International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 975-989
https://doi.org/10.70917/ijcisim-2026-3168
Article

IPMMSA: MODELLING AN IMPROVED PROMPT-BASED MULTI-MODAL SENTIMENT ANALYSIS OVER FASHION DATASETS

M. Yuvaraja[1], C. Kumuthini[2]

[1] Department of Computer Science Dr.N.G.P Arts and Science College Coimbatore Tamilnadu, India.
yuvaraja0105@gmail.com
[2] Department of Computer Applications Dr.N.G.P Arts and Science College Coimbatore-641048 Tamilnadu, India.
kumuthini@drngpasc.ac.in

Abstract: To enhance pertinent decision-making in a variety of applications, prompt-based sentiment analysis attempts to leverage cross-modal opinion signals to analyze users' attitude direction regarding the specific attribute. Even though many techniques have been created, they are unable to use multiple knowledge types at once and are unable to successfully eliminate unwanted signals from various viewpoints, which can impair multimodal representations' discriminative power and keep models from performing better. To fill up the research gaps, this work suggests an improved prompt-based multi-modal sentiment analysis (IPMMSA) strategy that incorporates multi-view and diversified knowledge augmentation. In particular, it implements fine-grained image-aspect interactions by transforming the image into an underlying sequence of embedding's which makes filtering easier from a visual semantic standpoint. Attribute-guided vision-language interactions are then used to pull out important extract affective signals and suppress irrelevant content within a multimodal semantic framework, while the network structure is formulated to effectively exploit context-informed semantic fusion, syntactic relations, and sentiment-aware domain knowledge. Ultimately, the model yields robust and expressive multimodal embedding's to boost aspect-based sentiment analysis performance during multimodal integration with multi-modal fashion dataset. Finally, to show the superiority along with efficacy of our suggested approach extensive experiments were conducted on two widely used multi-modal fashion datasets.

Keywords: multi-modal analysis, sentiment analysis, prompt analysis, syntactic relation, domain knowledge

1. Introduction

On social media sites like Facebook, Twitter, and Weibo, a considerable volume of user-generated data has emerged that either directly or indirectly reflects users' thoughts about the specified topics [1]. It is anticipated that ABSA which involves mining and analyzing the explicit sentiment orientation for each aspect with polarity labels of positive, negative, or neutral will support decision-making in a variety of applications, including political elections, applications such as encompassing purchase guidance, service optimization, intelligent education and related domains [2]. Nevertheless, because it is extremely difficult to extract enough sentiment indicators in informal or noisy data, fragmentary, or brief online content, the majority of ABSA research now in existence are text-based approaches and have inadequate analytical results [3]. The aspect-level sentiment orientation for the target (i.e., Cason Olson) cannot be accurately identified by ABSA algorithms that only use the text data (i.e., "Cason Olson #stud"). To make a comprehensive sentiment prediction for the targeted aspect, it is essential to extract complementary sentiment cues from multiple modalities, as humans typically express their opinions and sentiments through a variety of channels, including text, images, and/or videos [4]. Consequently, multimodal ABSA (MABSA), which

draws inspiration from human sentiment identification has emerged as a viable approach to ABSA tasks.

The importance and necessity of extracting and combining complementing sentiment cues extracted from the textual input and the corresponding image have been highlighted by a number of techniques that have been developed recently [5]. The following restrictions, however, continue to keep them from performing to a higher standard. During the implementation of image interactions, granular aspect–visual relations, precise aspect–visual correlations, aspect-specific visual interactions early research rarely consider removing the possible interference caused by the image [6]. It is well known that just a portion of an image's information can help with sentiment detection for the specified aspect; the rest of the image's information will obstruct the model's ability to make the right decision [7]. Although the phrase clearly expresses disapproval of the element (i.e., Donald Trump), the image's multiple cheerful faces could easily lead the models to generate the wrong sentiment judgments. Therefore, it is very advantageous to remove image noise for the purpose of improving the robustness of fused multimodal features and enhance sentiment forecast accuracy [8]. Consequently, recent studies have attempted to mitigate visual noise using gating mechanisms or resemblance correspondence between the input text along with the corresponding image. Nevertheless, in modeling detailed image–aspect interactions alongside high-level image sentence interactions, existing approaches frequently disregard the disparity between image and aspect representations and the dominant contribution of text-based input for sentiment analysis [9]. This leads to poor noise filtering performance. Therefore, one of MABSA's biggest challenges is figuring out how to successfully lessen the detrimental effect of noise in the image brings. To integrate the complimentary information that has been retrieved from several perspectives, multimodal fusion is an essential step. However, the features gathered from various perspectives could be noisy and redundant, which would lessen the importance of important information in the sentiment classification [10]. This also leads to a high risk of over-fitting and inadequate robustness. Thus, it is essential to explore methods for generating robust multimodal fusion representations while retaining critical task-relevant information.

This existing work provide an improved MABSA method with cross-perspective noise suppression and knowledge-augmented learning to address the aforementioned problems [11]. To feature discrepancy representative and strong cross-modal attributes and improve emotion detection for the focused element the approach seeks to completely leverage various forms of knowledge (In particular syntactic, contextual, semantic, and along with outward emotional information) while efficiently filtering disturbances examined in three dimensions: visual–semantic, visual–textual and multimodal fusion [11]. To share the shared feature space together with the aspect expressions, the module, which models visual semantic information noise filtering initially encodes the image into a latent series of token vector using an encoder-decoder architecture and an image feature extractor. When modeling fine-grained image–aspect interactions, it is anticipated to reduce the disparity in representation between the aspect and its associated image [12]. This will help to preserve aspect-specific visual semantic information and simplify filtering disturbances through the visual–semantic lens. NFVTS is a different module that first extracts the framework employs object-level visual features as a visual cue, which is then concatenated using the sentence embedding to serve as keys and merits in the multi-modal attention computation, while accounting for the dominant function of the input sentence in sentiment analysis [13]. To produce context-aware visual-promoted textual semantic representations, it seeks to acquire complementing visual features and lessen

the detrimental effects of picture irrelevant information. NFVTS additionally builds a multi-knowledge graph to jointly utilize diverse knowledge sources, enabling the model to learn richer representations of the targeted aspect. This is followed by aspect-focused attention to reduce noise at the visual–textual semantic level. This is because multiple types of knowledge such as syntactic, contextual, semantic, and external sentiment information are crucial to improve the performance of aspect-based sentiment analysis (ABSA) [14].The initial multimodal fusion characteristics are then obtained by concatenating the outputs from these two perspectives. These features are further processed using the information bottleneck concept principle to suppress irrelevant and overlapping information, thereby producing richer multimodal representations for accurate sentiment prediction, particularly considering the possible noise introduced during the multi-angle multimodal integration process [15]. The following is a summary of this paper's contributions:

1. To produce stronger and more distinctive multi-modal features, we provide an improved prompt-based multi-modal sentiment analysis (IPMMSA) technique that can efficiently remove noise from multiple viewpoints while simultaneously utilizing numerous types of information. As far as we know, we are the initial to combine various-perspective with multiple-knowledge promotion to enhance IPMMSA performance.

