International Journal of Social Science and Education Research Volume 4 Issue 12, 2021
ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0046

Research on Text Classification of Cyber Violence Speech Based on LSTM

Meilin Tian

Beihang University, Beijing, 100191, China

Abstract

Nowadays, with the development of technology, network communication has gradually become one of the most important ways for people to obtain and exchange information. However, due to the anonymity of the network speech, cyber violence is not a novelty for people. Especially, these years teenagers have become a larger part of Internet users. Therefore, the supervision of comments of social media is very important and necessary. This research analyzes the comments of several controversial topics of Weibo, which is one of the most popular social media in China. It also distinguishes normal comments from some hash comments which attacks the others. The research also trains the dataset after data crawling, cleaning, and labeling. It not only identifies some comments with obvious rude remarks, but also distinguishes other speech with obvious personal attack although it does not contain any dirty words. After training based on the LSTM mode, this study finally evaluates the model which could classify different type of speech. The result of the research shows that it is feasible to supervise violent speech through training the LSTM model. Compared with the previous method of speech supervision, which simply blocks the impolite words, this method also has high accuracy and sensitivity to some novel network terms and abbreviations.

Keywords

LSTM; Social media; Cyber violence; Speech supervision.

1. Introduction

With the development of cyber violence, China has gradually valued the importance of building a green network environment. In 2020, Chinese government has established Regulations on Ecological Governance of network information content [1]. It proposed that China should strengthen the importance of establishment and improvement of a comprehensive network governance system and supervision of cyber violence, cyber manhunt and information forgery. It is pointed out in the regulations that teenagers Internet users are exposed to an unregulated network system, and 28.89% of them have been suffering from violent and abusive comments, most of which are ridicule, abusive or insulting remarks. Therefore, it is very important to establish a green network ecology, which not only requires the supervision of relevant departments, but also require the controlling of remarks and behaviors by the social media.

Though the social media in China are not completely lack of supervision, the traditional methods are generally based on blocking some abusive keywords, or on deleting those comments which contains the intention of insulting others. However, the Internet speech are not formal and change quickly, these methods are not flexible and robust enough.

However, with the development of Artificial Intelligence, especially machine learning, we have several new ideas of solving such problems. By building a model based on machine learning of supervision of insulting remarks, we can quickly and easily identify and screen the rude comments. Through training the LSTM model, we can solve the problems of the quick alternations of network speech and recognize the abbreviations and emotional expressions in

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0046

the comments. The supervision of illegal speech by deep learning can reduce the workload of censors and greatly reduce the time of supervision.

2. Research Methods

2.1. Research Process

2.1.1. Establishment and Acquisition of Dataset

Weibo is one of the most popular social media with a large quantity of users in China. According to the Weibo User Report in 2020 [2], the monthly live users reached 523 million, of which the proportion of post-90s and post-00s in close to 80%, and the trend of increasing proportion of younger user is obvious. Weibo is one of the most popular software for Chinese Internet users to obtain news and discuss social issues. Therefore, we chose the comments of several controversial social issues as the dataset of the research.

In many controversial issues in Weibo, the people in the event will be discussed and might be attacked by the Internet users. Sometimes, even the people who make their own comments would be abused and even be hunted by others. This study focuses on the supervision and classification of cyber violence speech, so it drives importance on the comments of several controversial past events in China, including five events. One of those people who had been discussed have committed suicide due to cyber violence. Those people involved in such incidents suffered from those intolerable insults and even the cyber manhunt. Most of them suffered from depression and decided never to express their feelings on the Internet.

2.1.2. Data Crawling and Cleaning

In this research, we use Python to crawl the comments of Weibo, the steps and basic ideas are as follows. We obtain the ID of a comment first, construct the request URL in the headers, and then obtain the cookies after logging in to turn the pages in Weibo. There is no need to analyze the users, so we only save the users' comments without paying attention their users' names and the number of likes.

While crawling comments, clean the data by clearing all <a> hyperlink labels, and convert some escape characters, such as   to spaces, etc. In addition, there are a lot of abbreviations and emoticons in Weibo comments. We can convert those expressions into words through Unicode expression transcoding.

