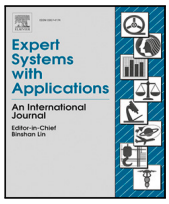
Expert Systems With Applications 264 (2025) 125818

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



TCMT: Target-oriented Cross Modal Transformer for Multimodal Aspect-Based Sentiment Analysis

Wang Zou [a], Xia Sun [a,*], Wenhuan Wu [b], Qiang Lu [a], Xiaodi Zhao [a], Qirong Bo [a,*], Jianqiang Yan [a,*]

[a] School of Information Science and Technology, Northwest University, Xi'an, 710127, China
[b] School of Electrical and Information Engineering, Hubei University of Automotive Technology, Shiyan, 442002, China

ARTICLE INFO

Keywords:
Multimodal aspect-based sentiment analysis
Cross-modal transformer
Object detection

ABSTRACT

Multimodal Aspect-Based Sentiment Analysis (MABSA) technology aims to utilize both textual and visual modalities to achieve Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect Sentiment Classification (MASC) in tweets. Current research has overlooked the impact of noise from irrelevant regions in images on model performance. Additionally, there has been insufficient utilization of the textual information contained within images and the syntactic features of sentences. In this paper, we propose a Target-oriented Cross Modal Transformer (TCMT) for MABSA. The model consists of a textual auxiliary module, a visual auxiliary module, and a main module: the textual aspect-sentiment extraction module, the visual aspect-sentiment prediction module, and the textual-visual alignment cross-modal module. In the textual auxiliary module, we utilize syntactic features to assist the model in identifying the boundaries of multi-word aspect terms and employ Optical Character Recognition (OCR) technology to capture textual information contained within images. In the visual auxiliary module, we employ Adjective-Noun Pairs (ANPs) detection for supervised training of images. Additionally, we have improved the cross-modal Transformer structure by designing a GCN-based Transformer in the textual auxiliary module to learn syntactic graphs, and a CNN-based Transformer in the visual auxiliary module to focus more on important information in images. In the cross-modal MABSA module, we design a target-oriented interaction component to facilitate modal interaction learning and mitigate the impact of image noise, along with an alignment auxiliary component to optimize modal alignment training. We conducted extensive experiments on two publicly available benchmark datasets. The results demonstrate that the performance of the TCMT model is significantly superior to that of the baseline model, achieving state-of-the-art results. Both the textual auxiliary module and the visual auxiliary module effectively assist the cross-modal MABSA module in completing the task more efficiently.

1. Introduction

With the rapid development of the Internet and information technology, an immense volume of information is generated daily on social platforms like Twitter and Facebook. This information contains valuable insights. For example, in the fields of production and consumption, mining user comments not only offers manufacturers valuable feedback to enhance their products but also provides consumers with guidance when selecting goods. Analyzing comments about politicians can reveal insights into public support, while mining comments on sports news can illuminate the level of audience affection for sports stars. Therefore, the effective management and utilization of online comment information have garnered significant attention from scholars.

Recently, researchers have proposed utilizing Aspect-Based Sentiment Analysis (ABSA) (Do et al., 2019) methods for mining comment information. This method effectively extracts aspect terms and sentiment polarity from sentences. However, online review information not only includes text; users also upload corresponding images to express their emotions. Traditional Multimodal Sentiment Analysis (MSA) (Gandhi et al., 2023) techniques can capture coarse-grained sentiment polarity but often struggle to extract specific aspect sentiments effectively. In recent years, some researchers have proposed Multimodal Aspect-based Sentiment Analysis (MABSA) (Zhou et al., 2021) techniques that build upon MSA and ABSA. This technique effectively utilizes images to enhance the extraction of fine-grained information

* Corresponding authors.
E-mail addresses: zouwang@stumail.nwu.edu.cn (W. Zou), raindy@nwu.edu.cn (X. Sun), wuwenhuan5@163.com (W. Wu), nwulq@stumail.nwu.edu.cn (Q. Lu), zhaoxiaodi@stumail.nwu.edu.cn (X. Zhao), boqirong@nwu.edu.cn (Q. Bo), yanjq@nwu.edu.cn (J. Yan).

https://doi.org/10.1016/j.eswa.2024.125818
Received 18 June 2024; Received in revised form 30 October 2024; Accepted 15 November 2024
Available online 23 November 2024

Table 1
Summary of previous main works on the MABSA task.

Framework	Methodology	Strengths	Weaknesses
Pipeline	JML (Ju et al., 2021)	Image-text relationship detection.	Error propagation of the two subtasks.
BART	VLP-MABSA (Ling et al., 2022)	BART encoder achieves representation of images and text, while the BART decoder implements multiple subtasks.	Lack of alignment between image and text modalities.
	AoM (Zhou et al., 2023)	Text-image correlation matrix effectively aligns the modal information.	Lack of independent supervised learning optimization for both images and text.
Cross-modal Transformers	CMMT (Yang et al., 2022)	Supervised learning of images and text, with text guiding cross-modal interaction.	Irrelevant regions in images will introduce noise to the text.
	Atlantis (Xiao et al., 2024)	Using image aesthetics to evaluate visual sentiment.	Ignoring the syntactic features of the text and the textual information from images.

Fig. 1. Examples of MABSA. Where aspect terms are highlighted in yellow and red, respectively.

from text. As illustrated in Fig. 1, MABSA aims to achieve Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-based Sentiment Classification (MASC).

The previous main work on the MABSA task can be summarized as shown in Table 1, which can be categorized into three types of methods: Pipeline, Bidirectional Autoregressive Transformer (BART) (Lewis et al., 2020), and Cross-modal Transformers. Ju et al. (2021) first proposed a Joint Multimodal Learning (JML) method with a relation detection module to evaluate the contribution of images to text. However, their method is based on a pipeline framework, which leads to error propagation between the two subtasks. The BART encoder–decoder framework effectively represents both modalities simultaneously and performs multiple subtasks in the decoder, thereby avoiding the issue of error propagation. Ling et al. (2022) proposed a Vision-Language BART (VLP-MABSA) to integrate image and text information, but they did not address the alignment issue between the two modalities. Zhou et al. (2023) proposed an Aspect-oriented Method (AoM) to detect semantic and sentiment information related to aspects in images. However, this method lacks independent supervised learning optimization for both images and text. To better leverage the semantic information from both text and images, researchers have designed a three-module structure based on cross-modal Transformers (Tsai et al., 2019): a text auxiliary module, a visual auxiliary module, and a cross-modal MABSA module. Yang et al. (2022) proposed a Cross-Modal Multi-Task Learning (CMMT) framework and designed a dynamic gating mechanism to adjust the contribution of images to text. However, they overlooked the noise introduced by irrelevant regions in the images. Xiao et al. (2024) proposed the Atlantis model to explore image sentiment from an aesthetic perspective, but they overlooked the syntactic features of the text and the textual information from images.

However, their research overlooks some important factors: **(i) Irrelevant information in images may introduce noise to the text, thus affecting the model's performance.** While the aforementioned methods proposed relation detection, dynamic gate control guidance, and aspect-related detection mechanisms, they did not fundamentally address the issue of reducing the impact of image noise on the text. In Fig. 1(a), although there is considerable noise in the image, the most significant information lies in the presence of the Eiffel Tower and the church. To address the issue of image noise, we propose a target-guided cross-modal MABSA method, which utilizes target detection technology to focus on the important information within the images. **(ii) Neglecting the utilization of syntactic features from text and textual information from images.** The above method directly inputs text into the model, while ignoring syntactic features of the text, such as part-of-speech and dependency features. Part-of-speech features aid the model in identifying the boundaries of multi-word aspect terms, while dependency features assist the model in better understanding sentence structure. Additionally, images often contain text information of significant value; mining this information helps models better understand the input sentences. In Fig. 1(b), identifying the text "TWO OF THESE PRESIDENTS" in the image would aid the model in extracting "Bill Clinton" and "Barack Obama" more effectively. Therefore, we employ Optical Character Recognition (OCR) (Memon et al., 2020) technology in our model to recognize text in images. **(iii) The cross-modal Transformer model architecture needs further optimization.** The above study is based on the original encoder structure of a cross-modal Transformer, which effectively captures sequence relationships in sentences. However, a large number of parameters in the Transformer structure can lead to model overfitting and increased computational resource requirements. Therefore, we further optimized the Transformer

structure in the text auxiliary module, visual auxiliary module, and cross-modal MABSA module. We replaced the Feedforward layers of the Transformer in each of the three modules with a Graph Convolutional Network (GCN) (Kipf & Welling, 2016) and a Convolutional Neural Network (CNN) (Połap & Jaszcz, 2024), along with an alignment auxiliary component.

In this paper, we propose a target-guided cross-modal Transformer method for MABSA, called TCMT.[1] The model consists of three modules: the textual aspect-sentiment extraction module, the visual aspect-sentiment prediction module, and the textual-visual alignment cross-modal MABSA module. In the textual aspect-sentiment extraction module, we utilize a GCN-based cross-modal Transformer to combine textual syntactic features with text information detected in images via Optical Character Recognition (OCR). These syntactic features include word features, part-of-speech features, and dependency features. In the visual aspect-sentiment prediction module, we employ a CNN-based cross-modal Transformer to learn the feature information of images and utilize Adjective-Noun Pairs (ANPs) (Borth et al., 2013) to identify words present in the images, further facilitating the supervised training of these images. In the textual-visual alignment cross-modal MABSA module, we design two components: the target-oriented interaction component and the alignment auxiliary component. The target-oriented interaction component includes two cross-modal attention queries to facilitate learning between the target detection regions, visual module hidden vectors, and textual module hidden vectors. The alignment auxiliary component utilizes the fused hidden vector matrix to perform a dot product computation with the global text-image matrix, further optimizing the cross-modal alignment between text and images.

The main contributions of the paper are summarized as follows:

(1) We propose a TCMT model consisting of three improved cross-modal Transformer module structures. Among these, the textual aspect-sentiment extraction module and the visual aspect-sentiment prediction module serve as auxiliary modules, while the textual-visual alignment cross-modal MABSA is the main module. We conduct supervised learning separately on the two auxiliary modules, followed by cross-modal interaction and modality alignment in the main module.

(2) We introduce object detection regions as key target information in images to reduce the impact of noise from irrelevant areas. Additionally, we leverage the syntactic features of text to assist the model in better identifying the boundaries of multi-word aspect terms, as well as employ OCR technology to detect textual information contained in images.

(3) We conduct extensive experiments and visualization analyses on benchmark datasets. The experimental results indicate that the TCMT model outperforms current state-of-the-art baseline models, enabling better performance in MABSA tasks.

The structure of this paper is as follows: Section 2 introduces related work on ABSA, MSA, and MABSA. Section 3 presents the overall framework of the TCMT model. Section 4 elaborates on the experimental findings and analysis of the TCMT model. Finally, Section 5 summarizes the contributions of the paper.

