Aspect Enhancement and Text Simplification in Multimodal Aspect-Based Sentiment Analysis for Multi-Aspect and Multi-Sentiment Scenarios

Linlin Zhu[1], Heli Sun[1]*, Qunshu Gao[1], Yuze Liu[2], Liang He[1]

[1]College of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2]College of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
zhulinlin@stu.xjtu.edu.cn, hlsun@xjtu.edu.cn, 13673559526@163.com, yuzeliu@stu.xjtu.edu.cn, lhe@xjtu.edu.cn

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) plays a pivotal role in the advancement of sentiment analysis technology. Although current methods strive to integrate multimodal information to enhance the performance of sentiment analysis, they still face two critical challenges when dealing with multi-aspect and multi-sentiment data: i) the importance of aspect terms within multimodal data is often overlooked, and ii) models fail to accurately associate specific aspect terms with corresponding sentiment words in multi-aspect and multi-sentiment sentences. To tackle these problems, we propose a novel multimodal aspect-based sentiment analysis method that combines **A**spect **E**nhancement and **T**ext **S**implification (**AETS**). Specifically, we develop an aspect enhancement module that boosts the ability of model to discern relevant aspect terms. Concurrently, we employ text simplification module to simplify and restructure multi-aspect and multi-sentiment texts, accurately capturing aspects and their corresponding sentiments while reducing irrelevant information. Leveraging this method, we perform three tasks including multimodal aspect term extraction, multimodal aspect sentiment classification, and joint multimodal aspect-based sentiment analysis. Experimental results indicate that our proposed AETS model achieved state-of-the-art performance on two benchmark datasets.

Introduction

Multimodal aspect-based sentiment analysis is a complex and challenging research area that aims to address the problem of accurately capturing and interpreting the sentiment responses of users from diverse sources of information (Gandhi et al. 2023; Zhou et al. 2023). It is usually classified into three categories: Multimodal Aspect Term Extraction (MATE), Multimodal Aspect Sentiment Classification (MASC) and Joint Multimodal Aspect-based Sentiment Analysis (JMASA). MATE aims at identifying and extracting specific aspect terms from given multimodal data. MASC aims at determining the sentiment propensities of these aspects. JMASA is an integrative task that aims to simultaneously extract and categorize the aspects and their sentiments. As a hot challenge in the field of multimodal learning, MABSA has attracted wide attention.

*Corresponding author.

(a) Aspect terms in the image. (b) An illustration of text simplification.

Figure 1: An example of multimodal data with multi-aspect and multi-sentiment.

Previous research widely relies on the integration of visual and textual information to carry out sentiment analysis. Ju et al. (2021) are the first to define the concept of the JMASA task and propose a joint learning approach incorporating the mechanism of image-text relations to evaluate the contribution of visual content to the extraction of aspect-sentiment pairs. Subsequently, Ling et al. (2022) propose a task-specific Vision-Language Pre-training framework for MABSA (VLP-MABSA), which is a unified multimodal encoder-decoder architecture for all the pretraining and downstream tasks. More recently, Yang et al. (2022) design a multitask learning architecture named Cross-Modal Multitask Transformer (CMMT) for the End-to-End MABSA task. Zhou et al. (2023) propose an Aspect-oriented Method (AoM) to detect aspect-relevant semantic and sentiment information. Xiao et al. (2023) present Cross-modal Fine-grained Alignment and Fusion Network (CoolNet) to boost the performance of visual-language models in seamlessly integrating vision and language information. Recently, to align images and text within the language modality across multiple granularities, Xiao et al. (2024) introduce Atlantis, an innovative aesthetic-oriented methodology for JMASA task.

Although current methods have made impressive strides, they still encounter several challenges: **1) How to accurately identify all aspects?** Figure 1 presents a multimodal example with multi-aspect and multi-sentiment. In MABSA, extracting all aspect terms from multimodal data is essential but challenging. For example, the VLP-MABSA (Ling, Yu, and Xia 2022) successfully identified the "*Daniel Radcliffe*" aspect in another example, yet overlooked the "*Elijah*

Wood" aspect. Extracting comprehensive aspect terms from multimodal data is fundamental for MABSA. **2) How to correlate specific aspects with their corresponding sentiments in multi-aspect and multi-sentiment sentences?** Figure 1(b) demonstrates a text with two aspects, "*food*" and "*service*", associated with the sentiments "positive" and "negative", respectively. In multi-sentiment analysis, accurate aspect-sentiment alignment and reduction of word interference are crucial for improving MABSA performance.

