Advances in Computer, Signals and Systems (2020) 4: 19-24
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2020.040104
ISSN 2371-8838

Online Active Learning for Offensive Language Detection

Yixuan Chai[a,*], and Guohua Liu[b]

School of Computer Science and Technology, Donghua University, Shanghai, China
[a]chaiyixuan@mail.dhu.edu.cn, [b]ghliu@dhu.edu.cn
**corresponding author*

Keywords: Offensive language detection, Online learning, Active learning.

Abstract: Recent success in deep learning-based offensive language detection, however, the off-line learning strategy cannot cope with the challenge of rapid growth and evolution of offensive language. In this paper, we introduce an online active learning method to offensive language detection model. Online learning makes use of the user's report feedbacks to continue training the detection model. In addition, active learning methods can filter out the mislabeled feedbacks to ensure the safety of the online learning process. Extensive experiments demonstrate that the proposed method can achieve promising results on the offensive response dataset.

1. Introduction

Offensive language refers to rude or disgusting language. It can be expressed as depreciation of someone or something. Offensive language usually appears in large numbers on social media (such as Twitter, MySpace, etc.) or website reviews (such as YouTube video sites, etc.), and if it is not cleared, it will seriously affect the user experience. The offensive language detection model faces the challenge of infringing on the rapid growth and evolution of offensive utterances in practical applications. Traditional offline training strategies cannot deal well with this problem. Therefore, the detection model needs to introduce a continuous learning mechanism. In this paper, we introduce an online active learning algorithm to the detection model. Online learning is a pipeline where users can report offensive language that is not filtered out by the detection model. The samples of the user report set are used to continue training the detection model. However, there may be normal language in the user report set. If we regard the offensive language as positive samples, then all samples in the user report set are ideally true positive (TP) samples. However, in actual applications, there may be users intentionally feeding back false samples, that is, there are a large number of true negative (TN) samples in the user report set. When a large number of TN samples appear in the report set, it will seriously affect the classification accuracy of the detection model. To deal with this problem, we introduce a confidence-based active learning method to filter the error feedback from the report set. Since the samples reported by users are normal ones predicted by the detection model, under the premise that the detection model is well trained, the higher the confidence value of the reported sample, the greater the possibility that the sample is mislabeled.

The main contribution of this work is as follows: our work proposed an online active learning method to address the evolution of offensive language problem, and to the best of our knowledge,

this work is the first one to make use of the active learning algorithm to filter out the mislabeled sample to ensure the safety of online learning process.

2. Related Work

2.1. Offensive Language Detection

Offensive language detection task is usually cast into the text classification work. In the earlier work, researchers optimized the feature extraction process according to the characteristics of the offensive language. For example, [1] observes that offensive words often appear behind the second-person pronouns. Therefore, they will give a higher weight to the words behind the window-sized second-person pronouns. However, the words within a fixed window size are not necessarily all offensive words. [2] further uses part-of-speech [3] to accurately extract features. First, the sentence is marked as a personal pronoun, copula, and other nouns. Then feature extraction is based on common infringement speech patterns, such as giving other nouns + personal pronoun after copula a higher weight: "He is a bad guy" gives higher weight to "bad". [4] introduces syntactic dependence to enrich the feature extraction of offensive language. Syntactic dependency can capture the relationship between two words in a sentence, such as the subject-predicate relationship between "he" and "is", and the verb-object relationship between "is" and "bad guy". In addition, [5] uses the topic model to further enrich text features. Using the topic model can make the model more focused on the classification of certain topics (e.g. ethnicity or religion), and can also alleviate the problem of semantic sparseness in the short text classification task. In recent years, deep learning methods have developed rapidly and achieved remarkable results in various tasks. The end-to-end training form can automatically extract text features, which makes researchers just need to focus on the model structure. [6] uses the long-short-term memory-based (LSTM) model with emotion and word embedding for hate speech detection. [7] uses a bi-directional recurrent network (Bi-RNN) with attention mechanism to detect cyberbullying. Bi-RNN is used to integrate bidirectional context information. The attention mechanism reflects the contribution of different words. [8] proposed a hierarchical attention networks architecture to capture the hierarchical structure of social media conversations.

2.2. Active Learning

Active learning [8] is one of the supervised learning algorithms. In traditional supervised learning, it must provide labeled data for training. However, in some scenarios, data labeling is time-consuming or difficult to obtain (such as CT images, etc.). Active learning is proposed to alleviate this problem. The active learning algorithm will select valuable samples from the unlabeled samples to be annotated by experts. Then, adding these labeled samples to the training set to train the model. Therefore, higher accuracy can be obtained even the training dataset is small. The key of the active learning method is the query algorithm [9]. The query algorithm determines which samples are valuable.

