Knowledge-Based Systems 226 (2021) 107136

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Positionless aspect based sentiment analysis using attention mechanism

Rohan Kumar Yadav *, Lei Jiao, Morten Goodwin, Ole-Christoffer Granmo

Centre for Artificial Intelligence Research (CAIR), University of Agder, 4879 Grimstad, Norway

ARTICLE INFO

Article history:
Received 10 November 2020
Received in revised form 26 January 2021
Accepted 11 May 2021
Available online 13 May 2021

Keywords:
Aspect based sentiment analysis
Opinion lexicon
LSTM/GRU
Position embedding

ABSTRACT

Aspect-based sentiment analysis (ABSA) aims at identifying fine-grained polarity of opinion associated with a given aspect word. Several existing articles demonstrated promising ABSA accuracy using positional embedding to show the relationship between an aspect word and its context. In most cases, the positional embedding depends on the distance between the aspect word and the remaining words in the context, known as the position index sequence. However, these techniques usually employ both complex preprocessing approaches with additional trainable positional embedding and complex architectures to obtain the state-of-the-art performance. In this paper, we simplify preprocessing by including polarity lexicon replacement and masking techniques that carry the information of the aspect word's position and eliminate the positional embedding. We then adopt a novel and concise architecture using two Bidirectional GRU along with an attention layer to classify the aspect based on its context words. Experiment results show that the simplified preprocessing and the concise architecture significantly improve the accuracy of the publicly available ABSA datasets, obtaining 81.37%, 75.39%, 80.88%, and 89.30% in restaurant 14, laptop 14, restaurant 15, and restaurant 16 respectively.

1. Introduction

Aspect-based sentiment analysis (ABSA) is one of the sentiment analysis that aims to identify the polarity of aspect word associated with its context. It has been categorized as a standard evaluation framework for fine-grained sentiment analysis [1]. Among various aspect-based sentiment classification problems, we focus, in this study, on the task that is to map the polarity of the opinion on a aspect word into one of the following potential sentiments, namely, positive, neutral, or negative. For instance, the sentence "great food but the service was dreadful". has an aspect word "food" having a positive polarity and another aspect word "service" having a negative polarity in this context. ABSA includes various tasks including identification, classification, and aggregation. Most of the existing studies formulate ABSA as a classification problem where the information of aspect word is integrated [2]. Following the same stream of research, we also focus on the sentiment classification [3] in this article.

ABSA can be a quite challenging classification problem because of the ambiguity of sentiment in the sentence. The context-based feature usually plays an important role in the classification of sentiment, which introduces the hypothesis that the understanding of a word is mostly dependent on the context words and their locations. Hence both context words, as well as the position of the aspect word, become important features for sentiment classification [4,5]. Understandably, even human beings spontaneously search for context words to evaluate the sentiment of a word when we read an article. This naturally makes the context and the position information vital features to be embedded into the deep learning model for better performance.

Various neural network architectures, from simple to complex ones, have been developed for position-aware sentiment classifications with a focus on the aspect word [6,7]. A position encoding vector developed in [8] has been a popular choice for embedding positional information in Long Short term Memory (LSTM) based models. There the position index of the surrounding words is represented by the relative distance to the aspect word. Such position embedding creates a probability distribution among the context that is then embedded along with the word embedding of each word for classification of sentiments. However, a sophisticated neural network architecture is required for good performance because of the lack of sentiment lexicon knowledge with the integration of positional embedding [9]. Even a slight

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.

E-mail addresses: rohan.k.yadav@uia.no (R.K. Yadav), lei.jiao@uia.no (L. Jiao), morten.goodwin@uia.no (M. Goodwin), ole.granmo@uia.no (O.-C. Granmo).

https://doi.org/10.1016/j.knosys.2021.107136

increment of accuracy in ABSA task usually requires a more complex architecture [10].

In this paper, we propose a very simple preprocessing of ABSA task by using sentiment lexicons and a masking technique that removes complex positional embedding thereby requiring a very straightforward architecture to obtain the state-of-the-art performance. As we know, human being usually makes sentiment classification of a particular word based on the surrounding words. Besides, human being, most probably, understands the meaning of each word and the sentiment associated with it as a priori. On the contrary, a neural network does not have this inbuilt knowledge. Even though various pre-trained word embedding captures the semantic relationship among the words, they are usually complicated. Therefore, it is important to find an efficient way to offer the model necessary knowledge, as priori, as much as possible. To give the model extra knowledge about sentiment in a simple way, we employ Opinion Lexicon [11] that has a list of positive and negative sentiment words. We use these lexicons and replace all the possible positive words with the "positive" tokens and negative words with "negative" tokens. The words that are not in the Opinion Lexicon will be left as they are. Additionally, to avoid complex positional embedding, the aspect word is masked with a common token, making it a Masked Aspect Embedding and the original sentence as Sentence Embedding. Then, we adopt the Attention-based Bidirectional Gated Recurrent Unit (BiGRU) to train both the input to classify the sentiment of the masked aspect word. To evaluate the performance of the proposed methodology, we experiment with all available restaurant and laptop datasets of the ABSA task [12–14]. The numerical results show that the proposed scheme obtains either similar or higher accuracy compared with the state-of-the-art solutions that use positional embedding and a complex architecture.

The main contributions of the paper are summarized as follows:

1. Mask aspect words with common token and use it as aspect embedding along with original sentence embedding thereby removing the complex positional embedding.
2. Propose a very straightforward Attention based BiGRU architecture that performs either similar or better than the comparable state-of-the-art solutions.

The rest of the paper is organized as follows. We review related studies in Section 2. The proposed preprocessing and deep learning architecture are described in detail in Section 3. In Section 4, we show the experiment results and reveal the benefits of proposed schemes before concluding the paper in Section 5.

2. Related work

This section consists of three parts. The first part includes the related studies on sentiment analysis in general. The second part surveys, in brief, ABSA tasks based on LSTM as encoder [15]. The last part reviews the attention based ABSA models that highly depend on the positional embedding [16].

2.1. Sentiment analysis

Sentiment Analysis is a task involving polarity detection, subjectivity/objectivity identification as well as multi-modal fusion [17]. Sentiment analysis can be carried out in different levels, such as in document, sentence, or aspect level [18]. For document-level sentiment analysis, the goal is to detect the polarity of the whole document irrespective of any mentioned aspects. Tripathy et al. explored various machine learning algorithms on IMDB and polarity dataset demonstrating document level sentiment classification [19]. Other several large dataset has been explored to show that the character-level convolution networks could achieve state-of-the-art result [20]. A Linguistically Regularized LSTM is another variant of deep neural networks that can achieve competitive performance [21]. On the other hand, for sentence-level sentiment analysis, it has been developed in [22] a Bidirectional Emotional Recurrent Unit (BiERU) for Conversational Sentiment Analysis using generalized neural tensor block followed by a two-channel classifier. Various sentiment analysis tasks usually focus on analyzing data at the aggregate level, merely providing a binary classification (positive vs. negative), which does not account for finer characterization of emotion involved. On the contrary, a Multi-Level Fine-Scaled Sentiment Sensing with Ambivalence Handling is proposed for analyzing fine scale of both positive or negative sentiments [23]. Such fine-grained sentiment classification mostly relies of the weightage of word in the context.

