Knowledge-Based Systems 245 (2022) 108586

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning

Guixian Xu, Zixin Zhang, Ting Zhang *, Shaona Yu, Yueting Meng, Sijin Chen

College of Information Engineering, Minzu University of China, 100081, Beijing, China

ARTICLE INFO

Article history:
Received 24 July 2021
Received in revised form 10 March 2022
Accepted 10 March 2022
Available online 23 March 2022

Keywords:
Sentiment classification
BiLSTM
Attention
Transfer learning

ABSTRACT

Aspect-level sentiment classification, a fine-grained sentiment analysis task which provides entire and intensive results, has been a research focus in recent years. However, the performance of neural network models is largely limited by the small scale of datasets for aspect-level sentiment classification due to the challenges to label such data. In this paper, we propose an aspect-level sentiment classification model based on Attention-Bidirectional Long Short-Term Memory (Attention-BiLSTM) model and transfer learning. Based on Attention-BiLSTM model, three models including Pre-training (PRET), Multitask learning (MTL), and Pre-training & Multitask learning (PRET+MTL) are proposed to transfer the knowledge obtained from document-level training of sentiment classification to aspect-level sentiment classification. Finally, the performance of the four models is verified on four datasets. Experiments show that proposed methods make up for the shortcomings of poor training of neural network models due to the small dataset of the aspect-level sentiment classification.

1. Introduction

With the rapid development of e-commerce, more and more consumers are commenting on the Internet platform. Sentiment analysis of comments is an important task in natural language processing (NLP) area which obtains much attention recently [1]. Aspect-level sentiment analysis is a fine-grained sentiment classification task, aiming to judge the sentiment polarity of the opinion target in a sentence [2]. The main purpose is to give a series of concise expressions according to these comments to show how much the consumer prefers something. For example, in the sentence"The food is delicious while the price is expensive", the user speaks positively and negatively towards two aspect targets, i.e., "food" and"price".

With the rise of deep learning, neural network models have shown state-of-the-art performance in many NLP tasks such as machine translation [3], semantic recognition [4], question answering [5], and text summarization [6]. Recurrent Neural Network (RNN) based on attention mechanism has become a mainstream approach in aspect-level sentiment analysis, where RNN is designed to capture local correlation and the attention mechanism is used to assign different weights to each vector to make the result of text representation more reasonable [7–10]. Attention-based long short-term memory (LSTM) network is the best neural network model [11] for obtaining the context features of each word in a long distance, with the goal of adjusting weights to obtain the overall features of the text.

Generally, neural network models can show excellent performance only under the training of large-scale corpus. However, the generation of aspect-level training data needs to manually label aspect targets, which is very challenging in practice. Therefore, there are a few aspect-level corpora at present, which greatly limits the performance of neural network models. Since a large number of corpora with document-level sentiment labels are easy to obtain, such as Amazon comments, collecting users' preferences about different aspect categories become practicable. Inspired by the observation, transfer learning methods are proposed to improve the performance of aspect-level sentiment classification by transferring knowledge from the domain of coarse-grained document-level task to the domain of fine-grained aspect-level sentiment classification task.

In this paper, the main contributions of our work can be summarized as follows:

1. Attention-based BiLSTM method is applied to aspect-level sentiment classification. The attention mechanism can be used to extract important features from the sequence according to the weight distribution.
2. Aspect plays a key role in aspect-level sentiment classification. However, aspect-level datasets are small. Thus, three methods are proposed to consider the information of

* Corresponding author.
E-mail addresses: guixian_xu@muc.edu.cn (G. Xu), 20301825@muc.edu.cn (Z. Zhang), zhang-ting@muc.edu.cn (T. Zhang), Yusn@muc.edu.cn (S. Yu), 347397914@qq.com (Y. Meng), 18631333368@163.com (S. Chen).

https://doi.org/10.1016/j.knosys.2022.108586
0950-7051/

aspect in the training process: PRET, MTL, and the combination of PRET and MTL.

3. In this paper, four public datasets are used to verify the performance of the proposed models, including SemEval 2014 Laptop, SemEval 2014 Restaurant, SemEval 2015 Restaurant, and SemEval 2016 Restaurant. At the same time, the performance of the models is compared with the current mainstream neural network models. Experiments show that the proposed aspect-level sentiment analysis methods are effective in the sentiment polarity judgment of a given opinion target in the sentence.

2. Background

Sentiment analysis is a study of people's sentiment or attitude towards something such as organizations, individuals, products, services, events or topics. The extensive research of sentiment analysis is bounded up with the development of social networks, such as Weibo,[1] Twitter,[2] and so on. Opinion information is significant to businesses because they can understand consumers' preferences for their products. The information is also important to the government, because they want to know the opinions of the public about existing policies and upcoming policies.

Aspect-level sentiment classification is usually regarded as a classification problem with the purpose of judging the sentiment polarity of the opinion target in the sentence [12]. It consists of three steps: identification, classification, and aggregation [13]. Firstly, sentiment–target pairs are identified in the text. Secondly, expressed sentiments are classified according to a set of sentiment values predefined, such as positive and negative. Finally, sentiment values are aggregated for each aspect to provide a concise overview.