2. To execute fine-grained multi-modal-aspect interactions and enable disturbances through the visual-semantic perspective, the IPMMSA component transforms the visual input converted into a hidden series of representations. After building the network structure to make by leveraging the majority of fused context-sensitive semantics, syntactic dependencies, and sentiment knowledge tailored to the domain, the IPMMSA module carries out aspect-focused

image–sentence interactions to reduce disturbances at the visual–textual semantic level and extract key sentiment cues. Furthermore, multimodal-level prompt-based sentiment analysis is thought to produce more reliable and superior cross-modal fused representations for sentiment prediction in the end.

3. To confirm the efficacy, superiority, and rationality of our suggested approach, we carry out comprehensive tests using two standard benchmark datasets, namely deepfashion_1 and fashion model dataset.

This paper's remaining sections are arranged as follows: The most relevant books are briefly reviewed in the following section. The next section provides a detailed introduction to our suggested approach. The experimental setups, baselines, findings, and the corresponding discussion is provided in the next section. The study is finally determined and future work is outlined in the final part.

2. Literature review

Sentiment analysis (SA) tries to allow devices to identify and distinguish the sentiment label of the text (e.g., positive, negative, or neutral). This is often referred to as opinion mining or emotion identification [11]. However, because of the features from online social media text (e.g., non-standard, fragmentary, or brief expressions), it is challenging to reliably determine the emotion polarity determined only from the textual modality [12]. Multimodal SA has lately gained popularity since users of online social platforms frequently share their thoughts in a variety of formats, including text, images, and/or videos. In order to have a thorough grasp of the sentiment polarity, MSA is committed to investigating complementing data from several modalities [13]. Previous MSA research has shown that in order to achieve more

accurate sentiment analysis performance, textual modality integration with other modalities is absolutely essential. Effectively fusing data from several modalities has emerged as a crucial problem for MSA tasks [14]. Generally speaking, there are three methods for creating multimodal representations: early-fusion, late-fusion, and hybrid fusion to derive single-modality features from multiple modalities, early-fusion techniques typically use numerous off-the-shelf models (e.g., RoBERTa and ResNet) [15]. These models are then fused to provide the fused multimodal features in the model's shallow layer. To illustrate, the author extracted textual and visual features using the LSTM network and the Inception architecture, respectively, and then fused these two feature types using a dense layer. Using voting or weighted algorithms, late fusion techniques combine the predictions of several classifiers constructed from the uni-modal data [16]. To create the final forecast, for example, initially acquired modality-specific prediction scores that are combined via weighted averaging. However, the sentiment prediction ability is limited since neither early nor late fusion can fully examine complimentary information. As a result, hybrid fusion MSA research has become more and more popular [17]. Attention-driven fusion, graph-based fusion, and gated fusion and other methods can be used to implement it. For instance, the author investigated multimodal integration using graph fusion technique after using adversarial learning to extract unimodal features [18]. Based on various designs and motivations, the hybrid fusion may generally combine information from several modalities more successfully. The hybrid fusion approach will also be used in this study to incorporate important multimodal data [19].

To successfully set up the relationship between the specified feature and its corresponding opinion term, researchers have attempted to take use of the sentence's syntactic dependencies [20]. They additionally known as syntax-driven ABSA approaches that typically use the syntactic structure as the adjacency matrix and the sentence's words as graph nodes to learn aspect-aware representations solely a small number of syntax-based studies were part of MABSA, while the majority of prior syntax-based studies relied exclusively on the textual modality, without incorporating information from other modalities [21]. AMIFN, for instance, regarded the sentence's syntactic dependencies as the graph's adjacency matrix after first integrating the sentence representation combined with global visual information using attention [22]. The graph nodes were then initialized using the multimodal fusion information. Given the textual input and image description, and the visual input, correspondingly, CoolNet built graph networks [23]. Nevertheless, prior research has typically overlooked the function of the textual modality in facilitating visual–textual interactions and has not included noise filtering from various angles. Furthermore, prior research has asserted the advantages of applying several forms of information, but few studies have concurrently utilized multiple-knowledge to learn more representative features [24] – [25].

3. Methodology

Using a single GeForce RTX 3090 GPU processor unit, we used MATLAB 2020a to train the network for 100 epochs. To compare the suggested model with the most advanced models, the same hyper-parameters and settings were used in every experiment. One was chosen as the batch size. We used the Adam optimizer with a starting learning

rate of 0.001 and applied a cosine annealing scheduler for learning rate adjustment. The data were partitioned into training, validation, and test sets in an 80:10:10 ratio. The dataset was obtained from five centers, but this was not taken into consideration when the data were randomly permuted and partitioned. Random rotations and flips were used to enrich the data during training. Fig 1 illustrates a diagram of the suggested framework utilizing U-Net as its foundation [26]. The

suggested framework includes three key elements: region attention (RA), multi-scale fusion (MSF), and squeeze and excitation norm (SEN). To obtain attribute extracted out of the fashion images, feature maps from the backbone encoder are combined with the RA output and passed to the decoder. The MSF and SEN norm are incorporated within the main network. The MSF utilizes various contextual details derived from abstract feature maps are positioned between the encoder and decoder. Most neural network-based learning networks employ the rectified Linear activation function along with normalization in sequence following each feature extraction layer, regarding this method as a single block [27]. The suggested component employs the SEN Norm to normalize this block.