Recently, the attention mechanism has shown promising performance in natural language processing (NLP) tasks, which improves deep neural network by letting them learn about where to focus. Recent studies on attention-based sentiment analysis are exampled by [24–26]. One of the applications of attention mechanism is the Attention-based Bidirectional CNN–RNN Deep Model (ABCDM) that extracts both past and future contexts by considering temporal information flow [27]. The attention mechanism in ABCDM is applied to the outputs of the bidirectional layers to shift the emphasis more or less on various words. On the other hand, some sentiment analysis studies focus not only on language modeling but also on common sense knowledge. For example, in [28], SenticNet 6 is proposed, which integrates top-down and bottom-up learning via an ensemble of symbolic and subsymbolic AI tools. However, these sentence-level sentiment analyses cannot be directly applied to the ABSA task where the sentiment of the sentence holds different opinions for distinct aspect words.

2.2. ABSA based on LSTM

ABSA tends to infer the polarity of a sentence's sentiment towards a particular aspect word. The sentiment may change throughout the sentence based on the context words. Hence, the main task is to model the relationship between the aspect word and the context words in an efficient manner. It is explained in [29] that around 40% classification error in this task is due to the ignorance of the aspect word. This significantly increased interests in the studies including early work on machine learning algorithms [30] that extracts a set of features to demonstrate the relationship between them.

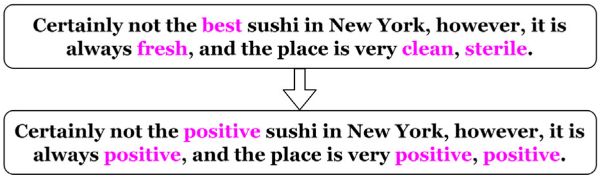
Neural networks, such as the LSTM network, can encode sentences without feature engineering, and have been implemented in many (NLP) tasks [31–33]. In [34], TD-LSTM is proposed, which consists of two dependent LSTMs to model the left and the right contexts divided by the aspect word, where the aspect word is also input into the model as word embedding. Similarly, Gated Neural Networks is designed to control the importance of left and right context [35]. However, these methods do not capture the relationship between the context and the aspect word because the divided sentence most probably contains only one aspect word. Since the introduction of attention network on translation task [36,37], many NLP tasks are interested to employ attention mechanism to model the relationship between the words in the sentence that seems very relevant to ABSA tasks. AE-LSTM adopts the attention mechanism to shift the focus of the aspect word towards relative context words [38]. Another work, similar to AE-LSTM, was proposed in [2] and it learns to interact between context words and aspect word based on associative relationships. In this way, the model can resiliently focus on the correct context words given the aspect word.

Although the attention mechanism has enhanced the efficiency of ABSA tasks, it simply processes the aspect word using the average pooling method while computing the attention score for the context. A typical drawback is that the performance suffers if the aspect consists of multiple words. To solve this problem, it is designed in [39] an Interactive Attention Network (IAN) to learn the attention representation for the context and the aspect word based on two different attention models in parallel and combining them eventually for sentiment classification. While IAN is an important work that considered context and aspect words's interactive learning, it still utilizes average vectors to calculate the attention score for both aspect word and context words. In [40], it is presented a hierarchical model of attention for the task of aspect-based sentiment analysis that included both attention at the aspect level and attention at the sentence level, where the attention used in the aspect-level is a self-attention that takes the output of hidden layers as the input. Moreover, several other studies use knowledge-based approaches to tackle the ABSA task. It is proposed in [41] methods of rule-based ontology that constructed ontologies to help improve the outcome of ABSA by using common domain information. Additionally, to guide the model to learn relevant rules so that the model can capture more useful features, it is pertinent to incorporate external information. Such rule based models are highly interpretable compared to the neural network [42]. Those knowledge-based approaches, however, are very dependent on the knowledge that they possess, which may be difficult to construct, and the knowledge rule may also be intricate to design effectively through neural networks.

2.3. Positional embedding based ABSA

To enhance the classification accuracy in the ABSA task, the position information of the given aspect word is integrated into the models [4,5,43]. These methods utilize position between the aspect word and the context words either by counting the number of words between them or using tree structure dependency as relevant information. With the concept that the context word closer to the aspect word would be more important, attention mechanisms are preferred in the memory-based models [44]. Similarly, it is employed in [4] the word distance between the aspect word and the context word to mitigate the disadvantage of memory network [45]. The performance is further improved by scaling the input representation of the convolutional layer with the positional relevance between the contexts and the aspect word, which helps CNN feature extractor easily locate the sentiment indicators more accurately [5]. Another position based ABSA task is represented in [6] where position-aware sentence representation is applied by concatenating position embedding and word embedding. Similarly, it is proposed in [46] a position-dependent method using position-aware attention and a deep bidirectional LSTM (DBi-LSTM).

Despite the promising performance enhancement using positional embedding in ABSA tasks, we observed that the model is usually very complex to obtain the best range of classification accuracy. Additionally, since the RNN models are good enough to capture the time series information, encoding extra dimension as a position embedding seems a complicated preprocessing scheme. Therefore, we propose a masking technique that replaces the aspect word with a common token in the aspect embedding so that it creates different order information for distinct aspect words. We incorporate this masked aspect embedding along with the original sentence embedding into attention based Bi-GRU to capture the context-dependent sentiment, which is to be detailed in the next section.


Certainly not the best sushi in New York, however, it is always fresh, and the place is very clean, sterile.

Certainly not the positive sushi in New York, however, it is always positive, and the place is very positive, positive.

Fig. 1. Replacement of sentiment carrying word with a common tag using Opinion Lexicon.

3. Proposed method

In this section, we describe in detail the proposed prepossessing and architecture for the ABSA task.

3.1. Preprocessing

As mentioned earlier, the sentiment of aspect word highly depends on the context words surrounding it. Human beings can understand the meaning and the sentiment of context words that describe the aspect word. That is why human being can easily extract the sentiment of any particular word. On the contrary, a neural network does not have the knowledge that shows the semantic and syntactic relationship between words. Word2vec and Glove embedding [47,48] capture the semantic relationship between the words but they are still far from human efficiency. Hence, we try to reduce the gap between the human knowledge and word embedding by making the semantically related word as the same token. To simplify the problem so that the neural networks can solve it better, we replace the sentiment-carrying words with the tag “positive” or “negative” based on Opinion Lexicon [11] as shown in Fig. 1. Opinion Lexicon is a list of English words with positive or negative sentiment. Such use of external resource in preprocessing not only integrates sentiment knowledge but also reduces the vocabulary size that is a substantial concern by itself in NLP [49]. Altogether this process replaces around 550 words with the token “positive” and “negative”.