The improvement of traditional sentiment classification methods is mainly based on feature engineering. Kiritchenko, Zhu, Cherry, & Mohammad (2014) [14] first obtained n-gram features, analytic features, and sentiment dictionary-based features. Then the features were sent to Support Vector Machines (SVM) for classification. Finally, the algorithm was applied to customer reviews. Wagner, Arora, Cortes, Barman, & Tounsi (2014) [15] judged the sentiment polarity based on various features extracted by manual design. Yi, Yi, & Zhou (2016) [16] used the positive and negative sentiment value of words to generate sentiment feature vectors. Then the model was trained by Naive Bayes and SVM methods. Finally, the method was applied to Twitter tweets for sentiment analysis. Although traditional methods have achieved good results in aspect-level sentiment classification, the performance of these methods is highly dependent on the quality of the designed features, which results in a lot of efforts and time spent on feature designing task. In addition, these models cannot capture semantic correlations between aspects and their contexts, which limits the generalizability of such methods [17]. To overcome the problem, neural network models have been successfully applied in many NLP tasks [11,18–20], which can help the model capture semantic relevance between aspects and contexts [4,8,21–23]. Dong, Wei, Tan, Tang, Zhou, and Xu (2014) [21] put forward to apply RNN to aspect-level sentiment classification for the first time. Sentiment polarity information was obtained from a text by an adaptive RNN. Syntactic structure information of the sentence was used to assist the model to improve the accuracy of sentiment classification. The model has attracted academic attention. However, subsequent studies have found that such models are strongly dependent on syntax and susceptible to parsing errors [24]. In particular, the performance is unstable when dealing with non-written expressions, such as Twitter data. In recent years, most of the aspect-level sentiment classification models with better performance are based on RNN. For example, Tang, Qin, Feng, and Liu (2016) [25] used two LSTMs to encode sentences from the left side and right side of the word. Then, the final output of the two networks was connected as the ultimate sentiment expression. Zhang, Zhang, and Vo, D.-T. (2016) [24] used a threshold neural network to model the syntactic and semantic information of sentences and the contextual information of aspect words. These RNN-based models have achieved good classification results. However, due to the characteristics of the RNN model itself, RNN cannot capture the hidden association between sentiment words which are relatively far away in complex sentences. For example, the LSTM network and the threshold network tend to pay attention to shorter-term input. Besides, Tai, Socher, & Manning (2015) [26] proposed Tree-Structured LSTM Networks, which has been proven to be very effective in many NLP tasks by exploiting the grammatical structure of sentences.

In recent years, with the successful application of the attention mechanism in image processing, some researchers have begun to apply the attention mechanism to NLP and achieved excellent results. For example, Wang, Huang, Zhao, & Zhu (2016) [10] proposed attention-based LSTM model with aspect embedding. Hidden layer outputs of each word and vector representation of the aspect word were obtained by using LSTM to encode the input sentence and the given aspect word, respectively. Then the output of the hidden layer was processed by attention mechanism. The sentiment polarity expression of aspect words was obtained by stitching the attention vector and the aspect word vector. Tang, Qin, & Liu (2016) [8] proposed a MemNet model based on attention mechanism. In the model, attention learning was performed by the external memory which is composed of word vectors of input sentences. Each layer of the model recalculates the attention distribution based on the results of the previous layer output. Attentional fine-tuning was achieved with the feature abstraction capabilities of deep networks. Finally, the sentiment polarity of the words in a given aspect was obtained. Chen, Sun, Bing, & Yang (2017) [9] proposed an attention-based Recurrent Attention on Memory (RAM) model based on MemNet. Gated Recurrent Unit (GRU) network is used to realize multi-layered abstraction of attention. Then the sentiment polarity of a given aspect word is obtained through nonlinearly combining the information captured by different attention layers. The three models mentioned above are the models with better comprehensive performance in aspect-level sentiment analysis tasks in the past two years. The commonality is the utilization of attention-based mechanism with vector similarity to obtain the sentence sentiment expression vector of a given aspect word, from where sentiment polarity classification is performed. Due to the impact neutrality and ambivalence have on sentiment analysis, many recent works focused on neutrality detection and ambivalence handling where sentiment-based deep models can be applied [27,28]. Wallaart & Frasincar (2019) [29] proposed a two-stage sentiment analysis algorithm. In the method, a lexicalized domain ontology was used to predict the sentiment and a neural network with a rotatory attention mechanism was employed. Recent research shows that figurative language poses a challenge to sentiment analysis [30]. Zhang S., Zhang X., Chan J.& Rosso P. (2019) [31]proposed three new methods based on attention models and applied sentiment-based transfer learning to irony detection. Based on the systematic research on Affective Computing and Sentiment Analysis, many new methods have been proposed to apply deep models to sentiment systems [32,33], such as symbolic and subsymbolic AI [34] and Convolutional Stacked Bidirectional LSTM with a Multiplicative Attention Mechanism [35]. It has been demonstrated that the context attention

[1] https://weibo.com/.

[2] http://www.twitter.com/.

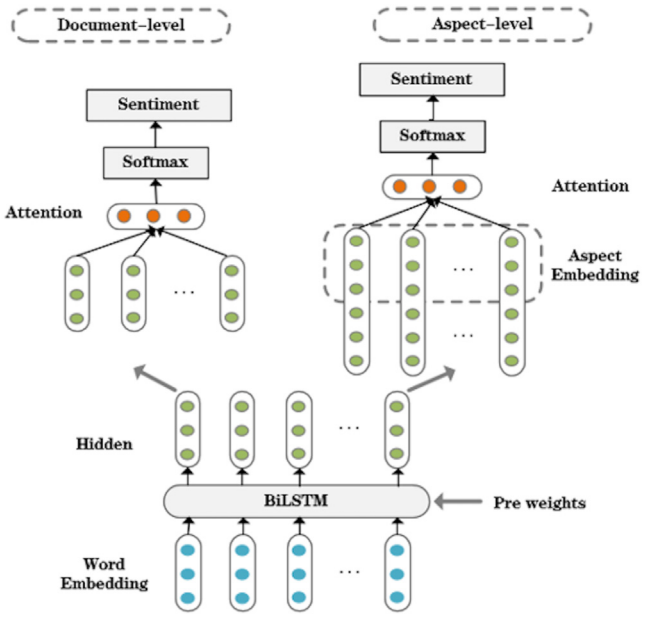

Fig. 1. The frame of Pre-training & Multitask learning model.

mechanism is effective in capturing the important information about given aspects [36,37], and the Hybrid Approaches have been proven to be effective in many aspect-based sentiment analysis tasks [38,39], which are only based on aspect-level data sets.

