PROCEEDINGS OF SCIENCE

# Attention-based BiLSTM Neural Networks for Sentiment Classification of Short Texts

**Xianglu Yao**[1]
*School of Math and Computer Department, Wuhan Polytechnic University*
*Wuhan, 430040, Hubei, China*
*E-mail:* collinyao@foxmail.com

Sentiment analysis of short texts such as single sentence has been a research hotspot of natural language processing (NLP). There still exists the challenge of effectively handling the problem with limited contextual information and semantic features. Hence, in this paper, an attention-based bidirectional LSTM neural network (AB-BiLSTM) is proposed to solve the problem. The proposed model can attend the qualitative and informative parts and learn semantic features from both directions of a sentence to perform sentiment analysis of short texts. The proposed model mainly contributes to take advantage of the attention mechanism to capture the informative parts of a sentence without any syntactic features and lexicon features. The model is conducted on the Stanford Sentiment Treebank dataset and the Movie Review Data provided by Cornell University for single sentence sentiment analysis of binary classification. The experimental results indicate that the proposal in this study has superior performances over the existing methods, without taking the attention mechanism into account.

*ISCC2017*
*16-17 December 2017*
*Guangzhou, China*

1Speaker





PoS(ISCC 2017)014

## 1.Introduction

In recent years, with the popularization of the internet technology and the ever-increasing scale of internet users, user text data has been accumulating rapidly. Studies on sentiment classification of short texts have become the main stream in the field of machine learning.

Sentiment analysis [1], also known as sentiment mining, has become an increasingly important research direction, is mainly used for analyzing the point of view, emotion, attitude, and evaluation of user comments, etc. In addition, the major part of common sentiment analysis is to predict polarity classification for user comments. In this paper, the stress has been mainly put on the sentiment polarity (such as positive, negative, neutral, and etc.) at sentence-level prediction.

Traditional machine learning method and deep learning method are the leading research approaches for sentiment classification nowadays. The traditional machine learning models include Support Vector Machine, Naive Bayes, Max Entropy and etc. Experiment [2] indicates that while combined with semantic features such as unigram, bigram, and position of words, the traditional machine learning methods can achieve certain effect, however with the pretreatment process being complex and dependent on a large amount of manual annotation datasets or lexical resources. Meanwhile, the final model is with difficulty in learning the deep semantic characteristics of the short texts.

Deep learning has showed an ideal effect on various tasks such as machine translation [3], question-answering system [4] and text summarization [5]. As a special RNN structure, LSTM [6] can learn the sentence representation of any lengths and dependencies, as well as overcome the problem of gradient explosion. The attention mechanism has been successfully applied in image recognition [7] and machine translation [8]. Experiments demonstrate that the attention model can measure the probability of attention distribution and also can effectively prevent the loss of information. Additionally, the accuracy of image recognition and machine translation can be distinctly improved.

Inspired by the successful application of the neural network based on attention mechanism in NLP [8], this paper proposes the attention-based BiLSTM model for sentiment classification of short texts. The BiLSTM model performs outstandingly to deal with information redundancy and long-term dependency problems, and the attention mechanism provides an optimized feature vector for sentiment classification, which plays a significant role in improving classifiers' performance.

In this papar, specific chapters are arranged as follows. Section 2 reviews the related work about sentiment classification. Section 3 describes the details of our attention-based BiLSTM model. Section 4 presents the experiment settings and results. Finally, the conclusion is drawn to as a reference for future studies.

## 2. Related Work

Since Nasukawa [1] put forward the concept of sentiment analysis in 2003, researchers have carried out deep and extensive researches on it. The research methods about sentiment classification mainly consists of supervised learning and unsupervised learning, and most of the current researches are based on supervised learning algorithm.

For example, via combining the advantages and disadvantages of Support Vector Machine and Naive Bayes, Wang demonstrated a synthesis of both methods, NBSVM model [9], which

had a good performance in many datasets. Rink and Harabagiu [2] utilized many semantic features and SVM classifier for sentiment analysis.

