Text-image semantic relevance identification for aspect-based multimodal sentiment analysis

TianZhi Zhang [1], **Gang Zhou** [Corresp., 1], **Jicang Lu** [1], **Zhibo Li** [1], **Hao Wu** [1], **Shuo Liu** [1]

[1] Information Engineering University, Zhengzhou, Henan, China

Corresponding Author: Gang Zhou
Email address: gzhougzhou@126.com

Aspect-Based Multimodal Sentiment Analysis (ABMSA) is an emerging task in the research of multimodal sentiment analysis, which aims to identify the sentiment of each given aspect in text and image. Although recent research on ABMSA has achieved some success, most existing models only use attention mechanism to interact aspect with text and image respectively and obtain sentiment output through multimodal concatenation, they often neglect to consider that some samples may not have semantic relevance between text and image. In this paper, we propose a Text-Image Semantic Relevance Identification (TISRI) model for ABMSA to address the problem. Specifically, we introduce a multimodal feature relevance identification module to calculate the semantic similarity between text and image, and then construct an image gate to dynamically control the input image information. On this basis, an image auxiliary information is provided to enhance the semantic expression ability of visual feature representation to generate more intuitive image representation. Furthermore, we finally employ attention mechanism to obtain the text-aware image representation through text-image interaction to prevent irrelevant image information interfering our model. Experiments demonstrate that TISRI achieves competitive results on two ABMSA Twitter datasets, and then validate the effectiveness of our methods.

Text-Image Semantic Relevance Identification for Aspect-Based Multimodal Sentiment Analysis

Tianzhi Zhang[1], Gang Zhou[1], Jicang Lu[1], Zhibo Li[1], Hao Wu, Shuo Liu[1]

[1] Information Engineering University, Zhengzhou, Henan, China

Corresponding Author:
Gang Zhou[1]
Science Avenue, Zhengzhou, Henan, 450000, China
Email address: gzhougzhou@126.com

Abstract

Aspect-Based Multimodal Sentiment Analysis (ABMSA) is an emerging task in the research of multimodal sentiment analysis, which aims to identify the sentiment of each given aspect in text and image. Although recent research on ABMSA has achieved some success, most existing models only use attention mechanism to interact aspect with text and image respectively and obtain sentiment output through multimodal concatenation, they often neglect to consider that some samples may not have semantic relevance between text and image. In this paper, we propose a Text-Image Semantic Relevance Identification (TISRI) model for ABMSA to address the problem. Specifically, we introduce a multimodal feature relevance identification module to calculate the semantic similarity between text and image, and then construct an image gate to dynamically control the input image information. On this basis, an image auxiliary information is provided to enhance the semantic expression ability of visual feature representation to generate more intuitive image representation. Furthermore, we finally employ attention mechanism to obtain the text-aware image representation through text-image interaction to prevent irrelevant image information interfering our model. Experiments demonstrate that TISRI achieves competitive results on two ABMSA Twitter datasets, and then validate the effectiveness of our methods.

Introduction

With the rapid development of Internet technology, online social and service platforms have gradually become an important part of people's lives(Q. Yu et al., 2019; Fuji and Matsumoto, 2017). Nowadays, the content posted by users is gradually diversified with the prevalence of social media and various service products, and people are more inclined to express their sentiment in multimodal ways such as text and image for different topics and events. Therefore, Multimodal Sentiment Analysis (MSA) task is becoming increasingly important in research communities. Sentiment Analysis (SA) is an effective method to extract valuable information

from massive data(Zhu et al., 2022). As an important fine-grained task in sentiment analysis, Aspect-Based Sentiment Analysis (ABSA) has attracted extensive attention from both academia and industry in the past decade for its ability to detect the sentiment polarity of the specific aspect in data(Zhang et al., 2018; Cao and Huang, 2023).

Aspect-based Multimodal Sentiment Analysis (ABMSA) is a new subtask of ABSA(Pontiki et al., 2016). In this paper, we introduce image as another modality to assist the text semantic expression, and then predict the sentiment polarity of the aspect involved in text and image. Table 1 shows two representative examples: Table 1(a) chooses “Taylor” as the aspect, the text aims to describe a detail about the event of Taylor’s award, and the smiling face in image also helps to identify the aspect “Taylor” as positive sentiment. Table 1(b) chooses “Percy Harvin” and “JETS” as aspects, the text provides a statement of an objective event and focuses all sentimental expressions on the image, we can infer the aspect “Percy Harvin” as a negative sentiment by his shocked expression in image, but “JETS” as an objectively existing organization has no image reflection, so the aspect “JETS” is assigned a neutral sentiment. Overall, ABMSA is a more refined and challenging task compared with global multimodal sentiment analysis, which can capture the sentiment polarity of text internal entities that cannot be obtained in the global tasks.

Given the importance of this field, researchers have proposed numerous methods for ABMSA. For example, Xu et al.(N. Xu et al., 2019) adopted attention mechanism to model interactions between aspect and text, as well as image. Yu et al.(Yu and Jiang, 2019), Yu et al.(J. Yu et al., 2019), and Wang et al.(Wang et al., 2021) further modeled the interactions of text-image, aspect-text, and aspect-image by employing pre-trained language and visual models. These research results demonstrate that integrating image into traditional text sentiment analysis can utilize more comprehensive sentiment information to achieve better sentiment identification effect.

Although MSA and ABSA are already popular research fields today, ABMSA is still a relatively new research task. Employing MSA and ABSA research methods to the ABMSA task may present the following challenges: (1) Some samples in dataset have no semantic relevance between text and image. (2) Compared to text, visual feature representation extracted from image is more difficult to perform semantic expression intuitively. (3) In text-relevant images, there may also be regions that are irrelevant to the text semantics and may introduce additional interference to the model.

To address the above challenges, we propose a general multimodal architecture named Text-Image Semantic Relevance Identification (TISRI) for ABMSA. Compared with traditional ABMSA models, our main contributions to TISRI are summarized as follows:

- To improve the interaction between aspect and text as well as image, we propose a Multimodal Feature Relevance Identification (MFRI) module, which determines the relevance between text and image semantics. Since image is only used as auxiliary information for text here, we construct an image gate to implement dynamic input for image information to prevent irrelevant interference for the model.

• To enhance the semantic expression of the visual feature representation, we construct an Image Feature Auxiliary Reconstruction (IFAR) layer that introduces Adjective-Noun Pairs (ANPs) extracted from each image in our datasets as image auxiliary information. By fine-tuning the semantic bias between image visual representation and image auxiliary information, we can improve the image visual representation in terms of sentiment from a text level.

• To prevent the model being influenced by irrelevant image regions, we further interact text and image representation through attention mechanism in the final multimodal feature fusion, and then obtain text-relevant image representation to achieve Image Feature Filtering (IFF).

Experimental results demonstrate that TISRI outperforms most existing advanced unimodal and multimodal methods, and achieves competitive results on two ABMSA Twitter datasets.

Related Work

Early research on sentiment analysis mainly focused on unimodal sentiment analysis of text(Chen, 2015; Li and Qian, 2016; Shin et al., 2016) and image(You et al., 2017; Li et al., 2018; Wu et al., 2020). In recent years, MSA has gradually become an important focus in sentiment analysis research, and ABMSA has further developed and improved on the basis of ABSA research.

Multimodal Sentiment Analysis (MSA)

In recent years, MSA task has attracted widespread attention in academic community(Cambria et al., 2017; Poria et al., 2020), which aims to model text and other non-text modalities (e.g., visual and auditory modalities), and mainly focuses on two subtasks: MSA for conversation and MSA for social media. In MSA for conversation, existing methods mainly focus on adopting different deep learning models (e.g., Long Short-Term Memory Network(Hochreiter and Schmidhuber, 1997), Gate Recurrent Unit(Chung et al., 2014), and Transformer(Vaswani et al., 2017)) to model the interaction between different modalities, which have demonstrated better performance in various MSA tasks (e.g., sentiment analysis(Zadeh et al., 2017; Poria et al., 2015, 2017; Liang et al., 2018), emotion analysis(Busso et al., 2004; Lee et al., 2011), and sarcasm detection(Castro et al., 2019; Cai et al., 2019)). In MSA for social media, it mainly includes sentiment analysis of social media image(Chen et al., 2014b; You et al., 2015; Yang et al., 2018a, 2018b) and multimodal sentiment analysis of text-image integration(You et al., 2016; Kumar and Garg, 2019; Kumar et al., 2020; Xu et al., 2018). However, the above research methods mainly focus on coarse-grained sentiment analysis (i.e., identifying the global sentiment reflected by each sample) and cannot be directly employed for fine-grained ABMSA tasks.

Aspect-Based Sentiment Analysis (ABSA)

As an important fine-grained sentiment analysis task, ABSA has been widely researched and applied in NLP field over the past decade(Cambria et al., 2017), and its current methods can be broadly divided into two categories: discrete feature-based method and deep learning-based method. Discrete feature-based method focuses on designing multi-specific features to train

learning classifiers for sentiment analysis(Vo and Zhang, 2015; Pontiki et al., 2016). Deep learning-based method mainly adopts various neural network models to encode aspects and corresponding context information, including the method based on Recursive Neural Network(Dong et al., 2014), Convolutional Neural Network(Xue and Li, 2018), Recurrent Neural Network(Ma et al., 2018; Chen et al., 2017), Attention Mechanism(Wang et al., 2018; Yang et al., 2019; Meškelė and Frasincar, 2020; Zhao et al., 2021), Graph Convolutional Network(Wang et al., 2020; Zhang and Qian, 2020), and pre-trained BERT model that has achieved great success in recent years(H. Xu et al., 2019; Sun et al., 2019). However, the above research methods mainly focus on text-based unimodal information, but do not take into account the fact that relevant information from other modalities (e.g., visual modality) can also contribute to sentiment analysis.

Aspect-Based Multimodal Sentiment Analysis (ABMSA)

To conduct research on ABSA utilizing information from different modalities, researchers have developed numerous models for ABMSA over the past three years by employing various effective methods in different tasks. Xu et al.(N. Xu et al., 2019) first explored the ABMSA task and proposed a multi-interactive memory network model MIMN based on BiLSTM for text-image interaction, while also constructed an e-commerce comment dataset for ABMSA. Yu et al.(Yu and Jiang, 2019) proposed an ABMSA model TomBERT based on the BERT architecture, and manually constructed two ABMSA Twitter datasets. Yu et al.(J. Yu et al., 2019) proposed an ABMSA model ESAFN based on entity-sensitive attention and fusion network. Khan et al.(Khan and Fu, 2021) proposed a novel model CapBERT that employs cross-modal transformation to convert the image content into text caption, and performs final sentiment analysis solely based on text modality. Wang et al.(Wang et al., 2021) proposed a recurrent attention network SaliencyBERT also based on BERT, the network effectively captures both intra-modal and inter-modal dynamics by designing a recurrent attention mechanism. Although the above research methods have been validated to be effective in the ABMSA task, they often neglect to identify whether the semantics between modalities are relevant or not. To address this problem, our model captures the semantic relevance between modalities by calculating the similarity between text and image features, which facilitates the effective development of its subsequent work.

Methodology

In this chapter, we first formulate our task, and introduce the overall architecture of our Text-Image Semantic Relevance Identification (TISRI) model, then delve into the details of each module in TISRI.

