RESIDUAL ATTENTION BI-DIRECTIONAL LONG SHORT-TERM MEMORY FOR VIETNAMESE SENTIMENT CLASSIFICATION

Nguyen Hoang Quan[1], *Vu Ly*[1], *Nguyen Quang Uy*[1]

Abstract

Sentiment classification is a problem of assessing and estimating values of people's opinions, sentiments, and attitudes to products, services, individuals, and organizations. Sentiment analysis helps companies understand their customers for improving marketing strategies in e-commerce, manufacturers decide how to improve their products, or people adjust behavior in their lives.

In this paper, we propose a deep network model to classify the reviewed product in Vietnamese. Specifically, we develop a new deep learning model called the Residual Attention Bidirectional-Long Short Term Memory (ReAt-Bi-LSTM) model. First, the residual technique is used in multiple layers Bidirectional Long Short Term Memory (Bi-LSTM), to enhance the model's capability in learning high-level features from input documents. Second, the attention mechanism is integrated after the last Bi-LSTM layer to assess each word's contribution to the context vector to the document's label. Last, the document's final representation is the combination of the context vector and output of Bi-LSTM. This representation captures both context information from the context vector and sequence information from the Bi-LSTM network. We conducted extensive experiments on four common Vietnamese sentiment datasets. The results show that our proposed model improves the accuracy compared with some baseline methods and one state of the art model for the sentiment classification problem.

Index terms

Sentiment classification, Neural network, Bi-LSTM, Attention, Residual.

1. Introduction

THE sentiment classification aims to identify and categorize opinions expressed in a text to determine the polarity of attitude (positive, negative, or neutral) of the customers towards companies' topics or products. Understanding the customer's personal opinion is of paramount importance in marketing strategies since this information helps to improve the products and services of businesses [1]. Moreover, obtaining product reviews of customers allows the companies to identify the new opportunities, predict sales trends, or manage reputation. However, reading a large number of comments is a

[1]Le Quy Don Technical University

tedious task [2]. Therefore, automatically sentiment classification is a vital task due to its ability to analyze massive comments or reviews.

Recently, deep neural networks have received great attention for sentiment classification [3]. An important step for applying deep learning models for sentiment classification is data representation. Many deep learning techniques use word embedding (WE) vectors as input features. We aims to transforming words in the document to a continuous vector [4]. Subsequently, deep neural networks, such as Convolutional Neural Networks (CNN) [5], Recurrent Neural Networks (RNN) [6], Long Short Term Memory (LSTM) [7] are used to learn from WE vectors to classify sentiment. Among these techniques, LSTM-based techniques are more commonly used for sentiment classification due to their ability to keep the long sequence dependency of the words in documents [8]. Additionally, Bidirectional-Long Short Term Memory (Bi-LSTM) [9], an extension of LSTM, has been proved to be a useful technique for sentiment classification. An appealing property of Bi-LSTM is that this model can represent a long dependency of words in both directions. However, in sentiment analysis, the performance of Bi-LSTM is not always satisfactory since the dependence of the sentiment to the words in sentences is varied significantly.

Moreover, there has not been any high quality and widely accepted pre-trained word embedding model published for the community in the Vietnamese language. Therefore, the WE vector in Vietnamese is not always a good representation of the semantics of words. To learn a good representation of a document from an embedding layer, researchers tend to add more stacked layers to the Bi-LSTM network. The drawback of this approach is that adding more neural network layers can lead to a degradation problem [10], and the network model is more challenging to train due to the accumulation of errors in multiple layers and the vanishing gradient [11].

In this paper, we propose the Residual Attention-based Bi-LSTM architecture (ReAt-Bi-LSTM) for Vietnamese sentiment classification. The main idea in ReAt-Bi-LSTM includes two folds. First, the residual technique is used in multiple layers of the Bi-LSTM network to handle the degradation problem in the training process. Second, the attention layers are added to the last layer of the Bi-LSTM model to exploit the core components that have a decisive influence on the sentiment of the document. We concatenate the output of the Bi-LSTM with the attention vector to create the final representation. By doing this, ReAt-Bi-LSTM is able to capture both context information from the context vector and sequence information learned by the Bi-LSTM. The main contributions of the paper are as follows:

- We propose a deep network model, i.e., ReAt-Bi-LSTM, to enhance the accuracy of the sentiment classification problem.
- We conduct an intensive experiment to evaluate the proposed model on four Vietnamese sentiment datasets.

The rest of the paper is organized as follows. Section 2 briefly reviews the previous works in applying deep learning to sentiment classification. Section 3 presents

the fundamental background of our paper. The proposed method is then described in Section 4. The tested datasets and experimental settings are provided in Section 5. Section 6 presents experimental results and the analysis. The conclusions and future work are discussed in Section 7.

2. Related Work

The CNNs and RNNs are the main techniques to represent words, sentences, and documents in sentiment classification [12], [4], [13], [14], [15], [16]. Among the two methods, RNN-based models are more widely used in the sentiment classification problem due to the ability to capture the semantic relation in documents [6].

Two extensions of RNN, including LSTM [7] and Bi-LSTM, are well-known for representing the long dependency of the words in sentences. Moreover, the Bi-LSTM network can provide information in both directions forward and backward at every sequence step [17]. Therefore, this model is usually used to learn the representation of documents in the sentiment classification problem.

An attention mechanism, which is a way to add inter-pretability in a neural network [18], has received much attention from the machine learning research community to improve the accuracy of neural networks for sentiment classification. The attention mechanism has achieved remarkable success in many fields, such as machine translation, text summary, image captioning. Chen et al. [19] used word and sentence level attention to classify product and user information in a document in which an attention vector represents both word and sentence. Yang et al. [20] proposed a hierarchical attention network to predict reviews. They also combine the word attention level and the sentence attention level to construct a document's representation. Zhou et al. [21] applied an attention-based LSTM network for cross-lingual sentiment classification. Their model includes two attention-based LSTM networks, and each network is trained on one language. This method's downside is that it requires a dictionary, which is expensive to build, to map the sentiments from one language to the other. Lin et al. [22] represented attention weights by a matrix in which each sentence is represented by a matrix, and each word in the matrix represents an aspect of the attention weights. They also added penalization terms to the loss function to improve the effectiveness of the attention weights.

