



ARTICLE

# Text-Image Feature Fine-Grained Learning for Joint Multimodal Aspect-Based Sentiment Analysis

Tianzhi Zhang<sup>1</sup>, Gang Zhou<sup>1,\*</sup>, Shuang Zhang<sup>2</sup>, Shunhang Li<sup>1</sup>, Yepeng Sun<sup>1</sup>, Qiankun Pi<sup>1</sup> and Shuo Liu<sup>3</sup>

<sup>1</sup>School of Data and Target Engineering, Information Engineering University, Zhengzhou, 450001, China

<sup>2</sup>Information Engineering Department, Liaoning Provincial College of Communications, Shenyang, 110122, China

<sup>3</sup>School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450000, China

\*Corresponding Author: Gang Zhou. Email: gzhougzhou@126.com

Received: 10 July 2024 Accepted: 05 October 2024 Published: 03 January 2025

## ABSTRACT

Joint Multimodal Aspect-based Sentiment Analysis (JMASA) is a significant task in the research of multimodal fine-grained sentiment analysis, which combines two subtasks: Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). Currently, most existing models for JMASA only perform text and image feature encoding from a basic level, but often neglect the in-depth analysis of unimodal intrinsic features, which may lead to the low accuracy of aspect term extraction and the poor ability of sentiment prediction due to the insufficient learning of intra-modal features. Given this problem, we propose a Text-Image Feature Fine-grained Learning (TIFFL) model for JMASA. First, we construct an enhanced adjacency matrix of word dependencies and adopt graph convolutional network to learn the syntactic structure features for text, which addresses the context interference problem of identifying different aspect terms. Then, the adjective-noun pairs extracted from image are introduced to enable the semantic representation of visual features more intuitive, which addresses the ambiguous semantic extraction problem during image feature learning. Thereby, the model performance of aspect term extraction and sentiment polarity prediction can be further optimized and enhanced. Experiments on two Twitter benchmark datasets demonstrate that TIFFL achieves competitive results for JMASA, MATE and MASC, thus validating the effectiveness of our proposed methods.

## KEYWORDS

Multimodal sentiment analysis; aspect-based sentiment analysis; feature fine-grained learning; graph convolutional network; adjective-noun pairs

## 1 Introduction

With the wide application of intelligent mobile terminals around the world, more and more people are inclined to publish information that includes multiple modalities, such as text and image, to express their opinions and sentiments in response to various events and topics [1]. This situation and trend make sentiment analysis become one of the most popular research tasks at present. However, capturing and fusing the above different modal information has created new challenges for sentiment analysis, thus giving birth to the emerging research area of Multimodal Sentiment Analysis (MSA) [2,3]. MSA



has been widely concerned by academics, businesses, governmental organizations, and public services in recent years due to its improved sentiment analysis accuracy and enhanced sentiment understanding comprehensiveness. While Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to analyze the sentiment of aspect terms in each sample, which select the specific noun phrases in text as aspect terms, and each text includes an indefinite number of aspect terms.

Nowadays, researchers have further proposed Joint Multimodal Aspect-based Sentiment Analysis (JMASA) based on the MABSA task, which can be divided into two subtasks: Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-oriented Sentiment Classification (MASC). Table 1 shows two representative examples of JMASA: Table 1(a) extracts the aspect terms “Ashford Town Ladies” and “Newquay” by combining text and image semantics, and predicts their positive and neutral sentiments, respectively, through the context in text and the scene in image. Table 1(b) also extracts the aspect terms “Stephen Curry” and “NBA” through text and image semantics, and infers their positive and neutral sentiments, respectively, by combining the words in text, the facial expression as well as the NBA scene in image.

**Table 1:** Two representative examples of JMASA

Image		
Text	(a) Congratulations to Ashford town ladies winners of the 2016 Newquay tournament # footballtour # newquaysixes	(b) Stephen Curry just played the best overtime in # NBA history - SB Nation
Output	(Ashford town ladies, Positive) (Newquay, Neutral)	(Stephen Curry, Positive) (NBA, Neutral)

As a significant multimodal fine-grained sentiment analysis task, JMASA has received extensive academic attention, and numerous research methods have been proposed in these years. For example, Ju et al. [4] controlled the rational utilization of visual information by employing a text-image relation detection method, Ling et al. [5] simplified all pretraining and downstream tasks by designing a unified multimodal encoder-decoder architecture, Yang et al. [6] enhanced model performance on the target tasks by setting auxiliary supervision for text and image, respectively, and Wang et al. [7] bridged the semantic gap between text and image representations by optimizing the scalar weight of balancing their features. However, most current models only perform the basic text and image feature encoding and often neglect the further analysis of unimodal intrinsic features, which may lead to the low accuracy of aspect term extraction and the poor ability of sentiment prediction due to the insufficient learning of unimodal features. Furthermore, JMASA treats text as the dominant modality, and image as an additional modality that assists with its semantic expression can provide important clues for text to some extent, but there may also be extra noise introduced by the image information unrelated to text semantics, so the feature fusion strategy between modalities is also a key factor that affects the overall model performance.

Given the above problems, we propose a JMASA model Text-Image Feature Fine-Grained Learning (TIFFL). Firstly, we adopt pretrained text and image encoders for the text-image multimodal sample to obtain unimodal feature representations. Secondly, a gating mechanism is constructed to prevent the visual features unrelated to text semantics from interfering with our model. Then, we introduce Graph Convolutional Network (GCN) [8] and Adjective-Noun Pairs (ANPs) [9] to better learn and represent the intrinsic features of text and image, respectively. Finally, an effective inter-modal fusion strategy is designed to generate the final representations of text and image features to further achieve aspect term extraction and sentiment polarity prediction. Our contributions to TIFFL are as follows:

- To promote the effective fusion of text and image information, a multimodal feature correlation discrimination module is proposed, which constructs a gating mechanism for the dynamic input of visual features by calculating the correlation degree of text and image semantics, while prevents the image information unrelated to text semantics from introducing extra noise.
- To further enhance the learning and representation of text and image intrinsic features, we adopt GCN to obtain the syntactic structure features of text, which addresses the context interference problem of identifying different aspect terms by raising the attention to noun phrases and calculating the sentiment scores between dependent words, while introduce image ANPs to enable visual semantic representation more intuitive, thus addressing the ambiguous problem of image semantic extraction.
- Experimental results on two Twitter benchmark datasets show that our model outperforms most unimodal and multimodal associated studies, with competitive results on the JMASA task as well as the two subtasks MATE and MASC.

## 2 Related Work

Early sentiment analysis studies were mostly conducted on unimodal forms such as text and image [10–12]. Over the years, MSA has emerged as a crucial research area in sentiment analysis, while JMASA has been further advanced and refined based on the MABSA task.

### 2.1 Graph Neural Network (GNN)

Graph Neural Network (GNN) has previously achieved excellent results in many Natural Language Processing (NLP) tasks including aspect-based sentiment analysis. Zhang et al. [13] proposed a syntactic dependency tree based GCN to obtain the contextual syntactic information and word dependencies of aspect term. Huang et al. [14] proposed a target-dependent Graph Attention network (GAT) to learn the sentiment information of aspect term by exploring contextual word dependencies. Sun et al. [15] stacked a GCN layer on Long Short-Term Memory (LSTM) network [16], which employs Bidirectional LSTM (BiLSTM) network to learn the contextual features of text and further perform convolutions over a dependency tree to extract the richer representations. Tang et al. [17] jointly considered the flat and graph-based representations in an iterative interaction manner by a dependency graph enhanced dual-transformer network. Wang et al. [18] proposed an aspect-oriented tree network that focuses on aspect terms by reshaping and pruning ordinary dependency trees. However, the above methods ignore the sentiment information between context words and aspect terms, which can directly demonstrate the sentiment expression for a specific aspect term of text.

## **2.2 *Multimodal Sentiment Analysis (MSA)***

MSA has received considerable academic attention over the last few years [19,20], which implements model construction by combining text with non-text information and is typically divided into two subtasks: conversation MSA and social media MSA. In conversation MSA, current studies mainly model information interactions in different modalities by adopting various deep learning methods such as LSTM, Gated Recurrent Unit (GRU) [21], Convolutional Neural Network (CNN) [22] and Transformer [23], which have been demonstrated superior performance in multiple sentiment related tasks such as sentiment analysis [24–26], emotion analysis [27,28] and sarcasm detection [29,30]. In social media MSA, major studies include social media image sentiment analysis [8,31,32] and text-image integrated sentiment analysis [33–35]. While these studies are applicable to coarse-grained global sentiment analysis, they cannot provide direct utilization for the fine-grained tasks.

## **2.3 *Multimodal Aspect-Based Sentiment Analysis (MABSA)***

To effectively exploit different modal information for aspect-based sentiment analysis, researchers have proposed numerous MABSA models in the past few years by utilizing different methods in various text and image tasks. Xu et al. [36] first discussed the multimodal fine-grained sentiment analysis task and proposed a text-image interaction model MIMN based on multi-interactive BiLSTM network, which can also be extended to the MABSA task, while also built an e-commerce comment dataset ZOL for the experiments. Yu et al. [37] proposed an entity-sensitive attention and fusion network model ESAFN that captures the aspect-text and aspect-image relations, then also constructed two Twitter benchmark datasets Twitter-15 and Twitter-17. Yu et al. [38] proposed an architecture improved model TomBERT based on the pretrained language model BERT [39] and achieved significant performance enhancement, which has been further cited and refined by multiple subsequent studies. Khan et al. [40] proposed a cross-modal translation model CapBERT by converting the image semantics into captions and captured the sentiment polarity only through text information. Zhao et al. [41] proposed an ANPs-based knowledge enhancement framework KEF and incorporated it into various models to improve their visual attention and sentiment prediction capabilities. Although these methods are applicable to the sentiment analysis of given aspect terms, the aspect terms are generally not directly given in practice, so multimodal aspect term extraction has become a prerequisite for the corresponding sentiment analysis.