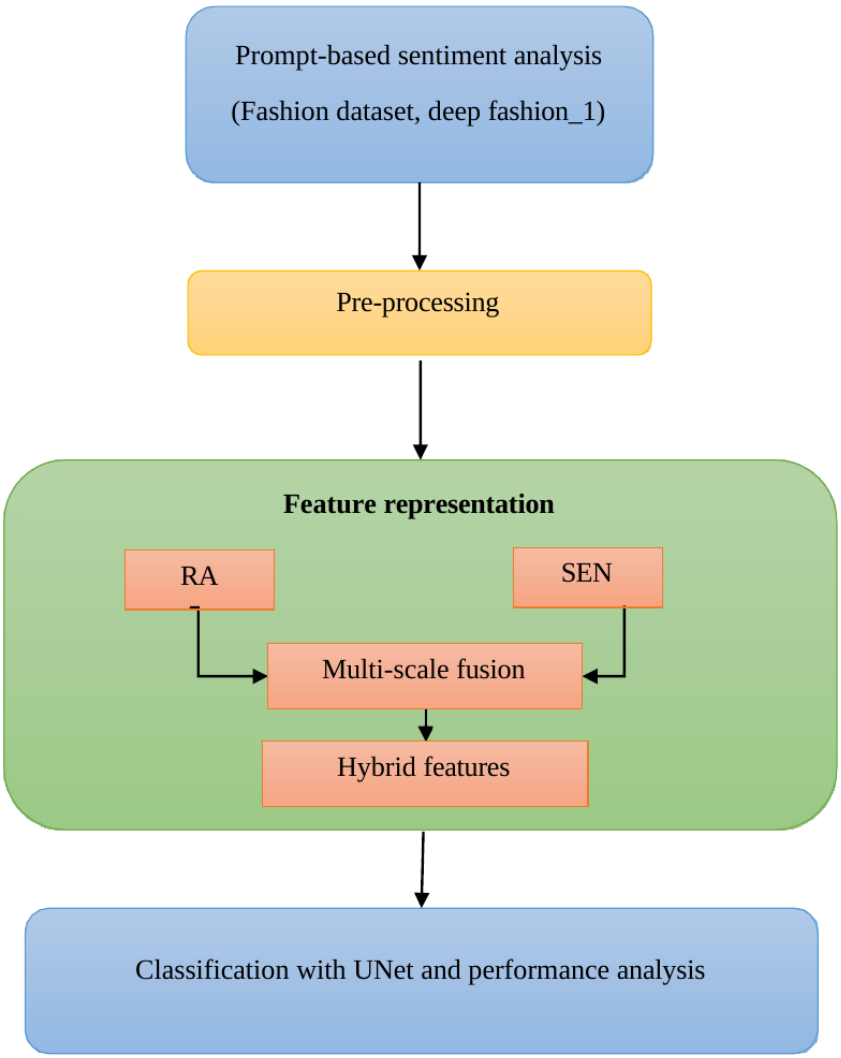

Figure 1: Flow diagram

3.1 Feature representation

Two categories of blocks were used in the encoder of the main U-Net. The first handles adjustments between input and output channel sizes, whereas the second applies when the channel sizes are the same. Each block contains 3 × 3 × 3 convolution, ReLU activation, SEN normalization, and a remaining connection.

3.2 Attention module

Given the increased sensitivity of input images, the first RA sub-network is suggested for use. To help with tumor localization, RA creates a tumor attention map $A \in R1 \times D \times H \times W$ from the input image $P \in R1 \times D \times H \times W$. The segmentation backbone incorporates a tumor attention map. The RA sub-network consists of multiple layers, normalization, and activation functions and is structured on a modified U-Net. A 3 × 3 × 3 convolution with stride two is used in the contracting path to reduce the input size by half replacing traditional 2 × 2 × 2 pooling layers. Batch normalization and ReLU are applied subsequently. Trilinear interpolation was used by the expanding layer to double the spatial dimensions of the image in the expanding path, and 3 × 3 × 3 convolution layer was used to cut the number of channels in half. To decrease over-fitting and introduce stochasticity into the network, we used dropout to concatenation in the decoder. There is a 0.5 dropout rate. As the last layer, we employed the tanh activation function to suppress less valuable features and highlight important ones [28].

3.3 Squeeze and excitation normalization

On input images, the fashion typically exhibit substantial uptake. Nonetheless, high physiological image uptake is also present in some non-sensitivity areas. As a result, it may be inaccurate to differentiate image and video using solely the values of imaging. To successfully connect data obtained from imaging with functional information from images, we employed SEN Norm in our network to strengthen relationships between channels within the combined text and video feature maps. The SEN block the interactions between channels, serves as the foundation for this module. Concatenated audio and video images served as the two channels' inputs to the backbone network. Each layer was then subjected to SEN norm, and the following three procedures were carried out. Initially, instance normalization was used to normalize the input features [29]. Second, as this approach is performed per channel, the results of instance normalization ignore information dependencies across channels. To address this issue, we used SEN blocks to construct a set of two parameters, γi, βi based on

input features. Lastly, $\gamma i \; and \; \beta i$ were employed to adjust and shift the normalized values

$$y_i = \gamma_i x_i + \beta_i$$

Where, yi is the SEN norm's final output and x' is the instance normalization's output.

