

Singapore Management University

## Institutional Knowledge at Singapore Management University

---

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

---

1-2020

### Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification

Jianfei YU

Jing JIANG

*Singapore Management University, [jingjiang@smu.edu.sg](mailto:jingjiang@smu.edu.sg)*

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

---

#### Citation

1

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [cherylds@smu.edu.sg](mailto:cherylds@smu.edu.sg).

# Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification

Jianfei Yu , Jing Jiang, and Rui Xia 

**Abstract**—Entity-level (aka target-dependent) sentiment analysis of social media posts has recently attracted increasing attention, and its goal is to predict the sentiment orientations over individual target entities mentioned in users' posts. Most existing approaches to this task primarily rely on the textual content, but fail to consider the other important data sources (e.g., images, videos, and user profiles), which can potentially enhance these text-based approaches. Motivated by the observation, we study *entity-level multimodal sentiment classification* in this article, and aim to explore the usefulness of images for entity-level sentiment detection in social media posts. Specifically, we propose an Entity-Sensitive Attention and Fusion Network (ESAFN) for this task. First, to capture the intra-modality dynamics, ESAFN leverages an effective attention mechanism to generate entity-sensitive textual representations, followed by aggregating them with a textual fusion layer. Next, ESAFN learns the entity-sensitive visual representation with an entity-oriented visual attention mechanism, followed by a gated mechanism to eliminate the noisy visual context. Moreover, to capture the inter-modality dynamics, ESAFN further fuses the textual and visual representations with a bilinear interaction layer. To evaluate the effectiveness of ESAFN, we manually annotate the sentiment orientation over each given entity based on two recently released multimodal NER datasets, and show that ESAFN can significantly outperform several highly competitive unimodal and multimodal methods.

**Index Terms**—Natural language processing, fine-grained sentiment analysis, multimodal sentiment analysis, neural networks, social media analysis.

## I. INTRODUCTION

**I**N THE age of social media, a large number of public multimodal posts are generated by users on platforms such as Twitter, Facebook and Instagram. It is quite useful to analyze this stream of data to study users' sentiment orientations towards a person, an organization or a location. Entity-level sentiment classification, also known as target-dependent sentiment classification, is the problem of identifying sentiment polarities towards

entities mentioned in an input sentence. For example, given the sentence “*The campus of UTEC University in Peru won @RIBA’s inaugural International Prize!*,” the user expresses *positive*, *neutral*, and *neutral* sentiment over “*UTEC University*,” “*Peru*,” and “*RIBA*” respectively. This problem has been receiving increasing attention from both academia and industry in the last decade [1].

In the literature, many methods have been proposed to perform sentiment classification for target entities. Traditional approaches to this problem focused on designing extensive hand-crafted features followed by feeding them to linear classifiers [2]–[5]. With the wide application of deep learning in NLP, different neural network architectures have also been proposed for entity-level sentiment classification, such as Recursive Neural Networks (ReNNs) [6], Convolutional Neural Networks (CNNs) [7] and Recurrent Neural Networks (RNNs) [8]. Recently, to better capture the semantic interactions between context words and target entities, many studies attempted to employ various attention mechanisms on top of RNNs, which have been shown to achieve state-of-the-art results in most benchmark datasets [9].

However, as multimodal data become increasingly popular on social media platforms, entity-level sentiment classification should no longer be based on textual content alone, as the aforementioned previous methods are. Multimodal posts usually come with images, and these images can often provide valuable insights into users' sentiment for a couple of reasons. On the one hand, for user posts only mentioning single entity, the sentiment towards the entity sometimes largely relies on its associated image due to the short and informal nature of textual contents in these posts. Take Table I.A and Table I.B for instance. Without taking the associated image into consideration, one would predict the sentiment towards “*Randy Johnson*” and “*Aston Villa*” to be *neutral*. However, in these posts, the two users respectively express *positive* and *negative* sentiment towards “*Randy Johnson*” and “*Aston Villa*,” since they post a pleasant image of *Randy Johnson* and an unpleasant image of the football manager of *Aston Villa*, respectively. On the other hand, for user posts mentioning multiple entities, it is often the case that the textual contents only focus on one entity by expressing subjective sentiment towards it, and the accompanying images can also help highlight this focused entity. For example, in Table I.C and Table I.D, we can see that the two images respectively focus on “*UTEC University*” and “*Chuck Bass*,” which are also the focus of the textual contents. Moreover, for the other mentioned entities (e.g., *Peru*, *RIBA*, and *MCM* in Table I.C and Table I.D), the textual




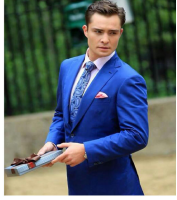
Manuscript received May 2, 2019; revised September 29, 2019; accepted November 26, 2019. Date of publication December 6, 2019; date of current version December 24, 2019. The work was supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative, in part by the Natural Science Foundation of China under Grant 61672288, and in part by the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars under Grant BK20160085. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Carlos Busso. (*Corresponding author: Jing Jiang.*)

J. Yu and R. Xia are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jfyu@njust.edu.cn; rxia@njust.edu.cn).

J. Jiang is with the School of Information Systems, Singapore Management University, 188065 Singapore (e-mail: jingjiang@smu.edu.sg).

Digital Object Identifier 10.1109/TASLP.2019.2957872

TABLE I  
REPRESENTATIVE EXAMPLES FOR ENTITY-LEVEL MULTIMODAL SENTIMENT CLASSIFICATION IN OUR DATASET. NAMED ENTITIES AND THEIR CORRESPONDING SENTIMENTS ARE HIGHLIGHTED

User Posts with Single Target Entity		User Posts with Multiple Target Entities	
			
A. <b>[Randy Johnson]</b> <sub>positive</sub> has arrived @Dbacks.	B. @dreamteamfc How many players will <b>[Aston Villa]</b> <sub>negative</sub> have left by the start of the season?	C. The campus of <b>[UTEC University]</b> <sub>positive</sub> in <b>[Peru]</b> <sub>neutral</sub> won <b>@[RIBA]</b> <sub>neutral</sub> 's inaugural International Prize!	D. <b>[Chuck Bass]</b> <sub>positive</sub> is everything <b>#[MCM]</b> <sub>neutral</sub>

contents often indicate neutral sentiment towards them, and the images also tend to pay less or even no attention to them. Hence, to further improve the performance of entity-level sentiment classification for these multimodal posts, it is crucial to develop an end-to-end model to effectively capture the intra-modality interactions including entity-text and entity-image alignments as well as the inter-modality interactions between the textual context and the visual context.

To capture the intra-modality and inter-modality interactions, we propose an entity-sensitive attention and fusion network (ESAFN). Specifically, to obtain the entity-sensitive textual representation, we first split the textual content of each input post into three components: left context, right context, and the target entity, and then utilize three separate LSTMs to obtain their initial representations. Based on this, we represent the target entity by averaging the values of its hidden states, and then determine the left and right contextual representations by employing the widely-used attention mechanism to capture the most important context information with respect to the target entity. Next, we design a textual fusion layer to aggregate the left and right contextual representations as well as the entity representation. Moreover, to obtain the entity-sensitive visual representation, we propose an entity-oriented visual attention mechanism to extract the important visual blocks that are closely related to the target entity, followed by designing a gated mechanism to eliminate the noise brought by the visual context. Finally, we apply another multimodal fusion layer to model the interactions between the textual and visual representations, and obtain the multimodal representation, followed by feeding it to a softmax layer for sentiment classification.

