Abstract

Malicious and offensive expressions in social media have become a serious social problem in recent years. To protect people from the damage caused by such offensive expressions, technology to automatically detect offensive posts in social media is in a huge demand. Supervised learning of a model to classify whether a text is offensive or not requires a labeled dataset consisting of offensive and non-offensive texts. Most previous studies constructed a dataset by collecting offensive texts using a pre-defined list of offensive words or expressions and manually annotating them with gold labels. Since it is rather difficult to create a comprehensive list of offensive words and expressions, however, texts including implicit offensive expressions or expressions containing unknown offensive words may not be included in the dataset, and a model trained from it may fail to classify such texts correctly. In addition, manual annotation of a large number of offensive texts requires much human labor.

This study proposes a method to automatically construct a dataset annotated with offensive/non-offensive labels that includes a wider variety of offensive expressions, and to train a model to predict intensity of offensiveness of a text using this dataset. Specifically, since bashing in social media, where users are accused of their immoral or unethical behavior by others, is likely to receive offensive responses, replies to a bashed post are collected as offensive texts. This method might be able to collect a wide variety of offensive expressions, since collected offensive texts are not restricted to ones that contain a limited number of offensive words. In addition, the quality of the constructed dataset is improved by automatically finding wrong offensive labels and fixing them.

First, users who have a large number of followers and often post about various topics related to bashing on Twitter (currently X) are selected. From their posts, tweets that receive a particularly large number of responses (called "bashed tweet") are manually selected, and tweets replying to those bashed tweets are collected as offensive texts. Similarly, non-offensive texts are obtained by collecting tweets replying to "non-bashed tweet", which is considered unlikely to attract criticism. Then, the validity of the constructed dataset was manually verified. It was found that only about 30 percent of the responses to bashed tweets were actually offensive texts. Therefore, we propose methods to obtain a more accurate model for prediction of the offensive intensity by iteratively correcting errors in the dataset and training the model. Our method consists of two kinds of modules: "initial data construction" and "model training".

The methods of "initial data construction" are methods to construct initial training data that is used at the beginning of the iterative training of the model. We propose the following three methods. Method i(intact) uses the aforementioned

dataset consisting of responses to bashed tweets and non-bashed tweets without any modification. Method ii(PtoN) calculates the similarity between tweets by Sentence BERT and modifies the labels of responses to bashed tweets that are similar to responses to non-bashed tweets from positive (offensive) to negative (non-offensive). Method iii(scoring) assigns an offensiveness score to each text, which is calculated by considering the bias of the frequency of the word bi-grams in the sets of responses to bashed tweets and non-bashed tweets.

As for "model training", three methods are proposed. In all these methods, BERT is used as the model for prediction of offensive intensity. Method A(vanilla) is a method to fine-tune a BERT model only once using an initial data. Method B(bootstrap) incrementally increases labeled samples in a bootstrapping manner. Method C(labeling) repeats re-estimation of offensive scores of all texts in the dataset and training the model using the updated dataset alternatively until the predicted offensive scores are converged.

In the experiments, we compare nine proposed methods obtained by combining three methods of initial data construction and three methods of model training, as well as a baseline model. The baseline model is a fine-tuned BERT model using a dataset constructed as follows. A list of 38 offensive words suggested by the previous studies was created, then the tweets including one of those words were excerpted from our dataset as offensive texts, while the same number of tweets not including any pre-defined offensive words are excerpted as non-offensive texts. As test data, we use 273 tweets annotated by three males in their 20s. We use "BERT base" and "BERT large" as pre-trained BERT models. The task in this experiment is a classification problem to determine whether a text is offensive or not, and the evaluation criteria are ROC-AUC and PR-AUC.

In the evaluation of the models using BERT base, the best method was the combination of Method ii(PtoN) and Method C(labeling) with respect to ROC-AUC and the combination of Method i(intact) and Method C(labeling) with respect to PR-AUC. Comparing methods of initial data construction i, ii, and iii, Method i and Method ii performed better than Method iii. Comparing the methods of model training A, B, and C, Method B(bootstrap) entirely poorly preformed; their ROC-AUC and PR-AUC were worse than the baseline. It indicated that Method B was ineffective. Seeing the PR curve, the proposed methods achieved higher precision than the baseline when the recall was low, lower when the recall was medium, and became higher again when the recall was high. The fact that the proposed methods outperformed the baseline when the recall was high indicated that the proposed methods were capable of detecting various offensive expressions. However, it was unclear why the proposed methods performed better even when the recall was low.

In the evaluation of the models using BERT large, the combination of Method i(intact) and Method A(vanilla) achieved the best ROC-AUC and PR-AUC. In a comparison of methods of initial data construction i, ii, and iii, on the one side, Method i performed the best in both ROC-AUC and PR-AUC, on the other hand, Method iii performed worse than the other methods. These results were consistent with the experiment using BERT base. Therefore, we can conclude that the methods of initial data construction can be ordered as i>ii>iii in terms of their effectiveness. The performance of Method A and Method C for model training were almost the same, and Method B poorly performed in the experiment using BERT base. Therefore, the methods of model training can be ordered as A$\simeq$C>B. Comparing the proposed methods to the baseline, both ROC-AUC and PR-AUC of our methods i$\times$A and method i$\times$C were higher than the baseline.