Adapting BERT for Target-Oriented Multimodal Sentiment Classification

Jianfei Yu[1,2] and **Jing Jiang**[2,*]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]School of Information Systems, Singapore Management University, Singapore
{jfyu, jingjiang}@smu.edu.sg

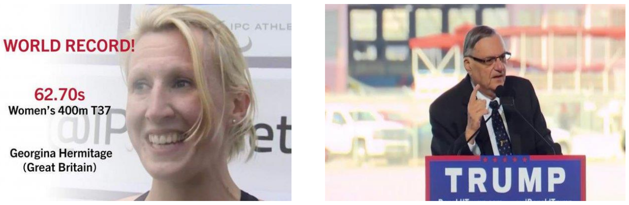
Abstract

As an important task in Sentiment Analysis, Target-oriented Sentiment Classification (TSC) aims to identify sentiment polarities over each opinion target in a sentence. However, existing approaches to this task primarily rely on the textual content, ignoring the other increasingly popular multimodal data sources (e.g., images), which can enhance the robustness of these text-based models. Motivated by this observation and inspired by the recently proposed BERT architecture, we study Target-oriented Multimodal Sentiment Classification (TMSC) and propose a multimodal BERT architecture. To model intra-modality dynamics, we first apply BERT to obtain target-sensitive textual representations. We then borrow the idea from self-attention and design a target attention mechanism to perform target-image matching to derive target-sensitive visual representations. To model inter-modality dynamics, we further propose to stack a set of self-attention layers on top to capture multi-modal interactions. Experimental results show that our model can outperform several highly competitive approaches for TSC and TMSC[1].

1 Introduction

Target-oriented Sentiment Classification (TSC) is a fundamental task in sentiment analysis, which aims to detect sentiment orientations over individual opinion targets mentioned in a sentence [Liu, 2012]. For example, given a tweet "*Georgina Hermitage is a #one2watch since she broke the 400m T37 WR.*", the user expresses **positive** and **neutral** sentiments towards *Georgina Hermitage* and *400m T37*, respectively.

To address this TSC problem, various supervised learning techniques empowered with both shallow and deep textual features have been proposed [Jiang *et al.*, 2011; Kiritchenko *et al.*, 2014; Dong *et al.*, 2014; Vo and Zhang, 2015; Zhang *et al.*, 2016]. With the recent trend of attention mechanism [Tang *et al.*, 2016b], many studies have proposed different attention-based neural architectures to model the interactions between opinion targets and their surrounding context words [Wang *et al.*, 2016; Wang *et al.*, 2018], which can further improve the state-of-the-art performance on several benchmark datasets [Li *et al.*, 2018].



(a). **[Georgina Hermitage]**$_{positive}$ is a #one2watch since she broke the **[400m T37]**$_{neutral}$ WR.

(b). **[Arizona]**$_{neutral}$ sheriff **[Joe Arpaio]**$_{negative}$ found in contempt in racial profiling case.

Figure 1: Representative examples for Target-Oriented Multimodal Sentiment Classification in our Twitter datasets. Opinion targets and their corresponding sentiment polarities are highlighted.

However, all these aforementioned approaches suffer from two common limitations. First, most of them only randomly initialize their model parameters, which may lead to sub-optimal solutions by only optimizing them with a small, task-specific corpus. With the recent trend of pre-training model parameters with unsupervised language models for various NLP tasks [Peters *et al.*, 2018; Howard and Ruder, 2018], it is natural to expect that these well initialized models can capture each word's semantic and syntactic meaning in different contexts and lead to better solutions for TSC.

More importantly, these existing methods primarily rely on textual content, and fail to consider the other associated data sources (e.g., images), which may potentially complement the textual content and enhance these text-based models. As the user-generated content on the Web (e.g., tweets, reviews) are increasingly multimodal, we observe that the associated images are generally useful for TSC for a couple of reasons. First, a user-generated sentence often focuses on one opinion target, and the associated image tends to highlight the focused target (e.g., *Georgina Hermitage* in Fig.1.a and *Joe Arpaio* in Fig.1.b). Second, it is sometimes hard to detect the sentiment over the focused target due to the short and informal natural of the sentence, but the associated image may help reflect a user's sentiment over the focused target (e.g., in Fig.1, the

[*] The corresponding author.

[1] We make our annotations for the two TMSC datasets publicly available via the link: https://github.com/jefferyYu/TomBERT.

two users respectively posted a pleasant image of *Georgina Hermitage* and an unpleasant image of *Joe Arpaio*). Third, for those remaining targets, the sentence often expresses *neutral* sentiment towards them, and the image also tends to pay less or even no attention to them (e.g., *400m T37* and *Arizona* in Fig.1). Therefore, it would be interesting to explore how to construct the alignment between opinion targets and textual/visual contents to model the intra-modality dynamics, and then fuse the textual and visual representations to uncover their inter-modality alignments in a unified model for Target-oriented Multimodal Sentiment Classification (TMSC).