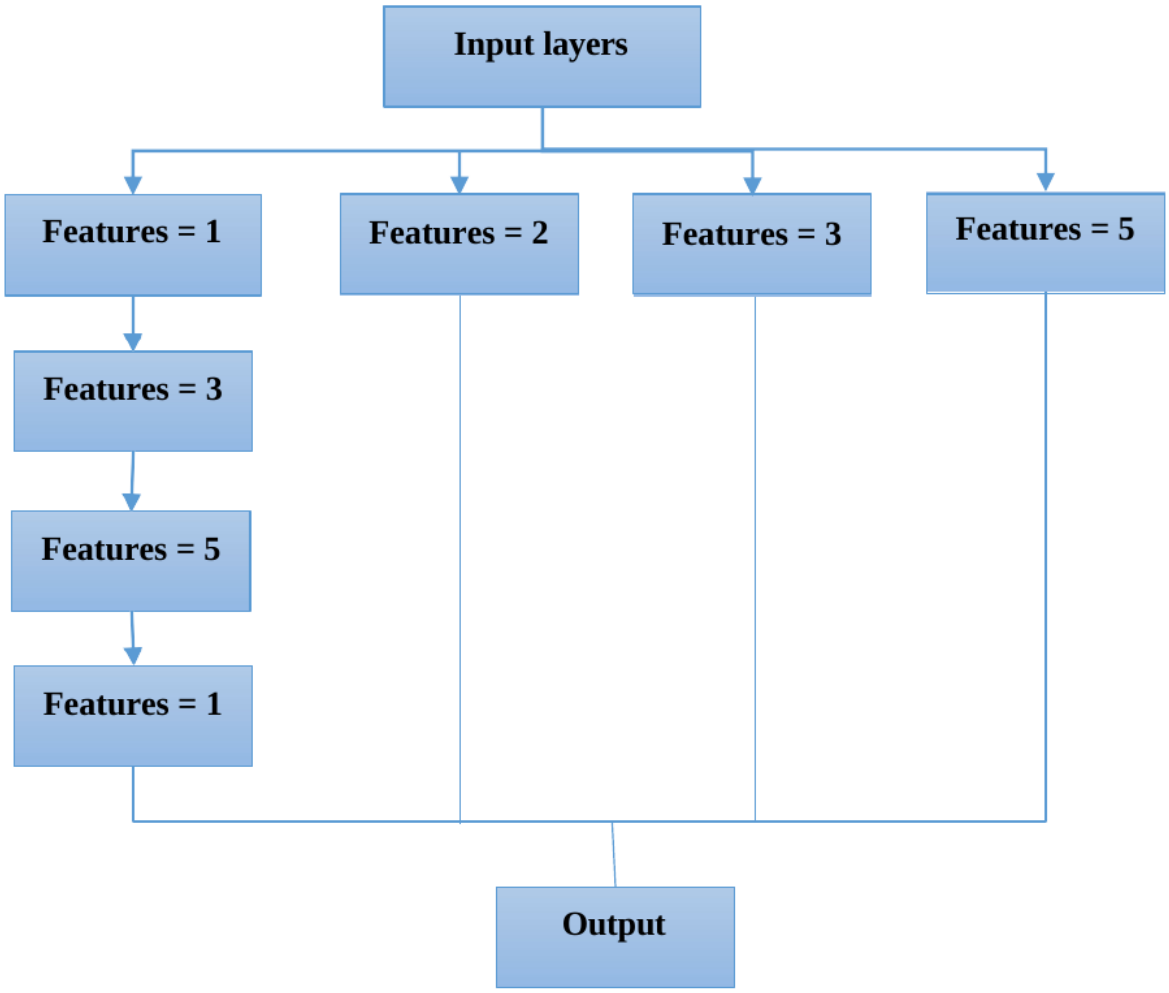

Figure 2: Fusion model

3.4 Fusion mechanism

The work created an MSF that combines contextual information on many scales to improve the network's generalization capability for a variety of geometries. The work may solve the problem of spatial information loss brought on by the encoder's successive pooling operations by placing the MSF in between the encoder and decoder. In this module, convolutions with varying kernel sizes are used to extract features at four different scales, expanding the kernels' receptive field while avoiding the computational overhead of larger kernel sizes. Fig 2 illustrates the kernel sizes used for each convolution with corresponding receptive field sizes of 3, 7, 11, and 19.The four elements are then combined into one to create the contextual information.

3.5 Decoder

In the decoder path, the attention from the RA and SEN norm were utilized to produce precise, high-resolution segmentation results. To direct the localization, the skip connection which conveys encoder features is modulated by the attention map through multiplication. A 3 × 3 × 3 convolution, ReLU and SEN norm make up each block. To transmit low-resolution features, we additionally included three upsampling paths in the decoder. These paths were made up of using tri-linear interpolation to extend A 1 × 1 × 1 convolution is applied to the feature map to decrease the number of channels. Lastly, the sigmoid function was used to produce a probabilistic segmentation map [29].

3.6 Loss analysis

The logistic loss function is most frequently utilized for segmentation problems since segmentation is typically thought of as pixel-wise classification. However, this loss of function would not be optimal in our investigation because of the large range of image sizes on dataset. Thus, rather than using the logistic loss function, we employed the soft dice loss function. In computer vision, the Dice similarity coefficient is the predominant metric to determine how similar two images are, particularly when it comes to input images. The definition of the soft dice loss is:

$$L_{dice\ loss}\ (y,\ \hat{y}) = 1 - \frac{2\ \sum_i^N y_i \hat{y}_i + \in}{\sum_i^N y_i^2 + \sum_i^N \hat{y}_i^2 + \in}$$

Without the need of extra restrictions, our method adopts end-to-end optimization guided by the soft dice loss and is regarded as a single integrated model.

3.7 Ablation study

The impact of applying the three suggested techniques to enhance image segmentation accuracy was examined by comparing various strategy combinations. Performance was superior to native U-Net when RA, SEN Norm, or MSF were used. RA combined with SEN norm achieved the highest performance the best out of all the two strategy combinations. The suggested model that combined the three suggested tactics fared better than the other combinations. The suggested model's DSC, which combined the three methods, was 8.7% higher than the U-Net's. Furthermore, as demonstrated by its superior precision over U-Net, The model lowered the incidence of false positives. Fig 3 presents segmentation results for the baseline U-Net, each individual improvement method and their combined application. Only the suggested approach was able to differentiate the images in this instance, distinguished from other areas exhibiting high uptake. In tumor delineation, Fig 3 demonstrates that the proposed model perform well than native U-Net [30]. Because the information contained in each image modality is distinct and complementary, combining image characteristics is helpful in the context of multi-modal imaging. Effective methods for combining the image features include co-learning the two features and using attention modules that concentrate on target structures. If image is accessible for segmentation, it's also critical to use the uptake data appropriately.

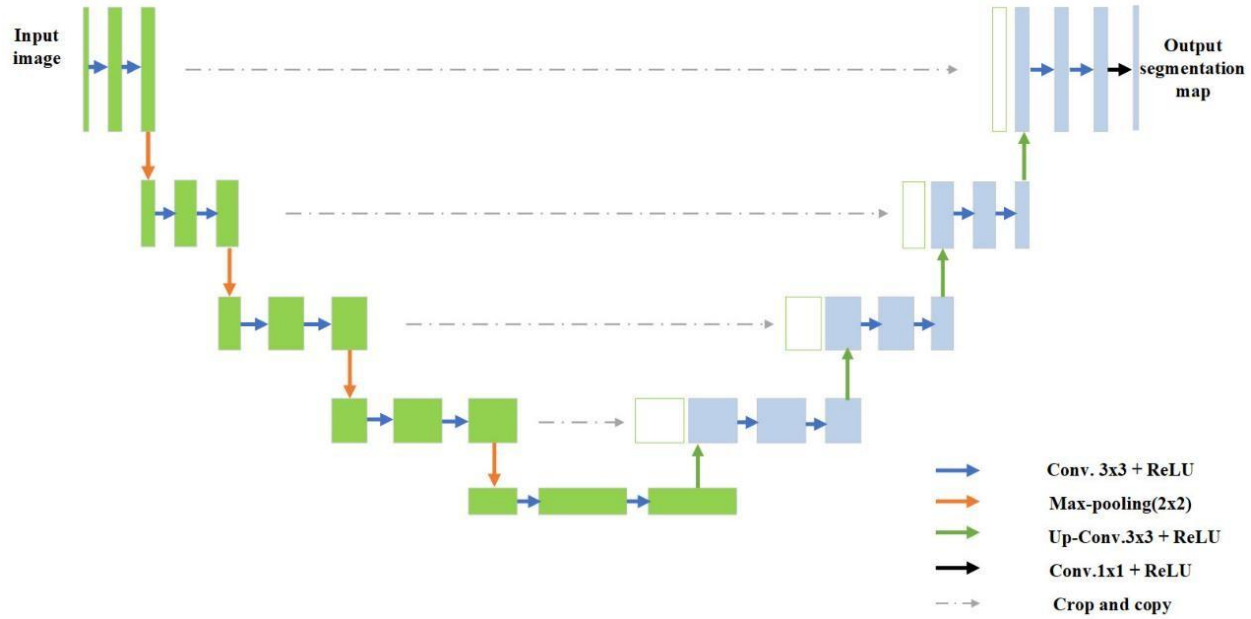

Figure 3: UNet model

Considering all of these factors, we created a new deep learning-based method to enhance segmentation of the images. By using RA to guide localization and SEN norm to co-learn multi-modal features, the suggested network can more successfully supplement. By using extra attention modules following the convolution procedure,

