International Core Journal of Engineering Volume 7 Issue 7, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

Deep Learning based Text Sentiment Analysis During Epidemic

Hui Li[a], Lin Zhang
College of Software Engineering, Shanghai Maritime University, Shanghai, 201306, China.
[a]2894637854@qq.com

Abstract

The outbreak of COVID-19 in 2019 is a heart-rending event for both countries and individuals. In this paper, the corpus of Chinese epidemic comments on Weibo was extracted and analyzed, and the Bert-Bilstm model was used to train the corpus. A comparative experiment was conducted with the deep learning models of BILSTM-ATTENTION and BILSTM, and the results showed that: Compared with the BILSTM-ATTENTION model and the BILSTM model, the accuracy of P was increased by 5.93% and 11.25%, the recall rate of R was increased by 4.2% and 8.98%, and the F1 value was increased by 5.91% and 11.09%, respectively. It is proved that the Bert-BILSTM model has better adaptability under the COVID-19 comment corpus of Weibo.

Keywords

Bert; COVID-19; BILSTM; Deep Learning.

1. Introduction

With the advent of the era of big data information and the development of network technology, the use of social media in people's lives more frequently, the user can through the Internet or various social platform can real time to share their life and the life of rich emotional color information, the timeliness and interaction also led to the rapid growth of Internet information. Effectively analyzing the value contained in this information will have different values and meanings for both society and individuals.

The COVID-19 [1] at the end of 2019 is an extremely painful event for both countries and individuals. The people of the whole country are in the same boat to fight the epidemic together. Outbreak in spreading, brings to the national people's panic at the same time, also in all aspects has brought the influence to the national people's life and make social and economic threat, as well as various aspects of the country's development has brought great challenges, and with the rapid development of the new champions league outbreak and diffusion as well as the personnel flow, the epidemic situation of severe constantly at home and abroad. Based on this, the mood of the people across the country has been greatly affected, and the COVID-19 has also caused strong public mood swings on the Internet.

2. Relevant model technology introduction

2.1 Bert

In October 2018, Google AI Research Institute put forward a pre-training model called Bert, whose appearance brings a new perspective and changes to the vector construction technology of words in the field of natural language processing. Bert's proposal is considered to be the best breakthrough technology in recent years. It has greatly improved the accuracy in eleven directions in the field of natural language processing, and has been proved to have very good performance. The structure of the Bert model is shown in the figure. see Fig. 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

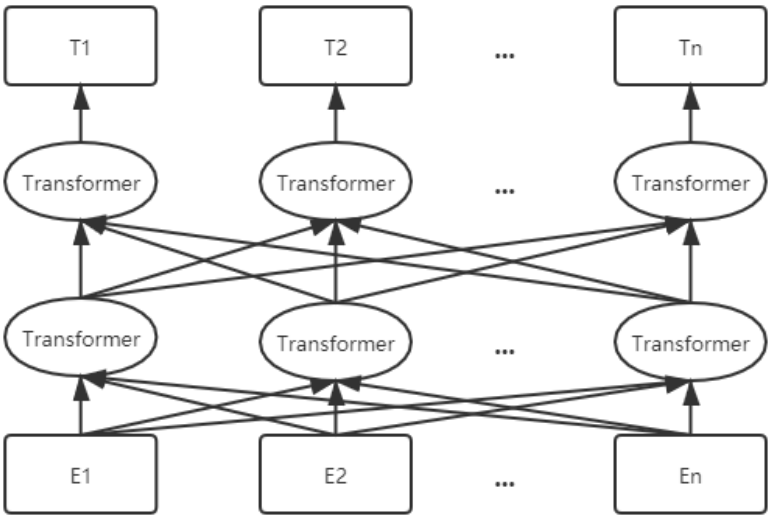

Figure 1. Bert model structure

Bert [2] pre-training model is mainly based on a language model of Transformer structure, which consists of three parts: input layer, coding layer and output layer. E1,..., En represents the input vector of the model, and in the middle is a multi-layer bi-directional Transformer feature extractor, T1..., Tn represents the output vector of the model.

In Bert, TokenEmbeddings, SegmentEmbeddings and PositionEmbeddings constitute the input layer of Bert. Transformer [3], as the feature extractor of the Bert pre-trained language model and the core component of Bert, is a serial-to-sequence model based on the self-attention mechanism, in which the main structure is the Encoder of the coding part. The diagram of the Bert coding layer is shown below. see Fig. 2.