Another important aspect of the preprocessing is to embed the position information of the aspect word. Traditional positional embedding considers the relative distance between aspect words and the context words in a sentence. Such embedding creates a probability distribution over the sentence with respect to the aspect word. However, such position embedding integrated with the input sentence is often initialized with trainable weights that increase the complexity of the model [50]. To mitigate this problem, we propose a simple masking technique that is based on the pattern learning behavior of the neural network. Usually, an ABSA task has two inputs: Sentence Embedding carrying the original sentence where position information is integrated and Aspect Embedding carrying aspect or aspect word. Here, we modify Aspect Embedding as Masked Aspect Embedding that carries the sentence with the aspect word masked by a common tag (here we call the common tag as “MASK”). We propose this preprocessing to remove the positional embedding required by Sentence Embedding. The modification between existing positional embedding and proposed masking technique is shown in Fig. 2, where $pos(w^1)$ is the relative positional encoding of the first word with respect to the aspect word and the total number of words in the sentence is n. We hypothesize that the masked token is a common token present in every sample at a different location, creates a positional pattern. Since any machine learning model tries to capture the repetitive patterns, we hypothesize that the model will pick the masked token and its necessary context words around it to classify the sentiment. In all brevity, we propose a model that learns sentiment patterns for the position of the masked token. The overall preprocessed input is shown in Fig. 3.

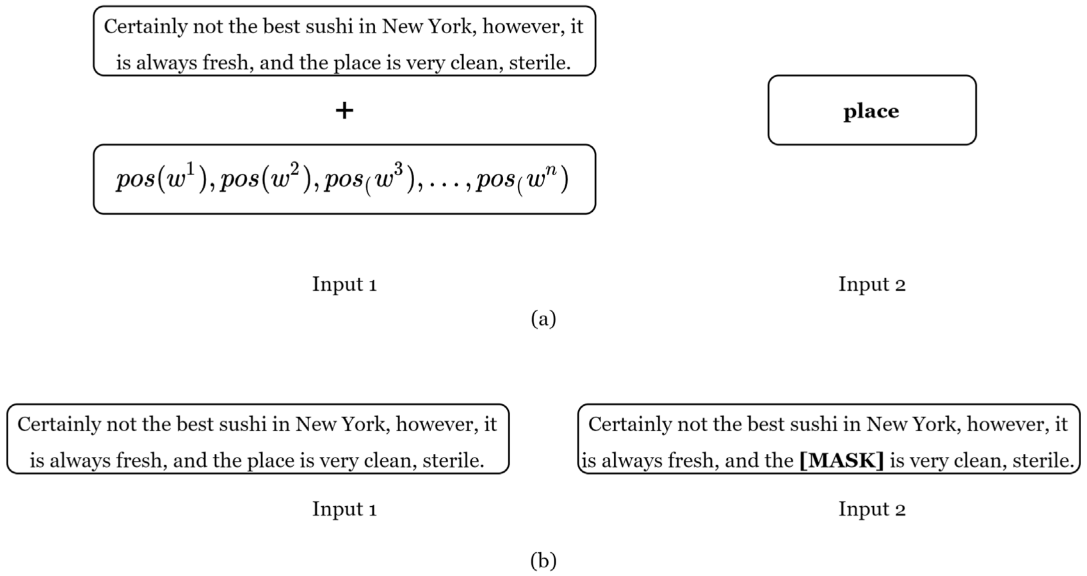

Certainly not the best sushi in New York, however, it is always fresh, and the place is very clean, sterile.
+
$pos(w^1), pos(w^2), pos_(w^3), \ldots, pos_(w^n)$
place
Input 1
Input 2
(a)
Certainly not the best sushi in New York, however, it is always fresh, and the place is very clean, sterile.
Certainly not the best sushi in New York, however, it is always fresh, and the [MASK] is very clean, sterile.
Input 1
Input 2
(b)

Fig. 2. (a) Existing approach of position embedding. (b) Proposed masking technique to learn pattern for the position.

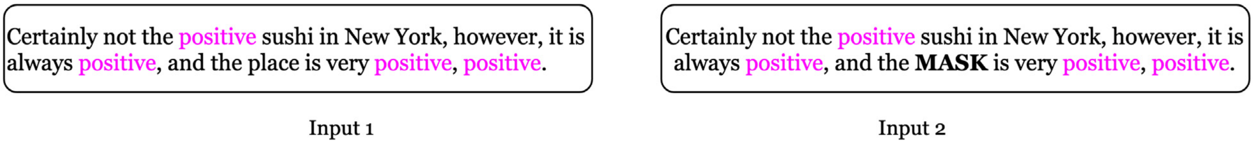

Certainly not the positive sushi in New York, however, it is always positive, and the place is very positive, positive.
Certainly not the positive sushi in New York, however, it is always positive, and the MASK is very positive, positive.
Input 1
Input 2

Fig. 3. Proposed preprocessed input.

3.2. Architecture description

The overall architecture of proposed model is shown in Fig. 4, which consists 3 sections: Input Embedding, Bi-GRU, and Attention Layer. As the input embedding has been explained in the preprocessing part, we will focus on the latter two in the following paragraphs.

3.2.1. Bidirectional Gated Recurrent Unit (Bi-GRU)

Recurrent neural network (RNNs) [51] have been the baseline for NLP recently, where the internal states are utilized to process data sequentially. However, RNNs have certain limitations that lead to the development of their variants, such as LSTM and GRU. Here, we have explored both LSTM and GRU for sequencing modeling. Since we aim at developing a very concise and efficient model, we opt for GRU in our final architecture. The GRU controls the flow of information like the LSTM unit without employing a memory unit, which makes it more efficient with uncompromised performance compared to LSTM [52]. In addition, GRU mitigates the problem of vanishing gradients and gradient explosions in vanilla RNN.

Our proposed model consists of two Attention-based Bi-GRUs: GRU_1 for Sentence Embedding and GRU_2 for Masked Aspect Embedding. Both of them are identical in architecture that has similar learning pattern with the same hyperparameters. The only difference is how the preprocessed input data is passed to these two separate Bi-GRUs. We assumed that GRU_2 captures the position of the masked token. Additionally, attention layer 2 gives the highest weightage to the masked token wherever it presents in the sentence. Similarly, GRU_1 is supposed to capture the context features from Sentence Embedding with attention layer 1, assigning higher weightage to the necessary context words. This hypothesis seems quite similar to how human being operates to understand the aspect-based sentiment.

Define $X = [x_1, x_2, x_3, \cdots, x_k]$ the Sentence Embedding (or Input 1), where k is the padded length of the sentence embedding to the forward layer of the GRU. There are two kinds of gates in GRU: the update gate and the reset gate. The update gate decides the amount of past information that needs to be brought into the current state and how much the new information is added. On the other hand, the reset gate takes care of how much information about the previous steps is written into the current candidate state h_t. Here, h_t is the output of the GRU at time step t and z_t represents the update gate. At a particular time step t, the new state h_t is given by:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t, \tag{1}$$

where $\odot$ is the element-wise multiplication and $\hat{h}_t$ is candidate activation. To update z_t, we have

$$z_t = \sigma\left(W_{z_t} x_t + U_{z_t} h_{t-1} + b_{z_t}\right). \tag{2}$$

Here, x_t is the word of the sentence at time step t that is plugged into the network unit and it is multiplied with its own weight W_{z_t}. Similarly, h_{t-1} holds the information of the previous unit and is multiplied with its own weight U_{z_t} and b_{z_t} is the bias associated with update state. The current state h_t can be updated using reset gate r_t by

$$\hat{h_t} = \tanh\left(W_h x_t + r_t \odot (U_h h_t) + b_h\right). \tag{3}$$

where W_h and U_h are weights associated with the candidate activation along with bais b_h.