## **2.4 *Joint Multimodal Aspect-Based Sentiment Analysis (JMASA)***

JMASA is a significant research task proposed during these years, which is generated by integrating MATE into MABSA. For the MATE task, Wu et al. [42] proposed a region-aware alignment network model RAN that extracts aspect term by aligning text-image entity regions. Yu et al. [43] proposed a Multimodal Named Entity Recognition (MNER) model UMT based on entity span detection, which dynamically captures the text-image information association through cross-modal feature interaction. Wu et al. [44] proposed a MNER model OSCGA based on text-image entity alignment, which achieves entity prediction by designing a neural network that combines image object and text character information. Jia et al. [45] proposed a MNER model MNER-QG based on an end-to-end machine reading comprehension framework, which provides prior knowledge of entity types and visual regions to enhance the text and image representations. For the JMASA task, Ju et al. [4] first proposed an auxiliary cross-modal relation detection model JML, which controls the rational utilization of visual information by designing a text-image relation detection method. Ling et al. [5] proposed a task-specific vision-language pretraining model VLP-MABSA, which simplifies all pretraining and downstream tasks by introducing a unified multimodal encoder-decoder

architecture. Yang et al. [6] proposed a multi-task learning cross-modal Transformer model CMMT, which enhances the model performance by constructing auxiliary supervision modules for text and image, respectively. Wang et al. [7] proposed a self-adaptive attention fusion model SAAF, which bridges the semantic gap between text and image representations by adjusting the scalar weight of balancing their features. Despite the above studies have been demonstrated to be effective on JMASA, they typically only employ pretrained language and vision encoders to achieve text and image feature encoding from a basic level, often neglecting the further analysis of unimodal intrinsic features. Therefore, we propose a novel model to address the low accuracy of aspect term extraction and poor ability of sentiment prediction problems due to the insufficient unimodal feature learning by performing intra-modal feature fine-grained analysis.

### 3 Methodology

In this chapter, we define the JMASA task and provide the overview of our proposed Text-Image Feature Fine-Grained Learning (TIFFL) model, then introduce the specific workflow of each component in TIFFL.

**Task Definition:** Motivated by most previous studies on joint aspect-based sentiment analysis [46–48], we describe the JMASA task as a text sequence labeling, which adopts the *BIO2* tagging schema [49] as aspect term extractor with seven classifications for each token. Specifically, given a set of input multimodal samples  $D = (x_1, x_2, \dots, x_d)$ , each sample  $x_i \in D$  contains an  $m$ -word text  $S = (w_1, w_2, \dots, w_m)$  and a corresponding image  $I$ . Our target task is to obtain the text label sequence  $y = (y_1, y_2, \dots, y_m)$  for each sample, where  $y_i \in \{O, B\text{-POS}, I\text{-POS}, B\text{-NEU}, I\text{-NEU}, B\text{-NEG}, I\text{-NEG}\}$ , O denotes the non-aspect term token label, B denotes the beginning token label of aspect term, I denotes the remaining token label, and POS, NEU and NEG denote the positive, neutral and negative sentiment labels, respectively.

#### 3.1 Overview

Fig. 1 shows the overall architecture of TIFFL, which is divided into four components: (1) Unimodal feature encoding. (2) Multimodal feature correlation discrimination. (3) Intra-modal feature fine-grained analysis. (4) Inter-modal feature fusion and output.

#### 3.2 Unimodal Feature Encoding

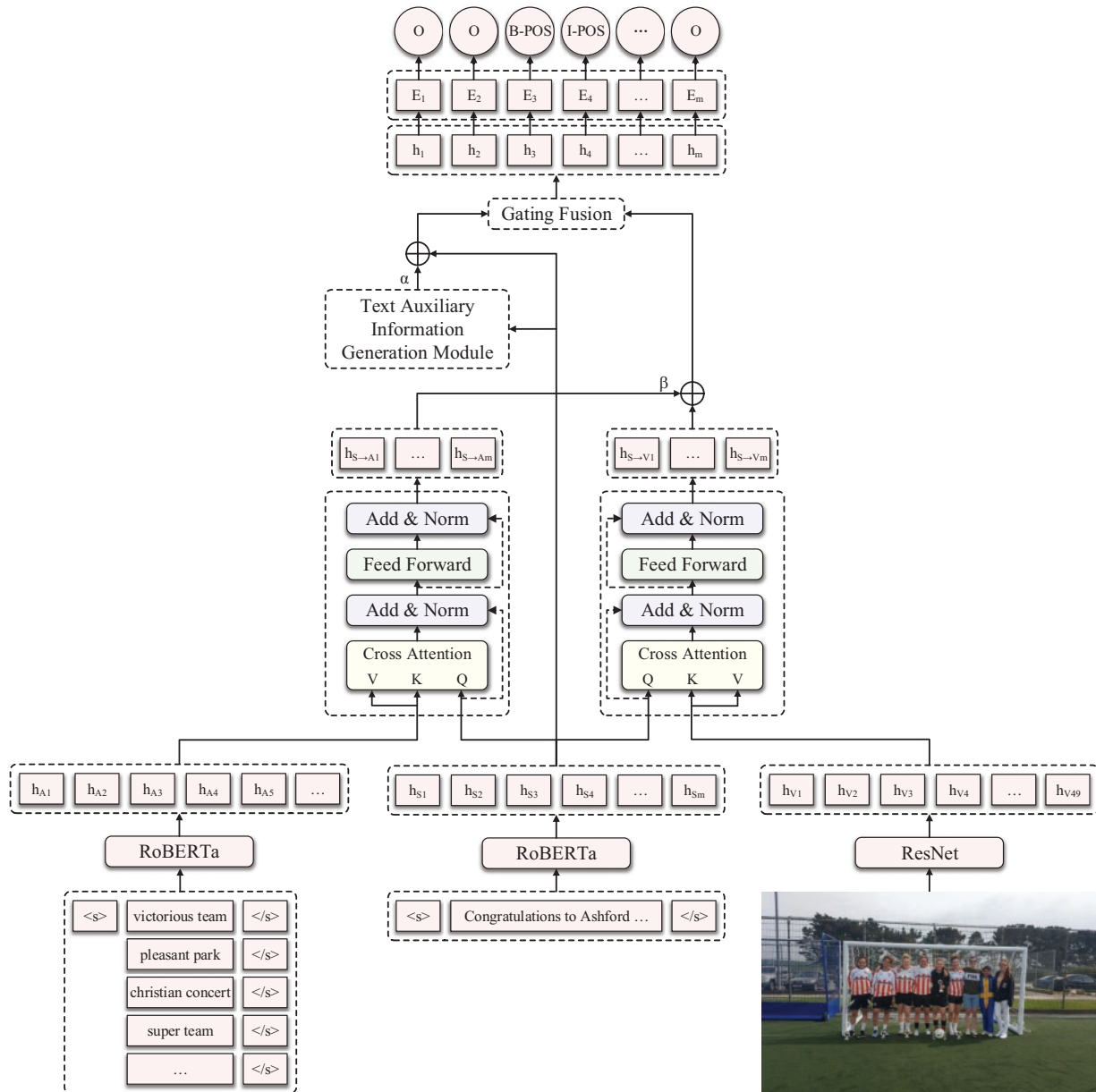
Unimodal feature encoding is the basis of all subsequent tasks in multimodal work, and the extracted feature vectors enable further cross-modal interaction and fusion between them. In this section, we employ pretrained language and vision models to obtain the input text and image feature representations, respectively.

##### 3.2.1 Text Feature Encoding

For the input text feature extraction, we designate the text encoder as the pretrained language model RoBERTa [50], which has produced more favorable results in multiple NLP tasks as an extension and enhancement of BERT. Specifically, we insert two specific tokens “<s>” and “</s>” at the beginning and end of the input text  $S$  as  $S'$ , then feed  $S'$  into RoBERTa to extract the text token representation that incorporates context information:

$$H_S = \text{RoBERTa}(S') \tag{1}$$

where  $H_S \in \mathbb{R}^{d \times m}$ ,  $d$  is the hidden dimension of text representation, and  $m$  is the length of  $S'$ .



**Figure 1:** The overall architecture of TIFFL model

### 3.2.2 Image Feature Encoding

For the input image feature extraction, we designate the image encoder as the pretrained vision model Residual Network (ResNet) [51], which avoids gradient disappearance with the increasing number of layers by employing residual connections. Compared with the VGG [52] network that has been widely adopted in earlier associated studies, ResNet enables a deeper extraction of image semantic information. Specifically, we resize the input image  $I$  to  $224 \times 224$  pixels as  $I'$ , then obtain the image



vision representation from the last convolution layer of the pretrained 152-layer ResNet:

$$H_I = ResNet(I) \quad (2)$$

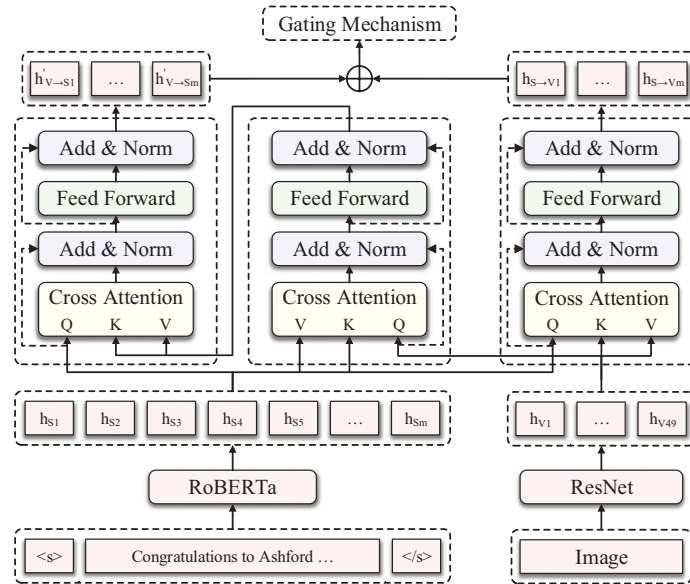
where  $H_I \in \mathbb{R}^{2048 \times 49}$ , 49 is the count of  $7 \times 7$  equal-size vision blocks, and 2048 is the dimension of a vision block. Considering the subsequent cross-modal interactions, we map text and image representations into the same semantic space and perform a linear transformation on  $H_I$  to generate the final image representation:

$$H_V = W_V H_I + b_V \quad (3)$$

where  $W_V \in \mathbb{R}^{d \times 2048}$  and  $b_V \in \mathbb{R}^d$  are the learnable linear transformation parameters.

