International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

Bidirectional Complementary Correlation-Based Multimodal Aspect-Level Sentiment Analysis

Jing Yang, Shanghai University of Engineering and Technology, China
Yujie Xiong, Shanghai University of Engineering and Technology, China*

ABSTRACT

Aspect-based sentiment analysis is the key to natural language processing, and it focuses on the polarity of emotions associated with specific text aspects. Traditional models that combine text and visual data tend to ignore the deeper interconnections between patterns. To solve this problem, the authors propose a multimodal sentiment-oriented analysis (BiCCM-ABSA) model based on bidirectional complementary correlation. The model utilizes text-image synergy through a novel cross-modal attention mechanism to align text with image features. With the transformer architecture, it is not only a simple fusion, but also ensures the complex alignment of multi-modal features and gating mechanisms. Experiments were conducted on the Twitter-15 and Twitter-17 datasets, achieving 69.28 accuracy and 67.54% F1 score, respectively. The experimental results demonstrate the advantages of BiCCM-ABSA, the bidirectional approach of the model and the effective cross-modal correlation set a new benchmark in the field of multimodal emotion recognition, providing insights beyond traditional single-modal analysis.

KEYWORDS

aspect-based sentiment analysis, attention mechanism, Index Terms-multimodal, multimodal fusion, social media

INTRODUCTION

In recent years, sentiment analysis has emerged as one of the most vibrant research areas in natural language processing. Its primary focus lies in analyzing people's emotional tendencies toward specific topics and events (Su et al., 2023; Yen et al., 2021). Aspect-level sentiment classification, a fundamental task in sentiment analysis, aims to discern the emotional polarity of different aspects within a text (Singh & Sachan, 2021). For example, in the sentence "Congratulations to Sean Harris, who wins the leading actor award," two aspects are mentioned: Sean Harris and the leading actor award. Based on the context, it can be inferred that the sentiment toward Sean Harris is positive, while it remains neutral toward the leading actor award.

However, texts on social media, often containing opinions on various subjects, pose a challenge in determining the sentiment polarity of multiple aspects from a single sentence (Tobaili et al., 2019;

DOI: 10.4018/IJSWIS.337598 *Corresponding Author

Sahoo & Gupta, 2021). Zhou et al. (2019) noted that lots of errors in sentiment classification arise from not considering the aspect words in sentences. To address this issue, Tang et al. (2016) introduced an attention mechanism to capture the semantic relationship between aspect words and their context, which aligns with the findings of Ismail et al. (2022). Recently, there has been research on aspect-level sentiment analysis based on pretrained language models (Song et al., 2019; Mohammed et al., 2022; Zhang et al., 2023). However, these approaches tend to overlook the integration of textual data with other modal data, which is increasingly relevant in today's social media landscape (Al-Qerem et al., 2020).

As the combination of textual descriptions with corresponding images has become the predominant way for users to express their views on social media platforms, multimodal aspect-based sentiment analysis (MABSA) has emerged as a new trend (Ren et al., 2021; Al-Ayyoub et al., 2018). In literature (Ling et al., 2022), MABSA is also referred to as target-oriented multimodal sentiment analysis or entity-based multimodal sentiment analysis. This task encompasses three subtasks: multimodal aspect term extraction, multimodal aspect sentiment classification, and multimodal aspect sentiment joint extraction. Specifically, multimodal aspect term extraction is to identify aspect terms in a text that are also linked to the visual content. Multimodal aspect sentiment classification is to classify the sentiment polarity of each identified aspect considering both textual and visual contexts. Multimodal aspect sentiment joint extraction is to simultaneously perform aspect term extraction and sentiment classification in a unified framework (Barbosa et al., 2022; Salhi et al., 2021). But it struggles with the complexity of effectively integrating and interpreting textual and visual data, and faces challenges in accurately capturing the nuanced semantic relationships between these modalities. Xu et al. (2019) introduced a multi-interaction memory network for multimodal target sentiment classification. Similarly, TomBERT (J. Yu & Jiang, 2019), building upon the bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2018), incorporated a target-sensitive visual attention mechanism. However, these methods have performed a simple fusion of data from different modalities, without delving deeply into the intrinsic correlations between these modalities. They do not adequately explore the deep, intrinsic correlations between textual and visual data. This superficial integration limits the depth and accuracy of sentiment analysis, failing to fully leverage the rich, nuanced interplay between text and image content that is characteristic of modern social media communication.

In the evolving landscape of natural language processing, aspect-based sentiment analysis (ABSA) has emerged as a crucial tool for understanding nuanced sentiments in textual data. Traditional ABSA models, however, often fall short in comprehensively analyzing the increasingly multimodal nature of social media content, where text and visual elements interplay to convey complex sentiments. In this paper, we introduce the BiCCM model, a groundbreaking bidirectional complementary correlation based multimodal target sentiment analysis model, redefining the standards in sentiment analysis. The cornerstone of the model is its advanced attention mechanism, uniquely engineered to pinpoint and extract emotional cues from images that directly correlate with aspect terms in the accompanying text. This pioneering approach not only achieves seamless and intuitive integration of textual and visual data but also pioneers a symbiotic interplay between these two modalities. The culmination of this innovation is the sophisticated amalgamation of image-responsive text analysis and text-responsive image processing, a breakthrough accomplishment that sets BiCCM apart from traditional sentiment analysis models. By transcending the limitations of conventional methodologies, the model paves the way for a more nuanced and contextually aware multimodal sentiment analysis.