However, neural network models are data-driven, needing a lot of aspect-level data for training effective neural network models. To solve the problem of the shortage of aspect-level datasets, we proposed models in the paper that can train effective models based on a large number of document-level data.

3. Research methods

The frame diagram is shown in Fig. 1.

3.1. Attention-based BiLSTM method

The model consists of five layers:

(a) Input layer: input sentence to the model.
(b) Embedding layer: express each word in word vector.
(c) BiLSTM layer: obtain high-level features from embedding layer by BiLSTM.
(d) Attention layer: by adjusting the weight, let the model focus on useful information in the data to improve the quality of output.
(e) Output layer: output sentiment polarity.

The model structure is shown in Fig. 2:

3.1.1. Embedding layer

For a given sentence with T words $S = \{x_1, x_2, \ldots, x_T\}$, every word x_i is converted into a real-valued vector e_i. Then the sentence will be transformed into a real-valued vector $emb_s = \{e_1, e_2, \ldots, e_T\}$.

3.1.2. BiLSTM layer

The main idea of LSTM is to introduce the gate mechanism and memory unit, which can control each LSTM unit to extract important features and discard unimportant feature. LSTM unit is shown in Fig. 3. Table 1 shows the symbols in Fig. 3.

Table 1
Symbol description.

Symbol	Description
x_t	Current input
h_{t-1}	Previous step generated
c_{t-1}	Current state of this cell
i_t	Input gate
f_t	Forget gate
c_t	Cell state
o_t	Output gate

Input gate i_t with corresponding weight matrix W_{xi}, W_{hi}, W_{ci}, b_i, forget gate f_t with corresponding weight matrix W_{xf}, W_{hf}, W_{cf}, b_f, output gate o_t with corresponding weight matrix W_{xo}, W_{ho}, W_{co}, b_o. Finally, the current hidden state of the output h_t is obtained by multiplying the current cell state by the weight matrix of the outputs:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = i_t g_t + f_t c_{t-1} \quad (3)$$

$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (5)$$

$$h_t = o_t \tanh(c_t) \quad (6)$$

For the task of sequence modeling, the information in the context of future and past is equally important. However, the future context is ignored by standard LSTM networks because LSTM processes sequences in a temporal order. Then unidirectional LSTM networks are extended in Bidirectional LSTM networks. The basic structure of the BiLSTM network is shown in Fig. 4. As shown in Fig. 4, the output of the ith word is shown in the following formula:

$$h_i = \underset{h_i}{\rightarrow} \oplus \underset{h_i}{\leftarrow} \quad (7)$$

3.1.3. Attention layer

Attention mechanism is proposed for classification tasks. The calculation method of attention is applied to both document-level and aspect-level classification tasks, where the attention calculation for the document-level classification task considers the output of the hidden layer of BiLSTM without considering the aspect embedding. H_h is a matrix consisting of output vectors $\{h_1, h_2, \ldots, h_T\}$ that the BiLSTM layer produced with corresponding weight matrix W_h. T is the sentence length. H_a represents the embedding of aspect. α is an attention weight vector and the representation r of the sentence is formed by a weighted sum of these output vectors:

$$H = \tanh\begin{pmatrix} W_h H_h \\ W_a H_a \otimes e_T \end{pmatrix} \quad (8)$$

$$\alpha = softmax(w^T H) \quad (9)$$

$$r = H_h \alpha^T \quad (10)$$

where $H \in R^{(d+d^a)\times T}$, $H_a \in R^{d^a}$, $H_h \in R^{d\times T}$, $e_T \in R^T$, $\alpha \in R^T$, $r \in R^d$. $W_h \in R^{d\times d}$, $W_a \in R^{d^a \times d^a}$, $w \in R^{d+d^a}$ are projection parameters. d is the size of hidden layers and d^a is the dimension of the aspect embedding. e_T is a column vector of 1s and $H_a \otimes e_T$ represents that the vector H_a is repeatedly concatenated T times, for which we adopt the operator used in Wang, Huang, Zhao, & Zhu (2016) [10]. w is a trained parameter vector and w^T is a transpose matrix. The hidden state $h^* \in R^d$ and the final sentence-pair representation is obtained to use for classification from:

