Interactive POS-aware network for aspect-level sentiment classification

Kai Shuang [a,*], Mengyu Gu [a], Rui Li [a], Jonathan Loo [b], Sen Su [a]

[a] State Key Laboratory of Networking & Switching Technology, Beijing University of Posts and Telecommunications, 100876 Beijing, PR China
[b] School of Computing and Engineering, University of West London, W55RF, UK

ARTICLE INFO

Article history:
Received XXX
Revised XXX
Accepted XXX

Keywords:
Aspect-level sentiment classification
Part-of-speech
Gating mechanism
Attention mechanism

ABSTRACT

Existing aspect-level sentiment-classification models completely rely on the learning from given datasets. However, these are easily misled by biased samples, resulting in learning some ill-suited rules that limit their potential. The information of some specific part-of-speech (POS) categories often indicates the word sentiment polarity, which can be introduced as prior knowledge to facilitate prediction of the model. Accordingly, we propose an interactive POS-aware network (IPAN) that explicitly introduces the POS information as reliable guidance to assist the model in accurately predicting sentiment polarity. We distinguish the information of different POS categories using a POS-filter gate and reinforce the features extracted from adjectives, adverbs, and verbs via a POS-highlighting attention mechanism. This enables the model to concentrate on the words that contain significant sentiment orientations and to obtain the most practical learning experience. To emphasize the target information, we construct a target-context gate that enables the interaction of the target information with contexts; consequently, the model considerably focuses on target-related sentiment features. The experiments on SemEval2014 and Twitter datasets verify that our IPAN consistently outperforms the current state-of-the-art methods.

1. Introduction

Sentiment analysis, also known as opinion mining [1,2], is an essential task in natural language processing (NLP). Notably, aspect-level sentiment classification, as a fine-grained task in sentiment analysis [2], has received considerable attention in recent years. Specifically, given a sentence and a target that appears in it, the task aims to determine the sentiment polarity of the sentence toward the target. For example, given the targets {*place, food*} and the sentence "*While this is a pretty place in that overly cute French way, the food was insultingly horrible*." For both the targets, the sentiment polarities are *positive* and *negative*, respectively.

In recent years, with the advancement of deep-learning methods, various neural models have performed notably in aspect-level sentiment-classification tasks [3–7]. However, the noises in some biased training samples has limited the effectiveness of these models. Supposedly, if we can provide some strong prior knowledge to a model during its learning, it would gain clear indications regarding which words or spans are critical to expressing the sentiment, thereby improving its prediction accuracy. Accordingly, the part-of-speech (POS) information, as the basic building block of grammar, becomes an appropriate choice as prior knowledge, as it helps us analyze a sentence and satisfactorily understand the structure thereof [8]. Importantly, some previous researches in subjective text analysis and sentiment analysis [9–18] concluded that certain POS categories would be the strong indicators of sentiments.

However, in introducing the POS information, two technical challenges arise. First, how to reasonably model the POS information for indicating the word sentiment polarity; second, how to adequately emphasize the target information to extract its relevant sentiment features. Some attempts were made to consider the POS information as a feature for sentiment analysis, and they can be broadly classified as follows: (i) One part of the relevant researches did not deeply analyze the specific influence of each POS category on the sentiment expressions of the contexts. Some of them [9,13,19] only intuitively introduced the information of a few POS categories that contributed to the sentiment, without any theoretical support or experimental proof. However, other studies [12,20,21] blindly used information from all the POS categories and unavoidably introduced noises into the model. (ii) Although the other part of the relevant researches [14,17,18,22] explored the relationship between the POS information and emotional

* Corresponding author.
E-mail addresses: shuangk@bupt.edu.cn (K. Shuang), pattygu0622@bupt.edu.cn (M. Gu), lirui@bupt.edu.cn (R. Li), jonathan.loo@uwl.ac.uk (J. Loo), susen@bupt.edu.cn (S. Su).

expressions, their modeling methods for POS information were significantly rough, and thus the useful features contained in this kind of information could not be completely utilized, significantly limiting the final model performance. For the second challenge, many works [3,4] merely used the target embedding to capture important context words by using an attention mechanism. However, the target information could not completely interact with the contexts only in this manner, resulting in suboptimal performance.

Words with different POS categories diversely contribute to the semantic expression of sentences; therefore, only a limited number of POS categories exist with special meanings toward expressing sentiments. If all the kinds of POS are included in the modeling process, redundant information would be introduced, distracting the model during training. Intuitively, adjectives, adverbs, and verbs significantly contribute toward expressing the sentiment of a context. For example, in the sentence "*The sashimi portion are big enough to appease most people, but I did not like the fact they used artificial lobster meat*," the key phrases that indicate the sentiment polarity of targets *sashimi portion* and *artificial lobster meat* are "*big enough*" and "*did not like*, " respectively, which comprise adjective, verb, and adverb. This intuition is supported by many previous achievements: adjectives always contain a certain sentiment orientation [9–14]. Other POS categories have also showed some relevance to the emotional expressions: nouns were used in [14,15], verbs in [13,14]; additionally, adverbs proved to contribute to sentiments and subjectivity in [11,14,16]. Therefore, to give an active play to the POS information, we distinguish these three kinds of POS categories from other categories and emphasize the information of them during the modeling process. Particularly, a novel POS-highlighting attention mechanism with a limiting condition is designed for restricting the model to significantly concentrate on these three specific POS categories. To address the final target of this study, focusing only on how to determine the sentiment polarity of the entire sentence is not sufficient. Therefore, we must address the second technical challenge to further adapt the model to the aspect-level classification task. Our solution is to apply a fine-grained and element-wise gating mechanism, so that the target information completely interacts with each context word.

The proposed interactive POS-aware network (IPAN) comprises five components: the embedding layer, POS-filter gate, sentiment feature extraction layer, target-context gate, and POS-highlighting attention mechanism. The POS-filter gate and POS-highlighting attention mechanisms are designed for differentiating adjectives, adverbs, and verbs from other POS categories to prevent interferences from these useless words. In the target-context gate, we implement a new form of gate unit to filter out target-unrelated sentiment features, whose effectiveness has been experimentally proved.

The following are the main innovations and contributions of this study:

- We assert that the POS information can be used as prior knowledge to perform the aspect-level sentiment-classification task by guiding the learning process of the model. Accordingly, we propose a network (IPAN) that explicitly introduces and models the information of some specific POS categories that facilitate the model in extracting sentiment features. We have also conducted a set of heuristic experiments to explore the contribution of different POS categories to sentiment expressions;
- We designed a target-context gate to enable the complete interaction of the target information with each context word; the gate units in the target-context gate are of a new form called gated ReLU Units (GReU). To verify the effectiveness of this gating mechanism and novel gating units, we experimentally compared the novel gating units with the existing ones;
- We designed an innovative POS-highlighting attention mechanism to guarantee that the model significantly focused on adjectives, adverbs, and verbs. Similarly, we implemented and compared three other alternative attention mechanisms based on the POS information. The experimental results illustrated the effectiveness and rationality of this mechanism;
- Experimental results indicated that our IPAN consistently surpassed the existing state-of-the-art baselines on the widely used SemEval2014 and Twitter datasets;

The rest of this paper is organized as follows: Section 2 discusses the related works. Section 3 details the proposed IPAN. Section 4 provides the investigation results of the POS information on sentiment analysis and presents extensive experiments to verify the effectiveness and performance of IPAN. Finally, Section 5 summarizes our work and provides insights into future research.

2. Related work

2.1. Overview of some relevant works on sentiment analysis

Although most existing approaches regard sentiment analysis as a simple categorization task, it is a suitcase research problem that requires tackling many NLP tasks [23,24]. In particular, some studies [23] have counted and proved that at least 15 subproblems involved in achieving human-like performance in sentiment analysis. In this section, we list some common relevant works on sentiment analysis and present several latest achievements for them.

2.1.1. Affective computing and sentiment analysis

Affective computing is trying to assign computers the human-like capabilities of observation, interpretation, and generation of affect features [25]. Affective computing and sentiment analysis is the basis for realizing the emotional intelligence of machines. Existing approaches of this filed fall into three main categories: knowledge-based techniques, statistical methods [21,12,26–28], and hybrid approaches [29–38]. Knowledge-based techniques classify the text into some categories according to some affect words, whose sources of affect words or multiword expressions include the affective lexicon [39], linguistic annotation scheme [40], WordNet-Affect [41], SentiWordNet [42], SenticNet [43], and other probabilistic knowledge bases trained from linguistic corpora [44–46]. Hybrid approaches exploit both knowledge-based techniques and statistical methods to perform sentiment analysis on text or multimodal data and finally obtained better model performance.

2.1.2. Social data analysis

Online Social Network (OSN) is considered a spark that bursts the Big Data era. Dealing with the increasing amount of information present on the Web is a critical task and requires efficient models developed by the emerging field of sentiment analysis [47–49]. To this end, some current researchers [36] proposed efficient approaches to support polarity, emotion, and strength oriented sentiment analysis in natural language text. Main approaches to big social data analysis can be broadly grouped into two categories: knowledge-based techniques [50] and statistical methods [51]. While the former mainly leverage on ontologies [52], lexicons [41], semantic networks [43], or patterns [53], the latter are gradually shifting to the adoption of ELM, deep learning [54] and convolutional neural network [55,56].

2.1.3. Word sense disambiguation

The word sense disambiguation (WSD) task aims at identifying the meaning of words in a given context for specific words conveying multiple meanings. Turney and Littman et al. [57] claimed that

sentiment-ambiguous words cannot be avoided easily in a real-world application. To conclude the previous work on WSD, most of them made use of term-level contexts, such as words and patterns, and resolved the polarity with a range of rule-based [58], lexicons-based [59–61], statistics-based or machine learning methods [62–64]. Besides, neural architectures are the current state of the art in WSD. Duque et al. [65] presented a new graph-based unsupervised technique, in this work the knowledge base took the context of the ambiguous terms into account but only adapted to the specific domain. Bevilacqua et al. [66] proposed a neural supervised architecture that was able to tap into this wealth of knowledge by embedding information.

2.1.4. Sarcasm detection

Sarcasm is a nuanced form of language where usually the speaker explicitly states the opposite of what is implied. In general, approaches to sarcasm detection can be classified into rule-based, statistical, and deep learning-based approaches. Rule-based approaches attempt to identify sarcasm through specific evidence. Bharti et al. [67] presented two rule-based classifiers, the first used a parser-based lexicon generation algorithm and the other aimed to capture both hyperbolic sarcasm and intensifiers (such as "absolutely"). Most works in statistical sarcasm detection rely on different forms of Support Vector Machines (SVM) [68–70], some of them also make use of Naive Bayes and Decision Trees [71], binary Logistic Regression [72] and fuzzy Clustering [73]. Amir and Wallace et al. [74] presented a novel Convolutional Network, and Ghosh et al. [75] utilized a combination of a Convolutional Neural Network, a Recurrent Neural Network to finish this task. Majumder et al. [76] argued that sarcasm detection and sentiment analysis were correlated, and trained a multitask learning-based framework modeling this correlation to further improve the sentiment analysis performance.