We conduct comprehensive experiments on one benchmark dataset and two manually annotated multimodal datasets from Twitter, and make a couple of observations. First, compared with several state-of-the-art text-oriented methods, our base model without incorporating the associated image is able to achieve indistinguishable or even better performance on all the datasets. Second, our full model ESAFN can consistently outperform the base model and highly competitive multimodal methods on the two multimodal datasets. Finally, since ESAFN is sensitive to the query entity, it can perform significantly better than baseline methods when the multimodal user posts

mention multiple entities. We will release our sentiment annotation as well as our code for research purpose via the link: <https://github.com/jefferyYu/ESAFN>.

## II. RELATED WORK

### A. Entity-Level Sentiment Classification

Entity-level (also known as target-dependent or aspect-level) sentiment classification is an important subtask in sentiment analysis and has been extensively studied in recent years [10]. Most existing methods can be generally grouped into the following two branches.

One line of work focuses on leveraging external resources (including Part-of-Speech Tagger, dependency parser, and sentiment lexicons, etc) to manually design a set of task-specific features, followed by applying traditional statistical learning methods over the features for sentiment prediction [4], [5], [11], [12]. Despite achieving respectable results on different benchmark datasets, they suffer from the heavy reliance on feature engineering.

Another line of work centers around incorporating target information into various neural network (NN) models. Dong *et al.* [6] designed a target-dependent recursive NN model on top of dependency parse trees. Later, Tang *et al.* [8] proposed to split each input sentence into two parts and use two LSTM models to respectively model the left context and the right context. Recently, Xue *et al.* [7] designed a gated convolutional NN model to select related sentiment features with respect to the target entity. Inspired by the advantages of attention mechanisms in capturing long-range context information in other NLP tasks [13]–[15], many recent studies have devised different attention mechanisms to model the interactions between the target entity and the context [16]–[22].

However, most of these studies only focused on modeling the textual context based on its relevance with the target entity, but did not consider visual features that are increasingly common in this age of social media. Different from them, the goal of our work is to effectively model the interactions between the target entity, the textual context, and the associated image context to accurately detect the sentiment over entities mentioned in multimodal social media posts.

## B. Multimodal Sentiment Classification

With the growth of multimodal data in social media, information from different modalities (visual, acoustic, etc.) has been leveraged to provide complementary sentiment signals to the traditional textual features in recent years [10]. The majority of these studies could be categorized into two lines.

The first line of work is designed for coarse-grained sentiment analysis of multimodal conversations. One group of work only focuses on integrating the corresponding acoustic information with textual features. Among them, a representative study by Bertero *et al.* [23] proposed a hierarchical CNN approach, which first performs speech recognition followed by classifying the emotion and sentiment for each utterance in interactive spoken dialogue systems. Another group of work centers on incorporating the associated audio and visual information in addition to textual features. In particular, an earlier study by Poria *et al.* [24] first employed a pre-trained CNN model to extract the textual features, and then applied multiple kernel learning to fuse textual, visual, and audio features for sentiment prediction of the last utterance. Later, they extended this work by proposing an LSTM-based architecture to capture the sequential structure of the historical conversational information [25]. Based on this work, Zadeh *et al.* [26] and Zadeh *et al.* [27] respectively designed a tensor fusion network and a memory fusion network to better capture the interactions between different modalities for each historical utterance. However, these coarse-grained methods may not perform well when directly applied to our fine-grained sentiment classification task. Therefore, in this paper, we are interested to propose a task-specific neural network model for entity-level multimodal sentiment classification.

The second line of work focuses on sentiment analysis of multimodal social media contents. Specifically, Borth *et al.* [28] first proposed to use textual tag to aid visual sentiment analysis, where they adopted a sentiment lexicon to detect the sentiment of textual tag, and treated this as its associated image's sentiment label. Based on this, Chen *et al.* [29] proposed a hierarchical system to first detect objects in images and then build object-based sentiment concept models for identifying visual sentiment concepts. Since the sentiment lexicon-based algorithm for automatic image annotation is noisy, You *et al.* [30] designed a progressive CNN architecture to eliminate the noise. However, these studies mainly focus on visual sentiment analysis, where the textual content is simply leveraged to generate sentiment labels for images with textual tags, and only the visual content is considered in their models. In contrast, our work focuses on multimodal sentiment analysis, where the textual and visual contents are fully utilized in our model. In addition, some previous work also considered textual and visual contents for multimodal sentiment analysis [31]. But their work is designed for coarse-grained tweet-level sentiment analysis, whereas our work targets at fine-grained entity-level sentiment analysis. Moreover, their proposed early fusion and late fusion approaches to integrate textual and visual contents are based on linear operators, while we propose a bilinear fusion layer, which can provide richer interactions between textual features and visual features.

## III. METHODOLOGY

In this section, we will describe our proposed entity-sensitive attention and fusion network (ESAFN) in detail.

### A. Notation and Problem Formulation

Given a sentence  $S$  with  $n$  words as well as its associated image  $\mathbf{V}$ , we assume that all the target entities in  $S$  (i.e., words or phrases) have been provided, which follows the standard setting for entity-level text sentiment classification [6]. With the  $(S, \mathbf{V})$  pair and one of its target entities  $T$  as inputs, our goal is to predict the sentiment orientation  $y$  over the target entity  $T$ , where  $y$  can be either *positive*, *negative*, or *neutral*.

**Textual Inputs:** It is shown in previous work that splitting textual inputs into the left context, the right context, and the target entity can better differentiate the input sequence when the sentence has multiple entities [32], [33]. Inspired by this, we also split the input sentence  $S$  into three parts. Formally, let us use  $\mathbf{s}_l = (\mathbf{x}_1^l, \mathbf{x}_2^l, \dots, \mathbf{x}_L^l)$ ,  $\mathbf{s}_r = (\mathbf{x}_1^r, \mathbf{x}_2^r, \dots, \mathbf{x}_R^r)$ , and  $\mathbf{t} = (\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_C^t)$  to denote the left context, the right context and the target entity, respectively, where  $\mathbf{x}_i$  is an  $e$ -dimensional word vector from a word embedding lookup matrix  $\mathbf{E} \in \mathbb{R}^{e \times |\mathcal{V}|}$  with a vocabulary size of  $|\mathcal{V}|$ , and  $L$ ,  $R$ , and  $C$  are the input lengths of each component.