### 3.3 Multimodal Feature Correlation Discrimination

Although image can provide the model with information other than text, our aim is to extract aspect terms and predict sentiment polarities with the image assistance, and the image information unrelated to text semantics not only fails to assist text in accomplishing the target task but may also introduce extra noise. In this component, we design a Multimodal Feature Correlation Discrimination (MFCD) module to better achieve the text and image information fusion by constructing an Image Gating Mechanism (IGM) for visual features. The internal architecture of MFCD is shown in Fig. 2, which consists of two layers: (1) Cross-modal feature interaction. (2) Image gating mechanism construction. Compared with mostly existing methods that directly integrate inter-modal information, our MFCD promotes the effective fusion of text and image information by calculating the correlation degree of their semantics to perform filtering.



**Figure 2:** The workflow of MFCD module

#### 3.3.1 Cross-Modal Feature Interaction

With the purpose of learning text feature representation in image, we employ the Multi-head Cross-modal Attention (MCATT) [53] mechanism that has multiple attention heads focusing on

different features to capture multimodal complex associations, and treat the image representation  $H_V$  as Query (Q), the text representation  $H_S$  as Key (K) and Value (V), then obtain the image-aware text representation  $H_{V \rightarrow S}$  by two Layer Normalization (LN) [54] and one Feed-Forward Network (FFN) [17]:

$$Z_{V \rightarrow S} = LN(H_V + MCATT(H_V, H_S)) \quad (4)$$

$$H_{V \rightarrow S} = LN(Z_{V \rightarrow S} + FFN(Z_{V \rightarrow S})) \quad (5)$$

where  $H_{V \rightarrow S} \in \mathbb{R}^{d \times 49}$ . However,  $H_V$  is treated as Q in this MCATT and the individual vector of  $H_{V \rightarrow S}$  is represented in form of a vision block. Considering the subsequent construction of gating mechanism, it is necessary to convert each vector into a token representation. Therefore, we employ another MCATT by treating  $H_S$  as Q,  $H_{V \rightarrow S}$  as K and V to generate the final image-aware text representation  $H'_{V \rightarrow S}$ , where  $H'_{V \rightarrow S} \in \mathbb{R}^{d \times m}$ .

For learning the image feature representation of each token in text, we employ the same cross-modal interaction method as the MCATT described above, and treat  $H_S$  as Q,  $H_V$  as K and V to generate the text-aware image representation  $H_{S \rightarrow V}$ , where  $H_{S \rightarrow V} \in \mathbb{R}^{d \times m}$ .

### 3.3.2 Image Gating Mechanism Construction

In a previous MNER study, Yu et al. [43] controlled the visual feature contribution to each token in text by constructing an image gate and achieved effective results in a series of experiments. Motivated by this work, we also decide to introduce an image gating mechanism, which serves to dynamically control the contribution of image information by assigning correlation weights to its features in  $[0, 1]$  with corresponding text. Specifically, we concatenate the above  $H'_{V \rightarrow S}$  and  $H_{S \rightarrow V}$ , then construct the gating mechanism by linear transformation and nonlinear activation:

$$g = \sigma(W_g [H'_{V \rightarrow S}; H_{S \rightarrow V}] + b_g) \quad (6)$$

where  $W_g \in \mathbb{R}^{d \times 2d}$  and  $b_g \in \mathbb{R}^d$  are the learnable linear transformation parameters, and  $\sigma$  is the element-wise nonlinear activation that controls the gating mechanism output in  $[0, 1]$ . The generated  $g$  is a weight vector where all element values are between 0 and 1, with element values close to 1 in the regions of high text-image correlation and close to 0 in the regions of low correlation, thus its subsequent multiplication with image associated representations can filter out the image information unrelated to text semantics.

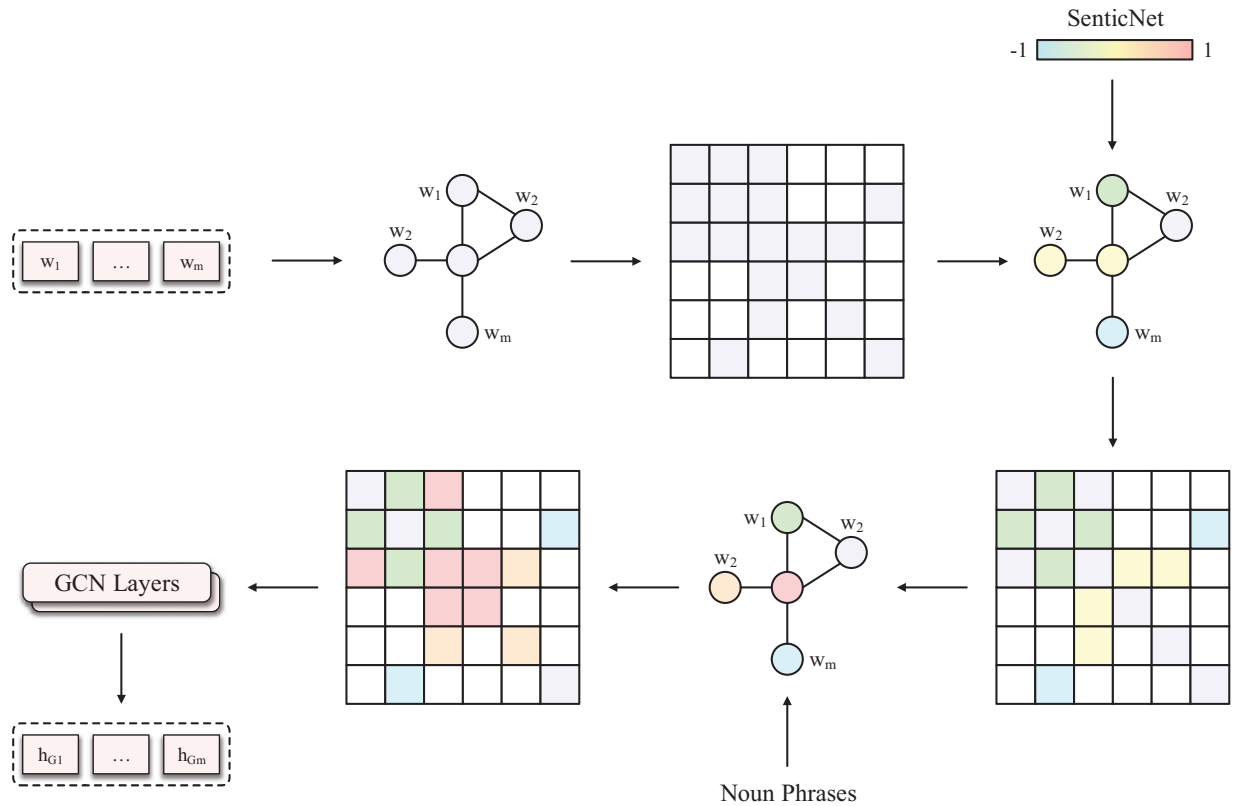
### 3.4 Intra-Modal Feature Fine-Grained Analysis

For the previous studies on JMASA, most methods typically implement text and image feature encoding from a basic level by employing pretrained language and vision encoders, which only guarantee the coarse-grained learning and representation of individual modal features but are not sufficient for a deeper understanding of the text internal structure and image semantic information, thus may lead to some impact on the model performance. To address this problem, we perform a more in-depth analysis and research on the unimodal intrinsic features, including: (1) Constructing a Text Auxiliary Information (TAI) based on sentiment-enhanced GCN to learn the syntactic structure features of text; (2) Constructing an Image Auxiliary Information (IAI) by adopting ANPs to assist the image semantic representation from the text level.



### 3.4.1 Text Auxiliary Information Based on Sentiment-Enhanced Graph Convolutional Network

Fig. 3 illustrates the generation process of TAI based on sentiment-enhanced GCN. For the text in a sample, it may contain one or more aspect terms, but different aspect terms involve different valid context information. For example, given an input text “Hosted the @MLBPDP event today with Mother Nature on our side! Dayton Moore was in the house. #TBones #FunWellDone”, the valid context information of the aspect term “Dayton Moore” is “was in the house” rather than “on our side” or other context information before it, and only coarse-grained learning of text may introduce unrelated context information for aspect terms. In view of this problem, we adopt GCN to learn the syntactic structure features of text to filter unrelated context information that may interfere with aspect term extraction and sentiment polarity prediction. Compared with other existing GNN methods, GCN has higher efficiency in processing graph data by extracting the spatial features of graph structure through convolution to learn the complex relationships between nodes.