2.1.5. Sentiment lexicons

Lexicons have been widely used for sentiment analysis, as they represent a simple, yet effective way to build rule-based opinion classifiers. For example, OpinionFinder distribution [77] compiled from manually developed resources augmented with entries learned from corpora, and SentiWordNet was a resource for opinion mining built on top of WordNet. As opposed to earlier lexicons, researchers have proposed some semantically rich lexicons, such as SenticNet [78,29] which modeled the sentiment of multiword expressions using commonsense knowledge derived from ConceptNet [79]. EmoSenticNet [80] implements to assign WordNet-Affect emotion labels to concepts in SenticNet with the assist of further fuzzy clustering and machine learning techniques. Different from the general-purpose emotion lexicons, Esuli et al. [42] proposed a generative unigram mixture model to learn a word-emotion and domain-specific emotion lexicons that offer more fine-grained estimates for word-emotion associations.

2.1.6. Other relevant works

There are also some existing works that incorporated two or more other NLP tasks to assist the model in detecting the sentiment polarity. Dragoni et al. [24] presented a commonsense ontology for sentiment analysis called OntoSenticNet, which can detect subtly expressed sentiments by enabling the analysis of multiword expressions. To address the problem that many applications of sentiment analysis obtained labeled data from multiple source domains, Xu et al. [81] adapted the source-domain training data to the target domain via a framework of multiclustering logistic approximation. Besides, the topics that evoke a certain emotion in readers are often context-sensitive, Rao [82] proposed a multi-label sentiment topic model, which can distinguish context-independent topics from both a background theme and a contextual theme. Yang and Rao et al. [83] introduced the assumption of "one segment expresses one sentiment" and proposed a segment-level joint topic-sentiment model to estimate the sentiment polarity of a document by capturing the topic-sentiment correlation.

Sentiment analysis task entails not only the NLP subtasks introduced above, but also all the subproblems of extracting semantic and emotional polarity from the text. In our real life, the categories of emotion are complex and variety, it is extremely challenging for machines to achieve human-like performance in identifying this. Therefore, most of the existing sentiment analysis tasks are to generally label the sentiment polarity in the text into fixed categories, and the models are finally completing a classification task. As we mentioned in the task definition in the first paragraph of Section 1, our research is also based on the assumption that the emotional polarity in the sentence is only positive, negative or neutral. Nonetheless, the above-mentioned related studies enable us to have a more comprehensive understanding of the real needs of sentiment analysis, and their techniques and modeling ideas also inspired our research.

2.2. POS information for sentiment analysis

For performing the sentiment analysis task or subjective text classification task (which classifies the pieces of a text as subjective or factual), there are several methodologies [9–12,19,84,85] that considered only certain POS categories as opinionated words, such as adjectives. Hatzivassiloglou et al. [11] developed an unsupervised learning system to learn the semantic orientation (positive or negative) of adjectives. Their system was based on the idea that adjectives connected by conjunctions likely had the same orientation, except for the ones connected by "but," as those connected by "but" likely had opposite orientations. In [11], the authors selected the sentences that contained either a gradable adjective from their list or an adjective identified in [13], without any classifier; the classification accuracy of the subjective sentences in their dataset was 72%.

Some works [12–18,20–22,87,86] did not discriminate among different POS categories, as they assumed that opinionated words could also occur in other POS categories, such as adverbs [16], verbs [13,86], and nouns [15]. Benamara et al. [16] used adverbs as clues for sentiment analysis. They proposed three alternative algorithms of scoring adverb–adjective combinations (AACs), and the average strength of the sentiment measure was obtained by summing the score (positive or negative) of all the AACs. Chesley et al. [13] used verbs to analyze the sentiments of blog posts. They classified verbs into four different sentiment classes and then used these classes, as well as some other features, in a support vector machine classifier to analyze the sentiment orientation of the blog posts. In [15], Riloff et al. presented an unsupervised method using extraction patterns to identify subjective nouns. The learned nouns, along with the adjectives from [10,11], were used as features for a naïve Bayes classifier, which classified subjective sentences with 81% precision and 77% recall. Pang et al. [21] appended POS tags to every word and used this information as a feature to the traditional n-gram approach. Because not all POS categories contribute to sentiment expressions, the results showed this method could not appropriately use the information. Yi et al. used all the nouns, verbs, adjectives, and adverbs in [14]. Their method involved a manually developed lexicon of sentiment expressions, and it achieved high precision but low recall. [17] used a maximum entropy modeling text classifier to classify the overall sentiment of documents and presented a method for weighting words based on POS to improve the classification performance; the experimental results showed that although their method boosted the baseline of sentiment analysis, it was limited

by the low learning ability of the model. [22] combined both the POS information and BERT model [88]; however, they did not have any selection process based on the contribution of POS categories to emotional expressions, resulting in significant redundant information and interference with the final prediction. [18] investigated the impact of POS tags on sentiment analysis; however they simply fed POS tags into the model rather than designing a process of transforming effective POS information into features that could be learned by the model.

2.3. Recent researches on aspect-level sentiment classification

Aspect-level sentiment analysis refers to three key issues: target extraction [89,90], sentiment analysis (the targets are provided), and joint the above two tasks [91]. In this study, we research only the second issue, which also consists of two main technical lines, namely, rule-based [59] and machine learning-based. The variants of recurrent neural networks (RNNs), such as long short-term memory (LSTM) [92], gated recurrent unit (GRU) [93], and the capsule network [94,95] have been widely used for aspect-level sentiment classification [96–100]. Tang et al. [96] employed a forward LSTM and a backward LSTM to model the left and right contexts of the aspect separately and then concatenated the context representations for prediction of the sentiment polarity. Yang et al. [97] jointly learned two subtasks: a domain classification and aspect-level sentiment classification to leverage the benefits of the supervised deep neural network as well as the unsupervised probabilistic generative model and further strengthen the representation learning in domain adaptation scenario. Yang et al. [98] proposed a new approach with the guidance of contextual, lexical, and syntactic cues, in which a new target representation sub-network was used to capture the semantic and contextual information of targets and a new dependence attention mechanism was utilized to model the syntactic dependency cues between targets and other words. Wang et al. [99] proposed the aspect-level sentiment capsules model (AS-Capsules), which utilized the correlation between aspect and sentiment to perform aspect detection and sentiment classification simultaneously, in a joint manner. Chen and Qian [100] proposed a Transfer Capsule Network (TransCap) model, which utilized an aspect routing approach and the dynamic routing approach to transfer document-level knowledge to aspect-level sentiment classification.

2.4. Pre-trained language models in aspect-level sentiment classification

Recently, pre-trained language models have proved effective in many NLP tasks by learning universal semantic rules using a significant amount of unlabeled data. Some of the prominent examples are ELMo [101], GPT [102], and BERT [88]. Especially, BERT has achieved remarkable performance in most relative downstream tasks. It is based on a multilayer bidirectional transformer and is trained on plain text for performing masked word prediction and next-sentence prediction tasks. A pre-trained BERT model can be fine-tuned for a downstream task using task-specific training data. Sun et al. [103] utilized BERT for the aspect-level sentiment-classification task by constructing an auxiliary sentence. Xu et al. [104] proposed a post-training approach for the ABSA task. Liu et al. [105] combined multi-task learning and pretrained BERT to improve the performance of various NLP tasks. Song et al. [106] used the intermediate layers of BERT to fine-tune it and applied this method in aspect-level sentiment classification.

2.5. Attention mechanism for aspect-level sentiment classification

The attention mechanism enables a model to learn which part of the text to focus on, and because the target information is significantly beneficial, some recent works directly used the target embedding to capture the importance of context words. Wang et al. [4] designed an attention-based LSTM to learn the target embedding and used it to compute attention weights. Ma et al. [5] interactively gained attention in the contexts and targets and separately generated representations for targets and contexts. Gu et al. [6] modeled the relation between a target and sentence by employing the bidirectional attention mechanism, as well as considering the position information of the target. Fan et al. [7] proposed a multi-grained attention network, which leveraged both fine- and coarse-grained attention mechanisms to compose its framework.

2.6. Gating mechanism for aspect-level sentiment classification

The gating mechanism controls the path through which information flows in the network, and it is efficient for recurrent networks [92,107–109]. Zhang et al. [110] generated the representations of contexts on both the sides of the target by employing GatedRNN and utilized gate units to model the relation between the target and both the side contexts. Inspired from this method, Liu et al. [111] replaced the vanilla RNN with the outputs of a contextualized attention model and further improved the model effectiveness. Xue and Li et al. [112] verified that the variants of these gate units, i.e., gated Tanh-Relu units (GTRU) outperformed the other gates in this task.

In conclusion, the previous works that used POS information to assist the model with sentiment analysis lacked in-depth research on the contribution of different POS categories toward the sentiment expressions of contexts, which blindly introduced several specific POS information selected only via human experiences. Additionally, most of these works used only POS tags as a part of the input without embedding it into the entire modeling process, wasting the information so that the final effects cannot meet expectations. Moreover, only a small number of models were used, and their performance improvements were hindered because of their limited learning abilities. After observing these problems, we not only explored the benefits of both a single POS category and POS-category combinations to the sentiment polarity classification task, but also designed a series of processes of converting the useful POS information into features that can be learned by neural networks with high fitting capacities. This makes the POS information a reliable guidance throughout the modeling process. Most existing BERT-based studies only appended an additional classifier after its original structure. However, considerable attempts have not been devoted toward performing aspect-level sentiment-classification tasks by combining the pretrained BERT model with other useful techniques. To exploit the semantic feature extraction ability of networks (such as BERT) and the ability of other effective NLP mechanisms to fuse and model various features, we utilized a gating mechanism to explicitly incorporate the POS information into the embedding of each word before the sentiment feature extraction layer. Subsequently, we designed another gate unit to enable the complete interaction of the target with contexts. Additionally, our POS-highlighting attention mechanism is designed on the idea that various POS categories differently contribute to the sentiment polarity, and a reasonable condition is proposed to guarantee that the weights of other POS categories are strictly lower than those of adjectives, adverbs, and verbs. Therefore, our mechanism is completely different from the previous attention mechanisms.