At r_t, the candidate state of step t can get the information of input x_t and the status of h_{t-1} of step $t-1$. The update function of r_t is given by

$$r_t = \sigma\left(W_{r_t} x_t + U_{r_t} h_{t-1} + b_{r_t}\right), \tag{4}$$

where W_{r_t} and U_{r_t} are the weights associated with the reset state and b_{r_t} is the bias.

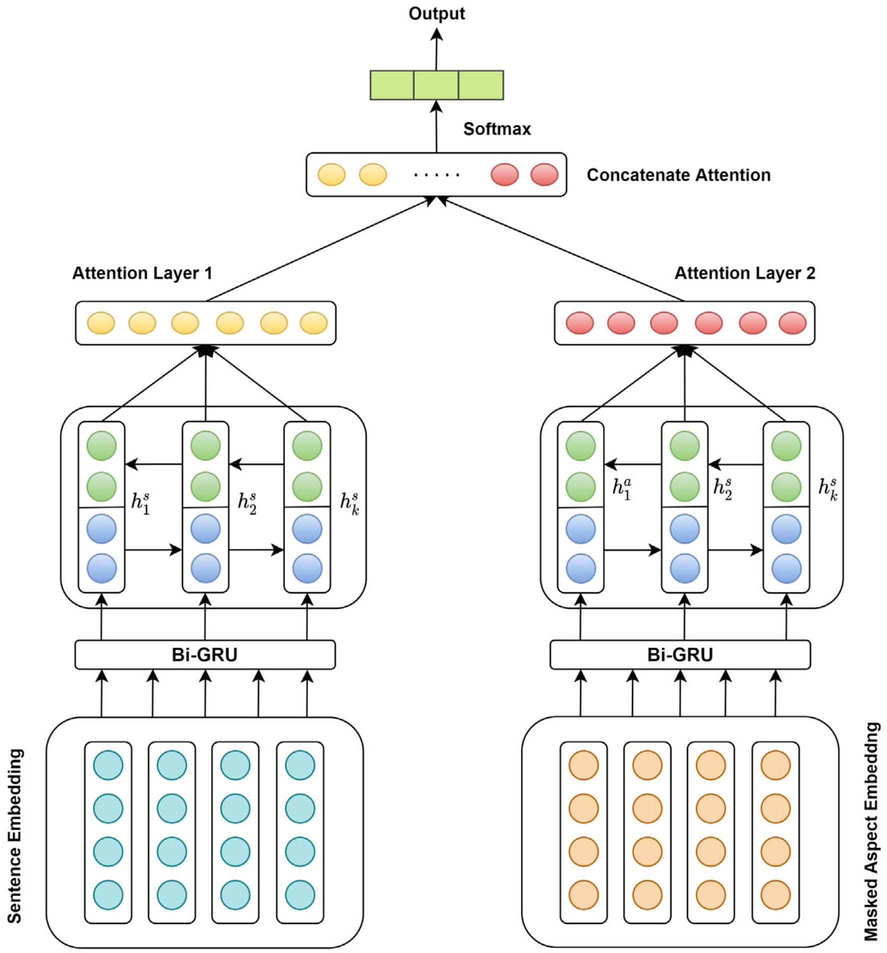


Fig. 4. Proposed attention based Bi-GRU architecture.

The Bi-GRU contains the forward GRU layer ($\overrightarrow{h_t}$) that reads the input sentence from step 0 to t and the backward GRU ($\overleftarrow{h_t}$).

$$\overrightarrow{h_t} = \overrightarrow{GRU}(x_t),\ \ t \in [1, T], \tag{5}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(x_t),\ \ t \in [T, 1], \tag{6}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]. \tag{7}$$

3.2.2. Attention layer

As we know that not all the words in the context have equal contribution for sentiment classification, an attention layer is assigned to prioritize important words in the context. Attention layer 1 is wrapped on top of GRU_1 to learn a weight α_t^1 for each hidden state h_t obtained at time step t. Since there are k inputs in the padded sequences, time step t will be from 1 to k. The weighting vector for attention layer 1, $\alpha_t^1 = [\alpha_1^1, \alpha_2^1, \alpha_3^1, \cdots, \alpha_k^1]$ is calculated based on the output sequence $H = [h_1, h_2, h_3, \cdots, h_k]$. The attention vector s_1 for attention layer 1 is calculated based on the weighted sum of these hidden states, as:

$$s_1 = \sum_{t=1}^{k} \left(\alpha_t^1 h_t\right), \tag{8}$$

where the weighted parameter α_t^1 is calculated by:

$$\alpha_t^1 = \frac{\exp\left(u_t^T u_w\right)}{\sum_t \exp\left(u_t^T u_w\right)}, \tag{9}$$

and $u_t = \tanh\left(W_w h_t + b_w\right)$. Here W_w and h_t are the weight matrices and b_w represents the bias. The parameter u_w represents context vector that is different at each step, which is randomly initialized and learned jointly during the training process.

Similarly, the attention layer 2 is wrapped on top of GRU_2 for assigning weightage to the masked token based on its position. The attention vector s_2 for attention layer 2 is given by:

$$s_2 = \sum_{t=1}^{k} \left(\alpha_t^2 h_t\right). \tag{10}$$

Finally, both of the attention layers are concatenated

$$s = Concatenate\left(s_1, s_2\right). \tag{11}$$

The concatenated layer is then sent to a fully connected layer and the softmax function generates a probability over c class labels.

4. Experiment results and evaluations

In this section, we will present the experiment results of the proposed scheme in detail. We conduct experiments on SemEval 2014 "restaurant" and "laptop", SemEval 2015 "restaurant" and SemEval 2016 "restaurant" dataset to verify the proposed hypothesis of lexicon addition and position-less masking. Additionally, we will also show the analysis of how the lexicon information and masking technique enhance the performance individually. We employ Keras [53] to implement our model. Adam [54] is adopted as the models' optimization method with the learning rate of $1 \times e^{-3}$. We also utilize Dropout [55] as the regularization

Table 1
Details of ABSA datasets.

Dataset	Train			Test		
	Pos	Neu	Neg	Pos	Neu	Neg
Rest 14	2164	637	807	728	196	196
Lap 14	994	464	870	341	169	128
Rest 15	948	34	269	432	38	257
Rest 16	1289	63	457	474	29	123

strategy and the probability of Dropout is kept 0.6. Words are initialized with Glove [48] of 300-dimension word embedding. The batch size is 128 and is run for 100 epochs in the test datasets for obtaining the best results.