Recently, Bi-LSTM has been widely used to improve the accuracy of sentiment classification. Li et al. [23] proposed a self-attention mechanism based on Bi-LSTM for the sentiment analysis task. They combined the word vector and the part-of-speech vector to form the feature channel inputs and then used a Bi-LSTM to learn each channel vector. After that, the self-attention model is used to discover important information from the output of the Bi-LSTM. Guan et al. [24] proposed an attention-based Bi-LSTM model for sentiment analysis. This model's attention mechanism directly learns the weight distribution of each word. Bahdanau et al. [25] proposed an attention mechanism for the neural machine translation problem. They added an attention layer at the end of

the Bi-LSTM layer to parameterize attention weights by a simple feed-forward neural network.

Another approach to improve the effectiveness of deep neural networks in sentiment classification is adding more layers to increase the capability of learning higher-level features. Moreover, the residual learning technique is also used in these multiple layers models to alleviate the degradation problem. Wang et al. [11] introduced a residual connection to every LSTM layer to construct an 8-layer neural network. The residual connection helps the deeper network to be easier to optimize. Yang et al. [26] used the residual connection in a deep RNN for sentiment classification. The residual technique improves the training process of the deep RNN and hence improves its accuracy in sentiment classification.

Our proposed model in this paper differs from the previous models [18] in several ways. First, we employ the residual connection into multiple layers Bi-LSTM to improve its ability to learn the document's high-level features. Second, the attention mechanism is integrated after the last layer of the Bi-LSTM to evaluate the contribution of different words in the document to its label. Third, we combine the context vector and the output of Bi-LSTM layers to represent the document that covers both context information and sequence information.

3. Background

This section presents the fundamental techniques used in the paper. These techniques include the Bi-LSTM model, the attention technique, and the residual technique.

3.1. Bi-LSTM model

LSTM was first introduced by Hochreiter et al. [7]. The heart of the LSTM network is its cell. The cell state provides a bit of memory to the model to remember the past. LSTM has three gates, i.e., input gate, forget gate, and output gate. The gate in LSTM is a simple sigmoid function with the output range from 0 to 1.

Let i_t, f_t, and o_t be the input gate, forget gate, and the output gate, respectively, and w_i, b_i, w_f, b_f, and w_t, b_t be the weight metrics and biases of neurons in the input gate, forget gate, and the output gate, respectively; Let σ be the *Sigmoid* function, h_{t-1} be the output of the previous LSTM block in the time sequence $t-1$ and x_t be the input at the current time sequence t, then the equations for the gates in the LSTM cell are as follows:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i), \tag{1}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f), \tag{2}$$

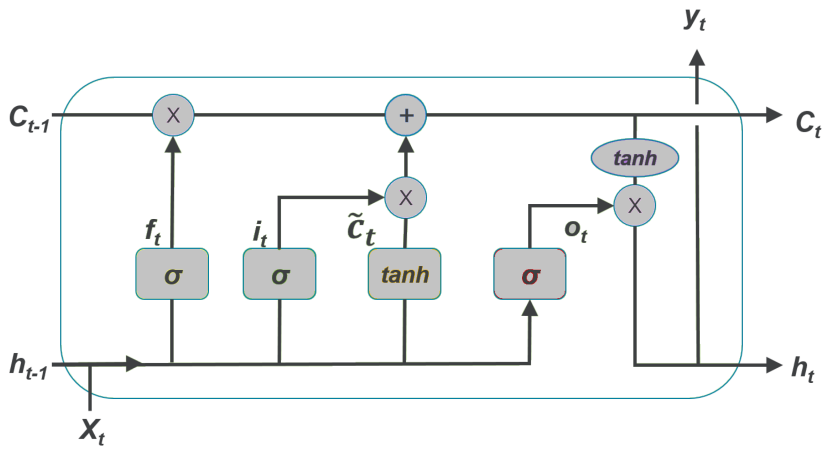

Figure 1. Architecture of a LSTM block.

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o), \tag{3}$$

The input gate Eq. 1 presents what information will be stored in the cell state. The forget gate in Eq. 2 presents what information can be thought away from the cell state, and the output gate Eq. 3 presents which information is used to provide the activation to the final output of the LSTM block at timestamp t.

Let define $\tilde{c}_t$ to be the candidate for cell state c_t at the timestamp t and h_t be the output of LSTM block at the timestamp t, the equations for cell state are as follows:

$$\tilde{c}_t = tanh(w_c[h_{t-1}, x_t] + b_c). \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t. \tag{5}$$

$$h_t = o_t * tanh(c_t). \tag{6}$$

Eq. 4 to Eq. 6 allow the cell state to decide the information to forget from the previous one (i.e., $f_t * c_{t-1}$) and the information to consider from the current timestamp (i.e., $i_t * \tilde{c}_t$). The output h_t of the current LSTM block is used to predict the output corresponding to the timestamp t and the output of the last sequence is used to predict the label of the input. For the sentiment classification problem, the training process is to predict the label of input sequence $x_1, x_2, \ldots x_n$ given the final hidden state h_n.