**Figure 3:** The generation process of TAI based on sentiment-enhanced GCN

Considering that aspect terms are typically noun phrases in text, we utilize the NLP tool Spacy (<https://spacy.io>, accessed on 20 August 2024) to extract noun phrases from text as the candidates for aspect terms, next continue to utilize Spacy to construct the syntactic dependency tree of text and obtain the corresponding adjacency matrix  $D \in \mathbb{R}^{m \times m}$  based on the dependency relationship between the words of each node:

$$D_{ij} = \begin{cases} 1, & w_i \text{ and } w_j \text{ are adjacent nodes} \\ 0, & w_i \text{ and } w_j \text{ are non-adjacent nodes} \end{cases} \quad (7)$$

where  $w_i$  and  $w_j$  are the  $i$ th and  $j$ th words, then introduce a sentiment dictionary named SenticNet to generate the sentiment score  $S$  between each adjacent node that enhances the adjacency matrix representation:

$$S_{i,j} = \text{SenticNet}(w_i) + \text{SenticNet}(w_j) \quad (8)$$

where  $\text{SenticNet}(w_i) \in [-1, 1]$  is the sentiment score of the word  $w_i$  in SenticNet,  $-1$  means the sentiment is negative,  $1$  is positive, and  $\text{SenticNet}(w_i) = 0$  indicates that the sentiment polarity of  $w_i$  is neutral or the word does not exist in SenticNet. Furthermore, we expect the noun phrases that may be aspect terms to receive more attention in the adjacency matrix, so the enhancement matrix  $T$  continues to be constructed for each noun phrase:

$$T_{i,j}^k = \begin{cases} 1, & w_i \text{ or } w_j \text{ is noun phrase} \\ 0, & w_i \text{ or } w_j \text{ is non - noun phrase} \end{cases} \quad (9)$$

where  $k \in n$  is the  $k$ th noun phrase,  $n$  is the count of noun phrases in text, then construct the sentiment-enhanced text adjacency matrix  $A$ :

$$A_{i,j} = \frac{1}{n} \sum_{k=1}^n [D_{i,j} \times (S_{i,j} + T_{i,j}^k + 1)] \quad (10)$$

After obtaining this adjacency matrix, we introduce GCN to learn the syntactic structure features and sentiment dependencies of the above noun phrases based on syntactic dependency tree:

$$h_i^l = \text{ReLU} \left( \frac{\sum_{j=1}^m A_{i,j} W_l h_j^{l-1}}{d_i + 1} + b_l \right) \quad (11)$$

where  $i$  is the current node,  $j$  is the adjacent node of  $i$ ,  $h_j^{l-1} \in \mathbb{R}^d$  is the representation of  $j$  in layer  $l$  that is generated by the previous GCN layer,  $h_j^0 \in \mathbb{R}^d$  is the initial node representation in GCN and also the representation of the  $j$ th token in  $H_s$ ,  $d_i = \sum_{j=1}^m A_{i,j}$  is the degree of  $i$ ,  $W_l$  and  $b_l$  are the learnable linear transformation parameters that map the current node features to adjacent nodes. Finally, we treat the output of the last GCN layer as sentiment-enhanced text representation and as TAI:

$$H_G = \{h_1^{\text{last}}, h_2^{\text{last}}, \dots, h_m^{\text{last}}\} \quad (12)$$

where  $H_G \in \mathbb{R}^{d \times m}$ .

### 3.4.2 Image Auxiliary Information Based on Adjective-Noun Pairs

To enhance the sentiment information expression in visual features, we adopt ANPs extracted from image to enhance the image semantic representation from another level. Unlike the image representation described above, ANPs can extract nouns such as people or objects appearing in image and adjectives modifying these nouns, which enable image semantics to be understood from the text level. Specifically, we employ an existing visual concept detector library DeepSentiBank [55] that can detect 2089 ANPs and their corresponding confidence scores for each image, then choose  $K$  ANPs with high confidence scores (Top- $K$  ANPs) for concatenation and feed them into RoBERTa to obtain

the ANPs representation:

$$H_A = RoBERTa(ANPs_1; ANPs_2; \dots; ANPs_K) \quad (13)$$

However, ANPs may contain the image region content unrelated to text semantics, and directly utilizing these ANPs to assist image representation would introduce extra noise to a large extent. Given this problem, we employ MCATT to perform an interaction between  $H_S$  and  $H_A$  to generate the text-aware ANPs representation  $H_{S \rightarrow A}$  as IAI, thus filtering the image content unrelated to text semantics as possible:

$$Z_{S \rightarrow A} = LN(H_S + MCATT(H_S, H_A)) \quad (14)$$

$$H_{S \rightarrow A} = LN(Z_{S \rightarrow A} + FFN(Z_{S \rightarrow A})) \quad (15)$$

where  $H_{S \rightarrow A} \in \mathbb{R}^{d \times m}$ .

### 3.5 Inter-Modal Feature Fusion and Output

In this component, we employ the image gating mechanism constructed in [Section 3.3](#) to perform feature fusion on the text representation  $H_S$ , the text-aware image representation  $H_{S \rightarrow V}$ , the sentiment-enhanced text representation  $H_G$  and the text-aware ANPs representation  $H_{S \rightarrow A}$  through an effective strategy to generate the inter-modal fusion representation of text and image features.

Since image is introduced as an additional modality to assist with text semantic expression in this study, we multiply the generated value of gating mechanism with the corresponding elements of image associated representations, thus dynamically controlling the input image information with the text word-level intensity, and filtering the image information unrelated to text semantics to prevent extra interference with subsequent work. Furthermore,  $H_G$  and  $H_{S \rightarrow A}$  as auxiliary text and image enhancement information are not in the same magnitude as  $H_S$  and  $H_{S \rightarrow V}$ , so we set weights named  $\alpha$  and  $\beta$  for  $H_G$  and  $H_{S \rightarrow A}$ , respectively, to control the TAI and IAI contribution to the inter-modal fusion representation. Finally, we fuse all corresponding text and image representations through the gating mechanism  $g$  to obtain the inter-modal fusion representation of text and image features:

$$H = (1 - g)(H_S; \alpha H_G) + g(H_{S \rightarrow V}; \beta H_{S \rightarrow A}) \quad (16)$$

Conditional Random Field (CRF) [\[56\]](#) is a discriminative undirected graph model that can effectively model the constraint relationships between sequence labels, so we adopt CRF to accomplish the aspect term extraction and sentiment polarity prediction tasks in our study. Specifically, we feed the above inter-modal fusion representation  $H$  into CRF to achieve text label sequence prediction:

$$P(y) = \frac{\exp(\text{score}(H, y))}{\sum_{y' \in Y} \exp(\text{score}(H, y'))} \quad (17)$$

$$\text{score}(H, y) = \sum_{i=0}^m T_{y_i, y_{i+1}} + \sum_{i=1}^m E_{h_i, y_i} \quad (18)$$

where  $T_{y_i, y_{i+1}}$  is the transfer score of the  $y_i$  label and the  $y_{i+1}$  label (i.e., the probability that  $y_i$  and  $y_{i+1}$  appear together), and  $E_{h_i, y_i}$  is the emission score of  $y_i$  (i.e., the probability that the output label is  $y_i$  when the input is  $h_i$ ).

For optimizing all model parameters, we employ the cross-entropy loss constructed between the predicted text label sequence and the real text label sequence as the training loss function on JMASA:

$$\mathcal{L} = -\frac{1}{|D|} \sum_{j=1}^{|D|} \log(P(y^j|H^j)) \quad (19)$$

## 4 Experiment

In this chapter, we perform a series of experiments with two Twitter benchmark datasets to demonstrate the effectiveness of our Text-Image Feature Fine-Grained Learning (TIFFL) model, then compare its performance with some representative methods in recent years.

### 4.1 Experimental Setup

**Datasets:** Considering the increasing application of social software such as Twitter and Facebook in people’s daily life, analyzing social media data has become a major research trend in academics. Twitter-15 and Twitter-17 are two benchmark datasets for JMASA built by Yu et al. [37], which are sampled from tweets containing text and image posted on the Twitter social media platform in 2014–2015 and 2016–2017 with the *BIO2* tagging schema described in Task Definition for labeling. There are 3502 and 2910 total texts as well as 8288 and 4819 total images for Twitter-15 and Twitter-17, respectively. The detailed statistics for the Twitter datasets are shown in Table 2 (where Pos, Neu and Neg are the counts of aspect terms as positive, neutral and negative, Total aspects is the count of aspect terms, and Sentence is the count of texts).

**Table 2:** The basic statistics for two Twitter benchmark datasets

	Twitter-15			Twitter-17		
	Train	Dev	Test	Train	Dev	Test
Pos	928	303	317	1508	515	493
Neu	1883	670	607	1638	517	573
Neg	368	149	113	416	144	168
Total aspects	3179	1122	1037	3562	1176	1234
Sentence	2101	727	674	1746	577	587

**Implementation Details:** For TIFFL, pretrained RoBERTa-base [50] and ResNet-152 [51] are employed as text and image encoders. In the process of parameter optimization, we adopt the AdamW learner with a weight attenuation of 0.01. Specifically, we set the batch size to 32 during training phase as well as 16 during development and testing phases, the training epoch to 25, the weight values  $\alpha$  and  $\beta$  to 0.6 and 0.5 on Twitter-15 as well as 0.7 and 0.4 on Twitter-17, the  $K$  value to 5, the number of GCN layers to 2, and the learning rate to  $3e-5$ . The final experimental results are chosen as the average scores of three independent trainings for all models. Our experiments are implemented based on PyTorch and run on an NVIDIA Tesla V100 GPU.

#### 4.2 Compared Baselines

Considering that JMASA consists of two subtasks, MATE and MASC, we compare TIFFL with various existing methods on the JMASA, MATE and MASC tasks to achieve the performance evaluation of our model. Tables 3–5 show the unimodal and multimodal compared baselines selected for the three tasks.

**Table 3:** Compared baselines for the JMASA task

Method	Description
SPAN [57]	SPAN is a method for extracting aspect terms and predicting sentiment polarities through LSTM-based multi-span decoding algorithm for the text modality only.
D-GCN [58]	D-GCN is a method for extracting aspect terms and predicting sentiment polarities through dependencies between words for the text modality only.
RoBERTa [50]	RoBERTa is a pretrained language model for BERT enhancement by employing better training strategies and larger corpus for the text modality only.
UMT [43] + TomBERT [38]	UMT is a MNER model based on span detection, and TomBERT is a MABSA model based on the BERT architecture, UMT + TomBERT combines the two models to accomplish the JMASA task for the text and image modalities.
OSCGA [44] + TomBERT	OSCGA is a MNER model based on entity alignment, OSCGA + TomBERT combines OSCGA with TomBERT to accomplish the JMASA task for the text and image modalities.
UMT-collapse	UMT-collapse applies UMT to the JMASA task for the text and image modalities.
OSCGA-collapse	OSCGA-collapse applies OSCGA to the JMASA task for the text and image modalities.
UMT-RoBERTa	UMT-RoBERTa replaces BERT in UMT-collapse with RoBERTa for the text and image modalities.
JML [4]	JML is a JMASA model based on auxiliary cross-modal relation detection for the text and image modalities.
VLP-MABSA [5]	VLP-MABSA is a JMASA model based on a unified multimodal encoder-decoder architecture for the text and image modalities.
CMMT [6]	CMMT is a JMASA model based on text-guided cross-modal interaction for the text and image modalities.
SAAF [7]	SAAF is a JMASA model based on a text-image selective fusion mechanism for the text and image modalities.