To tackle these challenges, we propose a multimodal aspect-based sentiment analysis method based on aspect enhancement and text simplification (AETS). Firstly, we utilize RoBERTa (Liu et al. 2019) and ViT (Dosovitskiy et al. 2020) respectively to obtain feature representations for text and image. Subsequently, we design an aspect enhancement module that constructs masked vector to mask aspect terms in both text and image, thereby increasing the sensitivity of model towards aspect terms. Following this, we devise a text simplification module, which constructs an undirected graph based on the syntactic dependency relationships within the sentences of the text. Employing the Breadth-First Search (BFS) (Kurant, Markopoulou, and Thiran 2010) algorithm, we extract core aspect terms along with their corresponding sentiment words, thereby simplifying the original text into a short text in order to eliminate redundant information. Finally, we perform MATE, MASC and JMASA tasks on two datasets. The contributions are summarized as follows:

- We are the first to offer an innovative solution for multi-aspect multi-sentiment scenarios, introducing a multimodal sentiment analysis model based on aspect enhancement and text simplification.
- We design an aspect enhancement module based on a masking mechanism for enhancing the sensitivity of the model to aspect terms. Additionally, we develop a syntax-based text simplification module to facilitate the alignment of multi-aspect with multi-sentiment, thereby reducing the interference of redundant information on aspect-sentiment associations.
- Experiments on two benchmark datasets (Twitter-2015 and Twitter-2017) that perform three tasks show that our model typically outperforms state-of-the-art methods.

Related Work

This section primarily introduces the research pertaining to the MATE, MASC, and JMASA tasks.

Multimodal Aspect Term Extraction (MATE). MATE focuses on extracting opinion-related aspect terms from multimodal data. To combat noise and modality heterogeneity, Wang et al. (2023) suggest the BFCL approach for entity recognition, while Wu et al. (2023) develop the MCG-MNER framework for refined multimodal representations. Liu et al. (2022) use a Bayesian Neural Network for accurate entity identification with visual cues, and Guo et al. (2023) introduce MGICL for MATE with multi-granularity contrastive learning. Wang et al. (2022) propose a MoRe-based framework for enhanced multimodal entity recognition. Despite progress in aspect term extraction, aspect reinforcement has been neglected. Our proposed aspect enhancement module addresses this by increasing the aspect sensitivity of model.

Multimodal Aspect Sentiment Classification (MASC). MASC focuses on recognizing and classifying the sentiment categories expressed within multimodal datasets. Yu and Jiang et al. (2019) investigate ESAFN for entity-level sentiment detection using images in social media. Although these advancements, textual data refinement is often ignored, risking noise sensitivity. Yu and Wang et al. (2022) develop the Image-Target Matching Network (ITM) for multitask learning, addressing relevance, object alignment, and sentiment classification. Yu and Chen et al. (2022) propose the Hierarchical Interactive Multimodal Transformer (HIMT) to account for semantic disparities between text and images. Yang et al. (2022) integrate facial emotions as affective cues in the FITE method for multimodal sentiment analysis. Zhao et al. (2022) create a Knowledge-Enhanced Framework (KEF) leveraging adjective-noun pairs for improved visual attention and sentiment prediction. Xiao et al. (2023) introduce CoolNet, which aligns and fuses modalities using semantic and syntactic text features with visual pixel-level details. We propose a sentiment-focused text simplification module to retain sentiment essence and discard irrelevant details, maintaining expression integrity.

Joint Multimodal Aspect-based Sentiment Analysis (JMASA). JMASA targets the identification of all aspect-sentiment pairs in multimodal data. Ju et al. (2021) introduce a Multimodal Joint Learning (JML) approach for simultaneous MATE and MASC, enhancing cross-modal relation detection. Yang et al. (2022) propose the Cross-modal Multitask Transformer (CMMT) to weigh visual contributions to text aspects. Ling et al. (2022) introduce VLP-MABSA, a unified encoder-decoder model for MABSA. Zhou et al. (2023) create an Aspect-oriented Method (AoM) for aspect-related semantic and sentiment detection. Xiao et al. (2024) develop Atlantis, an Aesthetic-oriented Multiple Granularities Fusion Network for JMASA, with a trident-shaped structure for comprehensive analysis. Zhu et al. (2024) introduce AESAL, a sentiment analysis model that fully captures the syntactic associations between different nodes in the text. Building on these studies in JMASA, we propose a novel multimodal sentiment analysis model with aspect enhancement and text simplification, designed to concurrently address MATE, MASC, and JMASA tasks.

Methodology

The AETS framework, depicted in Figure 2, includes feature extractor, aspect enhancement, text simplification , aspect term extraction, sentiment classification module, and a joint training module. Its key innovations are the aspect enhancement module and text simplification module, which respectively highlight aspect terms and simplify text complexity through syntactic dependencies, facilitating aspect and sentiment association. Each component is detailed in the subsequent sections.