3. The proposed model : IPAN

The training process of the model is analogous to the learning process of human beings: if one has considerable experience, he will have some useful and straightforward guidance on how to accomplish a task while avoiding detours. Similarly, if we provide the POS information as prior knowledge to the model, it can gain some additional knowledge of determining how important each word is to emotional expressions. In this section, we detail the structure of our proposed IPAN, which explicitly models the POS information and emphasizes the target information, and its overall architecture is depicted in Fig. 1. As shown in Fig. 1, the main body

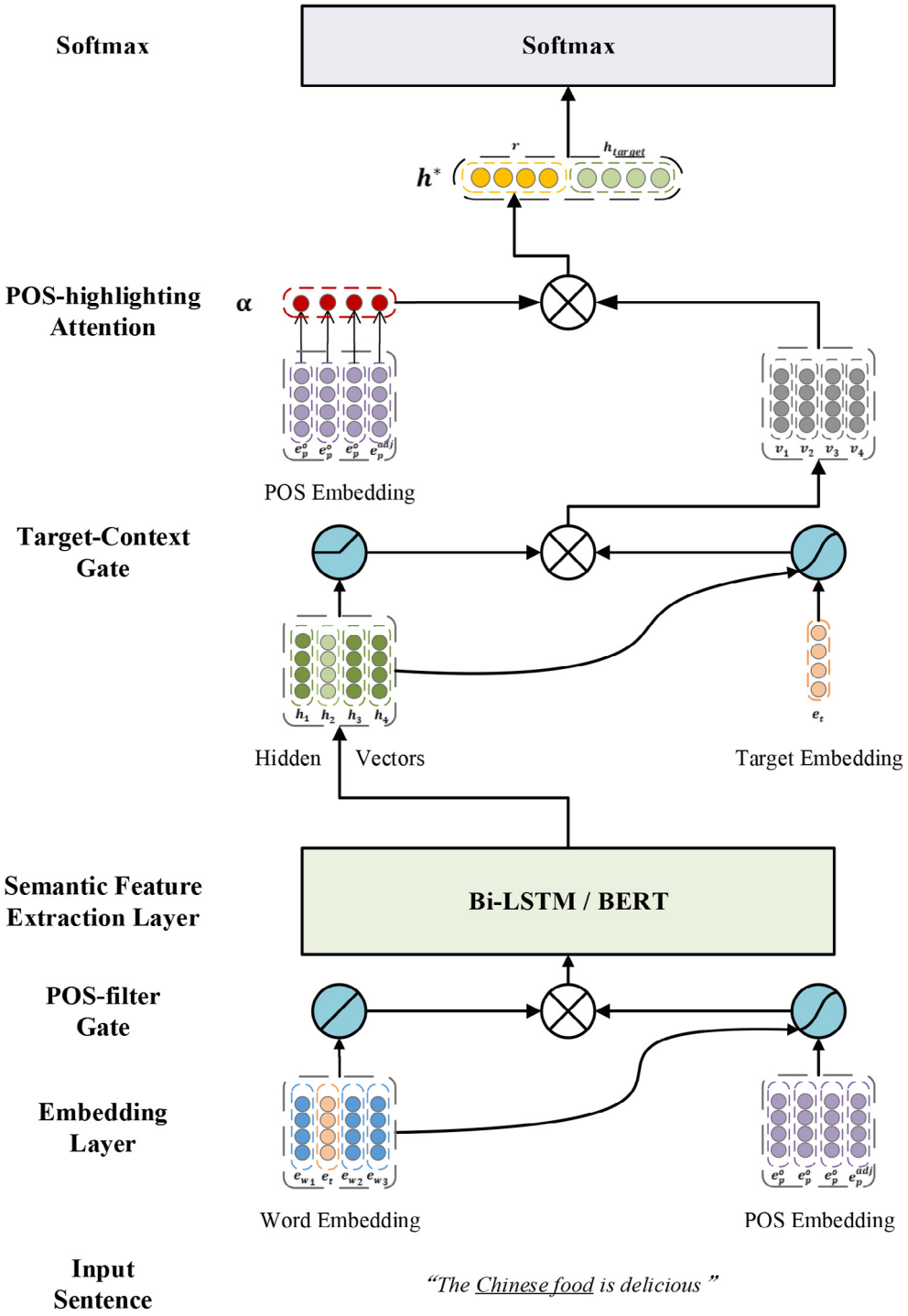

Fig. 1. Overall architecture of IPAN. The input sentence is "*The Chinese food is delicious*," and the target span is "*Chinese food*." After feeding into the embedding layer, the corresponding sequences of both the word and POS embeddings are obtained. The POS-filter gate generates a kind of POS-aware word embedding, and the semantic feature extraction layer extracts more advanced semantic features from its output. The target-context gate is a new form of gate unit, which emphasizes the target information and enables the interact thereof with each context word, and finally obtains target-related features. Accordingly, the novel POS-highlighting attention mechanism assigns different weights to words according to the contribution of their POS categories to emotional expressions, and thus the model can significantly focus on useful features while generating the final sentence representation. To adapt to the aspect-level classification task satisfactorily, the final sentence representation comprises the concatenation results of the weighted sentence vector and hidden vector h_{target} (h_3) that corresponds to the target.

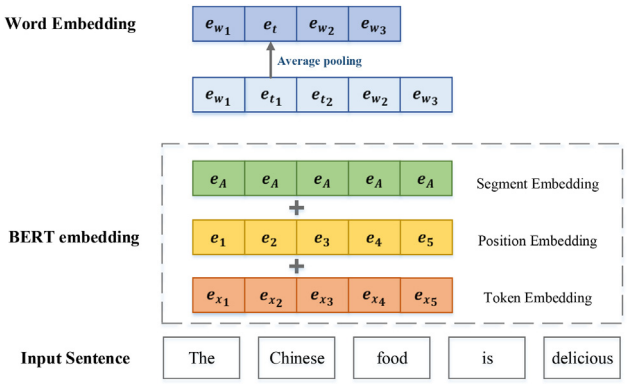

Fig. 2. IPAN-BERT word embedding. The original input representation are the sum of the token embeddings, the segment embeddings and the position embeddings, and segment embeddings are all represented by the same vector e_A.

of the network comprises the following five components, in the order of input to output: embedding layer, POS-filter gate, semantic feature extraction layer, target-context gate, and POS-highlighting attention mechanism. In the semantic feature extraction layer, bidirectional long short-term memory (Bi-LSTM) and BERT are applied as the base model, and we denote the corresponding two architectures as IPAN-LSTM and IPAN-BERT, respectively.

3.1. Embedding layer

Similar to most recent works in sentiment classification, we first map each word into a low-dimensional continuous vector. Specifically, let $E \in \mathbb{R}^{d \times v}$ be an embedding matrix that comprises all the word embeddings, where d denotes the dimensionality of word embedding and v the vocabulary size. Let a sentence contain $L+M-1$ words $[w_1, w_2, \ldots, t_1, t_2, \ldots, t_M, \ldots, w_{L-1}]$, and $[t_1, t_2, \ldots, t_M]$ represents the target spans (the target is usually not a single word). For the IPAN-LSTM, we obtain a sentence embedding $E_w \in \mathbb{R}^{d \times L}$ and target embedding $E_t \in \mathbb{R}^{d \times M}$ by looking up the embedding matrix E. Compared with the traditional Word2Vec- and GloVe-based embedding layers, which only provide a single context-independent representation of each token, the IPAN-BERT embedding layer takes a sentence as input and calculates the token-level representations using the information from the entire sentence. For a given token w of the input sentence, its input representation $e_w (w \in [1, L+M-1])$ is constructed by summing the corresponding token, segment, and position embeddings. A visualization of this construction is depicted in Fig. 2. Via this process, we can obtain the sentence embedding E_w and target embedding E_t of IPAN-BERT. To facilitate the subsequent operations, we compress $E_t \in \mathbb{R}^{d \times M}$ into a single target vector $e_t \in \mathbb{R}^{d \times 1}$ via average pooling, and the word-embedding vector of the input sentence is denoted by $[e_{w_1}, e_{w_2}, \ldots, e_t, \ldots, e_{w_{L-1}}]$. Regarding how to model the POS information of each word in the corresponding sentence, inspired by the target embedding in Wang et al. [4] and position embedding in Gu et al. [6], we designed and implemented a set of POS embeddings as follows.

In the processing of the input sentence, the POS of each piece of the original sentence in the datasets is first marked using the natural language toolkit (NLTK).[1] In our proposed model, POS is divided into four different categories: adjective (P_{adj}), adverb (P_{adv}), verb (P_{verb}) and others (P_{others}) categories. We used a POS-indexing dictionary to convert the POS tagging sequence obtained using NLTK into a series of POS-index sequences, whose lengths were equal to those of the input sentences (including punctuations), thereby facilitating the subsequent POS-embedding operation. For example, given a sentence "*The place is so cool and the service is prompt and thoughtful*," the target is *place*, and the POS-index sequence is represented as $p = [P_{others}, P_{others}, P_{verb}, P_{adv}, P_{adj}, P_{others}, P_{others}, P_{others}, P_{verb}, P_{adj}, P_{others}, P_{adj}, P_{others}]$.

The corresponding POS embedding of each element in p can be obtained by looking up a POS-embedding matrix $P = [e_p^{adj}, e_p^{adv}, e_p^{verb}, e_p^o] \in \mathbb{R}^{d_p \times N}$, which is randomly initialized and updated during the training process. For the target spans in p, their POS embedding is compressed to be able to use a uniform one-dimensional column vector $e_p^o \in \mathbb{R}^{d_p \times 1}$, which corresponds to the "others" category (P_{others}) in the POS-embedding matrix P. Here, d_p denotes the dimension of a POS embedding and N the four POS categories previously defined. The POS embedding (denoted as $E_p \in \mathbb{R}^{d_p \times L}$, where L equals the length of the word-embedding vector of the input sentence) aims to model the weight of each word with different POS in a sentence and make the model satisfactorily identify the distinctions between the words with different POS to more accurately predict results.

3.2. Proposed POS-filter gate

Before extracting high-level semantic features from the word embedding, we built a gating mechanism to explicitly incorporate the POS information into the embedding of each word, and we used gated linear units (GLUs) [108] to implement this gating mechanism.

$$S = E_w^T W_s + b_s \tag{1}$$

$$M = \text{sigmoid}\left(E_w^T W_m + E_p^T W_{m'} + b_m\right) \tag{2}$$

$$X = S \times M \tag{3}$$

where $W_s \in \mathbb{R}^{d \times d}, W_m \in \mathbb{R}^{d \times d}$, and $W_{m'} \in \mathbb{R}^{d_p \times d}$ denote weight matrixes, and $b_s \in \mathbb{R}^d$ and $b_m \in \mathbb{R}^d$ denote biases. Here, E_w^T and E_p^T denote the word and POS embeddings of the input sentence, respectively. The term S denotes the linear transformation of the word-embedding sequence, and M receives additional POS information with sigmoid activation function. The sigmoid gate is a common S-shaped function that maps inputs to $[0, 1]$; therefore, it can output a score regarding the correlation of the current POS and its corresponding word embedding. If this score is significantly high, the transformed word embedding $s_i \in \mathbb{R}^{1 \times d}$ $(i = 1, \ldots, L)$ would be accordingly amplified; otherwise, it would be blocked at the POS-filter gate.

A single word often belongs to multiple parts of speech, and thus its POS category can be uniquely determined only in a specific context. Therefore, the POS-filter gate mechanism was designed to filter out the irrelevant POS options for each text word. In other words, the gating mechanism selects the current specific type of POS for each word and then incorporates this additional information, which changes adaptively according to contexts, into the initial word representation. After passing through the POS-filter gate, the original word embedding E_w^T is converted into a set of tailor-made word vectors combined with the POS information, following which it is fed into the next layer to extract the semantic features.