2.1.3. Data Labelling

This study uses manual ways to annotate Weibo comments. The comments of Internet are full of satire, metaphor, and weird comments, so it is difficult to implement machine annotation. In the process of date labelling, we need to adopt a strict standard of judging the insulting speech, which might lack certain rationality. However, this standard is clearly defined in this research. The criteria are as follows:

Table 1. Text classification criteria

Personal Abusing	Those comments generally contain rude or insulting words. For example, the comments which compare people to other animals, the comments which abuse or curse the relatives or friends of others, the comments which abuse others for mental illness, etc.
Malicious comments	Those comments do not contain obvious rude or insulting words, but with a malicious intention. For example, the comments which casually criticize on others' bodies, etc.
Normal speech	Normal and rational comments.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0046

2.2. Model Establishment

2.2.1. Data Preprocessing

First, remove the stop words and punctuation. Stop words refer to words that occur frequently in natural languages but contain less information. Generally, they do not have semantic nor emotional meaning. Therefore, such words have little significance in text classification and could be deleted before modelling. In this study, we chose the stop word list provided by the natural language processing laboratory of Harbin University of technology.[3] However, in this study, due to the particularity of online speech, some stop words are deleted and 732 stop words are left. In addition, the punctuation marks are deleted, and some popular abbreviations are transformed, such as “srds” as “although…but”. Moreover, Chinese sords segmentation is carried out in this part and the vocabulary is divided and stored by using the Jieba extension.

2.2.2. LSTM Modelling

LSTM(Long-Short Term Memory) is a special RNN network, proposed by Hochreiter and schmidhuber in 1997 [4], performing well in text classification [5]. After data preprocessing, we train the LSTM model. The total number of comments is 3001, including 491 personal insulting comments, 689 malicious comments and 1821 normal comments. One-hot representation is used as words vectorization. Then, the training set and the test set are split with the proportion of 10% of test data. Finally, there are 2700 comments in training set and 301 in test set.

Then, the modelling of the sequence model of LSTM is carried out. The first layer is the embedding layer, which uses a vector with a length of 100 to define each word. Secondly, when the spacialdropout1d layer is updated each time, the input unit is set to 0 in a specific proportion, which can effectively prevent over fitting. The LSTM model is equipped with 100 memory units and the output layer is a full connection layer containing three categories. Because this study is a text multi-classification, the categorical cross entropy is used as the loss function. After defining the LSTM model, we start to train the data with 5 training cycles and the batch size as 64.

3. Results

This study adopts five learning cycles. With the increase of learning, the test accuracy gradually improves, and the test loss gradually decreases. In the fourth cycle, the loss rate of the test set increases and reaches the platform, which means the model is over fitting. At the meantime, the accuracy rate of the test set also increases slowly, or even decreases.

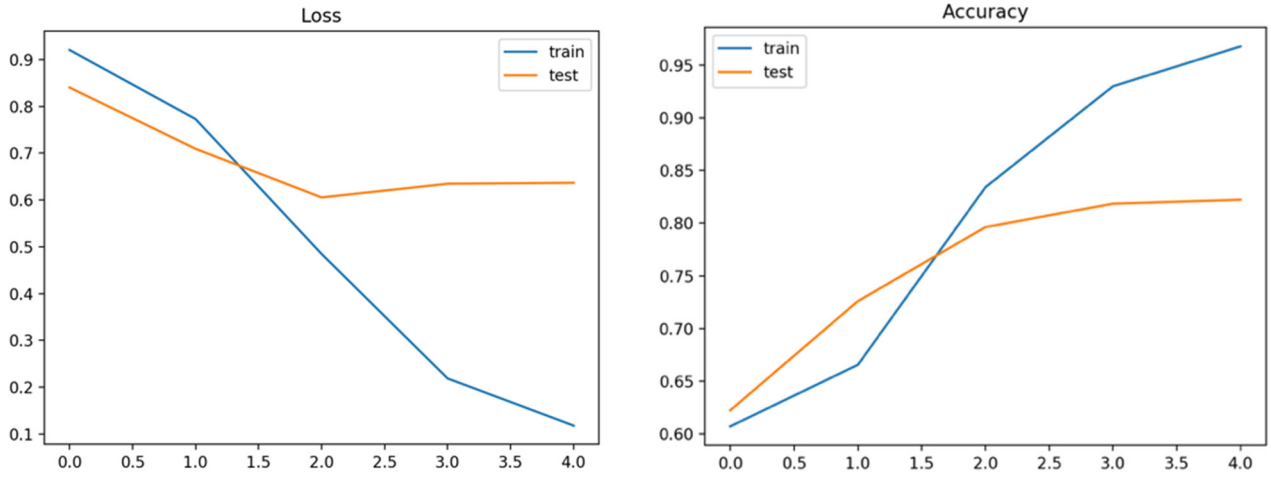

Figure 1. Loss and accuracy tendencies

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0046

In addition, the multi classification task can use the confusion matrix to represent the accuracy of different classes. In this figure, the vertical axis is the actual value and the horizontal axis is the predicting value. All values except the value in diagonal are the number of errors in prediction.

