# Network News Sentiment Analysis Based on BERT

**Shiwang Huang(23020211153937)[1], Xiaoyu Wang(31520211154087)[2], Xiaohan Ji(23020211153942)[1], Jingjing Xie(23020211153944)[1], Qingxian Tang(23020211153966)[1]**

[1]Department of Computer Science, Xiamen University, China
[2]Department of Artificial Intelligence, Xiamen University, China
23020211153966@stu.xmu.edu.cn

### Abstract

With the booming development of the Internet, users have produced a large amount of text information, such as news, microblog, blog, and so on. By exploring the emotional attitudes behind these texts, we can grasp the attitudes and mood changes of public opinion, and then help relevant personnel understand the overall evaluation and make appropriate adjustments. The topic of sentiment analysis is to analyze subjective and unstructured texts with emotional colors and identify the attitude of users. In this project, we collected texts from the Internet (mainly from the news), used the BERT model and the LSTM model to judge emotions expressed in texts (positive, neutral, or negative), and drew the following conclusions: The accuracy of Bert Base Chinese Model based on pre-training is much higher than LSTM model.

## Introduction

With the rise of various social platforms, more and more content is generated by users on the Internet, resulting in a large amount of text information, such as news, micro-blogs, and blogs. Faced with such a large and emotional text information, we can fully consider tapping its potential value to serve the people. For example, understanding the emotions in news content can grasp the attitudes and emotional changes of public opinion, which helps the government and enterprises realize public opinion analysis and monitoring, and respond to sudden public opinion events in a timely manner.

Therefore, in recent years, sentiment analysis has received close attention from researchers in the field of computer linguistics and has become a research hotspot. In today's era of information explosion, these news texts are growing exponentially every day, and manual analysis alone requires a lot of manpower and time. Therefore, automatic identification of positive and negative emotions expressed in online news has important theoretical significance and practical value.

Sentiment analysis is to identify the user's views and attitudes about a thing or a person, such as a movie review, a product evaluation, an experience and so on. According to the analysis of the subjective text with emotional color, the user's attitude is recognized, whether it is like, hate, or neutral. There are many applications in real life. For example, through the sentiment analysis of Weibo users(Sun et al. 2015) to predict stock trends, predict movie box office, election results, etc., it can also be used to understand users' preferences for companies and products, and the analysis results can be used Improve products and services, you can also discover the strengths and weaknesses of competitors and so on.

This task aims to accurately distinguish the emotional polarity of the text in the big data set. Emotions can be divided into three types: positive, negative, and neutral. In the face of massive news information, accurately identifying hidden emotional tendencies is of great significance for the effective monitoring, early warning, and guidance of public opinion and the healthy development of the public opinion ecosystem.

There are many difficulties in the topic of text emotion analysis. Text features are difficult to extract and standardize, which is more prominent in Chinese text. The same word may express different emotions in different contexts. For example, ”wear as much as you can in winter and as little as you can in summer” can be expressed in the same way in Chinese, which is not conducive to text emotion analysis. The use of non-emotional stop words will affect the emotional score. Besides, more and more new words on the Internet are also a challenge for text emotion analysis, and text sentiment analysis has certain subjective judgments. The same sentence may have different feelings for different people, and this ambiguity is very difficult to distinguish.

## Related Work

With the development of deep learning research, the application of the neural network model to emotion classification has achieved remarkable results. At present, emotion classification based on deep learning can be divided into emotion classification of single neural network, emotion classification of hybrid neural networks, emotion classification with attention mechanism, and emotion classification of the pre-training model.

Emotion classification of single neural network usually adopts a typical neural network model (such as CNN (Dos Santos and Gatti 2014), LSTM (Wang et al. 2016), etc.) to apply to text emotion classification. (Dos Santos and Gatti 2014) applied CNN model to pre-trained word vector for sentence-level classification to achieve emotion classification chest. (Melamud, Goldberger, and Dagan 2016) use two-way LSTM to effectively learn generic sentence context representations from large corpora. (Qian et al. 2016) define four rules to combine linguistic knowledge (including emotion dictionary, negative words, and adverbs of degree) with LSTM, and propose that LR-LSTM is applied to sentence level emotion classification, which has a much higher effect than the standard LSTM. (Jelodar et al. 2020) used the advantage of LSTM to better capture the semantic meaning of above and below to classify the comments on novel Coronavirus by emotion, to guide the related public opinion problems caused by novel Coronavirus.

