Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241268

Chinese offensive language analysis based on Bidirectional Encoder Representation Transformer (BERT)

Haomin Liu[1,4,5], Xiaozhou Wen[2], Kangkai Yan[3]

[1]Institute Computer Science and Engineering, Southeast University, Wuxi, 214135, China
[2]Institute of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China
[3]Institute of Automation / College of Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

[4]213201355@seu.edu.cn
[5]corresponding author

Abstract. Offensive language detection is important in maintaining civilized social media platforms and training language models. However, this task remains to be developed in Chinese due to the lack of datasets. For this reason, in this paper, we introduce the Bidirectional Encoder Representation Transformer (BERT) to construct a Chinese Offensive Language Detection (COLD) model. We find that the race category has similar representation features as the region category. We investigate both the lack of training set and data poisoning attacks. Specifically, first, specific labeled data is analyzed for its role in the dataset. In addition, the experiments investigate the effect of the type of attack on the results. Second, the BERT model is used to test the results. BERT is a context-based embedding model that generates word embeddings based on context. In addition, the encoder understands the context of each word through a multi-head attention mechanism. The experimental results show that the model has the ability to handle handicapped data. The training results for the defective data were not as good as for the original data, and the model was more unstable for the training set with racial labeling and the test set with regionality. This study evaluated the importance of theme as a detection label. Poisoning attacks in BERT by changing labels were found to improve the accuracy of the test. This study can provide valuable thoughts to the community.

Keywords: Offensive Language Detection, Chinese, COLDataset, BERT

1. Introduction

Offensive language detection task plays an important role in maintaining social platforms and promoting civilized communication [1]. With the rise of large-scale language models, the safety issues due to offensive generation continue to be exposed, attracting widespread attention from researchers and pushing the research boom on this task to new heights. To tackle the problem of offensive language detection, a reliable benchmark is a needed basis to accelerate in-depth research. People can freely engage in discussion on social media platform. However, there are many problems nowadays, such as online harassment. A project, called Perspective, is carried out by Google and Jigsaw, which uses

DOI: 10.54254/2755-2721/45/20241268

machine to automatically detect toxic language [2]. However, most of works focus on English while few for Chinese. With the development of social media, online harassment becomes one of critical social problems [3]. A survey of automated abuse detection methods, investigated by the Natural Language Processing (NLP), highlights main trends of harassment in United States [4]. The damage caused by personal attacks to online discourse has prompted many platforms to try to suppress this social phenomenon. However, it is still very difficult to understand the prevalence and impact of offensive language on the Internet platform. [5]. Compared with English, Chinese has richer semantics and more complex understanding of text content, which makes it more difficult and challenging to identify Chinese offensive language. In daily communication, the standardization and rationality of language are often the main points that people care about. The quality of language can affect the results of communication and the feelings of both parties. On Chinese social platforms, verbal abuse, verbal violence and other types of aggression often occur in the discussion of racial, gender and regional issues. Therefore, it is necessary to identify Chinese offensive language. If the platform can effectively identify such language and shield it immediately, it can effectively prevent the spread and influence of offensive language. The identification of Chinese offensive language helps to create a harmonious communication environment and social atmosphere.

To solve the issue of Chinese Offensive Language Detection (COLD), COLDataset is proposed with 37,480 comments with binary aggressive labels and covers diverse topics of race, gender, and region. With the proposed dataset, its limitation has not yet been thoroughly studied. The purpose of COLDataset is to explore the factors that trigger the aggressiveness of the language and the kind of model used to train disabled data. Diving deeper into these questions will facilitate understanding the dataset and pointing out some disadvantages while training model. Besides, Bidirectional Encoder Representations from Transformer (BERT) model is adopted to test the result [6]. BERT is a context-based embedding model, which can generate word embedding according to the context and give different embedding vectors to the words in the statement. BERT is based on the Transformer model, but is only the encoder in it, when the sentence is fed into the Transformer, the encoder understands the context of each word through the multi-head attention mechanism, and then outputs the embedding vector of each word [7][8]. In this experiment, the language we use is Chinese, and the same Chinese words may have different meanings in different languages, so we use BERT, which has certain advantages for the training of the model compared with other models [9]. And the BERT model in the experience is also pre-trained in Chinese. We add labels to the training set to distinguish between aggressive language related to region and race, and trained the model a lot. During the test, we also test the accuracy of the model by changing the label in the test set. With COLDataset, we check the performance of deleting one exact topic comments, to investigate their degree of dependence on a particular labeled data. Besides, we also set the topics as label, and test the ability to process poison data. The experiment result shows that the model's ability to deal with disabled data. The training results for the defective data are not as good as the original data, and the models where the training set is racially labeled and the test set is regional are more unstable. This study evaluates the importance of topic as a label to detect. It is found that the poison attack by changing the label in BERT not only does not reduce the accuracy of the test, but also improves it.