2. Related work

In this section, we first introduce textual aspect-based sentiment analysis. With the rapid growth of social media, sentiment analysis techniques now extend beyond text alone. Next, we briefly review the development of multimodal sentiment analysis. The shift from coarse-grained to fine-grained multimodal techniques enables more effective information utilization. Finally, we provide a detailed overview of related work in multimodal aspect-based sentiment analysis.

[1] Code is available at: https://github.com/ZouWang-spider/TCMT.

2.1. Textual aspect-based sentiment analysis

Early Sentiment Analysis (SA) (Chauhan et al., 2023; Du et al., 2024) primarily focuses on determining the overall sentiment polarity of paragraphs and sentences, falling under coarse-grained sentiment analysis. However, SA techniques often struggle to effectively handle the coexistence of multiple emotions within a sentence or capture the opinions of entities and sentiments expressed in the text. Therefore, researchers proposed Aspect-Based Sentiment Analysis (ABSA) technology (Zou et al., 2024a) to achieve fine-grained exploration of textual information. Currently, researchers commonly use pre-trained models and neural networks to address ABSA tasks. Commonly employed pre-trained models include Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and BART (Lewis et al., 2020). The neural network models include Long Short-Term Memory (LSTM) (Behera et al., 2021), Text Convolutional Network (TextCNN) (Kim, 2014), and Graph Convolutional Network (GCN) (Shang et al., 2024). In the ABSA task, multiple subtasks need to be implemented, and some researchers have proposed pipeline frameworks (Wang et al., 2017). However, this approach faces the issue of error propagation, where the performance of preceding subtasks directly impacts that of subsequent ones. To mitigate this issue, some scholars have proposed end-to-end frameworks (Peng et al., 2020), implementing all subtasks within a unified structured model. Recently, some scholars have proposed using table-filling methods (Zhai et al., 2023; Zhang et al., 2023b) to convert sequence labeling tasks into two-dimensional tables for completion. Additionally, other researchers have suggested employing multi-task learning (Zheng et al., 2024) and Machine Reading Comprehension (MRC) methods (Mao et al., 2021; Zou et al., 2024b), which utilize a question-answering approach to train all subtasks simultaneously. However, the aforementioned studies primarily focus on the textual modality and do not consider information from other modalities. Users often upload corresponding images in online comments, and the image modality can contain rich information that can enhance text mining.

2.2. Multimodal sentiment analysis

Multimodal Sentiment Analysis (MSA) techniques, which integrate multiple modalities of information to determine sentiment polarity, have garnered widespread attention from researchers. Current research on MSA focuses on conversational MSA and social media MSA. Conversational MSA primarily concentrates on exploring various neural network models to facilitate interaction and fusion between multiple modalities. Models such as LSTM, Gated Recurrent Unit (GRU), and Transformer have demonstrated good performance in tasks such as emotion detection (Liu et al., 2024), sarcasm detection (Yue et al., 2023), and implicit sentiment detection (Wang & Hou, 2023). Some studies also explore fusion paradigms between different modalities, including early and late fusion (Dixit & Satapathy, 2024) as well as temporal attention mechanisms (He et al., 2022). However, these methods are designed for modeling coarse-grained sentiment in conversations and may not effectively capture fine-grained sentiment. In social media MSA, researchers have proposed various models to perform visual sentiment analysis of social images (Chen et al., 2023; Zhang et al., 2023a). Additionally, some scholars integrate text and image information to conduct holistic sentiment analysis of multimodal social posts (Mittal et al., 2024; Xu et al., 2022). Unlike these studies that focus on coarse-grained MSA, our research primarily employs visual modalities to assist in extracting fine-grained textual information.

2.3. Multimodal aspect-based sentiment analysis

Multimodal Aspect-Based Sentiment Analysis (MABSA) (He et al., 2024; Jin et al., 2024) is a newly emerged research task in recent years, comprising two sub-tasks: Multimodal Aspect Term Extraction (MATE)

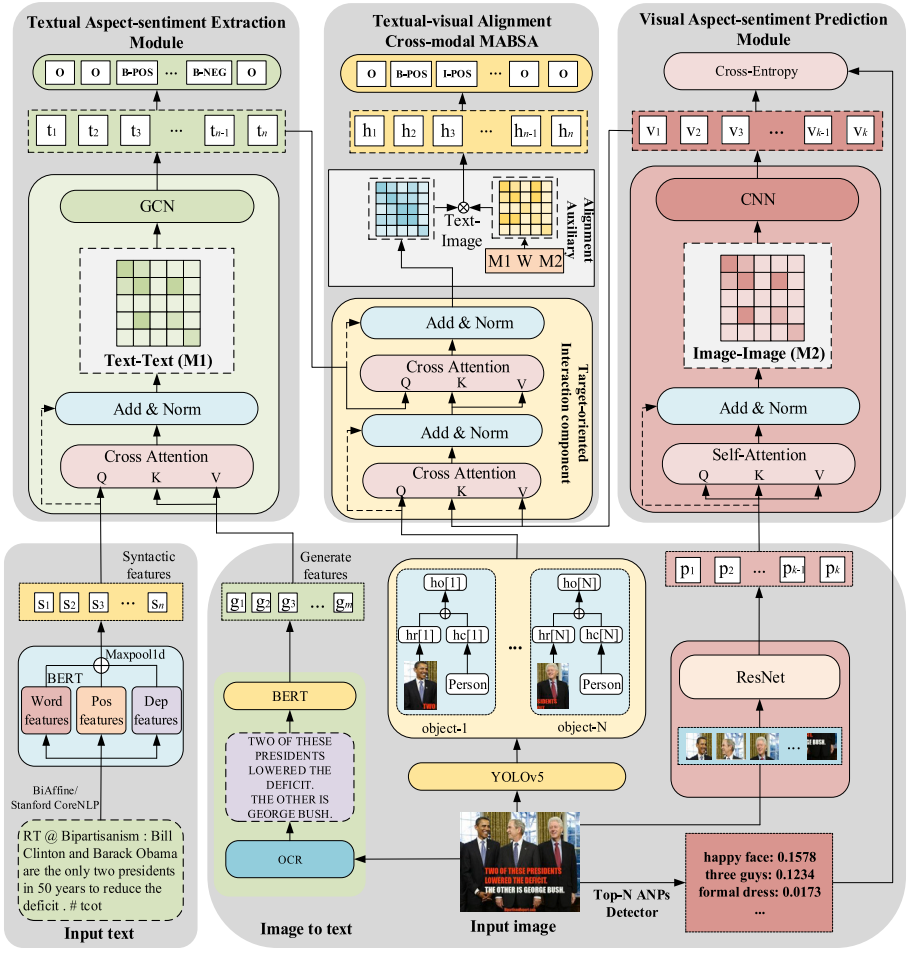


Fig. 2. Overall structure of the TCMT framework. Where M1 represents the text-text correlation matrix, M2 represents the image-image correlation matrix, and W represents the trainable weight matrix.

and Multimodal Aspect-based Sentiment Classification (MASC). MATE is a sequence labeling task, predicting corresponding labels for aspect terms and extracting them, while MASC is a classification task that determines the sentiment polarity of aspect terms. The current challenges of the MABSA task include: (i) leveraging images to enhance the exploration of textual information, and (ii) establishing effective connections between the two sub-tasks of MATE and MASC. In response to the aforementioned challenges, Ju et al. (2021) proposed an end-to-end ABSA framework to jointly address the MATE and MASC tasks. Yang et al. (2022) proposed utilizing both textual and visual modules to enhance the end-to-end implementation of MABSA. Some researchers have introduced pre-trained BART models to fuse textual and image information. Ling et al. (2022) proposed using a BART encoder for textual and image inputs and fine-tuning the BART decoder to accomplish the MATE and MASC subtasks. Zhou et al. (2023) introduced a MABSA model that detects aspect-guided information. This model also employs a BART encoder for textual and image inputs, aligns hidden vectors modally, and utilizes a GCN to learn the relationship between the two modalities. Some scholars explore sentiment information from the perspective of images. Xiao et al. (2024) constructed a model based on image aesthetics, utilizing aesthetic scoring and perception to jointly accomplish the MABSA task. However, effectively utilizing the visual modality to assist the textual modality, while bridging the semantic gap between the two modalities, remains a challenge for the aforementioned methods. Furthermore, irrelevant information in images may act as noise, impeding the extraction of aspect terms from the text. To tackle these challenges, we propose the TCMT framework. This model utilizes object detection regions to guide the visual modality and further aligns it with the text model. Moreover, to leverage the information within images, the model incorporates OCR techniques to capture the textual information contained in the images.

3. Methodology

In this section, we first introduce the formulation of the MABSA task. Then, we provide a detailed explanation of the overall structure of the MABSA framework. Finally, we present detailed introductions to the three modules of the TCMT model. The overall structure of the TCMT model is illustrated in Fig. 2.

3.1. Task formulation

The labels provided by benchmark datasets for the MABSA task follow a unified tagging schema $y_i \in$ {B-POS, I-POS, B-NEU, I-NEU, B-NEG, I-NEG}∪{O}. "B" indicates the starting position of the aspect term, "I" represents other words within the aspect term, "O" denotes irrelevant words, and "POS", "NEU", and "NEG" correspond to the three sentiment polarities of aspect terms. Therefore, the model only needs to perform a 7-class classification task for each word's corresponding label. Given the multimodal tweet dataset D, consisting of text-image pairs where the sentences are represented as $X_i = (x_1, x_2, x_3, \ldots, x_n)$ and their corresponding images as V_i.