Task Formulation: Given a set of multimodal samples $D = (x_1, x_2, \ldots, x_d)$ as input, each sample $x_i \in D$ contains an m-word text $S = (w_1, w_2, \ldots, w_m)$, an associated image I, and an n-word aspect $T = (w_1, w_2, \ldots, w_n)$ that is a word subsequence of S. Our task is to predict the sentiment

label $y \in Y$ of each given aspect, where Y consists of three categories: positive, negative, and neutral.

Overview

Figure 1 illustrates the overall architecture of TISRI, which contains the following modules: (1) Unimodal Feature Extraction Module. (2) Multimodal Feature Relevance Identification Module. (3) Aspect-Multimodal Feature Interaction Module. (4) Multimodal Feature Fusion Module. As shown at the bottom of Fig. 1, for a given multimodal sample, we first extract word feature representations from the input text and aspect, respectively, and visual feature representation from the input image, then aspect representation interacts with text and image representation to generate aspect-aware text representation and aspect-aware image representation. Next, we obtain the semantic similarity between text and image by constructing a multimodal feature relevance identification module. The overall method is shown in Fig. 2, where the fusion representations of text and image are obtained through cross-modal interaction, and then an image gate is constructed in a specific way to dynamically control the input image information. To enable better semantic expression of image feature, we propose an image feature auxiliary reconstruction layer. As shown in Fig. 3, the image visual representation is fine-tuned by introducing Adjective-Noun Pairs (ANPs) extracted from each image in our datasets as image auxiliary information to minimize their representation differences. Finally, to prevent the model being influenced by irrelevant image regions, we interact aspect-aware text representation with aspect-aware image representation, and then generate the final image representation. As shown at the top of Fig. 1, we further concatenate the aspect-aware text representation and the final image representation, and obtain the final sentiment label through a sentiment analysis linear layer.

Unimodal Feature Extraction Module

In this module, we adopt two pre-trained models to extract unimodal feature representations from aspect, text and image, respectively.

Aspect and Text Representation

Given an input text, we divide it into two parts: aspect T and its corresponding context C, and replace the aspect position in C with a special character "T". For text encoding, we employ pre-trained language model RoBERTa(Liu et al., 2019) as the text encoder of our model, which has been proven to achieve competitive performance in various NLP tasks including ABSA(Dai et al., 2021). For T and C, we follow the implementation mechanism of RoBERTa by inserting two special tokens into each input (i.e., "<s>" at the beginning and "</s>" at the end), and then feeding them into text encoder to obtain the hidden representations of aspect: $H_T = \text{RoBERTa}(T)$ and context: $H_C = \text{RoBERTa}(C)$, respectively, where $H_T \in ¡^{d \times t}$ and $H_C \in ¡^{d \times c}$, d is the hidden dimension, t is the length of aspect, and c is the length of context.

Next, we concatenate C with T as sentence S. For S, we use the token "</s>" to separate C from T, and then obtain the hidden representation of sentence: $H_S = \text{RoBERTa}(S)$ through RoBERTa implementation mechanism, where $H_S \in \mathbb{R}^{d \times s}$, $s = c + t$ is the length of sentence. The implementations of aspect, context, and sentence encoding are shown at the bottom of Fig. 1 and Fig. 2.

Image Representation

For image encoding, we employ Residual Network (ResNet)(He et al., 2016) as the image encoder of our model. Compared to the previous VGG network(Simonyan and Zisserman, 2014), ResNet uses residual connections to avoid gradient vanishing problems as the number of layers increases, which allows for deeper extraction of semantic information in image recognition tasks. Specifically, given an input image I, we first resize it to I' with 224×224 pixels, and then take the output of the last convolutional layer in pre-trained 152-layer ResNet as the image visual representation: $H_I = \text{ResNet}(I')$, where $H_I \in \mathbb{R}^{2048 \times 49}$, 49 is the number of visual blocks with the same size by dividing I' into 7×7, and 2048 is the vector dimension of each visual block. Since we will conduct cross-modal interaction with text and image to obtain the feature representation of text and image fusion, it is necessary to project image representation to the same semantic space as text representation. We employ a linear transformation function for H_I to obtain the final image representation: $H_V = W_I^T H_I$, where $W_I^T \in \mathbb{R}^{2048 \times d}$ is the learnable parameter. The implementation of image encoding is shown at the bottom of Fig. 1.

Multimodal Feature Relevance Identification Module

For images in multimodal samples, while they can provide information beyond text for sentiment analysis, our purpose is to use image to assist in analyzing the sentiment polarity of aspect in text, and images that are irrelevant to text semantics may lead to misalignment of aspects and introduce additional interference to the model. Therefore, we propose a Multimodal Feature Relevance Identification (MFRI) Module, which provides an image gate when integrating text and image, and dynamically controls the input image information based on its relevance to text semantics. MFRI is divided into two layers: (1) Text-Image Cross-Modal Interaction Layer. (2) Image Gate Construction Layer. As shown in Fig. 2, we provide a detailed introduction to the implementation methods of these two layers in the following sections.

Text-Image Cross-Modal Interaction Layer

To better learn sentence feature representation in image, we introduce a multi-head cross-modal attention mechanism (MC-ATT)(Tsai et al., 2019), which treats image representation H_V as query, and sentence representation H_S as key and value, then involves two layer normalization (LN)(Ba et al., 2016) and a feedforward network (FFN)(Vaswani et al., 2017) as follows:

$$Z_{V \to S} = \text{LN}(H_V + \text{MC-ATT}(H_V, H_S)) \quad (1)$$

$$H_{V\to S} = \mathrm{LN}(Z_{V\to S} + \mathrm{FFN}(Z_{V\to S})) \quad (2)$$

where $H_{V\to S} \in ¡^{d\times 49}$ is the image-aware sentence representation generated by MC-ATT layer. However, image representation is treated as query in the above MC-ATT layer, and each vector in the generated $H_{V\to S}$ represents a visual block rather than a word representation in sentence. We expect that image-aware sentence representation can reflect on each word in sentence. Given this problem, we introduce another MC-ATT layer that treats H_S as query, and $H_{V\to S}$ as key and value, then generates the final image-aware sentence representation $H'_{V\to S}$, where $H'_{V\to S} \in ¡^{d\times s}$.

To obtain the image representation for each word in sentence, we adopt the same method as above for cross-modal interaction, treating H_S as query, and H_V as key and value, then generating the sentence-aware image representation $H_{S\to V}$, where $H_{S\to V} \in ¡^{d\times s}$.

Image Gate Construction Layer

Yu et al.(Yu et al., 2020) introduced visual gate to dynamically control the contribution of image visual features to each word in text in the multimodal named entity recognition work and achieved effective experimental results. Inspired by this work, we construct a gate for the input image information, which is responsible for dynamically controlling the contribution of image information in our model by assigning a weight in $[0,1]$ to each image based on its relevance to corresponding sentence, preserving the higher relevance image by assigning a higher weight, and filtering the lower relevance image by assigning a lower weight. Specifically, we first concatenate $H'_{V\to S}$ and $H_{S\to V}$, and then construct the gate based on text-image relevance weight through linear transformation and nonlinear activation function:

$$g = \sigma(W_{S\to V}[H'_{V\to S}; H_{S\to V}]) \quad (3)$$

where $W_{S\to V} \in ¡^{d\times 2d}$ is the learnable parameter, σ is the element-wise nonlinear activation function, which is used to control the output of g in $[0,1]$.

Based on the above image gate g, we can obtain the final image representation that assigns relevance weight:

$$H'_V = g \cdot H_{S\to V} \quad (4)$$

Aspect-Multimodal Feature Interaction Module

After obtaining the feature representations of aspect, context, and image, we analyze the relationships between aspect and image as well as context, respectively. Furthermore, we design an Image Feature Auxiliary Reconstruction (IFAR) Layer, which serves as an auxiliary supervision for visual representation. The specific technical scheme of this module is shown at the middle part of Fig. 1, and the internal architecture of IFAR Layer is shown in Fig. 3. We provide a detailed implementation methods for them in the following sections.

Aspect Interaction Layer

The main purpose of this layer is to obtain aspect-aware image representation and aspect-aware context representation, so we employ MC-ATT layer to interact with aspect and image as well as context, respectively, to promote information integration between modalities. Specifically, we first conduct cross-modal feature interaction between aspect and image, treating aspect representation H_T as query, and image representation H_V as key and value:

$$Z_{T\to V} = \mathrm{LN}(H_T + \text{MC-ATT}(H_T, H_V)) \tag{5}$$

$$H_{T\to V} = \mathrm{LN}(Z_{T\to V} + \mathrm{FFN}(Z_{T\to V})) \tag{6}$$

where $H_{T\to V} \in ¡^{d\times t}$ is the aspect-aware image representation generated by MC-ATT layer. Similarly, we can also obtain the aspect-aware context representation $H_{T\to C}$, where $H_{T\to C} \in ¡^{d\times t}$.

Image Feature Auxiliary Reconstruction Layer

To improve the effectiveness of visual feature representation, we introduce Adjective-Noun Pairs (ANPs) extracted from the image in each sample. Since the nouns and adjectives in ANPs can reflect real content and sentiment in image to some extent, we employ them as auxiliary supervision for visual representation to obtain a more intuitive image semantic expression. Specifically, we adopt DeepSentiBank(Chen et al., 2014a) to generate 2089 ANPs for each image and select the top k ANPs as image auxiliary information.

However, the extraction of image ANPs is essentially a coarse-grained extraction method, so extracted ANPs may be the content of image regions that are irrelevant to aspect or may be semantic information that is incorrectly recognized for image, and directly using these ANPs can significantly introduce additional interference to the model due to their inaccuracy. Zhao et al.(Zhao et al., 2022) obtained nouns relevant to aspect by calculating semantic similarity between aspect representation and ANPs noun representation in the construction of ABMSA knowledge enhancement framework, and achieved excellent alignment effect in their experiment. Inspired by this work, we concatenate the above k ANPs and their corresponding nouns, respectively, and feed them into text encoder to obtain the ANPs representation H_{ANPs} and the noun representation H_N, and then we employ cosine similarity to calculate the semantic similarity between H_T and H_N to achieve the aspect alignment:

$$\alpha = \frac{H_T^T \cdot H_N}{\|H_T\| \cdot \|H_N\|} \tag{7}$$

where α is the similarity score between H_T and H_N, which we use as a weight vector representing the semantic relevance of ANPs to the aspect expressed in image. Next, we assign each individual in ANPs representation with its corresponding relevance weight to obtain the image auxiliary information representation:

$$H'_{ANPs} = \alpha \cdot H_{ANPs} \tag{8}$$

Furthermore, based on the construction of image gate g in the above Multimodal Feature Relevance Identification Module, we also treat g as image auxiliary information gate to dynamically control the contribution of ANPs to the model, and then obtain the final image auxiliary information representation:

$$H''_{ANPs} = g \cdot H'_{ANPs} \tag{9}$$

To enable visual attention to be more intuitive and accurate in representing the visual features of aspect in image, we introduce a reconstruction loss function based on mean square error (MSE) to minimize the difference between aspect-aware image representation $H_{T\to V}$ and final image auxiliary information representation H''_{ANPs}:

$$\mathsf{L}_R = \frac{1}{|D|}\sum_{i=1}^{|D|}(H''_{ANPs} - H_{T\to V})^2 \tag{10}$$

Multimodal Feature Fusion Module

In this module, we fuse aspect-aware context representation $H_{T\to C}$ and aspect-aware image representation $H_{T\to V}$ with our Image Feature Filtering (IFF) method to obtain the final aspect output representation. The implementation is shown at the top of Fig. 1. First, we employ MC-ATT to implement the interaction between $H_{T\to C}$ and $H_{T\to V}$ to obtain the visual feature representation corresponding to aspect-aware context in aspect-aware image as the final aspect-aware image representation, and then filter the irrelevant regions in image:

$$Z_{T\to C\to V} = \mathrm{LN}(H_{T\to C} + \text{MC-ATT}(H_{T\to C}, H_{T\to V})) \tag{11}$$

$$H_{T\to C\to V} = \mathrm{LN}(Z_{T\to C\to V} + \mathrm{FFN}(Z_{T\to C\to V})) \tag{12}$$

Next, we concatenate $H_{T\to C}$ and $H_{T\to C\to V}$, and then feed them into a multimodal self-attention layer based on Transformer for feature fusion between modalities:

$$H = \text{Transformer}(H_{T\to C}; H_{T\to C\to V}) \tag{13}$$

Finally, we feed the first token representation H^0 into Softmax layer to obtain the final sentiment label:

$$P(y|H) = \text{Softmax}(W^T H^0) \tag{14}$$

We adopt the cross entropy loss constructed by predicted values of aspect-based sentiment labels and their true values as the training loss function for model sentiment analysis task:

$$\mathsf{L}_S = \frac{1}{|D|}\sum_{j=1}^{|D|}\log P(y^j|H^0) \tag{15}$$

To further optimize all parameters of our model, we train the loss function for sentiment analysis jointly with image auxiliary reconstruction, and then construct a final training loss function combining the two tasks:

$$L = L_S + \lambda L_R \quad (16)$$

Where λ is the tradeoff hyper-parameter used to control the contribution of reconstruction loss.