Emotion classification of hybrid neural networks usually refers to combining the advantages of different network models and applying them to emotion classification. (Sun et al. 2019) integrated dependency tree and neural network for representation learning and proposed a method combining LSTM and GCN for emotion classification.

Emotion classification with attention mechanism usually refers to the introduction of attention mechanism into deep learning model for emotion classification. Since the attention mechanism can pay attention to a specific part of the input, it can effectively improve the effect of emotion classification. (Mnih et al. 2014) uses the attention mechanism on the RNN model to classify images. Subsequently, (Bahdanau, Cho, and Bengio 2014) uses attention-like mechanisms to perform translation and alignment simultaneously in machine translation tasks, and their work is the first to apply attention-like mechanisms to NLP. Then the attention mechanism is widely used in various NLP tasks based on NEURAL network models such as RNN or CNN. In 2017, (Vaswani et al. 2017) made extensive use of self-attention mechanism to learn text representation. The mechanism of self-attention has also become the focus of recent research and has been explored in various NLP tasks.

Emotion classification using the pre-training model means that using the trained model of the data set and fine-tuning the training model can achieve a better effect of emotion classification. (Yin, Meng, and Chang 2020) combined context representation with binary unit parsing tree and proposed sentiBERT model for emotion classification, which was mainly based on the improvement of Bert model (Devlin et al. 2018). (Ke et al. 2020) added language knowledge into the pre-training model and proposed a new pre-training language representation model sentiLARE by using Sentiwordnet (Baccianella, Esuli, and Sebastiani 2010) to extract the emotional polarity of each word perceived above and below. (Tian et al. 2020) used a semi-supervised method to mine emotional words and attribute-emotional word pairs, and integrated the two kinds of emotional knowledge into BERT, and proposed an emotional knowledge enhancement pre-training model called SKEP.

In our works, we choose document-level sentiment analysis. This problem is basically a text classification problem. Here in general it is assumed that the document is written by a single person and expresses an opinion about a single entity. One of the major challenges in the document-level classification is that all the sentences in a document may not be relevant in expressing the opinion about an entity. We've tried to solve this problem in two methods —— the LSTM method and the BERT method. In the following part, we will first introduce the two methods we used, then introduce the data set, and give the corresponding experiments, results, and conclusion.

## Proposed Method

Categorizing the emotional polarity of the news data. Positive emotions correspond to 0, neutral emotions correspond to 1 and negative emotions correspond to 2. According to the training data, the emotional polarity of the news in the test set should be judged by algorithm or model. The model is base on BERT model ,and the upper frame of BERT is altered. In addition, we also used LSTM to compare with Bert.Specific methods are as follows:

## BERT

### Pre-training

In this text classifification task, we choose to use the pre-training model. The Pre-training model makes good use of transfer learning, so that the model of natural language processing can be applied to different language tasks. The emergence of pre-training makes natural language processing enter a new era.

It has the following advantages:

- It can transfer the knowledge learned from the open domain to downstream tasks to improve low resource tasks, which is also very //beneficial to low resource language processing.
- Be able to learn the context sensitive representation of each word in the input sentence.
- The pre-training model has achieved the best results in almost all NLP tasks. In addition, the pre-training model+fine tuning mechanism has good scalability. When apply a new task, you only need to fine tuning for the labeled data of the task.

The pre-training model has three key technologies.

**The first key technology is transformer**. It is the core network of Pre-training language model. Like most seq2seq models, the structure of transformer is also composed of encoder and decoder. Encoder and decoder are composed of multi-head-self-attention and feedforward neural network. The context of words is obtained by calculating the degree of dependence or correlation between two words and then normalizing these scores.

**The second key technology is self supervised learning**. The most commonly used are AR (autoregressive) LM and AE (automatic coder). ARLM mainly uses the ahead word sequence to predict the occurrence probability of the next word. The automatic coder mainly processes the damaged input sentence (such as masking a word in the sentence, or disrupting the word order) to reconstruct the original data.