**Visual Inputs:** Due to the promising performance of deep CNN models in many image recognition tasks, we adopt one of the state-of-the-art CNN models called Residual Network (ResNet) [34] to extract visual features of different visual blocks. Given an input image  $\mathbf{V}$ , we first resize it to  $224 \times 224$  pixels as in [34], and denote the new image as  $\mathbf{V}'$ . Since the deeper layers in pre-trained ResNet have been shown to capture meaningful high-level features which may be potentially useful for our task (e.g., human faces, gestures, etc), we keep the output from the last convolutional layer in a pre-trained 152-layer ResNet model,<sup>1</sup> and obtain the visual feature representation as follows:

$$\mathbf{R} = \text{ResNet}(\mathbf{V}'), \quad (1)$$

where  $\mathbf{R}$  is a tensor with a dimension of  $2048 \times 7 \times 7$ . Note that here  $7 \times 7$  refers to the number of  $32 \times 32$  visual blocks in  $\mathbf{V}'$ , and 2048 refers to the dimension of the feature vector learned for each  $32 \times 32$  visual block.

We further assume that we have a set of manually labeled samples as our training set, which is denoted by  $D = \{(\mathbf{s}_l^{(j)}, \mathbf{s}_r^{(j)}, \mathbf{t}^{(j)}, \mathbf{R}^{(j)}, y^{(j)})\}_{j=1}^N$ .

### B. Overview of Our Approach

Our proposed method essentially consists of three components: Entity-Sensitive Textual Representation, Entity-Sensitive Visual Representation, and Multimodal Representation. The whole architecture is illustrated in Fig. 1.

For the target entity, we first employ a Long Short-Term Memory (LSTM) network to obtain the hidden state for each word

<sup>1</sup>We use the ResNet model pre-trained on ImageNet with 1000 classes via the link: [Online]. Available: <https://download.pytorch.org/models/resnet152-b121ed2d.pth>.



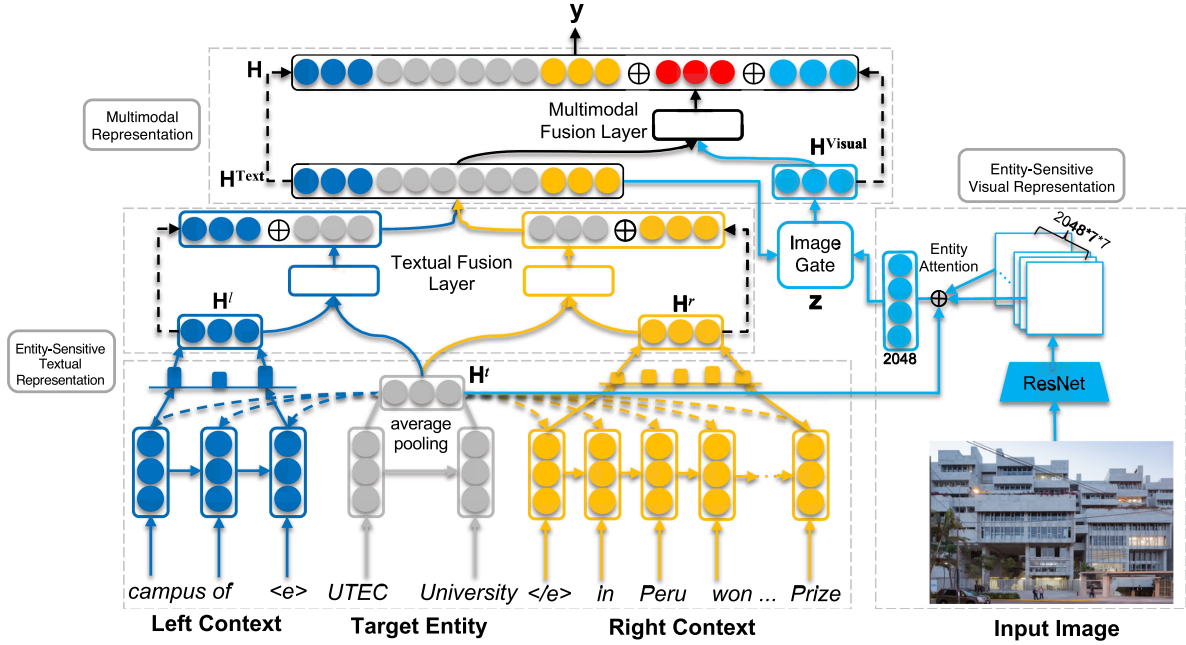


Fig. 1. The Proposed Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification.

in the target entity, followed by an average pooling operation to obtain its representation. For the **textual context**, we use another two LSTMs to get the hidden state for each word in both the left and the right contexts, and then utilize the target entity to help generate appropriate attention weights for each context word. Based on these attention weights, the weighted sum of the left and right context words will be considered as the left context representation and the right context representation, respectively. Moreover, we obtain the final textual representation by fusing the representations of the target entity, the left and right contexts with a textual fusion layer. For the **visual context**, after getting the feature vectors for each visual block, we compute the attention weights for each block based on their relatedness with the target entity. On top of the generated visual attention vectors, we further design a visual gated mechanism to dynamically control the contribution of the visual information, since in some cases the associated image might be noisy or even irrelevant to the textual content. Finally, we aggregate the textual and visual representations with a multimodal fusion layer, and feed the **multimodal representation** to a softmax function to perform sentiment classification at the entity level.

In the following subsections, we will introduce the details of learning the representations for these three components.

### C. Target Entity Representation

As mentioned above, we employ a standard LSTM network to sequentially read all the words in the input target entity to form the hidden state  $\mathbf{h}_i^t$  for each word  $\mathbf{x}_i^t$ :

$$\mathbf{h}_i^t = \text{LSTM}_{\Theta}(\mathbf{h}_{i-1}^t, \mathbf{x}_i^t), \quad i \in [1, C] \quad (2)$$

where  $\mathbf{h}_i^T \in \mathbb{R}^d$  and  $\Theta$  represents all the parameters in LSTM. After getting the hidden states of all the words in the target entity

$[\mathbf{h}_1^t, \mathbf{h}_2^t, \dots, \mathbf{h}_C^t]$ , we employ the average value of the hidden states as the final representation of the target entity:

$$\mathbf{H}^t = \sum_{i=1}^C \frac{1}{C} \mathbf{h}_i^t. \quad (3)$$

### D. Entity-Sensitive Textual Representation (ESTR)

**Entity Position Indicator (EPI):** Since the position of target entities can intuitively reflect the importance of context words with respect to it, many previous methods [16], [19], [22] proposed to pay higher attention to the context words closer to the target entity, where various position weighting strategies are designed to upweight the context words closer to the entity. However, these fixed position weighting strategies may suffer from diverse cases in social media posts, especially when the corresponding sentiment words or emojis have a long distance to the target entity. To tackle this limitation, we propose a simple but flexible strategy to indicate the entity position, which adds two indicator tokens (i.e.,  $\langle e \rangle$  and  $\langle /e \rangle$ ) before and after the target entity. For example, with “*UTEC University*” as the query, the textual input for Table I.C is “*The campus of  $\langle e \rangle$  UTEC University  $\langle /e \rangle$  in Peru won @RIBA’s inaugural International Prize!*,” and its left and rights contexts are illustrated in the bottom of Fig. 1.