The BiCCM-ABSA model distinguishes itself through its advanced integration of text and images, utilizing an innovative application of the transformer architecture. This approach allows for a more nuanced and effective combination of multimodal data, leveraging the intricate interplay between textual and visual elements. Key contributions of this research include:

The development of a novel bidirectional complementary correlation technique, enhancing the semantic analysis of text-image pairs.

Implementation of a customized transformer architecture, optimized for efficient and accurate multimodal sentiment analysis.

Demonstrated superiority of BiCCM-ABSA in handling complex, real-world social media data, evidenced by rigorous testing on diverse data sets.

Revolutionary BiCCM model: We introduce a cutting edge model that adeptly identifies and leverages the intricate relationships between aspect words and their corresponding visual representations, as well as the interplay between text and images. This model does not merely analyze; it discerns, producing feature vectors rich in semantic depth and fusing multimodal data to significantly elevate the precision of aspect level sentiment analysis.

Innovative cross-modal guided feature extraction: We have devised a pioneering method for feature extraction that establishes bidirectional links and complementary insights between textual and visual data. This methodology lays a solid foundation for the multimodal fusion process, marking a significant stride in the field.

Empirical validation with Twitter-15 and Twitter-17: Our comprehensive experiments on the widely recognized Twitter15 and Twitter-17 data sets underscore the superiority of the BiCCM model. It not only achieves but exceeds existing benchmarks in terms of accuracy and F1 score, emphatically validating the model's efficacy.

In conclusion, the BiCCM model is more than an incremental advancement; it represents a paradigm shift in how we approach sentiment analysis, especially in the context of social media data. By intricately weaving together the textual and visual threads, it offers a deeper understanding of sentiments, a development in the landscape of natural language processing.

RELATED WORK

Aspect-Level Sentiment Analysis in Text

Current research in text-based aspect-level sentiment analysis primarily focuses on extracting hidden semantics using deep neural networks. Recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks and gated recurrent units (GRUs) are used to encode sentences, capturing more implicit semantic information and making the encoding more effective. Murthy et al. (2020) employed LSTM to encode text from both sides of an aspect term, analyzing sentiment based on the context related to the aspect term. Lou et al. (2020) proposed a specific target sentiment analysis method based on a multi-attention convolutional neural network. Li et al. (2021) introduced a dual-channel graph convolutional network that considers both syntactic structure and semantic associations, enhancing semantic representation through regularization. Qi et al. (2022) utilized a bidirectional graph convolutional network to integrate the syntactic dependency relationships between aspect words and context, combining syntactic-semantic features with textual context features.

Multimodal Sentiment Analysis

The goal of multimodal sentiment analysis is to classify the overall sentiment of a statement using both visual and linguistic elements. Compared to unimodal sentiment analysis using just text or images, multimodal sentiment analysis requires considering the semantic correlations between different modalities. Huang et al. (2019) proposed a deep multimodal attention fusion model, where two independent attention mechanisms based on textual and image information respectively perform sentiment analysis. The final sentiment analysis result is obtained by combining the outcomes of sentiment analysis from these joint text-image features. F. Chen et al. (2019) distinguished features of each modality into invariant and specific modalities. They captured the common features between invariant modalities, while specific modalities provided unique information from each data modality to supplement the common features. W. Yu et al. (2021) annotated the sentiment polarity for each individual modality based on a multimodal unified label, and conducted multi-task learning by

combining multimodal and unimodal approaches. Mao et al. (2022) released a multimodal sentiment analysis platform, M-SENA, which integrates data management, feature extraction, model training, and result analysis. Y. Wu et al. (2022) addressed the issue of errors in emotion word extraction caused by automatic speech recognition. They proposed a sentiment word-aware multimodal sentiment analysis model, SWRM, which detects the position of sentiment words in the text and reconstructs the textual emotional semantics by generating candidate sentiment words.

Multimodal Aspect-Level Sentiment Analysis

Multimodal aspect-level sentiment analysis is a novel field combining aspect-level sentiment analysis with multimodal sentiment analysis. The goal of this task is to analyze the sentiment polarity of aspect words mentioned in sentences by combining visual and linguistic elements. H. Wu et al. (2020) introduced a region-aware alignment network that aligns aspect words extracted from text with corresponding visual regions in images. This is achieved through an adaptive attention network learning shared information to extract aspect terms. Khan and Fu (2021) used a textual transformer to translate images into text in multimodal data, creating auxiliary sentences. This approach provides additional textual information for aiding in model training. Ju et al. (2021) designed an auxiliary cross-modal relation detection module to determine the relevance between images and text. This method retains only those visual regions mentioned in the text content as valid visual information. Ling et al. (2022) concatenated textual and image representations as multimodal inputs and proposed a pretrained framework suitable for multimodal aspect-level sentiment analysis tasks. Zhang et al. (2023) proposed a multimodal, multitask interactive graph attention network, termed M3GAT, to simultaneously solve the three problems; the model is a proposed interactive conversation graph layer containing local-global context connection, cross-modal connection, and cross-task connection.