3.2. Model overview

The overall framework of TCMT is shown in Fig. 2. The model primarily consists of three modules: the textual aspect-sentiment ex-

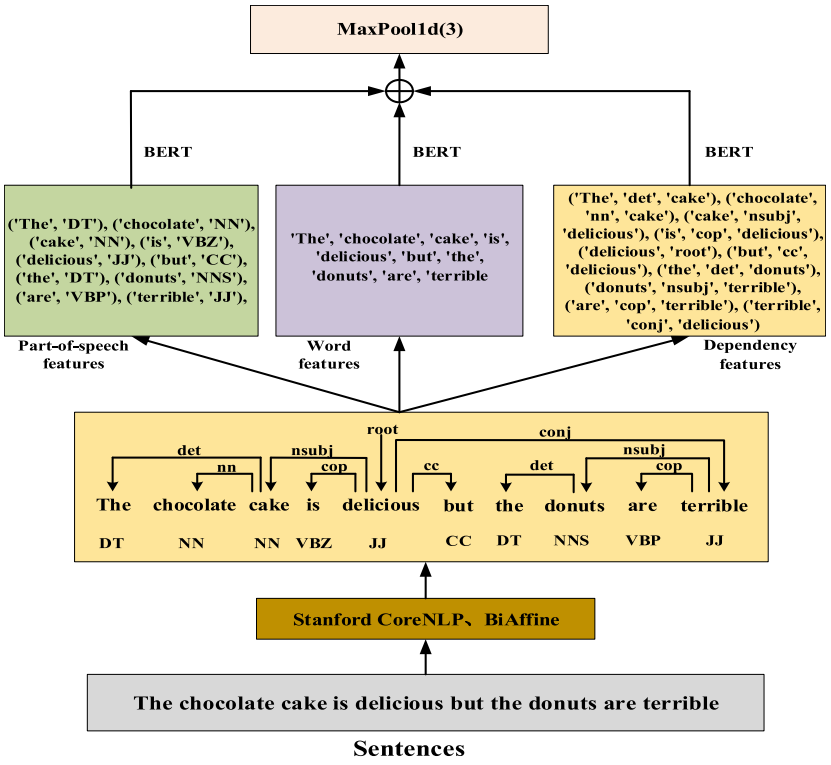


Fig. 3. Representation of syntactic features in text. Where "DT, NN, …" represents the part-of-speech of the words, and "det, nn, …" represents the dependencies between the words.

traction module, the visual aspect-sentiment prediction module, and the textual-visual alignment cross-modal MABSA module. The textual aspect-sentiment extraction module serves as an auxiliary for the supervised training of textual feature information. This module's inputs include the sentence with syntactic features and text captured from the image. The visual aspect-sentiment prediction module serves as an auxiliary for the supervised training of image information. Since the original dataset does not include labels for images, we utilize ANPs to obtain scored word labels from the images. The textual-visual alignment cross-modal MABSA module includes a target-guided interaction learning component and an alignment auxiliary component. Additionally, we employ object detection technology to guide cross-modal interaction learning, thereby reducing the impact of irrelevant noise in images on textual information.

3.3. Textual aspect-sentiment extraction module

Previous researchers transformed input sentences into word vector representations using BERT models, neglecting the syntactic feature information in the text. Therefore, we designed the textual aspect-sentiment extraction module to assist the cross-modal MABSA in better learning rich textual information representations.

Textual syntactic representation: Utilizing dependency features and part-of-speech (POS) features of sentences has been shown to improve model performance, as validated by extensive ABSA research. We utilized Stanford CoreNLP[2] (Manning et al., 2014) and BiAffine[3] (Dozat & Manning, 2016) tools to compute the part-of-speech features and dependency features of the input sentences. The process of handling syntactic features in the text is illustrated in Fig. 3. We represent the part-of-speech features as (word, POS) and the dependency features as (word1, dependency relation, word2). Then, we use the BERT model to separately convert the part-of-speech features, word features, and dependency features into word vectors. Finally, we apply a one-dimensional max-pooling layer with a window size of 3 to extract important features from the word vectors and reduce dimensionality. Given a sentence $X = \{x_i, i \in [1, n]\}$, the process of computing its syntactic feature information is as follows.

$$(x_i, d_i, x_j) = BiAffine(X) \tag{1}$$

$$(x_i, p_i) = \text{StanfordCoreNLP}(X) \tag{2}$$

$$h_i^d = BERT(x_i, d_i, x_j); h_i^w = BERT(x_i); h_i^p = BERT(x_i, p_i) \tag{3}$$

$$S_i = MaxPool1d(3, [h_i^d, h_i^w, h_i^p]) \tag{4}$$

Where d_i represents the dependency relationship, and p_i represents the part-of-speech features. $h_i^d, h_i^w, h_i^p \in \mathbb{R}^{n \times d}$ respectively represent the dependency feature vector, word feature vector, and part-of-speech feature vector. $d = 768$ is the dimension of the vector. $S_i \in \mathbb{R}^{n \times d}$ represents the feature vector after pooling calculation.

Image-to-text representation: The text contained in images is often crucial information, therefore we employ OCR[4] (Memon et al., 2020) technology to extract text content. When there is no text in the image, the OCR technology will output a special token "[PAD]", ensuring that the inputs Q, K, and V for cross-modal attention are all syntactic feature vectors. The detailed computation process for converting an image V to text is as follows.

$$R = OCR(V) \tag{5}$$

$$G_i = BERT(R) \tag{6}$$

[2] https://stanfordnlp.github.io/CoreNLP/
[3] https://github.com/chantera/biaffineparser
[4] https://github.com/madmaze/pytesseract

R represents the OCR text detected from the image. $G_i \in \mathbb{R}^{m\times d}$ represents the word vector representation of image-to-text. m represents the sum of the lengths of the OCR text.

Textual aspect-sentiment extraction: We use a GCN-based cross Transformer to query relevant information from the text generated by images. In this process, the input text serves as the query for cross attention (Jing et al., 2023), while the text generated from images serves as the key and value. Additionally, we convert the hidden vectors into matrices and utilize GCN to learn the syntactic structure of the sentences. The detailed computation process of the GCN-based cross Transformer is as follows:

$$CATT^i(G,S,S) = softmax(\frac{[GW_q^i][SW_k^i]^T}{\sqrt{d/h}})[SW_v^i] \tag{7}$$

$$\bar{H} = W_c[CATT^1(G,S,S);...;CATT^h(G,S,S)] \tag{8}$$

$$\tilde{H} = LayerNorm(\bar{H} + S) \tag{9}$$

$$H^t = GCN(H * H^T, g) \tag{10}$$

Where $\{W_q^i, W_k^i, W_v^i\} \in \mathbb{R}^{d/h\times d}$ is the learnable weight matrix corresponding to query, key, and value in cross Attention. $i \in [1,h]$, h is the number of attention heads. $W_c \in \mathbb{R}^{d\times d}$ represents the weight matrix of multi-head attention. g represents the dependency syntactic graph structure of the input sentence.

Textual auxiliary supervision: In the supervised learning phase, we initially use softmax to compute the probability distribution of predicted labels for each word. Subsequently, we utilize the standard Cross-Entropy (CE) loss function to calculate the loss between the predicted tensor and the true label tensor.

$$y_i^t = softmax(W^T H_i^t + b) \tag{11}$$

$$\mathcal{L}_{text} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n_j} l(y_i^t, y_i) \tag{12}$$

Where N represents the number of samples per training batch, and y_i represents the labels corresponding to each word in the sentence.

3.4. Visual aspect-sentiment prediction module

The benchmark dataset for MABSA does not provide labels for the images. To better capture the feature information from images, we designed the visual aspect-sentiment prediction module to assist in the supervised learning of image information for cross-modal MABSA.

Image representation learning:We employ a ResNet (He et al., 2016) model to convert image patches (Dosovitskiy et al., 2020) into feature vectors. When inputting an image patches $V = (v_1, v_2, v_3, \ldots, v_k)$, the process of computing the representation of the image is as follows.

$$P_i = \text{ResNet}(v_i) \tag{13}$$

Where $P_i \in \mathbb{R}^{k\times d}$ represents the feature vector computed for each patch. k represents the number of patches in an image and $d = 2048$ is the dimensionality of each patch's vector.

Image label detection: Given that the benchmark datasets for MABSA lack corresponding labels for images and a semantic gap exists between images and text, further improvement is needed in the model's learning ability to learn from images. We adopt the ANPs detection method proposed by Chen et al. (2014) in DeepSentiBank.[5] This method can detect aspect-opinion information present in images along with their corresponding probabilities, such as "happy face: 0.157" and "cross arms: 0.0173". This detection computes the probability distribution of 2,089 aspect-opinion pairs in the image. To mitigate noise interference, we select the top 4 as the detected ANPs. The computation process for detecting ANPs in image V is as follows.

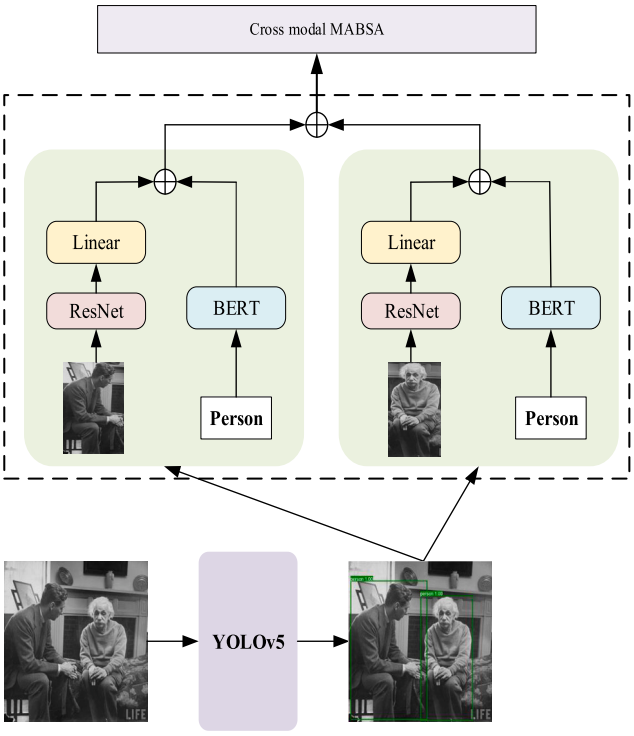


Fig. 4. Extraction of target detection regions in images.

$$A = ANPs(V) \tag{14}$$

$$F = softmax(W^T BERT(A) + b) \tag{15}$$

Where A represents the top 4 aspect-opinion pairs detected computation results.

Visual aspect-sentiment prediction: We use a CNN-based Transformer model to train the feature vectors of images. The computation process is as follows.

$$\bar{H} = W_v[ATT^1(P,P,P);...;ATT^h(P,P,P)] \tag{16}$$

$$\tilde{H} = LayerNorm(\bar{H} + P) \tag{17}$$

$$H^v = CNN(\tilde{H} * \tilde{H}^T) \tag{18}$$

Where ATT represents the self-attention mechanism, and W_v and W^T denotes the trainable weight matrices. V_i represents the probability distribution output by the Transformer.

Visual auxiliary supervision: We use the top 4 ANPs results as labels and apply softmax to convert them into probability distributions. Subsequently, we employ the cross-entropy loss function to compute the difference between the two probability distributions.

$$y^v = softmax(W^T H^v + b) \tag{19}$$

$$\mathcal{L}_{image} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n_j} l(y_i^v, F_i) \tag{20}$$

Where N represents the number of image samples, and n_j denotes the number of image patches in the jth sample.