**Context Representation:** Next, to better capture the semantic meanings and the long-range word dependencies of the left contexts and the right contexts, we leverage two separate LSTM networks to produce their  $d$ -dimensional hidden states:  $[\mathbf{h}_1^l, \mathbf{h}_2^l, \dots, \mathbf{h}_L^l]$  and  $[\mathbf{h}_1^r, \mathbf{h}_2^r, \dots, \mathbf{h}_R^r]$ .

Based on these hidden states, we further adopt the widely used attention mechanism [13] to learn the semantic representations of the left and right contexts. Intuitively, given different target

entities as queries, the importance of each context word should be different. For example, in Table I.C, given “*UTEC University*,” the context words “*won*” and “*Prize*” tend to receive more attention; whereas with “*Peru*” as the query, the preposition word “*in*” tend to receive more attention. Therefore, with the target entity representation  $\mathbf{H}^t$  as input, we compute the attention weights for each hidden state in the left context based on their bilinear interactions with  $\mathbf{H}^t$ :

$$u_i^l = \tanh(\mathbf{h}_i^l \cdot \mathbf{W}_H^l \cdot \mathbf{H}^t + b^l), \quad (4)$$

$$\alpha_i^l = \frac{\exp(u_i^l)}{\sum_{j=1}^L \exp(u_j^l)}, \quad (5)$$

where  $\mathbf{W}_H^l \in \mathbb{R}^{d \times d}$ , and  $b^l \in \mathbb{R}$  are learnable parameters. Based on this, the final representation of the left context can be derived as follows:

$$\mathbf{H}^l = \sum_{i=1}^L \alpha_i^l \mathbf{h}_i^l. \quad (6)$$

Similarly, we can derive the final representation of the right context  $\mathbf{H}^r$  based on Eq. (4) to Eq. (6).

**Textual Fusion Layer:** While many state-of-the-art methods used simple feature concatenation to integrate the information from the target entity and the textual context [18], [22], we argue that simply concatenating features will inevitably ignore the higher-order interactions between them. Therefore, we adopt the widely used bilinear models [35], which is expected to consider all the pairwise interactions between features. Specifically, instead of applying the standard bilinear operator introduced in [35] that leads to large parameter size and high model complexity, we propose to use a low-rank bilinear pooling operator, which has been demonstrated to retain the performance of the standard bilinear operator with much fewer parameters [36], to model the interactions between the entity and the left (or right) context as follows:

$$\mathbf{H}^{lt} = \mathbf{P}_l^\top (\sigma(\mathbf{U}_l^\top \mathbf{H}^l) \circ \sigma(\mathbf{U}_{lt}^\top \mathbf{H}^t)) + \mathbf{b}_l, \quad (7)$$

$$\mathbf{H}^{rt} = \mathbf{P}_r^\top (\sigma(\mathbf{U}_r^\top \mathbf{H}^r) \circ \sigma(\mathbf{U}_{rt}^\top \mathbf{H}^t)) + \mathbf{b}_r, \quad (8)$$

where  $\mathbf{U}_l, \mathbf{U}_r, \mathbf{U}_{lt}, \mathbf{U}_{rt} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{P}_l, \mathbf{P}_r \in \mathbb{R}^{d \times d}$ , and  $\mathbf{b}_l, \mathbf{b}_r \in \mathbb{R}^d$  are learnable parameters,  $\sigma$  is the non-linear transformation function  $\tanh$ , and  $\circ$  is the element-wise multiplication. To avoid information loss, we further combine the original context representation with the integrated representation as the final textual representation:

$$\mathbf{H}^{\text{Text}} = \mathbf{H}^l \oplus \mathbf{H}^{lt} \oplus \mathbf{H}^r \oplus \mathbf{H}^{rt}. \quad (9)$$

### E. Entity-Sensitive Visual Representation (ESVR)

Since the textual content in multimodal social media posts is naturally short and sometimes even incomplete, only learning the textual representation may still be insufficient to make correct sentiment predictions. Therefore, it is necessary and important to learn an effective visual representation to improve the model robustness.

Intuitively, given a specific target entity, it is often the case that only some parts of the accompanying image are related to it. Let

us again take Table I.C as an example. With “*UTEC University*” as the target entity, our model should only focus on the building, and ignore the other background. Similarly, in Table I.D, with “*Chuck Bass*” as the target entity, the background behind the person is not related to final sentiment predictions, and should be ignored.

Inspired by this, we apply the widely used visual attention mechanism [37], [38] over the visual feature representation  $\mathbf{R}$ . This is expected to help our model only focus on the visual blocks that are closely relevant to the target entity. We represent the  $2048 \times 7 \times 7$ -dimensional tensor  $\mathbf{R}$  as follows:

$$\mathbf{R} = \{\mathbf{r}_w | \mathbf{r}_w \in \mathbb{R}^{2048}, w = 1, 2, \dots, 49\}, \quad (10)$$

where  $\mathbf{r}_w$  is the feature representation of each block. Then, we calculate the attention weights for each block as follows:

$$u_w^v = \mathbf{q}^\top \tanh(\mathbf{W}_H^v \mathbf{H}^t + \mathbf{W}_R^v \mathbf{r}_w + \mathbf{b}^v), \quad (11)$$

$$\alpha_w^v = \frac{\exp(u_w^v)}{\sum_{j=1}^{49} \exp(u_j^v)}, \quad (12)$$

where  $\mathbf{W}_H^v \in \mathbb{R}^{d \times d}$ ,  $\mathbf{W}_R^v \in \mathbb{R}^{d \times 2048}$ ,  $\mathbf{q} \in \mathbb{R}^d$ , and  $\mathbf{b}^v \in \mathbb{R}^d$  are learnable parameters. Based on these visual attention weights, we can obtain the visual contextual representation with the weighted sum of all the visual blocks:

$$\mathbf{r}^v = \sum_{w=1}^{49} \alpha_w^v \mathbf{r}_w, \quad (13)$$

where  $\mathbf{r}^v$  is a 2048-dimensional image feature vector. To be consistent with the dimensions of textual context representations, we use a non-linear function to transform  $\mathbf{r}^v$  to a  $d$ -dimensional vector:

$$\mathbf{Q}^v = \tanh(\mathbf{W}^v \mathbf{r}^v + \mathbf{b}^v), \quad (14)$$

where  $\mathbf{W}^v \in \mathbb{R}^{d \times 2048}$  and  $\mathbf{b}^v \in \mathbb{R}^d$  are learnable parameters.

**Gated Mechanism:** Although incorporating  $\mathbf{Q}^v$  is expected to improve the model performance, it may also introduce some noise due to several reasons. First, in many multimodal social media posts, the input image may be only useful for inferring the sentiment of one target entities, but useless for the other target entities. For example, in Table I.C, given either “*Peru*” or “*RIBA*” as the target entity, it is unnecessary to incorporate the image feature vector into our model. Second, in some cases, the input image may be less relevant or even irrelevant to the textual context, and should largely be ignored by our model.