3.3. Semantic Feature Extraction Layer

To capture advanced relationships among words and obtain the contextual representation, we fed a new word representation X, which was received upon the interaction of the original word

[1] NLTK is a leading platform for building Python programs to work with human-language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources, along with a suite of text processing for tokenization, stemming, tagging, parsing, among other such tasks. More details are available at: https://www.nltk.org/.

embedding and POS information, into the semantic feature extraction layer, which was constructed via Bi-LSTM or the BERT model. We can simplify the process of this layer by adopting the following Formulation (4) or Formulation (5):

$$H = \mathrm{BERT(X)} \tag{4}$$
$$H = \mathrm{Bi-LSTM(X)} \tag{5}$$

where $H \in \mathbb{R}^{L\times 2d_h}$ in Formulation (4) and Formulation (5) denotes a matrix that comprises semantic features $[h_1, h_2, \ldots, h_L]$, which are considered the initial coarse-grained sentiment features of the input context, and we use them to perform the downstream processing. For Bi-LSTM, at time step t, its output h_t comprises two parts: forward hidden states $\vec{h}_t \in \mathbb{R}^{d_h}$ and backward hidden states $\overleftarrow{h}_t \in \mathbb{R}^{d_h}$, where d_h denotes the number of Bi-LSTM hidden units. For the BERT model, the stacked transformer layers are introduced to refine the contextual information layer-by-layer and form the original semantic features H.

3.4. Proposed Target-Context Gate

Considering that different targets in the same input sentence may always have various sentiment polarities, some mistakes are often introduced in the final decisions by blindly depending on the emotional polarity of the entire sentence. Accordingly, we must filter out some irrelevant features based on the current target information. Therefore, we used another gating mechanism that incorporated the current target information. Here, we first propose a new kind of gate unit, called GReU, to accurately select aspect-related high-level features from the outputs of the semantic feature extraction layer. The process can be formulated as follows:

$$\mathrm{A} = \mathrm{relu}(HW_a + b_a) \tag{6}$$
$$\mathrm{T} = \mathrm{sigmoid}\left(HW_t + e_t^T W_{t'} + b_t\right) \tag{7}$$
$$\mathrm{V} = \mathrm{A} \times \mathrm{T} \tag{8}$$

where $W_a \in \mathbb{R}^{2d_h\times 2d_h}$, $W_t \in \mathbb{R}^{2d_h\times 2d_h}$, and $W_{t'} \in \mathbb{R}^{d\times 2d_h}$ denote weight matrixes, and $b_a \in \mathbb{R}^{2d_h}$ and $b_t \in \mathbb{R}^{2d_h}$ denote biases. The term A denotes the essential semantic features contained in the hidden matrix, and T denotes the interdependence between the current target embedding and hidden vectors measured using the sigmoid activation function. Therefore, $V \in \mathbb{R}^{L\times 2d_h}$ in Formulation (8) is the result of semantic features filtering by using the target information.

3.5. Proposed POS-highlighting Attention Mechanism

In addition to the target information, each word in a sentence should be treated differently before generating the final sentence representation. This is because each kind of POS category distinctly contributes to the expression of the target sentiment polarity. We regard the POS category as the basis of calculating attention weights to assist the model in focusing on a word according to its relevance to the emotional expression. We apply the attention scores to the hidden features $V \in \mathbb{R}^{L\times 2d_h}$, which are the outputs of the target-context gate and own high relevance to the current target. The complete calculation algorithm of weighting the hidden vectors according to their different importance of the four POS categories toward expressing the sentiment is as follows:

$$q_{w_i} = \begin{cases} (e_p^o)^T W_q^o v_q^o, w_i \in \{P_{\text{others}}\} \\ (e_p^o)^T W_q^o v_q^o + \left|(e_p^{w_i})^T W_q v_q\right|, w_i \in \{P_{adj}, P_{adv}, P_{\text{verb}}\} \end{cases} \tag{9}$$

$$\alpha = \mathrm{softmax}(q) \tag{10}$$
$$\mathrm{r} = \alpha^T \mathrm{V} \tag{11}$$

where $W_q^o \in \mathbb{R}^{d_p\times d}$ and $W_q \in \mathbb{R}^{d_p\times d}$ denote weight matrixes, and $v_q^o \in \mathbb{R}^{d\times 1}$ and $v_q \in \mathbb{R}^{d\times 1}$ denote weight vectors. Additionally, $q = [q_{w_1}, \cdots, q_{w_L}] \in \mathbb{R}^{L\times 1}$, and $\alpha \in \mathbb{R}^{L\times 1}$ denotes a vector that comprises attention weights; $r \in \mathbb{R}^{2d_h}$ denotes the weighted representation of a sentence with the given target. According to our analysis, compared with other kinds of POS, adjectives, verbs, and adverbs are often more crucial in expressing emotions. Accordingly, we divided the calculation of q_{w_i} into two cases: if the POS of the current word belonged to the "others" category, we adopted the first method in Formulation (9). Here, $e_p^o \in \mathbb{R}^{d_p\times 1}$ denotes the POS embedding that corresponds to the "others" category; if the POS of the current word belonged to the adjective, adverb, or verb category, we used the second method in Formulation (13) to obtain the score q_{w_i}. Accordingly, the final results obtained by implementing Formulations (9) and (10) were entirely consistent with our expected goal: the attention score of a word that belongs to the "others" category is significantly lower than that of a word that belongs to the adjective, adverb, or verb category. For these three POS categories, the relative values of their attention scores depend on their own POS embedding, which is updated throughout the learning of the model. The POS-highlighting attention mechanism limits the generation of attention scores for different words, and the weights between the "others" POS category and another three POS categories always satisfy a fixed relative relationship. The advantage of this attention mechanism is that it assists the model in significantly focusing on the words (adjectives, adverbs, and verbs) that contain some emotional tendencies rather than being influenced by other words. In this attention mechanism, the POS information is the only factor we considered, which determines that the attention scores of the words that have the same POS attribute in a sentence are also the same. The attention weights vector α is applied to the sentiment feature V filtered through the target-context gate; therefore, the POS information does not entirely influence the weighted sentence representation. We will further explore the effectiveness of both the POS-highlighting attention mechanism and limiting condition in Section 4.5.

The final sentence representation is given as:

$$h^* = \left[r, h_{\text{target}}\right] \tag{12}$$

where $h^* \in \mathbb{R}^{4d_h}$ denotes the feature representation of a sentence that enters both target and POS information into consideration, and $h_{target} \in \mathbb{R}^{2d_h}$ denotes the hidden vector that corresponds to the target in the hidden matrix H. Finally, a softmax layer is employed to transform this real valued vector h^* into conditional probability distribution.

$$\mathrm{y} = \mathrm{softmax}(h^* W_s + b_s) \tag{13}$$

where $W_s \in \mathbb{R}^{4d_h\times|C|}$ and $b_s \in \mathbb{R}^{|C|}$ denote the parameters of the softmax layer and $|C|$ the final class number (equal to three in our model).

3.6. Model training

The model can be trained in an end-to-end manner in a supervised learning framework via backpropagation, and the objective function (loss function) is the cross-entropy loss. Let y be the correct target distribution for a sentence and $\hat{y}$ the predicted sentiment distribution. The model is trained by minimizing the cross-entropy error between the ground truth y and prediction $\hat{y}$ for all the data samples.

Table 1
Statistics of the three datasets.

	Restaurant		Laptop		Twitter	
	Train	Test	Train	Test	Train	Test
Positive	2,164	728	994	341	1,561	173
Neural	637	196	464	169	3,127	346
Negative	807	196	870	128	1,560	173
Total	3,608	1,120	2,328	638	6,248	692

$$loss = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2 \quad (14)$$

where i denotes the index of sentence and j the index of class. Additionally, λ denotes the L2-regularization factor and θ the parameter set that contains all the parameters.

4. Experiments

In this section, we apply both IPAN-LSTM and IPAN-BERT to three aspect-level sentiment-classification datasets and compare their results with those of some representative baselines. To further demonstrate the positive impact of the POS information for sentiment analysis and the effectiveness of each module designed for IPAN, we conducted some additional investigations, whose detailed processes and analyses are provided in Sections 4.3 to Section 4.7.[2]

4.1. Experiment setting

4.1.1. Parameter setting

IPAN-LSTM: In our experiments, all the word-embedding and target-embedding vectors were set as Glove vectors[3] [113], which were pre-trained on an unlabeled corpus of size approximately 840 billion, and out-of-vocabulary words were provided via sampling from a uniform distribution $U(-0.1, 0.1)$. All the weight matrixes and POS-embedding vectors obtained their initial values via sampling from another uniform distribution $U(-0.1, 0.1)$, and they could be updated during training. The dimensions of word embedding, target embedding, POS embedding, and the number of the hidden states of Bi-LSTM were 300. The length of the attention weights vector was the same as that of the input word-embedding vector. We used TensorFlow [114] to implement our proposed model and employed Adam [115] as the training method. We trained all the models using 60 epochs, batch size of 25 examples, the L2-regularization weight of 0.001, and the initial learning rate of 0.01.

IPAN-BERT: We adopted $\text{BERT}_{\text{BASE}}$ (uncased) as the basis for all the experiments of IPAN-BERT. All the weight matrixes and POS-embedding vectors were initialized and updated similar to IPAN-LSTM. The dimensions of word embedding, target embedding, POS embedding, and the number of the hidden states of BERT were 768. During the training, the coefficient λ of the L2-regularization item was 10^{-5} and the dropout rate was 0.1. The Adam optimizer with the learning rate of $2e-5$ was applied to update all the parameters. The maximum number of epochs was set to 10.

4.1.2. Dataset

We conducted experiments on the three most widely used datasets for the aspect-level sentiment-classification task. The first two datasets were from SemEval 2014 Task 4[4] [116], which contains some customers reviews of laptop and restaurant. The last dataset comprised many tweets collected by Dong et al. [117]. Each sample in all the datasets included a list of targets and their corresponding polarities, which were labeled as *positive, negative, or neural*. Additionally, the final goal of our model was to identify the aspect polarity of a sentence with the corresponding target in the most precise manner. The statistics of the three datasets are presented in Table 1.

4.2. Model comparison

To evaluate the performance of our proposed model completely, we compare it with the following baseline models.

- **LSTM** only uses a single LSTM network to model the input sentence and presents the average results of all output hidden states as the final representation of the context. Subsequently, it feeds it into a softmax layer to predict the sentiment polarity [4].
- **TD-LSTM** utilizes two LSTMs to model both the sides of the current target combined with the target information, respectively, and concatenates their outputs for making the prediction [3].
- **ATAE-LSTM** is developed based on AE-LSTM [4]. It further strengthens the effects of the target embedding and appends them with the original word embedding. Additionally, it designs an attention mechanism that completely uses the given target information [4].
- **IAN** introduces a method that separately models the target and sentences, and it proposes an interactive attention mechanism to learn the attention weights that reflect the relatedness between the target and input sentence. Finally, it concatenates the target and sentence representations for the prediction [5].
- **BiLSTM-ATT-G** models both left and right contexts by using two attention-based LSTMs and introduces three gate units to measure the importance of both the parts of the sentence and itself for the prediction [111].
- **RAM** is a multi-layer recurrent network that uses an attention-based Bi-LSTM to learn the sentence representation and position information to enhance the model performance [118].
- **GCAE** is built using convolutional layers and gate units, and each convolutional filter computes n-gram features using the embedding vectors and gate units with pooling layer to select aspect-related sentiment features [112].
- **TNet-AS** adapts a convolutional neural network (CNN) to perform target-level sentiment classification and also employs a Bi-LSTM to accumulate the context information for each word of the input sentence [119].
- **MGAN** builds a multi-grained attention layer behind a Bi-LSTM layer to capture the comprehensive sentimental information of the target [7].