**Table 4:** Compared baselines for the MATE task

Method	Description
RAN [42]	RAN is a MNER model based on region-aware alignment for the text and image modalities.
UMT	UMT is a MNER model based on span detection for the text and image modalities.
OSCGA	OSCGA is a MNER model based on entity alignment for the text and image modalities.
MNER-QG [45]	MNER-QG is a MNER model based on an end-to-end machine reading comprehension framework for the text and image modalities.

**Table 5:** Compared baselines for the MASC task

Method	Description
MIMN [36]	MIMN is a MABSA model based on multi-interactive Bi-LSTM for the text and image modalities.
ESAFN [37]	ESAFN is a MABSA model based on an entity-aware attention fusion network for the text and image modalities.
TomBERT	TomBERT is a MABSA model based on the BERT architecture for the text and image modalities.
CapBERT [40]	CapBERT is a MABSA model for converting image semantics into caption and encoding it in combination with input text for the text and image modalities.
KEF-TomBERT [41]	KEF-TomBERT is a MABSA model for applying a proposed knowledge enhancement framework KEF to TomBERT for the text and image modalities.
TomRoBERTa	TomRoBERTa replaces BERT in TomBERT with RoBERTa for the text and image modalities.
CapRoBERTa	CapRoBERTa replaces BERT in CapBERT with RoBERTa for the text and image modalities.
KEF-TomRoBERTa	KEF-TomRoBERTa replaces BERT in KEF-TomBERT with RoBERTa for the text and image modalities.

### 4.3 Experimental Results and Analysis

In this section, we perform experiments with TIFFL and corresponding compared baselines on the JMASA, MATE and MASC tasks, then analyze the generated results.

#### 4.3.1 Experiments for the JMASA Task

Table 6 shows the experimental results of TIFFL and compared baselines for JMASA on the Twitter-15 and Twitter-17 datasets, we choose Precision (P), Recall (R) and Macro-F1 (F1) as



evaluation metrics and mark the best score for each metric in bold. Moreover, the results with \* are produced with our implementation. Compared with the better performing methods CMMT and SAAF, TIFFL achieves competitive results on the Twitter datasets through multimodal feature correlation discrimination and intra-modal feature fine-grained analysis, essentially maintaining comparable model performance on Twitter-15 while improving precision, recall and Macro-F1 by about 0.9%, 0.6% and 0.8% over CMMT on Twitter-17, and about 0.3%, 1.0% and 0.7% over SAAF, respectively.

**Table 6:** Experimental results of TIFFL and compared baselines for the JMASA task

Modality	Method	Twitter-15			Twitter-17		
		P	R	F1	P	R	F1
Text	SPAN	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN	58.3	58.8	59.4	64.1	64.2	64.1
	RoBERTa	61.8	65.3	63.5	65.5	66.9	66.2
Text + Image	UMT + TomBERT	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA + TomBERT	61.7	63.4	62.5	63.4	64.0	63.7
	UMT-collapse	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapse	63.1	63.7	63.2	63.5	63.5	63.5
	UMT-RoBERTa	61.6	66.4	63.9	65.3	68.2	66.7
	JML	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA*	64.8	68.3	66.3	66.4	69.0	67.9
	CMMT	64.6	<b>68.7</b>	<b>66.5</b>	67.6	69.4	68.5
	SAAF	<b>65.6</b>	67.3	66.4	68.2	69.0	68.6
	TIFFL (Ours)*	65.0	68.3	<b>66.5</b>	<b>68.5</b>	<b>70.0</b>	<b>69.3</b>

According to the experimental results, we can conclude as follows: (1) Since RoBERTa is pre-trained based on BERT by optimizing training strategies and adopting larger corpus, its performance is much better than SPAN and D-GCN; (2) UMT + TomBERT and OSCGA + TomBERT are both pipeline methods that combine MATE and MASC, their performance is worse than UMT-collapse and OSCGA-collapse, which may be attributed to the propagation of error information between the two tasks; (3) UMT-RoBERTa performs better than UMT-collapse, which also proves that RoBERTa is more powerful than BERT; (4) JML, VLP-MABSA, CMMT and SAAF perform better than other compared baselines, which indicates that methods specifically designed for JMASA can better accomplish this target task; (5) Although JML employs auxiliary cross-modal relation detection to control the rational utilization of visual information, its lack of in-depth research on unimodal features leads to inferior model performance compared to TIFFL; (6) VLP-MABSA simplifies model complexity with a unified multimodal encoder-decoder architecture, but its direct integration of textual and image features may introduce extra interference information, which is a major factor that its model performance is not comparable to TIFFL.

However, TIFFL does not perform as well on Twitter-15 as on Twitter-17 compared to CMMT and SAAF, we speculate the possible reasons are as follows:

CMMT adopts all 2089 ANPs extracted from image as visual auxiliary supervision, which can guarantee the accurate recognition of image semantics from the quantitative level. Since our chosen Top- $K$  ANPs that might contain some error information such as misrecognized words, focusing only on these ANPs may introduce extra noise to some extent. While CMMT does not perform a correlation discrimination between text and image features, and Twitter-15 might have more high text-image correlation samples than Twitter-17, thus the model performance of TIFFL on Twitter-15 is less favorable. This argument is validated to be reliable in the ablation study in [Section 4.4](#) and the parameter setting component in [Section 4.5](#). Moreover, CMMT introduces another label sequence as text auxiliary supervision, thus achieving performance enhancement by constructing auxiliary supervision modules for both text and image modalities, but excessive space resource occupation and manual labeling cost are also major problems. TIFFL effectively avoids the problems by choosing a more cost-effective method, which is an additional advantage of our model.

SAAF adopts Beta distribution to adjust the scalar weight of balancing text and image features, which enhances the resilience of image representation incorporating text features to bridge the inter-modal semantic gap, thus resulting in superior model performance. While in the process of gate vector computation, SAAF only employs text-aware image representation and ignores that there might be some information unrelated to image semantics in text, which may also introduce extra noise to the gate vector construction and cause certain defects in image filtering. Similarly, the ablation study in [Section 4.4](#) validates that the proportion of samples with low text-image correlation in Twitter-17 is higher than Twitter-15. Therefore, the gate vector of SAAF might not play a significant role on Twitter-15, and TIFFL employs both image-aware text and text-aware image representations to construct a more effective gating mechanism that works better on Twitter-17. Meanwhile, we can learn that TIFFL achieves better results for the MATE task on Twitter-17 from [4.3.2](#), which is a key factor in the performance enhancement on this dataset.

#### 4.3.2 Experiments for the MATE Task

[Table 7](#) shows the experimental results of TIFFL and compared baselines for the MATE task on Twitter-15 and Twitter-17, we also choose P, R and F1 as evaluation metrics. Compared with other methods in the table, TIFFL achieves optimal experimental results on the Twitter datasets, improving Macro-F1 by about 0.9% and 1.9% over CMMT on Twitter-15 and Twitter-17, and about 1.1% and 2.0% over SAAF, respectively, which also validate that our proposed multimodal feature correlation discrimination and intra-modal feature fine-grained analysis methods can efficiently accomplish the aspect term extraction to further enhance the overall model performance on JMASA.

**Table 7:** Experimental results of TIFFL and compared baselines for the MATE task

Method	Twitter-15			Twitter-17		
	P	R	F1	P	R	F1
RoBERTa	84.0	87.1	85.5	92.1	93.4	92.7
RAN	80.5	81.5	81.0	90.7	90.0	90.3
UMT	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA	81.7	82.1	81.9	90.2	90.7	90.4
MNER-QG	82.7	81.2	81.7	88.3	86.8	87.3

(Continued)

**Table 7 (continued)**

Method	Twitter-15			Twitter-17		
	P	R	F1	P	R	F1
JML	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA*	83.1	88.2	85.5	90.2	92.5	91.3
CMMT	83.9	88.1	85.9	92.2	93.9	93.1
SAAF*	83.9	88.0	85.7	92.5	93.4	93.0
TIFFL (Ours)*	<b>84.7</b>	<b>89.0</b>	<b>86.8</b>	<b>94.5</b>	<b>95.5</b>	<b>95.0</b>

### 4.3.3 Experiments for the MASC Task

Table 8 shows the experimental results of TIFFL and compared baselines for the MASC task on Twitter-15 and Twitter-17, we choose Accuracy (Acc) and F1 as evaluation metrics. The experimental results of TIFFL on the Twitter datasets are not particularly favorable compared to other methods in the table, we speculate the possible reasons are the same as the inferences described in Section 4.3.1 that our Top-K ANPs may introduce extra noise due to error information, which is demonstrated to have a significant impact on sentiment prediction in this subsection. KEF as an enhancement framework for MASC also employs ANPs to enhance the image semantic representation, but it filters unrelated ANPs interference information by calculating the similarity between a specific aspect term and the noun in ANP that achieves better Accuracy on Twitter-15. However, TIFFL performs well on the MATE task, which compensates for its deficiencies on MASC, thus also enhancing the overall model performance on JMASA.

**Table 8:** Experimental results of TIFFL and compared baselines for the MASC task

Method	Twitter-15		Twitter-17	
	Acc	F1	Acc	F1
RoBERTa	76.3	71.4	69.8	68.0
MIMN	71.8	65.7	65.9	63.0
ESAFN	73.4	67.4	67.8	64.2
TomBERT	77.2	71.8	70.3	68.0
CapBERT	78.0	73.3	69.8	68.4
KEF-TomBERT	78.7	73.8	72.1	70.0
TomRoBERTa	77.6	73.2	71.3	70.1
CapRoBERTa	77.8	73.4	71.1	68.6
KEF-TomRoBERTa*	<b>78.8</b>	74.0	72.2	70.2
JML	78.7	–	72.7	–
VLP-MABSA*	78.5	73.8	73.2	71.4
CMMT	77.9	–	<b>73.8</b>	–
SAAF*	78.6	73.7	73.1	<b>71.6</b>
TIFFL (Ours)*	78.4	<b>74.5</b>	73.0	<b>71.6</b>

#### 4.4 Ablation Study

To further investigate the effectiveness of our proposed methods, we perform ablation analysis for three important units in TIFFL on Twitter-15 and Twitter-17: (1) Image Gating Mechanism (IGM). (2) Text Auxiliary Information (TAI). (3) Image Auxiliary Information (IAI). We first remove each unit individually and then remove all three units simultaneously to comprehensively demonstrate the contribution of each model unit. The ablation study results are shown in [Table 9](#).