Therefore, to dynamically eliminate the noise brought by the associated image, we propose to incorporate an image gate by combining  $\mathbf{H}^{\text{Text}}$  and  $\mathbf{r}^v$  as follows.<sup>2</sup>:

$$\mathbf{z} = \sigma(\mathbf{W}_H^z \mathbf{H}^{\text{Text}} + \mathbf{W}_R^z \mathbf{r}^v + \mathbf{b}^z), \quad (15)$$

<sup>2</sup>Note that visual gated mechanism has been introduced for dialogue-level multimodal sentiment classification in previous work [39] However, our proposed solution differs from theirs in two aspects: (1) Instead of purely relying on visual context, our gated output  $\mathbf{z}$  is determined by the textual context and the visual context together; (2) Instead of constraining each element of  $\mathbf{z}$  to be a binary value, we use  $\sigma$  to convert it to 0 to 1, which is more flexible and can dynamically learn the importance of visual context.

where  $\mathbf{W}_H^z \in \mathbb{R}^{d \times 4d}$ ,  $\mathbf{W}_R^z \in \mathbb{R}^{d \times 2048}$ , and  $\mathbf{b}^z \in \mathbb{R}^d$  are learnable parameters, and  $\sigma$  is the element-wise sigmoid function. Based on the gated output  $\mathbf{z}$ , the final visual context representation can be obtained as follows:

$$\mathbf{H}^{\text{Visual}} = \mathbf{z} \circ \mathbf{Q}^v. \quad (16)$$

### F. Multimodal Representation

**Multimodal Fusion Layer:** Based on the entity-sensitive textual representation  $\mathbf{H}^{\text{Text}}$  and visual representation  $\mathbf{H}^{\text{Visual}}$ , we apply another bilinear pooling operator to capture their inter-modality interactions as follows:

$$\mathbf{H}^{\text{MM}} = \mathbf{P}_m^\top (\sigma(\mathbf{U}_{\text{Text}}^\top \mathbf{H}^{\text{Text}}) \circ \sigma(\mathbf{U}_{\text{Visual}}^\top \mathbf{H}^{\text{Visual}})) + \mathbf{b}_m, \quad (17)$$

where  $\mathbf{U}_{\text{Text}} \in \mathbb{R}^{4d \times d}$ ,  $\mathbf{U}_{\text{Visual}} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{P}_m \in \mathbb{R}^{d \times d}$ , and  $\mathbf{b}_m \in \mathbb{R}^d$  are learnable parameters, and  $\sigma$  is the non-linear transformation function  $\tanh$ . Similar to the textual fusion layer as introduced before, we also combine it together with  $\mathbf{H}^{\text{Text}}$  and  $\mathbf{H}^{\text{Visual}}$  to form the final multimodal representation:

$$\mathbf{H} = \mathbf{H}^{\text{Text}} \oplus \mathbf{H}^{\text{Visual}} \oplus \mathbf{H}^{\text{MM}}. \quad (18)$$

Finally, we feed the multimodal representation  $\mathbf{H}$  to a softmax function for entity-level sentiment classification:

$$p(y|\mathbf{H}) = \text{softmax}(\mathbf{W}^\top \mathbf{H} + \mathbf{b}), \quad (19)$$

where  $\mathbf{W} \in \mathbb{R}^{6d \times 3}$  and  $\mathbf{b} \in \mathbb{R}^3$  are learnable parameters.

### G. Model Training

To optimize all the parameters in our method, our objective function is to minimize the standard cross-entropy loss:

$$\mathcal{J} = -\frac{1}{N} \sum_{j=1}^N \log p(y^{(j)}|\mathbf{H}^{(j)}). \quad (20)$$

## IV. EXPERIMENTS

In this section, we conduct different sets of experiments with the aim of answering the following research questions:

- **RQ1:** Can our entity-sensitive textual representation (*ESTR*) outperform the state-of-the-art text-oriented approaches on all the three datasets? (Section IV-B1)
- **RQ2:** Could our full model *ESAFN* bring significant improvements to *ESTR* and achieve the best performance on our two multimodal datasets? (Section IV-B2)
- **RQ3:** What is the effectiveness of different components in *ESAFN*, including the entity position indicator (EPI), splitting the text into three parts, the textual fusion layer (TFL), the gated mechanism, and the multimodal fusion layer (MFL)? (Section IV-C)
- **RQ4:** What is the advantage of *ESAFN* over other strong baseline approaches? (Section IV-D)

### A. Experiment Settings

**Datasets:** To evaluate the effectiveness of *ESAFN*, we conduct experiments on one unimodal benchmark dataset and two multimodal datasets. For the unimodal dataset, it is a benchmark constructed by [6], which mainly consists of pure textual user posts published between 2010 and 2014 in Twitter. For multimodal datasets, since there is no publicly available annotated Twitter corpus for our task, we choose to construct our datasets based on two publicly available multimodal named entity recognition (NER) datasets that were collected by [40] and [41] respectively. These two datasets respectively include multimodal user posts published during 2014-2015 and 2016-2017 in Twitter, and all the entities belong to four types: Person, Location, Organization, and Miscellaneous. The basic statistics of the three datasets are summarized in Table II. Note that in our two multimodal datasets, all the tweets contain textual contents and their associated images.

**Human Sentiment Annotation:** As the two multimodal datasets only contain manually annotated entities, we ask three domain experts to combine the textual content and the associated image to annotate the sentiment orientation towards each entity. In our initial annotation process (i.e., randomly choosing 200 tweets for annotation), we find that for tweets with single entity, the annotated sentiment has a high correlation, whereas for tweets with multiple entities, the agreement between annotators is relatively low. Therefore, we provide some guidelines to our annotators: for the entities highlighted by the image, the annotator should label their sentiment based on the text and image together; while for the other remaining entities, the annotator should label their sentiment mainly based on the text (in most cases, the sentiments towards them tend to be labeled as neutral). After obtaining the human annotation, we further use Cohen's kappa [42] to measure inter-annotator agreement. As shown in Table III, the agreement between most pairs of annotators in the two datasets is above 0.5, indicating that the sentiment can be agreed upon in a general sense. Among the disagreement cases, we find that most of them belong to neutral-subjective disagreement, and only less than 5% cases are related to positive-negative contradiction. We then take the majority label among the three annotators as the gold label, and filter the rare samples when there is no agreement between any two annotators. Finally, as shown in Table II, we further split the annotated data into training (60%), development (20%) and test (20%) sets.

**Pre-processing Details:** For all the three datasets, we follow most pre-processing rules used in [43] to tokenize the tweets, and the only difference is that we split the hashtags into two tokens: “#” and its following word.

**Parameter Settings:** For all the models, we set the word embedding size  $e$  to be 100 in the three datasets, and initialize the word embedding matrix  $\mathbf{E}$  using pre-trained word embeddings based on GloVe,<sup>3</sup> which will be fixed during the training process. The hidden dimension  $d$  and the number of LSTM layers in all

<sup>3</sup>[Online]. Available: <https://nlp.stanford.edu/projects/glove/>.