Experiment

In this chapter, we conduct extensive experiments on two ABMSA datasets to validate the effectiveness of our Text-Image Semantic Relevance Identification (TISRI) model.

Experimental Settings

Datasets: We adopt two benchmark datasets of ABMSA TWITTER-2015 and TWITTER-2017 proposed by Yu et al.(Yu and Jiang, 2019) that are composed of multimodal tweets posted on TWITTER in 2014-2015 and 2016-2017, where each sample consists of a text, an image, a given aspect, and the sentiment label (positive, negative, and neutral) corresponding to the aspect. The relevant information of these two datasets is shown in Table 2.

Implementation Details: For TISRI, we adopt RoBERTa-base(Liu et al., 2019) as the encoder for sentence, context, and aspect in text, and ResNet-152(He et al., 2016) as the image encoder. During alternating optimization process, we use AdamW as the learner to optimize parameters. Specifically, we set the batch size to 16, the training epoch to 9, the k value to 5, the λ value to 0.8, the model learning rate to 1e-5, the maximum length of sentence and context to 128, the maximum length of aspect to 32, and the hidden dimension to 768. We demonstrate the average results of three independent training runs for all our models. All the models are implemented based on PyTorch, and run on an NVIDIA Tesla V100 GPU.

Compared Baselines

In this section, we evaluate the performance of TISRI by comparing it with various existing methods. Specifically, we consider comparing the following unimodal and multimodal methods to our model:

- Res-Target: a baseline method for obtaining the visual feature representation of input image directly from the ResNet model.
- AE-LSTM(Wang et al., 2016): an attention-based LSTM model for obtaining important context relevant to aspect.
- MGAN(Fan et al., 2018): a multi-grained attention network that fuses aspect and context at different granularity.
- BERT(Devlin et al., 2018): a pre-trained language model with stacked Transformer encoder layers for the interaction between aspect and text.
- RoBERTa(Liu et al., 2019): further improvement of BERT model by adopting better training strategies and larger corpus.
- MIMN(N. Xu et al., 2019): a multi-interactive memory network for the interaction between aspect, text, and image.

- ESAFN(J. Yu et al., 2019): an entity-sensitive attention and fusion network for obtaining inter-modal dynamics of aspect, text, and image.
- ViLBERT(Lu et al., 2019): a pre-trained visual language model that takes aspect-text pairs as input text.
- TomBERT(Yu and Jiang, 2019): an aspect-aware ABMSA method based on multimodal BERT model architecture.
- SaliencyBERT(Wang et al., 2021): a recursive attention network based on multimodal BERT model architecture for ABMSA.
- CapBERT(Khan and Fu, 2021): a method of converting image into text caption and feeding it with the input text to a pre-processed BERT model.
- KEF-TomBERT(Zhao et al., 2022): an extended baseline to apply a proposed knowledge enhancement framework KEF to TomBERT.
- KEF-SaliencyBERT(Zhao et al., 2022): an extended baseline to apply a proposed knowledge enhancement framework KEF to SaliencyBERT.
- CapRoBERTa: an extended baseline that replaces BERT with RoBERTa in CapBERT.
- KEF-TomRoBERTa: an extended baseline that replaces BERT with RoBERTa in KEF-TomBERT.

Experimental Results and Analysis

Table 3 demonstrates the performance of our model and each compared baseline model on TWITTER-2015 and TWITTER-2017 datasets. We adopt Accuracy (Acc) and Macro-F1 as evaluation metrics and mark the best score for each metric in bold. As shown at the last five columns of Table 3, we compare our model with the latest proposed best performing KEF-TomBERT and KEF-SaliencyBERT last year. In addition, we also select the best performing CapBERT from original baseline model and better performing KEF-TomBERT from the above two models, and replace the BERT in them with RoBERTa to implement a more comprehensive and fair comparison of TISRI.

Based on all the experimental results in Table 3, we can conclude as follows: (1) The performance of Res-Target is lower than that of all text language models, which may be explained by the fact that image relevant to aspect mostly serve as an auxiliary role for text and do not perform well as an independent modality for sentiment prediction. (2) Most multimodal methods generally perform better than unimodal methods, which indicates that image information can complement text information to obtain a higher sentiment prediction ability. (3) TomBERT, SaliencyBERT and CapBERT perform much better than other multimodal models, and we speculate that adopting multi-head cross-modal attention with self-attention mechanism to do cross-modal interaction on aspect can obtain more robust feature representation. (4) Among all original baseline models, CapBERT achieves the best performance due to image caption, which indicates that text has a more intuitive semantic representation than image. (5) The performance of KEF-TomBERT and KEF-SaliencyBERT is better than that of other original

baseline models, which indicates that the knowledge enhancement framework KEF can improve the performance of original model by introducing image adjective and noun information to some extent and has excellent compatibility effect. (6) Since RoBERTa is more powerful than BERT, intuitively the overall performance of CapRoBERTa is generally better than that of CapBERT on the above evaluation metrics. (7) Compared to the best performing KEF-TomRoBERTa, our model achieves competitive results on the two datasets, which has about 0.5% higher Macro-F1 on TWITTER-2015 dataset, and about 0.4% and 1.2% higher Accuracy and Macro-F1 on TWITTER-2017 dataset, respectively.

For the slightly lower accuracy of our model on TWITTER-2015 dataset compared to KEF-TomRoBERTa, we speculate that the possible reason is that KEF-TomRoBERTa applies adjectives in the obtained ANPs directly to aspect-aware image representation, while the overall text-image relevance weights on TWITTER-2015 dataset may be relatively higher than those on TWITTER-2017 dataset, which is also validated in the TISRI w/o MFRI part of ablation study in Section 4.4. Therefore, the direct use of adjectives in this case can express the sentiment in image more intuitively to some extent. However, for the condition where ANPs identify semantic error in image or text-image relevance has a low weight, KEF-TomRoBERTa may introduce additional interference to the model by directly using irrelevant adjectives. Overall, we speculate that TISRI performs better on TWITTER-2017 dataset for this reason.

Ablation Study

To further investigate the impact of individual unit in TISRI on model performance, we perform ablation analysis on TWITTER-2015 and TWITTER-2017 datasets for several important units in the model: (1) Image Feature Filtering (IFF) method. (2) Image Feature Auxiliary Reconstruction (IFAR) Layer. (3) Multimodal Feature Relevance Identification (MFRI) Module. We first remove the above three units respectively, and then remove these units at the same time leaving only the bone framework, so we can have a clearer and more comprehensive understanding of the contribution of individual unit to the model performance improvement. The experimental results are shown in Table 4, where w/o represents the removal of the corresponding unit.

First, we can learn that removing IFF unit decreases Accuracy by about 1.9% and 1.5% on the two datasets, respectively, which validates that retaining useful information in image and implementing filtering on text-irrelevant image regions helps reduce the impact of interference on model performance. Next, removing IFAR unit decreases Accuracy by about 2.1% and 2.5% on the two datasets, respectively, which proves that the unit has a large contribution to model performance improvement and validates that adopting ANPs as image auxiliary information can be more intuitive for semantic expression of visual feature representation. Then, removing the MFCR unit decreases Accuracy by about 0.7% and 1.8% on the two datasets, respectively, which validates that assigning image to an inter-modal relevance weight can help to prevent additional interference to the model from text-irrelevant images. We can also learn that there are more images with higher text-image relevance weights in TWITTER-2015 dataset than in TWITTER-

2017 dataset, which validates the reason we inferred in Section 4.3. Finally, we remove all above units and observe that Accuracy decreases by about 2.2% and 4.5% on the two datasets, respectively, which validates the effectiveness of our proposed units in the model and also validates that these units contribute to model performance improvement to some extent from another perspective.

Parameter Analysis

In this section, we provide a detailed introduction and analysis of the process of evaluating optimal hyper-parameters. All of the above experiments are set based on optimized model hyper-parameters.

Values of Epoch and Batch Size

To analyze the impact of different epoch and batch size on model performance, we determine the final values of epoch and batch size through experiments in this subsection. Figure 4 and Figure 5 demonstrate the model performance of different epoch and batch size values on the two datasets, respectively, and we can draw the following inferences. First, we experiment with the value of epoch. We find that as the value of epoch increases, model performance shows an upward trend and then gradually stabilizes. The model performance is optimal when epoch equals 8, and then Accuracy and Macro-F1 of the model start to gradually decrease when epoch equals 9. Thus, we set the value of epoch to 9 in experiment. Accuracy and Macro-F1 corresponding to the epoch setting of TISRI on TWITTER-2015 and TWITTER-2017 datasets are shown in Fig. 4(a) and Fig. 4(b).

Then, we analyze the value of batch size using 8, 16 and 32, respectively, and experimental results on the two datasets are shown in Fig. 5(a) and Fig. 5(b). We can clearly find that the model achieves the best performance on both datasets when batch size equals 16. The possible reasons are speculated as follows: When batch size equals 8, it is small for the number of samples in the two datasets, and the training of model is not only time-consuming but also difficult to converge, which leads to the underfitting of model. In a certain range, the increase of batch size is conducive to the stability of model convergence. However, when batch size equals 32, the model may fall into local minimum because it is too large, which leads to the deterioration of model generalization performance. Thus, we set the value of batch size to 16 in experiment.

Value of k

To explore the impact of ANPs on model performance, we extract the top k ANPs for each image where k is set as each integer in $[1,10]$, and take values for them respectively to experiment. Figure 6(a) and Figure 6(b) demonstrate the model performance of k value on the two datasets, respectively, and we can draw the following inferences. First, the model performance is poor without ANPs as image auxiliary information, which indicates that combining ANPs can improve the performance of TISRI. Second, the model performance shows

a fluctuating upward trend as the number of ANPs increases and reaches the best state when k equals 5. However, the model performance no longer improves but shows a trend of decline when k is greater than 5. The possible reason is speculated as follows: The number of aspects involved in each text in the two datasets may not exceed 5, and when k is greater than it, image auxiliary information may introduce additional interference to the model. Therefore, we set the value of k to 5 in experiment.