To address these two limitations, in this paper, we build our model on top of the recent BERT architecture [Devlin *et al.*, 2018], whose pre-trained model parameters from a large corpus can help obtain contextualized word representations, and whose multi-head self-attention mechanism utilized in its Transformer encoder can automatically learn different levels of alignment between any two complex objects. Specifically, we first transform each input sentence into two sub-sentences: individual opinion target words and the remaining context words, and employ BERT to obtain target-sensitive textual representations. Moreover, inspired by the key idea of self-attention, we further design a target attention mechanism to automatically learn the alignment between opinion targets and images, where the targets are leveraged as queries to supervise the model to assign appropriate attention weights to different regions in the associated images to induce the target-sensitive visual representations. After modelling the intra-modality alignments, we further stack a set of self-attention layers on top of them to automatically capture their inter-modality interactions. We refer to this architecture as Target-oriented multimodal BERT or TomBERT for short.

Evaluations on three benchmark datasets for TSC and two manually annotated Twitter datasets for TMSC demonstrate the following: First, the fine-tuned BERT model outperforms the previously-reported best results on three benchmark datasets for TSC. Second, TomBERT can outperform both the state-of-the-art text-based approaches and highly competitive multimodal methods on the Twitter datasets for TMSC. Third, further analysis shows that due to the target-sensitive nature, TomBERT is especially useful when the input sentence has multiple opinion targets.

Our main contributions are summarized as follows:

- We devise a target-oriented multimodal BERT architecture for TMSC, where the two BERT-based modules at the bottom are used to capture intra-modality dynamics including target-text and target-image alignments, and another BERT-based module is stacked on top to capture inter-modality dynamics, i.e., text-image alignments.
- We propose to employ the standard BERT layer to model target-text and text-image alignments, and design a special target-image matching layer coupled with a target attention mechanism to model target-image alignments.

2 Related Work

2.1 Target-Oriented Sentiment Classification

As an important task in sentiment analysis, Target-oriented Sentiment Classification (TSC) has been extensively studied in recent years [Zhang *et al.*, 2018a]. One line of work focuses on leveraging external resources to manually design a set of task-specific features, followed by applying traditional statistical learning methods on the features for sentiment prediction [Kiritchenko *et al.*, 2014]. Another line of work centers on incorporating target information into various neural network (NN) models, including Recusive NN-based methods [Dong *et al.*, 2014], RNN-based methods [Tang *et al.*, 2016a] and CNN-based methods [Li *et al.*, 2018]. Inspired by the advantages of attention mechanisms in other NLP tasks, many recent studies design different attention-based methods to model the interactions between the target and the context [Wang *et al.*, 2018; Wang and Lu, 2018]. However, these studies fail to consider visual features that may boost these text-based approaches, which are the focus of this paper.

More recently, Xu et al. [Xu *et al.*, 2019] explored the task of aspect-level multimodal sentiment analysis by proposing a multi-hop memory network to model the cross-modality and single-modality interactions. Different from their work, we aim to explore the usefulness of the recent BERT model for both TSC and TMSC in this paper.

2.2 Multimodal Sentiment Classification

With the growth of multimodal data on the Web, information from different modalities (visual, acoustic, etc.) has recently been leveraged to provide complementary sentiment signals to traditional textual features [Zhang *et al.*, 2018a]. Most existing studies in this area focus on sentiment classification in a dialogue. Specifically, Poria *et al.* (2015) and Poria *et al.* (2017) respectively propose a multi-kernel learning approach and an LSTM-based sequential architecture to fuse the textual features, the visual features and the audio features. Following this line of work, Zadeh *et al.* (2017) and Zadeh *et al.* (2018) further designed a tensor fusion network and a memory fusion network to better capture the interactions between different modalities. However, these methods are designed for the coarse-grained dialogue sentiment classification, which might not be quite effective for our fine-grained target-oriented sentiment classification.

3 Methodology

In this section, we first formulate our task. We then review the standard BERT model, and present our two multimodal extensions of BERT, i.e., mBERT and TomBERT.

3.1 Task Definition

We are given a set of multimodal samples D. For each sample $c \in D$, it contains a sentence S with n words $(w_1, \ldots, w_n)$ and an associated image $\mathbf{I}$, as well as an opinion target T (a sub-sequence of words in S). For the opinion target T, it is also associated with a sentiment label y, which can be either *positive*, *negative*, or *neutral*. Our problem can be stated as follows: given D as training corpus, our goal is to learn a target-oriented sentiment classifier so that it can correctly predict sentiment labels for opinion targets in unseen samples.

3.2 Background

As mentioned before, since the BERT [Devlin *et al.*, 2018] model can help derive contextualized word representations

Opinion Target	Sentence Input Examples for BERT in TSC
Georgina Hermitage	**[CLS]** \$T\$ is a #one2watch since she broke the 400m T37 WR. **[SEP]** Georgina Hermitage **[SEP]**
400m T37	**[CLS]** Georgina Hermitage is a #one2watch since she broke the \$T\$ WR. **[SEP]** 400m T37 **[SEP]**

Table 1: BERT input for TSC. T indicates the target position.

with pre-trained model parameters from a large corpus and enjoy the capability of learning alignment between two arbitrary inputs, we leverage it as the base model for our task.