$$h^* = \tanh(r) \quad (11)$$

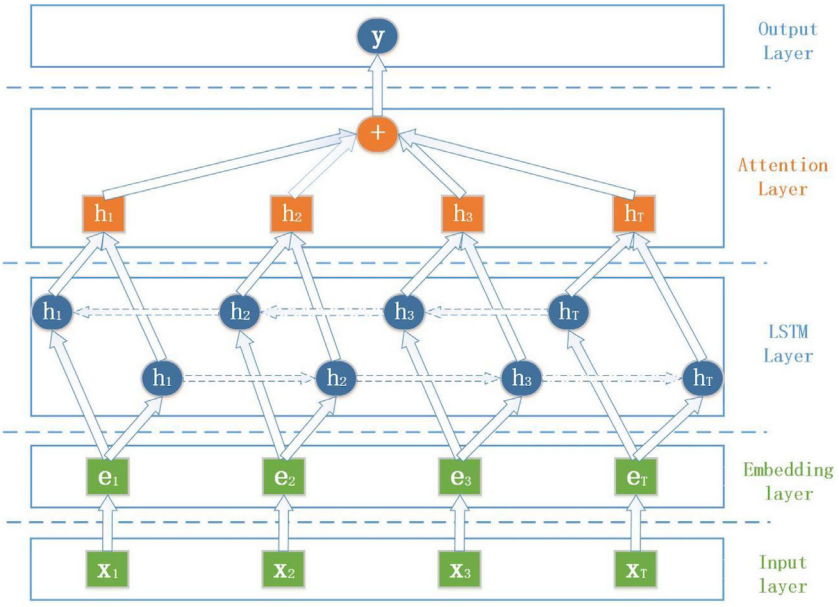

Fig. 2. Attention-based BiLSTM model.

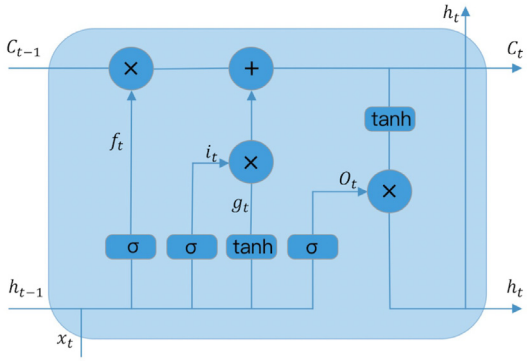

Fig. 3. Long short term memory unit.

3.1.4. Classification

In this setting, a softmax classifier is used to predict label $\hat{y}$ from a discrete set of classes Y for a sentence S. The classifier takes the hidden state h^* as input:

$$\hat{y}(y|S) = softmax(W^{(S)}h^* + b^S) \tag{12}$$

$$\hat{y} = argmax_y \hat{p}(y|S) \tag{13}$$

The cost function is the negative log-likelihood of the true class label $\hat{y}$:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda||\theta||_F^2 \tag{14}$$

where $t \in R^m$ is the one-hot represented ground truth and $y \in R^m$ is the estimated probability for each class by softmax. m is the number of target classes, and λ is an L2 regularization hyper-parameter. In this paper, L2 regularization is used to dropout to alleviate overfitting.

3.2. Transfer learning

Transfer learning: In traditional machine learning, to ensure the reliability and accuracy of the model, there are some conditions: (1) The training and the test data satisfy the conditions of independent and identical distribution. (2) A good classification model needs a lot of experimental data to be trained. However, these two conditions are often unsatisfactory in practical applications. Transfer learning uses existing knowledge to solve the problems of different but related fields, which loosens the hypotheses in traditional machine learning. Transfer learning can migrate models suitable for big data to the model for small data, which can avoid the work of annotating data tags and greatly improve the learning performance.

From a theoretical perspective, the following issues are studied in the transfer learning: condition for migration and migration algorithm.

PRET: To solve the problem of transfer learning above, PRET method is proposed to use for training model. Training a deep neural network model often requires a lot of labeled data for a long time. But getting a large annotated aspect-level data is often very expensive. It can be found that in Amazon reviews, enormous document-level labeled data online are easily accessible. The reviews contain abundant document-level data with sentiment labels. The purpose of PRET is to obtain a model which can extract universal features and apply the model to the network structure of various tasks. A pre-trained model is created to solve similar problems. In the field of NLP, PRET is mainly used for pre-training language models. And it is usually applied to tasks such as text classification, text similarity, text generation, sequence labeling and so on.

In this paper, the document-level dataset is first trained with BiLSTM model to get pre-trained weights. Related parameters are initialized with pre-trained weights. Then train them on the aspect-level dataset to fine-tune these weights.

MTL: To solve the problem of transfer learning above, a parameter multitasking learning method is proposed for shared space to train the model. MTL is a learning process, in which multiple related tasks are considered simultaneously. The idea of multitasking learning is to improve the generalization performance of single-task learning by utilizing the intrinsic relationship between tasks. In multitasking deep neural networks, computational complexity could be reduced by sharing low-level semantic information. Several common tasks could be combined correlation information by shared presentation layer well. The overall loss function is given by:

$$L(\theta) = J(\theta) + \lambda J^*(\theta) \tag{15}$$

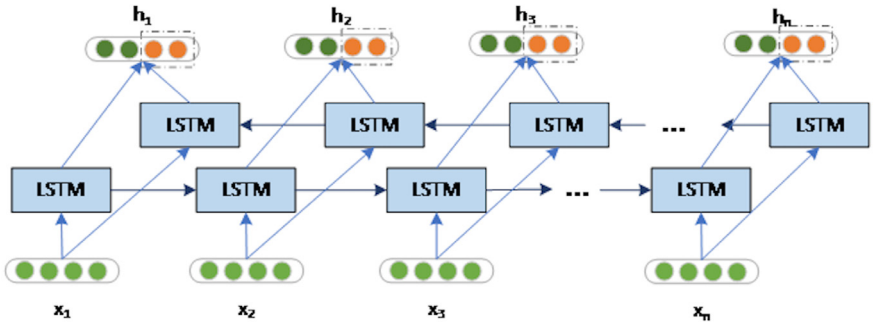

Fig. 4. Basic structure of the BiLSTM network.

Table 2
Dataset description.