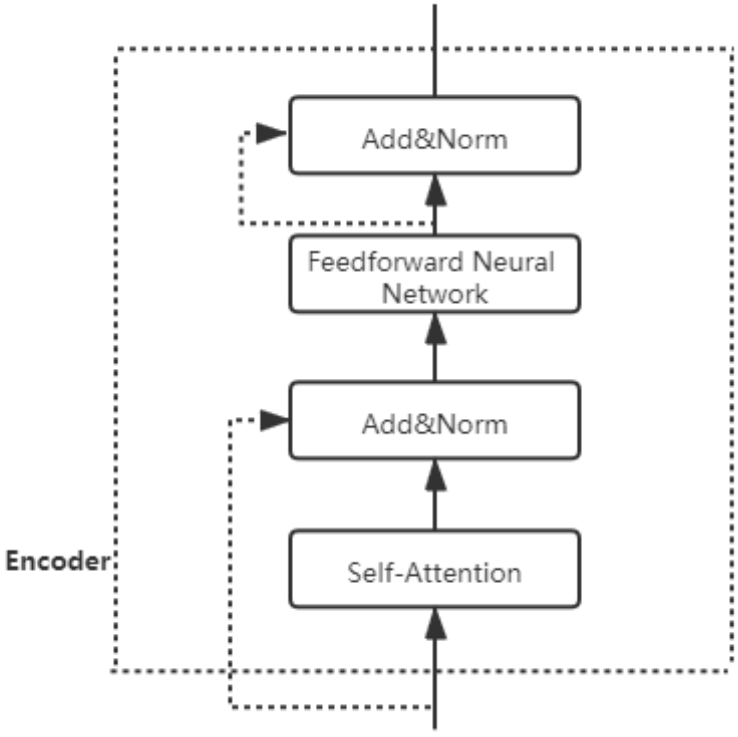

Figure 2. Bert coding layer structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

Bert is used to extract word vectors for the short text of Chinese Weibo comments with weak context and semantic connection. The bi-directional Transformer feature extractor can obtain more context information and extract more text features to obtain better word vector representation [4]. So this paper uses Bert model to extract word vectors.

2.2 BiLSTM

The BILSTM model is composed of a bidirectional LSTM network model, which takes into account both the previous information of the text and the future context information of the text, which can help improve the discriminant ability of the model. For example, if we encode the sentence "I love calligraphy", the forward LSTM input "I", "love", and "calligraphy" will get three vectors $\{h_{L0}, h_{L1}, h_{L2}\}$. Backward LSTM input "calligraphy", "love" and "I" to get the other three vectors $\{h_{R0}, h_{R1}, h_{R2}\}$. Then the forward and backward implicit vectors are concatenated to obtain $\{h_0, h_1, h_2\}$.

In the text to be analyzed, the information of the current word to be analyzed is not only related to the preceding part of the word, but also related to the following part of the word. In view of the characteristics of microblog comment text, this paper uses bidirectional short and short memory neural network BILSTM [5] to replace LSTM [6], making up for the following information. Since BILSTM is composed of bidirectional LSTM model, it trains from the front and back directions and then connects the final result to the same layer of output [7]. Due to its ability to remember past and future information, it has the following problems for the traditional LSTM model: The difficulty that the serialization processing problem can not capture the context information is solved, and the accuracy of classification can be improved theoretically. The BILSTM neural network model is shown in the figure below: see Fig. 3.