Value of λ

To investigate the effect of trade-off hyper-parameter λ that controls the auxiliary reconstruction loss contribution of IFAR layer on model performance, we set λ to a decimal number with an interval of 0.1 in the range of $[0,1]$ to experiment. Figure 7(a) and Figure 7(b) demonstrate the model performance of λ value on the two datasets, respectively. The model performance shows a fluctuating upward trend as λ increases, which has a more obvious effect on TWITTER-2017 dataset. When λ equals 0.8, the model performance reaches the best state, and then decreases gradually as λ increases. We speculate that the possible reason is that ANPs as image auxiliary information only serve to improve the semantic expression of image visual features. When the trade-off hyper-parameter λ exceeds a certain value, image auxiliary information plays a dominant role in image representation, but these ANPs may have semantic information of image recognition error, so the model will largely introduce additional interference when λ is too large and produce negative effect. Thus, we set the trade-off hyper-parameter λ to 0.8 in the error back propagation process.

Case Study

In this section, we provide an in-depth analysis of the results of different models on TWITTER-2015 and TWITTER-2017 datasets to better understand the advantages of our model. Specifically, we first select four samples from test datasets to compare the sentiment prediction performance of TISRI with other models, and then screen the samples for error analysis in TISRI and analyze the possible causes of their errors.

Prediction Results

Table 5 demonstrates the comparison results of the sentiment prediction performance of three models RoBERTa, CapRoBERTa, and TISRI on four samples where we have an advantage. Since our model uses image gate and ANPs to assist with image information, we demonstrate them in the table as well. Table 5(a) demonstrates that image has a higher relevance weight with text, and the noun “team” associated with the aspect “Thunder” in ANPs has positive words “excellent” and “victorious” as modifiers, so our model can accurately predict the sentiment polarity as positive, while CapRoBERTa gives a wrong prediction. In Table 5(b), the image also has a higher relevance weight with text, and the ANPs contain positive words such as “handsome” and “smile”, so our model also makes correct prediction, while CapRoBERTa predicts a wrong sentiment label. However, the image content in Table 5(c) is relatively

complex, and old black and white photo also has certain limitations in image recognition, so the text-image relevance weight is not high. Fortunately, the ANPs identify nouns such as “team” relevant to image content, and also have positive adjectives like “successful” as noun modifiers, so our model successfully predict correct sentiment label, while RoBERTa and CapRoBERTa give wrong predictions. In Table 5(d), the text-image relevance weight is slightly lower than the other three samples because the text is too short and less relevant to image content, but the ANPs have several positive words such as “hot”, “pretty”, and “sexy”, which help our model to make accurate prediction from another perspective, while CapRoBERTa predicts a wrong sentiment label.

However, we find that multimodal CapRoBERTa model makes all wrong predictions in these four samples, while unimodal RoBERTa model makes only one wrong prediction, which is unreasonable from model interpretability perspective. Through investigation, we learn that the image caption in Table 5(a) is “A man in a tennis outfit is jumping in the air.”, Table 5(b) is “A woman with a tie and a flower in her hand.”, Table 5(c) is “A group of baseball players standing next to each other.” and Table 5(d) is “Two women in a field with a dog.”. We find that these captions not only contain incorrect recognition, but fail to reflect the adjectives or nouns relevant to facial expression in Table 5(b) and Table 5(d). Furthermore, CapRoBERTa completely relies on image caption to obtain image representation but discards original image information, so it cannot accurately reflect the sentiment embodied in image to some extent, and then affect the final sentiment prediction.

Error Analysis

On the basis of the above experiments, we further perform error analysis on TISRI to deepen our understanding of model performance. Table 6 demonstrates three types of error prediction examples, including the following categories: (1) The ANPs are incorrect in image semantic recognition. (2) The aspect in text cannot find nouns with high similarity in ANPs. (3) The model cannot recognize deeper sentiment in text and image. First, the ANPs in Table 6(a) incorrectly recognize image semantics. For an image that does not reflect any positive sentiment, its ANPs produce positive words such as “laughing” and “funny” that completely hinder correct sentiment recognition, so the image representation is affected by these words. Then, the aspect “WILD Women” in Table 6(b) is actually an organization name. Since it cannot be represented intuitively in image causing ANPs recognizing some aspect-irrelevant nouns, the image cannot accurately express the sentiment semantics of aspect. Finally, the text in Table 6(c) states an objective event that Martin St. Louis announced his retirement, and the image demonstrates a moment he waved his hand on the field. However, our model can only identify semantic features on the surface of text and image but cannot feel Martin's unwillingness to leave the stadium, so this problem is also the difficulty for TISRI to further intelligently identify sentiment in the future.

Conclusions

In this paper, we propose an Aspect-Based Multimodal Sentiment Analysis (ABMSA) model TISRI. First, the model calculates text-image semantic relevance and constructs an image gate that dynamically controls the input of image information. Then, it introduces Adjective-Noun Pairs (ANPs) as image auxiliary information to enhance the semantic expression ability of image visual features. Finally, we adopt attention mechanism to interact with text and image representation to obtain filtered text-relevant image representation for the final sentiment prediction. Experimental results demonstrate that our proposed model outperforms the majority of existing advanced models on TWITTER-2015 dataset and all compared baseline models on TWITTER-2017 dataset, and validate the superiority of our model and the effectiveness of our methods.

We plan to expand our future research in the following directions. First, we aim to apply TISRI to more multimodal related tasks, where our inter-modal feature relevance identification and image feature auxiliary semantic enhancement units can be easily extended to other tasks such as multimodal event extraction and multimodal named entity recognition. Moreover, with the prevalence of large model, we aim to further explore how to effectively integrate large model into our work and achieve more specific multiclassification tasks in subsequent research.

Acknowledgements

This research is supported by the Science and Technology Research Program of the Department of Science and Technology of Henan Province (approval No.: 222102210081)

References

Yu Q, Zhou J, Gong W. 2019. A lightweight sentiment analysis method. *ZTE Communications* 17: 2–8. DOI: 10.12142/ZTECOM.201903002.

Fuji R, Matsumoto K. 2017. Emotion analysis on social big data. *ZTE Communications* 15: 30–37. DOI: 10.3969/j.issn.1673-5188.2017.S2.005.

Zhu L, Xu M, Bao Y, Xu Y, Kong X. 2022. Deep learning for aspect-based sentiment analysis: a review. *PeerJ Computer Science* 8: e1044. DOI: 10.7717/peerj-cs.1044.

Zhang L, Wang S, Liu B. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8: e1253. DOI: 10.1002/widm.1253

Cao F, Huang X. 2023. Performance analysis of aspect-level sentiment classification task based on different deep learning models. *PeerJ Computer Science* 9: e1578. DOI: 10.7717/peerj-cs.1578.

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, AL-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In: *ProWorkshop on Semantic Evaluation (SemEval-2016)*. 19–30. DOI: 10.18653/v1/S16-1002.

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

Yu J, Jiang J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Main track*. 5408-5414. DOI: 10.24963/ijcai.2019/751.

Yu J, Jiang J, Xia R. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28: 429–439. DOI: 10.1109/TASLP.2019.2957872.

Wang J, Liu Z, Sheng V, Song Y, Qiu C. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In: *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4. Springer*. 3–15. DOI: 10.1007/978-3-030-88010-1_1.

Chen Y. 2015. Convolutional Neural Network for Sentence Classification. *University of Waterloo*.

Li D, Qian J. 2016. Text sentiment analysis based on long short-term memory. In: *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*. 471–475. DOI: 10.1109/CCI.2016.7778967.

Shin B, Lee T, Choi J.D. 2016. Lexicon integrated CNN models with attention for sentiment analysis. In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.149-158. DOI: 10.18653/v1/W17-5220.

You Q, Jin H, Luo J. 2017. Visual sentiment analysis by attending on local image regions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 31(1). DOI: 10.1609/aaai.v31i1.10501.

Li Z, Fan Y, Liu W, Wang F. 2018. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications* 77: 1115–1132. DOI: 10.1007/s11042-016-4310-5.

Wu L, Qi M, Jian M, Zhang H. 2020. Visual sentiment analysis by combining global and local information. *Neural Processing Letters* 51: 2063–2075. DOI: 10.1007/s11063-019-10027-7.

Cambria E, Das D, Bandyopadhyay S, Feraco A. 2017. Affective computing and sentiment analysis. *A practical guide to sentiment analysis*. 1–10. DOI: 10.1109/MIS.2016.31.

Poria S, Hazarika D, Majumder N, Mihalcea R. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*. 108 - 132. DOI: 10.1109/TAFFC.2020.3038167.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural computation* 9: 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Chung J, Gulcehre C, Cho K, Bengio Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in neural information processing systems*. 6000–6010. DOI: 10.5555/3295222.3295349.

Zadeh A, Chen M, Poria S, Cambria E, Morency L-P. 2017. Tensor fusion network for multimodal sentiment analysis. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114. DOI: 10.18653/v1/D17-1115.

Poria S, Cambria E, Gelbukh A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2539–2544. DOI: 10.18653/v1/D15-1303.

Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P. 2017. Context-dependent sentiment analysis in user-generated videos. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 873–883. DOI: 10.18653/v1/P17-1081.

Liang PP, Liu Z, Zadeh A, Morency L-P. 2018. Multimodal language analysis with recurrent multistage fusion. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 150-161. DOI: 10.18653/v1/D18-1014.

Busso C, Deng Z, Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Lee S, Neumann U, Narayanan S. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Proceedings of the 6th International Conference on Multimodal Interfaces*. 205–211. DOI: 10.1145/1027933.1027968.

Lee C-C, Mower E, Busso C, Lee S, Narayanan S. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*. 53(9-10): 1162-1171. DOI: 10.1016/j.specom.2011.06.004.

Castro S, Hazarika D, Pérez-Rosas V, Zimmermann R, Mihalcea R, Poria S. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic*. 4619-4629. DOI: 10.18653/v1/P19-1455.

Cai Y, Cai H, Wan X. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2506–2515. DOI: 10.18653/v1/P19-1239.

Chen T, Yu FX, Chen J, Cui Y, Chen Y-Y, Chang S-F. 2014b. Object-based visual sentiment concept analysis and application. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. 367–376. DOI: 10.1145/2647868.2654935.

You Q, Luo J, Jin H, Yang J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 29(1). DOI: 10.1609/aaai.v29i1.9179.

Yang J, She D, Sun M, Cheng M-M, Rosin PL, Wang L. 2018b. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*. 20(9): 2513-2525. DOI: 10.1109/TMM.2018.2803520.