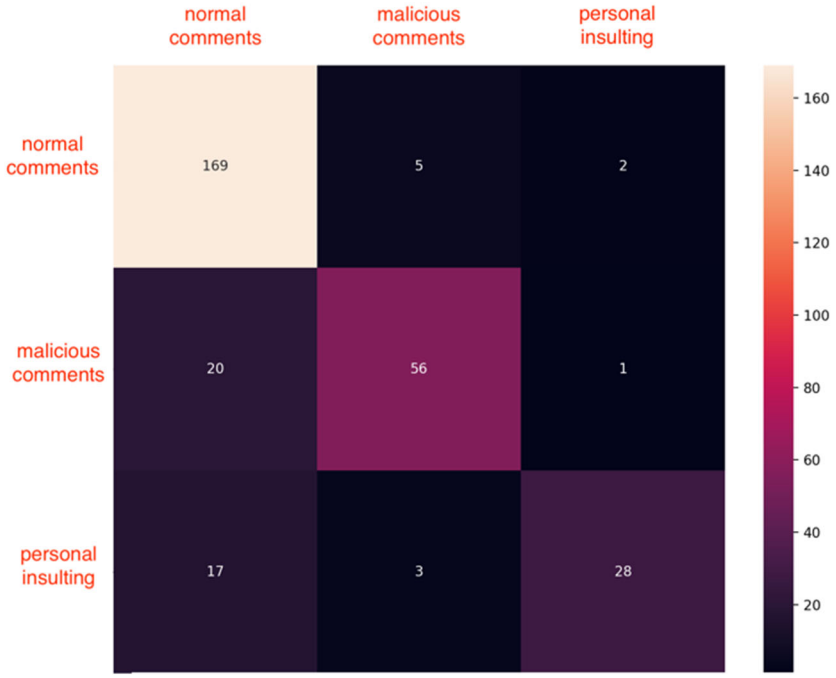

Figure 2. Confusion matrix of text classification

Accuracy indicates how many samples are accurately predicted for all samples. In text multi-classification, we also need to examine the precision, recall and F1 score.

Table 2. Precision, recall and F1-score

class	precision	recall	F1-score	support
Normal comments	82%	96%	88%	176
Malicious comments	88%	73%	79%	77
Personal insulting	90%	58%	71%	48
Accuracy: 84%				

According to the F1-scores, we can conclude that the F1-score of normal comments is the highest, 88%, while the F1-score of the personal insulting comments is the lowest, 71%, which may be due to few comments datasets of personal insulting comments, which means there are many mistakes in prediction.

4. Discussion

Concerning all the results above, the results are basically accords with the prediction and the expectation of the model. The precision rate of normal comments is low, but the recall rate is high, indicating that almost all normal comments are identified as normal and will not be reviewed as insulting ones. The precision of malicious comments and personal insulting comments are high, and the recall rates remain low, indicating there are some insulting comments are recognized as normal comments. The idea here is similar as our intention that we do not want to block too many impolite comments. If the comment is mild or ambiguous,

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0046

these comments should be remained. Therefore, this result could be meaningful in the real procession of surveillance in social media.

5. Conclusions

In conclusion, the application of artificial intelligence, especially deep learning, in surveillance of cyber violence speech not only has its social value, but also has operational feasibility. I believe that in the future, Chinese social media will strengthen the supervision of online violent comments, to create a more harmonious and healthier network environment.

References

[1] Regulations on Ecological Governance of network information content. (2021) Ecological Governance of network information. http://www.cac.gov.cn/2020-01/20/c_1581058057316205.htm

[2] Weibo User Report in 2020. (2020) 2020 Weibo User Report. https://data.weibo.com/report/

[3] HIT Stopwords. (2020) Stopwords Lists. https://github.com/goto456/stopwords

[4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[5] Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification * [C] // Neural Information Processing Systems. MIT Press, 2015, 27(6):649-657.