To employ BERT for TSC, we propose to transform each sentence S into two sub-sentences: the opinion target T and the remaining context C, and concatenate them as the input sequence for BERT. For example, the BERT input for Fig.1.a is given in Table 1. Formally, let us use $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ to denote the transformed input sequence, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input representation by summing up the word, segment, and position embeddings, and N is the maximum length of the sequence.

Next, let us briefly review the BERT model [Devlin *et al.*, 2018], which is essentially a multi-layer bi-directional Transformer encoder [Vaswani *et al.*, 2017] illustrated in the sentence encoder part of Fig.2.a.

To capture the global information, an m-head self-attention layer is first used to transform each position in the input sequence into a weighted sum of the input layer. Specifically, for the i-th head attention, the input layer $\mathbf{X} \in \mathbb{R}^{d \times N}$ is transformed based on the dot-product attention mechanism [Luong *et al.*, 2015] as follows:

$$\mathrm{ATT}_i(\mathbf{X}) = \mathrm{softmax}(\frac{[\mathbf{W}_{\mathbf{Q}_i}\mathbf{X}]^\top [\mathbf{W}_{\mathbf{K}_i}\mathbf{X}]}{\sqrt{d/m}})[\mathbf{W}_{\mathbf{V}_i}\mathbf{X}]^\top, \qquad (1)$$

where $\{\mathbf{W}_{\mathbf{Q}_i}, \mathbf{W}_{\mathbf{K}_i}, \mathbf{W}_{\mathbf{V}_i}\} \in \mathbb{R}^{d/m \times d}$ are learnable parameters corresponding to queries, keys and values respectively. Then, the outputs of the m attention mechanisms are concatenated together followed by a linear transformation as below:

$$\mathrm{MATT}(\mathbf{X}) = \mathbf{W}_m[\mathrm{ATT}_1(\mathbf{X}), \ldots, \mathrm{ATT}_m(\mathbf{X})]^\top, \qquad (2)$$

where $\mathbf{W}_m \in \mathbb{R}^{d \times d}$ is the parameter to learn[2].

Based on the output from the self-attention layer, BERT adds a residual connection from the input to the output, followed by a layer norm (LN) [Ba *et al.*, 2016] as follows:

$$\mathbf{Z} = \mathrm{LN}(\mathbf{X} + \mathrm{MATT}(\mathbf{X})). \qquad (3)$$

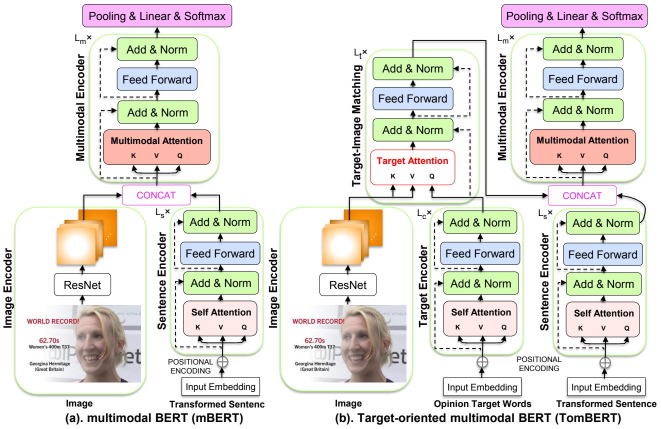
Moreover, a standard feed-forward network (a.k.a MLP) with GeLU [Hendrycks and Gimpel, 2016] as the activation function and another residual connection with layer norm are stacked on top to generate the output of the first BERT layer:

$$\mathrm{BT}(\mathbf{X}) = \mathrm{LN}(\mathbf{X} + \mathrm{MLP}(\mathbf{Z})). \qquad (4)$$

Finally, the entire model stacks L_s such BERT layers, and the final hidden state of the first token (i.e., [CLS]) is fed to a linear transformation function for classification.

3.3 Multimodal BERT (mBERT)

As illustrated in Fig.2.a, an intuitive but general solution to incorporate the associated image into the BERT architecture is to directly concatenate the image features with the final hidden states of the input sequence, followed by stacking additional BERT layers on top to model the inter-modal interactions between visual and textual representations.



(a). multimodal BERT (mBERT)

(b). Target-oriented multimodal BERT (TomBERT)

Figure 2: Overview of our multimodal BERT models for TMSC.

