v = a(W\vec{h}_i^v, W\vec{h}_j^v) \quad (7)$$

$$e_{ij}^{tv} = a(W\vec{h}_i^t, W\vec{h}_j^v) \quad (8)$$

This attention function is modelled as a single-layer feedforward neural network followed by a LeakyReLU activation:

$$e_{ij}^t = LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i^t \parallel W\vec{h}_i^t\right]\right) \quad (9)$$

$$e_{ij}^v = LeakyReLU\left(\vec{a}^T\left[W\vec{h}_j^v \parallel W\vec{h}_j^v\right]\right) \quad (10)$$

$$e_{ij}^{tv} = LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i^t \parallel W\vec{h}_j^v\right]\right) \quad (11)$$

where $\overrightarrow{a} \in \mathbb{R}^{2F'}$ is a learnable weight vector, $.^T$ represents transposition, and $\parallel$ denotes vector concatenation.

These raw attention scores are then normalised over the neighbourhood of node through the softmax function. i:

$$\alpha_{ij}^t = \frac{\exp\left(e_{ij}^t\right)}{\sum_{k \in N_i} \exp(e_{ik}^t)} \quad (12)$$

$$\alpha_{ij}^{v} = \frac{\exp\left(e_{ij}^{v}\right)}{\sum_{k\in N_i}\exp(e_{ik}^{v})} \quad (13)$$

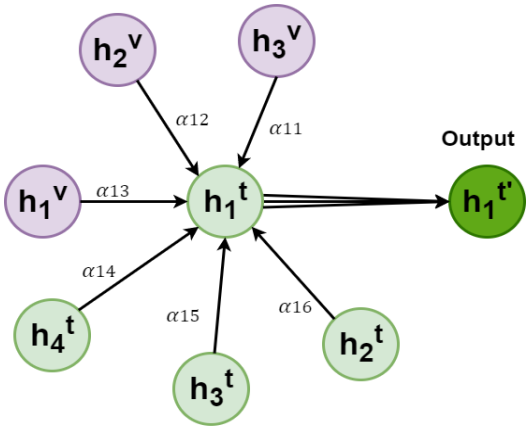


Figure. 5 UACMGA module

$$\alpha_{ij}^{tv} = \frac{\exp\left(e_{ij}^{tv}\right)}{\sum_{k\in N_i}\exp(e_{ik}^{tv})} \quad (14)$$

The final representation for each node is obtained by performing a weighted sum of the corresponding node features using the normalized attention coefficients, which are computed thereafter. A non-linear activation function (σ) may then be applied to this output.

$$\vec{h}_i^{t\prime} = \sigma(\sum_{j\in N_i}\alpha_{ij}^{t} W\vec{h}_j^{t}) \quad (15)$$

$$\vec{h}_i^{v\prime} = \sigma(\sum_{j\in N_i}\alpha_{ij}^{v} W\vec{h}_j^{v}) \quad (16)$$

$$\vec{h}_i^{tv\prime} = \sigma(\sum_{j\in N_i}\alpha_{ij}^{tv} W\vec{h}_j^{tv}) \quad (17)$$

As a next step, the UACMGA model incorporates uncertainty, to modulate the graph attention. It involves inter-modal and intra-modal graph attention operations by selectively regulating information interaction between a confident node and an uncertain node. For noise removal, the uncertainty technique employs two distinct strategies to incorporate uncertainty into the attention mechanism: Cut-off and Suppression.

3.4.1 Uncertainty-Aware Self Attention (UASA)

Let $G = (W, E)$ be a graph with nodes W, and attention computed from query $Q = XW^{Q}$, key $K = XW^{K}$, and value $V = XW^{V}$.