4.1. Datasets

Our experiments are conducted on four publicly available ABSA datasets. Each sample of every dataset is a single sentence of a product review with aspect word and the corresponding sentiment label associated. While the datasets are given in the laptop domain by SemEval 2015 and SemEval 2016, they only contain the aspect category without the aspect word. The "null" aspect terms are excluded from the datasets, and the "dispute" or more than one sentiment labels are also excluded from the aspect terms in the analysis. The remaining sentences contain at least one aspect of the word with a {positive, neutral, negative} sentiment tag. The numerical details of the datasets are shown in Table 1.

4.2. Compared methods

To evaluate the performance of the proposed model on ABSA datasets, we consider the following approaches as comparative models. These models are proper baselines for ABSA and they are as close as possible to this work. Additionally, we have used the models that have exactly been evaluated in these specific four datasets of ABSA.

- **Feature+SVM** extracts n-gram as a feature, parse feature, and lexicon features to train the classifier [30].
- **ContextAvg** averages the word embedding to form a context embedding and then it is feed to the softmax function along with aspect vector [45].
- **LSTM** uses the last hidden vector information of the LSTM as a sentence representation for classifying aspect level sentiment [15].
- **TD_LSTM** utilizes two LSTMs to learn the language model from left and right contexts of the aspect respectively [45].
- **ATAE_BiLSTM** This model is similar to our approach, which is an Attention-based LSTM architecture with Aspect Embedding. It computes the aspect-specific weighted score of each word according to the representation of the aspect. The sums of the LSTM hidden outputs based on the attention weights are utilized to generate the sentence representation for ABSA classification [38].
- **IAN** is an Interactive Attention Network model that calculates the attention weights of the word in sentiment and aspect interactively to generate aspect and sentence representations [39].
- **MemNet** integrates the content and the position of the aspect word into deep neural network [45].
- **RAM** is a multi-layer architecture where each layer consists of attention based aggregation of word features. A GRU cell is used to learn the sentence representation [4].
- **Ont+LCR-Rot-hop** uses a lexicon domain ontology and a rotatory attention mechanism to predict the sentiment of the aspect word [56].
- **PBAN** is a position-aware bidirectional attention network on bidirectional GRU. It also uses the mean pool and dot product to embed the position information of aspect word into sentence representation. It performs on par with the state of the art [6].
- **PAHT** is a position-aware hierarchical transfer model that models the position information from multiple levels to enhance the ABSA performance by transferring hierarchical knowledge from the resource-rich sentence-level sentiment classification (SSC) dataset [46].
- **MTKFN** is a Multi-source Textual Knowledge Fusing Network that incorporates knowledge from multiple sources to enhance the performance of ABSA. It uses pre-trained layers to extract contextual features and predicts the sentiment polarities. Additionally, it uses the information of conjunctions that captures the relationship between clauses and provides additional sentiment features [57].
- **BERTADA-base** is further trained on a domain-specific dataset and evaluated on the test set from the same domain [58].
- **XLNetADA-base** model is like BERTADA-base except for adopting XLNet [58].

4.3. Hardware configuration

The experiments are conducted on a Linux platform. The OS and GPU configuration of our server is NVIDIA DGX Server Version 4.6.0 (GNU/Linux 4.15.0-121-generic x86_64). We have used NVIDIA Tesla V100 SXM3 32 GB to train our model.

4.4. Performance comparison and analysis

The comparison of our proposed model with recent similar studies is shown in Table 2. These state-of-the-art studies are selected for comparison because they mostly depend on positional embedding as well as complex architecture, which are more relevant to our proposed model.

Among the baseline models that depend on language modeling, LSTM performs poorly on all four datasets. Above this lies ATAE_BiLSTM that has poor performance considering the fact that it utilizes the attention mechanism to model the aspect word. However, TD_LSTM performs better than the two models mentioned above because it considers both the left and right context as an aspect rather than the entire sentence. Similarly, MemNet performs better than IAN but it is not as good as RAM, because it does not use multiple attention mechanisms. Moreover, PAHT and MTKFN that utilize the hierarchical transfer model and external knowledge fusion respectively exhibit quite similar results as both of them are position-aware models. However, our proposed model surpasses both of these recent models with a significant margin using a much simpler architecture.

Among PBAN, PAHT and MTKFN, PBAN has higher accuracy than PAHT and MTKFN that is quite similar model to our proposed architecture. Hence, we focus on PBAN more than other listed models for comparison. PBAN that adopts two BiGRUs still falls behind in performance compared with our model. As shown in Table 2, PBAN achieves 81.16% accuracy on restaurant 14 and 74.12% on laptop 14 dataset. However, it uses traditional embedding with a more complex model than ours. On the other hand, by employing the Opinion lexicon and masked aspect embedding, our model gives 81.37% and 75.39% accuracy on restaurant 14 and laptop 14 datasets respectively. Additionally, the macro-F1 score also outperforms the above-mentioned models with a significant

Table 2
The state-of-the-art performance of ABSA on four datasets.

Dataset	Restaurant 14		Laptop 14		Restaurant 15		Restaurant 16	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Majority	65.00	26.26	53.45	23.22	54.74	23.58	72.36	29.99
Feature+SVM	80.16	–	70.49	–	–	–	–	–
ContextAvg	71.53	58.02	61.59	53.92	73.79	47.43	79.87	55.68
LSTM	74.49	59.32	66.51	59.44	75.40	53.30	80.67	54.53
TD_LSTM	78.00	68.43	71.83	68.43	76.39	58.70	82.16	54.21
ATAE_BiLSTM	77.63	64.97	69.61	63.04	77.40	54.29	86.01	60.32
IAN	77.35	64.77	69.58	61.08	78.07	51.89	85.44	56.51
MemNet	78.16	65.83	70.33	64.09	77.89	59.52	83.04	57.91
RAM	78.48	68.54	72.08	68.43	79.98	60.57	83.88	62.14
Ont+LCR-Rot-hop	–	–	–	–	80.60	–	88.00	–
PBAN	81.16	–	74.12	–	–	–	–	–
PAHT	79.29	68.49	75.71	69.55	80.86	60.76	85.81	67.11
MTKFN	79.47	68.08	73.43	69.12	80.67	58.38	88.28	66.15
BERTADA-base	84.92	76.93	77.69	72.60	–	–	–	–
XLNetADA-base	85.84	78.35	79.89	77.78	–	–	–	–
Proposed model	81.37	72.06	75.39	70.50	80.88	62.48	89.30	66.93

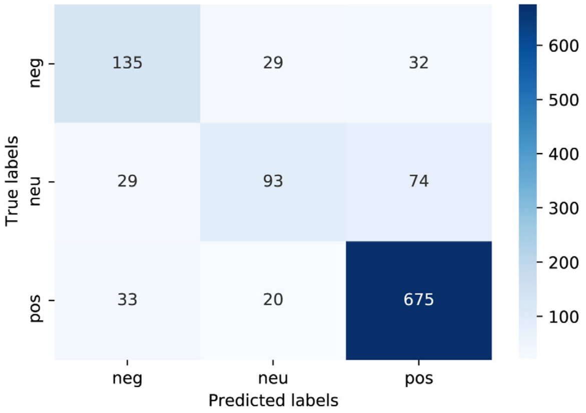


Fig. 5. Confusion matrix of restaurant 14 dataset.