Task Definition

Given a dataset $D = \{(T_i, V_i, A_i, S_i)\}_{i=1}^{K}$ containing K samples, each sample is composed of a text sequence

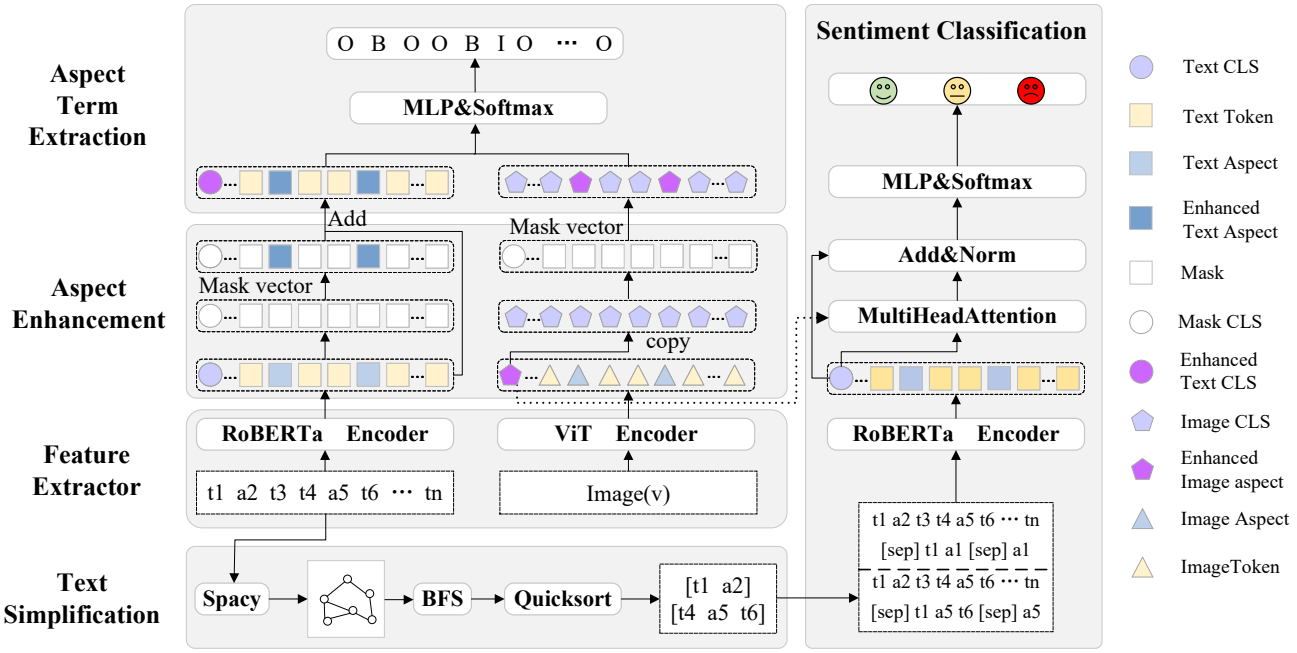

Figure 2: The overall architecture of AETS, mainly composed of an aspect enhancement module and a text simplification module, is designed to heighten the model's sensitivity to aspects and precisely capture the correspondence between aspects and sentiments.

$T = \{t_1, t_2, \ldots, t_n\}$ and a three-dimensional image matrix $V \in R^{3\times H\times W}$. Here, n represents the length of the text, while 3, H, and W respectively denote the number of channels, height, and width of the image. Our objective is to identify all aspect terms $A = \{a_1, a_2, \ldots, a_m\}$ within the text and their corresponding sentiment polarities $S = \{s_1, s_2, \ldots, s_m\}$, m refers to the number of aspect terms in the text sequence T. The MATE, MASC, and JMASA tasks are respectively defined as follows:

$$
\begin{aligned}
MATE: \quad & input = [\{t_1, t_2, \ldots, t_n\}, V]; \\
& output = [a_1, a_2, \ldots, a_m]. \\
MASC: \quad & input = [\{t_1, t_2, \ldots, t_n\}, V, \{a_1, a_2, \ldots, a_m\}]; \\
& output = [s_1, s_2, \ldots, s_m]. \\
JMASA: \quad & input = [\{t_1, t_2, \ldots, t_n\}, V]; \\
& output = [\{a_1, s_1\}, \{a_2, s_2\}, \ldots, \{a_m, s_m\}].
\end{aligned}
$$

We employ sequence labeling techniques to annotate aspect terms, where each element $a_i \in \{B, I, O\}$. Here, B denotes the beginning position of an aspect term, I represents an interior position within an aspect term that is not the starting position, and O signifies a part that is not an aspect term. For sentiment polarity , it is drawn from the set $\{POS, NEU, NEG\}$, where POS, NEU, and NEG respectively represent positive, neutral, and negative.

Feature Extractor

Given the superior performance of RoBERTa (Liu et al. 2019) for text representation and ViT (Dosovitskiy et al. 2020) for visual representation, we utilize RoBERTa and ViT to encode text and images, respectively. For the text sequence, special [CLS] and [SEP] tokens are inserted at the beginning and end, respectively, to denote the start and termination of the text. Correspondingly, for images, a [CLS] token is placed at the front as an identifier for the commencement of image information.

We obtain the text hidden states $H^t = \{h_1^t, h_2^t, ..., h_n^t\}$ and the image hidden states $H^v = \{h_1^v, h_2^v, ..., h_n^v\}$, as represented by the following equation:

$$H^t = RoBERTa(T) \tag{1}$$

$$H^v = ViT(V) \tag{2}$$

where $H^t \in R^{n\times d}, H^v \in R^{n\times d}$, n represents the number of words and d represents the hidden state dimension.