Our study advances beyond prior research that predominantly focused on the unidirectional influence of aspect words on either textual or visual elements. We acknowledge that, in practical applications, the interplay between text and images transcends mere aspect words. Our innovative model, BiCCM-ABSA, establishes a groundbreaking approach by integrating scene context from images to augment the understanding of textual features. This leads to a robust bidirectional semantic connection between textual and visual domains. Our methodology not only utilizes unimodal features derived from aspect words but also synergizes them with image-driven insights to enhance textual analysis. By implementing this strategy, we aim to markedly improve the precision of aspect-level sentiment analysis. This approach allows for a more intricate and context-aware evaluation than previously achievable. This development represents a significant stride in sentiment analysis, particularly in addressing the complexities of multimodal data. Our work aligns with recent advancements in the field, as seen in [reference recent studies], and contributes to the evolving narrative of how textual and visual data interconnect in sentiment analysis.

Model Description

In this section, the task of multimodal target sentiment analysis is defined, and the overall framework of the proposed bidirectional complementary correlation-based multimodal target sentiment analysis model (BiCCM) is introduced.

Task Definition

The multimodal target sentiment analysis task in this paper is defined as a classification task. For a given sample, the task is to determine the sentiment polarity (positive, negative, or neutral) of the aspect word(s) in the sample. Each sample D consists of a context S, aspect term T, and an image I. Here, S denotes the input sequence of the context excluding the aspect term, T is the aspect term sequence, and I is the associated image for the sample. For the training set, there is also an emotion label associated with the aspect term. As shown in Table 1, each aspect term has a corresponding sentiment label, with positions in the context S indicating the aspect term. The goal of multimodal

Table 1. Examples of model text input

Aspect Term T	Context S	Sentiment Label
Brent Seabrook	#celebrates his goal, but the good times didn't last. #Blackhawks fall	Positive
#Blackhawks	Brent Seabrook celebrates his goal, but the good times didn't last. #fall	Negative

target sentiment analysis is to learn a mapping function that can map sample D to the correct aspect term sentiment label.

Overall Framework

The BiCCM model, a cornerstone of this paper, represents a significant leap in sentiment analysis, distinguished by its three innovative components: the feature extraction module, the multimodal fusion module, and the classification layer. Each component is meticulously designed to synergize the processing of multimodal data:

Advanced feature extraction module: At the forefront of the model, this module adeptly extracts nuanced text features, encapsulating rich contextual semantic information and visual features, which precisely pinpoint aspect term details. This dual extraction from both text and images is a testament to the model's ability to delve into the complexities of multimodal data.

Sophisticated multimodal fusion module: The crux of the BiCCM model, this module, is an embodiment of innovation. It seamlessly integrates text and visual features through a multi-layered structure, comprising four submodules: image features guide text features (IGT), text features guide image features (TGI), multimodal fusion based on text features (MFBT), and a visual gating mechanism. Each submodule plays a pivotal role, orchestrating a harmonious fusion of modalities that surpasses traditional unimodal approaches.

Efficient classification layer: In the final stage, the outputs from the MFBT and the visual gate, representing a rich amalgamation of multimodal insights, are skillfully concatenated. This concatenated output is then processed through a softmax layer, ensuring accurate sentiment classification. This strategic layering not only exemplifies the model's comprehensive analytical ability but also its effectiveness in distilling complex multimodal data into actionable insights.

In essence, the BiCCM model, with its innovative architecture and integration of cutting-edge techniques, stands as a pioneering approach in the realm of sentiment analysis. It not only addresses the intricacies of multimodal data processing but also sets a new benchmark in the accurate classification of sentiments, showcasing the immense potential of this model in transforming the landscape of natural language processing and beyond.

Feature Extraction Module

1) Text feature extraction: In this study, BERT is used as the text encoder, capable of encoding words differently based on the context, utilizing pretrained model parameters from a large corpus provided by BERT.

The textual part of the dataset includes context *S* and aspect term *T*. For text encoding, the context *S* and aspect term *T* are concatenated with special tokens [CLS] and [SEP], forming the input sequence for BERT. The BiCCM model's text input is [*CLS*]*S*[*SEP*]*T*[*SEP*].

The input sequence for BERT is [*CLS*]*S*[*SEP*]*T*[*SEP*], where [*CLS*] is the inserted starting token and [*SEP*] is the separator token. The encoded output sequence is $\mathbf{H} = [h_0, h_1, ..., h_N]$, where each hi is a representation comprising word embedding, segment embedding, and position embed- ding, and

N is the maximum text length of the model input. Upon obtaining the text feature representation, an additional self-attention layer (Devlin et al., 2018) is added, as shown in Equation 1, to capture the correlation between context S and aspect term *T*, obtaining a hidden representation of the text, where hi represents the hidden representation of each word.

$$Att\left(X\right) = softmax\left(\frac{\left[W_Q X\right]^* \left[W_K X\right]}{\sqrt{d\ /\ m}}\right)\left[W_V X\right] \tag{1}$$

Here, W_Q, W_K, and W_V are the learnable parameters corresponding to Query, Key, and Value, respectively.