UASA (Cut-off): An uncertainty map $U \in [0,1]^{H\times W}$ is defined, the cut-off variant restricts outgoing connections from nodes with highly uncertain. The attention is computed as follows:

$$UASA_{cut}(Q,K,V) = softmax\left(\frac{QK^{T}}{\sqrt{d}} - M\right)V \quad (18)$$

where $M \in \mathbb{R}^{HW\times HW}$ and defined as:

$$M_{ij} = \begin{cases} 0, if\ U_j \geq \theta \\ \infty, if\ U_j < \theta \end{cases} \quad (19)$$

UASA (Suppression): Since the data from those doubtful nodes might still be valuable, the $UASA_{cut}$ procedure could be overly harsh. Uncertain nodes have had their uncertainties diminished because, as the input passes through several layers, they have previously been updated numerous times by characteristics from confident nodes. The use of a suppression mechanism allows for the gradual reduction, rather than outright blocking, of the impact of unknown nodes:

$$UASA_{sup}(Q,K,V) = softmax\left(\frac{QK^{T}.S}{\sqrt{d}}\right)V \quad (20)$$

where $S \in \mathbb{R}^{HW\times HW}$ and defined as:

$$S_{ij} = \frac{1}{T.U_{ij}(1-T)} \quad (21)$$

In Eq. (21), T signifies hyper-parameter which shows maximum influence.

3.4.2 Uncertainty-Aware Cross-Modal Attention (UACMA)

Let $H_t \in \mathbb{R}^{L\times d}$ and $H_t \in \mathbb{R}^{N\times d}$ illustrate the visual feature matrices and the textual feature matrices, respectively.

UACMA (Cut-off): In Cut-off variant of UACMA, attention from textual nodes to visual nodes is computed with hard masking based on uncertainty scores. The cross-modal attention weights are calculated by:

$$UACMA_{cut}(Q^t, K^v, V^v) = softmax\left(\frac{Q^v K^{v^T}}{\sqrt{d}} - M^{tv}\right)V^v \quad (22)$$

$$M_{ij}^{tv} = \begin{cases} 0, if\ U_j^v \geq \theta \\ \infty, if\ U_j^v < \theta \end{cases} \quad (23)$$

UACMA (Suppression): The calculation for the attention output is:

$$UACMA_{cut}(Q^t, K^v, V^v) = softmax\left(\frac{Q^v K^{v^T}.S_{tv}}{\sqrt{d}}\right)V^v \quad (24)$$

$$S_{ij}^{tv} = \frac{1}{T.U_{ij}^v(1-T)} \quad (25)$$

This selective attention mechanism mitigates the adverse impact of uncertain data, enhancing the robustness and reliability of the learned representations. Consequently, UACMGA enables more accurate and stable fusion of textual and visual modalities.

3.5 Modal training

The presented technique is trained using cross-entropy loss with L2 regularization:

$$\mathcal{L} = -\sum_{x=1}^{N}\sum_{z=1}^{C} y_x^z\, log\hat{y}_x^z + \lambda\|\theta\|^2 \quad (26)$$

In Eq. (27), N represents the total quantity of samples used for training, and C represents number of labels that represent sentiment. $\hat{y}_x^z = \mathbb{R}^C$ indicates the predicted sentiment probability distribution for x^{th} data and y_x^z corresponds to its true sentiment label.

3.6 Final classification using GCN

Finally, a SoftMax function is employed as a classifier to output sentiment label on fused representations:

$$\hat{y} = softmax\ (W_0 d + b_0) \quad (27)$$

Here, $\hat{y} \in \mathbb{R}^C$.

4. Result and discussion

4.1 Dataset description

This research makes use of two multimodal datasets, Twitter-15 and Twitter-17 [20]. These two databases consist of Twitter posts sent by users using many modes between 2014-2015 and 2016-2017. and preserve only posts containing information about people, places, organizations, and other four sorts of entities, including images and word-based context. Since there are no automatically annotated entities in these two multimodal benchmark datasets, three domain experts annotated the sentiment of each entity according to text and image. Subsequently, the data set was separated into train, development, and test sets randomly with a ratio of 60:20:20 respectively. For the experimental evaluations, the data is divided into 80:20 proportions and 80% will be used in training and para testing and 20 % used in testing. The arithmetical features of the Twitter-15 and Twitter-17 are presented in Table 1.

4.2 Implementation details and evaluation metrics

The implementation of both existing and proposed model was carried out in Python 3.11 and executed

Table 1. Descriptive data of Twitter-15 and Twitter-17

	Twitter-15	**Twitter-17**

.

	Training	*Testing*	*Training*	*Testing*
Positive	1231	317	2023	493
Neutral	2553	607	2155	573
Negative	517	113	560	168
Total	4301	1037	4738	1234

on a system with an Intel® CoreTM i5-4210 CPU @ 3GHz, 4GB RAM and a 1TB HDD running on Windows 10 64-bit with the same datasets which is illustrated in section 4.1. Fig. 6 and 7 confusion matrices for Twitter-15 and Twitter-17 for testing set.