[2] In these five subsections, we perform investigations based on only IPAN-LSTM. Accordingly, the IPAN mentioned in both the tables and the analysis parts refers to IPAN-LSTM.

[3] The pre-trained word vectors of Glove can be obtained from http://nlp.stanford.edu/projects/glove/.

[4] The detailed information of this dataset can be found at: http://alt.qcri.org/semeval2014/task4/.

- **BERT-base** directly uses pretrained $\text{BERT}_{\text{BASE}}$ embeddings on the down-stream task without performing any domain-specific language model finetuning [88].
- **BERT-PT** performs multi-task finetuning prior to downstream classification, where the BERT language model is jointly fine-tuned with a question-answering task [104].

Table 2 presents the performances of our model (includes IPAN-LSTM, IPAN-BERT, and IPAN-BERT-PT) and other baseline models on the three datasets, respectively. IPAN obtains the best results among all the methods, verifying the efficacy of our model. LSTM, which serves as the above-mentioned baseline model, obtains the worst classification accuracy because of its simple structure. Furthermore, our semantic feature extraction layer, which includes Bi-LSTM or BERT, can extract more abundant and more high-level semantic features compared with the original LSTM. Therefore, the Bi-LSTM-based model, like BILSTM-ATT-G, achieves an improvement of 6.3% and 5.2% points in terms of accuracy on Laptop and Restaurant, respectively. However, models that consider the importance of the target and employ attention mechanisms to assign attention weights to differentiate each word in a given sentence (such as ATAE-LSTM, IAN, and RAM), also improve remarkably in terms of the ultimate classification accuracy. Additionally, MGAN integrates both the coarse- and fine-grained attention mechanisms to make the sentiment information more sufficient compared with previous works, and the model effectiveness increases because of this adjustment. For the models with gating mechanisms, such as GCAE, their final performances are ideal, proving that gate units can moderately magnify useful sentiment features.

TNet-AS and MGAN, the two models with their final performances second only to that of our model, as presented in Table 2, still have some disadvantages compared with our IPAN. This is because TNet-AS employs a CNN as the feature extractor for this sentiment-classification task, and a CNN is less popular than LSTM for NLP tasks and still encounters some obstacles in extracting some long-term dependency features. Although MGAN designs a unique and sophisticated multi-grained attention mechanism, it does sufficiently use the target information to ensure that inappropriate words are assigned lower attention scores. Our POS-highlighting attention mechanism adds a limiting condition (attention score of the word in the "others" categories must be lower than those of adjective/adverb/verb categories) to guarantee the rationality of attention-weight distribution. However, the other two models ignore this potential help from the grammatical features of the input sentence during the sentiment-classification task, although they are more robust to the training noise.

For the comparisons among the BERT-based networks in Table 2, the final performances of the four BERT-based models, namely, BERT-base, BERT-PT, IPAN-BERT, and IPAN-BERT-PT, are significantly similar. Notably, IPAN-BERT-PT performs the best although with a low margin. Even if the IPAN architecture is combined with BERT-base model, whose generalization ability is slightly worse than that of BERT-PT, the resulting IPAN-BERT also achieves better accuracy and F1-score values than those of the other contrast models.

Finally, by applying the attention mechanism before generating the final sentence representation and artificially adding a limiting condition to maximize the useful prior knowledge contained in the POS information, IPAN proves to be a POS-aware neural network for effectively analyzing the sentiment polarity. Using the target-context gate mechanism, IPAN amplifies the current target information, making the entire model more appropriate for the aspect-level classification task, and IPAN-BERT-PT dominates the state-of-the-art performances in the above-mentioned three datasets.

4.3. Effect of POS information

To explore the benefits of a single POS category (POS category number $L = 2$) and different combinations of POS categories for the final model prediction, we design and train the other eight IPAN variants using different combinations of POS categories on the three datasets. The details and final performances of these variants are presented in Table 3. The comparative line graphs of the accuracies are depicted in Fig. 3.

Although there are similar investigations of the impact of POS information on sentiment analysis in [18], [18] emphasis more on various POS-category combinations and ignored the single POS category. Their results were also almost identical to those observed in our investigation. However, they surprisingly observed that considering the other POS categories (except for adjective, adverb, and verb categories) would improve the performance of the final sentiment classification. However, their experiment only regarded nouns as the representative of other POS categories and ignored the value of the remaining categories, which had some shortcomings. To compensate for this deficiency and maximize the value of all the POS categories, in all the experiments of this subsection the remaining categories are uniformly recorded as "others" category in addition to considering the POS categories that are currently most considered. Our experimental results confirm that the "others" category significantly improves the prediction accuracy of the classification model.

Table 2
Comparison with baselines. Accuracy and Macro-F1 as evaluation metrics on three-class on Laptop, Restaurant, and Twitter datasets. Here, "*" indicates the result of the original paper that proposed the corresponding model, and the best records are in bold. IPAN-BERT-PT is initialized with post-trained BERT [104] weights.

Models		Laptop		Restaurant		Twitter	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Baselines	LSTM	0.665*	–	0.743*	–	–	–
	TD-LSTM	0.721	0.669	0.788	0.691	0.673	0.643
	ATAE-LSTM	0.690	0.626	0.768	0.640	0.676	0.650
	IAN	0.721*	–	0.786*	–	–	–
	BiLSTM-ATT-G	0.728	0.691	0.795	0.698	0.710	0.684
	RAM	0.727	0.702	0.794	0.692	0.700	0.677
	GCAE	0.733	0.701	0.793	0.705	0.718	0.696
	TNet-AS	0.740	0.698	0.797	0.708	0.720	0.696
	MGAN	0.750	0.715	0.805	0.715	0.720	0.701
	BERT-base	0.752	0.717	0.815	0.716	0.727	0.711
	BERT-PT	0.781*	0.751*	0.850*	0.770*	–	–
Our Models	IPAN-LSTM	0.772	0.735	0.828	0.738	0.743	0.725
	IPAN-BERT	0.785	0.760	0.859	0.764	0.767	0.759
	IPAN-BERT-PT	**0.793**	**0.767**	**0.865**	**0.777**	**0.778**	**0.768**

Table 3
Details and results comparison of IPAN and its variants (accuracy as an evaluation metric). For simplicity, adjectives, adverbs, verbs, nouns, and pronouns are represented by their abbreviations in this table.

IPAN Variants	Number of POS (L)	Combination of POS	Laptop	Restaurant	Twitter
IPAN-I	2	(adj, others)	0.753	0.815	0.737
IPAN-II	2	(adv, others)	0.748	0.812	0.735
IPAN-III	2	(v, others)	0.747	0.812	0.732
IPAN-IV	3	(adj, adv, others)	0.765	0.815	0.740
IPAN-V	3	(adj, v, others)	0.755	0.817	0.739
IPAN-VI	3	(adv, v, others)	0.754	0.816	0.734
IPAN	**4**	**(adj, adv, v, others)**	**0.772**	**0.828**	**0.743**
IPAN-VII	5	(adj, adv, v, others, n)	0.745	0.818	0.735
IPAN-VIII	5	(adj, adv, v, others, pron)	0.758	0.817	0.735

The best records are in bold.

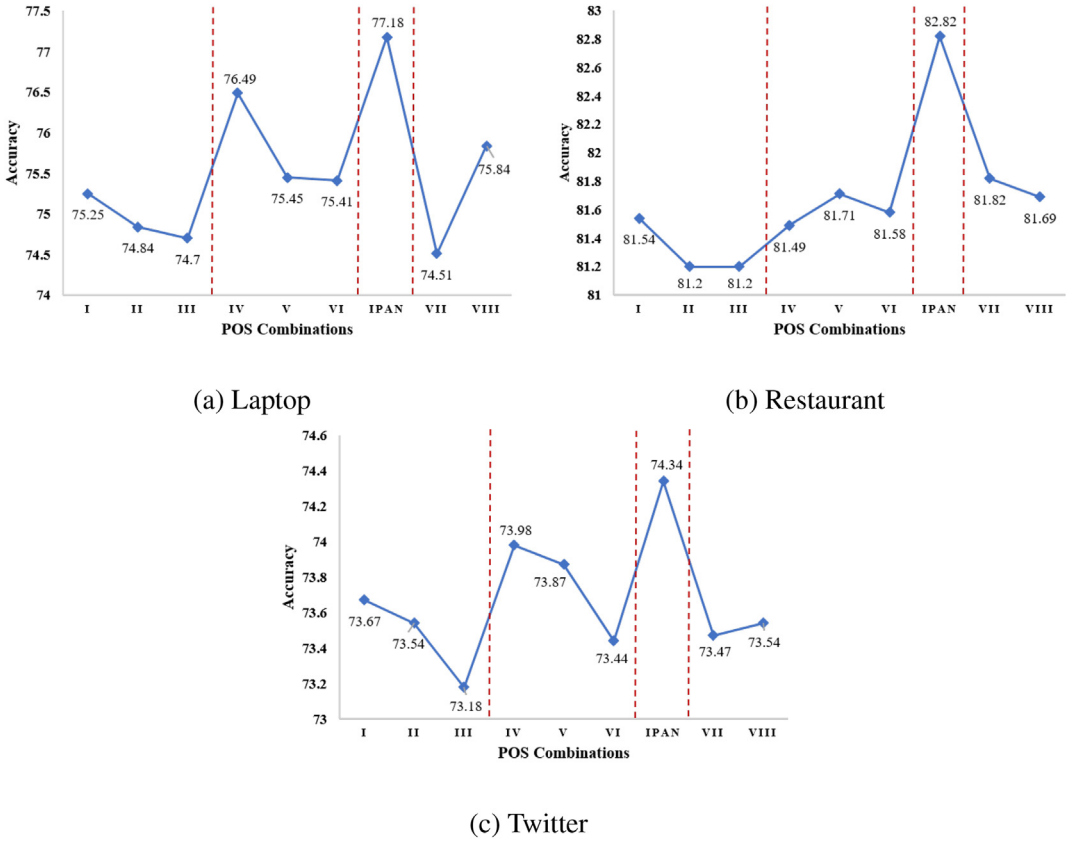

Fig. 3. (a), (b), and (c) denote the line graphs showing the prediction accuracies of IPAN and its eight variants on Laptop, Restaurant, and Twitter, respectively. Here, I–VIII on the horizontal axes correspond to IPAN-I through IPAN-VIII, respectively, which comprise different combinations of POS categories. The red dotted lines separate the groups with different POS category numbers L.