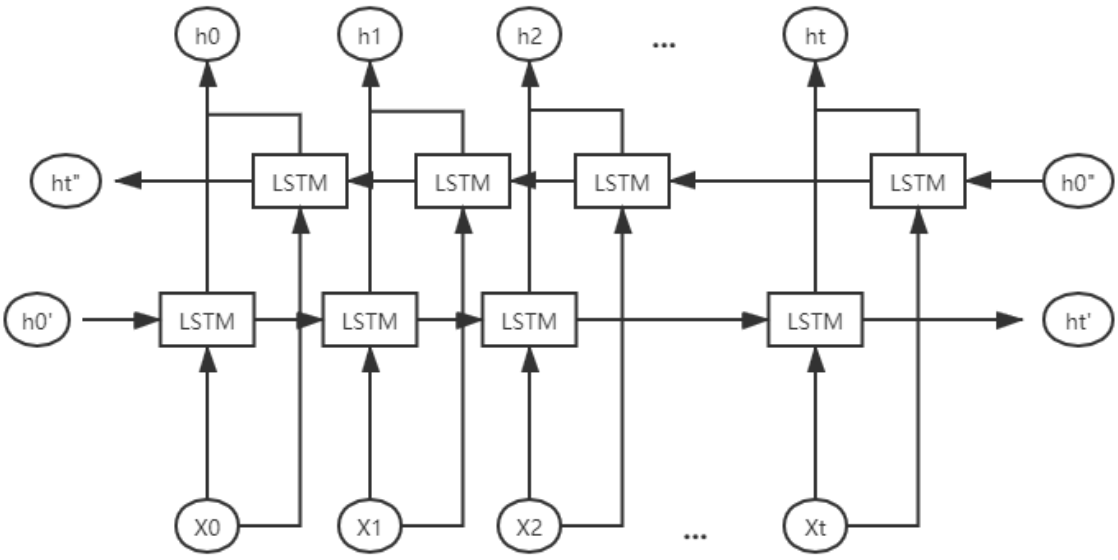

Figure 3. BiLSTM neural network model

2.3 Softmax function

Softmax function [8] is a normalized exponential function commonly used in multiple classification problems. It maps the output of neurons in multiple neural networks to the interval (0,1) to get the output value and calculate the probability of belonging to each sample. Its calculation formula is as follows:

$$S_i = \frac{e^i}{\sum_j e^j} \tag{1}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

3. The model based on Bert-Bilstm

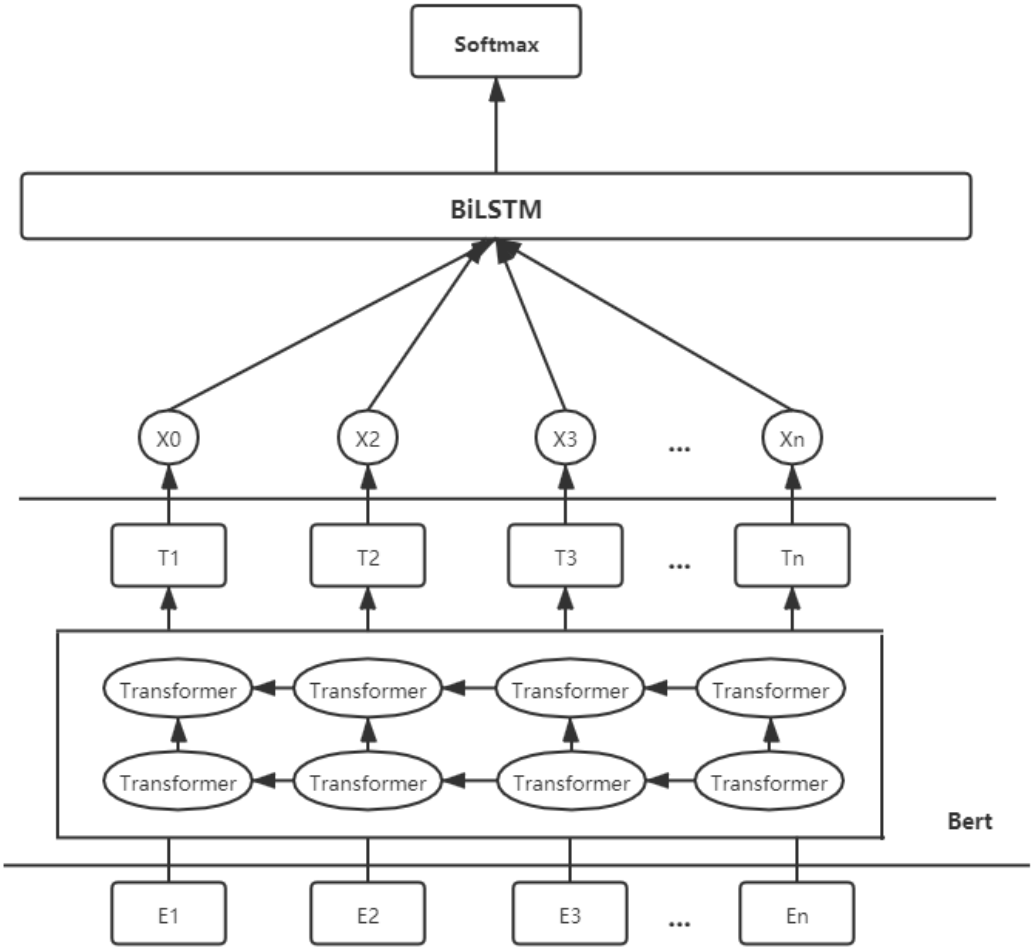

Figure 4. Bert - BiLSTM model

In this paper, Bert-Bilstm fusion neural network model is used to conduct sentiment analysis on the corpus of microblog epidemic comment texts. Microblog comment texts are short text data. For short text data, information expression will be more concise, which will lead to the lack of contextual information, and comments content will be random. This will result in poor expression of word vectors. According to the previous introduction, the core structure of Bert is a bi-directional Transformer, which can more thoroughly capture the context information of the text. To reduce because of the lack of semantic fuzziness in context in bad word expression vector is obtained, so the Bert - BiLSTM training models of neural network model with Bert as text information input, two-way short - and long-term network also contains double LSTM, able to capture better semantics of two-way dependence, so that the better training effect of the model. In the Bert- Bilstm neural network model, the word vector output by Bert is further extracted through the bidirectional neural network layer, and finally the Softmax function is used to judge the textual data bias. see Fig. 4.

