SI630 Project Update

Zhijun Cai

Abstract

This paper presents my approach to classify whether a tweet is offensive or not in a multilingual data set with tweets in five languages: Arabic, Turkish, English, Greek and Danish. To solve this problem, I first did some preprocessing on the raw data to remove unnecessary information like emojis, URLs, Phone numbers and etc, taking into consideration of uni-gram and bi-grams on the offensive phrases. My baseline is Logistic Regression classifier on Arabic tweets data set and the baseline has achieved 0.878 accuracy while a relatively low macro average F1 score at 0.72.

1 Introduction

Contemporary, the high accessibility to social mediaanonymity feature in commenting and more freedom of speech all contribute to more offensive comments online. Commenting behind the computer gives people more courage to say aggressive and offensive languages without being responsible for that. Cyberbully becomes a main issue on social media towards users. Because of the high accessibility to social media, those offensive tweets will have negative effect on the young generation. The detection of offensive tweets will build a better internet environment and protect social media users from being cyberbullied. This task is a binary classification problem and in experiment, I will try to use Multinormial Naive Bayes, Linear Support Vector, Random Forest Logistic regression classifiers and deep learning method like LSTM and BERT embedding. I vectorized my text using tf-idf vectorizor with uni-gram and bigram tokens, l2 norm distance and removed all the stop-words from five language. The outcome will be predicted as not offensive and offensive. The social media company and their users will interested in the outcome. The company can use the offensive language detection to block some intense posts.

2 Problem Definition and Data

The goal of this paper is to classify whether the tweet is offensive or not. The input data is lines of tweets and the output should be categorical: OFF(This post contains offensive language or a targeted (veiled or direct) offense) or NOT(This post does not contain offense or profanity.). Multinormial Naive Bayes, linear support vector, randorm forest, logistic regression, LSTM classifier are implemented. For each model, I calculated the Macro F1 score and utilize the result yield by the method that has the best score. If the macro f1 score is above 50, it is considered a success.

The datasets were collected by OffensEval 2020 which contains five languages: Arabic, Danish, English, Greek and Turkish. All of the datasets have hierarchical annotations which is the same as the the Offensive Language Identification dataset (OLID). In this paper, I will mainly focus on Subtask A - Offensive language identification.

The Arabic data set has already been separated into training set and validation set. There are 6839 tweets in the training data set and 1000 tweets in the validation data set. Both data sets has three columns: id, tweet, and the label for whether it is offensive. However by further looking at the training data, I found that the number of tweets for two classes are imbalanced. 5468 tweets are not offensive and 1371 tweets are offensive. The number of not offensive tweets is much higher than offensive tweets which indicates that accuracy score is not a proper evaluation score here. The other four data sets are not separated into training and testing already, so I will split them randomly into 0 percent training and 20 percent testing sets. The Danish data set which is the smallest data set, has 2961

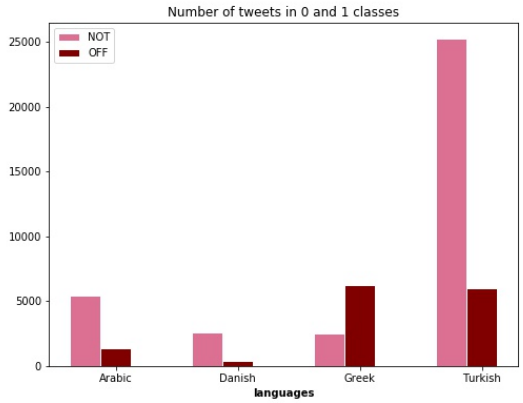

Figure 1: Imbalanced data between two classes

instances in which 2576 tweets are not offensive and 384 tweets are offensive. The turkish data set contains 31277 instances and 25231 of them are not offensive, 6046 are offensive. The Greek data set contains 8743 instances and 2486 are not offensive, 6357 are offensive. It is the only data set that the number of offensive tweets exceeds the number of not offensive tweets.