TABLE II  
THE BASIC STATISTICS OF TWITTER-14, TWITTER-15, AND TWITTER-17. POS AND NEG ARE SHORT FOR POSITIVE AND NEGATIVE CLASSES

	TWITTER-14			TWITTER-15					TWITTER-17								
	#POS	#NEG	#Neutral	#POS	#NEG	#Neutral	Total	#Avg Entities	Words	Len	#POS	#NEG	#Neutral	Total	#Avg Entities	Words	Len
Train	1561	1560	3127	928	368	1883	3179	1.348	9023	16.72	1508	416	1638	3562	1.410	6027	16.21
Dev.	-	-	-	303	149	670	1122	1.336	4238	16.74	515	144	517	1176	1.439	2922	16.37
Test	173	173	346	317	113	607	1037	1.354	3919	17.05	493	168	573	1234	1.450	3013	16.38

TABLE III  
AGREEMENT BETWEEN EVERY PAIR OF OUR THREE ANNOTATORS (A1, A2, A3)

	TWITTER-15			TWITTER-17		
	A1-A2	A2-A3	A1-A3	A1-A2	A2-A3	A1-A3
Cohen's kappa	0.550	0.490	0.558	0.587	0.540	0.575

the datasets are set to be 100 and 1. During training, Adam [44] is used to schedule the learning rate, where the initial learning rate is set to be 0.001. Also, the batch size and the dropout rate are set to be 10 and 0.5. Note that for all the parameters in *ResNet*, we initialize them with a pre-trained 152-layer model and keep them fixed in the training process. We implement all the models with PyTorch, and run experiments on a NVIDIA Tesla V100 GPU.

**Evaluation Metrics:** Following many previous studies for entity-level sentiment classification [9], we use the standard classification accuracy (ACC) and Macro- $F_1$  score as our evaluation metrics.

## B. Main Results

1) *Performance of ESTR (RQ1)*: To better show the effectiveness of our entity-sensitive textual representation model, we re-implemented the following pure text-oriented systems with the same parameter settings for comparison:

- *Majority*, a simple baseline method, which uses the majority sentiment label in the training set for prediction;
- *LSTM*, a standard sentence-level LSTM model without explicitly considering the target entity;
- *AE-LSTM*, an extension of *LSTM* proposed by [17], which utilizes the attention mechanism to capture the important context information related to the target entity;
- *TD-LSTM*, another extension of *LSTM* proposed by [8], which utilizes two LSTMs to model the left context and the right context of the target respectively;
- *MemNet*, a deep memory network [16], which applies a multi-layer attention mechanism on top of the common word embedding layer;
- *IAN*, one of the current representative systems [18], which proposes an interactive attention mechanism to better model the interactions between the target entity and the context words;
- *RAM*, another representative system [19], which builds up a deep neural architecture by applying a GRU model on top of the representations obtained from multi-hop attention mechanism;
- *MGAN*, the recent state-of-the-art system [22], which designs a multi-grained attention network for fusing the target and the context at various degrees of granularity;

TABLE IV  
EXPERIMENTAL RESULTS ON ENTITY-LEVEL TEXT SENTIMENT CLASSIFICATION. \* AND † RESPECTIVELY INDICATE THAT ESTR IS SIGNIFICANTLY BETTER THAN THE SECOND BEST COMPARED SYSTEM AND THE BEST COMPARED SYSTEM WITH P-VALUE < 0.05 BASED ON MCNEMAR'S SIGNIFICANCE TEST

Method	TWITTER-14		TWITTER-15		TWITTER-17	
	ACC	Macro- $F_1$	ACC	Macro- $F_1$	ACC	Macro- $F_1$
Majority	50.00	33.33	58.53	33.33	46.43	33.33
LSTM	66.50	64.70	67.98	57.30	55.92	51.69
AE-LSTM	67.34	65.72	70.30	63.43	61.67	57.97
TD-LSTM	70.80	69.00	70.67	63.58	64.66	60.65
MemNet	68.50	66.91	70.11	61.76	64.18	60.90
IAN	71.24	70.07	70.49	62.81	63.94	61.05
RAM	71.88	70.33	70.68	63.05	64.42	61.01
MGAN	<b>72.54</b>	70.81	71.17	64.21	64.75	61.46
ESTR	72.25*	<b>71.42*</b>	<b>71.36*</b>	<b>64.28*</b>	65.80†	<b>62.00†</b>

- *ESTR*, our entity-sensitive textual representation model as detailed in Section III-D, followed by a softmax layer for sentiment prediction.

Based on the results reported in Table IV, we can clearly observe that despite outperforming *Majority* with a large margin, the performance of *LSTM* is still relatively limited, which is intuitive because *LSTM* treats all the target entities in the same sentence equally and assigns them with the same sentiment labels. Moreover, due to the careful design of incorporating the target entity information into their models, *AE-LSTM*, *TD-LSTM*, *MemNet*, and *IAN* can consistently obtain respectable improvements over *LSTM*. In addition, the two state-of-the-art methods *RAM* and *MGAN* can further improve the results in most datasets, which may benefit from their deep architectures. Finally, compared with all the baseline methods, we can easily observe that our *ESTR* model can generally outperform a number of highly competitive approaches including *AE-LSTM*, *TD-LSTM*, *MemNet*, *IAN*, and *RAM*, and achieve indistinguishable or even better performance in comparison with the state-of-the-art approach *MGAN* on the three datasets. These observations demonstrate the effectiveness of the textual representation obtained by *ESTR*.

2) *Performance of ESAFN (RQ2)*: Since the focus of this paper is to incorporate the visual context into entity-level sentiment classification, here we consider several highly competitive multimodal approaches for comparison:

- *Res-Target*, a simple combination of the target entity representation and the visual context representation, which basically concatenates  $\mathbf{H}^t$  and  $\mathbf{Q}^v$  followed by feeding the resulting vector to a softmax layer for classification;
- *Res-RAM*, *Res-MGAN*, and *Res-ESTR*, three variants of a simple type of multimodal fusion method [45], which first apply max-pooling over the visual features to obtain



TABLE V  
EXPERIMENTAL RESULTS ON ENTITY-LEVEL MULTIMODAL SENTIMENT CLASSIFICATION. † INDICATES THAT OUR FULL MODEL *ESAFN* IS SIGNIFICANTLY BETTER THAN THE BEST COMPARED SYSTEM WITH P-VALUE < 0.05 BASED ON McNEMAR’S SIGNIFICANCE TEST

Modality	Method	TWITTER-15		TWITTER-17	
		ACC	Macro- $F_1$	ACC	Macro- $F_1$
Visual	Res-Target	59.88	46.48	58.59	53.98
Text	RAM	70.68	63.05	64.42	61.01
	MGAN	71.17	64.21	64.75	61.46
	ESTR	71.36	64.28	65.80	62.00
Text + Visual	Res-RAM	71.55	64.68	65.40	62.23
	Res-RAM-TFN	69.91	61.49	63.45	58.92
	Res-MGAN	71.65	63.88	66.37	63.04
	Res-MGAN-TFN	70.30	64.14	64.10	59.13
	MIMN	71.84	65.69	65.88	62.99
	Res-ESTR	72.03	63.98	66.13	63.63
	<i>ESAFN</i>	<b>73.38†</b>	<b>67.37†</b>	<b>67.83†</b>	<b>64.22†</b>