3. Proposed Method

The illustration of the online active learning algorithm is shown in Figure 1, there are three original contents, the first two records are offensive language. The detection model only filters the first offensive content and presents the rest to the user. The user can feedback the second piece to the pseudo dataset through the report pipeline. The samples of the pseudo dataset can filter out the mislabeled data through the active learning method.

Specifically, the active learning method is based on confidence. First, the user report samples are regarded as pseudo labeled data. Second, rank the D^P in ascending order by lc, where lc is defined as follows:

$$lc = p(y = normal|x, r, \theta) \quad (1)$$

Third, select samples from D^P to D^S where lc is less than a threshold δ. Finally, update the parameters of the detection model by D^S. In addition, we use BERT[10] as the detection model. The details of the algorithm are shown in Algorithm 1.

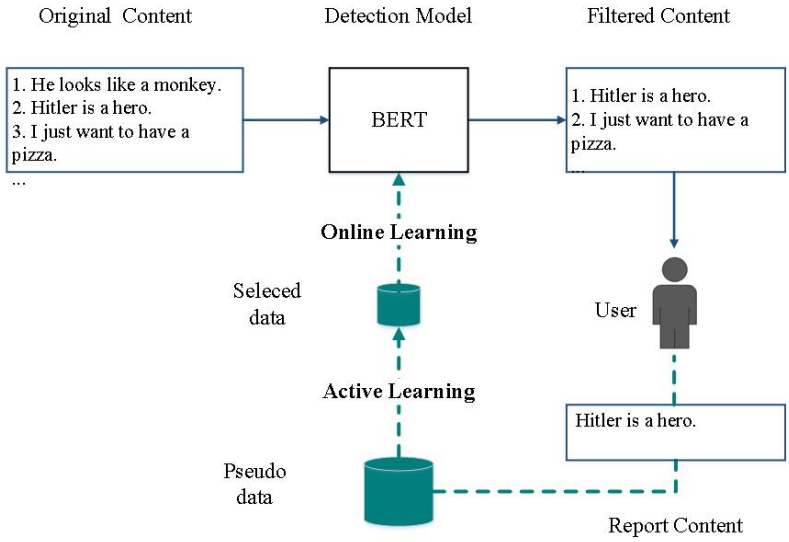

Figure 1: Illustration of online active learning of offensive language detection.

Algorithm 1:

Input: Pseudo-label set D^P, selected label set D^S, least confidence threshold δ, parameters of detection model θ.

1:**while** true **do**
2: *lcs* ← get least confidence value from D^P based Eq.(1)
3: **for** *sample* **in** D^P **do**
4: **if** *sample.lc* <δ **do**
5: add *sample* to D^S
6: delete *sample* from D^P
7: **end for**
8:Update θ with D^S
9:**end while**

Since the contents shown in the users are filtered out the $y = offensive$ samples by the detection model, all D^P samples are classified as $y = normal$ by the detection model. Under the premise that the detection model is well trained, the smaller $p(y = normal|x, r, \theta)$ means that the model is more uncertain about the classification result. The larger $p(y = normal|x, r, \theta)$ means that the model has more confidence in its own judgment. In this case, the user's feedback is more likely to be a TN sample (mislabeled).

4. Experiments

4.1. Datasets and Experiment Settings

The experiment uses offensive response dataset for evaluation. The dataset contains 110K chat records, 8% of them are offensive language. The dataset can available online[11].

We use the default vocabulary table of the BERT, the vocabulary size is 21128. the BERT model is initialized with Google BERT-base-Chinese parameters. The model has 12 heads and 12 self-attention layers. The hidden layer dimension of each attention layer is h=768. Word vector dimension E=756, the active learning rate is 5e-5, active learning sampling size $K$is 20% of the training dataset size. The code can available online[12].

4.2. Experimental Results

4.2.1. The Contribution of Online Active Learning to Classification

In order to evaluate the improvement of the classification accuracy of the online active learning algorithm, this section first selects randomly 20% of the training set and uses Stochastic Gradient Descent (SGD) training to simulate the first stage of offline supervised learning, the remaining 80% to simulate the online learning training process. 80% of the data will be ranked from smallest to largest according to the least confidence value. In the online learning process, active learning and SGD training methods will be compared to improve the model.