The performance measures used in this study are mentioned below.

Accuracy: The metric is defined as the proportion of sentiment labels that were accurately predicted compared to the aggerated of predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (28)$$

Precision: The accuracy rate is the proportion of actual positive estimates related to the sum amount of positive predictions.

$$Precision = \frac{TP}{TP+FP} \quad (29)$$

Recall: It is the fraction of true positives that were accurately appraised compared to the total amount of positives.

$$Recall = \frac{TP}{TP+FN} \quad (30)$$

F1-score: When comparing both recall and precision, it is necessary to find their harmonic mean.

$$F1 - score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (31)$$

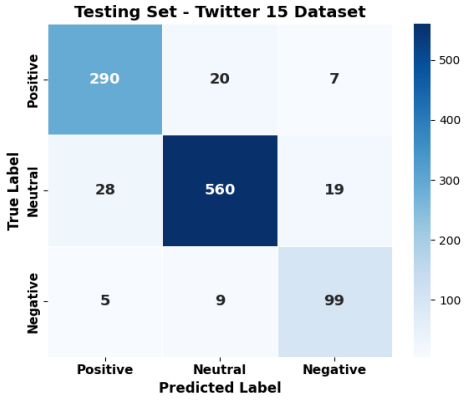


Figure 6. Confusion Matrix for Twitter-15 Dataset

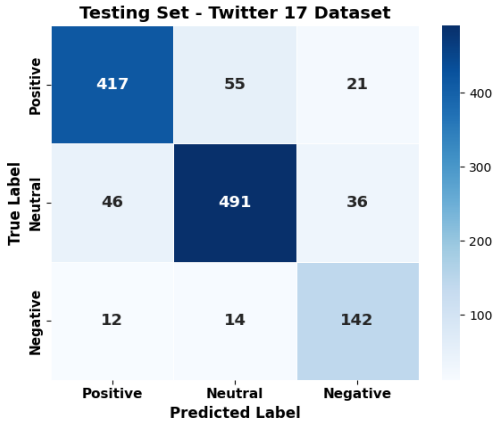


Figure 7. Confusion Matrix for Twitter-17 Dataset

4.3 Performance analysis

The performance of AGCMASN model is evaluated in this section. Using the dataset and metrics described in sections 4.1 and 4.2, the evaluation steps are divided into two sub tasks: MATE and MASC.

4.3.1 Performance of AGCMASN against standard models on MATE

In this subsection, the proposed model is evaluated by comparing it with standard models such as BERT + VGGNet and BERT + ResNet-50 for MATE on Twitter-15 and 17 datasets. In MATE, precision, recall and F1-score are the commonly used measures.

As shown in Fig. 8, AGCMASN consistently outperforms both BERT + VGGNet and BERT + ResNet-50 on the MATE task for all three metrics across the Twitter-15 and Twitter-17 datasets. On Twitter-15, AGCMASN improves precision by roughly 4–6% over the other two models and achieves a recall that is about 3–4% higher. This leads to an F1-score improvement of approximately 3–4%, indicating that AGCMASN captures aspect terms more comprehensively and with greater consistency.

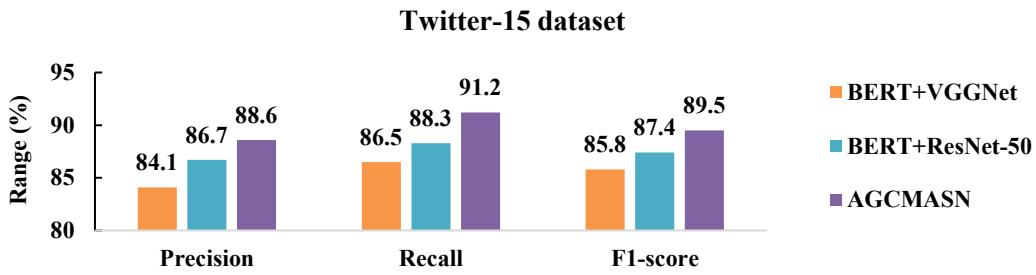


Figure. 8 Performance results of different methods for MATE on Twitter-15 dataset