[5] https://github.com/stephen-pilli/DeepSentiBank

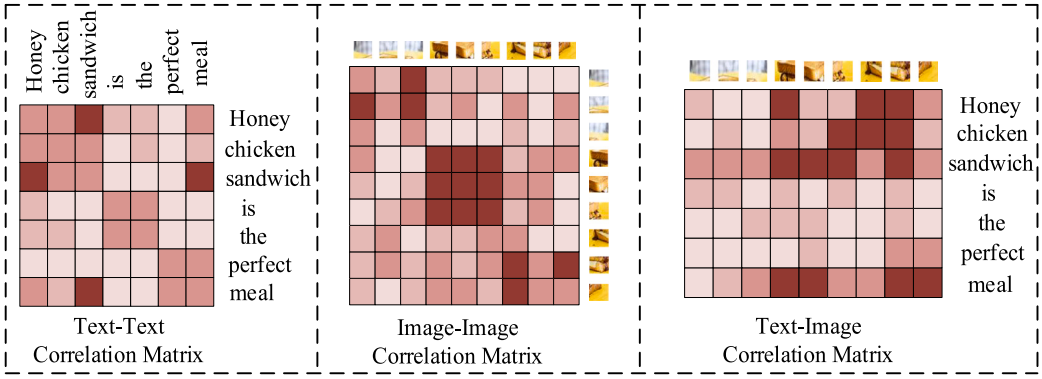


Fig. 5. Correlation matrix between text and image modalities.

3.5. Textual-visual alignment cross-modal MABSA

Previous research has overlooked the noise introduced by irrelevant information in images to text, thereby affecting the overall performance of the model. Therefore, we designed an object detection-guided textual-visual alignment cross-modal MABSA to achieve interaction and alignment between visual and textual modalities.

Target region representation: We utilized the state-of-the-art object detection algorithm You Only Look Once version 5 (Redmon et al., 2016) (YOLOv5[6]) in our model, which effectively detects target regions in the image and labels them. The process of representing these target regions in images is illustrated in Fig. 4. We extract regions with high confidence from the detection results using coordinate-based methods. Subsequently, we employ ResNet to convert these detection regions into feature vectors, and the BERT model to convert labels into word vectors. These vectors are then concatenated to form representations of the target regions. The detailed process of computing the representation of target regions is as follows.

$$(T_i, l_i) = YOLOv5(V) \tag{21}$$

$$h_i^r = Linear(ResNet(T_i)); h_i^c = BERT(l_i) \tag{22}$$

$$h_i^o = Concat(h_i^r, h_i^c) \tag{23}$$

$$O = Concat(h_1^o, \ldots, h_z^o) \tag{24}$$

Where T_i represents the extracted detection regions and l_i represents the corresponding labels, $i \in [1, z]$ and z represents the number of detection regions. $h_i^r \in \mathbb{R}^{1\times768}$ represents the feature vectors of the detection regions, and $h_i^c \in \mathbb{R}^{k\times768}$ represents the word vectors transformed from the labels. $h_i^o \in \mathbb{R}^{(k+1)\times768}$ represents the representation vector of an individual region, while $O \in \mathbb{R}^{z(k+1)\times768}$ represents the concatenated representation vectors of all detection regions.

Textual-visual alignment cross-modal Transformer: The cross-modal MABSA module includes a target-oriented interaction component and an alignment auxiliary component. In the interaction component, we first employ the detection regions to query the hidden vectors from the visual aspect-sentiment prediction module. Then, we utilize the hidden vectors from the textual aspect-sentiment extraction module to query the visual modality. The purpose of this design is to initially capture the contribution of the target regions to the overall image. Next, we establish alignment between textual and visual modalities to reduce the noise impact from irrelevant regions in the image. In the alignment auxiliary component, we adopt a text-image correlation matrix to optimize the hidden vector matrix after interactive learning. The text-image correlation matrix measures the relationship between words and image patches, facilitating modal alignment. An example of the correlation matrix between modalities is shown in Fig. 5.

$$CATT^i(O, H^v, H^v) = softmax(\frac{[OW_q^i][H^vW_k^i]^T}{\sqrt{d/h}})[H^vW_v^i] \tag{25}$$

$$\bar{V} = W_a[CATT^1(O, H^v, H^v); ...; CATT^h(O, H^v, H^v)] \tag{26}$$

$$\tilde{V} = LayerNorm(\bar{V} + O) \tag{27}$$

$$CATT^i(H^t, \tilde{V}, \tilde{V}) = softmax(\frac{[H^tW_q^i][\tilde{V}W_k^i]^T}{\sqrt{d/h}})[\tilde{V}W_v^i] \tag{28}$$

$$\bar{T} = W_b[CATT^1(H^t, \tilde{V}, \tilde{V}); ...; CATT^h(H^t, \tilde{V}, \tilde{V})] \tag{29}$$

$$\tilde{T} = LayerNorm(\bar{T} + H^t) \tag{30}$$

$$h_i = \tilde{T} * \tilde{T}^T \otimes \mathrm{M1} * W * \mathrm{M2} \tag{31}$$

Where $\{W_q^i, W_k^i, W_v^i\} \in \mathbb{R}^{d/h\times d}$ and $\{W_a, W_b\} \in \mathbb{R}^{d\times d}$ are the weight matrices, H^v represents the hidden vectors computed by the visual auxiliary module as per Eq. (20), and H^t denotes the feature vectors computed by the textual auxiliary module as per Eq. (12). $M1 \in \mathbb{R}^{n\times n}$ represents the hidden vector matrix of text-text, $M2 \in \mathbb{R}^{k\times k}$ represents the hidden vector matrix of image-image. $V \in \mathbb{R}^{n\times k}$ is a parameter matrix that can be learned. $h \in \mathbb{R}^{n\times d}$ represents the hidden vectors outputted and n is the number of words in the sentence.

Cross modal MABSA supervision: Similarly, we use the standard cross-entropy loss function along with L2 regularization to calculate the discrepancy between the probability distribution of the cross modal Transformer and the true labels.

$$H_i^c = softmax(W^c h_i + b) \tag{32}$$

$$\mathcal{L}_{MABSA} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n_j} l(H_i^c, y_i) + \lambda\sum_{\theta\in\Theta}\theta^2 \tag{33}$$

Where $W^c \in \mathbb{R}^{n\times n}$ is the learnable weight matrix, and b is the bias term. Θ is the parameter-set of the textual-visual alignment cross-modal MABSA module.

3.6. Model training

To better train the parameters in the model, we adopt a multi-task learning approach to jointly train the overall model. The computation process is as follows.

$$\mathcal{L} = \mathcal{L}_{MABSA} + \alpha\mathcal{L}_{text} + \beta\mathcal{L}_{image} \tag{34}$$

Where α and β are hyperparameters used to control the contributions of the two auxiliary modules, with initial values set to 1/2.

[6] https://github.com/ultralytics/yolov5

Algorithm 1 Inference process of the TCMT model.

Require: Sentence X and Image V
Ensure: Aspect term-sentiment pairs $A = \{(a, s)\}$
Initialize $A = \{\ \}$;
//The textual aspect-sentiment extraction module
Aggregate the three syntactic features of sentence X results in the feature vector S;
Use OCR to detect text in the image and embeding it as the feature vector g.
$H_s \Leftarrow CrossAttention(S, g, g)$;
$H_t \Leftarrow LayerNorm(S + H_s)$;
Use GCN to compute the vector matrix $H_t * H_t^T$ yields the text hidden vector T;
Perform supervised learning for the text;
//The visual aspect-sentiment prediction module
Utilize ANPs to detect the noun-probability pairs y_v in the image V;
$P \Leftarrow ResNet(V)$;
$H_v \Leftarrow Self - Attention(P, P, P)$;
$H_p \Leftarrow LayerNorm(P + H_v)$;
Use CCN to compute the vector matrix $H_p * H_p^T$ yields the image hidden vector F;
Use noun-probability pairs y_v to perform supervised training for the image;
//The textual-visual alignment cross-modal MABSA
Utilize Yolov5 for object detection and label representation as vectors O;
$H_o \Leftarrow Cross\ Attention(O, F, F)$;
$H_f \Leftarrow LayerNorm(O + H_o)$;
$H_m \Leftarrow CrossAttention(T, H_f, H_f)$;
$M_a \Leftarrow H_m * H_m^T$;
$M_b \Leftarrow H_t * H_t^T * V * [H_p * H_p^T]^T$;
Perform dot product computation between feature matrix M_a and matrix M_b to obtain matrix M_h;
$f \Leftarrow FC(M_h)$, where FC is the fully connected layer;
$(a_i, s_i) \Leftarrow softmax(f)$;
$A \Leftarrow \{(a_i, s_i)\}$;
Return A.

Table 2
Statistical information of the benchmark datasets.

Datasets	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234
#Sentence	2101	727	674	1746	577	587

4. Experiments

In this section, we conduct a comparison study, ablation study, attention visualization, in-depth performance analysis, error analysis, and case study to validate the overall performance of the TCMT model.

4.1. Experimental settings

Datasets: We utilize the Twitter2015 and Twitter2017 datasets,[7] which were originally provided by Zhang et al. (2018) for multimodal named entity recognition. Subsequently, Lu et al. (2018) annotated the sentiment polarity for each aspect, and Ling et al. (2022) made corrections to the datasets. Detailed statistical information about the datasets is presented in Table 2. The two datasets provide examples in the form of sentences, aspect terms, sentiment labels (−1, 0, 1 representing negative, neutral, and positive sentiment polarity), and image file names. All image information is stored separately in one file.

Evaluation Metrics: In the experiment, we used commonly used evaluation metrics in machine learning algorithms: Precision (P), Recall (R), F1 score (F1), and Accuracy.

Implementation Details: For text processing, we employ the BERT model to represent the textual syntactic features and the text detected by OCR. For image processing, we utilize pre-trained ResNet50 and YOLOv5 models, with the image patch size based on Dosovitskiy et al. (2020). The experimental hyperparameters are divided into static fixed parameters and dynamic adjustable parameters, with specific settings shown in Table 3.

Table 3
Detailed information of baseline models.

Parameter types	Hyperparameters	Setting
Static fixed parameters	BERT dim	768
	ResNet50 layers	50
	YOLOv5 confidence	0.75
	ANPs Top-N	4
	Cross-attention (Head)	12
	Self-attention (Head)	12
	GCN layers	2
	CNN kernel size	3 × 3
	Fully connected layer	256
	Epochs	50
	Batch size	32
	Learning rate	2e−5
	L2 regularization	0.01
	Optimizer	Adam
Dynamic adjustable parameters	Correlation weight matrix	Trainable
	Dropout rates	0.3; 0.5

4.2. Baseline models

We employed four groups of baseline data for extensive experimentation: MATE task baselines, MASC task baselines, text-based ABSA baselines, and MABSA baselines. The first two groups were used for in-depth performance analysis experiments, while the latter two were utilized for comparison study. A detailed introduction to the baseline models is shown in Table 4.

4.3. Main results

The precision, recall, and F1 scores achieved by our proposed TCMT framework and the baseline models on the two benchmark datasets are shown in Table 5. The experimental results indicate that the performance of the TCMT model significantly surpasses that of the baseline models, achieving F1 scores of 69.8% and 70.8% respectively.