**The third key technology is fine tuning**. When doing specific tasks, fine tuning aims to use its labeled samples to adjust the parameters of the pre-training network. Bert is a method based on fine tuning. In Bert, the input is two sentences, and the corresponding coding representation of each sentence is obtained through Bert. We can simply use the first hidden node of the pre-training model to predict the probability that the two sentences are synonymous sentences. At the same time, we need to add an additional linear layer and softmax to calculate the distribution of classification labels. The predicted loss can be transmitted back to Bert, and then the network can be fine tuned. Of course, we can also design a new network for specific tasks and take the results of pre-training as its input.

There are usually two pre-training methods. The trained network is used to extract features from new samples. Then, we input these features into a new classifier and train from scratch. Another fine-tuning strategy is to use the known neural network and known network parameters to modify the final output layer to its own required output layer.

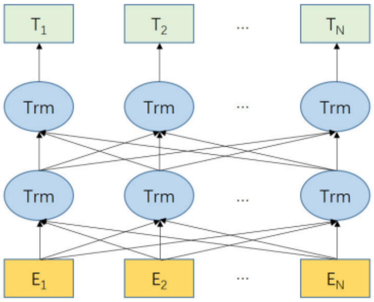

Figure 1: structures based on the BERT model

.

## The Architecture of BERT

Bert is a pre-trained language representation model, which emphasizes that the traditional one-way language model or the shallow splicing of two one-way language models are no longer used for pre-training. Instead, a new masked language model (MLM) is adopted to generate in-depth bi-directional language representation. Bert's model architecture (Figure 1) is actually a multi-layer biderectional transformer encoder. In this task, we first obtain a pre-training model based on Bert. Then, we can fine tune its output layer to obtain the output we want.

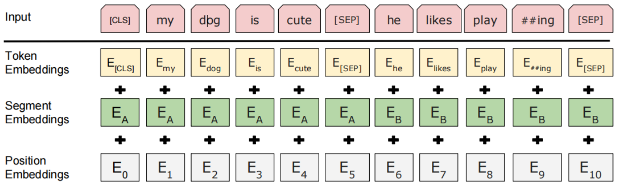

Figure 2: input based on the BERT model

.

## The Input of BERT

Although the input of Bert is the same as that of transformer, Bert is not a transformer embedded in a fixed position, and the embedding position of Bert is learnable. Moreover, the ”sentence” input by Bert can not only be a single sentence, but can be a combination of two or more sentences, but generally we only use two sentences. Figure 2 shows the input of Bert. The first tag of each sequence is always a special classification tag ([CLS]). The final hidden state corresponding to this tag is used as the aggregate sequence representation of the classification task. Sentence pairs are packaged into a sequence. We distinguish these sentences in two ways. First, we separate them with a special tag ([SEP]). Secondly, we add a segment embedding to each tag to indicate whether the word belongs to sentence A or sentence B. Bert's input is the sum of three embedded features. Among them, position embedding encodes the position information of words into feature vectors, and is an important part of introducing the position relationship of words into the model. The eigenvalue of segment embedding marked sentence A is 0 .The characteristic of sentence B is 1.

## Masked Language Model(MLM)

Before Bert, most models were only one-way training from left to right or right to left, without considering the context. The depth biderectional model takes this into account, so it is more powerful.In addition ,since bidirectional conditioning would allow each word to indirectly “see itself”, and the model could trivially predict the target word in a multi-layered context.

The first task of pre-training is MLM, which randomly masks a certain ratio (generally 15%) of tokens, and then determines the word at this position with a certain probability is replaces it with another word, replaces it with [mask] and remains unchanged. The specific strategies are as follows:

- 80% of the time: Replacing the word with the [MASK] token.
- 10% of the time: Replacing the word with a random word.
- 10% of the time: Keeping the word unchanged. The purpose of this is to bias the representation towards the actual observed word.

## Next Sentence Prediction

Many important downstream tasks such as question answering and natural language reasoning are based on understanding the relationship between two sentences. Language modeling does not capture these relationships directly. In order to get a model that can understand the relationship between two sentences, a unitary corpus is generally used to pre-train to get a binary next sense prediction. Therefore, next sentence prediction enables the model to understand the relationship between two sentences. We select sentences A and B for each pre-training sample. Then, with 50% probability, we regard B as A sentence after a, and with another 50% probability, we randomly select a sentence as a sentence after A.