**Table 9:** Ablation study of TIFFL

Method	Twitter-15			Twitter-17		
	P	R	F1	P	R	F1
TIFFL	<b>65.0</b>	<b>68.3</b>	<b>66.5</b>	<b>68.5</b>	<b>70.0</b>	<b>69.3</b>
TIFFL w/o IGM	62.9	66.5	64.7	66.4	67.6	67.0
TIFFL w/o TAI	62.4	65.9	64.3	66.8	67.6	67.2
TIFFL w/o IAI	63.2	66.9	65.0	67.2	68.5	68.0
TIFFL w/o IGM & TAI & IAI	61.9	65.2	63.5	65.1	67.5	66.3

First, we can inform that removing IGM decreases Macro-F1 by about 1.8% and 2.3% on the Twitter datasets, which indicates that the inter-modal fusion strategy in our model can effectively filter the image information unrelated to text semantics; Next, removing TAI decreases Macro-F1 by about 2.2% and 2.1% on the Twitter datasets, which indicates that our model introduces sentiment-enhanced GCN can deeply learn the syntactic structure features of text; Then, removing IAI decreases Macro-F1 by about 1.5% and 1.3% on the Twitter datasets, which indicates that our adopted ANPs can better assist the image semantic expression from the text level; Finally, removing all above units decreases Macro-F1 by about 3.0% on the Twitter datasets, which also indicates that our methods can contribute to the model performance on JMASA.

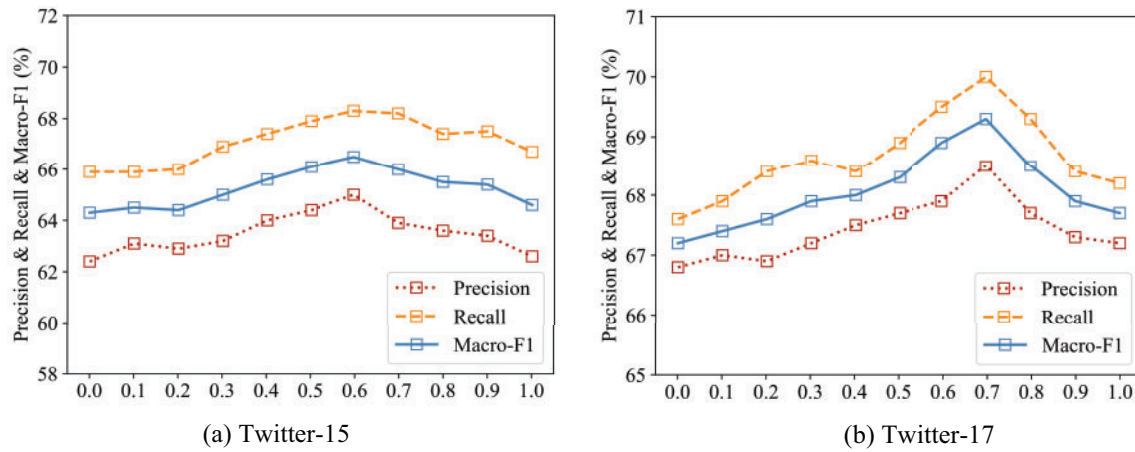
However, removing IAI shows a certain reduction in the decrease of Macro-F1 compared to IGM and TAI, and removing IGM decreases Macro-F1 less on Twitter-15 than Twitter-17, which also validates the argument in [4.3.1](#) that Top- $K$  ANPs might be interfered by error information and limit model performance, as well as Twitter-17 contains more samples with low text-image correlation than Twitter-15.

#### 4.5 Parameter Analysis

In this section, we detail the evaluation process of optimal hyper-parameters. The above experiments are all performed in TIFFL after hyper-parameter tuning.

##### 4.5.1 $\alpha$ Value

For testing the effect of the TAI weight  $\alpha$  on model performance during inter-modal feature fusion, we set  $\alpha$  as a decimal number with an interval of 0.1 in  $[0, 1]$ . [Fig. 4a,b](#) shows the model performance for  $\alpha$  values on Twitter-15 and Twitter-17, respectively.

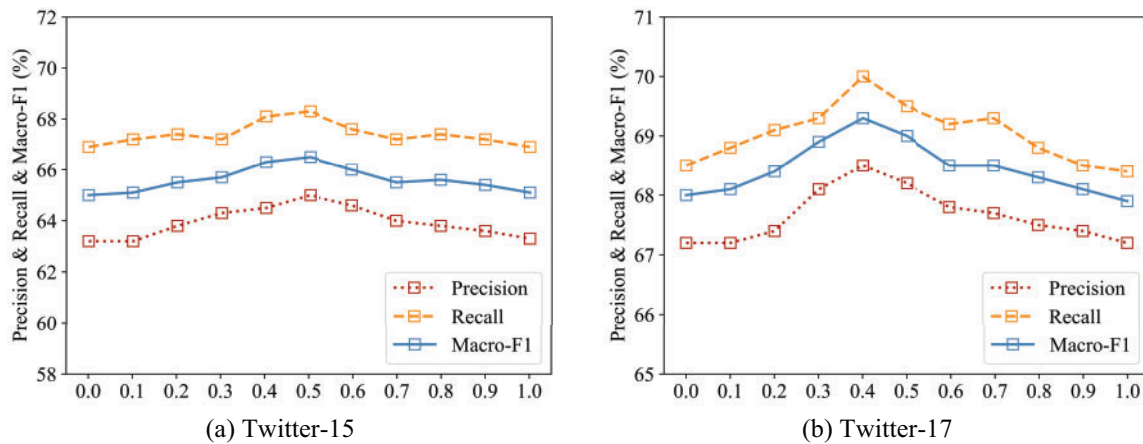


**Figure 4:** Effect of  $\alpha$  value on model performance

According to the test results, we can inform that our model performs worse without introducing TAI, which indicates that sentiment-enhanced GCN contributes to enhance the model performance to some extent. As  $\alpha$  increases, the performance shows an upward trend in fluctuation, with the best results when  $\alpha$  is 0.6 and 0.7 on Twitter-15 and Twitter-17, respectively. However, the performance starts to decrease as  $\alpha$  further increases, the possible reason is speculated as follows: In the process of constructing TAI, the extracted noun phrases can only be treated as undetermined aspect terms for reference because image information is not combined. Therefore, the model assigns a larger proportion to TAI as  $\alpha$  increases, so that the noun phrases of non-aspect terms might also receive too much attention and introduce extra noise.

#### 4.5.2 $\beta$ Value

For analyzing the value of the IAI weight  $\beta$  during inter-modal feature fusion, we also set  $\beta$  as a decimal number with an interval of 0.1 in  $[0, 1]$ . Fig. 5a,b shows the model performance for  $\beta$  values on Twitter-15 and Twitter-17, respectively.

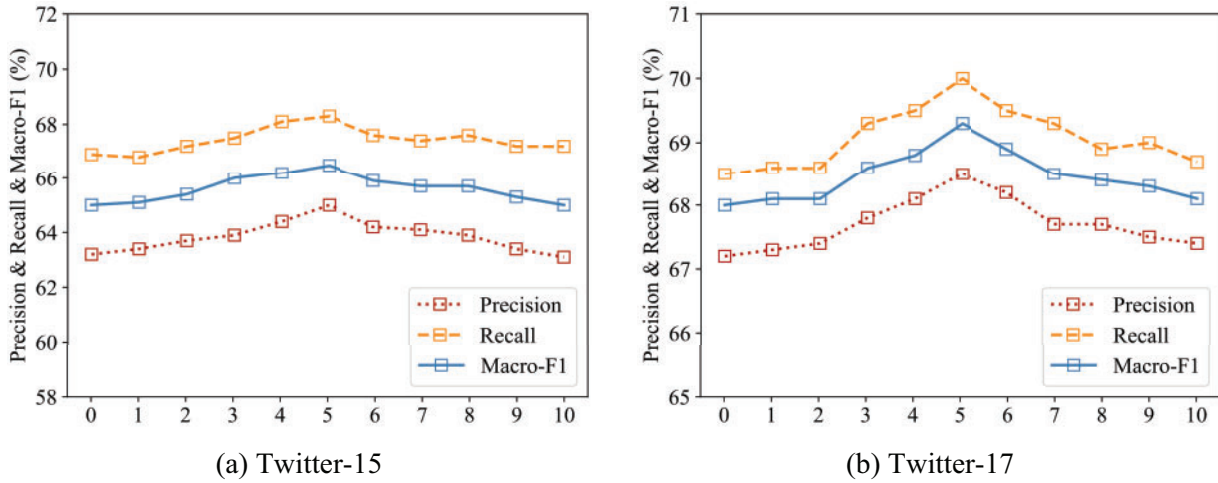


**Figure 5:** Effect of  $\beta$  value on model performance

According to the test results, we can inform as follows: Our model performance shows an upward trend in fluctuation as  $\beta$  increases, with the best results when  $\beta$  is 0.5 and 0.4 on Twitter-15 and Twitter-17, respectively. However, the performance starts to decrease as  $\beta$  further increases, the possible reason is speculated that IAI might contain some ANPs error information unrelated to image semantics. When  $\beta$  exceeds a certain range, the model pays more attention to IAI and weakens the dominance of original image features, so that the error information is also amplified resulting in a continuous degradation of the performance.

#### 4.5.3 $K$ Value

For exploring the optimal value of the ANPs number  $K$  in IAI, we set  $K$  to each integer in the range of  $[0, 1]$ . Fig. 6a,b shows the model performance for  $K$  values on Twitter-15 and Twitter-17, respectively.