Aspect Enhancement

In the MABSA, models often struggle to effectively distinguish all aspect terms. To address this issue, this chapter introduces an aspect enhancement module that helps enhance the ability of model to learn aspect terms in multimodal data.

Firstly, we leverage the spaCy[1] tool to meticulously segment nouns by their part-of-speech within sentences, thereby constructing a candidate aspect term set $CA = \{ca_1, ca_2, \ldots, ca_k\}$, where each ca_i represents the $i{-}th$

[1]https://spacy.io/

candidate aspect term, having a noun part-of-speech and belonging to the text sequence T. In order to enhance the attention of the model to these candidate aspect terms, we designed a dynamic masking mechanism shown in Equation (3). Specifically, we construct a masking vector $M = \{m_1, m_2, ..., m_n\}$, which masks all positions in the text sequence except those occupied by the candidate aspect term set CA. This results in obtaining a text hidden state H^t_{ca} that contains only the candidate aspects.

$$m_i = \begin{cases} 1, & t_i \in CA \\ 0, & others \end{cases} \tag{3}$$

If an element at position i in the text is in the set of candidate aspect terms, that position is taken to be 1, and the remaining positions are 0. The text hidden state H^t is conditionally filtered using M to form the enhanced candidate aspect text hidden state H^t_{ca} as shown in the following equation:

$$H^t_{ca} = H^t \odot M \tag{4}$$

where $\odot$ denotes a bitwise multiplication that preserves only the parts that are relevant to the candidate aspect. However, in order to maintain contextual integrity and information richness, we superimpose the augmented H^t_{ca} with the original textual hidden state H^t to generate a comprehensive augmented textual representation H^t_f:

$$H^t_f = H^t + H^t_{ca} \tag{5}$$

Secondly, in the multimodal fusion stage, the vector h^v_1 at the [CLS] position in the image hidden state H^v represents the global information of the image, which is extended by copying to a sequence of vectors of the same length as the textual representation $H^v_r = \{h^v_1, h^v_1, \ldots, h^v_1\}$, which is similarly filtered by applying a mask vector M to obtain the image representation H^v_f focusing on the candidate aspects:

$$H^v_f = H^v_r \odot M \tag{6}$$

Further, the augmented text representation H^t_f and the image representation H^v_f are spliced and integrated into a joint feature and decoded by a two-layer linear transformation and an activation function ReLU to get H^a:

$$H^a = ReLU((H^t_f \oplus H^v_f)W_1 + b_1)W_2 + b_2) \tag{7}$$

where $\oplus$ denotes the splicing operation, W_1, W_2 are trainable weight matrices and b_1, b_2 are bias terms.

Finally, by applying the Softmax function to the decoded H^T, we convert it into the form of a probability distribution for predicting the probability of the aspect term P_a:

$$P_a = Softmax(H^a) \tag{8}$$

Text Simplification

In aspect-based sentiment analysis, in order to eliminate redundant information and reduce the influence of other aspectual sentiment words on the sentiment judgment of the target aspect, this module is specially devised as a text simplification method that relies on syntactic dependency structure. Its primary goal is to accurately identify and extract the semantic core directly related to specific aspect terms, and to construct concise phrases for multi-aspect multi-sentiment judgment. The main steps are shown in Figure 3.

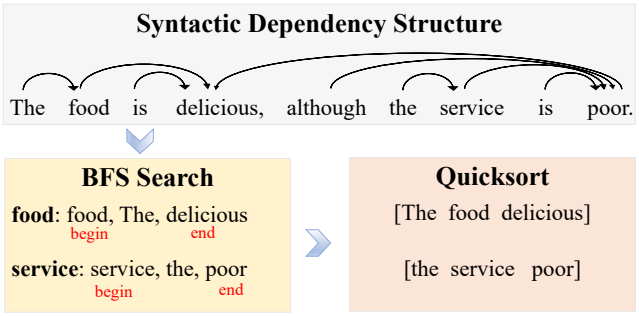

Figure 3: The process of text simplification.

Firstly, the spaCy tool is used to perform deep syntactic analysis of the text and abstractly model syntactic dependencies as an undirected graph $G = (V, E)$, where V is the set of word nodes, E is the set of edges, and the edge e_{ij} indicates that there is a syntactic relationship between node v_i and node v_j.

Algorithm 1: The procedure of text simplification.

Input: inputText T, targetAspect A
Output: simplifiedTexts ST

$G \leftarrow createDependencyGraph(T)$
foreach *aspect in* A **do**
  $node \leftarrow locateNode(G, aspect)$
  if *node* **then**
    queue.enqueue($node$)
    while *queue is not empty and not hasAdjective or hasNextaspect* **do**
      $currentNode \leftarrow queue.dequeue()$
      if $isAdjective(currentNode)$ *or* $isNextAspect(currentNode)$ **then**
        bag.append($currentNode$)
      else
        foreach *neighbor in getNeighbors(G,* $currentNode$*)* **do**
          if $neighbor \notin bag$ **then**
            queue.append($neighbor$)
            bag.append($neighbor$)
return ST

Next, for each target aspect term, we locate its corresponding node in the graph G and initiate a breadth-first search (BFS) from that node. BFS traverses neighboring nodes, prioritizing sentiment words (adjectives) closely related to the aspect. During the search, all connected nodes are added to a temporary vocabulary container until an adjective or another aspect term is found, as shown in Algorithm 1. Adjectives are considered key indicators of affective tendencies, and once found, the search stops. If no adjective

is found, it suggests that affective information is encoded in other lexical properties. When encountering another aspect term, all directly associated terms are assumed to be covered, and continuing the search may introduce noise from other aspects. Thus, the search terminates at this point.