693 Yang J, She D, Lai Y-K, Rosin PL, Yang M-H. 2018a. Weakly supervised coupled networks for
694 visual sentiment analysis. In: *Proceedings of the IEEE Conference on Computer Vision and*
695 *Pattern Recognition*. 7584–7592. DOI: 10.1109/CVPR.2018.00791.
696 You Q, Cao L, Jin H, Luo J. 2016. Robust visual-textual sentiment analysis: When attention
697 meets tree-structured recursive neural networks. In: *Proceedings of the 24th ACM International*
698 *Conference on Multimedia*. 1008–1017. DOI: 10.1145/2964284.2964288.
699 Kumar A, Garg G. 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and*
700 *Applications*. 78: 24103–24119. DOI: 10.1007/s11042-019-7390-1.
701 Kumar A, Srinivasan K, Cheng W-H, Zomaya AY. 2020. Hybrid context enriched deep learning
702 model for fine-grained sentiment analysis in textual and visual semiotic modality social data.
703 *Information Processing & Management*. 57: 102141. DOI: 10.1016/j.ipm.2019.102141.
704 Xu N, Mao W, Chen G. 2018. A co-memory network for multimodal sentiment analysis. In: *The*
705 *41st International ACM SIGIR Conference on Research & Development in Information*
706 *Retrieval*. 929–932. DOI: 10.1145/3209978.3210093.
707 Vo D-T, Zhang Y. 2015. Target-dependent twitter sentiment classification with rich automatic
708 features. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 1347–1353.
709 DOI: 10.5555/2832415.2832437.
710 Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K. 2014. Adaptive recursive neural network for
711 target-dependent twitter sentiment classification. In: *Proceedings of the 52nd Annual Meeting of*
712 *the Association for Computational Linguistics (Volume 2: Short Papers)*. 49–54. DOI:
713 10.3115/v1/P14-2009.
714 Xue W, Li T. 2018. Aspect based sentiment analysis with gated convolutional networks. In:
715 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*
716 *(Volume 1: Long Papers)*. 2514-2523. DOI: 10.18653/v1/P18-1234.
717 Ma Y, Peng H, Cambria E. 2018. Targeted aspect-based sentiment analysis via embedding
718 commonsense knowledge into an attentive LSTM. In: *Proceedings of the AAAI Conference on*
719 *Artificial Intelligence*. 32(1). DOI: 10.1609/aaai.v32i1.12048.
720 Chen P, Sun Z, Bing L, Yang W. 2017. Recurrent attention network on memory for aspect
721 sentiment analysis. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural*
722 *Language Processing*. 452–461. DOI: 10.18653/v1/D17-1047.
723 Wang S, Mazumder S, Liu B, Zhou M, Chang Y. 2018. Target-sensitive memory networks for
724 aspect sentiment classification. In: *Proceedings of the 56th Annual Meeting of the Association*
725 *for Computational Linguistics (Volume 1: Long Papers)*. 957–967. DOI: 10.18653/v1/P18-1088.
726 Yang C, Zhang H, Jiang B, Li K. 2019. Aspect-based sentiment analysis with alternating
727 coattention networks. *Information Processing & Management*. 56: 463–478. DOI:
728 10.1016/j.ipm.2018.12.004.
729 Meškelė D, Frasincar F. 2020. ALDONAr: A hybrid solution for sentence-level aspect-based
730 sentiment analysis using a lexicalized domain ontology and a regularized neural attention model.
731 *Information Processing & Management*. 57: 102211. DOI: 10.1016/j.ipm.2020.102211.

Zhao L, Liu Y, Zhang M, Guo T, Chen L. 2021. Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model. *Information Processing & Management*. 58: 102641. DOI: 10.1016/j.ipm.2021.102641.

Wang K, Shen W, Yang Y, Quan X, Wang R. 2020. Relational graph attention network for aspect-based sentiment analysis. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3229-3238. DOI: 10.18653/v1/2020.acl-main.295.

Zhang M, Qian T. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3540–3549. DOI: 10.18653/v1/2020.emnlp-main.286.

Xu H, Liu B, Shu L, Yu PS. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1. DOI: 10.18653/v1/N19-1242.

Sun C, Huang L, Qiu X. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 380-385. DOI: 10.18653/v1/N19-1035.

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*. 3034–3042. DOI: 10.1145/3474085.3475692.

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

Dai J, Yan H, Sun T, Liu P, Qiu X. 2021. Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1816-1829. DOI: 10.18653/v1/2021.naacl-main.146.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. DOI: 10.1109/CVPR.2016.90.

Simonyan K, Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R. 2019. Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the conference. Association for Computational Linguistics*. 6558-6569. DOI: 10.18653/v1/P19-1656.

Ba JL, Kiros JR, Hinton GE. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yu J, Jiang J, Yang L, Xia R. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3342-3352. DOI: 10.18653/v1/2020.acl-main.306.

Chen T, Borth D, Darrell T, Chang S-F. 2014a. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

Zhao F, Wu Z, Long S, Dai X, Huang S, Chen J. 2022. Learning from Adjective-Noun Pairs: A Knowledge-enhanced Framework for Target-Oriented Multimodal Sentiment Classification. In: *Proceedings of the 29th International Conference on Computational Linguistics*. 6784–6794.

Wang Y, Huang M, Zhu X, Zhao L. 2016. Attention-based LSTM for aspect-level sentiment classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 606–615. DOI: 10.18653/v1/D16-1058.

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*. 3433–3442. DOI: 10.18653/v1/D18-1380.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lu J, Batra D, Parikh D, Lee S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13-23. DOI: 10.5555/3454287.3454289.

Table 1(on next page)

Representative examples of ABMSA.

Given an image, a text, and an unspecified number of aspects, we aim to predict the sentiment polarity of each aspect.

Table 1: Representative examples of ABMSA. Given an image, a text, and an unspecified number of aspects, we aim to predict the sentiment polarity of each aspect.

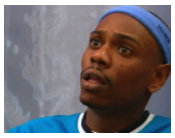
	Image	Text	Aspect	Output
(a)		**Taylor** posing with her Taylor Swift Award at the # BMIPopAwards	Taylor	(Taylor, Positive)
(b)		RT @ ESPN Numbers : Everyone reacting to **Percy Harvin** being traded to the **JETS** . . .	Percy Harvin JETS	(Percy Harvin, Negative) (JETS, Neutral)

Table 2(on next page)

The basic statistics of two TWITTER datasets.

Table 2: The basic statistics of two TWITTER datasets.

	TWITTER-2015			TWITTER-2017		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Positive	928	303	317	1508	515	493
Negative	368	149	113	416	144	168
Neutral	1883	670	607	1638	517	573
Total	3179	1122	1037	3562	1176	1234
Avg Aspects	1.348	1.336	1.354	1.410	1.439	1.450
Words	9023	4238	3919	6027	2922	3013
Avg Length	16.72	16.74	17.05	16.21	16.37	16.38

Table 3(on next page)

Experimental results on TWITTER-2015 and TWITTER-2017 datasets using different unimodal and multimodal methods in the ABMSA task.

Table 3: Experimental results on TWITTER-2015 and TWITTER-2017 datasets using different unimodal and multimodal methods in the ABMSA task.

	TWITTER-2015		TWITTER-2017	
Method	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
		Image Only		
Res-Target	59.88	46.48	58.59	53.98
		Text Only		
AE-LSTM	70.30	63.43	61.67	57.97
MGAN	71.17	64.21	64.75	61.46
BERT	74.15	68.86	68.15	65.23
RoBERTa	76.28	71.36	69.77	68.00
		Text and Image		
MIMN	71.84	65.69	65.88	62.99
ESAFN	73.38	67.37	67.83	64.22
ViLBERT	73.76	69.85	67.42	64.87
TomBERT	77.15	71.75	70.34	68.03
SaliencyBERT	77.03	72.36	69.69	67.19
CapBERT	78.01	73.25	69.77	68.42
KEF-TomBERT	78.68	73.75	72.12	69.96
KEF-SaliencyBERT	78.15	73.54	71.88	68.96
CapRoBERTa	77.82	73.38	71.07	68.57
KEF-TomRoBERTa	**78.75**	73.94	72.18	70.21
TISRI (Ours)	78.50	**74.42**	**72.53**	**71.40**

Table 4(on next page)

Ablation study of TISRI.

Table 4: Ablation study of TISRI.

	TWITTER-2015		TWITTER-2017	
Method	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
TISRI	**78.50**	**74.42**	**72.53**	**71.40**
TISRI w/o IFF	76.57	72.22	71.07	69.74
TISRI w/o IFAR	76.37	72.54	70.02	68.24
TISRI w/o MFRI	77.82	73.85	70.75	68.76
TISRI w/o IFF & IFAR & MFRI	76.28	71.79	68.07	66.86

Table 5(on next page)

Case study of RoBERTa, CapRoBERTa, and TISRI.

✓ and ✗ denote the correct and incorrect predictions, respectively.

Table 5: Case study of RoBERTa, CapRoBERTa, and TISRI. ✓ and ✗ denote the correct and incorrect predictions, respectively.

Image				
Text	(a) OKC evens the series ! Kevin Durant ' s 41 points lead **[Thunder]**$_{\text{Positive}}$ to 111 - 97 victory over Spurs in Game 4 .	(b) @ Soundkartell did an Interview with @ **[tomklose]**$_{\text{Positive}}$ at @ spotfestival and it was very kind . We talked about @ SpotifyDE	(c) RT @ juventusfcen : Two special memories # OnThisDay : a **[UEFA Cup]**$_{\text{Positive}}$ title in 1977 and our 16th Scudetto in 1975 .	(d) Some of that Dodger baseball ☀ □ @ **[alyssajacinto]**$_{\text{Positive}}$
Image Gate	0.703	0.648	0.559	0.478
Top-k ANPs	clean air holy cross excellent team tough race victorious team	handsome smile christian heritage stupid face handsome kid clean teeth	poor performance successful team holy angels fresh meat fancy dress	stunning beauty hot girls pretty girls dark skin sexy girls
Label	(Thunder, Positive)	(tomklose, Positive)	(UEFA Cup, Positive)	(alyssajacinto, Positive)
RoBERTa	(Thunder, Positive✓)	(tomklose, Positive✓)	(UEFA Cup, Neutral✗)	(alyssajacinto, Positive✓)
CapRoBERTa	(Thunder, Neutral✗)	(tomklose, Neutral✗)	(UEFA Cup, Neutral✗)	(alyssajacinto, Neutral✗)
TISRI (Ours)	(Thunder, Positive✓)	(tomklose, Positive✓)	(UEFA Cup, Positive✓)	(alyssajacinto, Positive✓)

Table 6(on next page)

Error cases of TISRI.

Table 6: Error cases of TISRI.

Image			
Text	(a) Petition to have [**Jessica Lange**]$_{\text{Neutral}}$ come back for American Horror Story season 6	(b) This morning @ SheilaGCraft hosted a brunch amp poured into our [**WILD Women**]$_{\text{Positive}}$ to honor them for their leadership in 2014 !	(c) RT @ NYRangers : OFFICIAL : [**Martin**]$_{\text{Negative}}$ St . Louis announces retirement from the National Hockey League . # NYR
Image Gate	0.590	0.494	0.433
Top-k ANPs	laughing baby crazy cat crazy face poor cat funny baby	awesome cake little tree colorful cake great food jolly christmas	excited crowd ill child excited student holy cross amazing race
Label	(Jessica Lange, Neutral)	(WILD Women, Positive)	(Martin, Negative)
TISRI (Ours)	(Jessica Lange, Positive ✗)	(WILD Women, Neutral ✗)	(Martin, Neutral ✗)

Figure 1

The overview of Text-Image Semantic Relevance Identification (TISRI) model architecture.

TISRI consists of four modules: Unimodal Feature Extraction Module, Multimodal Feature Relevance Identification Module, Aspect-Multimodal Feature Interaction Module, and Multimodal Feature Fusion Module.