3.2. Bi-LSTM with Attention

A bidirectional LSTM (Bi-LSTM) network combines two independent LSTM together [17]. This allows Bi-LSTM to have both backward and forward information of the time sequence rather than only backward information in LSTM. In Bi-LSTM, the inputs go in two directions, i.e., from past to future and from future to past. Thus, two hidden states are combined for the next Bi-LSTM layer. Fig. nm3 presents the

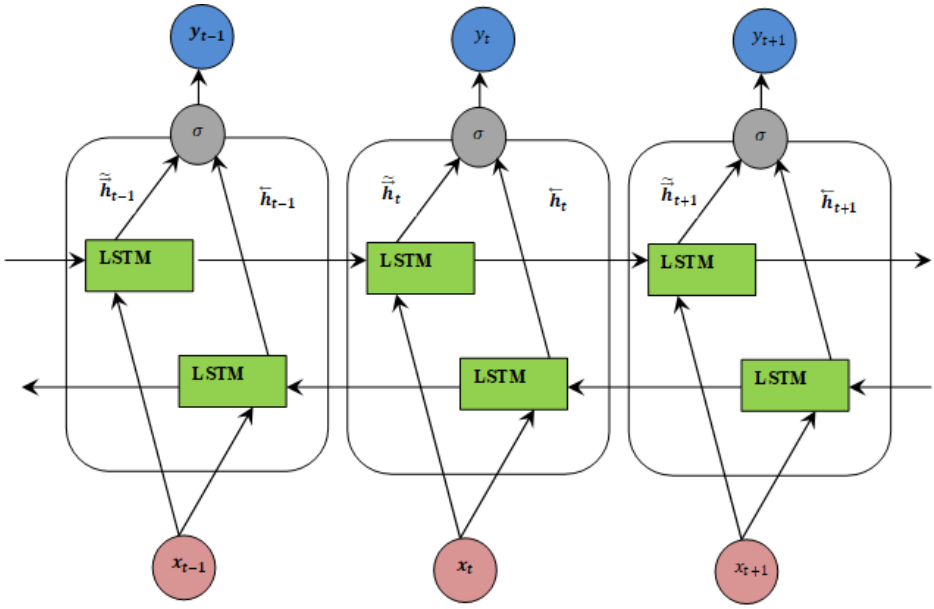

Figure 2. Architecture of a Bi-LSTM model.

architectures of a Bi-LSTM. Each word vector w_t of a document inputs to two LSTM cells (i.e., forward LSTM $\overrightarrow{h_t}$ in Eq. 7 and backward LSTM $\overleftarrow{h_t}$ in Eq. 8). The output of the hidden state is h_t which is the combination of $\overleftarrow{h_t}$ and $\overrightarrow{h_t}$ as in Eq. 9.

$$\overrightarrow{h_t} = LSTM\left(w_t, \overrightarrow{h_{t-1}}\right), \tag{7}$$

$$\overleftarrow{h_t} = LSTM\left(w_t, \overleftarrow{h_{t+1}}\right), \tag{8}$$

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right], t = 1, 2, ..., n \tag{9}$$

where $[;]$ is the concatenation of backward hidden state $\overleftarrow{h_t}$ and forward hidden state $\overrightarrow{h_t}$. The hidden state $H = h_1, h_2, \ldots, h_n$ is considered as the representation of the input sequence.

$$H = (h_1, h_2, ..., h_n) \tag{10}$$

3.3. Residual technique

The residual technique is used to alleviate the degradation problem in the training process of deep neural networks [10]. Fig. 3 presents the building blocks of a Bi-LSTM with residual technique in which a shortcut connection is added from one building block to another. One building block is formed by an identity mapping (a shortcut connection in Eq. 11).

$$y = \textit{Bi-LSTM}(x) + x, \tag{11}$$

where x and y are the input and output of the *Bi-LSTM* layer considered. The dimension of x and *Bi-LSTM*(x) must be equal.

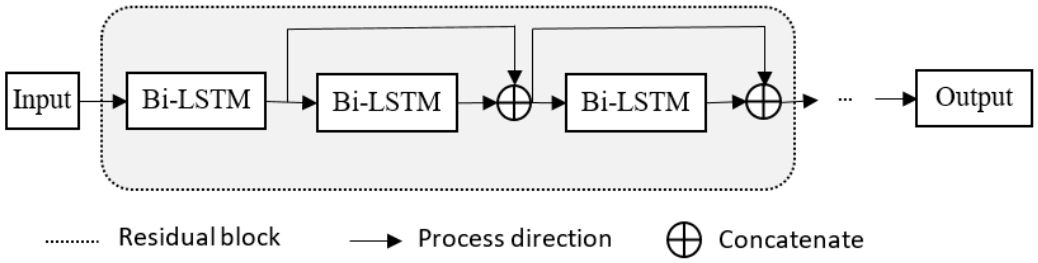

Figure 3. Residual learning architecture with increasing Bi-LSTM depth.

4. Proposed Methods

4.1. Pre-processing Vietnamese language

The first step in our proposed method is to pre-process the documents. Firstly, we removed redundant data from documents. Those redundant data are special characters, punctuation marks, or even symbols not included in linguistic conventions. Next, we replaced acronyms and emojis in reviews with similar meaning words. Next, we split the documents into sentences to avoid confusion in separating words in the next step. Finally, we tokenized words from the sentences we already have in the previous step.

In the Vietnamese language, a term has one or more words (called meaning word). Therefore, the tokenizing word is a challenging task. Moreover, the words tokenization problem also caused a certain error rate, which affects the selection of features for the sentiment analysis problem. For instances, the following words are meaningful in Vietnamese that consist of more than one word: thứ sáu, sinh viên, trung bình, etc. In this paper, we performed tokenization by using the Vietnamese language tool [27].

4.2. Proposed model

This subsection presents the structure of our proposed network. The proposed network is called Residual Attention with Bi-LSTM (ReAt-Bi-LSTM) that integrates the residual technique with multiple layers of Bi-LSTM network and an attention layer for sentiment classification. Finally, the output of the last Bi-LSTM is concatenated with the output of the attention layer to form the final document's representation. The training process attempts to map the representation vector $[c_t; h_t]$ to the label y_t by the *Softmax* function. The idea of using the residual technique in ReAt-Bi-LSTM is to alleviate the degradation problem when using many Bi-LSTM layers. Moreover, the attention mechanism allows the model to increase the weight of important words and decrease the weight of the unimportant words in the input documents. Subsequently, these two techniques help ReAt-Bi-LSTM enhance the accuracy of the sentiment classification problem.