Finally, in order to ensure the semantic coherence of the reconstructed phrases, a quicksort (Hoare 1969) algorithm was used to order the collected words, thus generating a streamlined sentiment phrase S_T that closely follows the target aspect terms and eliminates irrelevant interferences.

$$S_T = quicksort(Bag) \tag{9}$$

Sentiment Classification

In the aspect-based sentiment analysis task, we first extract phrases containing specific aspects and their syntactically related words from the original text in the text simplification phase, and splice the original text (T), the aspect-related phrases (S_T), and the aspect-specific terminology (A) into a new textual input $text = T + S_T + A$ by means of the special symbol [SEP], for example: "$t_1\ a_2\ t_3\ t_4\ a_5\ t_6$" becomes after splicing: "$t_1\ a_2\ t_3\ t_4\ a_5\ t_6\ [SEP]\ t_1\ a_2\ [SEP]\ a_2$" and "$t_1\ a_2\ t_3\ t_4\ a_5\ t_6\ [SEP]\ t_4\ a_5\ t_6\ [SEP]\ a_5$". Next, we select the vectors at the [CLS] position in the text and image hidden states as text and image feature representations, respectively, and fuse the text modal and image modal data using the multi-head attention mechanism to obtain the fused feature H^f.

$$H^f = MultiHeadAttention(h_1^t, h_1^v, h_1^v) \tag{10}$$

In this process, we adopt Transformer (Vaswani et al. 2017) decoder, treating image features as known information and text features as queries (q) that interact with image keys (k) and values (v). To enhance sentiment representation in the fused features, we use residual concatenation and stabilize the feature distribution with layer normalization.

$$H^{nf} = h_1^t + norm(H^f) \tag{11}$$

Next, we decode the above enhanced fusion feature H^{nf} by two fully connected layers, each containing a weight matrix w and a bias term b, and apply the ReLU activation function after the first layer.

$$H^s = ReLU(H^{nf}W_1 + b_1)W_2 + b_2) \tag{12}$$

Finally, we apply the Softmax function to normalize the decoded feature vector H^S to obtain the probability distribution of each sentiment category.

$$P_s = Softmax(H^s) \tag{13}$$

Training

Given that the aspect extraction and the sentiment classification utilize the same coding in the feature extraction phase, we integrate them into a single training framework for collaborative training. During training, the cross-entropy loss is computed independently for the aspect extraction subtask $\mathcal{L}_a$ and the sentiment classification subtask $\mathcal{L}_s$.

$$\mathcal{L}_a = -\sum_{i=1}^{m} y_i^a log(\hat{y_i}^a) \tag{14}$$

$$\mathcal{L}_s = -\sum_{i=1}^{m} y_i^s log(\hat{y_i}^s) \tag{15}$$

The predicted result for the aspect term extraction task is represented by $\hat{y_i}^a$, while the ground truth label for it is y_i^a. Similarly, the prediction outcome for the sentiment classification task is denoted by $\hat{y_i}^s$, with its corresponding actual label being y_i^s. During the training process, we use an alternating training strategy, which means that the two models train the encoder sequentially and share the parameters of the encoder, so as to achieve uniform training of the models.

Experiment

We compare our model with numerous methods on three tasks, including JMASA, MATE and MASC.

Experimental Settings

Datasets. We utilize two publicly available datasets, Twitter-2015 (Yu and Jiang 2019) and Twitter-2017 (Yu and Jiang 2019), which include multimodal (image-text) data with annotations for specific aspects and sentiments. Table 1 shows the distinct characteristics of datasets in aspect and sentiment diversity, with additional details in the appendix.

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1,508	515	493
Neutral	1,883	670	607	1,638	517	573
Negative	368	149	113	416	144	168
# sentence	3,502			2,910		
# one aspect	2,159(61.65%)			976(33.54%)		
# multi-aspect	1,343(38.35%)			1,934(66.46%)		
# multi-sentiment	1,257(35.89%)			1,690(58.08%)		

Table 1: The basic statistics of two Twitter datasets. " # sentence " indicates the number of sentences. # X in the last 3 lines denotes the number of sentences with such characteristics X. "multi" is an abbreviation for "multiple".

Implementation Details. We implement our method under Linux system, CUDA version 10.2, Pytorch version 1.12.0, Python version 3.8, and NVIDIA GeForce RTX 3090. In addition, we set the Learning rate to 2e-5, the dropout to 0.1, hidden sizes to 768.

