Identifying and Categorizing Offensive Language in Social Media
SemEval 2019 - Task 6
CSE556 Natural Image Processing Final Project Report
Group 12

Pranav Goyal
2017078
pranav17078@iiitd.ac.in

Tanish Jain
2017115
tanish17115@iiitd.ac.in

Vidit Jain
2017121
vidit17121@iiitd.ac.in

Abstract

We worked on the SemEval 2019 Task 6, i.e., Identifying and Categorizing Offensive Language on Social Media (Sem). We attempted 2 sub-tasks in this task, viz. **Sub-task A -** Offensive language identification, **Sub-task B -** Automatic categorization of offense types Transfer learning and domain adaptive learning have been applied to various fields including computer vision (e.g., image recognition) and natural language processing (e.g., text classification). One of the benefits of transfer learning is to learn effectively and efficiently from limited labeled data with a pretrained model. In the shared task of identifying and categorizing offensive language in social media, we preprocess the dataset according to the language behaviors on social media, and then adapt and fine-tune the Bidirectional Encoder Representation from Transformer (BERT) pre-trained by Google AI Language team

Keywords - Identifying Offensive Language, Offensive Language Analysis, Offensive Type Categorization, Language Processing

1 Introduction

Offensive language is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behaviour that many of them would not consider in real life. Their offensive remarks damage the social media website's reputation and also it can be hurtful for the person or group of persons on whom the offensive comment is targeted on. Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language to prevent abusive behaviour in social media. It has thus become a topic of interest for many researchers around the world. (Sem)

TIn SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (Zampieri et al., 2019b), the organizers collected tweets through Twitter API and annotated them hierarchically regarding offensive language, offense type, and offense target. The task is divided into three sub-tasks: a) detecting if a post is offensive (OFF) or not (NOT); b) identifying the offense type of an offensive post as targeted insult (TIN), targeted threat (TTH), or untargeted (UNT); c) for a post labeled as TIN/TTH in sub-task B, identifying the target of offense as individual (IND), group of people (GRP), organization or entity (ORG), or other (OTH). The three sub-tasks are independently evaluated by macro-F1 metric. The challenges of this shared task include: a) comparatively small dataset makes it hard to train complex models; b) the characteristics of language on social media pose difficulties such as out-ofvocabulary words and ungrammatical sentences; c) the distribution of target classes is imbalanced and inconsistent between training and test data. To address the problem of out-of-vocabulary words especially emoji and hashtags, we preprocess each tweet by interpreting emoji as meaningful English phrases and segmenting hashtags into space separated words.

One of the most effective strategies for tackling this problem is to use computational methods from the subject of Natural Language Processing, to identify offensive comments, aggression, and hate speech in user-generated content (e.g. posts, comments, microblogs, etc.). This topic has attracted significant attention in recent years as evidenced in recent publications (Waseem et al. 2017; Davidson et al., 2017, Malmasi and Zampieri, 2018, Kumar et al. 2018) and workshops such as AWL and TRAC.

2 Related Works

This paper (Effrosynidis et al., 2017) by Effrosynidis, Dimitrios and Symeonidis, Symeon and Arampatzis, Avi, analysed many pre-precessing techniques from many different research papers. This paper achieved an accuracy of about 60%. This paper also compares many studies. The related works in this paper have been described fully.

There have been many more research studies in this field. In (Tajinder Singh, 2016) , they analysed tweets' data which was full of abbreviations. slang words, words not in the dictionary, etc. ,i.e., basically words that make classification hard.

3 Methodology

The dataset (dat) that we used had around 13.5k training samples and 860 testing samples. It consists of tweet ids and the tweets itself with the labels of the three tasks in the training data and the labels of each task in testing data. The data of tweets is often uncertain and has many concerns worth worrying. The users on Twitter tend to use slang language, abbreviations, commit spelling mistakes, use an alphabet repeated times in a word to express an emotion, use capital letters for some or all words, etc. Working on such a dataset to identify offense becomes hard. That's what makes it exciting.

We used various pre-processing techniques. We use one online emoji project on github which could map the emoji unicode to substituted phrase. We treat such phrases into regular English phrase thus it could maintain their semantic meanings, especially when the dataset size is limited. The HashTag becomes a popular culture cross multi social networks, including Twitter, Instagram, Facebook etc. In order to detect whether the HashTag contains profanity words, we apply word segmentation using one open source on the github. One typical example would be 'LunaticLeft' is segmented as 'Lunatic Left' which is obviously offensive in this case.We then removed all the punctuation marks except the exclamation marks and the question marks as these often describe sentiments. Bert itself also performs many pre-processing steps including tokenization, converting the characters to lower case and many more.

First we explored the possible solutions for **Task A**, i.e., Offensive language identification. Finally we came across Bert-Large-Uncased. BERT-Large, Uncased (Whole Word Masking) has 24-layers, 1024-hidden, 16-heads and 340 million parameters (Research) (Ber) and it is the result of the research done by Google. We trained it using the GPU facilities provided by Google Colaboratory. It took 1-2 hours to train each model. We refined the model to work on our dataset, provided by SemEval (dat). We increased our accuracy from about 55% to about 81.9%.

For Task B, i.e. Automatic categorization of offense types, we finally used Bert sequence classifier and then CNN in sequence after searching and experimenting with other types of models. With this we got the best results. Here too, the Bert we used was Bert-large-uncased (Ber).

4 Results

The accuracy we obtained in Task A using Bert-large-uncased was about 84.63% on the testing data and F1-Score was 0.81925105. The state-of-the-art F1-Score for Task A is about 0.823, so we obtained an F1-Score which was quite close to the state-of-the-art F1-Score. But we could not improve it to more than that.

The state-of-the-art F1-Score for Task B is 0.755. The F1-Score that we obtained for Task B was near to 0.634.

5 Discussion & Conclusions

We used an advanced model ,i.e., Bert which was made by Google. It is an example of the most researched upon subject and with it we obtained good F1-Scores. We have already achieved a very high accuracy in identifying offensive language on tweets, that is something with many factors that are not there in normal natural language. More research that will be done in due time will definitely lead to a higher F1-Score and this technology would be used in various fields. The future of natural language processing would be even better.

References

http://alt.qcri.org/semeval2019/index.php?id=tasks.

https://sites.google.com/site/offensevalsharedtask/olid.

https://github.com/google-research/bert.

Dimitrios Effrosynidis, Symeon Symeonidis, and Avi Arampatzis. 2017. A comparison of pre-processing techniques for twitter sentiment analysis.

Google Research. Bert-large-uncased.

Madhu Kumari Tajinder Singh. 2016. Role of text pre-processing in twitter sentiment analysis. page 095.