According to the existing research conclusions associated with the effect of POS information on sentiment analysis in Section 2.2, it is reasonable to assume that adjectives, adverbs, and verbs vitally affect the emotional expression of contexts. Accordingly, we focused on these three POS categories to investigate the effect of the POS information. Additionally, as a supplement to [18], we explore the influence of other POS categories, represented by nouns and pronouns, on sentiment analysis. From Fig. 3, we can conclude that when $L = 2$ (the situation that corresponds to a single POS category), considering only the "adjective" category can achieve better model performance compared with considering either the "adverb" category or "verb" category. However, the trend in Fig. 3 (a), (c), and (b) is slightly different when $L = 3$: in (a) and (c). The peak value is observed at "IV" (corresponds to the combination of "adjective", "adverb", and "others" category) point. However, in (b), the combination of the "adjective" and "verb" categories is the best among the other choices. This may attributed to the samples in the three datasets having different semantic and grammatical characteristics. Regarding the addition of a new POS category ("noun" or "pronoun" category) based on our original four categories ($L = 5$), the results for the three datasets indicate that the model effectiveness would degraded upon this addition.

From the above-mentioned experimental phenomena, we can conclude that not all POS categories assist the model in analyzing the sentiment polarity. However, redundant information may counterproductively drop the final effectiveness of the model. Considering the single POS category, adjectives, adverbs, and verbs most significantly indicate sentiment polarity: adjectives, with the most positive influence on the model effectiveness, directly

guide the model to identify which words or spans are more meaningful for emotional expressions, while verbs have less emotional characteristics, and adverbs might modify the degree or express negation. The optimal POS strengths observed in the experiments of [17] can also be used to understand the relevance of each POS category toward sentiment analysis. Their results showed adjectives and adverbs to be highly relevant to the sentiment of the context, verbs as fairly relevant, and nouns as moderately relevant, completely consistent with our conclusions. Additionally, what matches the research results for a single POS category is that for POS combinations, when the above-mentioned three POS categories are separately considered and the remaining POS categories uniformly classified into "others" category, the POS information provides the largest assistance toward performing the accurate classification of sentiment polarity.

4.4. Investigation on two gating mechanisms

In this subsection, we compare GLU [108], GTU [120], GTRU [112], and GReU used in POS-filter and target-context gates, respectively. In this comparison experiment, we adopt the control-variable method; i.e., when we test the different performances of the following four gating mechanisms in POS-filter (target-context) gate, we need to keep POS-filter (target-context) gate optimal choice (GReU/GLU). Table 4 shows that all the four gate units applied in the POS-filter and target-context gates achieve significantly high accuracies on the Laptop dataset.

For the POS-filter gate, GLU outperforms the other gate units with a high margin, and GTU secures the second place. The experiment results strongly verify our hypothesis that the sigmoid function can control the flow of features associated with the POS information of each word. Although GLU is a simplified gating mechanism based on GTU, it can reduce the vanishing gradient problem coupling linear units to the gates, thereby retaining the non-linear capabilities of the layer while enabling the gradient to propagate through the linear unit without scaling [108]. GReU achieves the best model performance in the target-context gate, followed by GTRU, thereby almost consistent with the results in Xue and Li et al. [112]. Because different targets with different sentiment polarities can always appear in a sentence, ReLu function, which strictly limits the output of negative input samples to zero, becomes crucial in this condition. Consequently, both GReU and GTRU can output a similarity score according to the relevance between the given target information and semantic features, thereby preventing the interference from the features unrelated to the current target via the ReLu function.

Table 4
Accuracy and Macro-F1 of the different gate units used in POS-filter and target-context gates, respectively, on Laptop reviews.

Gate Units	POS-filter Gate		Target-Context Gate	
	Accuracy	Macro-F1	Accuracy	Macro-F1
GReU	0.758	0.711	**0.772**	**0.725**
GTU	0.761	0.719	0.750	0.702
GTRU	0.748	0.705	0.761	0.715
GLU	**0.772**	**0.725**	0.761	0.714

The best records are in bold.

4.5. Investigation on POS-highlighting attention mechanism

In this subsection, we design another three alternative attention mechanisms using the POS information to compare them with the POS-highlighting attention mechanism. We are the first to introduce an attention mechanism that integrates the POS information, which is completely different from the existing design intension (target-information/position-information). For fairness, instead of adopting any existing attention mechanism, we design three reasonable methods that calculate POS-attention scores as our comparison approaches. We test the performances of the following three kinds of attention mechanisms while keeping the other components of IPAN unchanged. The experiment results are listed in Table 5 (the three models are denoted as IPAN-POS Attention1#, IPAN-POS Attention2#, and IPAN-POS Attention3#, respectively).

POS Attention1#. The design of POS Attention1# aims to reduce the number of additional parameters added with the POS-highlighting attention mechanism to the maximum possible extent; accordingly, we adjust the original calculation method of q_{w_i} to the following process:

$$q'_{w_i} = \begin{cases} (e_p^o)^T v_q^o, w_i \in \{P_{\text{others}}\} \\ (e_p^o)^T v_q^o + \left|(e_p^{w_i})^T v_q\right|, w_i \in \{P_{\text{adj}}, P_{\text{adv}}, P_{\text{verb}}\} \end{cases} \tag{15}$$

where $v_q^o \in \mathbb{R}^{d_p \times 1}$ and $v_q \in \mathbb{R}^{d_p \times 1}$ denote weight vectors. Compared with Formulation (9), this calculation method omits weight matrixes W_q^o and W_q, and although it fits the purpose, it results in poor model performance (Table 5 shows that the average drop on the three datasets is 1.8%), indicating that sufficient number of parameters can moderately improve the model expressiveness.

POS Attention2#. In our POS-highlighting attention mechanism, we artificially add a limiting condition to guarantee that the attention scores of adjectives, adverbs, and verbs are always higher than the words of other kinds of POS category. To verify the rationality and validity of this strong limitation, we design another calculation method to obtain the attention scores without any restrictions.

$$q''_{w_i} = \begin{cases} (e_p^o)^T W_q^o v_q^o, w_i \in \{P_{\text{others}}\} \\ (e_p^{w_i})^T W_q v_q, w_i \in \{P_{adj}, P_{adv}, P_{verb}\} \end{cases} \tag{16}$$

where $W_q^o \in \mathbb{R}^{d_p \times d}$ and $W_q \in \mathbb{R}^{d_p \times d}$ denote weight matrixes, and $v_q^o \in \mathbb{R}^{d \times 1}$ and $v_q \in \mathbb{R}^{d \times 1}$ denote weight vectors. Compared with Formulation (9), although Formulation (16) still contains two different ways to compute q''_{w_i} according to the POS category of the current word, the restriction in Formulation (9) is discarded while calculating the attention scores of adjectives, adverbs, and verbs, respec-

Table 5
Comparison with three other attention mechanisms combined with the POS information. Accuracy and Macro-F1 are used as evaluation metrics on three-class about Laptop, Restaurant, and Twitter datasets.

Models	Laptop		Restaurant		Twitter	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
IPAN-POS Attention1#	0.745	0.697	0.814	0.723	0.729	0.708
IPAN-POS Attention2#	0.748	0.705	0.812	0.724	0.729	0.716
IPAN-POS Attention3#	0.741	0.694	0.813	0.718	0.729	0.713
IPAN-POS-highlighting Attention	**0.772**	**0.735**	**0.828**	**0.738**	**0.743**	**0.725**

The best records are in bold.

tively. The results show that the model that used POS Attention2# decreases the final classification accuracies by 2.4%, 1.6%, and 1.4% on the three datasets, respectively. This reflects that the model cannot conclude this limiting condition only via a learning process performed using a limited number of training samples. Therefore, the limiting conditions we added in the POS-highlighting attention mechanism are effective and significantly practical.

POS Attention3#. Some previous works [121] found that even if all of the words were the same POS category, their attention weights should be different. Hence, to explore whether we can enhance the effect of the attention mechanism by considering both the semantic features of the word and the POS attribute thereof, we design POS Attention3#, which combines word and POS embeddings. Additionally, the calculation method of q_{w_i} is adjusted to the following:

$$q'''_{w_i} = q_{w_i} + e^T_{w_i} W'''_q v'''_q \tag{17}$$

where $W'''_q \in \mathbb{R}^{d_p \times d}$ denotes a weight matrix and $v'''_q \in \mathbb{R}^{d \times 1}$ a weight vector. The result q'''_{w_i} in Formulation (17) is to add the biases calculated on the basis of the word embedding of each word to q_{w_i} in Formulation (9); however, the results in Table 5 indicate that introducing the original semantic information of a word into the calculation of attention score in this manner is less effective, which may be attributed to the interference due to this combination. Because there are many word-embedding varieties (equal to the vocabulary size v), Formulation (17) will have multiple possible outcomes, which are determined by both POS and word embeddings of a word; however, the model does not have sufficient learning ability to weigh the relationships between these two kinds of knowledge, and thus the final prediction deteriorates. Therefore, incorporating extra information into the attention mechanism via Formulation (17) introduces significant interference, distracting the model from its original goal.

4.6. Ablation study

To investigate the effect of three components, namely, POS-filter gate, target-context gate, and POS-highlighting attention, on our model effectiveness, we compared the full model (IPAN) with its three ablations. The results are presented in Table 6; the three ablated models are represented as IPAN w/o POS-filter Gate, IPAN w/o Target-Context Gate, IPAN w/o POS-highlighting attention, respectively.

From the results of these three ablated models and IPAN on the three datasets, we observe that removing any one of the three components deteriorates the final performance of the model. This phenomenon indicates that the integration of target and POS information is critical to achieving satisfactory model performance. Comparing the results of IPAN w/o POS-filter gate, IPAN w/o target-context gate, and IPAN w/o POS-highlighting attention, we observe that the final prediction accuracy and Macro-F1 of the model w/o POS-highlighting attention drop the most on the three datasets (accuracy drops are 4%, 3.5%, and 3.1% on the three datasets, respectively). This reconfirms the validity of the POS-highlighting attention mechanism and indicates that assigning a different weight to each POS category help the model focus on important words. The accuracy of the model w/o POS-filter gate drops by 2.6%, 1.8%, and 1.9% on the three datasets, respectively, suggesting that transforming the original word embedding into tailor-made vectors with POS information is significantly useful. All the data and analysis assert that using the POS information as prior knowledge to assist the modeling process significantly facilitates the learning of the neural network, thereby assisting with the sentiment-classification task. Additionally, the target-context gate, as a significant connection between the target and semantic information extracted from the original sentence via the semantic feature extraction layer, is critical to enhancing the model performance. The results of the model w/o target-context gate decrease by 3.1% points, 1.3% points, and 1.6% points on the three datasets, respectively, proving that the target-context gate prevents target-unrelated features from flowing into the next layer to affect the final prediction.