Fig. 4 presents ReAt-Bi-LSTM in detail. The input to ReAt-Bi-LSTM is the word vectors of the input document. We trained the word embedding model using the data collected on the wiki [28]. However, due to the lack of data, the word vector in

Vietnamese is often not as good as English. To address this problem, we add more Bi-LSTM layers to ReAt-Bi-LSTM to keep improving from the input word vectors. After several Bi-LSTM layers with the residual connection, the representation for each word x_t is formed by concatenating the forward hidden state and the backward hidden state h_t. Next, an attention mechanism is performed on the output of the hidden states H to determine the weight α_{tj} for each $h_j, j = 1, 2, \ldots, n$. Specifically, the cell state (or context vector) c_t is computed as a weighted sum of hidden states $h_1, h_2, \ldots h_n$:

$$c_t = \sum_{j=1}^{n} \alpha_{tj} h_j \tag{12}$$

The weight $\alpha_{t,j}$ of each hidden state h_j is computed by the following equation:

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{j=1}^{n} exp(e_{tj})}, \tag{13}$$

with

$$e_{tj} = tanh(W_h * h_j + b_h), \tag{14}$$

where W_h, b_h are the weight and bias of attention network.

Finally, the output of the last Bi-LSTM is concatenated with the output of the attention layer to form the final document's representation. The training process attempts to map the representation vector $[c_t; h_t]$ to the label y_t by the *Softmax* function.

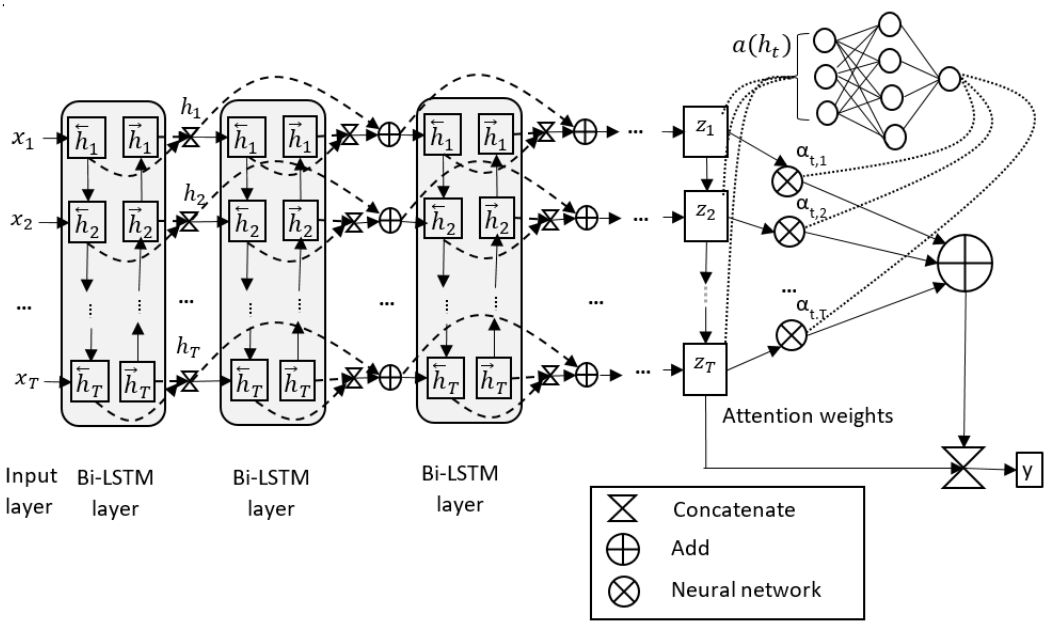

Figure 4. The structure of ReAt-Bi-LSTM.

5. Experimental Settings

This section presents the datasets used in the experiments and the parameter's setting for the tested models.

5.1. Datasets

We tested the proposed model using four Vietnamese sentiment datasets.

- VLSP dataset: This dataset contains user reviews of electronic products. This is a balanced dataset with 3 classes. It consists of 2 subsets: training set and test set. This dataset was developed by the Association of Vietnamese Language and Speech Processing (VLSP) [29].
- AiVN dataset: This dataset is used in the comment classification contest organized by aivivn.com[1]. This dataset is an imbalanced dataset of two classes.
- Foody dataset: The Vietnamese sentiment dataset for foods and services is collected and labeled by the Streetcodevn team. [2]
- VSFC dataset: Vietnamese Students' Feedback Corpus was developed by a team of authors at the University of Information Technology, Vietnam National University - Ho Chi Minh City, Vietnam [30]. It is an unbalanced dataset with three classes consisting of 3 subsets: train set, development set, and test set.

Table 1. Description of Vietnamese sentiment datasets.

Datasets	**VLSP**	**VSFC**	**Foody**	**AIVN**
Train	5100	11426	30000	16087
Development	No	1583	10000	No
Test	1050	3166	10000	No
Total	6150	16175	50000	16087

For the datasets without a test set, i.e., AiVN and Foody, we divided them into two sets: a training set and a test set with a ratio of 70:30. The number of samples on each dataset is described in Table 1.