In the past few years, deep learning has achieved impressive performance in various tasks of NLP. Hinton proposed a method based on word representation called word embedding [10], with the core idea that words will be mapped to a low dimensional space according to a distributed method, and the low-dimensional word vectors are used to represent the complicated and sophisticated semantic features of texts. The skip-gram model [11] proposed by Mikolov can conveniently and efficiently train word embeddings for one text dataset. Consequently, the method of deep learning can automatically capture the syntactic and semantic features of texts to some extent, which saves a lot of efforts and time.

The attention mechanism in early time was mainly adopted in the field of computer vision [12]. Bahdanau [8] proposed an encoder-decoder network based on the attention mechanism in the domain of machine translation. It can be assumed that the attention mechanism can attend the importance of different words in a sentence. Therefore, this paper puts forward a BiLSTM model based on attention mechanism for the purpose of sentiment analysis of short texts.

## 3.Attention-based BiLSTM Model

In this section, the architecture of Attention-based BiLSTM neural networks (AB-BiLSTM) is introduced for sentiment classification of short texts. The proposed model mainly consists of word encoder, attention layer and softmax layer, which are shown in Figure 1. The details of different parts of the model are presented as follows.

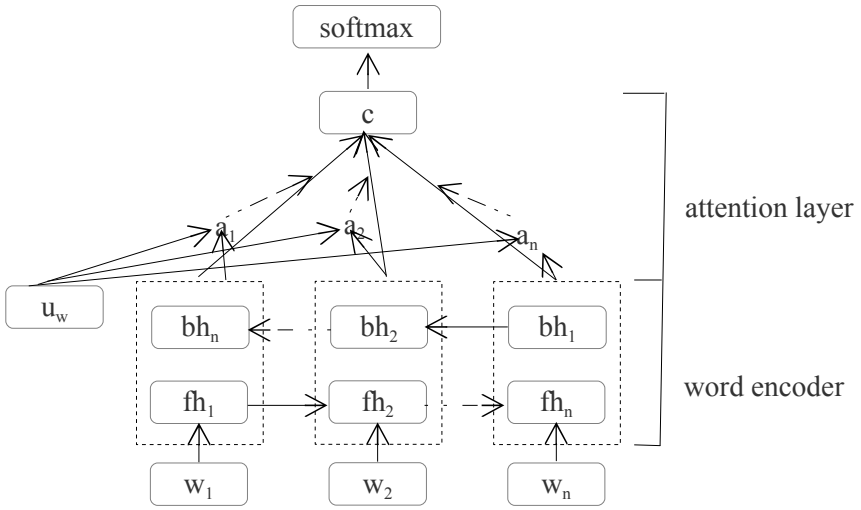


**Figure 1:** The architecture of Attention-based BiLSTM Model

### 3.1 Word Encoder

In this paper, we only concentrate on the sentence-level sentiment classification. It can be assumed that a sentence contains n words $[w_1,w_2,...,w_n]$, $w_k$ denotes the kth word in a sentence and n denotes the length of a sentence. Firstly, every word is embedded into a d dimensional vector, which is called as word embedding [13]. After all the word vectors are stacked, an embedding matrix $M^{n\times d}$ is generated, where d means the embedding size and n denotes the length of a sentence. Word embeddings can be regarded as parameters of neural networks or pre-trained from proper corpus via language model. And then, the embedding matrix is used as input for the BiLSTM networks to encode a sentence.

**LSTM:** Recurrent neural networks has a significant role for modeling sequential data through mapping a variable length vector of word to a fixed-length vector. However, the RNNs are incompetent to learn long-term dependencies on account of the gradient vanishing or exploding problems. The LSTM networks are proposed and developed based on the RNNs to address those weaknesses. The basic structure of LSTM consists of three gates and a cell memory state. A single LSTM cell is implemented by the following composite functions:

$$X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{3.1}$$

$$f_t = \sigma(W_f \cdot X + b_f) \tag{3.2}$$

$$i_t = \sigma(W_i \cdot X + b_i) \tag{3.3}$$

$$o_t = \sigma(W_o \cdot X + b_o) \tag{3.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot X + b_c) \tag{3.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.6}$$

$W_i$ , $W_f$ , $W_o$ are the weighted matrices and $b_i$ , $b_f$ , $b_o$ are biases of LSTM cell, which are all parameters of the input gates, forget gates and output gates respectively. σ is the sigmoid function and ⊙ is the element-wise multiplication. $x_t$ represents the word embedding of the input of LSTM cell and $h_t$ is the vector of the hidden state. The specific schematic is illustrated in Figure 2.

