Assignment 2: Abusive Language Classification

Wen Xie
University of Houston
wxie5@uh.edu

Abstract

Abusive language detection is a challenging but important work for online social media community. Abusive language is a quite general term, which includes kinds of subtasks. In this study, we try to classify comments from Facebook into three classes. They are non-aggressive (NAG) comments, overly-aggressive (OAG) comments, and covertly aggressive (CAG) comments. Our Training dataset includes 12,000 comments from different Facebook user ID while the validation data-set is comprised of 2,000 comments. Two bidirectional Long Short Term Memory (LSTM) network with pre-trained embeddings is trained to do the classification. The model performs well with macro-average F-1 score 0.5443.

1 Introduction

Social media platforms become a necessary part of everyday life, which takes an important role in shaping people's minds(Park and Fung, 2017). Our familiar platforms such as Facebook and Twitter have responded to criticism for not doing enough to prevent abusive language on their sites by instituting policies to prohibit the use of their platforms for attacks on people based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence towards (**?**). To protect the disadvantage groups, successful detection of online abusive language becomes more and more important.

Previous researchers try to increase the accuracy of the challenging work from a variety of ways. Traditional commercial methods make use of black-lists and regular expressions to detect the abusive language. However these measures fall short when contending with more subtle examples. To overcome these defaults, Nobata et al. (2016) consider a wide range of features in his article. The author used character N-gram features to model the types of conscious or unconscious bastardizations of offensive words. Specialized linguistic features (i.e. number of politeness words, number of punctuations) are included to cope with the noisy social media data. Besides, the authors also select syntactic features (i.e. Part-of-Speech tag) and distributional semantic features which includes pre-trained word embeddings, word2vec embeddings, and comments embeddings. Park and Fung (2017) make innovations on the methodology part of detecting abusive language and subtasks by considering CharCNN, WordCNN and HybridCNN. Apart form the model selection, the authors split the detection procedure into two steps.

In this study, the Recurrent Neural Network (RNN) models are utilized to classify the online comments. The structure of this report is as follows. In section one, it's introduction about abusive language. Methodology is documented in section two. After that, experiment results and interpretations are shown in section three. Conclusion follows at last.

2 Methodology

In this part, two kinds on content will be introduced. The first part is about the feature extraction. The second part is about the model selection.

2.1 Word Embeddings

The labeled data-set include three types of abusive language labels that are Non-aggressive (NAG) comments, Overly aggressive (OAG) comments, and covertly aggressive (CAG) comments. Table 1 summarize the overall information of the three labels. It's clear that the data-set is quite imbalanced. The comments of label CAG account for the minority of the data-set.

To figure out potential impacts of the imbalanced data-set on the model performance, oversampling methods are used first. Synthesis Minority Oversampling Technique (SMOTE) is selected to balance the train data-set. Consequently, the distributions of the three classes are evenly. The training data-set also includes more samples, which could enhance the performance of the model especially when detecting the minority class.

To increase the accuracy, pre-trained word embeddings based on Global Vectors for Word Representation (GloVe) dataset(Pennington et al., 2014)

Name of label	Number of comments
NAG	5052
OAG	4240
CAG	2708

Table 1: train data-set labels.

are utilized. There are 400,000 word in total in the data-set while our training data-set includes 1,126 different tokens.

2.2 Recurrent Neural Network

Recurrent Neural Network is good for processing sequence data for predictions. However, it suffers from short-term memory. If a sequence is long enough, they'll have a hard time carrying information from earlier time steps to later ones. LSTM and Gated Recurrent Units (GRU) are created as the solution to short-term memory. They have internal mechanisms called gates that can regulate the flow of information. LSTM and GRU are used in state of the art deep learning applications like speech recognition, speech synthesis, and natural language understanding. Besides, to consider the word relationship of the sequential data, bidirectional network is innovated. In this study, two layers of bidirectional LSTM is utilized after the embedding layer.

3 Experiment and Result

In this section, experiments including data pre-processing, word encoder, word embedding, and training the RNN model are documented first. Then the performance of the model is evaluated.