First, we analyze the performance of four baseline models on the text-only modality. Experimental results show that the BART model performs the best. The BART pre-trained model, proposed by the Facebook AI research team, is an encoder–decoder framework that utilizes BERT as the encoder and GPT as the decoder. Compared to the RoBERTa model, the BART model not only has text representation capabilities but also sequence generation abilities, enabling effective extraction of aspect terms and sentiment discrimination within sentences. The BART model achieved F1 scores of 63.9% and 65.4% on the two datasets, respectively. Our proposed TCMT model shows performance improvements of 5.9% and 5.4% compared to the BART model. This improvement is attributed to the BART model's lack of multimodal data learning, while the visual modality in our TCMT model better assists in understanding the semantic information of the text.

[7] Datasets can be downloaded at: https://github.com/NUSTM/VLP-MABSA.

Table 4
Detailed information of baseline models.

Tasks	Methods
MATE	**RAN** (Wu et al., 2020a) aligned text with target regions through a co-attention mechanism, achieving aspect extraction in multi-modal scenarios.
	UMT (Yu et al., 2020) employed a cross-modal Transformer to integrate textual and visual representations, accomplishing the MNER task.
	OSCGA (Wu et al., 2020) utilized the BIO-tagged visual object recognition as the representation of the image features.
MASC	**TomBERT** (Yu & Jiang, 2019) utilized BERT to capture intra-modal dynamics to address the MASC task.
	ESAFN (Yu et al., 2019) was based on LSTM for explicit modeling of text context and utilized image information to assist the model.
	CapTrBERT (Khan & Fu, 2021) translated images into captions as auxiliary sentences to achieve the MASC task.
Text-based ABSA	**SPAN** (Hu et al., 2019) adopted a hierarchical framework to transform the joint ABSA task into a span prediction problem.
	D-GCN (Chen et al., 2020) utilized the syntactic information learned by the GCN model and proposed a directional graph convolutional network to capture the correlations between words.
	RoBERTa (Liu et al., 2019) provided the contextual representation of the text to the self-attention based Transformer layer, and then utilized the CRF layer to accomplish sequence labeling.
	BART (Lewis et al., 2020) transformed the joint ABSA task into an index generation problem, enabling it to adapt the BART pre-training model to complete each sub-task.
MABSA	**UMT+TomBERT** and **OSCGA+TomBERT** improved upon the tasks mentioned above, implementing the MATE and MASC tasks using a pipeline framework.
	UMT-collapsed improved upon the UMT method by adopting collapsed labels (such as B-POS and I-POS) to accomplish the MABSA task.
	OSCGA-collapsed improved upon the OSCGA method by using collapsed labels.
	RpBERT-collapsed utilized a multi-task framework to train the model, achieving image-text relationship detection, and incorporating collapsed labels for improvement.
	CLIP (Radford et al., 2021) is a Transformer-based visual-text pretraining model that effectively represents the input textual and image information.
	JML (Ju et al., 2021) utilized a multi-task learning framework to accomplish the MABSA task, simultaneously addressing the MATE and MASC subtasks, wherein an auxiliary cross-modal relation detection module was established.
	CMMT (Yang et al., 2022) proposed a framework of cross-modal multi-task transformers, introducing a text-guided cross-modal interaction module that dynamically controls the contribution of visual information to the representation of each word in modal interaction.
	VLP-MABSA (Ling et al., 2022) was a pre-trained based MABSA approach that integrated modality information using the BART model, accomplishing the MATE and MASC tasks separately in downstream tasks.
	AoM (Zhou et al., 2023) designed an aspect-aware attention module to simultaneously select text tokens and image patches that are semantically relevant to the aspect. Additionally, graph convolutional networks were used to model interactions between visual-text and text-text.
	Atlantis (Xiao et al., 2024) designed three modules: text-visual alignment aspect-sentiment extraction, sentiment-aware image aesthetic evaluation, and aesthetics-aware joint MABSA. The first two modules are used to assist the latter module in achieving the MABSA task.

Second, we analyzed the performance of baseline models on both image and text modalities. Among the baseline models, the CMMT, AoM, and Atlantis models performed well. The **CMMT** model achieved F1 scores of 66.5% and 68.5% on the two datasets, respectively. In comparison, the TCMT model achieved F1 score improvements of 3.3% and 2.3% on the two datasets. The CMMT model is also designed with a text module and a visual module to auxiliary the end-to-end MABSA. However, it inputs the representation information of the text into the text auxiliary module and utilizes the ResNet model to extract features from images before feeding them into the visual auxiliary module. The CMMT model did not adequately consider the syntactic features of text and the textual information contained in images. Additionally, it overlooked the noise generated by irrelevant areas in images, which the cross-modal Transformer structure could further optimize. The **AoM** is the current state-of-the-art model, achieving F1 scores of 68.6% and 69.7% on the two datasets, respectively. The TCMT model improved upon these results with F1 score increases of 1.2% and 1.1% on the two datasets, respectively. The AoM model is designed based on the BART encoder–decoder framework. Initially, the image and text modalities are input into the BART encoder. Then, the GCN model is employed to learn the hidden vector of the image-text feature matrix and the text-text feature matrix. Finally, the BART decoder extracts aspect terms from the sentences. Although this model performs well, it does not utilize the textual information contained in the images. Additionally, AoM calculates feature matrices using cosine similarity, which makes the constructed matrices susceptible to noise from the modalities. For example, some unrelated regions in the images may have high cosine similarity with the vectors of keywords in the text. The **Atlantis** model achieved F1 scores of 67.3% and 69.4% on two datasets, respectively. Compared to the Atlantis model, the TCMT model achieved F1 score improvements of 2.5% and 1.4% on the same datasets. Similar to the CMMT model, the Atlantis model incorporates text and visual modules to assist the main module in performing MABSA tasks. However, the Atlantis model distinguishes itself by introducing image aesthetics awareness and aesthetic scoring in the visual module and by utilizing

Table 5
Main results (%) of the comparison study. The optimal results are marked in bold black.

Modality	Methods	Venue	Twitter-2015			Twitter-2017		
			P	R	F1	P	R	F1
Text	SPAN	ACL 2020	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN	COLING 2020	58.3	58.8	59.4	64.2	64.1	64.1
	RoBERTa	–	61.8	65.3	63.5	65.5	66.9	66.2
	BART	ACL 2021	62.9	65.0	63.9	65.2	65.6	65.4
Text & Image	UMT+TomBERT	ACL 2021	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT	ACM MM 2020	61.7	63.4	62.5	63.4	64.0	63.7
	UMT-collapsed	ACL 2020	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapsed	ACM MM 2020	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapsed	AAAI 2021	49.3	46.9	48.0	57.0	55.4	56.2
	CLIP	ICML 2021	44.9	47.1	45.9	51.8	54.2	53.0
	JML	EMNLP 2021	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA	ACL 2022	65.1	68.3	66.6	66.9	69.2	68.0
	CMMT	IPM 2022	64.6	68.7	66.5	67.6	69.4	68.5
	AoM	ACL 2023	67.9	69.3	68.6	68.4	71.0	69.7
	Atlantis	Inf. Fusion 2024	65.6	69.2	67.3	68.6	70.3	69.4
Our	TCMT	–	**69.3**	**70.4**	**69.8**	**70.2**	**71.5**	**70.8**

Table 6
Results (%) of the ablation study. The differences in model performance are highlighted in bold black.

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
TCMT	69.3	70.4	69.8	70.2	71.5	70.8
w/o Textual module	(64.6)↓**4.7**	(66.2)↓**4.2**	(65.4)↓**4.4**	(65.0)↓**5.2**	(66.7)↓**4.8**	(65.4)↓**5.4**
w/o Visual module	(65.5)↓**3.8**	(67.3)↓**3.1**	(66.2)↓**3.6**	(66.3)↓**3.9**	(67.0)↓**4.5**	(66.6)↓**4.2**
w/o Main module	(63.9)↓**5.4**	(65.3)↓**5.1**	(64.6)↓**5.2**	(64.4)↓**5.8**	(66.0)↓**5.5**	(65.2)↓**5.6**
w/o Interaction component	(64.8)↓**4.5**	(66.1)↓**4.3**	(65.4)↓**4.4**	(65.5)↓**4.7**	(67.0)↓**4.5**	(66.2)↓**4.6**
w/o Alignment component	(66.3)↓**3.0**	(67.6)↓**2.8**	(66.9)↓**2.9**	(66.7)↓**3.5**	(68.2)↓**3.3**	(67.4)↓**3.4**
w/o Syntactic features	(67.1)↓**2.2**	(67.9)↓**2.5**	(67.4)↓**2.4**	(67.6)↓**2.6**	(68.3)↓**3.2**	(68.0)↓**2.8**
w/o OCR-detected text	(66.7)↓**2.6**	(68.1)↓**2.3**	(67.3)↓**2.5**	(67.4)↓**2.8**	(68.1)↓**3.4**	(67.8)↓**3.0**

image information to introduce image captions in the text modality. Despite these enhancements, the Atlantis model overlooks OCR-detected text contained in images and fails to consider the syntactic features of text. Additionally, due to the diversity of image colors, the aesthetics of images are susceptible to noise from irrelevant elements.

Finally, we analyze the reasons for the superior performance of the TCMT model as follows: (1) It fully utilizes the syntactic features of text and textual information contained in the images, which the CMMT, AoM, and Atlantis models overlook. The syntactic feature information is crucial for accurately identifying boundaries of multi-word aspect terms, such as "liquid crystal display" and "wireless mouse". (2) It uses object detection to focus on important regions in the images, thereby reducing the impact of noise from irrelevant areas. In contrast, the aforementioned models input entire images without focusing on specific regions, which can degrade model performance due to image noise. (3) It is improved in the cross-modal Transformer structure. The specific improvements are as follows: (i) In the text auxiliary module, the model replaces the Feedforward layer with a GCN, designed to facilitate the model in further learning the syntactic graph structure of text and capturing correlations between words. (ii) In the visual assistance module, the model uses CNN to replace the Feed Forward layer, which design-wise helps reduce module parameters to further capture the inter-image patch correlations. (iii) In the cross-modal MABSA module, the model replaces the Feedforward layer with an alignment auxiliary component, which design-wise facilitates further alignment optimization of fused hidden vectors.

4.4. Ablation study

We conducted an ablation study to verify the impact of each component of the TCMT model on its overall performance. The specific components removed from the model during the experiments include the following: (1) "w/o Textual module" represents removing the textual aspect sentiment extraction module and using word vectors from the text. (2) "w/o Visual module" indicates removing the visual aspect sentiment prediction module and using image feature vectors as inputs. (3) "w/o Main module" indicates removing the textual-visual alignment cross-modal MABSA and fusing the hidden vectors outputted by the text and visual auxiliary modules as the final vector. (4) "w/o Interaction component" involves removing the target-oriented interaction component and the YOLOv5 object detection part from the main module. (5) "w/o Alignment component" indicates removing the alignment auxiliary component from the main module and using the hidden vectors obtained from the interaction component as the final vector. (6) "w/o Syntactic features" denotes not using part-of-speech features and dependency features of the text. (7) "w/o OCR-detected text" indicates not using OCR technology to detect text in images, and employing self-attention in the text auxiliary modality.