segmentation accuracy is improved by incorporating Attention U-Net and CBAM. When this technique is employed, each image's features are first extracted using convolution, and the extra attention modules then fuse the extracted features to carry out co-learning. Nevertheless, the computational complexity is increased by the independent extraction and fusion. A sequence of convolutional, activation, and normalization layers make up the convolutional operation in CNNs. Usually, batch or instance normalization are used in the normalizing layers. However, the normalizing layer we utilized, SEN Norm, feature representation and joint learning at the same time. However, RA employs acquiring knowledge that concentrates on uptake analyzed independently, without collaborative learning between the images that could lead to the incorrect classification of some areas exhibiting high resolution image uptake, potentially mistaken for prediction. Unlike other models, the approach combines data with multi-modal data, utilizing sensitivity and decreasing false-positive detections [30]. The employment of SEN norm with RA is more successful than without RA. Fashion image segmentation involves a variety of items. Images vary in size, form, and location, although normal structures commonly share similar morphologies. As a result, networks for segmentation need relevant contextual data.

To capture surrounding information and emphasize top level salient features developed a novel context feature extractor was incorporated between the U-Net encoder and decoder. Similar to this, we developed a model to use convolution to capture and learn surrounding information across multiple scales, which we then applied to RA and SEN. The usefulness of MSF in enhancing generalization ability is demonstrated, which shows that DSC increased by 0.97% when MSF was used in conjunction with RA and SEN. The impacts of merging the factors were examined. The two combinations outperformed learning with just one element, suggesting that each element helps to make up for the drawbacks of the other approach. Integration of RA and SEN performed the best out of the two combinations. In the same way, the performance improvement was increased when all three were combined. The findings of our investigation, alongside other techniques are summarized:

1) Background data: The suggested component outperformed existing model increasing accuracy by 10% and decreasing loss. Our model outperformed existing model in defining the image boundaries.

2) Sensitivity analysis: Through harnessing high sensitivity, our model achieved subsequent performance than model on all metrics, particularly achieving a 3.9% improvement in accuracy. Fig 4 demonstrates the model has more incorrect positives and incorrect negatives than the suggested method.

3) Co-learning: Attention model performed the best and existing model is the worst among the attention module-based techniques that concentrate on feature correlation. Our suggested model outperformed attention U-Net. The model achieved better performance compared with attention U-Net with respect to accuracy by 1.9% respectively. Attention U-Net produces more incorrect positives than the suggested component. But other approaches outperform the suggested approach in Table 1. To improve segmentation accuracy, RA creates a map that

considers high image samples. Our suggested method uses this information to separate tumors. The tumor attention map, which typically highlights a greater area than the images true size. Consequently, this is the cause of growth relative to other networks. According to this work, a successful approach to precise segmentation of images from input samples involves jointly learning features from multi-modal data images equipped incorporating an attention mechanism to makes use of images exhibiting higher sensitivity. To our knowledge, no previous study has taken use of great sensitivity while utilizing the complimentary information of two modalities.

4. Numerical results and discussion

To clearly demonstrate the efficacy of our suggested approach, this part first offers the experimental configuration and widely used baselines. Next, presents and analyzes the comparison and ablation results. Lastly, it displays the multimodal graphical illustration. In order to qualitatively demonstrate its superiority, a case study is also given.

Table 1: Statistics of Fashion model dataset and Deepfashion_1 dataset

Label	**Fashion model dataset**	**Fashion model dataset**	**Fashion model dataset**	**Deepfashion_1**	**Deepfashion_1**	**Deepfashion_1**
Label	Train	Dev	Test	Train	Dev	Test
Product title	929	304	318	1509	516	494
Product description	1884	671	608	1639	518	574

Label	Fashion model dataset	Fashion model dataset	Fashion model dataset	Deepfashion_1	Deepfashion_1	Deepfashion_1
Product category	369	150	114	415	145	169
Total	3182	1125	1040	3565	1179	1237

The comparison findings between our model and baselines are displayed in Table 2. The mean scores and standard deviations over five runs with random initialization are what we report. The baseline data are from their original articles because implementation details for certain experiments are unavailable. We performed a t-test, considering p-values < 0.05 as statistically significant shows that our approach is considerably superior to baselines. The greatest findings are bolded, while the most effective baseline findings are highlighted. Table 1 demonstrates which our suggested approach significantly improves on both datasets when compared to these baselines. It is also possible to get the following observations.

(1) Among all unimodal and multimodal methods, ResNet-Aspect performs the worst since it solely uses the visual modality. It implies that the visual modality alone is unable to supply sufficient data to enable meaningful sentiment analysis.

(2) Compared to ResNet-Aspect, the text-based baselines may perform better. This indicates that the visual modality is unable to be examined separately and that the textual modality is crucial for handling multimodal activities. Compared to earlier text-based methods, BERT-based unimodal approaches achieve markedly improved performance. It demonstrates that using a pre-trained language model instead of more conventional models can enhance

sentiment analysis performance. Furthermore, on both datasets, the text-only version of our model surpasses BERT-ATT by 2.6–3.5% in accuracy and 4.2–6.2% in F1. In order to improve sentiment analysis performance, it illustrates the advantages of building a multi-knowledge-depends IPMMSA and removing interference from several viewpoints.

3) Multimodal methods outperform their respective unimodal methods with the help of their related photos. It also illustrates the visual modality's supportive function. Therefore, while investigating coarse-grained image-sentence interactions, this is essential to understand promote textual qualities via visual cues and reduce the negative impact of visual noise.

(4) In general, multimodal strategies perform better than unimodal strategies. It suggests that there are important sentiment cues about the targeted aspect in both the input language and the image that goes with it. It is anticipated that combining sentiment cues from several modalities will enhance sentiment analysis outcomes. Thus, there has been a recent surge in interest in MABSA research. Recent multimodal baselines have made significant strides because of the pre-trained networks strong capabilities (like BERT or its variations). However, IPMMSA-based unimodal techniques perform better than BERT that disregards aspect-oriented detailed interactions. It illustrates the value and efficacy of investigating interactions guided by aspects for removing noise and obtaining aspect-related data. Therefore, in our suggested IPMMSA approach, the modules both use attention guided by aspects towards to weed out irrelevant data and preserve significant emotion signals related to the desired aspect.

(5) Our suggested IPMMSA strategy performs better than all current state-of-the-art multimodal techniques. On both datasets, its performance is 0.91–1.13% better on Acc and 0.74–0.88% better on F1 than the SOTA baseline. There are three potential causes. To begin with, current baselines employ using the extracted visual representation directly to explore fine-grained aspect-image interactions, without accounting for the representation gap, could lead to reduced effectiveness in noise filtering. The representation gap is unavoidable when different pre-trained models are used for different modalities. To close this gap, a number of baselines produce a supporting sentence for every image; nevertheless, they do not use visual meaning to investigate fine-grained aspect-image relationships correlations.