5.2. Parameters Settings

We use the pre-trained model based on a Vietnamese corpus collected from Wiki[28] to generate the embedding vectors of the tested methods. Each word is represented by a 300-dimensional vector. Moreover, these word vectors will be updated during the training process of ReAt-Bi-LSTM. The number of Bi-LTSM layers is five, and the number of neurons in the hidden layer of LSTM is 128, so each Bi-LSTM network has 256 units. The Adam algorithm, with the learning rate at 10^{-4}, is used to train the models. The early stopping technique is also used during the training process to avoid over-fitting. The training process is stopped when the model's accuracy on the

[1]https://www.aivivn.com/contests/1

[2]https://streetcodevn.com/blog/dataset

validation set is not reduced for five consecutive steps. We divide the training set into two parts for the datasets without the validation set, with a ratio of $8:2$. We also use the drop-out technique to prevent the model from over-fitting, in which 15% neurons in the Bi-LSTM layers are dropped during the training. The model with the highest accuracy on the validation is selected to be the final solution.

5.3. Evaluation Metrics

We use three metrics, i.e., accuracy (ACC), F-score (F1), and Area Under the Curve (AUC) score [31] to compare the tested methods. These metrics are calculated based on the four following definitions.

- True Positive (TP): A TP is an outcome where the model correctly predicts the positive class.
- True Negative (TN): A TN is an outcome where the model correctly predicts the negative class.
- False Positive (FP): An FP is an outcome where the model incorrectly predicts the positive class.
- False Negative (FN): An FN is an outcome where the model incorrectly predicts the negative class.

ACC is the most common criterion to compare classification algorithms. Formally, the ACC of a classifier method applied on a dataset is calculated as in Eq. 15.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

F1 score is the harmonic mean of *Precision* calculated by $\frac{TP}{TP+FP}$ and *Recall* calculated by $\frac{TP}{TP+FN}$.

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

AUC is the area under the Receiver Operator Characteristic (ROC) curve. The ROC curve is plotted with the True Positive Rate (*TPR*) against the False Positive Rate *FPR* where *TPR* is on the *y-axis* and *FPR* is on the *x-axis*. Here, TPR known as **sensitivity** measures the proportion of positive cases in the data that are correctly identified (Eq. 17). FPR known as **(specificity)** is the proportion of negative cases incorrectly identified as positive cases in the data (Eq. 18).

$$TPR = \frac{TP}{TP + FN}. \tag{17}$$

$$FPR = \frac{FP}{TN + FP}. \tag{18}$$

5.4. Experimental Setup

We carried out two sets of experiments to evaluate the proposed model. The first set is to compare the effectiveness of ReAt-Bi-LSTM with some previous deep neural network models in sentiment classification. The compared methods include three baseline models (i.e., CNN [32], [33], LSTM [32], and Bi-LSTM [34]), the attention models (i.e., CNN-attention [35], LSTM-attention [33], and Bi-LSTM-attention [36]), and a recently proposed model used Bi-LSTM with self-attention (i.e., SAN [37]). All methods are trained using the training sets and evaluated using the testing sets on four Vietnamese sentiment datasets, i.e., Foody, AiVN, VSFC, and VLSP.

The second set is to investigate the impact of increasing the depth in ReAt-Bi-LSTM. We compare the AUC score of the Bi-LSTM model with and without residual learning on the **VLSP** dataset when the number of Bi-LSTM layers increased from 1 to 7.

6. Results and Discussion

This section presents the result of the two experiment sets described in Subsection 5-D.

6.1. Performance Comparison

Table 2 presents the ACC, F1, and AUC score of ReAt-Bi-LSTM and the other tested models. First, we can observe that the LSTM-based models (i.e., LSTM, Bi-LSTM) often achieve higher accuracy than CNN-based models on almost datasets using three performance metrics. This result evidences for the effectiveness of the ability to compare semantic dependency of LSTM-based models in sentiment classification.

Second, it can be seen that using the attention mechanism helps to increase the accuracy of the LSTM and Bi-LSTM models. For example, the AUC scores of LSTM and Bi-LSTM-based models are increased from 80.49% and 80.82% with the baseline models to 81.32% and 81.01% with the attention mechanism-based models (i.e., LSTM-attention and Bi-LSTM-attention), respectively. Conversely, the self-attention mechanism in SAN does not help to improve the performance of the Bi-LSTM model in Vietnamese sentiment classification. Moreover, the Bi-LSTM-attention models usually perform better than the LSTM-attention models. The reason could be that in Bi-LSTM, both forward and backward sequences are considered when predicting the sentiment label while in LSTM only before words.

Last but most important, Table 2 shows that our proposed model, i.e., ReAt-Bi-LSTM, always achieves the best result among all tested models. For instance, the AUC scores are increased from 94.73%, 93.56%, 91.62%, and 83.80% with the Bi-LSTM-attention model trained on the Foody, AiVN, VSFC, and VLSP datasets to 95.18%, 95.70%, 95.11%, and 84.24%, respectively, with ReAt-Bi-LSTM. This result confirms that using a residual technique to alleviate the degradation problem and using the attention mechanism to emphasize the important parts in the input documents are beneficial for sentiment classification in the Vietnamese language.

Table 2. Results of metrics on datasets.

Metrics	Methods	Datasets			
		Foody	AiVN	VSFC	VLSP
ACC	CNN [33]	88.52	89.41	89.04	67.24
	CNN-Attention [35]	87.43	89.50	89.86	66.10
	LSTM [32]	88.53	89.35	90.21	68.29
	LSTM-Attention [33]	88.23	89.50	90.68	66.10
	Bi-LSTM [34]	88.11	89.48	89.92	67.81
	Bi-LSTM-Attention [36]	88.91	89.72	90.40	65.52
	SAN [18]	88.45	89.29	88.57	66.57
	ReAt-Bi-LSTM	**89.05**	**89.87**	**91.16**	**69.33**
F1	CNN [33]	88.51	89.45	88.29	66.36
	CNN-Attention [35]	87.43	89.52	89.35	66.11
	LSTM [32]	88.53	89.39	89.21	65.87
	LSTM-Attention [33]	88.22	89.52	90.03	64.56
	Bi-LSTM [34]	88.11	89.50	88.53	65.54
	Bi-LSTM-Attention [36]	88.91	89.74	89.21	65.77
	SAN [18]	88.44	87.17	86.18	61.68
	ReAt-Bi-LSTM	**89.05**	**89.89**	**90.42**	**67.74**
AUC	CNN [33]	94.90	95.57	93.46	83.13
	CNN-Attention [35]	93.54	95.35	93.52	82.50
	LSTM [32]	94.90	95.47	94.24	80.49
	LSTM-Attention [33]	94.77	95.60	94.46	81.32
	Bi-LSTM [34]	94.84	95.51	94.07	80.82
	Bi-LSTM-Attention [36]	95.17	95.58	94.77	81.01
	SAN [18]	94.73	93.56	91.62	83.80
	ReAt-Bi-LSTM	**95.18**	**95.70**	**95.11**	**84.24**