Dataset	Aspect	Positive	Negative	Neural	Entropy
D1	14Restaurant-Train	2164	807	637	1.37
	14Restaurant-Test	728	196	196	1.28
D2	14Laptop-Train	994	870	464	1.52
	14Laptop-Test	341	128	169	1.46
D3	15Restaurant-Train	1178	382	50	0.98
	15Restaurant-Test	439	328	35	1.20
D4	16Restaurant-Train	1620	709	88	1.08
	16Restaurant-Test	597	190	38	1.03

Table 3
Experimental parameter setting.

Parameter	Value
Embedding dimension	300
λ	0.1
Dropout probability	0.5
Epochs	15
Optimizer	RMSProp
Decay rate	0.9
Learning rate	0.001
Batch size	31

where $J(\theta)$ is the loss function of aspect-level classification and $J^*(\theta)$ is the loss function of document-level classification. $\lambda \in (0, 1)$ is a hyperparameter that controls the weight of $J^*(\theta)$.

In this study, document-level and aspect-level tasks are proposed to train simultaneously. In this setup, the Embedding layer and BiLSTM layer are shared by two tasks. The document is represented as an average vector on the BiLSTM output.

PRET+MTL: In this setup, PRET is first used to perform on the document-level dataset to get pre-trained weights. The pre-trained weights are used to initialize parameters for aspect-level model and document-level model. Then MTL is conducted.

4. Experiments

4.1. Dataset

For aspect-level datasets, the validity of the models is verified on four public datasets including SemEval 2014 Laptop, SemEval 2014 Restaurant,[3] SemEval 2015 Restaurant,[4] and SemEval 2016 Restaurant.[5] The following are represented by D1, D2, D3, and D4 respectively. The details of the datasets are shown in Table 2. Samples with conflicting polarities in the dataset have been removed.

For document-level datasets, two datasets from Yelp2014[6] and the Amazon Electronics[7] dataset were derived respectively. Since the Yelp dataset is from the similar domain as D1, D3, and D4, we used the Yelp dataset for PRET and MTL on D1, D3, and D4 datasets. The Electronics dataset is from the similar domain as D2, so the Electronics dataset is used on D2. The reviews are rated at five levels. To unify standards, 3-level sentiment classification is used in the experiments. In order to clearly distinguish sentiment polarity, reviews with a score of 1 are regarded as negative corpus, with a score of 3 are regarded as neutral corpus, with a score of 5 are regarded as positive corpus. There are 15000 data in each dataset, and the three categories are equally distributed.

4.2. Experiment setup

The 300-dimensional pre-trained word vector GloVe publicly released by Stanford University is used as the dictionary in this study [40]. GloVe vectors are used to train the initial parameters of the untrained word vector. There is no official development dataset in aspect-level dataset, so 20% of the initial training data were randomly sampled as the development set and the remaining 80% were used for training. For all neural network models, experimental parameters are shown in Table 3:

To guarantee the accuracy of experiments, average values of five randomly initialized experiments were used as results.

In this study, Accuracy and Macro-F1 are used as evaluation metrics.

4.3. Experiment

4.3.1. Model comparison

To further verify the effectiveness of aspect-level sentiment analysis methods proposed in this paper, four existing aspect-level sentiment analysis methods are compared with the proposed method based on the same corpus. The methods for comparison are as follows:

(a) TangA [8]: They developed two target dependent LSTM models.
(b) Wang [10]: They proposed an attention-based LSTM for aspect-level sentiment classification.
(c) TangB [25]: They introduced a deep memory network for aspect-level sentiment classification.
(d) Chen [9]: They adopted multiple-attention mechanism to capture sentiment features separated by a long distance.
(e) The proposed method: PRET, MTL, and Attention-BiLSTM model are simultaneously applied to aspect-level sentiment analysis tasks.

[3] http://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools.
[4] http://alt.qcri.org/semeval2015/task12/index.php?id=data-and-tools.
[5] http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools.
[6] https://www.yelp.com/dataset.
[7] http://jmcauley.ucsd.edu/data/amazon/.

Table 4
The results of four models of existing aspect-level sentiment analysis.

Methods	D1		D2		D3		D4	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
TangA	75.37	64.51	68.25	65.96	76.39	58.70	82.16	54.21
Wang	78.60	67.02	68.88	63.93	78.48	62.84	83.77	61.71
TangB	76.87	66.40	68.91	62.79	77.89	59.52	83.04	57.91
Chen	78.48	68.54	72.08	68.43	79.98	60.57	83.88	62.14
Proposed method	**80.89**	**72.17**	**72.57**	**69.34**	**82.91**	**70.54**	**86.94**	**71.25**

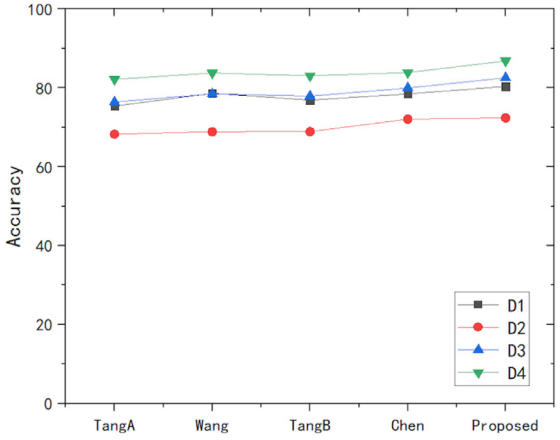

Fig. 5. Comparison of accuracy of various aspect-level sentiment experiments.