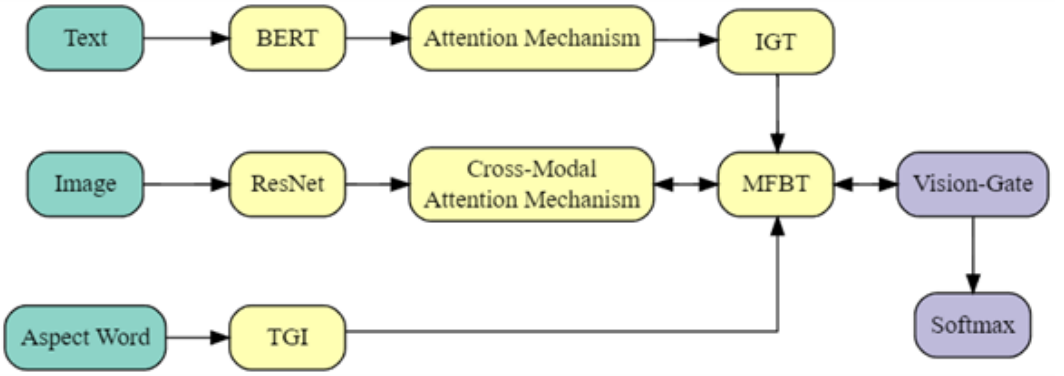
2) Image feature extraction: Considering the varying sizes of images in each sample, the image size is uniformly adjusted to 224x224 pixels. The output of the last convolutional layer of the Resnet-152 model is used as the deep feature representation of the image. In practice, each input image is divided into 7x7 equally-sized visual blocks, with each block represented by a 2048-dimensional vector. The image features obtained from Resnet-152 can be represented as $\mathbf{U} = [u_1, u_2, ..., u_{49}]$, where u_j is the 2048-dimensional vector representation of the j th visual block.

As the task of this study is to predict the sentiment polarity based on aspect terms, a cross-modal attention mechanism is added to capture region features in the image related to the aspect term. First, the feature representation of the aspect term is obtained using the BERT encoder, where *M* is the maximum length of the aspect term input to the model. The aspect term's feature vector *A* is used as the Query, and the image features *U* as Key and Value in the attention network. The calculation, as shown in Equation 2, obtains aspect term-sensitive image features *TI*.

$$Att\left(A,U\right) = softmax\left(\frac{\left[W_Q^{'} A\right]^* \left[W_K^{'} U\right]}{\sqrt{d\ /\ m}}\right)\left[W_V^{'} U\right]^* \tag{2}$$

Figure 1. The overview of BiCCM

BiCCM consists of an extraction module, a multi-modal fusion module (rounded rectangle as shown in the figure), and a classification module.



Where *W*, *W*, and *W* are the learnable parameters corresponding to Query, Key, and Value, respectively. Our decision to employ the transformer architecture in our model is rooted in its established efficacy for managing sequential data and its versatility across diverse natural language processing (NLP) tasks. The architecture's competence in processing extensive data sets, coupled with its sophisticated attention mechanism, positions it as particularly well-suited for discerning the subtleties inherent in sentiment analysis. This feature enables our model to selectively concentrate on pertinent aspects within the textual data while fluidly adjusting to varying contexts. Such dynamic adaptability is crucial for performing precise sentiment analysis within complex multimodal data sets. By leveraging the transformer architecture's strengths, our model harnesses these capabilities to deliver enhanced performance in sentiment analysis, reflecting the latest advancements in NLP methodologies.

Multimodal Feature Fusion Module

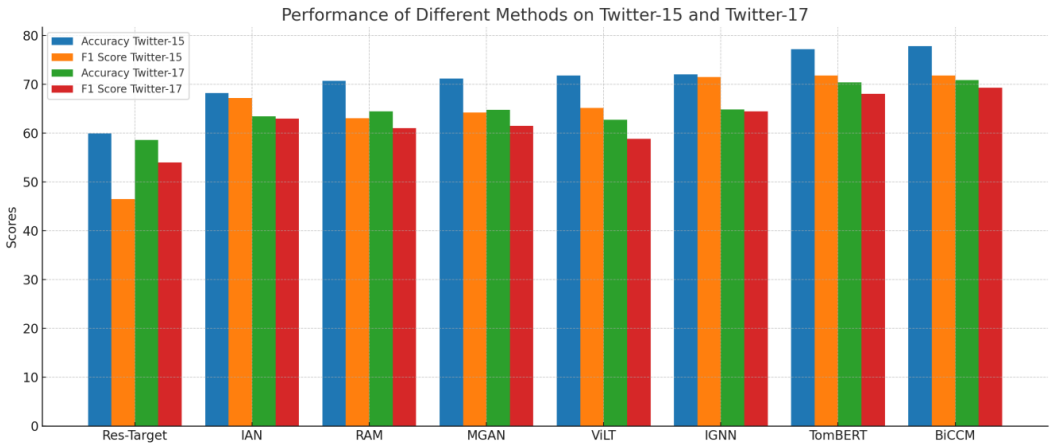
The multimodal feature fusion module consists of four submodules: image features guide text features (IGT), text features guide image features (TGI), multimodal fusion based on text features (MFBT), and the visual gate. The structure of the multimodal fusion module is illustrated in Figure 2. In the IGT module, the aspect term-sensitive image feature (TI) guides the generation of text features enriched with image information.

Then, feature fusion is performed using a transformer, aligning multimodal features with image features through a gating mechanism.

1) Image Features Guide Text Features (IGT) Module:

To explore the bidirectional relevance and complementarity between text and images, the image feature sensitive to aspect terms guides the learning of text representations. A multi-head attention mechanism (Murthy et al., 2020) is used, which increases the weight coefficients of sentiment words in the text based on image regions related to the aspect term's sentiment. The aspect term- sensitive image feature *TI* is used as the Query, and the hidden text representation R as the Key and Value in the attention network. The calculation, as shown in Equation 3, is used to obtain image-sensitive text features.