**Figure 6:** Effect of  $K$  value on model performance

According to the test results, we can inform that our model performs worse when  $K$  is equal to 0, which indicates that adopting ANPs as IAI can further enhance the model performance. As  $K$  increases, the performance shows an upward trend in fluctuation, with the best results when  $K$  is equal to 5 on the Twitter datasets. However, the performance starts to decrease when  $K$  exceeds 5, the possible reason is speculated as follows: Each text in the Twitter datasets involves up to five aspect terms, unlike CMMT that adopts all ANPs for the vision representation supervision, we treat ANPs as the candidates for aspect terms in image. When  $K$  is greater than the count of aspect terms, IAI might generate extra noise to the model by introducing too much misrecognized word information.

Moreover, by comparing the test results of the above three parts, we can learn that TAI tuning outperforms IAI on model performance, which also validates the argument in Section 4.3.1.

## 4.6 Case Study

In this section, we choose four representative samples to compare TIFFL with RoBERTa, CMMT and SAAF on the MATE and JMASA tasks to better demonstrate the superiority of our model. Meanwhile, we sequentially remove the three important units of TIFFL described in Section 4.4 to



prove the effectiveness of our designed methods through these samples. The case information and prediction results are shown in [Tables 10](#) and [11](#).

**Table 10:** Case study for the MATE task



Image		
Text	(a) Joe Maddon talks to # Cubs pitcher Jason Motte (30) as Jon Lester stretches on first day of spring training in Mesa	(b) RT @ WizardGirlsNBA: Excited to have Cheerleaders Gdynia here the way from Poland for # PolishHeritageNight # WizKings
Noun phrase	Joe Maddon, # Cubs, pitcher, Jason Motte, Jon Lester, day, spring training, Mesa	WizardGirlsNBA, Cheerleaders Gdynia, way, Poland
Top- <i>K</i> ANPs	holy cross, outdoor sports, mad dog, holy angels, classic race	holy child, excited crowd, proud student, talented kids, successful team
Label	(Joe Maddon, Neutral) (Jason Motte, Neutral) (Jon Lester, Neutral) (Mesa, Neutral)	(Cheerleaders Gdynia, Positive) (Poland, Neutral)
RoBERTa	(Joe Maddon, Neutral) ✓ (# Cubs, Neutral) ✗ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓ (Poland, Neutral) ✗
CMMT	(Joe Maddon, Neutral) ✓ (# Cubs, Neutral) ✗ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓ (Poland, Neutral) ✗
SAAF	(Joe Maddon, Neutral) ✓ (# Cubs, Neutral) ✗ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓ (Poland, Neutral) ✗
TIFFL (Ours)	(Joe Maddon, Neutral) ✓ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓

(Continued)

**Table 10 (continued)**

TIFFL w/o IGM	(Joe Maddon, Neutral) ✓ (# Cubs, Neutral) ✗ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓
TIFFL w/o TAI	(Joe Maddon, Neutral) ✓ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓
TIFFL w/o IAI	(Joe Maddon, Neutral) ✓ (Jason Motte, Neutral) ✓ (Jon Lester, Neutral) ✓ (Mesa, Neutral) ✓	(Cheerleaders Gdynia, Positive) ✓ (Poland, Neutral) ✗

**Table 11:** Case study for the JMASA task

Image		
Text	(a) Jean Marmoreo - ready to run ! # stwm	(b) I've just witnessed Wes Morgan lift a premier league title. Football really is bonkers
Noun phrase	Jean marmoreo, # stwm	Wes Morgan, premier league, Football, bonkers
Top- <i>K</i> ANPs	young driver, young fan, happy christmas, bad girls, fat pig	sexy halloween, ill child, excited crowd, fancy dress, great party
Label	(Jean Marmoreo, Positive)	(Wes Morgan, Positive)
RoBERTa	(Jean Marmoreo, Neutral) ✗	(premier league, Neutral) (Wes Morgan, Negative) ✗ (premier league, Neutral) ✓
CMMT	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Negative) ✗ (premier league, Neutral) ✓
SAAF	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Negative) ✗ (premier league, Neutral) ✓
TIFFL (Ours)	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Positive) ✓ (premier league, Neutral) ✓
TIFFL w/o IGM	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Positive) ✓ (premier league, Neutral) ✓

(Continued)

**Table 11 (continued)**

TIFFL w/o TAI	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Negative) × (premier league, Neutral) ✓
TIFFL w/o IAI	(Jean Marmoreo, Positive) ✓	(Wes Morgan, Positive) ✓ (premier league, Neutral) ✓

First, [Table 10](#) shows two samples where TIFFFL is dominant on the MATE task. In the sample of [Table 10\(a\)](#), RoBERTa, CMMT and SAAF extract one more incorrect aspect term “# Cubs” on top of the correct aspect term extraction, while TIFFFL achieves the accurate aspect term extraction and sentiment polarity prediction, we speculate the possible reasons are as follows: The inter-modal fusion strategy in TIFFFL can effectively combine text and image features to accomplish MATE. Although “# Cubs” is a description of the aspect term “Jason Motte”, it is not reflected in the image. However, CMMT dynamically controls the intervention of image information by predicting word confidence, which treats text information as the dominant role on the target task and underestimates the effect of image information, and SAAF does not consider the image-unrelated information from text during the construction of gate vector. Therefore, both CMMT that preferentially extracts the aspect terms in text and RoBERTa that only extracts the aspect terms in text make incorrect predictions, while SAAF also makes incorrect predictions by introducing extra noise due to the imperfect gate vector construction. Furthermore, the aspect term “# Cubs” is also extracted when TIFFFL removes IGM, we hypothesize the possible reason is that the image information unrelated to text semantics cannot be efficiently filtered without utilizing the gating mechanism, which introduces extra noise and degrades the model performance of aspect term extraction.

In the sample of [Table 10\(b\)](#), RoBERTa, CMMT and SAAF extract an incorrect aspect term “Poland” on top of the correct aspect term, while “Poland” is also extracted when TIFFFL removes IAI. The possible reasons are speculated as follows: TIFFFL constructs IAI to clearly understand that there is no reflection of “Poland” in the image from the nouns of ANPs, and SAAF does not conduct further analysis of image features. RoBERTa is only for text modal and cannot intervene with image information. Moreover, though CMMT adopts ANPs as visual auxiliary supervision as well, its utilization of all 2089 ANPs may contain unrelated interference information leading to the extraction of incorrect aspect term. Therefore, TIFFFL achieves accurate aspect term extraction and sentiment polarity prediction by effectively combining text and image features.

Then, [Table 11](#) shows two samples where TIFFFL has an advantage on the JMASA task. In the sample of [Table 11\(a\)](#), RoBERTa, CMMT, SAAF and TIFFFL all extract the correct aspect term, but only RoBERTa makes incorrect sentiment prediction for the aspect term “Jean Marmoreo”, the possible reasons are speculated as follows: RoBERTa can only analyze the text modality and cannot identify the facial expression feature in image, and SAAF can combine image features to make correct sentiment prediction. Although Top-*K* ANPs contain more misrecognized word information, CMMT and TIFFFL can also predict the correct sentiment by combining the smiling facial expressions in image.

In the sample of [Table 11\(b\)](#), the four models also extract the correct aspect terms, but RoBERTa, CMMT and SAAF give wrong sentiment predictions for the aspect term “Wes Morgan”, while the sentiment of “Wes Morgan” is also incorrectly predicted when TIFFFL removes TAI, the possible reasons are speculated as follows: The noun phrases extracted by TIFFFL contain the actual aspect terms “Wes Morgan” and “premier league”, which help our model to perform the MATE task better,

and TIFFL avoids the effect of the word “bonkers” on the previous sentence by introducing TAI to learn the syntactic structure features of text, so removing TAI is demonstrated to have an impact on model performance. However, RoBERTa, CMMT and SAAF cannot learn the intrinsic features of text from a deeper level during text feature encoding, so “bonkers” may interfere with the aspect term “Wes Morgan” to some extent resulting in its prediction as negative.

## 5 Conclusion

In this paper, we propose a text-image feature fine-grained learning model TIFFL for Joint Multimodal Aspect-based Sentiment Analysis (JMASA). For text feature learning, the model constructs an enhanced adjacency matrix of word dependencies and learns the syntactic structure features of text by employing Graph Convolutional Network (GCN), thus solving the context interference problem of identifying different aspect terms. For image feature learning, the model introduces image Adjective-Noun Pairs (ANPs) to represent visual feature semantics more intuitively, thus solving the problem of ambiguous image semantic extraction. Thereby, the model can further enhance the performance of aspect term extraction and sentiment polarity prediction. Experiments on two Twitter benchmark datasets demonstrate that TIFFL outperforms most advanced studies on Twitter-15 and all compared baselines on Twitter-17, thus validating the superiority of our adopted methods.

Since TIFFL performs less well on Twitter-15 than Twitter-17, we subsequently plan to implement a model optimization and conduct a specific investigation of the Twitter-15 dataset to figure out the cause of unsatisfactory model performance. Moreover, we determine the values of weight hyper-parameters through experimental tests, but the method relies too much on manual operation. Therefore, we plan to implement automatic learning of the above hyper-parameters to generate the proportions of text and image auxiliary information in the inter-modal fusion representation for the next research task, thus the model can assign more reasonable fusion weights according to the specific details of individual modalities in sample to achieve further performance enhancement.

**Acknowledgement:** This work was supported by the Science and Technology Project of Henan Province.

**Funding Statement:** This work was supported by the Science and Technology Project of Henan Province (No. 222102210081).