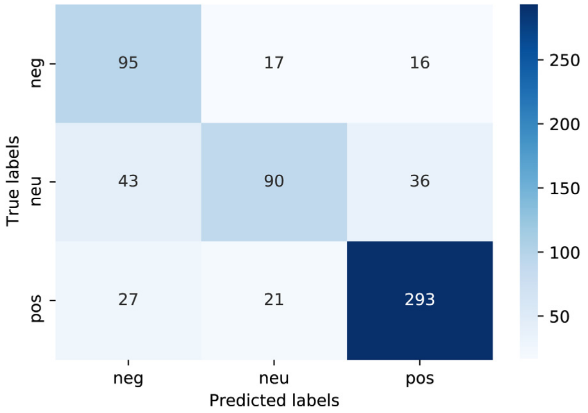


Fig. 6. Confusion matrix of laptop 14 dataset.

margin except for the restaurant 16 dataset where the macro-F1 score is slightly below PAHT.

Even though the main aim of this paper is to design a simple yet effective model, to make a comprehensive comparison, we would present the performance of some new transformer-based models, such as BERTADA-base and XLNetADA-base approaches. It is quite obvious that this sophisticated pre-trained contextualized embedding achieves the upper state-of-the-arts results by using a softmax classifier. Since our model does not apply position embedding thereby reducing the trainable parameters, our comparison is mostly focused on the position-dependent model as explained before.

4.5. *Error and sensitivity analysis*

Error analysis studies the impact of inaccuracy in the model for meaningful insight. In this study, we demonstrate error analysis using the confusion matrix. The confusion matrices show the relationship between the true and the predicted class as shown in Figs. 5, 6, 7, and 8. It can be seen from the confusion matrix that our model slightly suffers to identify the neutral sentiment. We believe that this may be because our model gives more focus to sentiment carrying lexicon tokens such as "positive" and "negative".

In addition to the accuracy of the model, we also explore the sensitivity of the model with respect to input representations, detailed in the subsections below.

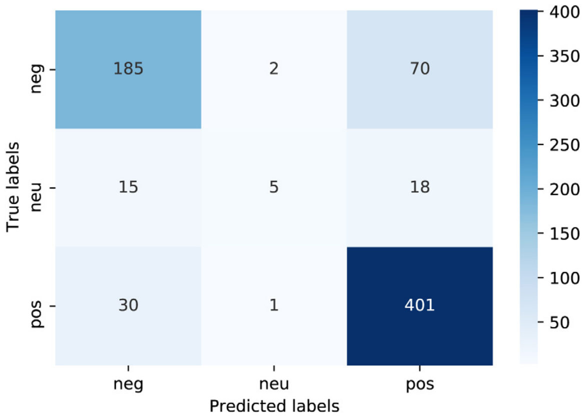


Fig. 7. Confusion matrix of restaurant 15 dataset.

4.5.1. *Effect of input representations*

We here demonstrate the sensitivity of the model by changing the dimension of Glove input representation as shown in Table 4. One can observe that the dimensions of the glove vectors have a significant impact on the performance of the model. However, the dimension of 200d and 300d does not have much difference except for restaurant 16 dataset. Hence, higher dimension representation has more semantic impact that boosts the accuracy of the model.

Table 3
Effect of the proposed preprocessing on all four datasets.

Dataset	Restaurant 14		Laptop 14		Restaurant 15		Restaurant 16	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Lexicon replacement	79.55	67.31	71.32	64.80	78.82	57.86	87.38	63.66
Masked aspect embedding	80.62	70.72	73.51	68.01	80.33	61.26	88.65	65.00

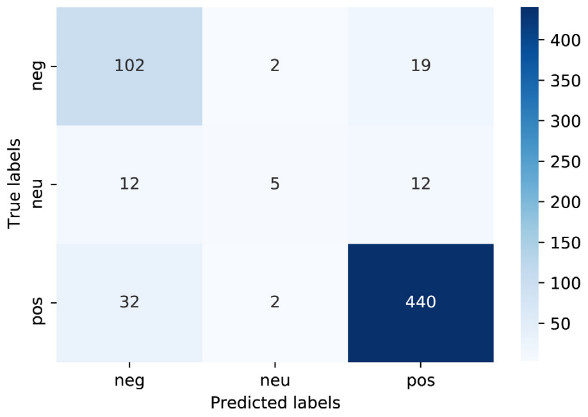


Fig. 8. Confusion matrix of restaurant 16 dataset.

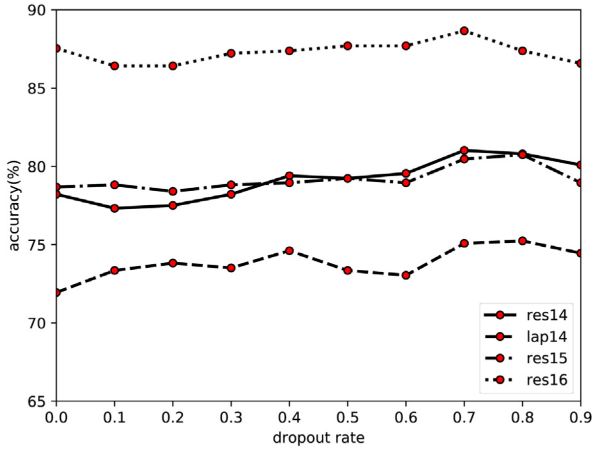


Fig. 9. Effect of dropout rate.

Table 4
Effect of various Glove vector for word representation on accuracy (%).

Glove	res14	lap14	res15	res16
glove.6B.50d	79.29	72.73	78.13	85.46
glove.6B.100d	80.18	72.29	78.82	86.42
glove.6B.200d	78.93	74.92	79.92	86.42
glove.6B.300d	79.73	73.67	80.33	86.26
glove.42B.300d	81.37	75.39	80.88	89.30
glove.840B.300d	80.00	74.14	79.09	87.54

4.5.2. Effect of dropout rate

We here explore the effect of the dropout rate that is applied to the input of each layer of BiGRU. For simplicity, we employ the same dropout for both BiGRU varying from 0.0 to 0.9. The results of various dropout rates on the accuracy are shown in Fig. 9. As we can see, the effect of dropout does not affect accuracy significantly. The change in accuracy is less than 3% in all 4 datasets. However, the best result always occurs at the dropout of range 0.6 to 0.8.

Table 5
Comparison of binary classification on ABSA datasets.