Evaluation Metrics. We assess AETS on JMASA, MATE and MASC task by Micro-F1 score (F1), Precision (P) and Recall (R), while on MASC task we use Accuracy (Acc) and F1 following previous studies (Zhou et al. 2023).

Baselines

We compare AETS model with four types of methods.

Methods for textual ABSA. 1) **SPAN** (Hu et al. 2019) is an end-to-end span-based framework. 2) **D-GCN** (Chen, Tian, and Song 2020) is a BERT-based graph convolutional

Modality	Methods	Venue	Twitter-2015			Twitter-2017		
			P	R	F1	P	R	F1
Text-based	SPAN*	ACL 2019	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	ARXIV 2019	61.8	65.3	63.5	65.5	66.9	66.2
	BART*	IJCNLP 2021	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	UMT+TomBERT*	ACL 2020, IJCAI 2019	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT*	MM 2020, IJCAI 2019	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse*	MM 2020	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse*	AAAI 2021	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse*	ACL 2020	61.0	60.4	61.6	60.8	60.0	61.7
	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
	Atlantis	Inf. Fusion 2024	65.6	69.2	67.3	68.6	70.3	69.4
	AoM	ACL 2023	67.9	69.3	68.6	68.4	71.0	69.7
	AETS	**Ours**	**69.7**	**74.7**	**72.1**	**72.6**	**73.7**	**73.1**

Table 2: Results of different methods for JMASA on the two Twitter datasets. * denotes the results from (Zhou et al. 2023). The best results are bold-typed and the second best ones are underlined.

network with sequence tagging. 3) **RoBARTa** (Liu et al. 2019) uses a Transformer encoder and CRF for aspect-sentiment pair identification. 4) **BART** (Yan et al. 2021) is a pre-trained model for seven ABSA tasks.

Methods for JMASA. 1) **UMT-collapse** (Yu et al. 2020), **OSCGA-collapse** (Yu, Jiang, and Xia 2019), and **RpBERT-collapse** (Wu et al. 2020b) utilize shared visual cues. 2) **UMT+TomBERT** and **OSCGA+TomBERT** (Yu and Jiang 2019) employ sequential pipelines. 3) **JML** (Ju et al. 2021). 4) **VLP-MABSA** (Ling, Yu, and Xia 2022). 5) **CMMT** (Yang, Na, and Yu 2022). 6) **Atlantis** (Xiao et al. 2024). 7) **AoM** (Zhou et al. 2023).

Methods for MATE. 1) **RAN** (Wu et al. 2020a) is a pioneering method that combines object and text features. 2) **UMT** (Yu, Jiang, and Xia 2019) offers a harmonized architecture to address visual context biases. 3) **OS-CGA** (Yu et al. 2020) is an object-sensitive model that integrates visual and textual information.

Methods for MASC. 1) **ESAFN** (Yu, Jiang, and Xia 2019) employs entity-level analysis. 2) **KEF** (Yu and Jiang 2019) uses image adjective-noun pairs. 3) **FITE** (Yang, Zhao, and Qin 2022) incorporates image sentiment. 4) **HIMT** (Yu, Chen, and Xia 2022) is a multimodal architecture. 5) **ITM** (Yu et al. 2022) is a multi-task learning model. 6) **CoolNet** (Xiao et al. 2024) is a high-performance network with cross-modal alignment.

Main Results

In this segment, we demonstrate how our method outperforms current state-of-the-art approaches.

Performance on JMASA. Table 2 details our JMASA results. **First**, AETS model outperforms text-only models, thanks to its text simplification module. **Second**, multimodal pipelines and adaptive methods lag due to their lack of semantic-sentiment integration. **Last**, AETS outperforms other multimodal approaches, achieving a substantial F1 gain of 3.5% on Twitter-2015 and 3.4% on Twitter-2017. This underscores the benefits of focusing on aspect term recognition and text simplification.

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
AETS(Ours)	**89.4**	**92.5**	**90.9**	**93.3**	**97.3**	**95.3**

Table 3: Results of different methods for MATE. * denotes the results from (Zhou et al. 2023).

Performance on MATE. Table 3 shows AETS excels in the MATE task, surpassing the second-best AoM model on all Twitter-2015 metrics: P up by 4.8%, R by 4.6%, and F1 by 4.7%. On Twitter-2017, it outperforms CMMT with improvements of 1.1% in P, 3.4% in R, and 2.2% in F1, demonstrating high sensitivity of AETS in detecting aspect terms.

Performance on MASC. As shown in Table 4, the AETS model achieves the optimal overall performance. Compared to the suboptimal AoM model, AETS exhibits a slight decrease of 0.7% in Acc on the Twitter-2015 dataset, while it reaches the highest levels in all other metrics. This may be due to the relatively small number of sentences with multiple-aspect and multiple-sentiment in the Twitter-2015 dataset (all below 40%), which may led to the model's capabilities not being fully utilized.