6.2. Impact of Residual Learning

This subsection analyses the effectiveness of using the residual mechanism in ReAt-Bi-LSTM. We compare the ACC score of ReAT-Bi-LSTM with and without using the residual technique on VLSP when the number of Bi-LSTM layers is increased from 1 to 7. This result is presented in Fig. 5 in which the x-axis shows the number of Bi-LSTM layers while the y-axis gives the accuracy of testing models corresponding to the number of Bi-LSTM layers.

We observed that adding more Bi-LSTM layers to ReAt-Bi-LSTM without using residual technique does not improve its accuracy in sentiment classification. The accuracy of ReAT-Bi-LSTM without residual learning is decreased when the number of Bi-LSTM layers is increased. The reason is that adding more stacked Bi-LSTM layers often leads to the degradation problem [10], and it makes the ReAt-Bi-LSTM model harder to train [11]. Conversely, adding more stacked Bi-LSTM layers with the residual technique enhances the accuracy significantly. This is because the residual method helps to alleviate the degradation problem in the training process of the Bi-LSTM model. The

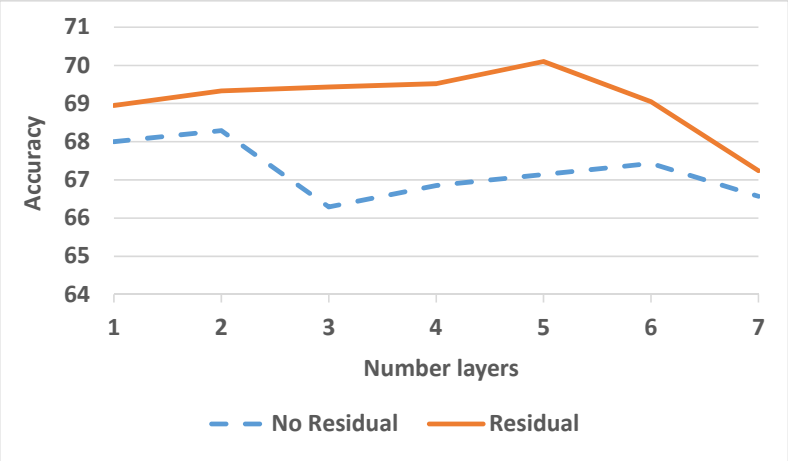

Figure 5. Compare accuracy of stacked layers with and without residual in the VLSP dataset.

figure shows that the ACC score of ReAt-Bi-LSTM reached the highest value when the number of Bi-LSTM layers increased to five, and then it decreased. Perhaps, adding too many Bi-LSTM layers enlarges the model size, and this leads to the network being overfitted on the training data. In our experiment, we choose the number of layers at 5 for the good performance of ReAt-Bi-LSTM on all tested datasets.

6.3. Impact of Attention Mechanism

This subsection examines the effectiveness of using the attention mechanism in ReAt-Bi-LSTM. We compare the ACC score of ReAT-Bi-LSTM with and without using the Attention technique on VLSP when the number of Bi-LSTM layers is increased from 1 to 6. This result is presented in Fig. 6. It can be seen that adding more Bi-LSTM layers to ReAt-Bi-LSTM without using the attention technique does not improve its accuracy in sentiment classification. The accuracy of ReAT-Bi-LSTM without attention fluctuates when the number of Bi-LSTM layers is increased. Conversely, adding more stacked Bi-LSTM layers with the attention enhances the accuracy significantly. The figure shows that the ACC score of ReAt-Bi-LSTM reached the highest value when the number of Bi-LSTM layers increased to five, and then it decreased. Overall, the results in this subsection and subsection IV.B shows that both residual and attention techniques are important for the performance of ReAt-Bi-LSTM. Therefore, combining two techniques in one model, i.e., ReAt-Bi-LSTM, helps to enhance the accuracy of the model in sentiment classification. These results provide some explanation for the better performance of ReAT-Bi-LSTM compared to the other tested methods.