The ablation experiment results are shown in Table 6. Analyzing the outcomes, it's evident that removing each component affects the overall performance of the model in the following order: main module > textual module > visual module > Interaction component > Alignment component > OCR-detected text > Syntactic features. Notably, the removal of the main module has the most significant impact on the TCMT model's performance. Additionally, the textual module and visual module each have a significant impact on the overall performance of the model, indicating that our improved cross-modal Transformer structure is effective. It is worth noting that the "Interaction component" and the "Alignment component" in the main module also have a certain impact on the model's performance. This suggests that the target-oriented interaction component facilitates the interaction and fusion of information between the two modalities, while the alignment auxiliary component aids in optimizing their alignment. Moreover, the impact of the "Interaction component" is significantly greater than that of the "Alignment component". This is because the target-oriented interaction module uses object detection to focus on important areas of the image, thereby reducing the impact of noise. Removing "syntactic features" and "OCR-detected text" led to a decrease in the overall performance

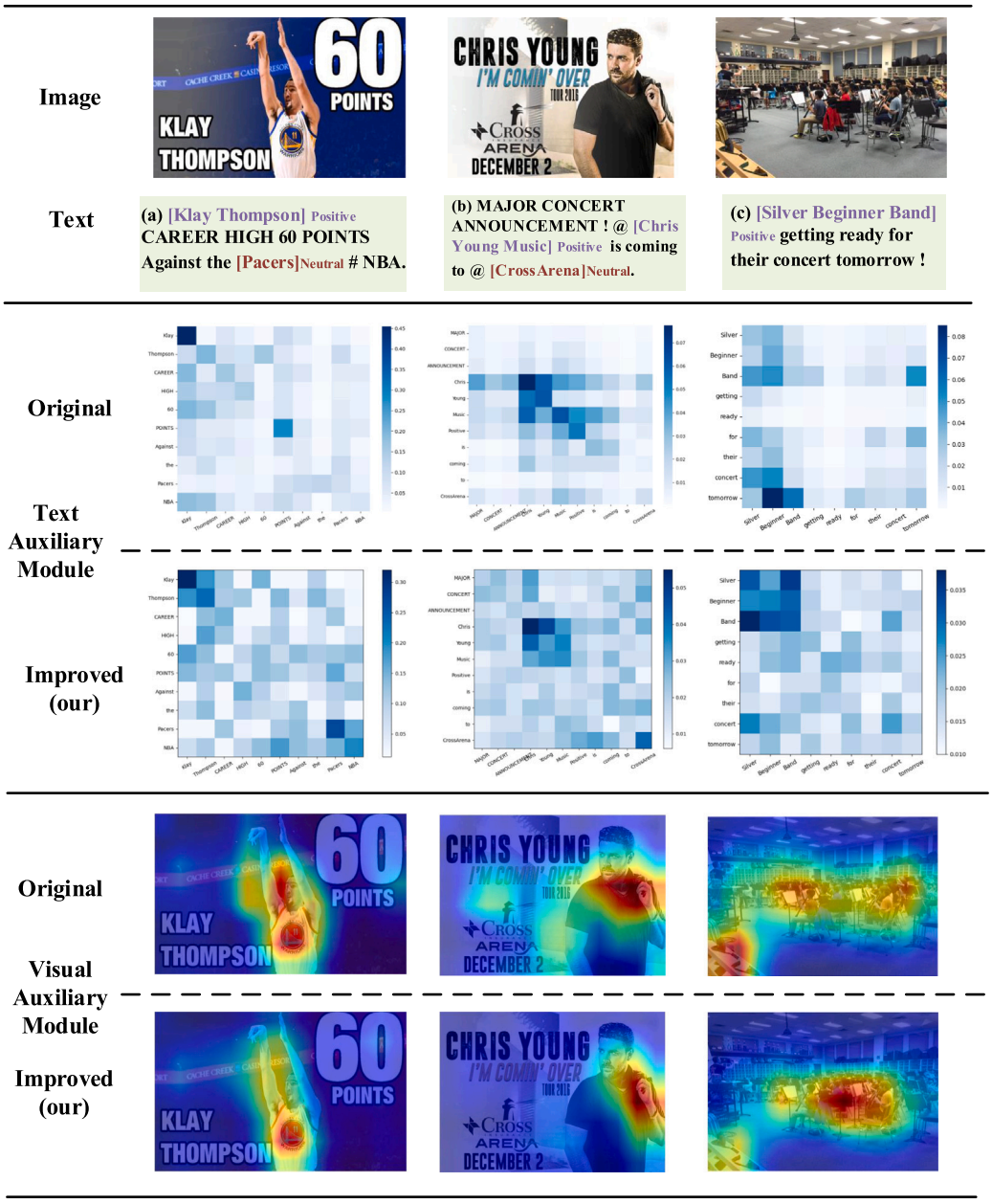


Fig. 6. Attention visualization of the visual auxiliary module and the text auxiliary module in the TCMT model. We highlight aspect terms expressing positive sentiment in red and those expressing neutral sentiment in purple. "Original" indicates the basic Transformer framework, while "Improved" represents our improved Transformer framework.

of the model. This indicates that syntactic features enhance the model's understanding of text semantics, while OCR-detected text provides important information and rich context.

4.5. Attention visualization

To validate the effectiveness of the text and visual auxiliary module for cross-modal MABSA, we conducted an attention visualization experiment to scrutinize the attention distribution of these modules. The results of the attention visualization experiment are shown in Fig. 6, displaying the attention distribution of the original Transformer-based module and the improved module for both text and images. From the attention visualization results of the text auxiliary module, it is evident that the improved GCN-based Transformer can capture aspect terms more effectively compared to the original structure. Furthermore, the improved module can more accurately identify multi-word aspect terms in sentences, such as "Klay Thompson", "Chris Young Music", and "Silver Beginner Band".

We believe the performance improvement of the text auxiliary module is due to leveraging the syntactic feature information from the text. Additionally, using GCN can learn syntactic graph structures, thereby accurately focusing on the aspect terms in the sentences. From the attention visualization results of the visual-assisted module indicate that our improved CNN-based Transformer focuses more attention on key information than the original structure. The original Transformer structure exhibits relatively dispersed attention values for the images, whereas our improved structure concentrates attention, making it easier for the model to capture key information. We believe the improvement in visual performance is due to the use of CNN in the module, which enhances the focus on image feature information, resulting in better attention visualization outcomes.

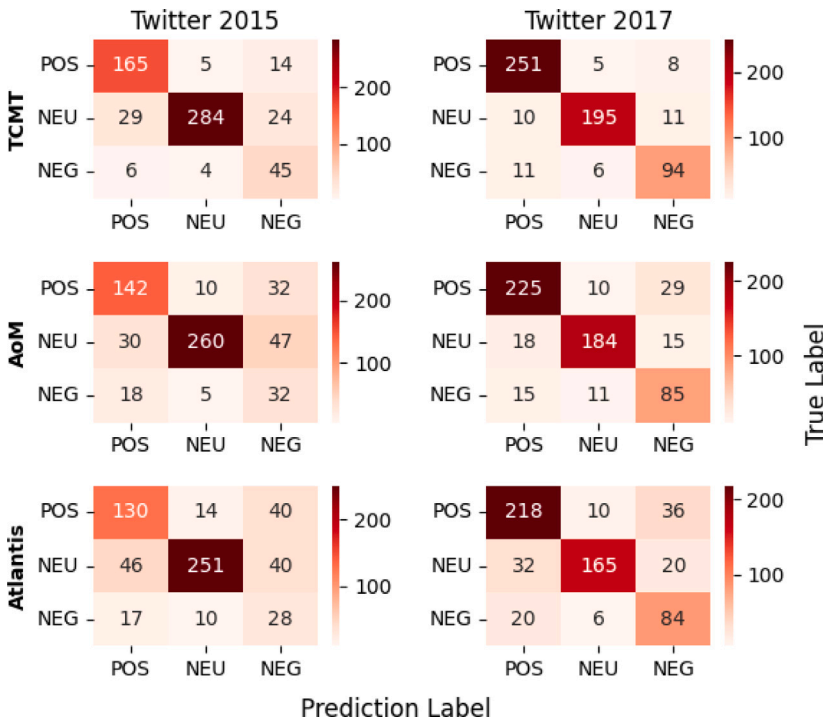


Fig. 7. Results of the confusion matrix for multi-word aspect term extraction.

Table 7
Performance breakdown by sentiment polarity is reflected in the F1 score (%). "POS, NEG, NEU" represent the positive, negative, and neutral sentiment polarities, respectively. "All" encompasses all sentiment polarities.

Methods	Twitter-2015				Twitter-2017			
	POS	NEG	NEU	All	POS	NEG	NEU	All
RoBERTa	58.4	54.5	67.7	63.5	66.3	63.1	66.9	66.2
UMT-RoBERTa	58.2	53.7	68.1	63.9	66.8	61.1	68.0	66.7
CMMT	63.9	55.2	69.6	66.5	70.4	62.9	68.5	68.5
Atlantis	63.3	**60.0**	70.4	67.3	70.0	**63.0**	70.7	69.4
TCMT	**64.2**	58.3	**71.5**	**68.2**	**71.2**	62.7	**71.7**	**70.5**

Table 8
Experimental results (%) for the MATE subtask of the model. The optimal results are marked in bold black.

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
RAN	80.5	81.5	81.0	90.7	90.7	90.0
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
JML	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA	83.6	87.9	85.7	90.8	92.6	91.7
CMMT	83.9	88.1	85.9	92.2	93.9	93.1
AoM	84.6	87.9	86.2	91.8	92.8	92.3
Atlantis	84.2	87.7	86.1	91.8	93.2	92.7
TCMT	**85.8**	**89.4**	**87.6**	**93.5**	**94.1**	**93.8**

4.6. In-depth performance analysis

In the in-depth performance analysis experiments, we conducted a performance breakdown by sentiment polarity to verify the model's performance on three emotions. Additionally, we conducted experiments on the MATE and MASC subtasks of MABSA, respectively.

Performance breakdown by sentiment polarity: Table 7 shows the performance comparison between the TCMT model and baseline models across sentiment categories on both datasets. The results confirm that the TCMT model significantly outperforms the baseline models in overall sentiment analysis. Specifically, in the positive sentiment category across both datasets, the TCMT model surpasses the Atlantis model by 0.9% and 1.2%, respectively. Comparable results are observed in the negative sentiment polarity. For neutral sentiment across the two datasets, the TCMT model achieved F1 score improvements of 1.1% and 1.0%, respectively, over the Atlantis model. Moreover, the TCMT model demonstrates overall sentiment improvements of 0.9% and 1.1% on the datasets, respectively. The experimental results on the Twitter-2015 and Twitter-2017 datasets show that the TCMT model effectively utilizes information from both visual and textual modalities to determine aspect sentiment polarity.