Model	Restaurant 14	Laptop 14
PBAN	91.67	87.81
Proposed model	**92.51**	**90.15**

4.5.3. Effect of opinion lexicon and masked aspect embedding

To verify the efficiency of proposed preprocessing, we further evaluate the performance of lexicon replacement and masking aspect words individually. The two main preprocessing used in our paper are :

- Lexicon replacement using Opinion Lexicon where Input 1 is Sentence embedding and Input 2 is aspect word embedding. (without masking aspect word)
- Masking aspect word where Input 1 is Sentence embedding and Input 2 is masked aspect embedding. (without lexicon replacement)

The performance of the scheme under these two conditions is evaluated for all four datasets and shown in Table 3. As one can see, the effect of masking aspect word is significantly higher than the lexicon replacement. This is because the masking technique carries the position information into the model. On the contrary, the lexicon information helps to generalize the word and reduce the vocabulary size hence it has lower influence on the accuracy compared with the masking technique. Therefore, the lexicon replacement just boosts the final accuracy by a small margin.

4.6. Two-class sentiment classification

Some of the positional embedding based studies on ABSA also focus on the performance in the binary classification of ABSA task, in which it neglects the neutral class and makes it a binary classification as to predict positive or negative sentiment. Hence, we evaluate here the performance of our proposed model in binary classification and compare it with a position embedding based model [6]. The comparison is shown in Table 5. We can see that our proposed method significantly outperforms the traditional positional embedding technique on both restaurant 14 and laptop 14 datasets.

4.7. Case studies

To have a detailed insight into why our proposed model performs better than the baselines with a straightforward architecture, we sample two examples from the restaurant 16 dataset and visualize attention heatmaps based on the trained model. Here we have two inputs for two separate BiGRUs: one being the sentence itself and the other being the masked sentence. As we have already discussed, the aspect words are masked with a common token, say "**MASK**". From Fig. 10, we can see that the sentiment of the aspect word "food" is classified as positive sentiment from our model. The original sentence is fed to GRU_1 that has an attention layer 1 and the Masked sentence is fed to GRU_2 that has an attention layer 2. Here, attention layer 1 captures the context words throughout the sentence whereas attention layer 2 pays high attention to the masked token with

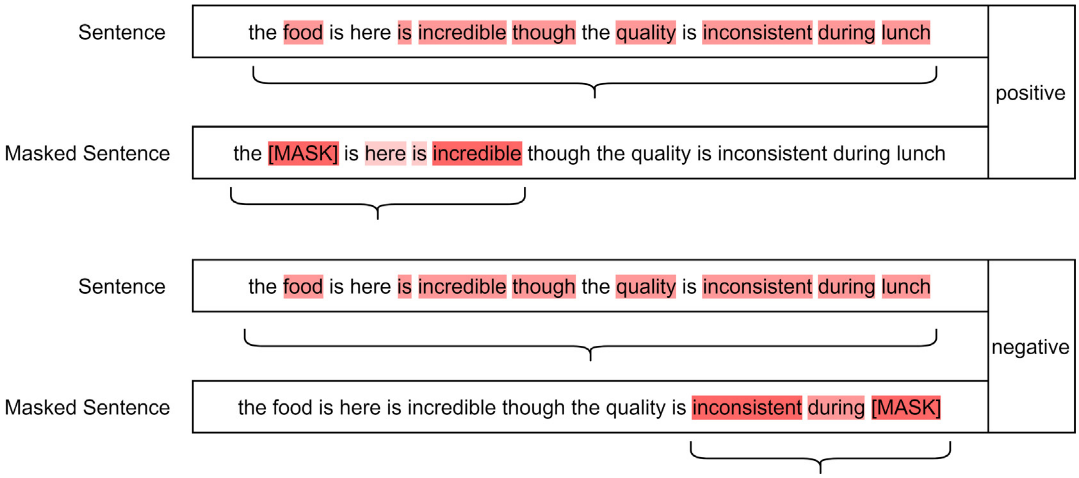

Sentence
the food is here is incredible though the quality is inconsistent during lunch
Masked Sentence
the [MASK] is here is incredible though the quality is inconsistent during lunch
positive
Sentence
the food is here is incredible though the quality is inconsistent during lunch
Masked Sentence
the food is here is incredible though the quality is inconsistent during [MASK]
negative

Fig. 10. Visualization of two typical examples. The red color represents the attentive weight of the word. A deeper color indicates a larger weight value.

weight narrowing down to important context words as shown in Fig. 10. In the first example, attention layer 2 shifts the attention weightage towards the masked token "food" (the first half of the sentence) that highly depends on the context "incredible" for positive sentiment. For the second example, attention layer 2 shifts the attention weightage towards the masked token "lunch" (the second half of the sentence) whose sentiment depends on context "inconsistent" for negative sentiment. This is a clear validation of our hypothesis that the masked token will hold the position information without using any additional trainable positional embedding. In all brevity, attention layer 1 assigns weightage to words in the sentence, and attention layer 2 narrows down the weightage from these selected words to important context words, carrying sentiment of the aspect word.

5. Conclusion

In this paper, we propose an efficient preprocessing scheme with an attention-based GRU model for aspect-based sentiment analysis. We first explore sentiment knowledge called Opinion Lexicon that is a list of positive, neutral, and negative sentiment words. In more detail, we replaced the words in ABSA dataset with a common tag, such as "positive" for positive sentiment words. This external input helps to bridge the gap from semantically related words to a certain extent and reduces the task's vocabulary. Since the ABSA is a position-dependent task, it requires the information of position along with the sentence embedding or aspect embedding. The extra trainable weights for position information increase the complexity of the model. To reduce the complexity, we proposed a masking technique that masks the aspect word in the sentence with a common token "MASK". This masked embedding is separately passed to the model along with the sentence embedding. Experimentally, we have shown that the proposed scheme performs at par with the state of the art and it outperforms several position-aware methods with very straightforward attention-based BiGRUs architecture.

CRediT authorship contribution statement

Rohan Kumar Yadav: Conceptualization, Methodology, Implementation, Writing. **Lei Jiao:** Supervision, Writing - review & editing. **Morten Goodwin:** Supervision, Writing - review & editing. **Ole-Christoffer Granmo:** Supervision, 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.