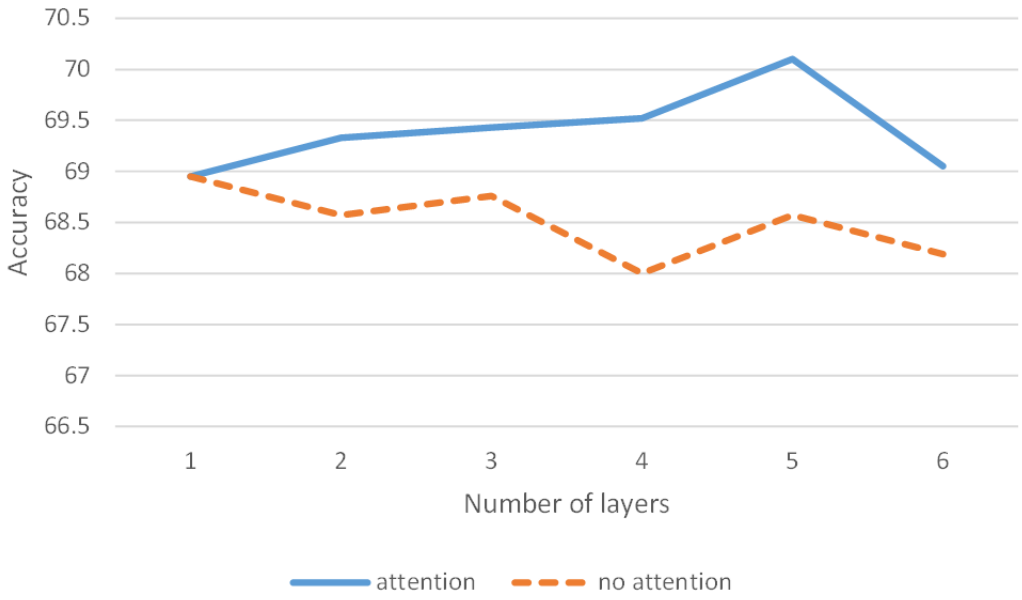

Figure 6. Compare models with and without attention.

6.4. Error Analysis

This subsection presents some samples which are incorrectly predicted by ReAt-Bi-LSTM. This result is listed in Table 3. It can be seen from this table that the incorrect cases often belong to two groups. The first group includes the samples which contain many negative words such as "không", "chưa", "chẳng", etc. These samples imposes the difficulty for the predictive model. The second group includes the samples containing a lot of abbreviations and the samples are written without accents. In the future, we plan to use some reprocessing techniques to address these problems.

7. Conclusion

In this paper, we proposed the ReAt-Bi-LSTM architecture for Vietnamese sentiment classification. In ReAt-Bi-LSTM, the input words are embedded using the Word2Vec technique and then encoded by a multiple layers Bi-LSTM model. The Bi-LSTM model captures the semantic information in both forward and backward directions at every word encoding. Moreover, the residual technique reduces the degradation problem in the training process once the number of Bi-LSTM layers increases. Finally, the attention layers are added to the last Bi-LSTM layer to exploit the core components that have a decisive influence on the sentiment of the document.

We have carried out extensive experiments on four Vietnamese sentiment analysis datasets to evaluate our proposed models' strength and property. The results of ReAt-Bi-LTSM are compared with various baselines and a recently proposed model using three

Table 3. Some errors from the proposed model.

Reviews	Predicted	True label
Mình đang rất cần 1 con SSD siêu như này,nhưng giá trên rõ ràng ko đủ khả năng để mua rồi	**0**	**1**
mình thích cái tích hợp thanh toán ghê tiết là ở vn cả Aw và Android wear vẫn chưa dùng dc.	**2**	**1**
em cứ thích nó lai lai chút, kiểu i5 lai với mi 1s ấy	**2**	**1**
Hỗ trợ TIếng Việt, nhưng nhập liệu giọng nói Tiếng Việt không được.	**0**	**1**
Chả thấy đẹp gì cả!	**2**	**0**
ZenFone 3 không làm tui thất vọng. Đang trên tay, ngất ngây con gà tây!	**0**	**2**
"Xuống"quá rồi BB ơi, anh hết muốn yêu em rồi !!!	**2**	**0**
Chưa ấn tượng lắm	**2**	**0**
nghe giang hồ đồn,là hiện thị đẹp lắm	**2**	**1**
Bạn tìm con nào cấu hình và giá rẻ như này được không.	**2**	**1**
Đã chuẩn bị từ lâu rồi. Hiện đủ tiền mua Iphone 6 Plus, từ giờ tới tháng 9 chắc gom đc thêm ít để tậu em này.	**1**	**2**
thấy gear s ngon hơn...	**2**	**0**
DT OBI SF1, 2 SIM, rat tot ban.	**1**	**2**
dù sao vẫn lag	**2**	**0**
Qua dep.... Gia tam 7 trieu la may com Sam, Apple, Asus, ... chet het.	**2**	**0**
Tui mua sony 50w800c thay chay cung nhanh lam day cac bac. Hinh anh thi sony moi la so 1	**0**	**2**
Dang sai ipad pro la da thay lon lam roi con ss nay lam ra chac de nhin choi thoi ai ma mua vi su qua kho cua no	**1**	**0**

performance metrics. The experimental results demonstrate that our proposed model considerably enhances the accuracy of the Vietnamese sentiment classification problem compared to the other tested models.

There are several research directions arising from our paper. First, we plan to enhance the attention architecture in this paper by using an attention matrix to represent a document [38]. In this method, each row of the matrix represents one level of attention for the document. Second, the trainable weights in ReAt-Bi-LSTM increase its training time. To reduce the training time, we plan to use a federated learning technique [39]. This technique distributes the training process into multiple devices to speed up the training process. Finally, we would like to apply our model to a broader range of datasets, including English datasets, to better understand its strengths and limitations.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2019.05.

References

[1] A. F. Agarap and P. Grafilon, "Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn)," *arXiv preprint arXiv:1805.03687*, 2018.
[2] D. Mali, M. Abhyankar, P. Bhavarthi, K. Gaidhar, and M. Bangare, "Sentiment analysis of product reviews for e-commerce recommendations," in *Proceedings of 44th IRF International Conference, 29th November*, 2015.
[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[4] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.
[5] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.
[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
[9] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[11] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual lstm model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, 2018.
[12] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
[13] G. U. Srikanth *et al.*, "Survey of sentiment analysis using deep learning techniques," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*. IEEE, 2019, pp. 1–9.
[14] A. Rios and R. Kavuluru, "Convolutional neural networks for biomedical text classification: application in indexing biomedical articles," in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, 2015, pp. 258–267.
[15] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
[16] K. Baktha and B. Tripathy, "Investigation of recurrent neural networks in the field of sentiment analysis," in *2017 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2017, pp. 2047–2050.
[17] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673 – 2681, 12 1997.
[18] G. Letarte, F. Paradis, P. Giguère, and F. Laviolette, "Importance of self-attention for sentiment analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 267–275.
[19] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1650–1659.
[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.
[21] X. Zhou, X. Wan, and J. Xiao, "Attention-based lstm network for cross-lingual sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 247–256.
[22] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
[23] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional lstm with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, 2020.