**Author Contributions:** Tianzhi Zhang wrote the main manuscript text, Gang Zhou and Shuang Zhang participated in the experiment, Shunhang Li analyzed the data, Yepeng Sun processed data, Qiankun Pi set up the experimental environment and Shuo Liu prepared tables and figures. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data that support the findings of this study are available from the corresponding author, Gang Zhou, upon reasonable request.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

- [1] W. Fan and Z. Shi, “Cross-modal consistency with aesthetic similarity for multimodal false information detection,” *Comput. Mater. Contin.*, vol. 79, no. 2, pp. 2723–2741, 2024. doi: [10.32604/cmc.2024.050344](https://doi.org/10.32604/cmc.2024.050344).
- [2] W. Liu, S. Cao, and S. Zhang, “Multimodal consistency-specificity fusion based on information bottleneck for sentiment analysis,” *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 2, 2024, Art. no. 101943. doi: [10.1016/j.jksuci.2024.101943](https://doi.org/10.1016/j.jksuci.2024.101943).
- [3] L. Deng, B. Liu, and Z. Li, “Multimodal sentiment analysis based on a cross-modal multihead attention mechanism,” *Comput. Mater. Contin.*, vol. 78, no. 1, pp. 1157–1170, 2024. doi: [10.32604/cmc.2023.042150](https://doi.org/10.32604/cmc.2023.042150).
- [4] X. Ju *et al.*, “Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection,” in *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.*, Punta Cana, Dominican Republic, 2021, pp. 4395–4405.
- [5] Y. Ling, J. Yu, and R. Xia, “Vision language pre-training for multimodal aspect-based sentiment analysis,” in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.*, Dublin, Ireland, 2022, pp. 2149–2159.
- [6] L. Yang, J. C. Na, and J. Yu, “Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis,” *Inf. Process. Manag.*, vol. 59, no. 5, 2022, Art. no. 103038. doi: [10.1016/j.ipm.2022.103038](https://doi.org/10.1016/j.ipm.2022.103038).
- [7] Z. Wang and J. Guo, “Self-adaptive attention fusion for multimodal aspect-based sentiment analysis,” *Math. Biosci. Eng.*, vol. 21, no. 1, pp. 1305–1320, 2024. doi: [10.3934/mbe.2024056](https://doi.org/10.3934/mbe.2024056).
- [8] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, *arXiv:1609.02907*.
- [9] D. Borth, R. Ji, T. Chen, T. Breuel, and S. F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proc. 21st ACM Int. Conf. Multimed.*, Barcelona, Spain, 2013, pp. 223–232.
- [10] Y. Chen, “Convolutional neural network for sentence classification,” M.S. thesis, Univ. of Waterloo, Waterloo, ON, Canada, 2015.
- [11] B. Shin, T. Lee, and J. D. Choi, “Lexicon integrated CNN models with attention for sentiment analysis,” 2016, *arXiv:1610.06272*.
- [12] Q. You, H. Jin, and J. Luo, “Visual sentiment analysis by attending on local image regions,” in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, vol. 31. doi: [10.1609/aaai.v31i1.10501](https://doi.org/10.1609/aaai.v31i1.10501).
- [13] C. Zhang, Q. Li, and D. Song, “Aspect-based sentiment classification with aspect-specific graph convolutional networks,” in *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 4567–4577.
- [14] B. Huang and K. Carley, “Syntax-aware aspect level sentiment classification with graph attention networks,” in *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 5469–5477.
- [15] K. Sun, R. Zhang, S. Mensah, Y. Mao, and X. Liu, “Aspect-level sentiment analysis via convolution over dependency tree,” in *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 5679–5688.
- [16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [17] H. Tang, D. Ji, C. Li, and Q. Zhou, “Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification,” in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 3229–3238.
- [18] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, “Relational graph attention network for aspect-based sentiment analysis,” in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 6578–6588.
- [19] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, “Tensor fusion network for multimodal sentiment analysis,” 2017, *arXiv:1707.07250*.
- [20] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, 2020. doi: [10.1109/TAFFC.2020.3038167](https://doi.org/10.1109/TAFFC.2020.3038167).
- [21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, *arXiv:1412.3555*.

- [22] D. Bertero, F. B. Siddique, C. S. Wu, Y. Wan, R. H. Y. Chan and P. Fung, “Real-time speech emotion and sentiment recognition for interactive dialogue systems,” in *Proc. 2016 Conf. Empir. Methods Nat. Lang. Process.*, Austin, TX, USA, 2016, pp. 1042–1047.
- [23] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000–6010
- [24] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process.*, Lisbon, Portugal, 2015, pp. 2539–2544.
- [25] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh and L. P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Proc. 55th Annu. Meet. Assoc. Comput. Linguist.*, Vancouver, BC, Canada, 2017, pp. 873–883.
- [26] P. P. Liang, Z. Liu, A. Zadeh, and L. P. Morency, “Multimodal language analysis with recurrent multistage fusion,” 2018, *arXiv:1808.03920*.
- [27] C. Busso *et al.*, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proc. 6th Int. Conf. Multimodal Interf.*, State College, PA, USA, 2004, pp. 205–211.
- [28] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1162–1171, 2011. doi: [10.1016/j.specom.2011.06.004](https://doi.org/10.1016/j.specom.2011.06.004).
- [29] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea and S. Poria, “Towards multimodal sarcasm detection (an \_obviously\_ perfect paper),” 2019, *arXiv:1906.01815*.
- [30] Y. Cai, H. Cai, and X. Wan, “Multi-modal sarcasm detection in twitter with hierarchical fusion model,” in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, 2019, pp. 2506–2515.
- [31] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y. Y. Chen and S. F. Chang, “Object-based visual sentiment concept analysis and application,” in *Proc. 22nd ACM Int. Conf. Multimedia*, New York, NY, USA, 2014, pp. 367–376.
- [32] J. Yang, D. She, M. Sun, M. M. Cheng, P. L. Rosin and L. Wang, “Visual sentiment prediction based on automatic discovery of affective regions,” *IEEE Trans. Multimed.*, vol. 20, no. 9, pp. 2513–2525, 2018. doi: [10.1109/TMM.2018.2803520](https://doi.org/10.1109/TMM.2018.2803520).
- [33] Q. You, L. Cao, H. Jin, and J. Luo, “Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks,” in *Proc. 24th ACM Int. Conf. Multimed.*, New York, NY, USA, 2016, pp. 1008–1017.
- [34] N. Xu, W. Mao, and G. Chen, “A co-memory network for multimodal sentiment analysis,” in *41st Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, New York, NY, USA, 2018, pp. 929–932.
- [35] A. Kumar and G. Garg, “Sentiment analysis of multimodal twitter data,” *Multimed. Tools Appl.*, vol. 78, no. 17, pp. 24103–24119, 2019. doi: [10.1007/s11042-019-7390-1](https://doi.org/10.1007/s11042-019-7390-1).
- [36] N. Xu, W. Mao, and G. Chen, “Multi-interactive memory network for aspect based multimodal sentiment analysis,” in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, vol. 33, pp. 371–378.
- [37] J. Yu, J. Jiang, and R. Xia, “Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 429–439, 2019. doi: [10.1109/TASLP.2019.2957872](https://doi.org/10.1109/TASLP.2019.2957872).
- [38] J. Yu and J. Jiang, “Adapting BERT for target-oriented multimodal sentiment classification,” in *Proc. Twenty-Eighth Int. Joint Conf. Artif. Intell.*, Macao, China, 2019, pp. 5408–5414.
- [39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.
- [40] Z. Khan and Y. Fu, “Exploiting BERT for multimodal target sentiment classification through input space translation,” in *Proc. 29th ACM Int. Conf. Multimed.*, New York, NY, USA, 2021, pp. 3034–3042.
- [41] F. Zhao, Z. Wu, S. Long, X. Dai, S. Huang and J. Chen, “Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification,” in *Proc. 29th Int. Conf. Comput. Linguist.*, Gyeongju, Republic of Korea, 2022, pp. 6784–6794.



- [42] H. Wu, S. Cheng, J. Wang, S. Li, and L. Chi, “Multimodal aspect extraction with region-aware alignment network,” in *Proc. Nat. Lang. Process. Chin. Comput.*, Zhengzhou, China, 2020, pp. 145–156.
- [43] J. Yu, J. Jiang, L. Yang, and R. Xia, “Improving multimodal named entity recognition via entity span detection with unified multimodal transformer,” in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, Seattle, WA, USA, 2020, pp. 3342–3352.
- [44] Z. Wu, C. Zheng, Y. Cai, J. Chen, H. F. Leung and Q. Li, “Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts,” in *Proc. 28th ACM Int. Conf. Multimed.*, New York, NY, USA, 2020, pp. 1038–1046.
- [45] M. Jia *et al.*, “MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding,” in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, 2023, vol. 37, no. 7, pp. 8032–8040. doi: [10.1609/aaai.v37i7.25971](https://doi.org/10.1609/aaai.v37i7.25971).
- [46] M. Zhang, Y. Zhang, and D. T. Vo, “Neural networks for open domain targeted sentiment,” in *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process.*, Lisbon, Portugal, 2015, pp. 612–621.
- [47] X. Li, L. Bing, P. Li, and W. Lam, “A unified model for opinion target extraction and target sentiment prediction,” in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, vol. 33, pp. 6714–6721.
- [48] Z. Chen and T. Qian, “Relation-aware collaborative learning for unified aspect-based sentiment analysis,” in *Proc. Conf. Assoc. Comput. Linguist.*, Seattle, WA, USA, 2020, pp. 3685–3694.
- [49] E. F. Sang and J. Veenstra, “Representing text chunks,” 1999, *arXiv:cs/9907006*.
- [50] Y. Liu *et al.*, “Roberta: A robustly optimized BERT pretraining approach,” 2019, *arXiv:1907.11692*.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
- [52] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.
- [53] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proc. Conf. Assoc. Comput. Linguist.*, Florence, Italy, 2019, pp. 6558–6569.
- [54] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.
- [55] T. Chen, D. Borth, T. Darrell, and S. F. Chang, “DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks,” 2014, *arXiv:1410.8586*.
- [56] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. ICML*, Williamstown, MA, USA, 2001, vol. 1, no. 2, p. 3.
- [57] M. Hu, Y. Peng, Z. Huang, D. Li, and Y. Lv, “Open-domain targeted sentiment analysis via span-based extraction and classification,” 2019, *arXiv:1906.03820*.
- [58] G. Chen, Y. Tian, and Y. Song, “Joint aspect extraction and sentiment analysis with directional graph convolutional networks,” in *Proc. 28th Int. Conf. Comput. Linguist.*, Barcelona, Spain, 2020, pp. 272–279.