Methods	Venue	Twitter-2015		Twitter-2017	
		Acc	F1	Acc	F1
ESAFN*	TASLP 2020	73.4	67.4	67.8	64.2
TomBERT*	IJCAI 2019	77.2	71.8	70.5	68.0
CapTrBERT*	MM 2021	78.0	73.2	72.3	70.2
JML	EMNLP 2021	78.7	-	72.7	-
VLP-MABSA	ACL 2022	78.6	73.8	73.8	71.8
CMMT	IPM 2022	77.9	-	73.8	-
KEF	COLING 2022	78.2	73.5	71.9	68.7
FITE	ACL 2022	78.5	73.9	70.9	68.7
HIMT	TAFFC 2023	78.1	73.7	71.1	69.2
ITM	IJCAI 2022	78.3	74.2	72.6	72.0
CooINet	IPM 2023	79.9	75.3	71.6	69.6
AoM	ACL 2023	**80.2**	75.9	76.4	75.0
AETS	Ours	79.5	**80.8**	**76.6**	**75.2**

Table 4: Results of different methods for MASC. * denotes the results from (Zhou et al. 2023).

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
Full	**69.7**	**74.7**	**72.1**	**72.6**	**73.7**	**73.1**
w/o Img	66.5	74.7	70.4	70.7	70.7	70.7
w/o Att	67.3	69.5	68.4	69.3	71.1	70.2
w/o CAT	65.6	64.3	64.9	62.0	67.3	64.5
w/o CAI	69.8	72.7	71.2	67.3	70.3	68.8
w/o Simp	66.3	74.7	70.2	63.7	69.9	66.7

Table 5: The performance comparison of our full model and its ablated methods.

Ablation Study

In this section, we investigate the impact of each AETS component on JMASA, with results presented in Table 5.

W/o Img is an AETS variant that solely relies on text data, omitting image information. Table 5 reveals that this variant underperforms across metrics compared to the Full method, underscoring the significance of image data.

W/o Att does not incorporate the multi-head attention mechanism, leading to a substantial F1 score decrease of 3.7% on Twitter-2015 and 2.9% on Twitter-2017. This highlights the importance of multi-head attention for effective information fusion in MABSA.

W/o CAT removes text-based candidate aspect attention from the aspect enhancement module, resulting in a notable performance decrease, especially noticeable on the Twitter-2017 dataset with its numerous aspect terms, indicating a reduced aspect sensitivity.

W/o CAI signifies the aspect enhancement module's neglect of image-based candidate aspects, leading to a widespread performance decline, emphasizing the value of image aspect candidates for optimizing the model.

W/o Simp represents the variant without text simplification. The absence of simplification significantly impacts performance, with P decreasing by 3.4% on Twitter-2015 and 8.9% on Twitter-2017, highlighting the importance of sentence simplification for aspect-term relationships.

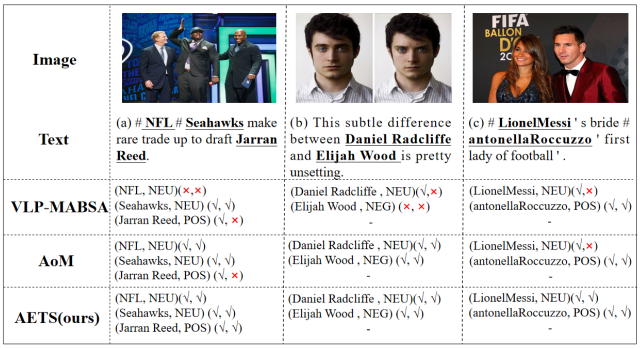

Figure 4: Predictions of different methods on three test samples. NEU, POS, and NEG denote neutral, positive, and negative sentiments, respectively.

Case Study

To further demonstrate the effectiveness of AETS, we present three test examples with predictions from different methods. As shown in Figure 4, for example (a), VLP-MABSA incorrectly predicts the aspect "*NFL*" and its associated sentiment. Moreover, while it correctly identifies "*Jarran Reed*" as an aspect, it misjudges its sentiment. Concurrently, the AoM model also makes an error in predicting the sentiment of the aspect term "*Jarran Reed*". However, AETS successfully detects three aspects and accurately predicts the sentiment corresponding to each one. In scenario (b), the VLP-MABSA model performs poorly, correctly predicting only one aspect, "*Daniel Radcliffe*", and incorrectly predicting the sentiment of aspect "*Daniel Radcliffe*". At the same time, the aspect "*Elijah Wood*" and its sentiment were both incorrectly predicted. In contrast, both AoM and our AETS model perform well, effectively detecting all aspects and sentiments. In case (c), both VLP-MABSA and AoM inaccurately predict the sentiment of "*LionelMessi*". But our AETS model succeeds in forecasting its correct sentiment. Across these three cases, our AETS model consistently demonstrates superior performance, adeptly capturing all aspect terms and precisely forecasting the sentiment orientation for each one. This outstanding performance is attributed to the meticulously designed aspect-enhancement module and the efficient text simplification module incorporated within our model.