[24] P. Guan, B. Li, X. Lv, and J. Zhou, "Attention enhanced bi-directional lstm for sentiment analysis," *J Chin Inform Proc*, vol. 33, no. 2, pp. 105–111, 2019.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[26] Y. Wen, A. Xu, W. Liu, and L. Chen, "A wide residual network for sentiment classification," in *Proceedings of the 2018 2Nd International Conference on Deep Learning Technologies*, 2018, pp. 7–11.

[27] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 56–60. [Online]. Available: https://www.aclweb.org/anthology/N18-5012

[28] H.-Q. Nguyen and Q.-U. Nguyen, "An ensemble of shallow and deep learning algorithms for vietnamese sentiment analysis," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 2018, pp. 165–170.

[29] H. Nguyen, H. Nguyen, Q. Ngo, L. Vu, V. Tran, N. Xuan Bach, and C. Le, "Vlsp shared task: Sentiment analysis," *Journal of Computer Science and Cybernetics*, vol. 34, pp. 295–310, 01 2019.

[30] K. Van Nguyen, V. D. Nguyen, P. X. Nguyen, T. T. Truong, and N. L.-T. Nguyen, "Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2018, pp. 19–24.

[31] M. Hossin and S. M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 01–11, 03 2015.

[32] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 573–580. [Online]. Available: https://www.aclweb.org/anthology/S17-2094

[33] M. Usama, B. Ahmad, E. Song, M. S. Hossain, M. Alrashoud, and G. Muhammad, "Attention-based sentiment analysis using convolutional and recurrent neural network," *Future Generation Computer Systems*, 2020.

[34] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 534–539. [Online]. Available: https://www.aclweb.org/anthology/D17-1056

[35] J. Zhang, Z. Wu, F. Li, J. Luo, T. Ren, S. Hu, W. Li, and W. Li, "Attention-based convolutional and recurrent neural networks for driving behavior recognition using smartphone sensor data," *IEEE Access*, vol. 7, pp. 148 031–148 046, 2019.

[36] C. Pandey, Z. Ibrahim, H. Wu, E. Iqbal, and R. Dobson, "Improving rnn with attention and embedding for adverse drug reactions," in *Proceedings of the 2017 International Conference on Digital Health*, 2017, pp. 67–71.

[37] A. Ambartsoumian and F. Popowich, "Self-attention: A better building block for sentiment analysis neural network classifiers," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium: Association for Computational Linguistics, November 2018.

[38] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[39] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3298981

Manuscript received: 01-10-2020; Accepted: 10-12-2020.

■

Nguyen Hoang Quan graduated from College of Natural Science HCM in 2001, received a master's degree from Le Quy Don Technical University in 2010. Currently a PhD student at the Faculty of Information Technology, Le Quy Don Technical University. Research field: NLP, sentiment analysis. E-mail: nghoangquan@gmail.com

Vu Ly received her MS degree at Inha University, Korea in 2014. She currently is a PhD student in the major of Mathematics theory for Information Technology from Le Quy Don Technical University, Vietnam. Her research interest includes data mining, machine learning, deep learning, network security.

Nguyen Quang Uy graduated from Le Quy Don Technical University in 2004. Received a doctorate from Dublin University - Ireland in 2011. Currently Associate Professor Ph.D at the Faculty of Information Technology - Le Quy Don Technical University. Research field: Artificial Intelligence, Machine Learning, Information Security. E-mail: quanguyhn@gmail.com

RESIDUAL ATTENTION BI-DIRECTIONAL LONG SHORT-TERM MEMORY CHO BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM TIẾNG VIỆT

Tóm tắt

Phân loại quan điểm là bài toán ước tính giá trị của các ý kiến, tình cảm và thái độ của mọi người đối với các sản phẩm, dịch vụ, cá nhân và tổ chức. Phân tích quan điểm giúp các công ty hiểu khách hàng của họ để cải thiện chiến lược tiếp thị trong thương mại điện tử, nhà sản xuất quyết định cách cải thiện sản phẩm hoặc mọi người tự điều chỉnh hành vi trong cuộc sống.

Trong bài báo này, chúng tôi đề xuất mô hình mạng học sâu để phân loại các đánh giá sản phẩm bằng tiếng Việt. Cụ thể, chúng tôi phát triển một mô hình học sâu mới được gọi là mô hình Residual Attention Bi-directional Long Short-Term Memory (ReAt-Bi-LSTM). Đầu tiên, kỹ thuật Residual được sử dụng trong nhiều lớp Bidirectional Long Short-Term Memory (Bi-LSTM) để nâng cao khả năng học đặc trưng của mô hình từ các tài liệu đầu vào. Thứ hai, cơ chế Attention được tích hợp sau lớp Bi-LSTM cuối cùng để đánh giá đóng góp của mỗi từ của vector ngữ cảnh. Cuối cùng, biểu diễn của các văn bản là sự kết hợp của vector ngữ cảnh và đầu ra của Bi-LSTM. Biểu diễn này nắm bắt cả thông tin ngữ cảnh từ vector ngữ cảnh và thông tin có trình tự từ mạng Bi-LSTM. Chúng tôi đã tiến hành nhiều thử nghiệm trên bốn bộ dữ liệu quan điểm tiếng Việt. Kết quả cho thấy rằng mô hình đề xuất của chúng tôi cải thiện độ chính xác hơn so với một số phương pháp cơ bản và một số mô hình hiện đại cho bài toán phân loại quan điểm.