## The Output of Bert

For the text classification task, we only need to insert the task specific input and output into the Bert, and then fine tune all parameters end-to-end. In this task, we use linear layer and

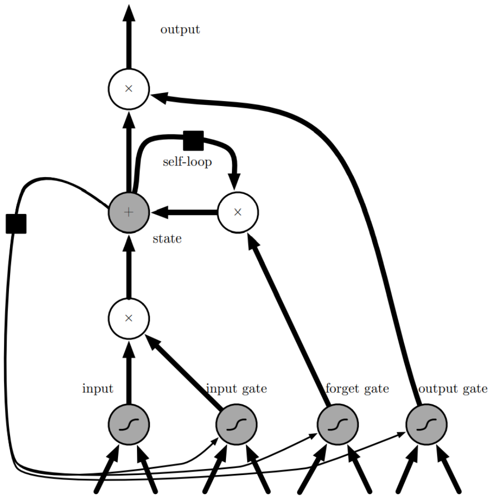

Figure 3: Block diagram of the LSTM recurrent network "cell."

.

activation function as our output layer, so the text is divided into three categories: 0, 1 and 2. 0 is positive,1 is neutral2 is negative

### BERT-Base-Chinese Model

In our works, we use the BERT-Base-Chinese model, which is an improved model based on Bert and can classify Chinese texts. The model download address is:httpshttps//huggingface.co/bert-base-chinese

## LSTM

Long short-term memory networks (Hochreiter and Schmidhuber 1997), which are also known as LSTM have a specific architecture to handle long-term dependencies with sequential data. In particular, LSTM aims to overcome the vanishing gradient and exploding gradient problem of RNN (Bengio, Simard, and Frasconi 1994). LSTM cells are more complicated and maintain two states namely cell state and hidden state. LSTM unit consists of four gates that interact to maintain the cell state and hidden state.

In particular, an LSTM network could control the information flow from current input and hidden state of previous time steps using the gated architecture. Figure 3 depicts a block diagram of the LSTM recurrent network "cell". The cell state is maintained to carry information about long-term dependencies. The LSTM decides which information to keep or dump using the forget gate, which is a function of the previous hidden state and current input. The input gate facilitates the functionality to decide what information should be preserved inside the cell state. The cell state is updated based on the old cell state in the previous time step and the candidate value for the current time step. The candidate value for a given time step is calculated based on the input of the current time step and the previous cell state. While the cell state in the previous time step produces an impact on the current cell state incorporated with the forget gate, the candidate value makes an impact on the cell state engaging with the input gate. Finally, the output gate decides which part of the cell state should be delivered as the output in a given cell state.

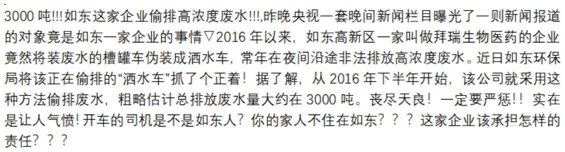
3000 吨!!!如东这家企业偷排高浓度废水!!!,昨晚央视一套晚间新闻栏目曝光了一则新闻报道的对象竟是如东一家企业的事情▽2016 年以来，如东高新区一家叫做拜瑞生物医药的企业竟然将装废水的槽罐车伪装成洒水车，常年在夜间沿途非法排放高浓度废水。近日如东环保局将该正在偷排的"洒水车"抓了个正着！据了解，从 2016 年下半年开始，该公司就采用这种方法偷排废水，粗略估计总排放废水量大约在 3000 吨。丧尽天良！一定要严惩!！实在是让人气愤! 开车的司机是不是如东人？你的家人不住在如东？？？这家企业该承担怎样的责任？？？

Figure 4: One sample of data

.

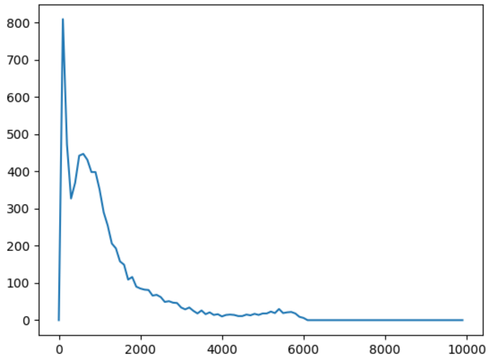

Figure 5: Data length distribution chart

.