Figure 2. Experimental results of model ablation



$$Att_i\left(TI,R\right)=softmax\left(\frac{\left[W_{q_iTI}\right]^*\left[W_{k_iR}\right]}{\sqrt{d/m}}\right)\left[W_{v_i}R\right]^* \tag{3}$$

Here, *W*, *W*, and *W* are the learnable parameters corresponding to Query, Key, and Value, respectively. The outputs of m attention mechanisms are concatenated, followed by a linear transformation as shown in Equation 4.

$$MATT\left(TI,R\right)=W_m\left[Att_1\left(TI,R\right),\ldots,Att_m\left(TI,R\right)\right] \tag{4}$$

Where W'' is the learnable parameter. Two layers of residual networks and layer normalization are stacked, followed by a feedforward network, as shown in Equation 5.

$$IT=LN\left(TI+MLP\left(LN\left(TI+MATT\left(TI,R\right)\right)\right)\right) \tag{5}$$

2) Text Features Guide Image Features (TGI) Module:

Similar to the aspect term-image feature guide text module, the hidden text representation *R* is used as the Query, and the aspect term-sensitive image feature *TI* as the Key and Value in a multi-head attention network to obtain text-sensitive image features *RI*.

3) Multimodal Fusion Based on Text Features (MFBT) Module:

As the aspect term-sensitive image feature *TI* was used as the Query in the aspect term-image feature guide text module, generating vectors for each visual block, these visual block information are fused with text information. The hidden text representation *R* is used as the Query, and the image-sensitive text feature *IT* as Key and Value to generate image-perceptive text features *O*. The calculation is shown in Equation 6.

$$Att_i\left(R,TI\right)=softmax\left(\frac{\left[W_{q_i}^{'}R\right]^*\left[W_{k_j}^{'}TI\right]}{\sqrt{d/m}}\right)\left[W_{v_i}^{'}TI\right]^* \tag{6}$$

Where *W*, *W*, and *W* are the learnable parameters corresponding to Query, Key, and Value, respectively. The outputs of *m* attention mechanisms are concatenated, followed by a linear transformation as shown in Equation 7.

$$MATT\left(R,TI\right)=W_m^{'}\left[Att_1\left(R,TI\right),\ldots,Att_m\left(R,TI\right)\right] \tag{7}$$

Again, two layers of residual networks and layer normalization are stacked, followed by a feedforward network, as shown in Equation 8.

$$MATT\left(R,TI\right)=W_{m}^{'}\left[Att_{1}\left(R,TI\right),\ldots,Att_{m}\left(R,TI\right)\right] \tag{8}$$

4) Visual Gating Mechanism:

Considering that some function and structural words in the text, such as "the," "for," and "to," cannot be aligned with visual blocks, a visual gate is used to adjust the final output of the text-sensitive image feature *RI*. The image-perceptive text feature *O* and text-sensitive image feature *RI* are combined according to Equation 9 to obtain the text-perceptive image feature.

$$g=\sigma\left(W_{a}O+W_{q}RI\right) \tag{9}$$

The image-perceptive text feature *O* and the text-perceptive image feature *P* are concatenated to obtain the final multi- modal representation *Q*. The motivation for implementing the cross-modal attention mechanism in the BiCCM-ABSA model is to bridge the gap in traditional sentiment analysis that often ignores the synergy between text and images. This mechanism is crucial for capturing the full spectrum of sentiments in multimodal data, ensuring that the model accurately reflects the complex interplay of visual and textual cues present in social media content. This mechanism adeptly synthesizes data from both modalities, ensuring that the sentiment analysis is comprehensive and contextually relevant. It excels in identifying key aspects within text and correlating them with corresponding visual elements, thereby providing a holistic sentiment analysis. This synthesis is particularly crucial in processing complex multimodal data, where traditional models may overlook the intricate relationship between text and images.

Classification Layer

The multimodal representation *Q* is input into a fully connected layer, followed by a softmax layer to obtain the probability distribution of the aspect term's sentiment label, as shown in Equation 10.

$$p\left(y\left|Q\right.\right)=softmax\left(W^{*}Q\right) \tag{10}$$

Where *W* is the learnable parameter. Cross-entropy is used as the loss function, as shown in Equation 11.

$$L=-\frac{1}{\left|D\right|}\sum_{j=1}^{|P|}\log p\left(y^{(j)}\left|Q^{(j)}\right.\right) \tag{11}$$

EXPERIMENTS AND RESULT ANALYSIS

Data Set

For performance analysis, this study employs the Twitter-15 and Twitter-17 data sets, which are publicly available and rich in multimodal content. These data sets, curated by Zhang et al. (2020) and Lu et al. (2021), encompass a diverse range of multimodal tweets from 2014-2017. They are particularly chosen for their comprehensive annotations, which include aspect terms and corresponding sentiment labels (positive, negative, or neutral), as further enriched by J. Yu et al. (2019). The inclusion

of both tweet text and accompanying images in these data sets makes them ideal for evaluating the effectiveness of our BiCCM-ABSA model in real-world, multimodal scenarios.

To assess the performance of our model, we utilize accuracy and F1 score as primary metrics. Accuracy is chosen for its straightforward representation of overall model performance, indicating the proportion of correct predictions. The F1 score, a harmonic mean of precision and recall, is particularly relevant for our study as it balances the trade-off between the model's ability to correctly identify sentiments (precision) and its capacity to cover the full range of relevant sentiments in the data set (recall). This combination of metrics provides a comprehensive evaluation of our model's performance in real-world, aspect-based sentiment analysis.