Specifically, for the associated image $\mathbf{I}$, we first resize it to 224×224 pixels, and then adopt one of the state-of-the-art image recognition models ResNet-152 (res5c) [He *et al.*, 2016] to obtain the output of the last convolutional layer:

$$\mathrm{ResNet}(\mathbf{I}) = \{\mathbf{r}_j | \mathbf{r}_j \in \mathbb{R}^{2048}, j = 1, 2, ..., 49\}, \qquad (5)$$

which essentially splits the original image into $7 \times 7 = 49$ regions and each region is represented by a 2048-dimensional vector $\mathbf{r}_j$. Next, we use a linear transformation function to project the visual features to the same space of textual features: $\mathbf{G} = \mathbf{W}_v \mathrm{ResNet}(\mathbf{I})$, where $\mathbf{W}_v \in \mathbb{R}^{d \times 2048}$ is the learnable parameter.

Furthermore, the image features $\mathbf{G} \in \mathbb{R}^{d \times 49}$ and the textual features $\mathbf{H_S} \in \mathbb{R}^{d \times N}$ (i.e., the hidden states in the final layer, $\mathrm{BT}_{L_s}(\mathbf{X})$) are concatenated together, followed by feeding them to the Multimodal Encoder, which contains another set of BERT layers to automatically model the rich interactions between textual and visual features:

$$\mathrm{ME}(\mathbf{G}, \mathbf{H_S}) = \mathrm{BT}_{L_m}([\mathbf{G}, \mathbf{H_S}]), \qquad (6)$$

where L_m is the number of layers in the Multimodal Encoder. Finally, the final hidden state of the "[CLS]" token is used for classification.

3.4 Target-oriented mBERT (TomBERT)

Although the above mBERT model is expected to well capture the inter-modality dynamics, its main limitation lies in the fact that its visual representation is insensitive to the opinion target, since the visual features for the same input sentence are always the same regardless of the target it considers.

Intuitively, with a specific opinion target as input, it is often the case that only some regions of the associated image are closely related to it, and the other regions should be ignored to eliminate the noise. For example, in Fig.1.a, with "*Georgina Hermitage*" as the target, we should mainly focus on her smiley face, and ignore the other background. In contrast, with "*400m T37*" as the target, we should only focus on the regions containing "*Women's 400m T37*". In this case, if our model takes the entire image into account and mistakenly

[2] We ignore the bias term to avoid confusion in this paper.

pays attention to the smiley face, it is highly probable that the model will make an incorrect prediction.

Inspired by this, we design a Target-Image (TI) Matching layer, which employs an m-head target attention mechanism to perform matching between the target and the image to obtain a target-sensitive visual representation. As shown in Fig.2.b, we first apply another BERT encoder to get the hidden representation of the opinion target: $\mathbf{H_T} = \text{BT}_{L_c}(\mathbf{T})$, where $\mathbf{T} \in \mathbb{R}^{d \times M}$ and M are respectively the target input representation and the maximum length of the target, and L_c is the number of layers. Next, we treat the hidden states of the target $\mathbf{H_T}$ as queries, and the regional image features $\mathbf{G}$ as keys and values, such that the target is leveraged to guide the model to align it with the appropriate regions, i.e., only assigning high attention weights to the image regions that are closely related to the target. Specifically, the i-th head target attention takes the following form:

$$\text{ATT}_i(\mathbf{G}, \mathbf{H_T}) = \text{softmax}(\frac{[\mathbf{W_Q}'_i \mathbf{H_T}]^\top [\mathbf{W_K}'_i \mathbf{G}]}{\sqrt{d/m}})[\mathbf{W_V}'_i \mathbf{G}]^\top, \tag{7}$$

where $\{\mathbf{W_Q}'_i, \mathbf{W_K}'_i, \mathbf{W_V}'_i\} \in \mathbb{R}^{d/m \times d}$ are parameters. Furthermore, similar as BERT, we adopt a feed-forward network and two layer norms with residual connections to obtain the target-sensitive visual output:

$$\text{TI}(\mathbf{G}, \mathbf{T}) = \text{LN}(\mathbf{H_T} + \text{MLP}(\text{LN}(\mathbf{H_T} + \text{MATT}(\mathbf{G}, \mathbf{H_T})))). \tag{8}$$

We then stack L_t such TI Matching layers to obtain the final visual representation: $\mathbf{H_V} = \text{TI}_{L_t}(\mathbf{G}, \mathbf{T})$, where $\mathbf{H_V} \in \mathbb{R}^{d \times M}$ and each hidden state in $\mathbf{H_V}$ is essentially a weighted sum of the 49 regions in the associated image.

Next, to form the multimodal input representations, we consider two concatenation types as below:

- **All-Text**: directly concatenate $\mathbf{H_V}$ and $\mathbf{H_S}$;
- **First-Text**: only consider the final state of the first element (i.e., the special [CLS] token in the target input) in $\mathbf{H_V}$, and concatenate $\mathbf{H_V^0}$ with $\mathbf{H_S}$.