Our suggested IPMMSA module solves this problem by transforming embedding the image into an implicit space token embedding ordered aspect-image representations interaction learning. Second, current research does not account for the possible interference introduced by the image; instead, it performs cross-modal image-sentence interactions by directly using the visual information. Furthermore, individuals hardly ever fully utilize a variety of knowledge categories to acquire more potent sentiment signals. In order to address this problem, our suggested first, the IPMMSA module learns textual representations enriched with visual cues while taking into account the supporting role of images. Next, it builds a multiple-knowledge-based attention to extract important sentiment information.

Lastly, it filters out interference and retains helpful sentiment signals related to the target aspect by performing aspect-oriented attention. Finally, from the standpoint of cross-modal fusion, noise filtration is not taken into account in many baseline models. To further eliminate noise and produce more reliable multimodal fusion features, our suggested IPMMSA component compresses the multimodal representations by following the information bottleneck. The fact that our suggested approach outperforms all of these baselines on both fashion datasets is therefore not surprising as shown in Fig 4 to Fig 6.

Table 2: Performance based on different approaches

Modal	**Fashion dataset**	**Fashion dataset**	**DeepFashion_1**	**DeepFashion_1**
Modal	Acc	F1	Acc	F1
	Visual model	Visual model	Visual model	Visual model
ResNet	59	47	57	53
Proposed	89	87	87	83
	Text model	Text model	Text model	Text model
LSTM	70	63	62	57
RAM	71	64	65	61
IAN	71	63	66	62
GAN	72	64	65	65
BERT	75	69	69	59
BERT-QA	76	67	63	66
ATT	77	68	68	70
Proposed	98.3	93.6	91.5	90.6
	Textual and visual model	Textual and visual model	Textual and visual model	Textual and visual model
ResIAN	71	63	66	63
ResGAN	72	64	64	64
MIM	73	65	65	63
AFN	74	67	68	64
Bert	75	69	67	65
Tom-Bert	77	71	70	68
CapBert	78	73	69	69
KEF-Bert	79	74	72	69
HMT	78	75	71	72
ITM	79	73	73	68
MSC	80	71	70	69
FTE	78	79	71	69
MPTFI	77	73	73	68
ABSA	78	71	72	69
SRI	79	75	70	72
MG	80	74	80	71
Proposed	92	95	93	92

Table 3: Ablation study

Model	Fashion dataset	Fashion dataset	DeepFashion_1	DeepFashion_1
Model	Acc	F1	Acc	F1
Without NFVS	80	74	71	69
Without image conversion	60	53	60	55
Without aspect	78	74	71	70
Without visual information	78.5	72	72	71
Without SentiNet	78.6	73	73	68
Without GCN	79	72	72	70
Proposed	89.9	85	84	82