Experimental Setup

Most parameters in the experiments follow the configurations of the BERT pretrained model. For text modality, a BERT-based pretrained model is used for initializing text features. The maximum length for text input is set to 128 and, for aspect term input, it is 16. For image modality, images are resized to 224x224, and pre-trained ResNet-152 is used to initialize image features. Each image is divided into 7x7 equally sized visual blocks, each represented by a 2048-dimensional vector, resulting in a 7x7x2048-dimensional image feature vector. Table 3 presents the important parameter settings used in the experiments.

Table 2. Statistics of data sets

Twitter-15					
	Negative	Neutral	Positive	Total	Average Length
Training Set	368	1883	928	3179	16.7
Validation Set:	149	679	303	1122	16.7
Test Set	113	607	317	1037	17
Twitter-17					
	Negative	Neutral	Positive	Total	Average Length
Training Set	416	1638	1508	3562	16.2
Validation Set	144	517	515	1176	16.4
Test Set	168	573	493	1234	16.4

Table 3. Model paraments

Modality	**Parameter**	**Parameter Value**
Text	Maximum Length of Text	128
	Maximum Length of Aspect Term	16
Image	Input Image Size	224*224
	Learning Rate	5e-5
	Dropout	0.1
Multimodal	Batch Size	8
	Epoch	24
	Attention Heads	12

Baseline Methods and Evaluation Metrics

To validate the effectiveness of the proposed method, comparative experiments are conducted with three categories of representative methods:

1) Vision-based methods like Res-Target, which uses Resnet-152 to extract image features and concatenates them with aspect term features for classification in the BERT model.
2) Text-based methods

IAN (Qi et al., 2022): Interactive attention network model, which generates representations for both the context and aspect terms separately. It utilizes an attention mechanism for interactive learning and concatenates these representations to predict the sentiment polarity of aspect terms.

RAM (Huang et al., 2019): Multi-layer attention and memory network model, which employs a bidirectional LSTM structure to establish a memory unit. It obtains weighted features from the memory unit through a multi-layer attention mechanism.

MGAN (F. Chen et al., 2019): Multi-granularity attention model, which concatenates the coarse-grained weights of the overall impact of aspect terms on the text and the fine-grained weights of each word within the aspect terms on the text.

3) Multimodal-based methods

ViLT (W. Yu et al., 2021): A vision and language pre-trained model, which concatenates text features and projected features of image slices and inputs them into a Transformer encoder.

IGNN (Devlin et al., 2018): Multimodal interactive graph model, which forms a multimodal interaction graph through a full connection of the aspect term-word dependency graph and the visual block position relationship graph, using a graph attention network for multimodal feature fusion.

TomBERT (J. Yu & Jiang, 2019): Aspect term-image matching attention model, which uses an attention mechanism to match aspect terms with images and concatenates this with text features for multimodal feature fusion using a transformer.

To effectively evaluate the performance of this model, accuracy and F1 score are chosen as evaluation metrics.

Experimental Results and Analysis

The results of the method proposed in this paper compared to baseline methods on the experimental datasets are presented in Table 4. The experimental results demonstrate that the bidirectional complementary correlation-based multimodal target sentiment analysis method proposed in this study shows improved accuracy and F1 scores on both the Twitter-15 and Twitter-17 data sets. Based on the results, the following conclusions can be drawn:

Comparison with text- based methods: The performance of the vision-based Res- Target method is considerably limited compared to text-based methods, indicating that target sentiment analysis primarily relies on textual information, and images are less suitable for independently analyzing the sentiment polarity of aspect terms.

Advantages of multimodal methods: Multimodal methods outperform unimodal methods, suggesting that images and text provide complementary information, with images playing a supportive role to text.

Efficiency of multimodal fusion: The Res-BERT+BL method, which merely concatenates feature vectors from different modalities, performs worse compared to other multimodal methods. Treating modalities separately leads to a fragmentation of inter-modal relations, failing to capture the semantic information. Multimodal feature fusion captures the interactive information between different modalities.

Aspect tem-sensitive features: Both mPBERT and TomBERT analyze aspect term sentiment after fusing text and image features. TomBERT's use of aspect tem-sensitive image features guided by aspect terms for attention weight allocation in image regions is more effective than mPBERT's approach. This attention to emotion-related regions in images under the guidance of aspect terms can significantly improve the accuracy of target sentiment analysis. CapTrBERT's ApproachCap TrBERT transforms visual modality into text modality. Its experimental results validate this approach and further confirm that in multimodal target sentiment analysis, the text modality, as the primary information source, provides rich semantic information.

The BiCCM model proposed in this paper shows improved accuracy and F1 values on the Twitter-15 data set compared to baseline models. The model outperforms most comparative methods on the Twitter-15 dataset and achieves the highest accuracy and F1 values on Twitter17.

This paper's approach of generating aspect term-sensitive image features, followed by cross-modal mutual guidance and feature fusion based on each modality, demonstrates that leveraging inter-modal correlations to generate cross-modal features can capture the shared sentiment polarity in text and images. Feature fusion effectively achieves inter-modal information complementarity, significantly enhancing the accuracy of sentiment analysis.