2. Methodology

2.1. Dataset Description and Preprocessing

The dataset used in this study, called COLDataset, is a comprehensive dataset specifically designed for training and evaluating models for Chinese offensive language detection [10-12]. It serves as a benchmark resource to address the growing concern about offensive language and hate speech in Chinese online communities. The COLDataset, containing quite a number of comments with binary aggressive labels and covers diverse topics of race, gender, and region, comprises a diverse collection of text samples containing offensive content, including vulgar language, hate speech, and abusive remarks. The dataset provides annotated labels indicating the presence or absence of offensive language

DOI: 10.54254/2755-2721/45/20241268

in each sample, enabling the development and evaluation of offensive language detection models. With its carefully curated and annotated data, the COLDataset offers researchers and developers a standardized and reliable benchmark for training and assessing the performance of offensive language detection systems in the Chinese language. It facilitates the comparison of different models and algorithms, fostering advancements in the field of natural language processing and promoting the development of effective strategies to combat offensive language online.

By leveraging the COLDataset, researchers can analyze patterns, perform experimental evaluations, and develop state-of-the-art models to accurately detect and mitigate offensive language in Chinese text, thus promoting a safer and more inclusive online environment. After looking at the dataset, we find that data labeled race had similar performance characteristics to data labeled region. Therefore, we divide the whole training set into the training set containing only the label race and another training set containing only the label region, whose offensive labels remain unchanged. The test set is treated the same as the training set. When we test, we perform pairwise training to verify the correlation between the two labels and the effect of the type of offensive language on the prediction of offensive labels by training the model with the training set containing only one label and then validating the results on the test set containing only the other label.

2.2. Proposed Approach

The overall process of this study is as shown in the figure 1, mainly including Determine research direction, Data cleaning, Model training, Result analysis, adjust the process, and Obtain conclusions.

Specifically, first, search for relevant literature, select and analyze the degree of correlation between aggressive vocabulary based on the adaptability and correlation of the model. Select the Bert model and other relevant software for this model analysis. Second, classify data based on its labels. Divide the dataset into two parts based on region and race, training and testing sets that only contain region and race. In addition, change the label of all data containing one region to race, and vice versa to form the corresponding dataset. Third, conduct matching training on the divided dataset. Match the training set that only includes regions with the test set that only includes races for testing, and the other two are the same. Change the label of a training set with the same name and a region as the label to race, only including the test labels on the test set with that name. Same as the corresponding race. Fourth, conduct data analysis based on the obtained results, and if the results meet expectations, conduct a conclusion analysis on the degree of correlation. If the results obtained do not match the expectations, the dataset is trained and adjusted, and a new round of model training is conducted until the results roughly match the expected results. Final, conduct final data integration to determine the degree of correlation between regional and racial aggressive vocabulary.

DOI: 10.54254/2755-2721/45/20241268

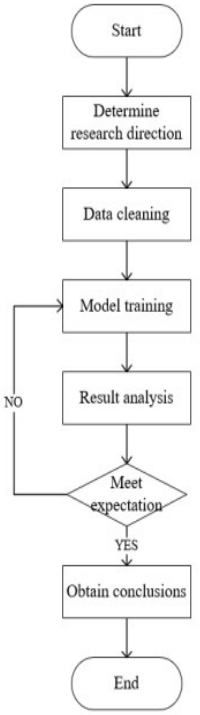

Figure 1. Overall flow chart

2.2.1. BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a method of pre-training language representations, come up by the google artificial intelligence (AI) language in October, 2018. An universal model with language understanding was trained on a large corpus, such as Wikipedia, and was mainly used in NLP task, such as question answering. By recognizing the context of each layer of content, BERT is able to represent unlabeled text bidirectionally with training. Unlike other model systems, BERT is the first unsupervised and deeply bidirectional system for pre-training NLP. Moreover, the concept of BERT is easy to understand and has strong ability in execution. BERT is a pre-traied language model that all developers can inherit it directly. Compared with Transformer, BERT, deep and shallow, has a better effect. This is the first convincing work to prove that extending to extreme model size can also lead to huge improvements in very small-scale tasks, provided that the model has been fully pre-trained.[9] One of BERT's innovations is the bidirectional Transformer. BERT automatically has bidirectional encoding and powerful feature extraction capabilities by referencing the Encoder module in the Transformer architecture and removing the Decoder module. So BERT emphasizes understanding language more than just generating language.

