Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network

Abstract—Cyberbullying can have a deep and long lasting impact on its victims, who are often adolescents. Accurately detecting cyberbullying helps prevent it. However, the noise and errors in social media posts and messages make detecting cyberbullying very challenging. In this paper, we propose a novel pronunciation based convolutional neural network (PCNN) to address this challenge. Upon observing that the pronunciation of misspelled words in informal online conversations is often unchanged, we used the phoneme codes of the text as the features for a convolutional neural network. This procedure corrects spelling errors that did not alter the pronunciation, thereby alleviating the problem of noise and bullying data sparsity. To overcome class imbalance, a common problem in cyberbullying datasets, we implement three techniques that include threshold moving, cost function adjusting, and a hybrid solution in our model. We evaluate the performance of our models using two cyberbullying datasets collected from Twitter and Formspring.me. The results of our experiment show that PCNN can achieve improved recall and precision compared to baseline convolutional neural networks.

Dataset:

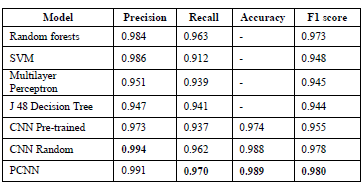
1) 1313 messages from twitter

2) 13,000 messages from formspring.me

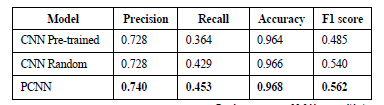
Methodology:

Convolutional neural networks (CNN), originally created for image processing, have performed very well in natural language processing (NLP), especially in sentiment analysis and question classification. Inspired by their powerful feature representation capability, flexible structure, and high efficiency for training using a GPU, we adopted CNN as the baseline classifier. To have a clear performance comparison between Point Convolutional Neural Networks (PCNN) and the baseline CNN, only one layer of convolution and max pooling was used with three different filter sizes. The sizes of the three convolutional filters were chosen to be 1, 2, and 3. The filter sizes were chosen based on how many consecutive words were necessary to detect bullying content. In addition to using neural network, classic machine learning classifiers were used too, such as Random Forest, SVM, Multilayer Perceptron which is also considered neural network and finally Decision Tree all had remarkable results.

Accuracy however wasn’t calculated to the Twitter dataset due to it being unbalanced. Approximately 55% of the words in the FormSpring dataset vocabulary cannot be found in the dictionary while only 15% of the words in the Twitter dataset are misspelled.



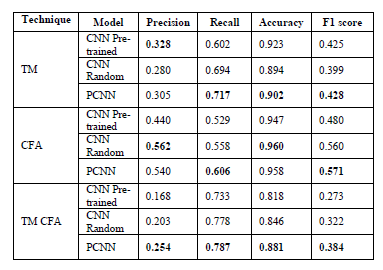
On the other hand the FormSpring dataset showed promising results.



Due to having an issue in dataset imbalance, three techniques was introduced in order to get a good reading on the results, they are threshold-moving, cost function adjusting, and a hybrid solution between the two.

HANDLING CLASS IMBALANCE ON THE FORMSPRING DATASET

It shows that all three techniques enhanced the recall at the cost of precision and even accuracy. Among them, CFA improved the overall classification performance the most, increasing recall and F1 score without hurting accuracy. Moreover, PCNN obtained the highest recall and F1 score than others when using cost function adjusting.



HANDLING CLASS IMBALANCE ON THE TWITTER DATASET

These techniques slightly improved recall, and the combination of TM and CFA performed the best out of the three. Furthermore, TM CFA PCNN can improve the overall performance and outperforms the two baseline CNN models. However, the improvement is insignificant since the degree of class imbalance in the Twitter dataset is low and the recall is already very high. Thus, these techniques need to be evaluated on a noisier and more imbalanced dataset.