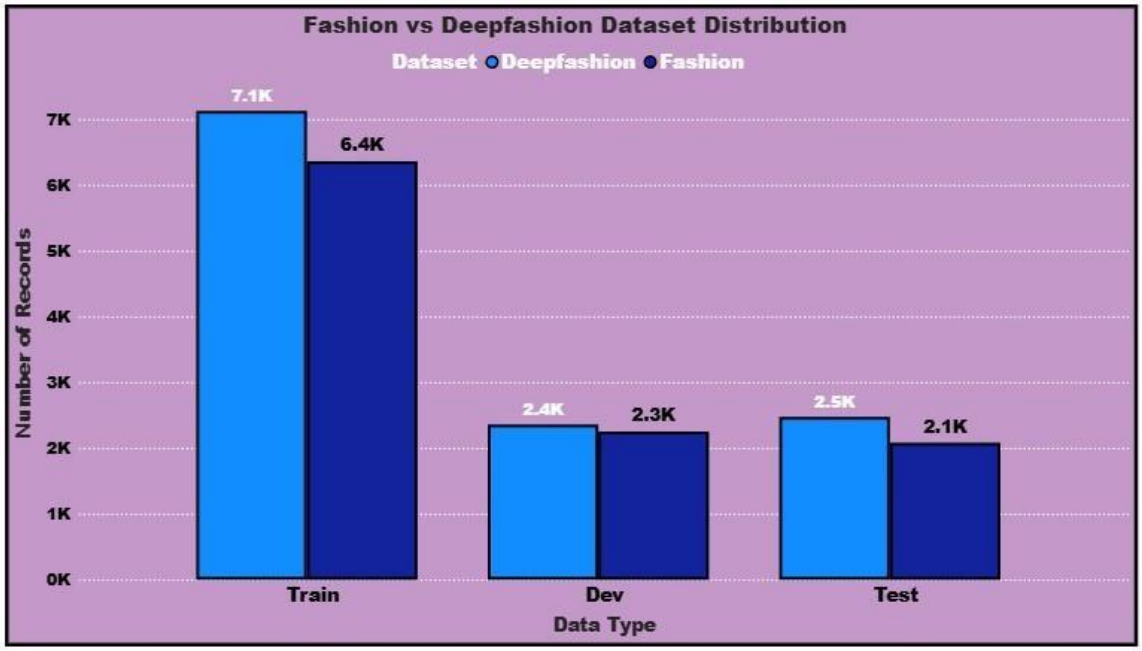

Figure 4: Dataset distribution

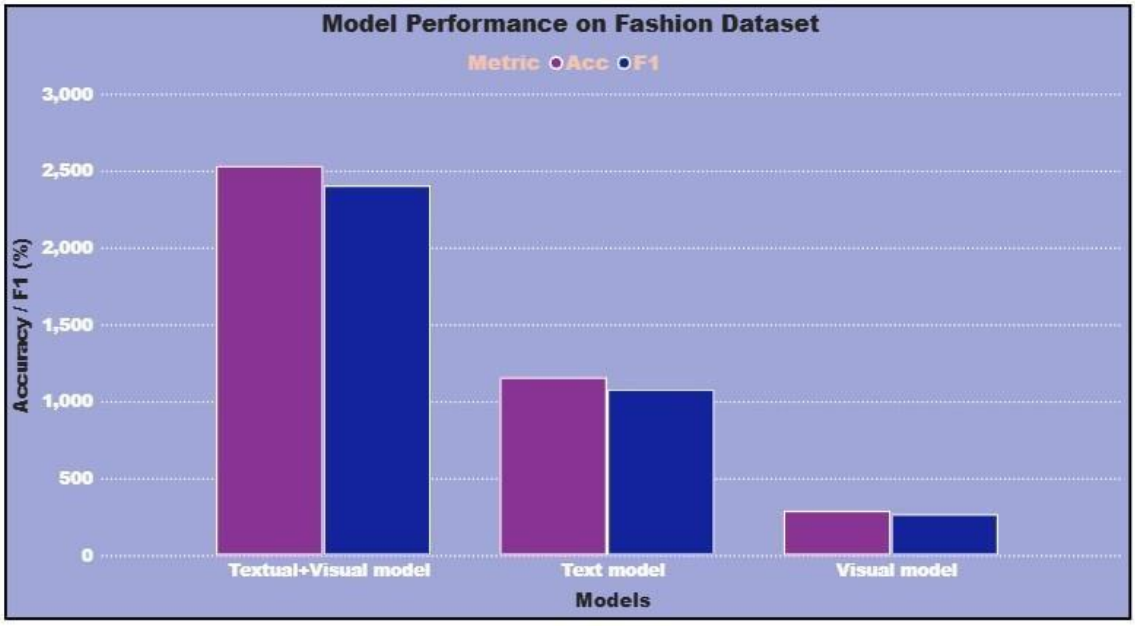

Figure 5: Accuracy and F1-score comparison with fashion dataset

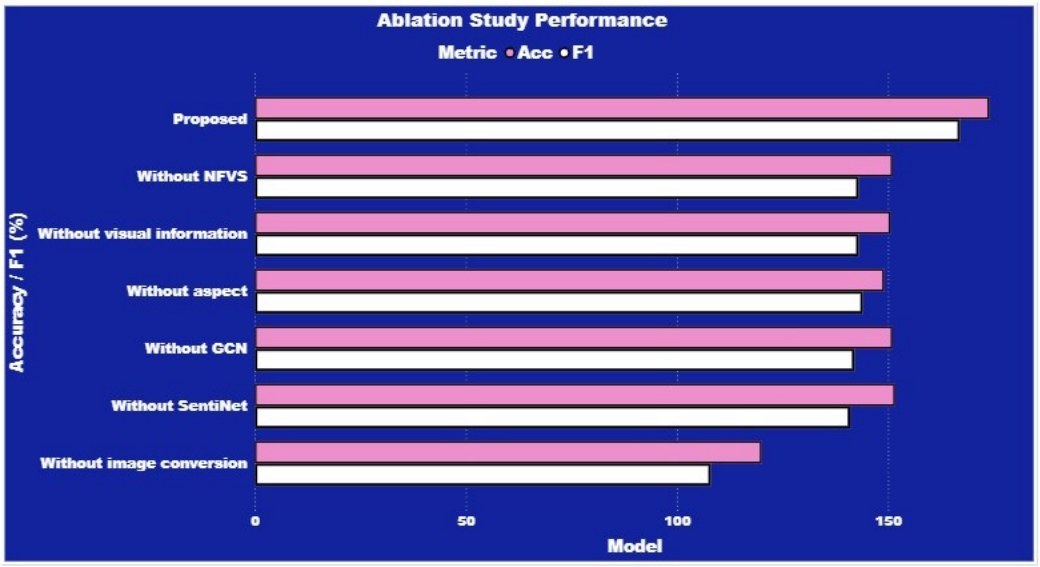

Figure 6: Ablation study based comparison

4.1 Complexity analysis

The model complexity of our suggested method is contrasted with a few example multimodal baselines in this subsection. The results summarize the trainable parameter counts (million), training duration (seconds), and inference time (seconds) on the fashion datasets. All approaches were trained under the experiments were conducted with the same batch size and number of epochs on an NVIDIA RTX 4090 GPU to ensure a fair evaluation. Out of all of these multimodal approaches, HIMT exhibits the highest model parameters and longest

learning duration, likely because of its use of multiple multi-head attention layers for auxiliary reconstruction and multimodal fusion. Table 4 compares the proposed model with our suggested technique has a comparatively larger parameter size and training duration. The observed improvement is most likely due to the incorporation enhancement through multiple knowledge sources and noise filtering from multiple perspectives. Table 4 further demonstrates the efficacy of our approach, indicating that it achieves superior performance compared to HMT and other baselines.

There is no difference between our approach, HMT, and BERT in terms of inference time. These findings lead us to conclude that our suggested approach can achieve a great balance between complexity and performance as in Fig 7.

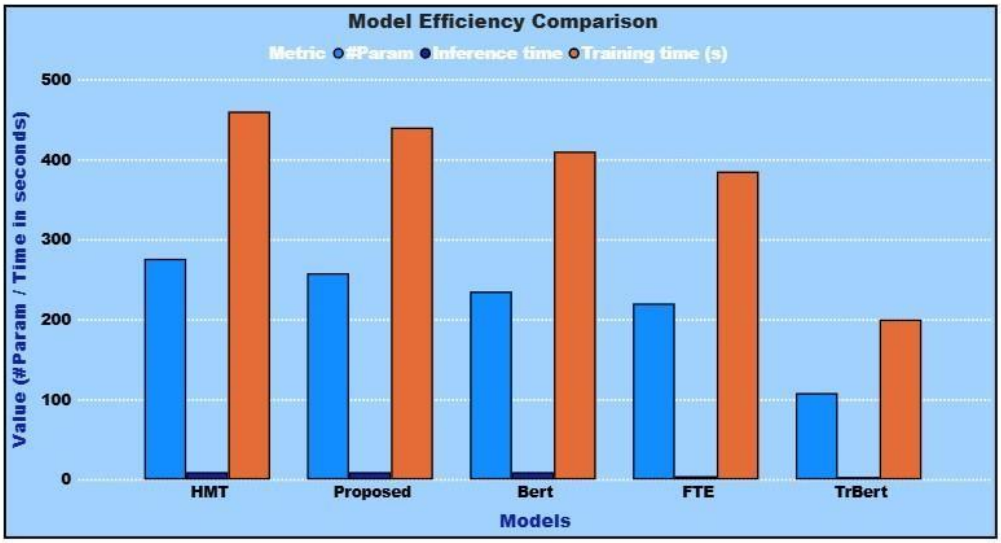

Figure 7: Complexity analysis

Table 4: Complexity analysis

Methods	#Param	Training time (s)	Inference time
TrBert	108	200	3
Bert	235	410	9
FTE	220	385	4
HMT	276	460	9
Proposed	258	440	9

5. Conclusion

To overcome the shortcomings of previous research, we present an improved prompt-based multi-modal sentiment analysis (IPMMSA) in this work that can simultaneously fully utilize multiple knowledge types and efficiently remove noise from three distinct viewpoints to

produce more robust and representative multimodal features. To remove visual noise and retrieve crucial visual semantic information, the IPMMSA component specifically transforms the image is transformed into a series of token embedding's, after which aspect-level visual attention is applied. Ahead of performing attention mechanism oriented toward the aspect eliminate disturbances and maintain important leveraging sentiment cues, the IPMMSA module learns textual semantics guided by visual information and builds a multi-knowledge graph convolutional network (GCN) to enhance image–text semantic representation. Furthermore, to produce more reliable and superior multimodal representations for improving sentiment analysis performance, NFMF seeks to eliminate noise from the multimodal fusion viewpoint. The two well-known fashion datasets have been the subject of several investigations. Relative to the baselines, our method achieves improvements of 98% in accuracy and 90% in F1 score. According to the ablation investigations, our suggested method will degrade performance on F1 by more than 1% for both datasets if any of its components are removed. Consequently, these outcomes show how much better and more successful our suggested method is. Although this study assumes that the previous work, ATE has been finished, many multimodal postings or reviews lack annotated aspect terms due to the time-consuming and arduous nature of human annotation. Our approach will concentrate in the context of joint IPMMSA research in the future that attempts into concurrently identify the aspect phrases and determine the attitudes that correlate to them. Additionally, our future work will include creating more potent unimodal feature extractors leveraging semi-supervised or self-supervised learning techniques as opposed to merely utilizing pre-trained models. It is anticipated that further discriminative unimodal characteristics will be explored by making full use of the vast amount of unlabeled data on online social sites.

