AUTOMATIC DETECTİON OF INSULTS IN SOCIAL COMMENTARY

Gönül Şakar, Serra Nur Ünal
Department of Computer Engineering
University of Hacettepe, Ankara,TURKEY
E-mails: {gonul.sakar, serra.unal}@hacettepe.edu.tr

1. Introduction

Social media connects people from all over the world. However, due to the lack of rules in social media and the convenience of being anonymous, people began to insult,harm, harass, humiliate, threaten and bully each other in their comments. Analyzing and detecting these abusive comments can prevent the psychological violence on the platform so that a healthy discussion can take place.

It is necessary to mention the differences between concepts like hate speech, offensive, profane,abusive languages, and cyberbullying. According to Wikipedia, hate speech is defined as "any speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, gender, disability,sexual orientation, or gender identity." Hate speech does not need to contain abusive words. For example: 'Send all Muslims back to their own countries.' On the other hand, offensive language is the text which uses abusive slurs or derogatory terms. For example: 'Bitch it is none of your business.' In cyberbullying, the purpose is harassing individuals.For example: 'You came and ruined everything.' In our project, we focus on offensive language, ie insulting.

Detecting offensive language is difficult for a variety of reasons. Using keyword spotting is not a feasible solution because a word can be written in multiple ways such as 'Nigger' or 'Ni9 9er'. Using a blacklist is also useless, since the list must be updated regularly. Moreover, comments may include sarcasm.

2. Related Work

Various approaches have been made in order to handle the problem of hate language detection. One of them is Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach [1]. They proposed an approach that differentiate between hate speech and offensive language on Twitter platform. They used publicly available Twitter dataset by using Twitter API for fetching public user tweets. They combined three different datasets : Crowdflower [2, 3] which are labeled as "Hateful", "Offensive" and "Clean" manually and [4].They used n-gram weighted with term frequency-inverse document frequency (TFIDF) values as their features.They implemented three models: Logistic Regression, Naive Bayes and Support Vector Machines as classifier models. They found that Logistic Regression performs better among the three models for n-gram and TFIDF features after tuning the hyperparameters. They preprossed the data first by removing unnecessary contents like spaces,urls etc.Then split the data to train (70%) and test(30%) sets. Then they extracted n-grams and made normalization. They performed 10 fold cross validation and compared the three models.The results showed that Logistic Regression performs better with the optimal ngram range 1 to 3 for the L2 normalization of TFIDF.

Some of the other works propose the multi-level classification approach [5, 6]. In [5], researchers are focused on flame detection. Firstly they started with preprocessing step. Then they used three-level classification method for flame detection using Insulting and Abusing Language Dictionary. They used Weka to chose fast algorithms for all the levels. They used the Naive Bayes in the first level, Multinomial Updatable Naïve Bayes classifier for the second and for the last level rule-based classifier (named DTNB - Decision Table/Naive Bayes hybrid classifier) was used. In each level, they gained more accuracy than previous levels.

In [6], they started with preprocessing the data like [5] than they passed this preprocessed data to two types of classifiers: the lexicon-based classifier and the n-gram SVM classifier. The lexicon classifier uses the word lists and calculates a lexical score. The n-gram SVM classifier was used to reduce the dimension of the attributes. The outputs of the SVM and lexical score was combined with other attributes as one attribute set and these set become to input of the neu-

ral network. They used the Impermium Kaggle Dataset. As a result of, the [5, 6] gained more accuracy with multi-level classification method than existing techniques.

3. Dataset

As far as we can see, in the related works, they used datasets which are created by combining multiple publicly available datasets. We decided to use Kaggle dataset instead of combining more than one dataset. In 2012, Kaggle hosted Detecting Insults in Social Commentary [7] task. They released a dataset contains posts on adult topics like politics, employment, military, etc for this task. The data consists of a label column followed by two attribute fields (neutral comment, or insulting comment). In addition to that, we decided to use Wikipedia abusive language data set includes approximately 115k labeled discussion comments from English Wikipedia.

4. Methodology

After analyzing the paper "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach " [1] we decided to used the similar methodology in our project. We decided to implement Logistic Regression model because it performs better according to their results. In addition to that, we decided to implement SVM model. Because according to "A Survey on Hate Speech Detection using Natural Language Processing" [8] paper, as classifiers for this purpose mostly Support Vector Machines are used.

We decide to follow data preprocessing and cleaning (noise reduction, normalization, get rid of the case sensitivity, tokenization, stemming, lemmatization, discarding URLs, punctuations and stop words), feature engineering and scaling, model building and learning, analyze and evaluate the results steps which are common steps for project implementation.

References

[1] Aditya Gaydhani , Vikrant Doma , Shrikant Kendre and Laxmi Bhagwat. *Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach.*

[2] Dataset1, `https://data.world/crowdflower/hate-speech-identification`

[3] Dataset2, `https://data.world/ml-research/automated-hate-speech-detection-data`

[4] Dataset3, `https://github.com/ZeerakW/hatespeech`

[5] Reinald Kim Te Amplayo, Jason Occidental. *Multi-level classifier for the detection of insults in social media. Conference Paper : 15th Philippine Computing Science Congress*.

[6] Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. *Multi-level classifier for the detection of insults in social media. Conference Paper : 15th Philippine Computing Science Congress*.

[7] Dataset4, `https://www.kaggle.com/c/detecting-insults-in-social-commentary`

[8] Anna Schmidt, Michael Wiegand. *A Survey on Hate Speech Detection using Natural Language Processing*.