References

[1] J. Zhao, K. Liu, L. Xu, Sentiment analysis: Mining opinions, sentiments, and emotions, Comput. Linguist. 42 (2016) 595–598.
[2] Y. Tay, L.A. Tuan, S.C. Hui, Learning to Attend via Word-Aspect Associative Fusion for Aspect-Based Sentiment Analysis, AAAI, New Orleans, Louisiana, USA, 2018.
[3] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2016) 813–830.
[4] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: EMNLP, ACL, Copenhagen, Denmark, 2017, pp. 452–461.
[5] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, in: ACL, ACL, Melbourne, Australia, 2018, pp. 946–956.
[6] S. Gu, L. Zhang, Y. Hou, Y. Song, A position-aware bidirectional attention network for aspect-level sentiment analysis, in: COLING, Santa Fe, New Mexico, USA, 2018, pp. 774–784.
[7] B. Xu, X. Wang, B. Yang, Z. Kang, Target embedding and position attention with LSTM for aspect based sentiment analysis, in: International Conference on Mathematics and Artificial Intelligence, in: ICMAI, ACM, New York, NY, USA, 2020, pp. 93–97.
[8] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: COLING, Dublin, Ireland, 2014, pp. 2335–2344.
[9] Y. Song, J. Wang, T. Jiang, Z. Liu, Y. Rao, Attentional encoder network for targeted sentiment classification, 2019, ArXiv abs/1902.09314.
[10] K. Xu, H. Zhao, T. Liu, Aspect-specific heterogeneous graph convolutional network for aspect-based sentiment classification, IEEE Access 8 (2020) 139346–139355.
[11] M. Hu, B. Liu, Mining and summarizing customer reviews, in: ACM SIGKDD, New York, NY, United States, 2004, pp. 168–177.
[12] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, .SemEval Dublin, Ireland, 2014, pp. 27–35.
[13] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, SemEval-2015 task 12: Aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation, .SemEval, Denver, Colorado, USA, 2015, pp. 486–495.
[14] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, G. Eryiğit, SemEval-2016 task 5: Aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation, .SemEval, San Diego, California, USA, 2016, pp. 19–30.
[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.

[16] J. Zhou, J. Huang, Q. Hu, L. He, Is position important? deep multi-task learning for aspect-based sentiment analysis, Appl. Intell. 50 (2020) 3367–3378.

[17] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.

[18] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowl.-Based Syst. 89 (2015) 14–46.

[19] A. Tripathy, A. Anand, S.K. Rath, Document-level sentiment classification using hybrid machine learning approach, Knowl. Inf. Syst. 53 (2017) 805–831.

[20] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: NIPS, Vol. 28, Curran Associates, Inc., Montréal CANADA, 2015, pp. 649–657.

[21] Q. Qian, M. Huang, J. Lei, X. Zhu, Linguistically regularized LSTM for sentiment classification, in: ACL, ACL, Vancouver, Canada, 2017, pp. 1679–1689.

[22] W. Li, W. Shao, S. Ji, E. Cambria, BIERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, 2020, arXiv:2006.00492.

[23] Z. Wang, S.-B. Ho, E. Cambria, Multi-level fine-scaled sentiment sensing with ambivalence handling, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 28 (2020) 683–697.

[24] Y. Lou, Y. Zhang, F. Li, T. Qian, D. Ji, Emoji-based sentiment analysis using attention networks, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 19 (5) (2020).

[25] M. Usama, B. Ahmad, E. Song, M. Hossain, M. Alrashoud, M. Ghulam, Attention-based sentiment analysis using convolutional and recurrent neural network, Future Gener. Comput. Syst. 113 (2020) 571–578.

[26] C. Xi, G. Lu, J. Yan, Multimodal sentiment analysis based on multi-head attention mechanism, in: International Conference on Machine Learning and Soft Computing, New York, NY, United States, 2020, pp. 34–39.

[27] 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.

[28] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis, ACM, New York, NY, USA, 2020, pp. 105–114.

[29] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent Twitter sentiment classification, in: ACL, ACL, Portland, Oregon, USA, 2011, pp. 151–160.

[30] 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), ACL, Dublin, Ireland, 2014, pp. 437–442.

[31] M. Sundermeyer, R. Schlüter, H. Ney, LSTM neural networks for language modeling, in: INTERSPEECH, Portland, OR, USA, 2012, pp. 194–197.

[32] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: NIPS, MIT Press, Cambridge, MA, USA, 2014, pp. 3104–3112.

[33] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: EMNLP, Lisbon, Portugal, 2015, pp. 1422–1432.

[34] D. Tang, W. Qin, X. Feng, T. Liu, Effective LSTMs for target-dependent sentiment classification, in: COLING, Osaka, Japan, 2016, pp. 3298–3307.

[35] M. Zhang, Y. Zhang, D.-T. Vo, Gated Neural Networks for Targeted Sentiment Analysis, AAAI, Phoenix, Arizona USA, 2016.

[36] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2016, arXiv:1409.0473.

[37] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, 2015, ArXiv abs/1508.04025.

[38] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: EMNLP, Austin, Texas, USA, 2016, pp. 606–615.

[39] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, in: IJCAI, Melbourne, Australia, 2017, pp. 4068–4074.

[40] Y. Ma, H. Peng, E. Cambria, Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM, AAAI, New Orleans, Louisiana, USA, 2018, pp. 5876–5883.

[41] K. Schouten, F. Frasincar, F. de Jong, Ontology-enhanced aspect-based sentiment analysis, in: J. Cabot, R. De Virgilio, R. Torlone (Eds.), Web Engineering, Springer International Publishing, 2017, pp. 302–320.

[42] R.K. Yadav, L. Jiao, O.-C. Granmo, M. Goodwin, Human-Level Interpretable Learning for Aspect-Based Sentiment Analysis, AAAI, Vancouver, Canada, 2021.

[43] X. Li, W. Lam, Deep multi-task learning for aspect term extraction with memory interaction, in: EMNLP, AACL, Copenhagen, Denmark, 2017, pp. 2886–2892.

[44] S. Sukhbaatar, a. szlam, J. Weston, R. Fergus, End-to-end memory networks, in: NIPS, 28, Curran Associates, Inc., Montréal CANADA, 2015, pp. 2440–2448.

[45] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, in: EMNLP, Austin, Texas, USA, 2016, pp. 214–224.

[46] J. Zhou, Q. Chen, X. Huang, Q. Hu, L. He, Position-aware hierarchical transfer model for aspect-level sentiment classification, Inform. Sci. 513 (2020) 1–16.

[47] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, Vol. 26, Curran Associates, Inc., Nevada, USA, 2013, pp. 3111–3119.

[48] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: MNLP, Doha, Qatar, 2014, pp. 1532–1543.

[49] W. Chen, Y. Su, Y. Shen, Z. Chen, X. Yan, W.Y. Wang, How large a vocabulary does text classification need? A variational approach to vocabulary selection, in: NAACL, ACL, Minneapolis, MN, USA, 2019, pp. 3487–3497.

[50] C.W. Wu, Prodsumnet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions, 2019, arXiv abs/1809.02209.

[51] T. Mikolov, M. Karafi, S. Khudanpur, Recurrent neural network based language model, in: INTERSPEECH, Makuhari, Chiba, Japan, 2010.

[52] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: Workshop on Deep Learning@NIPS, Montréal, Canada, 2014.

[53] F. Chollet, et al., Keras, 2015.

[54] D.P. Kingma, J. Ba, ADAM: A method for stochastic optimization, in: ICLR 2015, San Diego, CA, USA, Conference Track Proceedings, 2015.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958.

[56] O. Wallaart, F. Frasincar, A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models, in: ESWC, Portoroz, Slovenia, 2019.

[57] S. Wu, Y. Xu, F. Wu, Z. Yuan, Y. Huang, X. Li, Aspect-based sentiment analysis via fusing multiple sources of textual knowledge, Knowl.-Based Syst. 183 (2019) 104868.

[58] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification, 2020, ArXiv abs/1908.11860.