Similar to mBERT, we further add a multimodal encoder on top to obtain the final multimodal hidden representations:

$$\mathbf{H} = \text{ME}(\mathbf{H_V}, \mathbf{H_S}) \quad \text{or} \quad \mathbf{H} = \text{ME}(\mathbf{H_V^0}, \mathbf{H_S}). \tag{9}$$

To integrate the visual and textual representations for final classification, we consider the following three pooling types to obtain the final output:

- **FIRST**: The first token of the multimodal input sequence is always a weighted sum of the 49 regional image features. Its final hidden state is regarded as the aggregate multimodal sequence representation with visual representations as queries, and thus can be taken out as the output: $\mathbf{O} = \mathbf{H}^0$;
- **CLS**: Similarly, the final hidden state for the special token (i.e., [CLS] token in the sentence input) is the aggregate representation with textual representations as queries, and can also be used as the output: $\mathbf{O} = \mathbf{H}^{\text{[CLS]}}$;
- **BOTH**: We concatenate these two hidden states as the hybrid output: $\mathbf{O} = [\mathbf{H}^0, \mathbf{H}^{\text{[CLS]}}]$.

Finally, we feed $\mathbf{O}$ to a linear function followed by a softmax function for target-oriented sentiment classification:

$$p(y|\mathbf{O}) = \text{softmax}(\mathbf{W}^\top \mathbf{O}), \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{(2)d \times 3}$ is the learnable parameter. To optimize all the parameters in our TomBERT model, the objective is to minimize the standard cross-entropy loss function as below:

$$\mathcal{J} = -\frac{1}{|D|} \sum_{j=1}^{|D|} \log p(y^{(j)} | \mathbf{O}^{(j)}). \tag{11}$$

4 Experiments

In this section, we carry out extensive experiments to answer the following research questions:

- **RQ1**: Can the fine-tuned BERT model outperform state-of-the-art text-based approaches on both existing benchmark datasets and our two datasets? (Section 4.2)
- **RQ2**: Is the associated image generally useful for TSC? Could our TomBERT model bring significant improvements to BERT and achieve the best performance on our two multimodal datasets? (Section 4.2)
- **RQ3**: What is the effectiveness of TomBERT with respect to the pooling layer, the multimodal concatenation layer, and the number of hidden layers? (Section 4.3)
- **RQ4**: What is the key advantage of TomBERT over other highly competitive approaches? (Section 4.4)

4.1 Experiment Settings

To evaluate the effect of *BERT* and *TomBERT*, we adopt three benchmark datasets for TSC (i.e., `LAPTOP` and `REST` from SemEval-2014 Task 4 [Pontiki *et al.*, 2014] as well as `TWITTER-14` constructed by [Dong *et al.*, 2014]) and two multimodal datasets for TMSC (i.e., `TWITTER-15` and `TWITTER-17` respectively collected by [Zhang *et al.*, 2018b] and [Lu *et al.*, 2018]). `LAPTOP` and `REST` consist of Amazon customer reviews in laptop and restaurant domains respectively, and the three `TWITTER` datasets include user tweets posted during 2010-2014, 2014-2015 and 2016-2017, respectively. Since the two publicly available multimodal datasets `TWITTER-15` and `TWITTER-17` only provide annotated targets (i.e., entities) in each tweet, we ask three domain experts to annotate the sentiment towards each target, and take the majority label among the three annotators as the gold label. For space limitation, we only show the basic statistics of `TWITTER-15` and `TWITTER-17` in Table 2.

We build our *TomBERT* model on top of the pre-trained uncased *BERT*$_{base}$ model released by [Devlin *et al.*, 2018], and tune the hyper-parameters on the development set of each dataset. Specifically, for BERT-based models, we set the learning rate as 5e-5, the number of attention heads as m = 12, and the dropout rate as 0.1. The batch size is respectively set as 16 and 32 for all the models for TSC and TMSC, respectively. Besides, for *TomBERT*, the maximum length of the sentence input and the target input are respectively set as N = 64 and M = 16. The number of layers for encoding the sentence input and the target input are both set to be 12, i.e., L_s = L_c = 12, where the parameters are both initialized from the pre-trained *BERT*$_{base}$ model. All the models are fine-tuned for 8 epochs, and are implemented based on PyTorch with a NVIDIA Tesla V100 GPU.

	TWITTER-15							TWITTER-17						
	#POS	#NEG	#Neutral	Total	#Avg Targets	Words	Length	#POS	#NEG	#Neutral	Total	#Avg Targets	Words	Length
Train	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	317	113	607	1037	1.354	3919	17.05	493	168	573	1234	1.450	3013	16.38

Table 2: The basic statistics of our two multimodal Twitter datasets. *POS* and *NEG* are short for *positive* and *negative* classes.

Method	LAPTOP		REST		TWITTER-14	
	ACC	Mac-F_1	ACC	Mac-F_1	ACC	Mac-F_1
SVM	70.49	-	80.16	-	63.40	63.30
AE-LSTM	68.90	-	76.60	-	-	-
IAN	72.10	-	78.60	-	-	-
TD-LSTM	71.83	68.43	78.00	66.73	66.62	64.01
MemNet	70.33	64.09	78.16	65.83	68.50	66.91
RAM	75.01	70.51	79.79	68.86	71.88	70.33
TNet-LF	76.01	71.47	80.79	70.84	74.68	73.36
TNet-AS	76.54	71.75	80.69	71.27	74.97	73.60
MGAN	75.39	72.47	81.25	71.94	72.54	70.81
BERT	**76.96**	**73.67**	**84.29**	**77.22**	**75.14**	**74.15**

Table 3: Experimental results on three benchmark datasets for TSC. The results of compared systems are retrieved from previous papers.