4. The experiment design

4.1 Setting up the experimental environment

This paper uses the operating system under Windows 10, the framework is TensorFlow 1.14.0, the programming language is Python 3.6 version, and the development tool is JetBrainSpyCharm to conduct the experiment.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

4.2 Data acquisition

The data set of this paper adopts crawler technology to crawl the real comment text of the news microblog related to "COVID-19 epidemic" on the official website of People's Daily Online according to keywords such as "COVID-19" and takes it as the research object. The time span is from February 16, 2020 to March 22, 2020.The content captured by each microblog includes the microblog's blog content, comments, and the number of thumb ups. At the same time, due to the noise of the crawling comment data, the crawling data should be preprocessed, including data cleaning, bias labeling, word segmentation and stop word removal. In the end, 73,500 micro-blog comments were retained, of which 24,500 were positive, neutral and negative comments respectively, making the ratio of the three emotional tendencies close to 1:1:1 and ensuring the balance of data distribution. In order to facilitate the adequacy and accuracy of subsequent model experiments, the processed data sets that have been processed after a series of preprocessing are divided into training set, verification set and test set according to the ratio of 8:1:1 to conduct subsequent model experiments.

4.3 Evaluation standard

In this paper, the indexes of the performance evaluation methods mainly used in the experimental study include precision rate P, recall rate R and F1 value, and the performance of the experimental model can be comprehensively and effectively evaluated through these methods.

Accurate rate:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{2}$$

The recall rate:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{3}$$

F1 value:

$$F1 = \frac{(2 * \Pr ecision * \mathrm{Re}call)}{(\Pr ecision + \mathrm{Re}call)} \tag{4}$$

TP represents true cases, TN represents true negative cases, FP represents false positive cases, and FN represents false negative cases.

4.4 Experimental results and analysis

In this section, the Bert-Bilstm model in this paper was compared with several other models in terms of accuracy P, recall rate R and F1 [9] on the data set of the COVID-19 crawling, see Table 1.

Table 1. Comparative experimental results

Model	P	R	F1
BiLSTM	0.7128	0.7134	0.7012
BiLSTM-Attention	0.7660	0.7612	0.7530
Bert-BiLSTM	0.8253	0.8032	0.8121

It can be seen that Bert-BILSTM was compared with the BILSTM model and the BILSTM-ATTENTION model, and the results of the Chinese epidemic comment data on Weibo were improved in terms of accuracy rate, recall rate and F1 value. Therefore, it can be seen that the model used in this paper has a good effect on the epidemic comment corpus based on Weibo.

5. Conclusion

This paper adopts the Bert-Bilstm model and applies it to the micro-blog text sentiment analysis task during the epidemic period. Firstly, the dynamic word vector features of the text are extracted from

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0069

the Bert layer, which better combines the contextual information and context. Then, the output is used as the input of the BILSTM network layer to learn the contextual information and emotional features of the text and obtain the semantic representation of the text. The comparison between the Bert-Bilstm model and the other two groups of deep learning models proves the effectiveness of this model on the microblog epidemic comment corpus. In the future, other network structures, such as the Attention mechanism and the CNN network model, can be considered to further improve the performance of the model.

References

[1] Nanshan Chen, Min Zhou, Xuan Dong, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a de-scriptive study[J]. The Lancet, 2020 (prepublish).

[2] Devlin J, Chang M W, Lee K, et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[3] VASWANIA, SHAZEERN, PARMARN, etal. Attention is all you need[J]. Conference and Workshop on Neural Information Processing Systems, 2017: 6000-6010

[4] YANG P, DONG W Y. Recognition method of Chinese named entities based on BERT embedding [J/OL]. Computer Engineering:1-7. [2019-11-21].

[5] XIAO Z, LIANG P J. Chinese sentiment analysis using bidirectional LSTM with word embedding [C]// ICCCS. Proceedings of the 2nd International Conference on Cloud Computing and Security, Nanjing, China: Springer, 2016:601-610.

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term momory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[7] ULLAHA, AHMADJ, MUHAMMAD K, et al. Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features[J]. IEEEAccess, 2017, PP(99):1-1.

[8] HU T T, FENG Y Q, SHEN L J, etal. Selection of main features of LSTM speech emotion based on attention mechanism[J]. Acoustic Technology, 2019, 38(4):414-421.

[9] YANG P, YANG Z H, LUO L, etal. Recognition of chemical drug named entity basedon attention mechanism [J]. Computer Research and Development, 2008, 55(7):1548-1556.