Reference

1. Chen YZ, Shi LY, Lin JL, Chen JT, Zhong JY, Dong C (2025) Multi-granularity visual-textual jointly modeling for aspect-level multimodal sentiment analysis. J Supercomput 81(1):46
2. Zhang TJ, Zhou G, Lu JC, Li ZB, Wu H, Liu S (2024) Text-image semantic relevance identification for aspect-based multimodal sen timent analysis. PeerJ Comput Sci 10:e1904
3. Yang D, Li XH, Li Z, Zhou CY, Wang XF, Chen F (2024) Prompt fusion interaction transformer for aspect-based multimodal sen timent analysis. In: Proceedings of the 2024 IEEE international conference on multimedia and expo, pp 1–6
4. Su J, Tang J, Jiang H, Lu Z, Ge Y, Song L et al (2021) Enhanced aspect-based sentiment analysis models with progressive self supervised attention learning. Artif Intell 296:103477
5. SunC,HuangL,QiuX(2019)Utilizingbertforaspect-basedsenti ment analysis via constructing auxiliary sentence. In: Proceedings of NAACL, pp 380–385
6. 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
7. YangJ, XiongYJ (2024)Bidirectional complementary correlation based multimodal aspect-level sentiment analysis. Int J Semant Web Inf Syst 20(1):1–16
8. Nguyen DQ, Vu T, Nguyen AT (2020) Bertweet: a pretrained lan guage model for english tweets. In: Proceedings of EMNLP, pp 9–14
9. HuJ, YangCH, HuangK, WangHJ, PengB, LiTR (2024) Information bottleneck fusion for deep multi-view clustering. Knowl-Based Syst 289:111551
10. Wan Z, Zhang C, Zhu P, Hu Q (2021) Multi-view information bottleneck representation learning. In: Proceedings of AAAI con ference on artificial intelligence, pp 10085–10092
11. Xie JY, Zhu YC, Chen ZZ (2021) Micro-video popularity prediction via multimodal variational information bottleneck. IEEE Trans Multimed 25:24–37
12. Phan HT, Nguyen NT, Hwang D (2023) Aspect-level sentiment analysis: a survey of graph convolutional network methods. Inf Fusion 91:149–172
13. Xiao LW, Wu XJ, Yang SW, Xu JJ, Zhou J, He L (2023) Cross modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. Inf Process Manag 60:103508
14. ZhaoZG, TangMW, ZhaoFJ, ZhangZH, ChenXL(2020) Incorpo rating semantics, syntax and knowledge for aspect-based sentiment analysis. Appl Intell 53(12):16138–16150
15. Mai S, Hu H, Xing S (2020) Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion. In: Proceedings of the 34th AAAI conference on artificial intelligence, pp 164–172
16. Wei Y, Wang X, Guan W, Nie L, Lin Z, Chen B (2020) Neural multimodal cooperative learning toward micro-video understand ing. IEEE Trans Image Process 29:1–14
17. Kumar A, Vepa J (2020) Gated mechanism for attention based multi modal sentiment analysis. In: Proceedings of the IEEE inter national conference on acoustics, speech, and signal processing, pp 4477–4481
18. Wang D, Guo XT, Tian YM, Liu JH, He LH, Luo XM (2023) TETFN: a text enhanced transformer fusion network for multi modal sentiment analysis. Pattern Recogn 136:109259
19. Pandey A, Vishwakarma DK (2024) Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: a survey. Appl Soft Comput 152:111206
20. Liu HY, Chatterjee I, Zhou MC, Lu XYS, Abusorrah A (2020) Aspect-based sentiment analysis: a survey of deep learning meth ods. IEEE Trans Comput Soc Syst 7(6):1358–1375
21. Yang J, Dong XX, Du X (2023) SMFNM: semi-supervised mul timodal fusion network with main-modal for real-time emotion recognition in conversations. J King Saud Univ Comput Inf Sci 35:101791
22. Yu J, Wang J, Xia R, Li J (2022) Targeted multimodal sentiment classification based on coarse-to-fine grained image-target match ing. In: Proceedingsofthethirty-first international joint conference on artificial intelligence, pp 4482–4488
23. Zhao F, Wu Z, Long S, Dai X, Huang S, Chen J (2022) Learn ing from adjective-noun pairs: a knowledge-enhanced framework for target-oriented multimodal sentiment classification. In: Pro ceedings of the 29th international conference on computational linguistics, pp 6784–6794
24. Hou X, Huang J, Wang G, Qi P, He X, Zhou B (2021) Selective attention based graph convolutional networks for aspect-level sen timent classification. In: Proceedings of the ACM 15th workshop graph-based methods natural language processing, pp 83–93
25. An JY, Zainon WMNW, Hao Z (2023) Improving targeted multi modal sentiment classification with semantic description of images. CMC Comput Mater Continua 75(3):5801–5815
26. Yu J, Jiang J, Xia R (2020) Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE ACMTrans Audio Speech Lang Process 28:429–439

27. Liu M, Zhou F, Chen K, Zhao Y (2021) Co-attention networks based on aspect and context for aspect-level sentiment analysis. Knowl-Based Syst 217:106810
28. Bilal H, Ahmed F, Aslam MS, Li QM, Hou J, Yin BQ (2024) A blockchain-enabled approach for privacy-protected data sharing in internet of robotic thingsnetworks.Human-CentricComputInfSci 14:71
29. Bilal H, Obaidat MS, Aslam MS, Zhang J, Yin BQ, Mahmood K (2024) Online fault diagnosis of industrial robot using IoRT and hybrid deep learning techniques: an experimental approach. IEEE Internet Things J 11(19):31422–31437
30. Aslam MS, Bilal H, Band SS, Ghasemi P (2024) Modeling of nonlinear supply chain management with lead-times based on Takagi-Sugeno fuzzy control model. Eng Appl Artif Intell 133:108131