Modality	Method	TWITTER-15		TWITTER-17	
		ACC	Mac-F_1	ACC	Mac-F_1
Visual	Res-Target	59.88	46.48	58.59	53.98
Text	AE-LSTM	70.30	63.43	61.67	57.97
	MemNet	70.11	61.76	64.18	60.90
	RAM	70.68	63.05	64.42	61.01
	MGAN	71.17	64.21	64.75	61.46
	BERT	74.15	68.86	68.15	65.23
	BERT+BL	**74.25**	**70.04**	**68.88**	**66.12**
Text + Visual	Res-MGAN	71.65	63.88	66.37	63.04
	Res-MGAN-TFN	70.30	64.14	64.10	59.13
	Res-BERT+BL	75.02	69.21	69.20	66.48
	Res-BERT+BL-TFN	73.58	68.74	67.18	64.29
	mBERT (All-Text)	74.86	69.01	69.61	67.12
	mPBERT (FIRST)	69.62	63.67	65.56	63.20
	mPBERT (CLS)	**75.79**	**71.07**	68.80	67.06
	mPBERT (BOTH)	75.31	70.18	**69.61**	**67.12**
	TomBERT (All-Text)	76.37†	72.60†	69.61	67.48
	TomBERT (FIRST)	**77.15**†	**71.75**†	70.34†	68.03†
	TomBERT (CLS)	76.57†	71.17†	69.69	67.75
	TomBERT (BOTH)	76.18†	71.27†	**70.50**†	**68.04**†

Table 4: Experimental results on our two Twitter datasets for TMSC. The results of compared systems are based on our implementation. **BL** denotes for another BERT layer. For fair comparison with other BERT-based baselines, we only use one hidden layer in the TI Matching and Multimodal Encoder of m(P)BERT and TomBERT. † indicates that TomBERT is significantly better than all the compared methods with p-value $<$ 0.05 based on McNemar's significance test.

4.2 Main Results

Performance of Fine-Tuned BERT (RQ1)

To demonstrate the effect of the fine-tuned BERT for TSC, we first compare it with a number of representative approaches: 1). *SVM* [Kiritchenko *et al.*, 2014], including many carefully designed linguistic features; 2). *AE-LSTM* [Wang *et al.*, 2016], incorporating aspect embeddings and target-specific attention mechanism; 3). *TD-LSTM* [Tang *et al.*, 2016a], using two LSTMs to model the left context and the right context of the target respectively; 4). *IAN* [Ma *et al.*, 2017], proposing an interactive attention mechanism to model the interactions between the target and the context; 5). *MemNet* [Tang *et al.*, 2016b], applying a multi-hop attention mechanism on top of word embeddings and position embeddings with targets as queries; 6). *RAM* [Chen *et al.*, 2017], constructing a neural architecture by applying a GRU model on top of the representations obtained from multi-hop attention mechanism; 7). *TNet* [Li *et al.*, 2018], adapting CNN with target-specific transformation to integrate the target and the context; 8). *MGAN* [Fan *et al.*, 2018], building up a multi-grained attention network for fusing the target and the context.

We report the accuracy (ACC) and Macro-F_1 score of text-based methods on all the five datasets in Table 3 and the text modality part of Table 4. It is easy to find that *BERT* consistently outperforms all the baselines, which supports our first motivation that the pre-trained model can lead to better optimal solutions, and thus bring improvements for TSC.

Performance of TomBERT (RQ2)

We then consider the following highly competitive approaches for evaluating our *TomBERT* model: 1). *Res-Target*: concatenating $\mathbf{H_T}$ and the max-pooling of $\mathbf{G}$; 2). *BERT+BL*, adding another BERT layer on top of $BERT_{base}$, which is similar to *MBERT* but without visual features. 3). *Res-MGAN*, a simple combination of textual and visual contents by concatenating the max pooling of $\mathbf{G}$ with the hidden representation of *MGAN*; 4). *Res-MGAN-TFN*, using Tensor Fusion Network (TFN) [Zadeh *et al.*, 2017] to fuse the textual and visual representations in *Res-MGAN* with rich interactions; 5). *Res-BERT+BL* and *Res-BERT+BL-TFN*, replacing the textual encoder in *Res-MGAN* and *Res-MGAN-TFN* with *BERT+BL*; 6). *mBERT (All-Text)*, the model detailed in Section 3.3; 7). *mPBERT*, a variant of *mBERT*, which uses the max pooling of $\mathbf{G}$ as input visual features and one of the three pooling types to obtain the final output; 8). *TomBERT (All-Text)*, our model detailed in Section 3.4, where we use *All-Text* concatenation and *CLS* pooling for the final output; 9). *TomBERT (FIRST, CLS, or BOTH)*, our model with *First-Text* concatenation and one of the three pooling types for the final output.