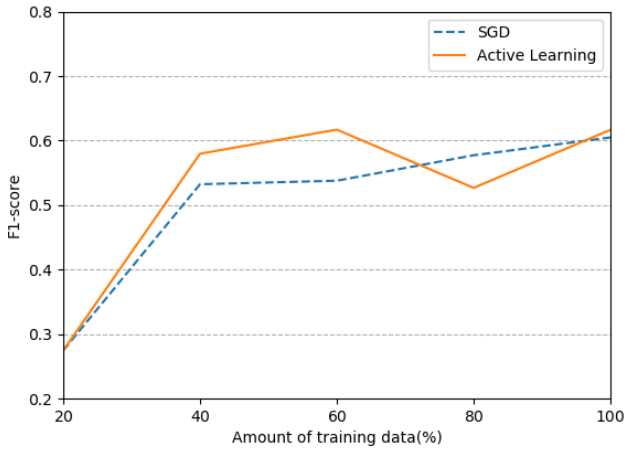

Figure 2: F1-score changes with the online active learning process

The experimental results are shown in Figure 2. Compared with the SGD algorithm, the active learning algorithm obtains a higher F1 value in the first 60% of the training data, and the F1 value in 80% of the training data is lower than the SGD algorithm. It shows that the samples with low confidence value in active learning have much improved classification accuracy, but the samples with high confidence value have little improvement in classification accuracy.

4.2.2. The Detection of Mislabeled Samples By the Least Confidence

In order to verify the effect of active learning on mislabeled samples detection, we use samples predicted as FN in the test set (785 in total) to simulate offensive contents that are not filtered by the detection model. Then randomly add the same number of TN samples to simulate the normal

contents. Then use the least confidence value to sort the samples in ascending order. The goal of the experiment is that the top 50% of the data does not contain TN samples.

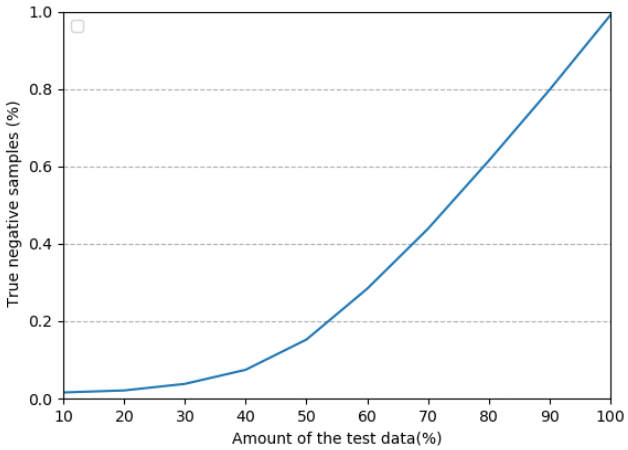

Figure 3: True negative samples distribution on least confidence ranked data

The experimental results are shown in Figure 3. 15.23% of the top 50% of the data are TN samples. In the case of random selection, the first 50% of the data should have 50% TN samples. Therefore, it can filter out mislabeled samples (TN sample in the experiment) by the least confidence value.

5. Conclusions

In this paper, we introduce an online active learning method to offensive language detection model. Online learning makes use of the user's report feedbacks to continue training the detection model. In addition, active learning methods can filter out mislabeled feedbacks to ensure the safety of the online learning process. Extensive experiments demonstrate that online active learning can improve the classification accuracy, and the active learning algorithm can filter out mislabeled samples.

References

[1] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in Dutch-Belgian Information Retrieval Workshop, DIR 2012, 2012, pp. 23–26.

[2] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 468–469.

[3] M. Dadvar, D. Trieschnigg, and F. De Jong, "Expert knowledge for automatic detection of bullies in social networks," in Belgian/Netherlands Artificial Intelligence Conference, 2013, pp. 57–63.

[4] Y. Chen, "DETECTING OFFENSIVE LANGUAGE IN SOCIAL MEDIAS FOR PROTECTION OF by," no. December, 2011.

[5] J. Li, Y. Song, C. Gao, M. R. Lyu, and I. King, "Topic Memory Networks for Short Text Classification," arXiv Prepr., vol. arXiv:1809, 2018.

[6] A. Cimino and F. Dell'Orletta, "Hate Me, Hate Me Not: Hate Speech Detection on Facebook," 2017.

[7] A. Zhang, B. Li, S. Wan, and K. Wang, "Cyberbullying Detection with BiRNN and Attention Mechanism," in Machine Learning and Intelligent Communications, 2019, pp. 623–635.

[8] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 1994, vol. 29, no. 2, pp. 3–12, 1994.

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1992, pp. 287–294.
[10] M. C. Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2017.
[11] https://github.com/chaiyixuan/Offensive-Responses-Dataset
[12] https://github.com/chaiyixuan/onlineAL