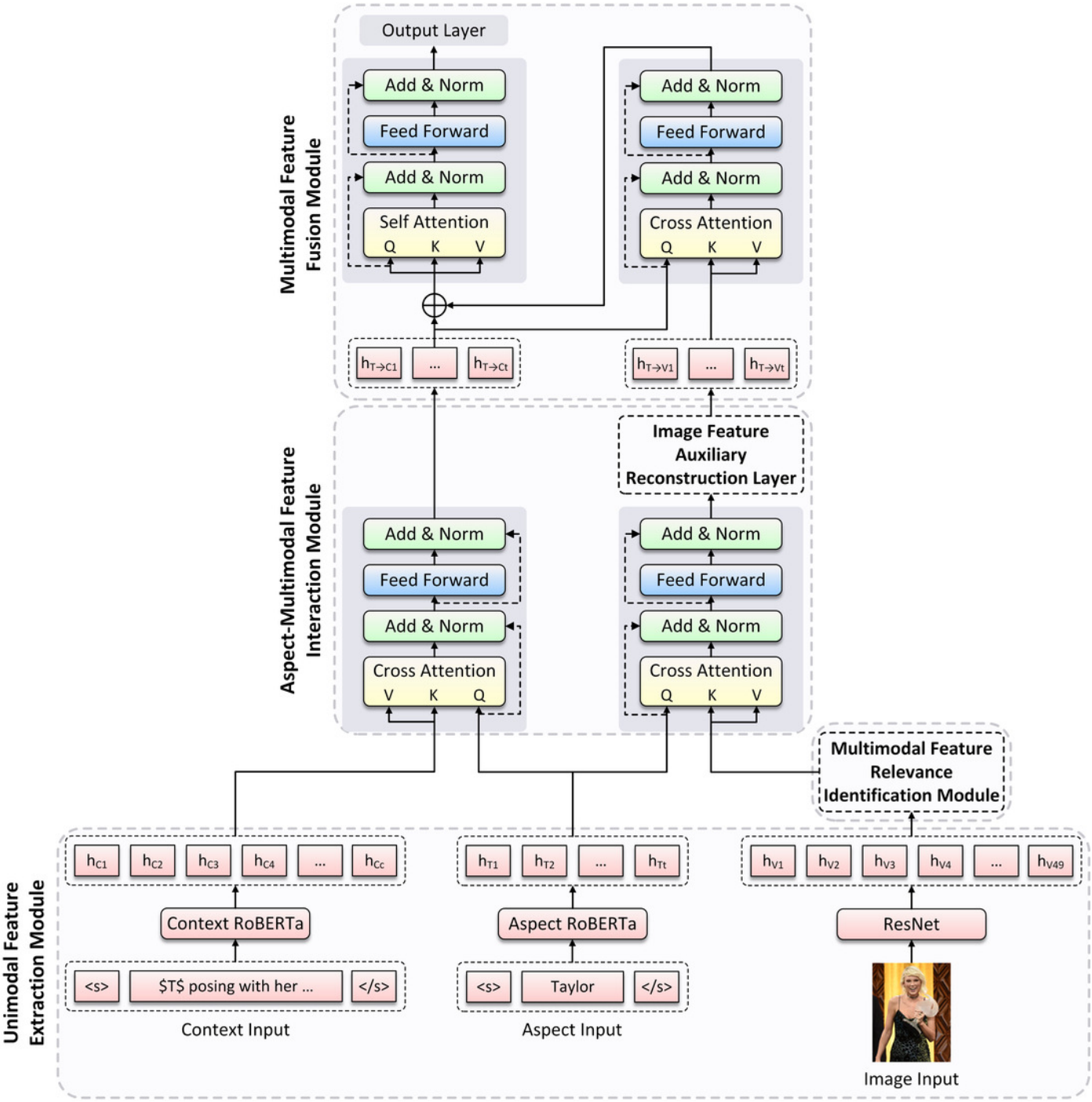

Output Layer
Multimodal Feature Fusion Module
Add & Norm
Feed Forward
Add & Norm
Self Attention
Q K V
Add & Norm
Cross Attention
Q K V
$h_{T\to C1}$ ... $h_{T\to Ct}$
$h_{T\to V1}$ ... $h_{T\to Vt}$
Aspect-Multimodal Feature Interaction Module
Image Feature Auxiliary Reconstruction Layer
Add & Norm
Feed Forward
Add & Norm
Cross Attention
V K Q
Add & Norm
Feed Forward
Add & Norm
Cross Attention
Q K V
Multimodal Feature Relevance Identification Module
Unimodal Feature Extraction Module
h_{C1} h_{C2} h_{C3} h_{C4} ... h_{Cc}
h_{T1} h_{T2} ... h_{Tt}
h_{V1} h_{V2} h_{V3} h_{V4} ... h_{V49}
Context RoBERTa
Aspect RoBERTa
ResNet
<s> T posing with her ... </s>
<s> Taylor </s>
Context Input
Aspect Input
Image Input

Figure 2

The overview of Multimodal Feature Relevance Identification (MFRI) Module architecture.

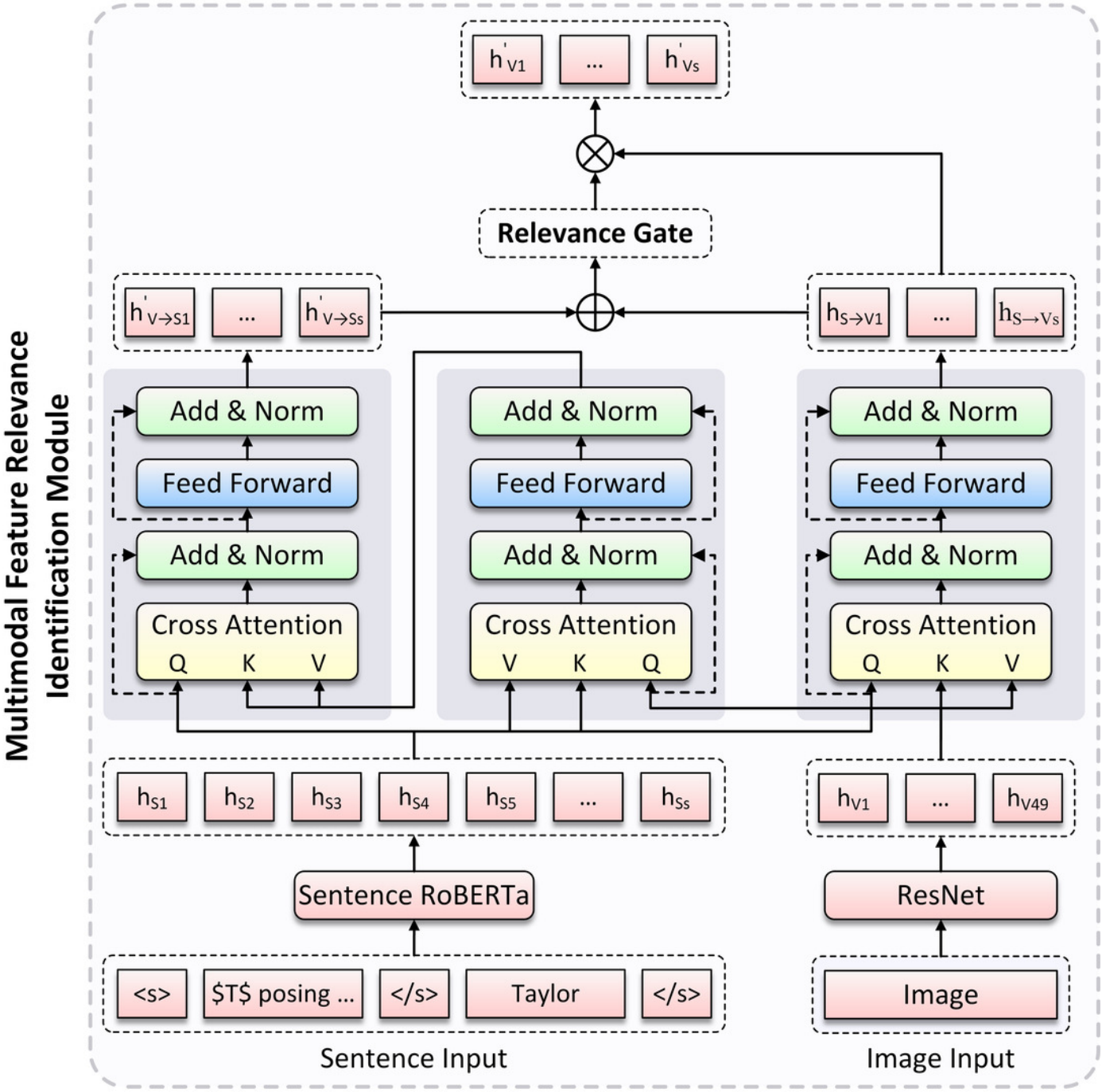


Figure 3

The overview of Image Feature Auxiliary Reconstruction (IFAR) Layer architecture.

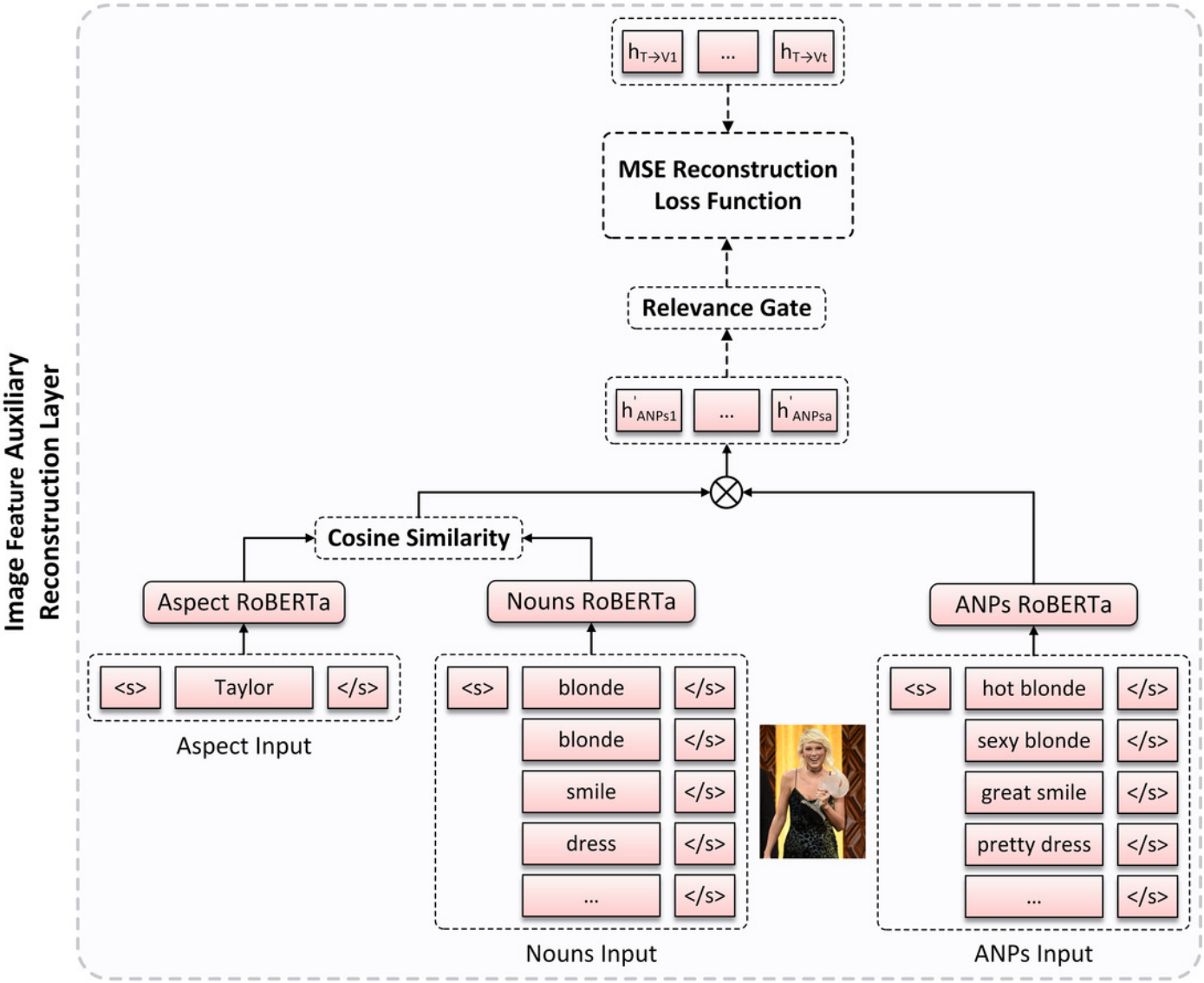


Figure 4

Effect of epoch on TWITTER-2015

Effect of epoch on model Accuracy and Macro-F1.