2.2.2. COLDataset

The emergence of pre-trained language models makes many generation tasks achieve good results. However, because these models are trained on large-scale data, they inevitably learn some biased or offensive content, resulting in no guarantee of security when the real scene is online. In order to solve the problem of offensive language in the system, dictionary rules alone are not enough, so many people have started language toxicity studies, and several data sets have appeared, such as WTC[5], OLID[10], BAD[11], and RealToxicPrompts[12]; However, these data are English data sets, which cannot be directly used in Chinese tasks, even if the use of machine translation methods, translated into Chinese, but language habits, language expression, data quality cannot be guaranteed. The development of detecting Chinese offensive language automatically was limited by the detoxification that relies mostly

DOI: 10.54254/2755-2721/45/20241268

on the blacklisting mechanism in online communities and language model generations. COLDataset is the first publicly available data set on abusive language in Chinese, covering topics such as race, gender and region. We changed labels in dataset to conduct experiments for different purposes.

2.3. Loss Function

Choosing the suitable loss function is an essential part in the training of deep learning models. For this classification task, the Cross-entropy loss function is optimal due to its effectiveness in multi-class classification problems. The Cross-entropy Loss is especially used in classification tasks because it measures the difference between the predicted value and the true value. The formula for the Cross-entropy Loss is as follows:

$$CrossEntropyLoss = -\sum(l * log(p) + (1 - l) * log(1 - p)) \quad (1)$$

In the above formula, l represents the true class label (0 or 1), and p represents the predicted class probability. The Cross-entropy Loss function computes the logarithmic loss for each class separately and sums them together. It penalizes the model more when the predicted probability diverges from the true label. A lower loss value indicates better alignment between predicted and true class probabilities. Cross-entropy Loss is often used in multi-class classification problems where each data sample belongs to one of the multiple classes. It provides a continuous and differentiable loss function that can be used to train models through gradient descent optimization. The goal of using Cross-entropy Loss is to optimize model parameters to minimize the loss, enabling the model to make accurate predictions by maximizing the likelihood of the true class for each input sample.

3. Result and Discussion

3.1. Normal circumstance

The analysis is mainly carried out from the macro and micro, as well as the reasons for the mutation of some graphs, to find the rules between the offensive languages. This is the topic of the experiment, assessing offensive language, and finding the relationship between the two by training the race and region related data sets and testing the normal data sets. TrainRaceTestRegion is a training race data set to test regional data sets. The results show that the race-labeled offensive language training set can respond correctly to the corresponding regional data set after training with the corresponding bert model. Similarly, TrainRegionTestRace is a training region data set and a test race data set. The results show that the offensive language training set labeled with region can respond correctly to the corresponding race data set after training with the corresponding bert model. Normal.csv is used for reference. The relationship between race and region obtained through Normal training process is used to prove the final conclusion.

As shown in the figure 2, in terms of the overall trend, the data sets of Normal and TrainRegionTestRace have higher accuracy and are relatively stable on this interval model. The dataset of TrainRaceTestRegion decreased sharply when the model number was about 50, and the accuracy and stability of the Models in the other regions were also relatively high. If the sudden decline in accuracy of TrainRaceTestRegion is ignored, the accuracy of Normal is the highest in the three cases, followed by TrainRaceTestRegion, and TrainRegionTestRace is the lowest. The models where the training set is racially labeled and the test set is regional are more unstable. The sudden decline in the accuracy of TrainRaceTestRegion may be due to an accidental network depth problem, and more repeated training may eliminate such mutants. For TrainRaceTestRegion and TrainRegionTestRace, the overall process in the middle is consistent, and the final data result image should theoretically be similar. As shown in the figure above, when the models were compared, the two were indeed in a similar trend for training and testing on the whole. However, when the model was around 50, the TrainRaceTestRegion was mutated, resulting in the accuracy of the two models after 50 losing the previous similar trend, and there was also a big difference in accuracy. In general, among the three, Normal.csv has the highest accuracy, TrainRaceTestRegion has moderate accuracy, and TrainRegionTestRace has the lowest accuracy, which

DOI: 10.54254/2755-2721/45/20241268

is in line with the overall expected results. For some mutated points, such as Normal.csv models between 20 and 40, and TrainRaceTestRegion around 50, the overall data is relatively unstable, and there may be some deviation in the data sets in these places, or due to the network depth and bandwidth, repeated training may eliminate this mutation.