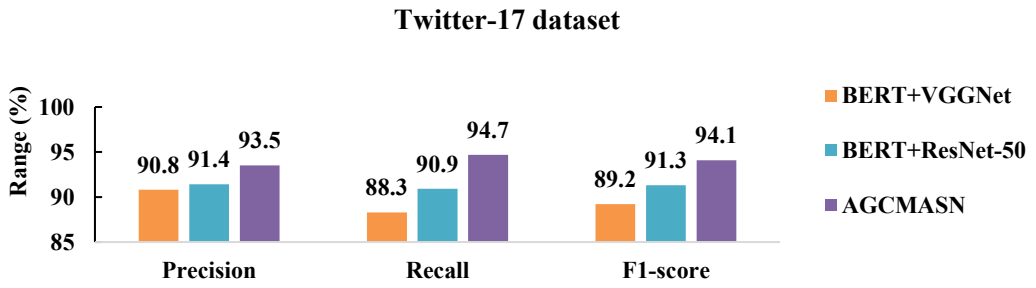


Figure. 9 Performance results of different methods for MATE on Twitter-17 dataset

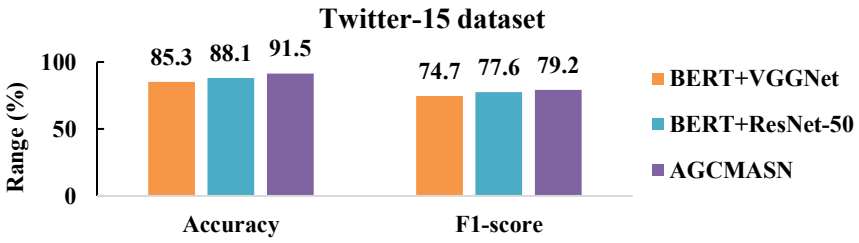
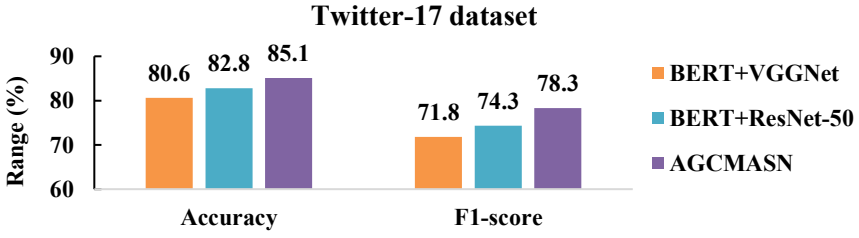


Figure. 10 Performance results of different methods for MASC on Twitter-15 dataset



.

Figure. 11 Performance results of different methods for MASC on Twitter-17 dataset

A similar performance pattern is observed on Twitter-17 which is illustrated on Fig. 9. AGCMASN again demonstrates higher precision with an increase of around 2–3%, while its recall surpasses the competing models by roughly 2%. These gains result in an F1-score advantage of about 2%, showcasing AGCMASN's stable improvements across datasets. The consistent percentage increases suggest that stronger multimodal alignment enables AGCMASN to extract aspect terms more reliably than both BERT + VGGNet and BERT + ResNet-50.

4.3.2 Performance of AGCMASN against standard models on MASC

In this subsection, the AGCMASN model is evaluated by comparing it with same existing model for MASC on Twitter-15 and 17 datasets. In MASC, accuracy and f1-score are the commonly used measures.

As illustrated in Fig. 10, AGCMASN shows clear performance gains over both BERT + VGGNet and BERT + ResNet-50 in the MASC task across the Twitter-15 dataset. AGCMASN achieves an accuracy increase of approximately 5–7% compared to and BERT + ResNet-50 and over 10% relative to BERT + ResNet-50. Its F1-score also improves by around 3–5% over both baselines, indicating more reliable sentiment prediction for extracted aspects. A similar improvement trend is observed Fig. 11 which shows performance on Twitter-17. AGCMASN attains an accuracy gain of roughly 7–10% over the competing models, while its F1-score increases by about 3–6%. These consistent gains across both datasets demonstrate AGCMASN's stronger capability in integrating multimodal cues for more accurate aspect-level sentiment classification, further supporting its overall effectiveness within the MABSA framework.