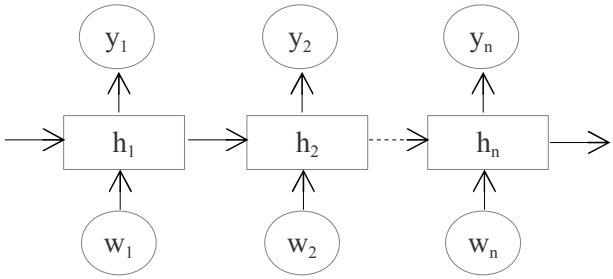


**Figure 2:** The architecture of LSTM network

**BiLSTM:** The bidirectional LSTM(BiLSTM) contains two independent LSTMs, which acquire annotations of words by summing up information from two directions of a sentence, and then merge the sentimental information in the annotation. Specifically, at each time step t, the forward LSTM calculates the hidden state $fh_t$ based on the previous hidden $fh_{t-1}$ state and the input vector $x_t$ , while the backward LSTM calculates the hidden state $bh_t$ based on the opposite hidden state $bh_{t-1}$ and the input vector $x_t$. Finally, the vectors of two directions are concatenated as the final hidden state of the BiLSTM model. The two LSTM neural network parameters in BiLSTM networks are independent of each other, and they share the same word embeddings of the sentence. The final output $h_t$ of the BiLSTM model at the step t is shown as the following equation:

$$h_t = [fh_t, bh_t] \tag{3.7}$$

### 3.2 Attention Layer

As we all known, the meanings and emotions usually vary according to different parts of the sentence . Some words in a sentence can be decisive while the others are irrelevant. Based on that, the attention mechanism is introduced to attend those informative words and aggregate

their representations to form a sentence vector. In effect, the attention mechanism is to compute a context vector in a sentence.

Concretely, based on the mentioned above, the LSTM or BiLSTM network will produce a hidden $h_t$ state at each time step. To begin with, the vector of $h_t$ is fed into a one-layer MLP to learn a hidden representation $u_t$. Then a scalar importance value is computed for $h_t$ given $u_t$ and a word-level context vector $u_w$. At last, the attention-based model computes the weighted mean of the state $h_t$ through a softmax function. The context vector $u_w$ can be perceived as a high-level representation [14] for distinguishing the importances of different words over the word sequences. The formulas can be described as follows:

$$u_t = \tanh(W_w h_t + b_w) \tag{3.8}$$

$$a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \tag{3.9}$$

$$c = \sum_t a_t h_t \tag{3.10}$$

### 3.3 Softmax Layer

In this paper, a fully connected softmax layer is used as a classifier. As the high-level representation of the sentence, the vector c can be adopted as the sentiment features for sentence classification:

$$\hat{y} = softmax(W_c c + b_c) \tag{3.11}$$

where $\hat{y}$ is the predicted value from the model, $W_c$ is the weighted matrix and $b_c$ is the bias.

### 3.4 Model Training

The cross-entropy error of sentiment classification is used as the loss function:

$$L = -\sum_i y_i \log \hat{y}_i \tag{3.12}$$

where $y_i$ is the gold-standard sentiment label and $\hat{y}_i$ is the predicted label from the model.

The back-propagation method [15] is utilized for the whole set of parameters to obtain the derivative of the loss function and then update all parameters of models with stochastic gradient descent.