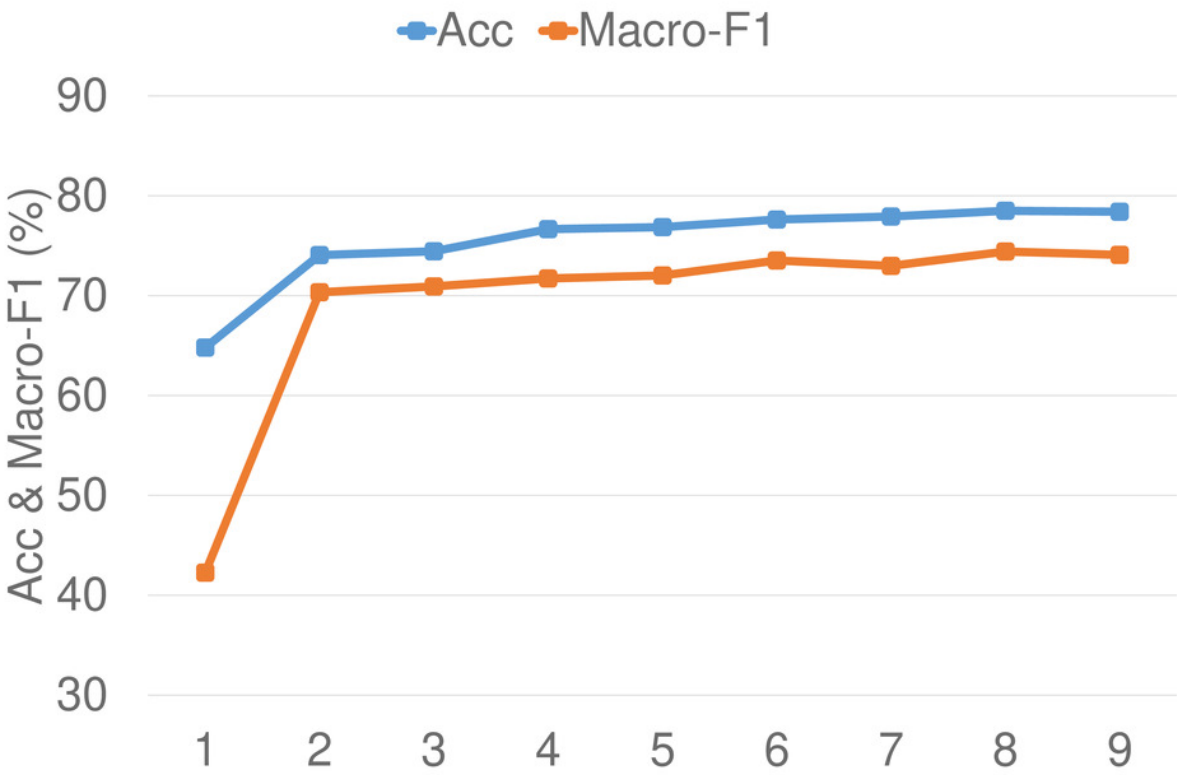


Figure 5

Effect of epoch on TWITTER-2017

Effect of epoch on model Accuracy and Macro-F1.

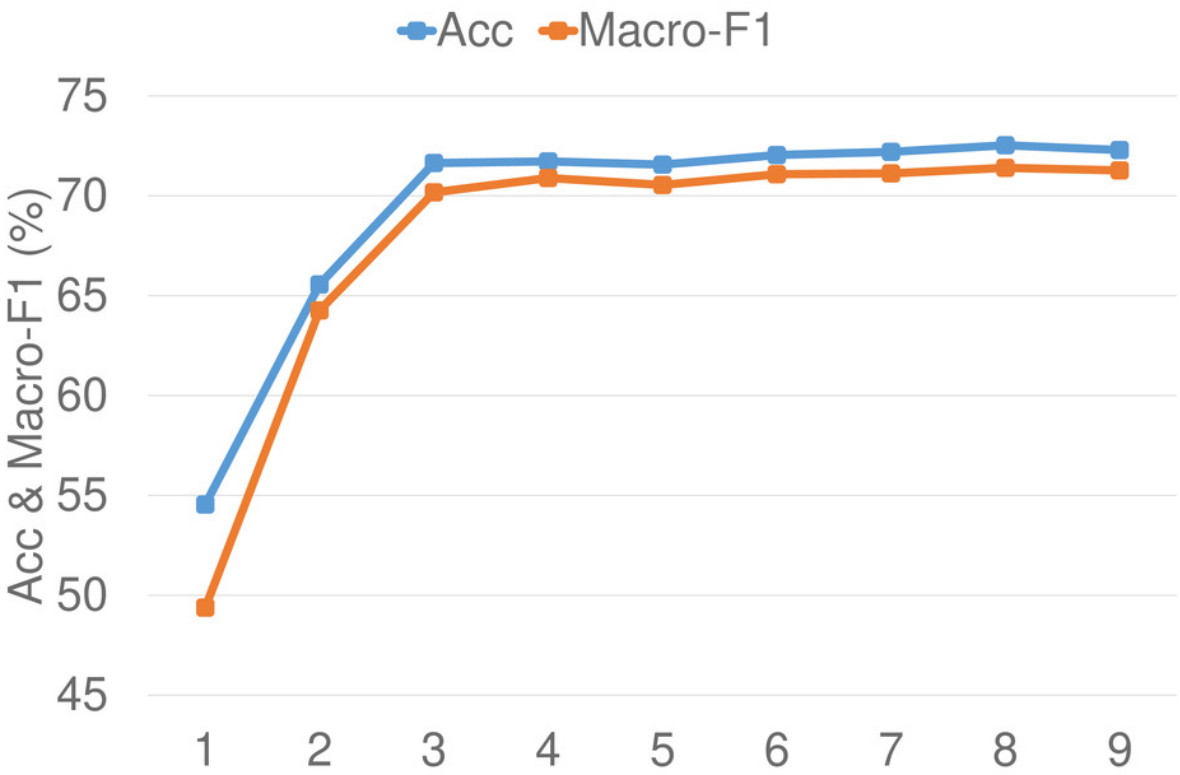


Figure 6

Effect of batch size on TWITTER-2015

Effect of batch size on model Accuracy and Macro-F1.

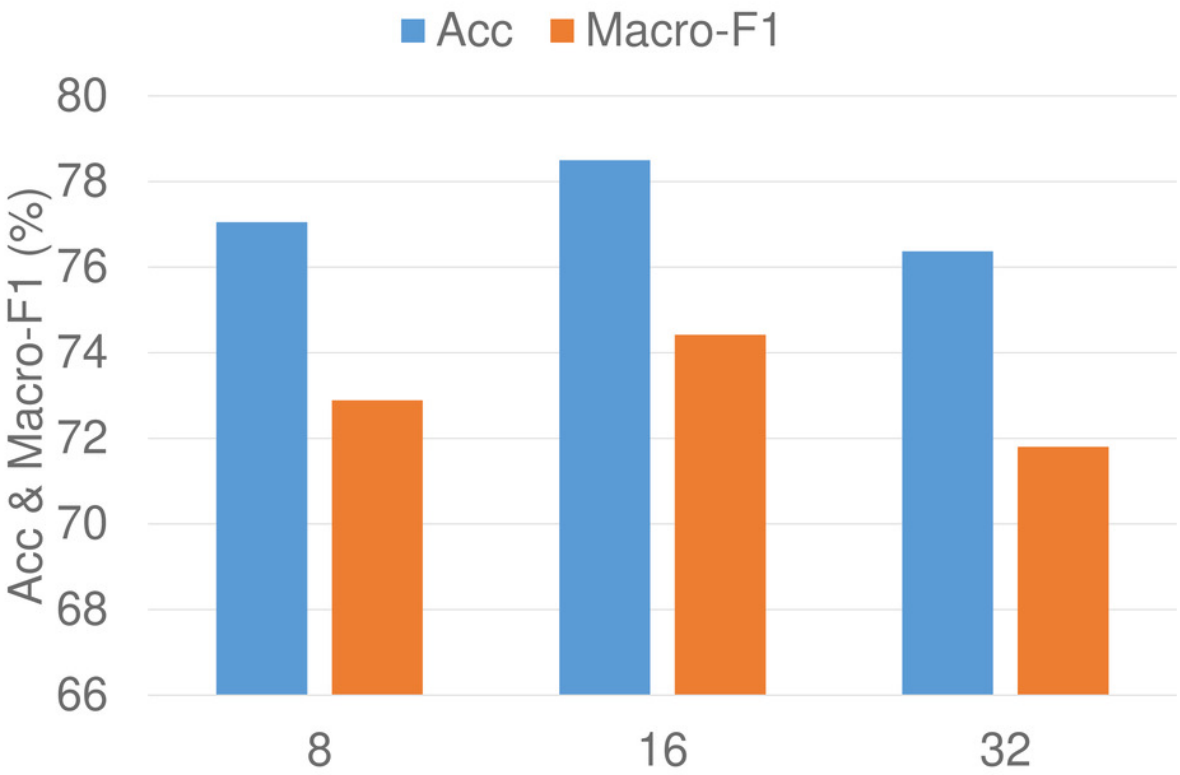


Figure 7

Effect of batch size on TWITTER-2017

Effect of batch size on model Accuracy and Macro-F1.