Conclusion

In this paper, we introduce a MABSA method that integrates aspect enhancement and text simplification. Our method includes a dedicated aspect enhancement module to enhance the ability of model to capture all aspect information from both textual and visual data. Moreover, we have developed a text simplification module that simplifies the text content based on syntactic dependencies to eliminate interference from redundant information, achieving precise judgments in multi-aspect, multi-sentiment scenarios. This methodology is applied to MATE, MASC, and JMASA tasks. Experiments conducted on two benchmark datasets demonstrate the effectiveness of AETS.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 62472348, in part by the Aviation Science Foundation under Grant 2023M071070002, in part by the Key Research and Development Program of Shaanxi under Grants 2022GY-332, 2023-YBGY-230, and 2024GX-YBXM-533, in part by the Innovation Capability Support Plan of Shaanxi under Grant 2022PT-33, in part by the Xi'an Science and Technology Plan for the Key Industrial Chain Technology Research Project under Grant 23ZDCYJSGG0007, in part by the Xi'an Science and Technology Plan for the Key Industrial Chain, Key Core Technology Research Project under Grant 23LLRH0022, in part by the Qinchuangyuan Construction of Two Chain Integration Important Project under Grant 23LLRHZDZX0006, and in part by the Xianyang City Major Science and Technology Innovation Special Project under Grant L2024-ZDKJ-ZDCGZH-0012. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Chen, G.; Tian, Y.; and Song, Y. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, 272–279. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; and 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.

Guo, A.; Zhao, X.; Tan, Z.; and Xiao, W. 2023. MGICL: Multi-Grained Interaction Contrastive Learning for Multimodal Named Entity Recognition. 639–648.

Hoare, C. 1969. Algorithm 63, partition; algorithm 64, quicksort. *algorithm*, 65: 321–322.

Hu, M.; Peng, Y.; Huang, Z.; Li, D.; and Lv, Y. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 537–546. Florence, Italy: Association for Computational Linguistics.

Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; and 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*, 4395–4405. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kurant, M.; Markopoulou, A.; and Thiran, P. 2010. On the bias of BFS (Breadth First Search). In *2010 22nd International Teletraffic Congress (lTC 22)*, 1–8.

Ling, Y.; Yu, J.; and 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 (Volume 1: Long Papers)*, 2149–2159. Dublin, Ireland: Association for Computational Linguistics.

Liu, L.; Wang, M.; Zhang, M.; Qing, L.; and He, X. 2022. UAMNer: uncertainty-aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 52: 1–17.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Chen, X.; Shang, Z.; and KE, W. 2023. Multimodal Named Entity Recognition with Bottleneck Fusion and Contrastive Learning. *IEICE Transactions on Information and Systems*, E106.D: 545–555.

Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; and Lu, W. 2022. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5925–5936. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wu, H.; Cheng, S.; Wang, J.; Li, S.; and Chi, L. 2020a. *Multimodal Aspect Extraction with Region-Aware Alignment Network*, 145–156. ISBN 978-3-030-60449-3.

Wu, J.; Gong, C.; Cao, Z.; and Fu, G. 2023. MCG-MNER: A Multi-Granularity Cross-Modality Generative Framework for Multimodal NER with Instruction. 3209–3218.

Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.-f.; and Li, Q. 2020b. 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*, 1038–1046.

Xiao, L.; Wu, X.; Xu, J.; Li, W.; Jin, C.; and He, L. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, 106: 102304.

Xiao, L.; Wu, X.; Yang, S.; Xu, J.; Zhou, J.; and He, L. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6): 103508.

Yan, H.; Dai, J.; Ji, T.; Qiu, X.; and Zhang, Z. 2021. A Unified Generative Framework for Aspect-based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2416–2429. Online: Association for Computational Linguistics.

Yang, H.; Zhao, Y.; and Qin, B. 2022. Face-Sensitive Image-to-Emotional-Text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3324–3335. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yang, L.; Na, J.-C.; and Yu, J. 2022. Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis. *Inf. Process. Manage.*, 59(5).

Yu, J.; Chen, K.; and Xia, R. 2022. Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, PP: 1–1.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Yu, J.; Jiang, J.; and Xia, R. 2019. Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28: 429–439.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3342–3352. Online: Association for Computational Linguistics.

Yu, J.; Wang, J.; Xia, R.; and Li, J. 2022. Targeted Multimodal Sentiment Classification based on Coarse-to-Fine Grained Image-Target Matching. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 4482–4488. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhao, F.; Wu, Z.; Long, S.; Dai, X.; Huang, S.; and Chen, J. 2022. Learning from Adjective-Noun Pairs: A Knowledge-enhanced Framework for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Zhou, R.; Guo, W.; Liu, X.; Yu, S.; Zhang, Y.; and Yuan, X. 2023. AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8184–8196. Toronto, Canada: Association for Computational Linguistics.

Zhu, L.; Sun, H.; Gao, Q.; Yi, T.; and He, L. 2024. Joint Multimodal Aspect Sentiment Analysis with Aspect Enhancement and Syntactic Adaptive Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6678–6686. International Joint Conferences on Artificial Intelligence Organization. Main Track.