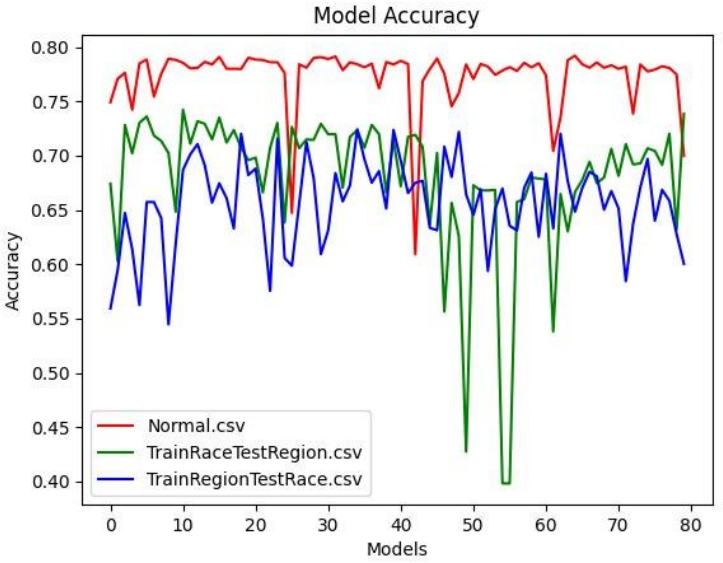

Figure 2. Normal circumstance Accuracy. Normal.csv represents the results of the original training set and the original test set. trainRaceTestRegion represents the results of the training set labeled with races and the test set labeled with regions. trainRegionTestRace represents the results of the training set labeled with regions and the test set labeled with races.

4. Conclusions

In this paper, we use the BERT model to construct a detector to solve the COLD problem. BERT is to generate word embeddings based on the context and assign different embedding vectors to the words in the utterance. An additional encoder understands the context of each word through the multi-head attention mechanism. Specifically, the BERT model in the experiments is pre-trained using Chinese. We added labels to distinguish offensive language related to region and race in the training set and trained the model extensively. The result of the model was tested by changing the labels in the test set during testing. Using COLDataset, we check the performance of deleting comments on an exact topic to investigate how much they depend on specific labeling data. The experimental results show that the model has the ability to handle disabled data.Meanwhile, this study evaluates the importance of topic as a detection label. In the future, we plan to create new datasets with a various kind of topics and use it to discover the correlations between each other, enabling natural language processing models to achieve better performance in detecting Chinese offensive language.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Deng J, Zhou J , Sun H ,et al.COLD: A Benchmark for Chinese Offensive Language Detection[J]. 2022.DOI:10.48550/arXiv.2201.06025.

[2] Hosseini H, Kannan S, Zhang B, et al. Deceiving google's perspective api built for detecting toxic comments[J]. arXiv preprint arXiv:1702.08138, 2017.

[3] M. Duggan, Online harassment. Pew Research Center, 2014.

DOI: 10.54254/2755-2721/45/20241268

[4] Mishra P, Yannakoudakis H, Shutova E. Tackling online abuse: A survey of automated abuse detection methods[J]. arXiv preprint arXiv:1908.06024, 2019.
[5] Wulczyn E, Thain N, Dixon L. Ex machina: Personal attacks seen at scale[C]//Proceedings of the 26th international conference on world wide web. 2017: 1391-1399.
[6] Koroteev M V. BERT: a review of applications in natural language processing and understanding[J]. arXiv preprint arXiv:2103.11943, 2021.
[7] Gillioz A, Casas J, Mugellini E, et al. Overview of the Transformer-based Models for NLP Tasks[C]//2020 15th Conference on Computer Science and Information Systems (FedCSIS). IEEE, 2020: 179-183.
[8] Acheampong F A, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches[J]. Artificial Intelligence Review, 2021: 1-41.
[9] Devlin J , Chang M W , Lee K ,et al.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[10] Zampieri M, Malmasi S, Nakov P, et al. Predicting the type and target of offensive posts in social media[J]. arXiv preprint arXiv:1902.09666, 2019.
[11] Xu J, Ju D, Li M, et al. Recipes for safety in open-domain chatbots[J]. arXiv preprint arXiv:2010.07079, 2020.
[12] Gehman S, Gururangan S, Sap M, et al. Realtoxicityprompts: Evaluating neural toxic degeneration in language models[J]. arXiv preprint arXiv:2009.11462, 2020.