## Experiment

### Data Set

Because news is often a matter of reporting and evaluation, our data set is not a traditional one sentence form, but a paragraph. As shown in figure 4. The corresponding label of this news paragraph is 2, which is negative.

As can be seen from figure 4, news headlines and news content are mixed together. In addition, there are many problems in the data set, such as more punctuation and useless symbols, stop words, separation of title and content fields, etc. So we process the data. This makes the data cleaner and the accuracy of the model will be improved accordingly.

Figure 5 shows our data length distribution chart (the x-axis is the data length and the y-axis is the corresponding number). From the chart, we can see that most of the data lengths are about 400. Figure 6 shows our data category distribution chart. About 10% are labeled 0 (positive), about 50% are labeled 1 (neutral), and about 40% are labeled 2 (negative). Then we roughly divide the data set into three categories: training set, verification set and test set, accounting for 60%, 20% and 20% respectively.

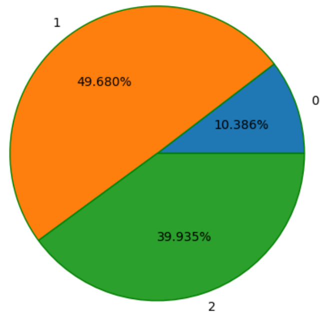

Figure 6: Data category distribution diagram

.

### The Result Based BERT linear layer and activation function

Table 1 shows the accuracy of four different pre-training models combined with the final linear layer and activation function. From the table, we can see that the accuracy of the main pre-training model BERT-base-Chinese used in this example is the highest, reaching 84.5%.

| Pre-Training Model | ACC |
|---|---|
| bert-base-chinese | 0.8452624403544649 |
| chinese-roberta-wwm-ext-pytorch | 0.825494205862304 |
| chinese-rbtl3-pytorch | 0.8070892978868439 |
| chinese-wwm-pytorch | 0.8064076346284935 |

Table 1: Accuracy based on four different pre-training models

### The Result Based LSTM

In this method, the network layer is embedded layer, LSTM layer, and full connection layer. At the embedding layer, the main work is to transform the words in the text into 300-dimensional word embedding. We use the pre-trained Chinese word vector list, and the word vector features include word, character, and n-gram.At the LSTM layer, we set the length of the hidden layer to 128, and the final full-connected layer output to 3 layers for 3 classifications. Finally, the accuracy of LSTM is 71%

## Conclusion

From the results of the above two models, it can be seen that the accuracy of Bert base Chinese model based on pre-training is much higher than that of LSTM model. Although LSTM solves the problem that DNN or bow word bag model can not distinguish timing, it has poor effect on the long dependence problem. After many steps of transmission, the information will be weakened. So that we can't better extract the features in the text.

In this paper, based on the Bert model, by introducing MLM, the learned representation can fuse the context in two directions. Next sentence prediction is also added to understand the relationship between two sentences. Through these two mechanisms, Bert can better extract the features in the text

# References

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2200–2204.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Y.; Simard, P. Y.; and Frasconi, P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dos Santos, C.; and Gatti, M. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 69–78.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9: 1735–1780.

Jelodar, H.; Wang, Y.; Orji, R.; and Huang, S. 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10): 2733–2742.

Ke, P.; Ji, H.; Liu, S.; Zhu, X.; and Huang, M. 2020. Sentilare: Linguistic knowledge enhanced language representation for sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6975–6988.

Melamud, O.; Goldberger, J.; and Dagan, I. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, 51–61.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, 2204–2212.

Qian, Q.; Huang, M.; Lei, J.; and Zhu, X. 2016. Linguistically regularized lstms for sentiment classification. *arXiv preprint arXiv:1611.03949*.

Sun, K.; Zhang, R.; Mensah, S.; Mao, Y.; and Liu, X. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5679–5688.

Sun, X.; Gao, F.; Li, C.; and Ren, F. 2015. Chinese microblog sentiment classification based on convolution neural network with content extension method. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 408–414. IEEE.

Tian, H.; Gao, C.; Xiao, X.; Liu, H.; He, B.; Wu, H.; Wang, H.; and Wu, F. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 606–615.

Yin, D.; Meng, T.; and Chang, K.-W. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.