TABLE VI  
ABLATION STUDY OF *ESTR*. † INDICATES THAT *ESTR* IS SIGNIFICANTLY BETTER THAN ALL ITS VARIANTS BY REMOVING ITS SUB-COMPONENTS

Method	TWITTER-14		TWITTER-15		TWITTER-17	
	ACC	Macro- $F_1$	ACC	Macro- $F_1$	ACC	Macro- $F_1$
<i>ESTR</i>	<b>72.25†</b>	<b>71.42†</b>	71.36	<b>64.28</b>	<b>65.80†</b>	<b>62.00†</b>
w/o EPI	71.38	69.75	<b>71.65</b>	63.52	64.18	59.37
w/o Split	70.80	70.15	70.33	61.54	64.18	60.65
w/o TFL	70.65	69.23	71.17	63.17	64.83	61.07

$\mathbf{g} = \text{MaxPool}(\mathbf{R})$ , and then concatenate  $\mathbf{g}$  and the textual representation from *RAM*, *MGAN*, and *ESTR*, followed by a softmax layer for classification;

- *Res-RAM-TFN* and *Res-MGAN-TFN*, two variants of another type of multimodal fusion method [26], which use a bilinear interaction operator to combine  $\mathbf{g}$  and the textual representation from *RAM* and *MGAN* through a complex fusion matrix, and feed the resulting matrix to Sentiment Inference Subnetwork for final classification;
- *MIMN*, the recent state-of-the-art multimodal approach for aspect-level sentiment classification proposed by Xu *et al.* [46], which adopts multi-hop memory network to model the interactive attention between the aspect word, the textual context, and the visual context;
- *EASFN*, our full model as introduced in Section III.

We report the results of all the compared methods in Table V. First, we can find that the performance of *Res-Target* is quite limited, which indicates that the textual content is quite important for entity-level sentiment classification, and should not be ignored. Second, it is easy to observe from Table V that with the help of the visual context, *Res-RAM*, *Res-MGAN*, and *Res-ESTR* can consistently bring improvements over their corresponding baseline approaches, which implies that the associated image can indeed play a supporting role to text and provide complementary information. Third, it is intuitive to see that by modelling the interaction between the target entity and the textual and visual contexts, *MIMN* can generally outperform most baseline approaches. But since it is mainly based on a relatively weak model *MemNet*, it still performs slightly worse than *Res-ESTR*.

TABLE VII  
ABLATION STUDY OF *ESAFN*. † SHOWS THAT *ESAFN* IS SIGNIFICANTLY BETTER THAN ALL ITS VARIANTS BY REMOVING ITS SUB-COMPONENTS

Method	TWITTER-15		TWITTER-17	
	ACC	Macro- $F_1$	ACC	Macro- $F_1$
<i>ESAFN</i>	<b>73.38†</b>	<b>67.37†</b>	<b>67.83†</b>	<b>64.22†</b>
w/o <i>ESVR</i>	72.03	63.98	66.13	63.63
w/o Gate	72.71	65.35	66.61	64.06
w/o MFL	71.94	65.86	66.77	63.09

TABLE VIII  
BREAKDOWN OF ACCURACY WITH RESPECT TO SENTENCES WITH SINGLE TARGET ENTITY AND MULTIPLE TARGET ENTITIES IN THE TEST SETS OF OUR TWO MULTIMODAL DATASETS. † DENOTES THAT *ESAFN* IS SIGNIFICANTLY BETTER THAN OUR TEXTUAL METHOD *ESTR*

Method	TWITTER-15		TWITTER-17	
	#Entities = 1 (566 samples)	#Entities ≥ 2 (471 samples)	#Entities = 1 (581 samples)	#Entities ≥ 2 (653 samples)
<i>ESTR</i>	72.97	69.43	67.47	64.32
Res- <i>ESTR</i>	72.79	71.13	64.72	67.38
<i>ESAFN</i>	<b>73.14</b>	<b>73.67†</b>	<b>68.16†</b>	<b>67.53†</b>

Finally, it is obvious that our full model *ESAFN* can outperform the second best results by 1.87% and 2.20% points in accuracy and 2.56% and 0.93% points in F1-score on TWITTER-15 and TWITTER-17 respectively. This demonstrates the usefulness of our multimodal attention and fusion approach for entity-level multimodal sentiment classification.





### C. Ablation Study

In this subsection, we investigate the effectiveness of different components in our proposed approach in order to answer the questions raised in **RQ3**.

**Components in *ESTR*:** As shown in Table VI, all the components contributing to *ESTR* play important roles to the final result. In particular, discarding the entity position indicator will generally drop the performance in all the datasets, which shows its usefulness. Besides, in comparison with merging the left and the right contexts, splitting them into two parts seems to perform better across all the three datasets, which is consistent with the findings in previous studies [32]. Moreover, the incorporation of the textual fusion layer demonstrates its indispensable effect to the final performance.

**Components in *ESAFN*:** In Table VII, we can find that replacing the entity-sensitive visual representation (*ESVR*) with the max-pooling value of its visual features will significantly drop the performance, which indicates the importance of aligning the target entity with the associated image. In addition, it is clear to see that removing the gated mechanism leads to a large performance drop, which also shows the usefulness of filtering the noisy visual features. Finally, the multimodal fusion layer, which combines the textual and visual representations with a bilinear operator, demonstrates its effectiveness in boosting the model performance.

TABLE IX  
COMPARISON OF PREDICTIONS FROM ESTR, RES-ESTR, AND ESAFN ON SEVERAL TEST SAMPLES. ✓ AND ✗ RESPECTIVELY DENOTE THE CORRECT AND INCORRECT PREDICTIONS

Associated Image	Input Sentence & Predicted Label	Associated Image	Input Sentence & Predicted Label
	<p>A. One Sunday closer to [Buffalo Bills]<sup>1</sup> football. #GoBills #FeelTheRush</p> <p>Human Label: (1-positive)</p> <p>ESTR: (1-neutral ✗)</p> <p>Res-ESTR: (1-positive ✓)</p> <p>ESAFN: (1-positive ✓)</p>		<p>B. [Jean Marmoreo]<sup>1</sup> — ready to run! #[STWM]<sup>2</sup>.</p> <p>Human Label: (1-positive, 2-neutral)</p> <p>ESTR: (1-neutral ✗, 2-neutral ✓)</p> <p>Res-ESTR: (1-positive ✓, 2-positive ✗)</p> <p>ESAFN: (1-positive ✓, 2-neutral ✓)</p>
	<p>C. Bought the radio staff [Burger King]<sup>1</sup> to show our support for [Tim Hortons]<sup>2</sup> - Wow fries good!</p> <p>Human Label: (1-positive, 2-positive)</p> <p>ESTR: (1-neutral ✗, 2-positive ✓)</p> <p>Res-ESTR: (1-positive ✓, 2-positive ✓)</p> <p>ESAFN: (1-positive ✓, 2-positive ✓)</p>		<p>D. These ladies in #[Knoxville]<sup>1</sup> know #RaiseTheWage creates better economic opportunity for all Americans.</p> <p>Human Label: (1-neutral)</p> <p>ESTR: (1-neutral ✓)</p> <p>Res-ESTR: (1-positive ✗)</p> <p>ESAFN: (1-positive ✗)</p>

#### D. Discussion

To answer the question raised in **RQ4**, we conduct additional experiments on the test set of our two datasets, and carefully choose several representative test samples to analyze our model predictions.