4.4 Ablation study

To analyse the contribution of individual core factor in the proposed AGCMASN framework, an extensive ablation study is conducted by selectively disabling specific modules and comparing the results with other models. This analysis helps to understand how each element contributes to the overall performance on AGCMASN. Table 2 reports the ablation results obtained by substituting the UACMGA module inside the proposed AGCMASN with several standard attention models, including Self Attention Mechanism (SAM), Channel Attention Mechanism (CAM), and Graph Attention Network (GAT). When the full AGCMASN model

Table 2. MASC results of proposed AGCMASN replaced with UACMGA and other standard attention models

Models	**Accuracy (%)**	
	Twitter-15	**Twitter-17**
AGCMASN with SAM	82.6	78.4
AGCMASN with CAM	85.2	80.3
AGCMASN with GAT	88.9	82.7
AGCMASN (with UACMGA)	**91.5**	**85.1**

Table 3. MASC performance of AGCMASN using different text, image, and multimodal feature extractors

Models	**Accuracy (%)**	
	Twitter-15	**Twitter-17**
Text		
AGCMASN with BoW	81.5	77.5
AGCMASN with Word2Vec	85.7	79.3
AGCMASN with BERT	88.4	82.6
Image		

.

AGCMASN with ResNet-18	84.9	78.4
AGCMASN with ResNet-34	87.1	80.7
AGCMASN with ResNet-50	89.5	83.3
Multimodal (Text + Image)		
AGCMASN	**91.5**	**85.1**

incorporates UACMGA, the performance surpasses all other configurations by 2–3% on both datasets. These improvements clearly indicate that UACMGA plays a crucial role in enhancing multimodal alignment and significantly contributes to the overall effectiveness of the AGCMASN model.

An ablation study to test the performance of the proposed AGCMASN framework with various text, image and multimodal feature extractors was conducted and the results are presented in Table 3. AGCMASN-BERT outperforms BoW and Word2Vec in both Twitter-15 and Twitter-17 datasets without visual content. The text-only setting, AGCMASN-BERT performs the best among the three models BoW, Word2Vec and AGCMASN for both Twitter-15 and Twitter-17 datasets without taking into account any visual details. Likewise, in the image-only setting, AGCMASN with ResNet-50 will perform better than ResNet-18 or ResNet-34. For the text-image multimodal integration task, the proposed AGCMASN obtains the best accuracy of 91.5% on Twitter-15 and 85.1% on Twitter-17 data, which shows that multimodal feature combination is effective to aspect-based sentiment classification.

4.5 Comparative Analysis of Existing and Proposed Models

Table 4 presents the comparative analysis of the proposed AGCMASN model with existing multimodal aspect-based sentiment analysis methods such as Atlantis, AESAL, TDSEN, AETS, and DaNet on the Twitter-15 and Twitter-17 data sets. The existing systems achieve accuracy of 79.3% to 81.3% at Twitter-15 and 74.2% to 79.0% at Twitter-17. Among these approaches, DaNet obtains the highest performance among existing methods. This proposed model of AGCMASN is able to achieve high accuracy for Twitter-15 (91.5%) and Twitter-17 (85.1%), outperforming its competitors using multimodal feature fusion and sentiment classification.

4.6 Aspect-level analysis

In this section, an aspect-level sentiment analysis results are presented using different ABMSA models across image–text instances. For each instance, multiple aspect entities are extracted, and each model predicts a sentiment label for every aspect. These predictions are compared against the ground-truth sentiment annotations.

Tables 5 and 6 present the aspect-level sentiment analysis results for the Twitter-15 and Twitter-17 datasets. The AGCMASN model consistently demonstrates superior performance in both the datasets. In the Twitter-15, AGCMASN correctly predicts all aspect sentiments for all three samples, whereas Atlantis, AESAL, TDSEN, AETS, and DaNet show errors in both neutral and negative classification. Similarly, in the Twitter-17 examples, AGCMASN aligns perfectly with the ground truth across all aspect entities, while the baseline models struggle particularly with distinguishing neutral versus positive sentiment and in handling complex political or event-related contexts.

Table 4. Comparative Analysis of AGCMASN and existing models using multimodal feature extractors

Models	**Accuracy (%)**	
	Twitter-15	**Twitter-17**
Text + Image		
Atlantis [21]	79.3	74.2
AESAL [22]	80.1	78.8
TDSEN [23]	80.3	76.3
AETS [24]	79.5	76.6
DaNet [25]	81.3	79.0

.