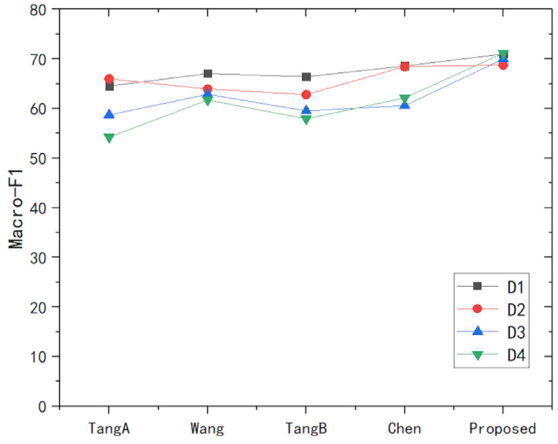

Fig. 6. Comparison of macro-F1 of various aspect-level sentiment experiments.

Experimental results of the four different methods and the proposed method are shown in Table 4, where data in bold are the best experimental results with the same dataset. It can be observed that the proposed method yields better performance on all datasets, especially on D1, D3, and D4. The proposed model achieves more significant improvements on F1, with a maximum improvement of 17% on F1 compared to the other four methods. The proposed method is helpful to capture the domain-specific opinion words through the knowledge learned from document-level datasets, and helps to better exploit the intrinsic relationship between document-level and aspect-level tasks to improve the generalization performance of aspect-level task learning.

According to the results in Table 4, a comparison of the accuracy of various aspect-level sentiment experiments is shown in Figs. 5, 6.

From Figs. 5 and 6, it can be seen that in general, the results of the proposed method are higher than those of the other four experiments on the four datasets, which illustrates the proposed method can effectively deal the task of aspect-level sentiment classification.

In addition, as can be seen from the experimental results, the performance of Chen's experimental method is significantly higher than that of the other three methods. Compared with the other three experiments, the most outstanding feature of the method is the adoption of multiple-attention mechanism to enhance the expressive ability for handling complications. At the same time, it can be seen that Wang's method shows good experimental results on D1 and D3. Compared with other experiments, the most prominent feature of this method is that they proposed attention-based LSTM in the task. Attention can extract important features from the sequence according to the weight distribution. It also illustrates that it is appropriate to introduce attention mechanism into our models.

4.3.2. Ablation tests

To better understand the effectiveness of Attention mechanism, PRET, and MTL on aspect-level sentiment classification, we performed an ablation test. In the experiment, SVM and BiLSTM are compared as baseline. Other compared methods are shown below.

(1) SVM: A standard SVM with RBF kernel. The parameter C is set to 1.0 and g is set to 0.5, the pre-trained GloVe vectors are used as input to the SVM classifier.
(2) BiLSTM: Bidirectional LSTM networks. The model is used as a baseline to validate the effectiveness of our proposed method.
(3) Attention-BiLSTM: BiLSTM model based on attention. The model is used to verify the effect of attention mechanism in aspect-level sentiment analysis.
(4) Attention-BiLSTM+PRET: PRET based on Attention-BiLSTM model. The model is used to verify the effect of PRET in aspect-level sentiment analysis.
(5) Attention-BiLSTM+MTL: MTL based on Attention-BiLSTM model. The model is used to verify the effect of MTL.
(6) Attention-BiLSTM+PRET+MTL: PRET and MTL based on Attention-BiLSTM model. This model is used to verify the effect of the combination of PRET and MTL in aspect-level sentiment analysis.

The experimental results are shown in Table 5, where the bold data is the best experimental result under the same dataset.

Based on the above results, the conclusions are as follows:

First of all, compared with traditional machine learning methods, the proposed method is better than the experimental results based on SVM on four data sets. Because SVM method cannot capture the semantic correlation between aspects and their context, it limits the generalizability of such methods. Then compared with neural network methods, BiLSTM model based on attention is introduced, and the results of Attention-BiLSTM are higher than that of the experiment based on BiLSTM. For all data sets, the accuracy of Attention-BiLSTM is improved. It shows that, on the one hand, BiLSTM enables every moment to contain context information before and after it and BiLSTM enables to understand

Table 5
The results of aspect-level sentiment analysis.

Methods	D1		D2		D3		D4	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
SVM	80.16	68.63	70.49	67.85	68.72	63.55	74.82	64.22
BiLSTM	76.52	65.36	68.06	65.21	76.34	55.38	83.21	59.45
Attention-BiLSTM	77.94	67.59	69.18	65.92	78.49	61.63	83.84	60.23
Attention-BiLSTM+PRET	79.83	70.23	**72.69**	**70.58**	81.76	69.93	85.60	69.24
Attention-BiLSTM+MTL	79.57	68.54	70.38	67.28	82.34	67.26	85.29	67.92
Attention-BiLSTM+PRET+MTL	**80.89**	**72.17**	72.57	69.34	**82.91**	**70.54**	**86.94**	**71.25**

the semantic information of the current vocabulary to the greatest extent. On the other hand, attention architecture can extract significant features from the sequence according to the weight distribution. The results show Attention-BiLSTM is very useful to improve the experimental performance in aspect-level sentiment analysis.

Secondly, it can be observed that PRET is very helpful for improving the accuracy and macro-F1 of aspect-level sentiment classification by comparing the results of Attention-BiLSTM and Attention-BiLSTM+PRET. Compared with baselines, improvements in macro-F1 are more, especially on the datasets of D3 and D4 because the distribution of D3 and D4 are extremely unbalanced. In Attention-BiLSTM+PRET, firstly, the document-level dataset is trained to get pre-trained weights, and then parameters of the task are initialized with the pre-trained weights. In this way, the model can be better pre-trained by the method. Then experimental results can be gotten by fine-tuning the parameters in the shortest time. It can be proved that the pre-trained method proposed in this paper is applicable to aspect-level sentiment analysis.