## 4. Experiments and Analysis

### 4.1 Datasets

In the mentioned experiments, the Stanford Sentiment Treebank dataset (SST) [11] and movie review data (MR) [16] provided by Cornell University are applied to evaluate the model.

**SST:** The dataset includes five kinds of labels, which contain the very negative label, negative label, neutral label, positive label, and very positive label. Besides, this dataset contains 11,855 sentences, with nineteen words in each sentence on average. After removing neutral reviews, abinary labeled version of SST is obtained. There are 11,123 sentences available, which contains 5,679 positive sentences and 5,444 negative sentences. A standard 10-fold cross validation (CV) is conducted for this dataset.

**MR:** It contains 10,662 sentences, in which the positive and negative sentences are in equal proportions. This corpus is split into 80% for training, 10% for validation and 10% for testing.

In Table 1, additional details about the two corpora are described.

| Dataset | Data Size | Vocab Size | Train | Dev | Test |
|---|---|---|---|---|---|
| SST | 11123 | 20757 | 10011 | 1112 | CV |
| MR | 11855 | 18758 | 8530 | 1066 | 1066 |

**Table 1:** Details of the Two Corpora

The CV stands for the 10-fold cross validation and the others are the size of the corresponding set.

### 4.2 Experimental Settings

The models are implemented based on the Google Tensorflow, which integrates the current relatively popular deep learning modes. Although the Tensorflow platform supports the application of GPUs, experiments are only conducted on CPUs and achieve good effect. The details settings of experiments are as follows:

**Word embeddings:** Word embeddings can be regarded as parameters of neural networks or pre-trained from proper corpus. In this paper, the word embeddings are trained with the models together and initialized by the uniform distribution [-0.1, 0.1]. They are also fine-tuned during the training procedure.

**Hyper parameters:** The hyper parameters of the models are tuned repeatedly several times. In order to achieve the ideal effects of the experiments, the experiment parameters are selected by grid search and cross validation. Finally, the size of word embeddings is set to be 50 and the size of LSTM cell is set to be 50. The attention vector has a dimension of 100 and is initialized at random, which is set to attend the importance of evey word. The dropout rate is set to 0.5 to prevent overfitting. A mini-batch size 128 and the Adam [17] optimization algorithm are utilized to train all models. Each model is trained for 15 epochs in the experiments. More details are shown in Table 2.

**Evaluation metric:** The accuracy metric is adopted as theevaluation metric. The definition is as follow:

$$Accuracy = \frac{T}{N} \tag{4.1}$$

in which T is the correct number of predicted samples, N is the total number of tested samples.

**Baseline methods:** The models are compared with the standard LSTM neural networks and bidirectional LSTM networks without using the attention mechanism. After obtaining the average values of all the hidden states of the LSTM and BiLSTM networks, they are fed into a softmax function to estimate the probability of each sentiment label.

| parameter | parameter name | value |
|---|---|---|
| rnn_size | LSTM cell size | 100 |
| num_rnn_layers | Layer size | 1 |
| embedding_size | Embed size | 50 |
| attention_dim | Attention dimension | 100 |
| dropout | Dropout | 0.5 |
| max_steps | Input size | 60 |
| batch_size | Batch size | 128 |
| lr | Learning rate | 1e-4 |

| | | |
|---|---|---|
| epoch_size | Epoch size | 15 |

**Table 2:** Details of Experiment Parameters Which Our Models Use.

### 4.3 Results and Model Anylysis

| Model | SST | MR |
|---|---|---|
| LSTM | 85.2% | 80.2% |
| BiLSTM | 87.9% | 83.7% |
| AB-LSTM | 88.2% | 85.6% |
| AB-BiLSTM | **90.3%** | **87.8%** |

**Table 3:** Accuracy of Different Models on Both Corpora for Binary Classification by Using the Accuracy Metric.

Table 3 clearly describes the results of the experiment. The scores of the proposed model are in bold. The LSTM and BiLSTM networks have achieved a good effect for the binary sentiment classification. And the BiLSTM obtains an improvement over LSTM when bidirectional semantic information is taken into consideration.