**Advantages in Posts With Multiple Entities:** As shown in Table VIII, we can easily find that by incorporating the entity-sensitive visual representation and the multimodal fusion layer into *ESTR*, our full model *ESAFN* can bring significant improvements, especially when the post contains multiple target entities. This observation is in line with our intuition that the associated image can help us better distinguish the entities focused by the textual content, and therefore improve the model performance.

**Case Study:** Table IX shows the comparison between the predictions of baseline methods and those of our *ESAFN* model on four samples. First, in Table IX.A, with the help of the associated image that shows a celebrating posture, the two multimodal approaches can correct the wrong prediction made by *ESTR*. Similarly, in Table IX.C, since the associated image posted by the user contains a smiley face, multimodal methods correctly predict the sentiment over the two entities as *positive*. Moreover, in Table IX.B, we can see that the image is about *Jean Marmoreo* instead of *STWM*. Although both *ESTR* and *Res-ESTR* gave a wrong prediction over one of the two entities, our *ESAFN* model may identify the alignments between *Jean Marmoreo* and the image, and therefore correctly predicts the sentiment over the focused entity as *positive*, and the sentiment over the other entity *STWM* as *neutral*. These three examples further confirm our motivations that our method is generally useful in two cases: 1). the input post has multiple entities; 2). the input post has only one entity but it is hard to infer its sentiment from the textual content alone, either due to long distance or incomplete context information.

**Error Analysis:** To show the negative effect of incorporating the associated image, we further analyze the cases that both *Res-ESTR* and *ESAFN* make wrong predictions, and find that most of these error cases are related to the *neutral* class (86.5% for TWITTER-15 and 84.3% for TWITTER-17). A representative example is given in Table IX.D. We can easily observe that here the sentiment over the entity *Knoxville* should be *neutral*, but the two multimodal models incorrectly predict its sentiment as *positive*, perhaps because the associated image posted by the user contains several smiley faces.

#### V. CONCLUSION

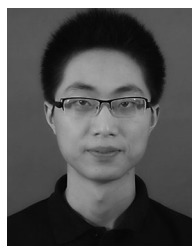
In this paper, we studied entity-level multimodal sentiment classification, and proposed an entity-sensitive attention and fusion network to effectively model the intra-modality interactions including entity-text and entity-image alignments, and the inter-modality interactions, i.e., text-image alignments. Experimental results on one unimodal benchmark dataset and the two multimodal datasets demonstrate that our method can achieve the best performance in comparison with a number of unimodal and multimodal approaches.

The main limitation of our approach is the assumption that the entities in each sentence have been provided or extracted by existing named entity recognition tools. As a follow-up work, we plan to propose an end-to-end multimodal architecture to jointly extract entities and assign corresponding sentiment to each entity. We believe that end-to-end entity-level multimodal sentiment analysis is a promising direction in the future.

#### REFERENCES

- [1] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016.

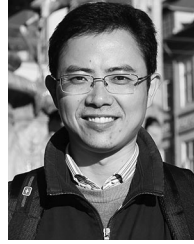
- [2] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2011, pp. 151–160.
- [3] J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua, "Aspect ranking: Identifying important product aspects from online consumer reviews," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2011, pp. 1496–1505.
- [4] D.-T. Vo and Y. Zhang, "Target-dependent Twitter sentiment classification with rich automatic features," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 1347–1353.
- [5] L. Deng and J. Wiebe, "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2015, pp. 179–189.
- [6] L. Dong *et al.*, "Adaptive recursive neural network for target-dependent Twitter sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2014, pp. 49–54.
- [7] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 2514–2523.
- [8] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. Int. Conf. Comput. Linguist., Tech. Papers*, 2016, pp. 3298–3307.
- [9] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 946–956.
- [10] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, 2018, Art. no. e1253.
- [11] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "NRC-canada-2014: Detecting aspects and sentiment in customer reviews," in *Proc. Int. Workshop Semantic Eval.*, 2014, pp. 437–442.
- [12] M. Pontiki *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.
- [13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [14] Z. Yang *et al.*, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2016, pp. 1480–1489.
- [15] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [16] D. Tang *et al.*, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2016, pp. 214–224.
- [17] Y. Yang *et al.*, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2016, pp. 606–615.
- [18] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 4068–4074.
- [19] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2017, pp. 452–461.
- [20] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, "Target-sensitive memory networks for aspect sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 957–967.
- [21] N. Majumder *et al.*, "IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2018, pp. 3402–3411.
- [22] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2018, pp. 3433–3442.
- [23] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2016, pp. 1042–1047.
- [24] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2015, pp. 2539–2544.
- [25] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2017, pp. 873–883.
- [26] A. Zadeh *et al.*, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2017, pp. 1103–1114.
- [27] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.
- [28] D. Borth, R. Ji, T. Chen, T. Breuel, and S. F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
- [29] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 367–376.
- [30] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.
- [31] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1071–1074.
- [32] M. Zhang, Y. Zhang, and D. T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3087–3093.
- [33] J. Liu and Y. Zhang, "Attention modeling for targeted sentiment," in *Proc. Conf. Eur. Ch. Assoc. Comput. Linguist.*, 2017, pp. 572–577.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
- [35] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.
- [36] J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha, and B. T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.
- [37] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [38] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 289–297.
- [39] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L. P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.
- [40] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 1990–1999.
- [41] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5674–5681.
- [42] J. Cohen, "A coefficient of agreement for nominal scales," *Edu. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [43] O. Owoputi *et al.*, "Improved part-of-speech tagging for online conversational text with word clusters," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2013, pp. 380–390.
- [44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [45] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguist., Hum. Lang. Technol.*, 2018, pp. 2122–2132.
- [46] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 371–378.



**Jianfei Yu** received the Ph.D. degree from Singapore Management University, Singapore, in 2018, and the B.Sc. and M.Eng. degrees from Nanjing University of Science and Technology, Nanjing, China, in 2012 and 2015, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include natural language processing, sentiment analysis, information extraction, and question answering.



**Jing Jiang** received the Ph.D. degree from the Department of Computer Science, University of Illinois at Urbana Champaign, IL, USA, in 2008. She is currently an Associate Professor with the School of Information Systems, Singapore Management University, Singapore. Her research interests include natural language processing, information extraction, and social media analysis.



**Rui Xia** received the B.Sc. degree from Southeast University, Nanjing, China, in 2004, the M.Sc. degree from East China University of Science and Technology, Shanghai, China, in 2007, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. His research interests include natural language processing, machine learning, and data mining.