Ablation Study

To analyze the impact of different model components on the experimental results, five variants of the BiCCM model were designed for ablation studies. The results of the ablation experiments are shown in Table 5. The variants are as follows: w/o Image: No visual modality, only textual modality input. w/o Cross-Att: No aspect term-guided image module, global image features are fed into the model for sentiment analysis. w/o IGT: No aspect term-image feature guide text module, text features are fused without the guidance of aspect term- sensitive image features. w/o TGI: No text-guided aspect term- image feature module, aspect term-sensitive image features are directly fed into the visual gate without textual guidance. w/o MFBT: No multimodal feature fusion-text module, image- sensitive text features generated under aspect term-sensitive image feature guidance are fed directly into the classification module without feature fusion. w/o VG: No multimodal feature fusion-image module, aspect term-sensitive image features are fed directly into the classification module without passing through the visual gate.

Based on Table 5, it can be observed that the model with only textual input and no visual module still maintains good performance, indicating that the visual modality only plays a supportive role.

Table 4. Experimental results

Modality Vision	Method Res-Target	Twitter-15		Twitter-17	
		Accuracy	F1 Score	Accuracy	F1 Score
		59.88	**46.48**	**58.59**	**53.98**
Text	IAN	68.18	67.14	63.41	62.94
	RAM	70.68	63.05	64.42	61.01
	MGAN	71.17	64.21	64.75	61.46
Multimodal	ViLT	71.80	65.10	62.70	58.80
	IGNN	71.98	71.43	64.83	64.42
	TomBERT	77.15	71.75	70.34	68.03
	BiCCM	77.82	71.77	70.82	69.30

Table 5 Experimental results of model ablation

Model	Twitter-15		Twitter-17	
	ACC	Mac-F1	ACC	Mac-F1
w/o Image	77.33	72.04	68.80	66.50
w/o Cross-Att	77.24	71.27	67.74	65.29
w/o IGT	76.75	71.20	69.50	66.94
w/o TGI	76.08	71.14	69.28	66.81
w/o MFBT	76.80	71.20	68.31	65.65
w/o VG	76.56	70.35	69.12	67.66
BiCCM	77.82	71.77	**70.82**	**69.30**

The performance decreases without the target- guided image module, especially on the Twitter-17 data set, which has shorter average text length and more aspect terms per tweet. Without this module, it is challenging to determine the image regions corresponding to multiple aspect terms, making it difficult to analyze the sentiment polarity of aspect terms accurately. Conversely, the absence of the cross-modal guidance module has a more significant impact on the Twitter- 15 data set, where the correlation between text and images in tweets is tighter. The performance drops significantly on both data sets without the multimodal feature fusion module, indicating that feature fusion establishes connections between textual and image features, achieving complementary information between the two modalities.

CONCLUSION

This paper presents a bidirectional complementary correlation-based multimodal target sentiment analysis model. The main idea is to capture emotion-related regions in images associated with aspect terms through the intrinsic correlation between text and images. The model guides feature representation between modalities, yielding image-sensitive text features and text-sensitive image features. Finally, features from both modalities are fused to obtain a multimodal feature representation containing complementary information from text and images. Additionally, considering the cross-modal alignment between text and images, a visual gate adjusts the output of image features. Experimental results on the Twitter-15 and Twitter-17 data sets validate the effectiveness of the proposed model. Future work will focus on deepening the research on text-image semantic alignment to enhance the intrinsic correlation between text and image features.

The BiCCM-ABSA model has a wide range of potential applications across various industries. In marketing, it can analyze customer feedback on social media, providing valuable insights into consumer sentiment toward products and services. In finance, it can be used for sentiment analysis of market-related discussions, aiding in predictive analysis of market trends. Additionally, in public-opinion monitoring, it can help governmental and nongovernmental organizations gauge public sentiment on policies or events. The broader implications for natural language processing include advancing the field of multimodal sentiment analysis and setting new benchmarks for accuracy in complex data environments.

While the BiCCM-ABSA model represents a significant advancement in sentiment analysis, it has certain limitations. One of the main limitations is its performance in scenarios with highly ambiguous or sparse aspect-sentiment pairs. Additionally, the model's reliance on high-quality, annotated data sets may limit its applicability in less structured environments. Future work should focus on addressing these challenges, improving the model's adaptability and accuracy in a wider range of contexts.

AUTHOR NOTE

Acknowledgements: The authors would like to thank the editor and anonymous reviewers for their contributions toward improving the quality of this paper.

Data availability: The data used to support the findings of this study are included within the article.

Conflicts of interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding statement: This research received no external funding.

Correspondence concerning this article should be addressed to School of Electronic and Electrical Engineering, Shanghai University of Engineering and Technology, Shanghai, 201620.

Corresponding author: Yujie Xiong, xiong@sues.edu.cn.