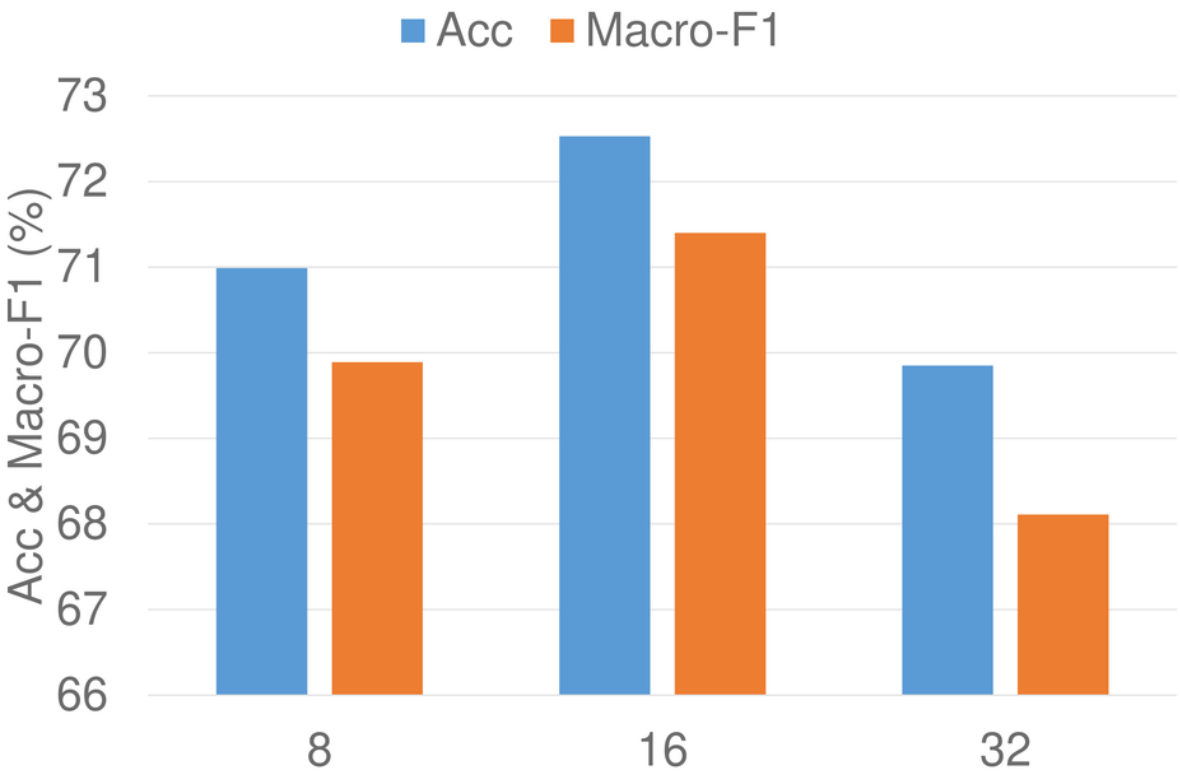


Figure 8

Effect of k on TWITTER-2015

Effect of k on model Accuracy and Macro-F1.

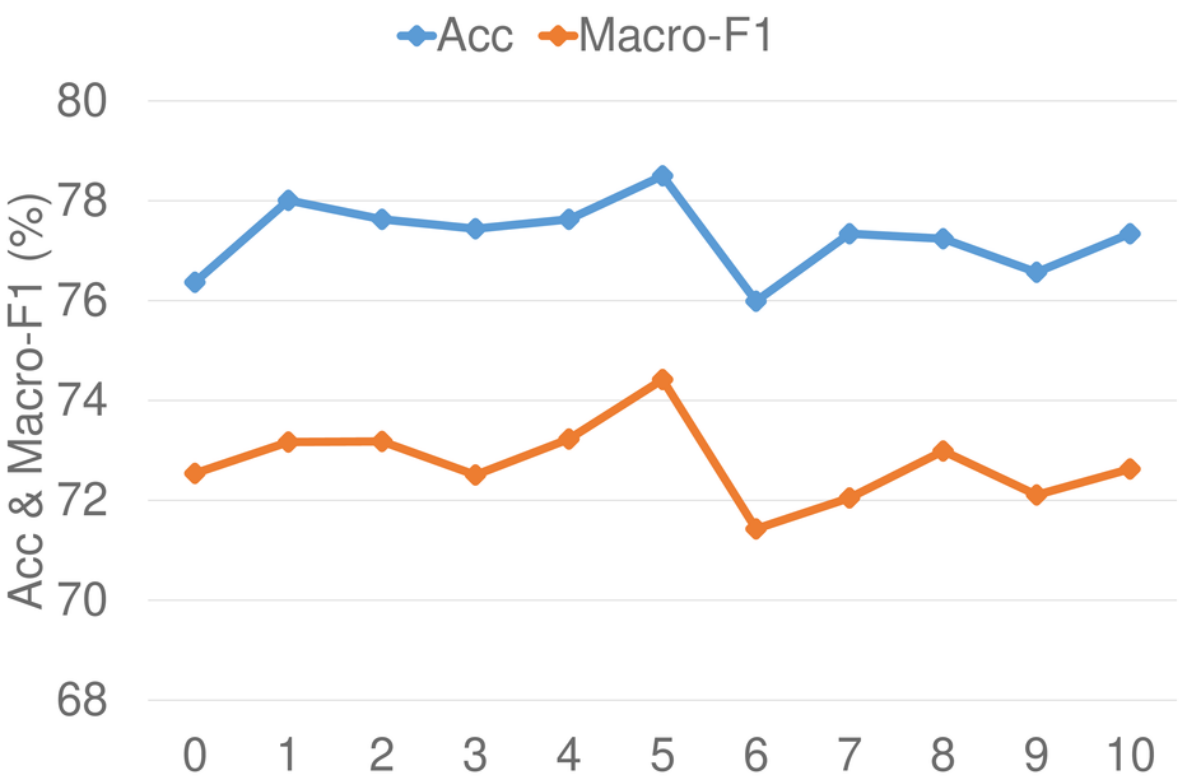


Figure 9

Effect of k on TWITTER-2017

Effect of k on model Accuracy and Macro-F1.

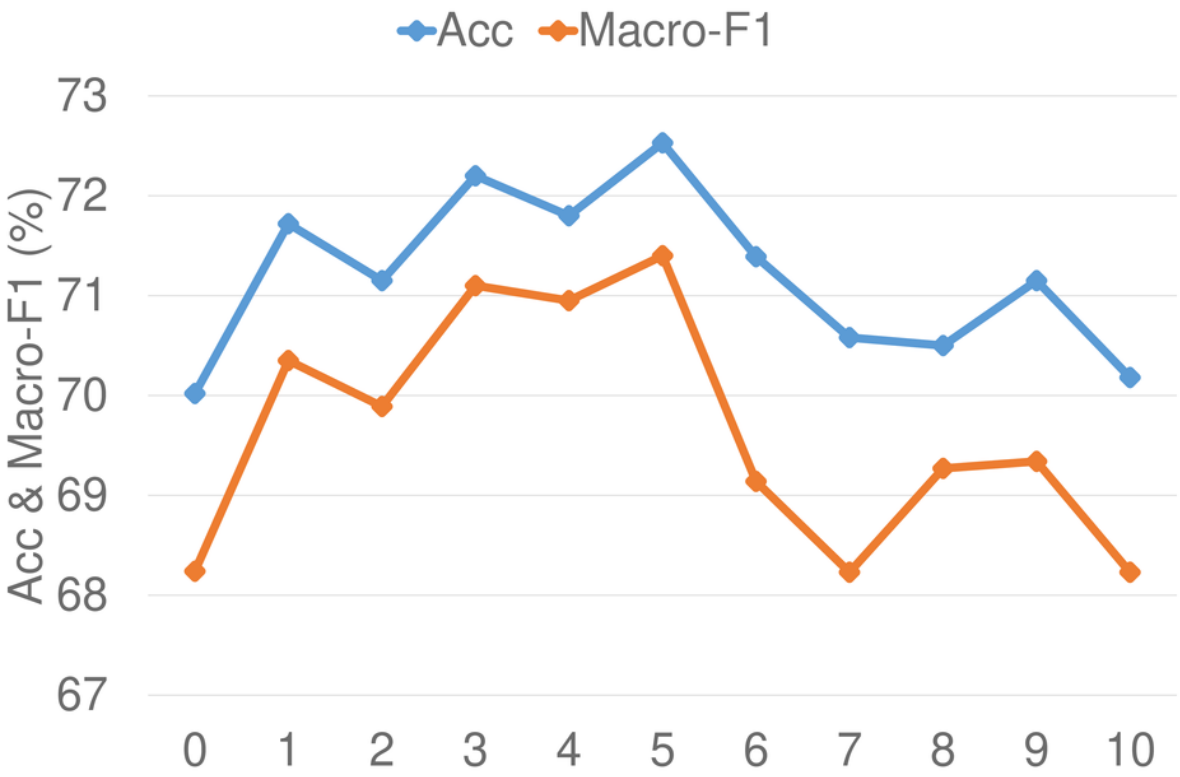


Figure 10

Effect of λ on TWITTER-2015

Effect of λ on model Accuracy and Macro-F1.

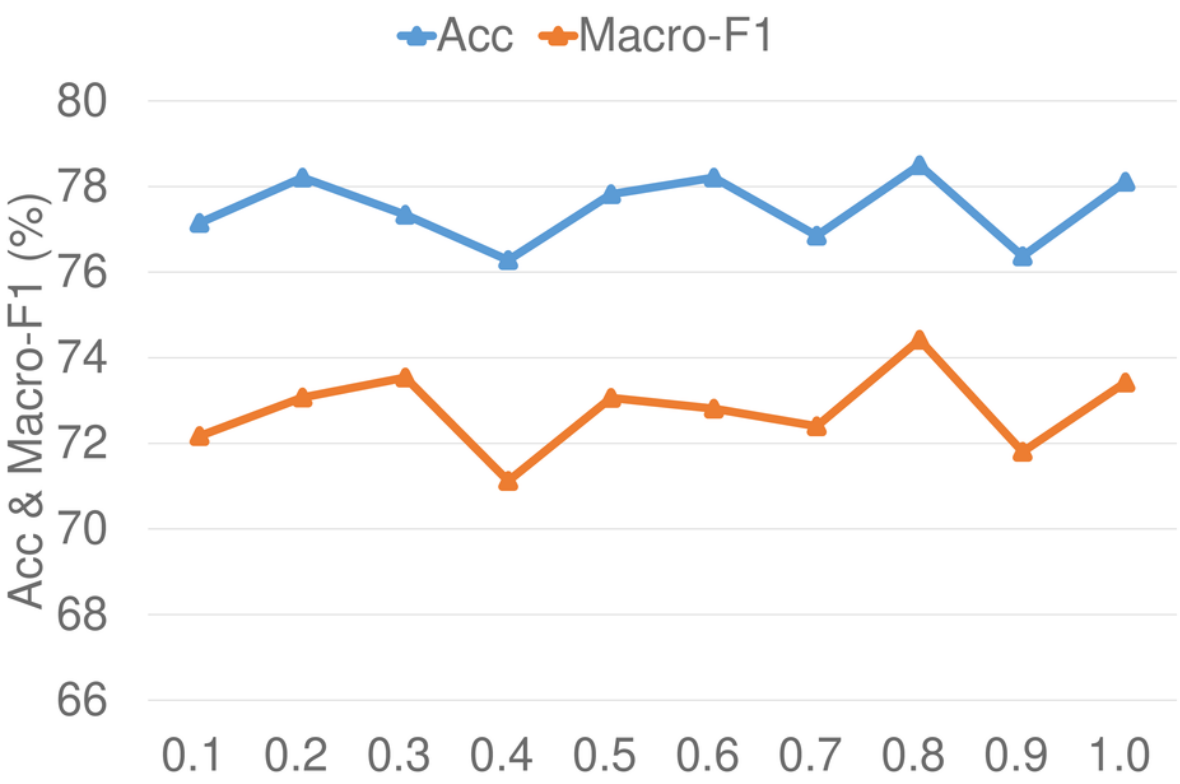


Figure 11

Effect of λ on TWITTER-2017

Effect of λ on model Accuracy and Macro-F1.