AGCMASN	**91.5**	**85.1**

Table 5. Aspect-level sentiment prediction results on the Twitter-15 dataset

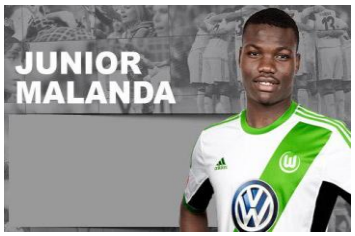
Image			
Text	RT @ MySportsLegion: **LeBron James** is already wearing a **Johnny Football Browns** jersey. (via @ WFNYScott)	BREAKING NEWS : **Wolfsburg** youngster **Junior Malanda** has died in a car accident , according to **BILD**.	Congrats to TWO # MariettaAward winners, **Ashley** amp **Christine** at **PF Indianapolis**, **IN**. Keep making members feel welcome!
Entity	{LeBron James, Johnny, Football Browns}	{ Wolfsburg, Junior Malanda, BILD}	{Ashley, Christine, PF Indianapolis, IN}
Ground Truth	**{NEU, NEU, NEU}**	**{NEU, NEG, NEU}**	**{POS, POS, NEU, NEU}**
Atlantis [21]	**{NEU, POS, NEU}**	**{NEU, NEG, POS}**	**{POS, NEU, NEU, NEU}**
AESAL [22]	**{NEU, NEU, NEG}**	**{POS, NEU, NEU}**	**{NEU, POS, NEU, NEU}**
TDSEN [23]	**{NEG, NEU, POS}**	**{POS, NEG, NEU}**	**{NEG, POS, NEU, POS}**
AETS [24]	**{NEU, POS, NEG}**	**{NEG, NEU, POS}**	**{POS, NEU, NEG, POS}**
DaNet [25]	**{POS, NEG, NEU}**	**{NEU, POS, NEG}**	**{NEU, POS, NEG, POS}**
Proposed AGCMASN	**{NEU, NEU, NEU}**	**{NEU, NEG, NEU}**	**{POS, POS, NEU, NEU}**

Table 6. Aspect-level sentiment prediction results on the Twitter-17 dataset

Image			
Text	# **NFL** Could # **Cowboys** ' **Elliott** be most sought - after player in fantasy football . . .	**Syria** conflict: **Russia** issues warning after **US** coalition downs jet	Enjoyed performing at **Royal Festival Hall**. #**London** #Alchemy2016 #**RoyalFestivalHall** #Weekend #dance #Classical #**Indian**
Entity	{NFL, Cowboys, Elliott}	{Syria, Russia, US}	{Ashley, Christine, PF Indianapolis, IN}
Ground Truth	**{NEU, NEU, POS}**	**{NEG, NEU, NEG}**	**{POS, NEU, POS, NEU}**
Atlantis [21]	**{NEU, NEU, NEU}**	**{NEU, NEG, NEG}**	**{POS, NEU, NEU, POS}**
AESAL [22]	**{NEU, NEG, NEU}**	**{NEG, NEU, NEU}**	**{NEU, POS, NEG, NEU}**
TDSEN [23]	**{NEU, NEU, POS}**	**{POS, NEG, NEU}**	**{POS, NEG, POS, NEU}**
AETS [24]	**{NEU, NEG, POS}**	**{NEG, NEG, NEU}**	**{POS, NEU, NEU, NEU}**
DaNet [25]	**{NEG, NEU, NEU}**	**{NEU, NEU, NEG}**	**{NEU, NEU, POS, NEU}**

Proposed AGCMASN	**{NEU, NEU, POS}**	**{NEG, NEU, NEG}**	**{POS, NEU, POS, NEU}**

5 Conclusion

This work presents the AGCMASN model to effectively address challenges in ABMSA. By leveraging uncertainty modelling during multimodal fusion, AGCMASN preserves informative features from both textual and visual inputs, extracted via BERT and ResNet-50 respectively. The fused representation, processed through a GCN, enhances sentiment classification performance while mitigating the limitations of prior attention-based graph methods. Experimental results demonstrate that AGCMASN achieves superior accuracy and robustness compared to existing approaches, validating its effectiveness in capturing nuanced sentiment indications across modalities.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Amitha and Priya; Methodology, Anitha; Software, Amitha; Validation, Amitha, Priya, Formal Analysis, Amitha; Investigation, Amitha; Resources, Amitha; Writing—Original Draft Preparation, Amitha; Writing—Review and Editing, Amitha; Visualization, Amitha; Supervision, Priya.