From the results of the LSTM and BiLSTM models, it is possible that the bidirectional LSTM model can get more semantic features, which is helpful for the sentiment clsssifiction.

The basic LSTM model is unable to attend any informative parts of a sentence, so that it is difficult for the LSTM model to improve the accuracy of sentiment classification.

Comparing with LSTM, the AB-LSTM model demonstates that the attention mechanism can improve the accuracy of the LSTM model for the sentiment clsssifiction by almost 2%~3%. By incorporating attention mechanism, the attention-based BiLSTM model achieves comparable performances on both corpus in comparison with many of the external baseline methods. The experiment results illustrate that the proposed model is more effective for the sentiment clsssifiction, compared with the above baseline methods.

## 5.Conclusion

In this paper, the attention-based BiLSTM neural networks is proposed to perform sentiment classification. The main contributions of the proposed model is to capture the informative parts of a sentence and learn the contextual information between non-consecutive words without any syntactic features and lexicon features. Experiments demonstrate that the proposed model, AB-BiLSTM, is more competitive and obtains superior performance over the baseline models for sentiment classification. In the experiment, only the attention mechanism is considered, while other factors and characteristics are taken into account in the future work.

## References

[1] T. Nasukawa, J. Yi. *Sentiment analysis: Capturing favorability using natural language processing*[C]. Proceedings of the 2nd International Conference on Knowledge Capture. ACM, 2003:70-77.

[2] B. Rink, S. Harabagiu. *UTD: Classifying semantic relations by combining lexical and semantic resources*[C]. Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010:256-259.

[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami , C. Dyer. *Neural Architectures for Named Entity Recognition*. arXiv preprint arXiv:1603.01360, 2016.

[4] D. Golub, X. He. *Character-Level Question Answering with Attention*. arXiv preprint arXiv:1604.00727, 2016.

[5] A.M. Rush, S. Chopra ,J. Weston. *A Neural Attention Model for Abstractive Sentence Summarization*. arXiv preprint arXiv:1509.00685, 2015.

[6] S. Hochreiter, J. Schmidhuber. *Long Short-Term Memory*[J]. Neural Computation, 1997, 9(8):1735–1780.

[7] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu. *Recurrent models of visual attention*[C]. In Advances in Neural Information Processing Systems, 2014:2204-2212.

[8] D. Bahdanau, K. Cho, Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv preprint arXiv:1409.0473, 2014.

[9] S. Wang, C.D. Manning. *Baselines and bigrams: simple, good sentiment and topic classification*[C]. Proceedings of ACL'12, Jeju Island, 2012: 90-94.

[10] G.E. Hinton. *Learning distributed representations of concepts*[C]. Proc of the 8th Annual Conference of the Cognitive Science Society, 1986:1-12.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. *Distributed Representations of Words and Phrases and Their Compositionality*[C]. Advances in Neural Information Processing Systemsm, 2013:3111-3119.

[12] M. Denil, L. Bazzani, H. Larochelle, F.N. De. *Learning where to attend with deep architectures for image tracking*[J]. Neural Computation, 2012, 24(8):2151-2184

[13] Y. Bengio, P. Vincent, C. Janvin. *A neural probabilistic language model*[J]. Journal of Machine Learning Research, 2003, 3 (6) :1137-1155.

[14] S. Sukhbaatar,A. Szlam,J. Weston,R. Fergus. *End-To-End Memory Networks*. arXiv preprint arXiv:1503.08895, 2015.

[15] A. Graves , J. Schmidhuber. *Framewise phoneme classification with bidirectional LSTM and other neural network architectures*[J]. Neural Networks the Official Journal of the International Neural Network Society, 2005,18 (5-6):602-610.

[16] B. Pang, L. Lee. *Seeing stars: Exploiting class relationships for sentiment categorization with respectto rating scales*[C]. In Proceedings of ACL 2005, 2005.

[17] D.P. Kingma and J.L. *Ba. Adam: a method for stochastic optimization*[C]. In Proc. International Conference on Learning Representations(ICLR), 2015:1-13.