REFERENCES

Al-Ayyoub, M., Rabab'ah, A., Jararweh, Y., Al-Kabi, M. N., & Gupta, B. B. (2018). Studying the controversy in online crowds' interactions. *Applied Soft Computing*, *66*, 557–563. doi:10.1016/j.asoc.2017.03.022

Al-Qerem, A., Alauthman, M., Almomani, A., & Gupta, B. B. (2020). Iot transaction processing through cooperative concurrency control on fogcloud computing environment. *Soft Computing*, *24*(8), 5695–5711. doi:10.1007/s00500-019-04220-y

Barbosa, A., Bittencourt, I. I., Siqueira, S. W., Dermeval, D., & Cruz, N. J. (2022). A context-independent ontological linked data alignment approach to instance matching. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–29. doi:10.4018/IJSWIS.295977

Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, (pp. 452-461). ACL. doi:10.18653/v1/D17-1047

Fan, F., Feng, Y., & Zhao, D. (2018). Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, (pp. 3433-3442). ACL. doi:10.18653/v1/D18-1380

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image-text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, *167*, 26–37. doi:10.1016/j.knosys.2019.01.019

Ismail, S., Shishtawy, T. E., & Alsammak, A. K. (2022). A new alignment word-space approach for measuring semantic similarity for arabic text. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–18. doi:10.4018/IJSWIS.297036

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). ACL. doi:10.18653/v1/2021.emnlp-main.360

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). ACM. doi:10.1145/3474085.3475692

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. (pp. 5583-5594). ACL.

Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., & Hovy, E. (2021). Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th annual meeting of the association for computational linguistics & the 11th international joint conference on natural language processing, 1*, (pp. 6319-6329). ACL. doi:10.18653/v1/2021.acl-long.494

Lou, Z., Wu, Y., Fan, C., & Chen, W. (2020). *Aspect-based sentiment analysis on convolution neural network & multi-hierarchical attention. In 2020 international conference on technologies & applications of artificial intelligence (TAAI)*. IEEE.

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, 1*, (pp. 1990-1999). ACL. doi:10.18653/v1/P18-1185

Mohammed, S. S., Menaouer, B., Zohra, A. F. F., & Nada, M. (2022). Sentiment analysis of covid-19 tweets using adaptive neuro-fuzzy inference system models. [IJSSCI]. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–20. doi:10.4018/IJSSCI.300361

Murthy, G., & Allu, S. R., & havarapu, B., Bagadi, M., & Belusonti, M. (2020). Text based sentiment analysis using lstm. *Int. J. Eng. Res. Tech. Res*, *9*(5).

Qi, S. Z., Xianying, H., Haidong, S., & Jiayan, L. (2022). Aspect based sentiment analysis with progressive enhancement & graph convolution. *Application Research of Computers*. *Jisuanji Yingyong Yanjiu*, *39*(7).

Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A survey of deep active learning. *ACM Computing Surveys*, *54*(9), 1–40. doi:10.1145/3472291

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983. doi:10.1016/j.asoc.2020.106983

Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using e-reputation for sentiment analysis: Twitter as a case study. [IJCAC]. *International Journal of Cloud Applications and Computing*, *11*(2), 32–47. doi:10.4018/IJCAC.2021040103

Singh, S. K., & Sachan, M. K. (2021). Classification of code-mixed bilingual phonetic text using sentiment analysis. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *17*(2), 59–78. doi:10.4018/IJSWIS.2021040104

Su, C. W., Liu, Y., Chang, T., & Umar, M. (2023). Can gold hedge the risk of fear sentiments? *Technological and Economic Development of Economy*, *29*(1), 23–44. doi:10.3846/tede.2022.17302

Tobaili, T., Fern&ez, M., Alani, H., Sharafeddine, S., Hajj, H., & Glavas, G. (2019). Senzi: A sentiment analysis lexicon for the latinised Arabic (arabizi). In *International conference recent advances in natural language processing 2019 natural language processing in a deep learning world: Proceedings*, (pp. 1204–1212). Springer. doi:10.26615/978-954-452-056-4_138

Vaswani, A., Shazeer, N., & Parmar, N. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wu, H., Cheng, S., Wang, J., Li, S., & Chi, L. (2020). Multimodal aspect extraction with region-aware alignment network. In *Natural language processing and Chinese computing: 9th CCF international conference, NLPCC 2020, Zhengzhou, China, October 14-18, proceedings, part I 9* (pp. 145-156). Springer. doi:10.1007/978-3-030-60450-9_12

Xu, N., Mao, W., & Chen, G. (2019). Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 371–378. doi:10.1609/aaai.v33i01.3301371

Yen, S., Moh, M., & Moh, T. S. (2021). Detecting compromised social network accounts using deep learning for behavior & text analyses. [IJCAC]. *International Journal of Cloud Applications and Computing*, *11*(2), 97–109. doi:10.4018/IJCAC.2021040106

Yu, J., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. *IJCAI (United States)*, 5408–5414. doi:10.24963/ijcai.2019/751

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 10790–10797. doi:10.1609/aaai.v35i12.17289

Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Advance online publication. doi:10.1609/aaai.v32i1.11962

Zhang, Q., Guo, Z., Zhu, Y., Vijayakumar, P., Castiglione, A., & Gupta, B. B. (2023). A deep learning-based fast fake news detection model for cyberphysical social services. *Pattern Recognition Letters*, *168*, 3138. doi:10.1016/j.patrec.2023.02.026

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*.

Zhou, J., Huang, J. X., Chen, Q., Hu, Q. V., Wang, T., & He, L. (2019). Deep learning for aspect-level sentiment classification: Survey, vision, & challenges. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 78454–78483. doi:10.1109/ACCESS.2019.2920075

Jing Yang holds a master's degree; she graduated from Shanghai University of Engineering Science University in 2023. Her research interests include multimodal sentiment analysis.

Yujie Xiong holds a doctoral degree; he graduated from East China Normal University in 2018. His research interests include pattern recognition and intelligent systems.