4.7. A case study

To understand our proposed POS-highlighting attention mechanism intuitively, we visualize the attention weight on two sentences, as depicted in Fig. 4. The color depth indicates the attention-score level: the darker is the color, the higher is the level. The samples in Fig. 4 are randomly selected from the test sets of Restaurant and Laptop. In Fig. 4, for the first sentence "*Works well, and I am extremely happy to back to an apple OS*," the polarity is *positive* for *apple OS*. Our model is inclined toward considering adjectives, verbs, and adverbs in the given sentence. Accordingly, it finally assigns the highest attention score to the adjective "*happy*," which is critical to judging the sentiment polarity of *apple OS*. However, the words that belong to the "others" category, such as "*and*," "*I*," and "*an*," obtain less attention, thereby demonstrating again that the limiting condition we designed for ensuring that

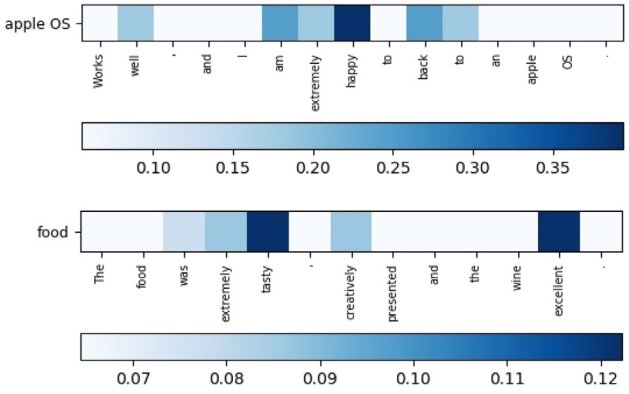

Fig. 4. Attention visualizations. The two samples are randomly selected from Laptop and Restaurant datasets, and their targets are "*apple OS*" and "*food*," respectively. The above two bars represent the final attention scores of each word in these two sentences calculated via the POS-highlighting attention mechanism.

Table 6
Comparison with the ablated models. Accuracy and Macro-F1 are used as evaluation metrics on three-class about Laptop, Restaurant, and Twitter datasets.

Models	Laptop		Restaurant		Twitter	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
IPAN w/o POS-filter Gate	0.746	0.703	0.810	0.722	0.724	0.706
IPAN w/o Target-Context Gate	0.741	0.688	0.815	0.731	0.727	0.707
IPAN w/o POS-highlighting Attention	0.732	0.685	0.793	0.683	0.712	0.691
IPAN	**0.772**	**0.735**	**0.828**	**0.738**	**0.743**	**0.725**

The best records are in bold.

the weight of other POS categories is strictly lower than that of the adjective, adverb, and verb categories satisfactorily works. Considering the second sentence (“*The food was extremely tasty, creatively presented and the wine excellent*”) in Fig. 4, distinctively, the word “*tasty*” is the most important word to express the sentiment of the given target *food*. From the second example, it is noteworthy that some words such as “*extremely*” and “*creatively*,” which belong to the adverb category, receive less attention compared with words such as “*tasty*” and “*excellent*,” which belong to the adjective category. This phenomenon indicates that the model has adaptively learned the different importance degrees of the four POS categories according to their influences toward expressing the sentiment polarity of the current target, during the training process. In the above-mentioned two sentences, the “adverb” and “verb” categories are less useful than the adjective category for the final model prediction and, hence, receive the lower attention scores.

Notably, in the second sentence, the attention scores of “*tasty*” and “*excellent*,” which do not target the same target *food*, are almost the same. This is consistent with the computational rules of the POS-highlighting attention mechanism, in which we did not consider any semantic factors but only the effects of different POS categories on emotional expressions. If these two words express opposite sentiment polarities, the model makes an incorrect decision. However, by incorporating the efficient target-context gate, this problem can be solved, as the final attention scores are applied to its filtered feature vectors. Additionally, Formulation (12) indicates that the final sentence representation comprises the concatenation results of both the weighted sentence vector and the hidden states from the semantic feature extraction layer corresponds to the target embedding, guaranteeing that the target information is sufficiently emphasized before the model makes the final decision. Together with these two components (the target-context gate and POS-highlighting attention), IPAN will eventually considerably focus on the words that are most relevant to the given target, and can therefore reflect their emotional orientations according to their POS attributes.

5. Conclusion

We proposed an IPAN for performing aspect-level sentiment classification. Instead of blindly relying on the learning knowledge acquired from a few existing datasets, IPAN introduced the POS information as prior knowledge to explicitly provide the model with more evident indications that would facilitate its final prediction of sentiment polarity. Based on our observations that adjectives, adverbs, and verbs often contain useful information related to emotional expressions, we only modeled the information of these three POS categories to avoid the noise due to knowledge redundancy. Notably, we not only used this POS information to construct a POS-filter gate to generate transformed word embeddings while considering the POS information but also designed a POS-highlighting attention mechanism with a limiting condition to ensure that these three POS categories were assigned high attention scores. To adapt to the aspect-level sentiment classification, we formulated a target-context gating mechanism to emphasize the target information and implement the interaction between the target and each context word, where the grating mechanism comprises a novel form of gated units called GReU. Our IPAN consistently outperformed the previous state-of-the-art methods on SemEval2014 and Twitter datasets.

CRediT authorship contribution statement

Kai Shuang: Conceptualization, Methodology, Software, Writing - original draft. **Mengyu Gu:** Investigation, Software, Methodology, Writing - original draft. **Rui Li:** Methodology, Validation. **Jonathan Loo:** Validation, Writing - review & editing. **Sen Su:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for the constructive comments. This work was supported in part by the National Key Research and Development Program of China (No. 2016QY01W0200). The work was also supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 61921003).

References

[1] B. Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies 5 (1) (2012) 1–167.
[2] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval 2 (1–2) (2008) 1–135.
[3] D. Tang, B. Qin, X. Feng, T. Liu, Effective lstms for target-dependent sentiment classification, arXiv preprint arXiv:1512.01100.
[4] Y. Wang, M. Huang, L. Zhao, et al., Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.
[5] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, arXiv preprint arXiv:1709.00893.
[6] S. Gu, L. Zhang, Y. Hou, Y. Song, A position-aware bidirectional attention network for aspect-level sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 774–784.
[7] F. Fan, Y. Feng, D. Zhao, Multi-grained attention network for aspect-level sentiment classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3433–3442.
[8] P.R. Kroeger, Analyzing Grammar: An Introduction, Cambridge University Press, 2005.
[9] J. Wiebe et al., Learning subjective adjectives from corpora, Aaai/iaai 20 (2000).
[10] V. Hatzivassiloglou, Effects of adjective orientation and gradability on sentence subjectivity, Proceedings of Coling 30 (3) (2000) 299–305.
[11] V. Hatzivassiloglou, K.R. McKeown, Predicting the semantic orientation of adjectives, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1997, pp. 174–181.
[12] M. Hu, B. Liu, Mining opinion features in customer reviews, in: AAAI, vol. 4, 2004, pp. 755–760.
[13] P. Chesley, B. Vincent, L. Xu, R.K. Srihari, Using verbs and adjectives to automatically classify blog sentiment, Training 580 (263) (2006) 233.
[14] J. Yi, T. Nasukawa, R. Bunescu, W. Niblack, Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, in: Third IEEE International Conference on Data Mining, IEEE, 2003, pp. 427–434.
[15] E. Riloff, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics, 2003, pp. 25–32.
[16] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, V.S. Subrahmanian, Sentiment analysis: adjectives and adverbs are better than adjectives alone, in: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007, short paper.
[17] C. Nicholls, F. Song, Improving sentiment analysis with part-of-speech weighting, in: 2009 International Conference on Machine Learning and Cybernetics, vol. 3, IEEE, 2009, pp. 1592–1597.
[18] W.-H. Khong, L.-K. Soon, H.-N. Goh, S.-C. Haw, Leveraging part-of-speech tagging for sentiment analysis in short texts and regular texts, in: Joint International Semantic Technology Conference, Springer, 2018, pp. 182–197.
[19] G. Fei, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, A dictionary-based approach to identifying aspects implied by adjectives for opinion mining, in: Proceedings of COLING 2012: Posters, 2012, pp. 309–318.
[20] S. Brody, N. Elhadad, An unsupervised aspect-sentiment model for online reviews, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 804–812.
[21] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on

Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.
[22] S. Pei, L. Wang, T. Shen, Z. Ning, Da-bert: Enhancing part-of-speech tagging of aspect sentiment analysis using bert, in: International Symposium on Advanced Parallel Processing Technologies, Springer, 2019, pp. 86–95.
[23] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intelligent Systems 32 (6) (2017) 74–80.
[24] M. Dragoni, S. Poria, E. Cambria, Ontosenticnet: a commonsense ontology for sentiment analysis, IEEE Intelligent Systems 33 (3) (2018) 77–85.
[25] J. Tao, T. Tan, Affective computing: a review, in: International Conference on Affective Computing and Intelligent Interaction, Springer, 2005, pp. 981–995.
[26] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: a deep learning approach, in: ICML, 2011.
[27] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.
[28] R.Y. Lau, Y. Xia, Y. Ye, A probabilistic generative model for mining cybercriminal networks from online social media, IEEE Computational Intelligence Magazine 9 (1) (2014) 31–43.
[29] E. Cambria, Affective computing and sentiment analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.
[30] E. Cambria, A. Hussain, Sentic computing, Marketing 59 (2) (2012) 557–577.
[31] Y. Xia, E. Cambria, A. Hussain, H. Zhao, Word polarity disambiguation using bayesian model and opinion-level features, Cognitive Computation 7 (3) (2015) 369–380.
[32] M. Dragoni, A.G. Tettamanzi, C. da Costa Pereira, A fuzzy system for concept-level sentiment analysis, in: Semantic Web Evaluation Challenge, Springer, 2014, pp. 21–27.
[33] M. Araújo, P. Gonçalves, M. Cha, F. Benevenuto, ifeel: a system that compares and combines sentiment analysis methods, in: Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 75–78.
[34] J.M. Chenlo, D.E. Losada, An empirical study of sentence features for subjectivity and polarity classification, Information Sciences 280 (2014) 275–288.
[35] J. K.-C. Chung, C.-E. Wu, R. T.-H. Tsai, Improve polarity detection of online reviews with bag-of-sentimental-concepts,in: Proceedings of the 11th ESWC. Semantic Web Evaluation Challenge, Crete. Springer, 2014, pp. 379–420.
[36] F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis, Knowledge-Based Systems 69 (2014) 86–99.
[37] G. Gezici, R. Dehkharghani, B. Yanikoglu, D. Tapucu, Y. Saygin, Su-sentilab: A classification system for sentiment analysis in twitter, in: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 471–477.
[38] D.R. Recupero, V. Presutti, S. Consoli, A. Gangemi, A.G. Nuzzolese, Sentilo: frame-based sentiment analysis, Cognitive Computation 7 (2) (2015) 211–225.
[39] A. Ortony, G.L. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge University Press, 1990.
[40] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation 39 (2–3) (2005) 165–210.
[41] C. Strapparava, A. Valitutti, et al., Wordnet affect: an affective extension of wordnet., in: Lrec, vol. 4, Citeseer, 2004, p. 40.
[42] A. Esuli, F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining., in: LREC, Vol. 6, Citeseer, 2006, pp. 417–422.
[43] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Twenty-eighth AAAI Conference on Artificial Intelligence, 2014.
[44] R.A. Stevenson, J.A. Mikels, T.W. James, Characterization of the affective norms for english words by discrete emotional categories, Behavior Research Methods 39 (4) (2007) 1020–1024.
[45] S. Somasundaran, J. Wiebe, J. Ruppenhofer, Discourse level opinion interpretation.
[46] D. Rao, D. Ravichandran, Semi-supervised polarity lexicon induction, in: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), 2009, pp. 675–682.
[47] A. Sapountzi, K.E. Psannis, Social networking data analysis tools & challenges, Future Generation Computer Systems 86 (2018) 893–913.
[48] L. Oneto, F. Bisio, E. Cambria, D. Anguita, Statistical learning theory and elm for big social data analysis, IEEE CompUTATionAl inTelliGenCe mAGAzine 11 (3) (2016) 45–55.
[49] E. Cambria, D. Rajagopal, D. Olsher, D. Das, Big social data analysis, Big Data Computing 13 (2013) 401–414.
[50] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Knowledge-based approaches to concept-level sentiment analysis, IEEE Intelligent Systems 28 (2) (2013) 12–14.
[51] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Statistical approaches to concept-level sentiment analysis, IEEE Intelligent Systems 28 (3) (2013) 6–9.
[52] A. Gangemi, V. Presutti, D.R. Recupero, Frame-based detection of opinion holders and topics: a model and a tool, IEEE Computational Intelligence Magazine 9 (1) (2014) 20–30.
[53] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, A. Hussain, Sentiment data flow analysis by means of dynamic linguistic patterns, IEEE Computational Intelligence Magazine 10 (4) (2015) 26–36.
[54] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 1555–1565.
[55] D. Tang, F. Wei, B. Qin, T. Liu, M. Zhou, Coooolll: A deep learning system for twitter sentiment classification, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 208–212.
[56] C. Dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69–78.
[57] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems (TOIS) 21 (4) (2003) 315–346.
[58] S.-C. Yang, M.-J. Liu, Ysc-dsaa: an approach to disambiguate sentiment ambiguous adjectives based on saaol, in: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 440–443.
[59] X. Ding, B. Liu, P.S. Yu, A holistic lexicon-based approach to opinion mining, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 231–240.
[60] L. Qiu, W. Zhang, C. Hu, K. Zhao, Selc: a self-supervised model for sentiment classification, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 929–936.
[61] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 347–354.
[62] B. Lu, B.K. Tsou, Cityu-dac: disambiguating sentiment-ambiguous adjectives within context, in: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 292–295.
[63] A. Pak, P. Paroubek, Twitter based system: using twitter for disambiguating sentiment ambiguous adjectives, in: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 436–439.
[64] Y. Wu, M. Wen, Disambiguating dynamic sentiment ambiguous adjectives, in: Proceedings of the 23rd International Conference on Computational Linguistics (coling 2010), 2010, pp. 1191–1199.
[65] A. Duque, M. Stevenson, J. Martinez-Romo, L. Araujo, Co-occurrence graphs for word sense disambiguation in the biomedical domain, Artificial Intelligence in Medicine 87 (2018) 9–19.
[66] M. Bevilacqua, R. Navigli, Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2854–2864.
[67] S. K. Bharti, K. S. Babu, S. K. Jena, Parsing-based sarcasm sentiment recognition in twitter data, in: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2015, pp. 1373–1380.
[68] D. Davidov, O. Tsur, A. Rappoport, Semi-supervised recognition of sarcasm in twitter and amazon, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, 2010, pp. 107–116.
[69] O. Tsur, D. Davidov, A. Rappoport, Icwsm—a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews, in: Fourth International AAAI Conference on Weblogs and Social Media, 2010.
[70] R. Kreuz, G. Caucci, Lexical influences on the perception of sarcasm, in: Proceedings of the Workshop on computational approaches to Figurative Language, 2007, pp. 1–4.
[71] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in twitter, Language Resources and Evaluation 47 (1) (2013) 239–268.
[72] D. Bamman, N.A. Smith, Contextualized sarcasm detection on twitter, in: Ninth International AAAI Conference on Web and Social Media, 2015.
[73] S. Mukherjee, P.K. Bala, Sarcasm detection in microblogs using naïve bayes and fuzzy clustering, Technology in Society 48 (2017) 19–27.
[74] S. Amir, B.C. Wallace, H. Lyu, P.C.M.J. Silva, Modelling context with user embeddings for sarcasm detection in social media, arXiv preprint arXiv:1607.00976.
[75] A. Ghosh, T. Veale, Fracking sarcasm using neural network, in: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2016, pp. 161–169.
[76] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intelligent Systems 34 (3) (2019) 38–43.
[77] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2005, pp. 486–497.
[78] E. Cambria, S. Poria, R. Bajpai, B. Schuller, Senticnet 4: a semantic resource for sentiment analysis based on conceptual primitives, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers, 2016, pp. 2666–2677.
[79] H. Liu, P. Singh, Conceptnet—a practical commonsense reasoning tool-kit, BT Technology Journal 22 (4) (2004) 211–226.
[80] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, Emosenticspace: a novel framework for affective common-sense reasoning, Knowledge-Based Systems 69 (2014) 108–123.
[81] F. Xu, J. Yu, R. Xia, Instance-based domain adaptation via multiclustering logistic approximation, IEEE Intelligent Systems 33 (1) (2018) 78–88.
[82] Y. Rao, Contextual sentiment topic model for adaptive social emotion classification, IEEE Intelligent Systems 31 (1) (2015) 41–47.

[83] Q. Yang, Y. Rao, H. Xie, J. Wang, F.L. Wang, W.H. Chan, E.C. Cambria, Segment-level joint topic-sentiment model for online review analysis, IEEE Intelligent Systems 34 (1) (2019) 43–50.
[84] H. Jin, M. Huang, X. Zhu, Sentiment analysis with multi-source product reviews, in: International Conference on Intelligent Computing, Springer, 2012, pp. 301–308.
[85] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Computational Linguistics 37 (1) (2011) 9–27.
[86] W. X. Zhao, J. Jiang, H. Yan, X. Li, Jointly modeling aspects and opinions with a maxent-lda hybrid, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 56–65.
[87] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[88] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
[89] T.U. Tran, H.T.-T. Hoang, H.X. Huynh, Bidirectional independently long short-term memory and conditional random field integrated model for aspect extraction in sentiment analysis, in: Frontiers in Intelligent Computing: Theory and Applications, Springer, 2020, pp. 131–140.
[90] A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko, A. Scharl, Aspect-based extraction and analysis of affective knowledge from social media streams, IEEE Intelligent Systems 32 (3) (2017) 80–88.
[91] M. Mitchell, J. Aguilar, T. Wilson, B. Van Durme, Open domain targeted sentiment, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1643–1654.
[92] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.
[93] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555.
[94] G.E. Hinton, A. Krizhevsky, S.D. Wang, Transforming auto-encoders, in: International Conference on Artificial Neural Networks, Springer, 2011, pp. 44–51.
[95] S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.
[96] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, arXiv preprint arXiv:1605.08900.
[97] T. Yang, Q. Yin, L. Yang, O. Wu, Aspect-based sentiment analysis with new target representation and dependency attention, IEEE Transactions on Affective Computing.
[98] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, X. Chen, Neural attentive network for cross-domain aspect-level sentiment classification, IEEE Transactions on Affective Computing.
[99] Y. Wang, A. Sun, M. Huang, X. Zhu, Aspect-level sentiment analysis using as-capsules, in: The World Wide Web Conference, 2019, pp. 2033–2044.
[100] Z. Chen, T. Qian, Transfer capsule network for aspect level sentiment classification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 547–556.
[101] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365.
[102] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.
[103] C. Sun, L. Huang, X. Qiu, Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, arXiv preprint arXiv:1903.09588.
[104] H. Xu, B. Liu, L. Shu, P. S. Yu, Bert post-training for review reading comprehension and aspect-based sentiment analysis, arXiv preprint arXiv:1904.02232.
[105] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, arXiv preprint arXiv:1901.11504.
[106] Y. Song, J. Wang, Z. Liang, Z. Liu, T. Jiang, Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference, arXiv preprint arXiv:2002.04815.
[107] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu, Exploring the limits of language modeling, arXiv preprint arXiv:1602.02410.
[108] Y. N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 933–941.
[109] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1243–1252.
[110] M. Zhang, Y. Zhang, D.-T. Vo, Gated neural networks for targeted sentiment analysis, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[111] J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 572–577.
[112] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, arXiv preprint arXiv:1805.07043.
[113] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[114] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.
[115] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[116] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: aspect based sentiment analysis, Proceedings of International Workshop on Semantic Evaluation at (2014) 27–35.
[117] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, Adaptive recursive neural network for target-dependent twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational linguistics L(volume 2: Short papers), vol. 2, 2014, pp. 49–54.
[118] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 452–461.
[119] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, arXiv preprint arXiv:1805.01086.
[120] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, K. Kavukcuoglu, Neural machine translation in linear time, arXiv preprint arXiv:1610.10099.
[121] Y. Zou, T. Gui, Q. Zhang, X. Huang, A lexicon-based supervised attention model for neural sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 868–877.

Kai Shuang received the master's and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT) in 2003 and 2006, respectively. He is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology, BUPT. His research interests include deep learning, natural language processing, image processing, cloud computing, and big data technology.

Mengyu Gu received the bachelor's degree in telecommunications engineering from the Shanghai University. She is currently a research assistant and a Master Degree Candidate in Beijing University of Posts and Telecommunications. Her research interests include deep learning, natural language processing, language modeling and text classification.

Rui Li received the bachelor's degree in information and computing science from the Dalian University of Technology. He is currently a research assistant and a Doctor Degree Candidate in Beijing University of Posts and Telecommunications. His research interests include deep learning, natural language processing, language modeling and text classification.

Jonathan Loo received his M.Sc. degree in Electronics (with Distinction) and the Ph.D. degree in Electronics and Communications from the University of Hertfordshire, Hertfordshire, U.K., in 1998 and 2003, respectively. Between 2003 and 2010, he was a Lecturer in Multimedia Communications with the School of Engineering and Design, Brunel University, Uxbridge, U.K. Between June 2010 and May 2017, he was an Associate Professor in Communication Networks at the School of Science and Technology, Middlesex University, London, U.K. From June 2017, he is a Chair Professor in Computing and Communication Engineering at the School of Computing and Engineering, University of West London, United Kingdom. His recent research interests include deep learning, natural language and image processing, cloud computing, wireless/mobile communication and networks, cyber security. He has successfully graduated 18 Ph.D. students as their principal supervisor, and has co-authored more than 250 journal and conference papers in the aforementioned specialized areas. Dr. Loo has been an Associate Editor for Wiley International Journal of Communication Systems since 2011. He was the Lead Editor of the book entitled Mobile Ad Hoc Networks: Current Status".

Sen Su is a professor in Beijing University of Posts and Telecommunications. His research interests include deep learning, natural language processing, computer vision, cloud computing and big data technology. Contact him at susen@bupt.edu.cn.