Based on Table 4, we can make a couple of observations: 1). The perforamnce of *Res-Target* is quite limited, which implies that the associated images only play a supporting role to text, and cannot be treated independently for target-oriented sentiment prediction; 2). Due to the higher capacity of model parameters, *BERT+BL* brings minor improvements over *BERT*; 3). *Res-MGAN* and *Res-BERT+BL* can generally boost the performance of *MGAN* and *BERT+BL*, indicating that the associated images are generally useful to enhance text-based approaches; 4). Interestingly, although *TFN* learns rich interactions between modalities, it even drops the performance of *Res-MGAN* and *Res-BERT+BL*. This suggests that it is hard for the complex fusion matrix to directly capture the interactions between two modalities; 5). *mBERT*

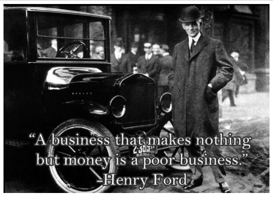
Associated Image	Input Sentence & Predicted Label	Associated Image	Input Sentence & Predicted Label
	Join **[iKON]**[1], at their 1 night only, first ever **[SG]**[2] concert. Go at 3pm! Human Label: (**1-positive**, **2-neutral**) BERT+BL: (**neutral** ✗, **positive** ✗) mPBERT: (**neutral** ✗, **positive** ✗) TomBERT: (**positive** ✓, **neutral** ✓)	 "A business that makes nothing but money is a poor business." -Henry Ford	Happy birthday to **[Henry Ford]**[1], who was born 150 years ago today ! (Photo: **[Getty Images]**[2]) Human Label: (**1-positive**, **2-neutral**) BERT+BL: (**positive** ✓, **positive** ✗) mPBERT: (**positive** ✓, **positive** ✗) TomBERT: (**positive** ✓, **neutral** ✓)

Table 5: Predictions of BERT+BL, mPBERT, and TomBERT on several test samples. ✗ and ✓ denote incorrect and correct predictions.

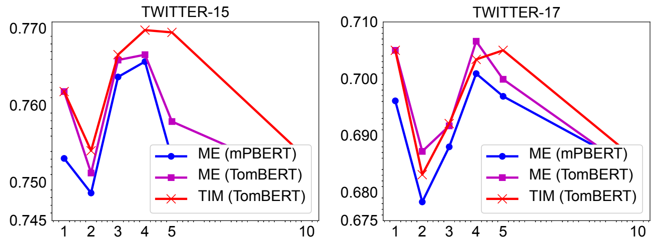


Figure 3: Comparisons of the number of hidden layers for TI Matching (TIM) and Multimodal Encoder (ME) Layers in mPBERT and TomBERT. For both methods, we use BOTH pooling in the top layer.

and *mPBERT* can outperform *BERT+BL* in most cases, and generally have better performance than *Res-BERT+BL-TFN*, which shows that the top multimodal encoder can well capture the inter-modality interactions; 6). Finally, regardless of the concatenation and pooling types we use, *TomBERT* consistently achieves the best results on the two datasets, and most of these gains are significant with $p<0.05$. These observations support our second motivation that *TomBERT* can well capture intra-modality and inter-modality dynamics.

4.3 Further Analysis of TomBERT (RQ3)

To answer the questions in **RQ3**, we investigate the effect of different components in both *mBERT* and *TomBERT*.

First, comparing the three pooling types in Table 4, we observe the following: 1). no matter which pooling type we use, *TomBERT* can generally perform better than *mPBERT*. Since the only difference between them is the TI Matching module, this suggests that our target attention mechanism is able to generate target-sensitive visual representations, which can lead to significant performance gain; 2). For *mPBERT*, since its visual representations are not target-sensitive, it is reasonable that using its final hidden state (i.e. *FIRST*) results in limited performance. In contrast, using *CLS* and *BOTH* can learn to pay more attention to textual representations, and have much better results; 3). For *TomBERT*, since visual and textual representations are both target-sensitive, it is intuitive that all the three pooling types can result in promising results.

Moreover, for the two concatenation types, we can see from Table 4 that incorporating all the visual features (i.e., *All-Text*) generally leads to slightly worse performance than abstracting all the visual features into a single vector except *mPBERT (FIRST)*. This further confirms that in TMSC, the images are used to support the text for target sentiment detection, and paying too much attention to the visual features may bring some noise and drop the performance.

Finally, since the Multimodal Encoder (ME) and TI Matching (TIM) modules in *mPBERT* and *TomBERT* may stack multiple layers, we analyze the impact of their layer number L_m and L_t. As shown in Fig. 3, for ME, *mPBERT* and *TomBERT* can achieve the best results when $L_m = 4$; while for TIM, *TomBERT* performs the best when L_t is around 5. When further increasing L_m and L_t, the result becomes worse probably due to the increase of model parameters.