Performance breakdown by MATE and MASC subtasks: To further validate the model's performance on both MATE and MASC subtasks, we conducted performance breakdown experiments for the two subtasks. The experimental results for the MATE subtask are shown in Table 8, while the experimental results for the MASC subtask are shown in Table 9. Analysis of Table 8 reveals that our TCMT model significantly outperforms the baseline models in the MATE subtask. Among the baseline models, the AoM and Atlantis models show relatively strong performance. The TCMT model achieves F1 score improvements of 1.4% and 1.5% compared to the AoM model on both datasets. Compared to the Atlantis model, our model achieves F1 score improvements of 1.5% and 1.1% in the MATE subtask. The superior performance on the MATE task demonstrates that the TCMT model can better utilize information from both text and visual modalities to extract aspect terms from sentences. Leveraging the information contained in images enables the model to better understand the semantics of sentences.

From the experimental results in Table 9, it is evident that the performance of the TCMT model on the MASC subtask is also significantly better than that of the baseline models. Among the baseline models, the AoM and Atlantis models perform relatively well. Compared to the AoM model, our TCMT model achieved accuracy improvements of 1.2% and 0.9%, and F1 score increases of 0.8% across both datasets on the MASC subtask. Compared to the Atlantis model, the TCMT model achieved accuracy improvements of 2.1% and 3.1% on the MASC subtask across both datasets. The MASC subtask results demonstrate that the TCMT model effectively leverages information from both modalities to accurately determine the sentiment polarity of aspect terms. This further

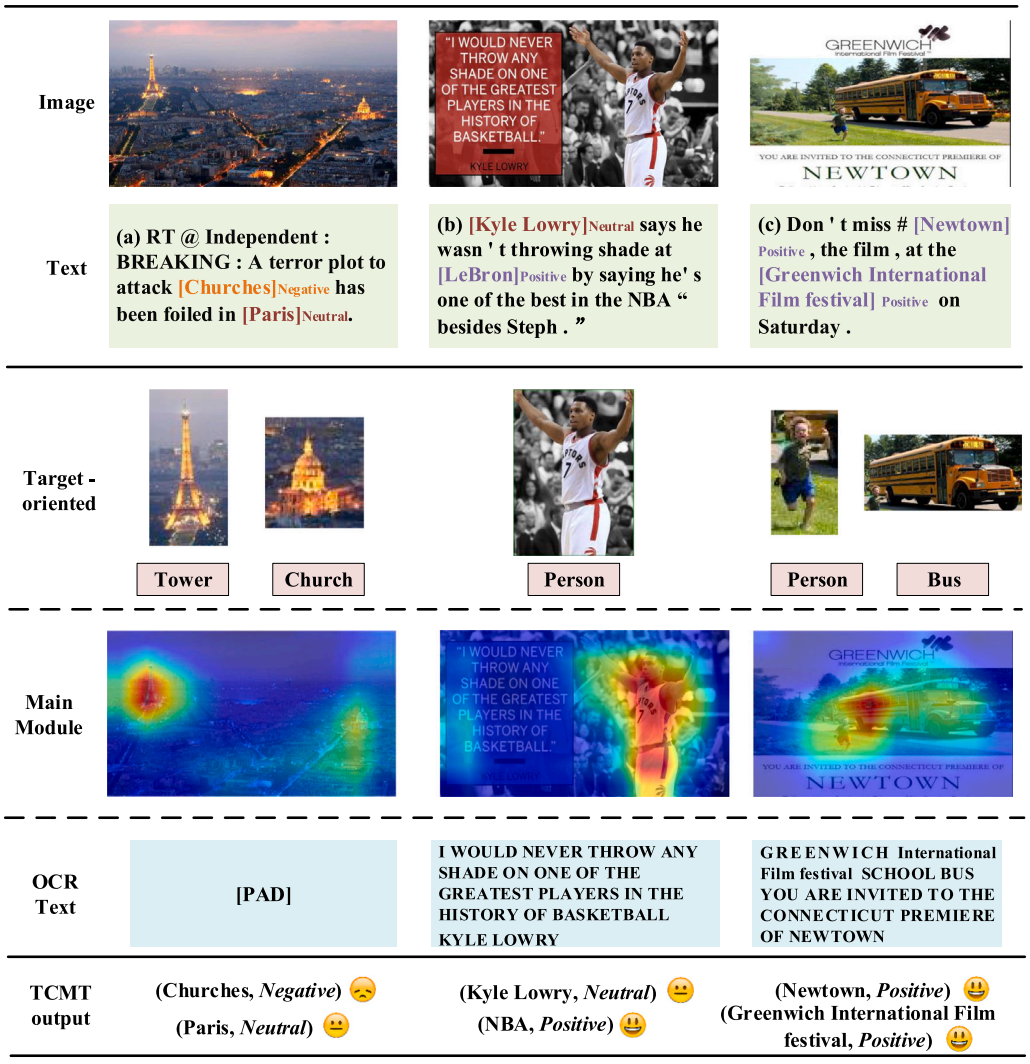


Fig. 8. Error analysis experimental results of the technical components in the TCMT model. The positive aspect terms are highlighted in purple, neutral aspect terms are highlighted in red, and negative aspect terms are highlighted in yellow.

Table 9
Experimental results (%) for the MASC subtask of the model. The optimal results are marked in bold black.

Methods	Twitter-2015		Twitter-2017	
	Accuracy	F1	Accuracy	F1
TomBERT	77.2	71.8	70.5	68.0
ESAFN	73.4	67.4	67.8	64.2
CapTrBERT	78.0	73.2	72.3	70.2
JML	78.7	–	72.7	–
VLP-MABSA	78.6	73.8	73.8	71.8
CMMT	77.9	–	73.8	–
AoM	80.2	75.9	76.4	75.0
Atlantis	79.3	–	74.2	–
TCMT	**81.4**	**76.7**	**77.3**	**75.8**

validates the effectiveness of the model's image target guidance, text auxiliary module, and visual auxiliary module.

4.7. Error analysis

To further analyze the performance of the TCMT model, we conducted error analysis experiments, including confusion matrices for multi-word aspect terms extraction and analysis of model technical components. The experimental results of the error analysis are shown in Figs. 7 and 8.

Confusion matrix for multi-word aspect terms extraction: Many aspect terms in sentences consist of multiple words, such as names and locations. We counted the number of multi-word aspect terms in the test sets of both datasets.[8] In the Twitter 2015 test set, there are 184 multi-word aspect terms with positive sentiment polarity, 337 with neutral sentiment polarity, and 55 with negative sentiment polarity. In the Twitter 2017 test set, there are 264 multi-word aspect terms with positive sentiment polarity, 217 with neutral sentiment polarity, and 111 with negative sentiment polarity. The confusion matrix for multi-word aspect terms extraction is shown in Fig. 7. The figure illustrates that the TCMT model significantly outperforms the AoM and Atlantis models in extracting multi-word aspect terms across both datasets. As both datasets contain numerous multi-word aspect terms, the AoM and Atlantis models overlooked these terms in the sentences. Extracting multi-word aspect terms requires accurate identification of word boundaries, increasing the complexity of the MATE subtask. To address this challenge, we utilized the part-of-speech and dependency features of the sentences. Additionally, textual information from the images provided rich context for the sentences.

Analysis of model technical components: We conducted error analysis experiments on the main module of the TCMT model and

[8] https://github.com/ZouWang-spider/TCMT/blob/main/TCMT/CountSentiment.py

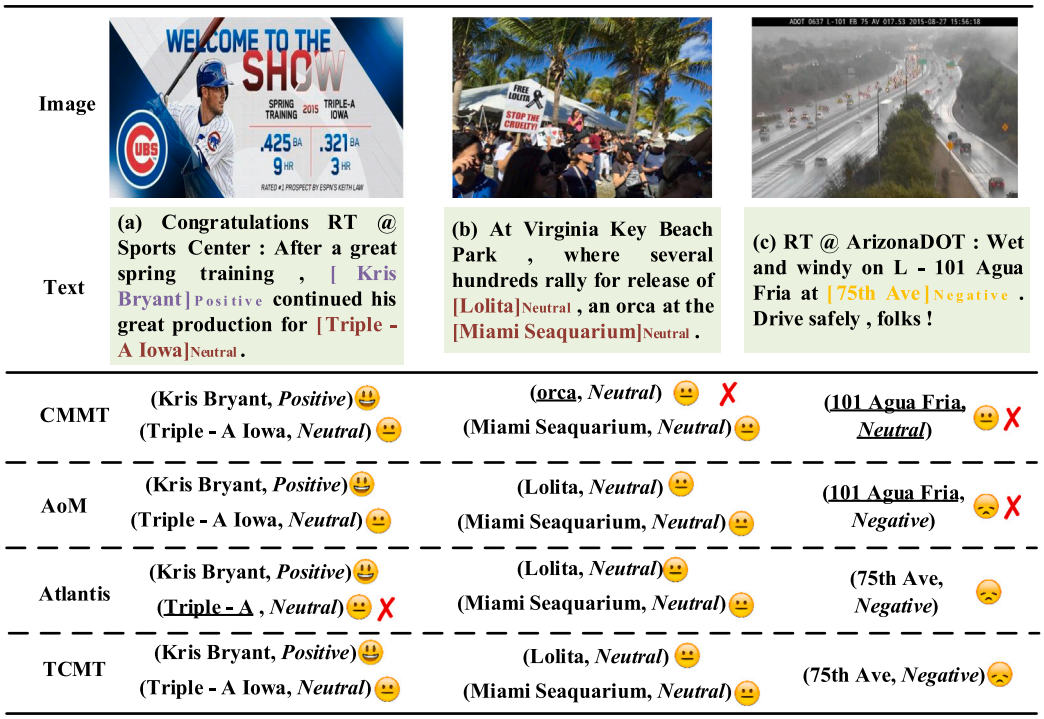


Fig. 9. Case study results. We underline the incorrectly predicted results.

OCR-detected text. The experimental results are presented in Fig. 8. The experimental results clearly show that the target-oriented interaction component in the main module effectively detects target regions in images that correspond to aspect term information in the text. The visualized attention results of the main module indicate that the model's visual attention will be concentrated around the target under the influence of the target-oriented interaction component. This enables the model to effectively focus on the target information, reducing the impact of noise from irrelevant areas in the images. With the support of the target-oriented interaction component and the alignment auxiliary component in the main module, the TCMT model can accurately extract aspect terms from sentences and determines their corresponding sentiment polarity. Additionally, the OCR-detected text reveals that OCR technology can extract key information from images. This information is crucial for extracting aspect terms, such as "KYLE LOWRY", "NEWTOWN", and "GREENWICH International Film Festival". The error analysis results indicate that the target-oriented interaction and alignment auxiliary modules in TCMT enhance the interaction and alignment of textual and visual modalities. Additionally, the introduction of OCR technology enriches the context of sentences by enabling the extraction of crucial information from images.