References

1. B. Liu, Sentiment Analysis and Opinion Mining. *Springer Nature*, 2022.
2. A. Yessenalina, Y. Yue, and C. Cardie, "Multi-level structured models for document-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.,* 2010, pp. 1046–1056.
3. B. Abimbola, E. de La Cal Marin, and Q. Tan, "Enhancing legal sentiment analysis: A convolutional neural network–long short-term memory document-level model*," Mach. Learn. Knowl. Extr.,* vol. 6, no. 2, pp. 877–897, 2024.
4. K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.,* vol. 28, no. 3, pp. 813–830, 2016.
5. H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu, "A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction," *Neurocomputing*, vol. 419, pp. 344–356, 2021.
6. G. Xu, Z. Zhang, T. Zhang, S. Yu, Y. Meng, and S. Chen, "Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning*," Knowl.-Based Syst.,* vol. 245, p. 108586, 2022.
7. R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Comput. Surv.,* vol. 55, no. 13s, pp. 1–38, 2023.
8. M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.,* vol. 55, no. 7, pp. 5731–5780, 2022.
9. P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining,* vol. 11, no. 1, p. 81, 2021.
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," *J. Comput. Social Sci.,* vol. 7, no. 3, pp. 2767–2838, 2024.
11. H. Li, H. Ji, H. Liu, D. Cai, and H. Gao, "Is a picture worth a thousand words? Understanding the role of review photo sentiment and text-photo sentiment disparity using deep learning algorithms," *Tourism Manage.,* vol. 92, p. 104559, 2022.
12. J. Ren, "Multimodal sentiment analysis based on BERT and ResNet," *arXiv preprint arXiv:*2412.03625, 2024.
13. B. Xu, C. Xu, and B. Su, "Cross-modal graph attention network for entity alignment," *in Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3715–3723.
14. J. Yu, J. Wang, R. Xia, and J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching," *in Proc. Int. Joint Conf. Artif. Intell.* (IJCAI), Jul. 2022, pp. 4482–4488.
15. S. Wang, G. Cai, and G. Lv, "Aspect-level multimodal sentiment analysis based on co-attention fusion*," J. Image Graph.,* vol. 28, no. 12, pp. 3838–3854, 2023.
16. X. Yang et al., "Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt," *in Findings Assoc. Comput. Linguistics: ACL 2023*, Jul. 2023, pp. 11575–11589.
17. L. Xiao et al., "Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis," *IEEE Trans. Affect. Comput.,* 2025.
18. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv*:1810.04805, 2019.
19. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), 2016, pp. 770–778.
20. https://www.researchgate.net/figure/Statistics-of-Twitter-15-and-Twitter-17-datasets-Aspects-avg-number-of-aspects-per_tbl1_393783988 3.
21. L. Xiao, X. Wu, J. Xu, W. Li, C. Jin, and L. He, "Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis," *Information Fusion*, vol. 106, p. 102304, 2024.

.

22. L. Zhu, H. Sun, Q. Gao, T. Yi, and L. He, “Joint Multimodal Aspect Sentiment Analysis with Aspect Enhancement and Syntactic Adaptive Learning,” In *IJCAI*, pp. 6678-6686, 2024.
23. X. Huang, H. Sun, X. Liu, J. Wu, and L. He, “Text-dominant speech-enhanced for multimodal aspect-based sentiment analysis network,” *Information Fusion*, p. 103543, 2025.
24. L. Zhu, H. Sun, Q. Gao, Y. Liu, and L. He, “Aspect enhancement and text simplification in multimodal aspect-based sentiment analysis for multi-aspect and multi-sentiment scenarios,” In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 39, No. 2, pp. 1683-1691, 2025.
25. A. Zhu, M. Hu, X. Wang, J. Yang, Y. Tang, and N. An, “DaNet: Dual-Aware Enhanced Alignment Network for Multimodal Aspect-Based Sentiment Analysis,” In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14369-14381, 2025.