Results show that the performance of Attention-BiLSTM+MTL on D1 and D2 is similar to that of Attention-BiLSTM model. But for D3 and D4, the improved effect of MTL can be observed. The improvement is much higher than that of the experiment with BiLSTM. The reason is that the performance of learning can be improved through the internal relationship among multiple tasks, especially when the datasets are unbalanced.

Finally, it can be seen that the combination of PRET+MTL yielded better results as a whole. Among them, the best experimental results of each dataset are obtained on D1, D3, and D4. The experimental result of using PRET + MTL on D2 is only slightly lower than the model only using PRET. There are some reasons: Firstly, only one loss function was defined in the experiments for multiple tasks. Ideally, each task should have a well-defined loss function. However, the scale of loss for different tasks is very different. Thus, there should be multiple losses for multiple tasks. Secondly, as one of the most critical hyperparameters, learning rate may not have been adjusted to the optimal coefficient.

At the same time, as can be seen from Table 5, when Macro-F1 is observed separately, the maximum improvement of Attention-BiLSTM+PRET+MTL on Macro-F1 in the four datasets is 6.81%, 4.10%, 15.61%, 11.80% compared with other methods. The improvement of Macro-F1 on the D3 and D4 datasets is significantly higher than that on the D1 and D2 datasets. There are two main reasons to explain. Firstly, there are more neutral data in D1 and D2, but less in D3 and D4. Therefore, neutral features can be learned by classifiers from D1 and D2. Secondly, the number of neutral samples in the D3 and D4 test groups is very small. When the data is minimal, accuracy and F1 will be greatly affected. For example, if five neutral samples are correctly identified, the recall rate of two datasets will increase by more than 10%. Therefore, Macro-F1 on D3 and D4 is more affected.

4.3.3. Effect of pre-training

To illustrate that the proposed method is really useful in aspect-level sentiment classification, experiments are conducted to change the percentage of document-level training datasets for PRET from 0% to 100%. The experiments are shown in Tables 6 and 7.

Table 6
Accuracy with different percentages of document-level datasets.

Percentage (%)	0	20	40	60	80	100
D1	78.31	78.65	79.07	79.46	79.92	80.89
D2	68.93	68.47	69.91	71.52	71.95	72.57
D3	78.28	80.86	80.41	81.27	82.49	82.91
D4	84.81	85.38	85.97	86.43	86.08	86.94

Table 7
Macro-F1 with different percentages of document-level datasets.

Percentage (%)	0	20	40	60	80	100
D1	67.51	67.38	68.59	69.02	69.63	72.17
D2	65.94	64.78	66.82	67.53	67.80	69.34
D3	61.93	65.25	68.61	68.74	68.62	70.54
D4	60.39	66.17	69.84	70.39	69.27	71.25

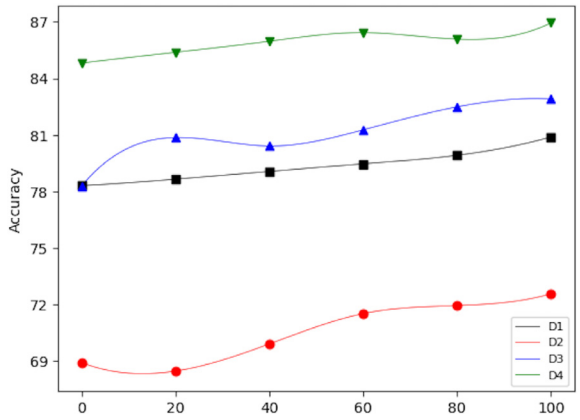

Fig. 7. The change of accuracy with different percentages of document-level datasets.

According to the data in Tables 6 and 7, the change with different percentages of document-level training examples is shown in Figs. 7 and 8.

It can be seen that with the step size increases, accuracy changes less, and Macro-F1 changes more. However, they are all in a steady upward trend. For accuracy, with the number of document examples increasing, the improvement in accuracy is stable. The results illustrate that the use of document-level data can effectively improve the performance of the model, which also preliminarily proves that the models are helpful to improve the performance. For macro-F1, it is found that the improvement is stable on D1 and D2, but rough on D3 and D4 datasets. The reason is probably that the data of D3 and D4 is very unbalanced. However, the situation can be greatly improved when using pre-trained knowledge. When the percentage increases to 80%, the performance fluctuations occur. It is possible that the model is over-fitting. Finally, when the percentage increases to 100%, the

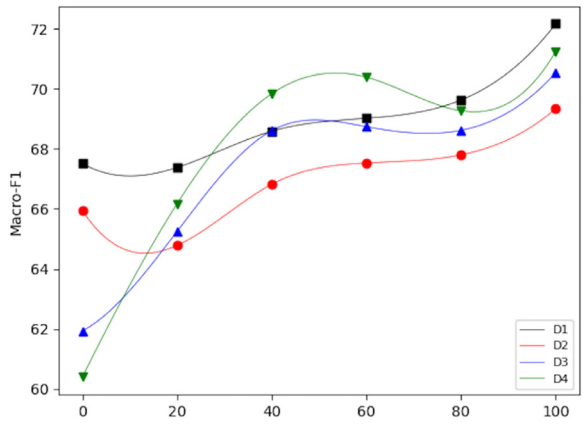

Fig. 8. The change of Macro-F1 with different percentages of document-level datasets.

performance is optimal. It also illustrates that the proposed methods are very effective for improving the accuracy of aspect-level sentiment classification.