3 Related Work

In this section, I looked through some related work on offensive or aggressive detection done by other researchers. (Gómez-Adorno et al., 2018) implemmented logistic regression for detecting aggressive tweets in spanish. They have incorporated lingustic features and patterns and several types of n-grams and achieved a 42.85 f-score on aggressiveness class. (Chen et al., 2012) proposed the Lexical Syntactic Feature (LSF) architecture to detect and incorporated user's writing style, as well as some cyberbully language features. Their LSF framework achieves precision of 98.24 percent. (Razavi et al., 2010) applied multi-level classification in the detection of offensive language. They used complement Naive Bayes for the first level and multinomial Naive Bayes for the second level. The result did not predict the sentence well. (Symeonidis et al., 2017) has considered number of question marks, exclamation marks, number of capitalized words, number of elongated words as features. However, they found that number of exclamation marks under-performed as a feature. Symeonidis et al. (2017) noted that

> The three classifiers with the best results were the Bernoulli Naiıve Bayes, the Stochastic Gradient Descent (SGD), and the Linear SVC.The final step was to use the majority voting classification method that combines three different classifiers and outputs the class that the majority of them agreed.

4 Methodology

Because of the noise in the data set will affect the result negatively, I first pre-processed the data by removing unnecessary information like URL, emoji, phone number, stop words of five languages, special characters, and getting lowercase to extract useful tweet text for predicting. Then, I transformed my text to vector using TF-IDF embedding with a logarithmic form for frequency, minimum numbers of documents a word must be kept set to 5, a l2 norm distance to make sure all of the feature vector has an euclidean distance, and uni-gram or bi-gram considered. Then I trained the logistic regression classifier with response factorized to 0 and 1. 0 for not offensive and 1 for offensive.

Next I will implement other classifiers: multinormial naive bayes, Linear support vector classifiers, bagging model like Random forest, boosting model like Xtereme Gradient Boosting model. However, different languages have different semantic features. I will try to embed with different hyperparameters or try different embedding methods like bag of words or GloVe. To make things more efficient, I will build a voting machine which will train all kinds of classifier model that I want and vote out the one with highest marco F1 score for the specific input language data.

Last but not the least, because it is a multilingual context, I think it is also a good choice to implement BERT to help embed my data and perform models.

5 Evaluation and Results

	precision	recall	f1-score	support
0	0.87	1.00	0.93	821
1	0.94	0.34	0.50	179
accuracy			0.88	1000
macro avg	0.91	0.67	0.72	1000
weighted avg	0.89	0.88	0.85	1000

Figure 2: Baseline Logistic regression model score

The classification report of my baseline on the Arabic tweets is in Figure 2. Looking merely at the

accuracy score, I found that the classification accuracy rate is 0.88 which seems quite good. However, accuracy rate is not an appropriate evaluation metric in this case because of the imbalance data between two classes. In other words, if a model predicts all the instance to be not offensive, it will get a 0.79 percent accuracy rate. The recall for class 1 is 0.34 which is relatively low. The baseline has only caught 34 percent of the offensive tweets. There is still a lot to improve. I decided to using macro F1 score as my evaluation metrics because it takes in to account both accuracy and recall, which is a balance between the two.

6 Discussion

As discussed in the previous part that the high accuracy of my baseline is due to imbalanced data, there are still large percent of offensive tweets that I have not detected. I plan to investigate more on the semantic feature of offensive tweets, extract more features like pos-tags and try different embedding methods to include in my model to improve my recall for offensive tweets.

7 Work Plan

I am planning to build a voting classifier and try embed with different hyper-parameters, or use BERT embedding in the next few days. Then for each data sets I will train the model and select the best one based on the evaluation metrics. Next, I will visualize my results in graphs and other forms to make sure it is clear to understand. Finally, I will organize my graphs, results and insights in the paper and the blog. If I got extra time, I will try to detect whether the offensive tweets have targeted or not and what kind of target it aimed for.

Acknowledgments

If you got help from anyone or had substantive discussions, please acknowledge those people here and describe how they contributed. The work you do for your project should be entirely your own.

References

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, pages 71–80.

Helena Gómez-Adorno, Gemma Bel Enguix, Gerardo Sierra, Octavio Sánchez, and Daniela Quezada. 2018. A machine learning approach for detecting aggressive tweets in spanish. In *IberEval@ SEPLN*. pages 102–107.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*. Springer, pages 16–27.

Symeon Symeonidis, Dimitrios Effrosynidis, John Kordonis, and Avi Arampatzis. 2017. Duth at semeval-2017 task 4: a voting classification approach for twitter sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 704–708.

A Supplemental Material