3.1 Data Pre-processing

Three same distribution data-sets about Facebook data are given in this task. Training data-set includes 12,000 different comments and three kinds of labels while validation data-set includes 2,000 samples. The test data-set includes 1001 samples without labels. First, the maximum length of the comments is found to be 1,126. And there are 26,359 unique tokens contained in the training and validation data-sets. Second, each comment is encoded into one vector in which each distinct word is represented by one unique digit. The length of each comment vector is 1,126. There are zero paddings if the comment does not have 1,126 word.

After encoding the comments, oversampling is conducted to balance the training data-set. SMOTE method is selected in this study. After the over-sampling step, each class contains 5052 samples. As a result, the training data-set includes 15,156 samplings. It's believed that balanced data-set can make the performance better. Pre-trained word embeddings GloVe including 400,000 words is downloaded from the official website. We search each word included in our training data-set in the GloVe data and get the weight of the corresponding word. At last, the weights matrix is obtained.

3.2 Training RNN model

Word embedding layer is followed by two layers of bidirectional LSTM. And the it's the one layer of dense layer. The classification output is regularized by one softmax layer. The embedding layer is set to non-trainable. Figure one shows the details of the whole model.

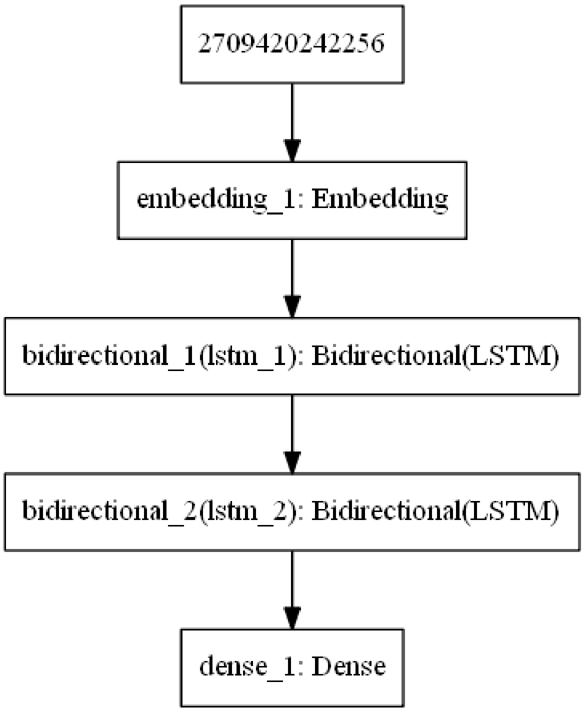
2709420242256
embedding_1: Embedding
bidirectional_1(lstm_1): Bidirectional(LSTM)
bidirectional_2(lstm_2): Bidirectional(LSTM)
dense_1: Dense

Figure 1: Model structure

The training optimizer is set as Adam optimizer. Batch size is 256 considered the large number of training samples. Sparse categorical cross entropy loss is used to evaluate the performance. Patience with 5 epoch is set to early stop the training procedure if the validation loss does not decrease in a five consecutive epochs. The training step stops at the 27 epoch. And with the trained model, the performance is evaluated based on the validation data-set. Table 2 summarize the performance results.

Class	**precision**	**recall**	**f1**	**supp**
NAG	0.70	0.63	0.66	815
OAG	0.46	0.48	0.47	485
CAG	0.48	0.52	0.50	700
accuracy			0.56	2000
macro avg	0.55	0.54	0.54	2000
weighted avg	0.57	0.56	0.56	2000

Table 2: validation results.

NAG	**OAG**	**CAG**
511	102	202
57	233	195
158	175	367

Table 3: confusion matrix.

3.3 Prediction results

With the best model, abusive language category is predicted. Totally, there are 1001 comments and predictions.

4 Conclusion

From the table 2, we can see that the overall performance macro-average F-1 score is 0.54. The F-1 score of Non-aggressive comments is highest, which is reasonable given the training samples of non-aggressive comments are the most. The overtly aggressive classification results are quite worse than the other two even through the oversampling method is used to balance the training data-set. In this study, the complicated and useful features such as lexicon, syntactic, linguistic features are not considered. However, given the performance results at present, we can try to add these features together to increase the accuracy of the model.

In the future, more useful features should be considered such as chunking features and parsing features.

References

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.