4.8. Case study

To verify the performance of the TCMT model on online Twitter comments, we conducted a case study experiment, selecting CMMT, AoM, and Atlantis as baseline models. The experimental results are shown in Fig. 9. The case study results demonstrate that the TCMT model can accurately extract aspect terms from tweets and determine their corresponding sentiment polarities. In the sample from Fig. 9(a), the Atlantis model fails to accurately extract the aspect term "Triple-A Iowa". In contrast, the TCMT model employs OCR technology to capture the text "Triple-A Iowa" from the image, thus successfully completing the MABSA task. In the example shown in Fig. 9(b), the CMMT model incorrectly extracts the aspect word "Lolita". Our TCMT model combines object detection results to identify crowds in images, while OCR accurately detects the "FREE LOLITA" text in images. In the example illustrated in Fig. 9(c), the image contains cluttered traffic information that affects the baseline model's ability to extract aspect terms. The CMMT and AoM incorrectly extracted "101 Agua Fria". However, the TCMT model accurately extracted the aspect term "75th Ave" from the sentence. This is attributed to the TCMT model's object detection, which identifies multiple "cars" in the image, facilitating further learning of traffic-related information from the text. Additionally, OCR technology effectively captured the text "75th Ave" within the image. The case study results demonstrate that the TCMT model, utilizing target-oriented multimodal information interaction alignment learning along with OCR for capturing key information from images, effectively handling complex and diverse multimodal social media data, thereby improving MABSA task performance.

5. Conclusions

This paper proposes a target-oriented cross-modal multimodal sentiment analysis method (TCMT). The model consists of textual and visual auxiliary modules and a main module: the textual aspect-sentiment extraction module, the visual aspect-sentiment prediction module, and the textual-visual alignment cross-modal MABSA module. In the textual auxiliary module, we fully utilize the syntactic features of the text to assist the model in identifying the boundaries of multi-word aspect terms, as well as employing OCR technology to capture textual information contained within images. Additionally, we have improved the textual auxiliary module by incorporating a GCN-based Transformer structure and enhanced the visual auxiliary module by using a CNN-based Transformer structure. In the cross-modal MABSA module, we design a target-oriented interaction component to mitigate the influence of image noise and an alignment auxiliary component to further optimize cross-modal training. We conducted extensive experiments on the Twitter-2015 and Twitter-2017 datasets. The main results demonstrate that TCMT significantly outperforms baseline models. Findings from ablation and visualization experiments indicate that the text and visual auxiliary modules effectively enhance the model's overall performance. In-depth performance analysis and error analysis experiments confirm that leveraging syntactic features aids the model in extracting

multi-word aspect terms from sentences. Additionally, ORC technology provides critical contextual knowledge for the model. Case study results indicate that TCMT performs the MABSA task more effectively on the tweets.

However, the TCMT method has some limitations. The three-module structure based on the cross-transformer framework inevitably leads to a larger number of parameters. Additionally, the model's utilization of visual aspects from images can be further improved, such as the colors of the images background, facial expressions of individuals, and spatial relationships between objects. In future work, we intend to reduce the parameters of the TCMT framework by applying a multi-task parameter-sharing approach. Furthermore, we plan also to design separate aspect-visual perception and sentiment-visual perception mechanisms, utilizing the sentiment-visual perception mechanism to better capture the emotional polarity embedded in images. We believe this work can inspire more researchers' interest and creativity.

CRediT authorship contribution statement

Wang Zou: Conceptualization, Methodology, Writing. **Xia Sun:** Methodology, Data curation, Funding acquisition, Review & editing. **Wenhuan Wu:** Conceptualization, Resources, Funding acquisition. **Qiang Lu:** Review. **Xiaodi Zhao:** Review. **Qirong Bo:** Review & editing. **Jianqiang Yan:** Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by the National Natural Science Foundation (No. 61877050), the Hubei Province Education Department Science and Technology Research Project, China (No. Q20201801), and the Hubei University of Automotive Technology PhD Research Start-up Fund Project, China (No. BK202004).

Data availability

Data will be made available on request.

References

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing and Management*, *58*(1), Article 102435.

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on multimedia* (pp. 223–232).

Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, *49*, 100576.

Chen, T., Borth, D., Darrell, T., & Chang, S. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586.

Chen, D., Su, W., Wu, P., & Hua, B. (2023). Joint multimodal sentiment analysis based on information relevance. *Information Processing and Management*, *60*(2), 103193.

Chen, G., Tian, Y., & Song, Y. (2020). Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics* (pp. 272–279).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4171–4186).

Dixit, C., & Satapathy, S. M. (2024). Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications*, *240*, Article 122579.

Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, *118*, 272–299.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. In *International conference on learning representations*.

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial sentiment analysis: techniques and applications. *ACM Computing Surveys*.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444.

He, Y., Huang, X., Zou, S., & Zhang, C. (2024). PSAN: Prompt Semantic Augmented Network for aspect-based sentiment analysis. *Expert Systems with Applications*, *238*, 121632.

He, Y., Sun, L., Lian, Z., Liu, B., Tao, J., Wang, M., & Cheng, Y. (2022). Multimodal temporal attention in sentiment analysis. In *Proceedings of the 3rd international on multimodal sentiment analysis workshop and challenge* (pp. 61–66).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, M., Peng, Y., Huang, Z., Li, D., & Lv, Y. (2019). Open-domain targeted sentiment analysis via span-based extraction and classification. In *ACL* (pp. 537–546). Association for Computational Linguistics.

Jin, W., Zhao, B., Zhang, Y., Huang, J., & Yu, H. (2024). WordTransABSA: Enhancing Aspect-based Sentiment Analysis with masked language modeling for affective token prediction. *Expert Systems with Applications*, *238*, Article 122289.

Jing, P., Cui, K., Guan, W., Nie, L., & Su, Y. (2023). Category-aware multimodal attention network for fashion compatibility modeling. *IEEE Transactions on Multimedia*, *25*, 9120–9131.

Ju, X., Zhang, D., Xiao, R., Li, J., Li, S., Zhang, M., & Zhou, G. (2021). Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4395–4405).

Khan, Z., & Fu, Y. (2021). Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3034–3042).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. In *International conference on learning representations*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7871–7880).

Ling, Y., Yu, J., & Xia, R. (2022). Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics* (pp. 2149–2159).

Liu, Y., Ott, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, *107*, Article 102300.

Lu, D., Neves, L., Carvalho, V., Zhang, N., & Ji, H. (2018). Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics* (pp. 1990–1999).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System demonstrations* (pp. 55–60). Association for Computational Linguistics.

Mao, Y., Shen, Y., Yu, C., & Cai, L. (2021). A joint training dual-mrc framework for aspect based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13543–13551).

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review. *IEEE Access*, 142642–142668.

Mittal, T., Chowdhury, S., Guhan, P., Chelluri, S., & Manocha, D. (2024). Towards determining perceived audience intent for multimodal social media posts using the theory of reasoned action. *Scientific Reports*, *14*(1), 10606.

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., & Si, L. (2020). Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8600–8607).

Połap, D., & Jaszcz, A. (2024). Decentralized medical image classification system using dual-input CNN enhanced by spatial attention and heuristic support. *Expert Systems with Applications*, Article 124343.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Shang, W., Chai, J., Cao, J., Lei, X., Zhu, H., Fan, Y., & Ding, W. (2024). Aspect-level sentiment analysis based on aspect-sentence graph convolution network. *Information Fusion*, *104*, Article 102143.

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics, Meeting. NIH Public Access* (p. 6558).

Wang, H., & Hou, M. (2023). Quantum-like implicit sentiment analysis with sememes knowledge. *Expert Systems with Applications*, *232*, Article 120720.

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Coupled multi-layer attentions for co-extraction of aspect and opinion terms*.

Wu, H., Cheng, S., Wang, J., Li, S., & Chi, L. (2020). Multimodal aspect extraction with region-aware alignment network. In *NLPCC* (pp. 145–156).

Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H., & Li, Q. (2020). Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1038–1046). Association for Computing Machinery.

Xiao, L., Wu, X., Xu, J., Li, W., Jin, C., & He, L. (2024). Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, Article 102304.

Xu, B., Huang, S., Du, M., Wang, H., Song, H., Sha, C., & Xiao, Y. (2022). Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the 29th international conference on computational linguistics* (pp. 1855–1864).

Yang, L., Na, J. C., & Yu, J. (2022). Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing and Management*, *59*(5), Article 103038.

Yu, J., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*.

Yu, J., Jiang, J., & Xia, R. (2019). Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 429–439.

Yu, J., Jiang, J., Yang, L., & Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL* (pp. 3342–3352).

Yue, T., Mao, R., Wang, H., Hu, Z., & Cambria, Z. (2023). KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, *100*, Article 101921.

Zhai, Z., Chen, H., Li, R., & Wang, X. (2023). USSA: A Unified Table Filling Scheme for Structured Sentiment Analysis. In *Proceedings of the 61st annual meeting of the association for computational linguistics* (pp. 14340–14353).

Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*.

Zhang, Y., Jia, A., Wang, B., Zhang, P., Zhao, D., Li, P., Hou, Y., Jin, X., Song, D., & Qin, J. (2023). M3GAT: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems*, *42*(1), 1–32.

Zhang, M., Zhu, Y., Liu, Z., Bao, Z., Wu, Y., Sun, X., & Xu, L. (2023). Span-level aspect-based sentiment analysis via table filling. In *Proceedings of the 61st annual meeting of the association for computational linguistics* (pp. 9273–9284).

Zheng, Y., Gong, J., Wen, Y., & Zhang, P. (2024). DJMF: A discriminative joint multi-task framework for multimodal sentiment analysis based on intra-and inter-task dynamics. *Expert Systems with Applications*, *242*(2024), Article 122728.

Zhou, R., Guo, W., Liu, X., Yu, S., Zhang, Y., & Yuan, X. (2023). AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics* (pp. 8184–8196).

Zhou, J., Zhao, J., Huang, J. X., Hu, Q. V., & He, L. (2021). MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, *455*, 47–58.

Zou, W., Zhang, W., Tian, Z., & Wu, W. (2024). A syntactic features and interactive learning model for aspect-based sentiment analysis. *Complex & Intelligent*, 1–19.

Zou, W., Zhang, W., Wu, W., & Tian, Z. (2024). A multi-task shared cascade learning for aspect sentiment triplet extraction using BERT-MRC. *Cognitive Computation*, 1–18.