5. Conclusion

In this paper, aspect-level sentiment classification based on transfer learning and Attention-BiLSTM model is proposed. Based on Attention-BiLSTM model, PRET, MTL, and PRET+MTL are proposed to transfer the knowledge obtained from document-level training of sentiment classification to aspect-level sentiment classification. Our contribution is twofold: First, BiLSTM model based on attention is applied to aspect-level sentiment classification to extract significant features from the sequence according to the weight distribution. Second, the proposed methods using PRET and MTL solve the problem that the neural network model training performs not well due to the lack of aspect-level sentiment classification dataset. However, there are still many problems needing to be considered. In our future work, we will try to solve the problems of parameter training and selection, as there are many factors that affect the performance of aspect-level sentiment classification, such as syntactic structure and semantic information. In addition, we will explore ways to improve the fine-grained scale of sentiment categories by investigating ambivalence handling methods, that could facilitate to reveal fine-scaled emotions, potentially enhancing the model's performance.

CRediT authorship contribution statement

Guixian Xu: Conceptualization, Methodology, Software. **Zixin Zhang:** Data curation, Writing – reviewing and editing. **Ting Zhang:** Visualization, Validation. **Shaona Yu:** Supervision. **Yueting Meng:** Writing – original draft, Investigation. **Sijin Chen:** Software.

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

This work was supported by the Chinese National Funding of Social Sciences [grant number 19BGL241], in part by the Humanities and Social Science Fund of Ministry of Education of the People's Republic of China [grant number 18YJA740059].

References

[1] T. Nasukawa, J. Yi, Sentiment analysis: Capturing favorability using natural language processing, in: Proceedings of the 2nd International Conference on Knowledge Capture, 2003, pp. 70–77.
[2] B. Pang, L. Lee, Opinion mining and sentiment analysis, Comput. Linguist. 35 (2) (2009) 311–312.
[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270.
[4] W. Yin, H. Schütze, B. Xiang, B. Zhou, Abcnn: Attention-based convolutional neural network for modeling sentence pairs, Trans. Assoc. Comput. Linguist. 4 (2016) 259–272.
[5] X. He, D. Golub, Character-level question answering with attention, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1598–1607.
[6] A.M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 379–389.
[7] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, in: IJCAI'17, AAAI Press, 2017, pp. 4068—4074.
[8] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 214–224.
[9] 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.
[10] Y. Wang, M. Huang, X. Zhu, L. Zhao, 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.
[11] 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.
[12] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2015) 813–830.
[13] M. Tsytsarau, T. Palpanas, Survey on mining subjective data on the web, Data Min. Knowl. Discov. 24 (3) (2012) 478–514.
[14] S. Kiritchenko, X. Zhu, C. Cherry, S. Mohammad, NRC-Canada-2014: Detecting aspects and sentiment in customer reviews, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 437–442.
[15] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, L. Tounsi, DCU: Aspect-based polarity classification for SemEval task 4, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEVal 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 223–229.
[16] Y. Shun-Ming, et al., Twitter sentiment classification with sentimental feature vector, J. Chin. Comput. Syst. 37 (11) (2016) 2454–2458.
[17] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent twitter sentiment classification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 151–160.
[18] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.
[19] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, in: NIPS'15, MIT Press, Cambridge, MA, USA, 2015, pp. 2440–2448.
[20] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 946–956.
[21] 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 (Volume 2: Short Papers), 2014, pp. 49–54.
[22] M. Zhang, Y. Zhang, D.-T. Vo, Gated neural networks for targeted sentiment analysis, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, in: AAAI'16, AAAI Press, 2016, pp. 3087—3093.
[23] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32 (1), 2018.

[24] D.-T. Vo, Y. Zhang, Target-dependent Twitter sentiment classification with rich automatic features, in: Proceedings of the 24th International Conference on Artificial Intelligence, in: IJCAI'15, AAAI Press, 2015, pp. 1347—1353.
[25] D. Tang, B. Qin, X. Feng, T. Liu, Effective LSTMs for target-dependent sentiment classification, 2015, arXiv preprint arXiv:1512.01100.
[26] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1556–1566.
[27] A. Valdivia, M.V. Luzón, E. Cambria, F. Herrera, Consensus vote models for detecting and filtering neutrality in sentiment analysis, Inf. Fusion 44 (2018) 126–135.
[28] Z. Wang, S.-B. Ho, E. Cambria, Multi-level fine-scaled sentiment sensing with ambivalence handling, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 28 (04) (2020) 683–697.
[29] O. Wallaart, F. Frasincar, A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models, in: European Semantic Web Conference, Springer, 2019, pp. 363–378.
[30] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, A. Reyes, SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEVal 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 470–478.
[31] S. Zhang, X. Zhang, J. Chan, P. Rosso, Irony detection via sentiment-based transfer learning, Inf. Process. Manage. 56 (5) (2019) 1633–1644.
[32] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[33] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis, Future Gener. Comput. Syst. 115 (2021) 279–294.
[34] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, in: CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 105—114.
[35] A.K. Ja, T. Esther, A convolutional stacked bidirectional LSTM with a multiplicative attention mechanism for aspect category and sentiment detection.
[36] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of the 2018 World Wide Web Conference, in: WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, pp. 1023–1032.
[37] B. Xing, t. nguyen, D. Song, J. Wang, F. Zhang, Z. Wang, H. Huang, Earlier attention? Aspect-aware LSTM for aspect-based sentiment analysis, 2019, pp. 5313–5319.
[38] D. Meškelė, F. Frasincar, ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model, Inf. Process. Manage. 57 (3) (2020) 102211.
[39] M.M. Truşc, D. Wassenberg, F. Frasincar, R. Dekker, A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention, in: International Conference on Web Engineering, Springer, 2020, pp. 365–380.
[40] J. Pennington, R. Socher, C.D. 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.