Method	TWITTER-15		TWITTER-17	
	#targets $= 1$	#targets ≥ 2	#targets $= 1$	#targets ≥ 2
	(566 samples)	(471 samples)	(581 samples)	(653 samples)
BERT+BL	75.62	72.61	69.02	68.76
mPBERT	76.50	73.88	**69.88**	69.37
TomBERT	**78.80**	**75.16**	69.02	**71.52**

Table 6: Breakdown of Accuracy with respect to sentences with single opinion target and multiple opinion targets in our test set.

4.4 Case Study (RQ4)

To better understand the advantage of *TomBERT*, we further group test sentences with single target and multiple targets into two categories, and report the results on them in Table 6. It is easy to observe that *TomBERT* performs significantly better than *BERT+BL* and *mPBERT* when input sentences have multiple targets, which is in line with our motivation.

Furthermore, we select several representative test samples to compare the predictions of different methods. In the left side of Table 5, we can see that although *BERT+BL* and *mPBERT* incorrectly predict the sentiment over *IKON* and *SG*. With the help of the image, our *TomBERT* model can identify that the focus of the tweet is the band *IKON* other than *SG*, and therefore predict the sentiment over the focused target *IKON* as *positive*, and the other target *SG* as *neutral*. Similar observations can be made on the targets *Henry Ford* and *Getty Images* in the right side of Table 5.

5 Conclusion

In this paper, we study Target-oriented Multimodal Sentiment Classification (TMSC), and propose a Target-oriented multimodal BERT (TomBERT) architecture to effectively capture the intra-modality and inter-modality dynamics. Extensive evaluations on five datasets for TSC and TMSC demonstrate the effectivenss of BERT and our TomBERT model in detecting the sentiment polarity for individual opinion target.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Chen *et al.*, 2017] Peng Chen, Zhongqian Sun, et al. Recurrent attention network on memory for aspect sentiment analysis. In *EMNLP*, pages 452–461, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dong *et al.*, 2014] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL*, pages 49–54, 2014.

[Fan *et al.*, 2018] Feifan Fan, Yansong Feng, et al. Multi-grained attention network for aspect-level sentiment classification. In *EMNLP*, pages 3433–3442, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*, 2016.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, pages 328–339, 2018.

[Jiang *et al.*, 2011] Long Jiang, Mo Yu, et al. Target-dependent twitter sentiment classification. In *ACL*, pages 151–160, 2011.

[Kiritchenko *et al.*, 2014] Svetlana Kiritchenko, Xiaodan Zhu, et al. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *SemEval 2014*, pages 437–442, 2014.

[Li *et al.*, 2018] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. Transformation networks for target-oriented sentiment classification. In *ACL*, pages 946–956, 2018.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[Lu *et al.*, 2018] Di Lu, Leonardo Neves, et al. Visual attention model for name tagging in multimodal social media. In *ACL*, pages 1990–1999, 2018.

[Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[Ma *et al.*, 2017] Dehong Ma, Sujian Li, et al. Interactive attention networks for aspect-level sentiment classification. In *IJCAI*, pages 4068–4074, 2017.

[Peters *et al.*, 2018] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, pages 2227–2237, 2018.

[Pontiki *et al.*, 2014] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval*, pages 19–30, 2014.

[Poria *et al.*, 2015] Soujanya Poria, Erik Cambria, et al. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, pages 2539–2544, 2015.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, et al. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883, 2017.

[Tang *et al.*, 2016a] Duyu Tang, Bing Qin, et al. Effective LSTMs for target-dependent sentiment classification. In *COLING*, pages 3298–3307, 2016.

[Tang *et al.*, 2016b] Duyu Tang, Bing Qin, Ting Liu, et al. Aspect level sentiment classification with deep memory network. In *EMNLP*, pages 214–224, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Vo and Zhang, 2015] Duy-Tin Vo and Yue Zhang. Target-dependent twitter sentiment classification with rich automatic features. In *IJCAI*, pages 1347–1353, 2015.

[Wang and Lu, 2018] Bailin Wang and Wei Lu. Learning latent opinions for aspect-level sentiment classification. In *AAAI*, pages 5537–5544, 2018.

[Wang *et al.*, 2016] Yequan Wang, Minlie Huang, Li Zhao, et al. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, pages 606–615, 2016.

[Wang *et al.*, 2018] Shuai Wang, Sahisnu Mazumder, et al. Target-sensitive memory networks for aspect sentiment classification. In *ACL*, pages 957–967, 2018.

[Xu *et al.*, 2019] Nan Xu, Wenji Mao, and Guandan Chen. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *AAAI*, 2019.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, et al. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114, 2017.

[Zadeh *et al.*, 2018] Amir Zadeh, Paul Pu Liang, et al. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641, 2018.

[Zhang *et al.*, 2016] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. Gated neural networks for targeted sentiment analysis. In *AAAI*, pages 3087–3093, 2016.

[Zhang *et al.*, 2018a] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page p.e1253, 2018.

[Zhang *et al.*, 2018b] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*, pages 5674–5681, 2018.