Toxic Comment Classification on Twitter Data

Divya Rajendran
divrajen@iu.edu

1 Introduction

The increasing online social media presence has opened doors for comments which sometimes hurt the sentiments of the individuals viewing them. Since every positive or negative experience or thoughts are posted on the internet for every to see and comment upon, abuse crops up in these media in the form of non-verbal communication. The main effort of this project is to be able to classify data in one particular social media and filter out the abusive comments from it.

1.1 Objective and Significance

Identify and classify tweets based on six classes - toxic, severe toxic, identity hate, threat, insult, and obscene. These levels of toxicity are obtained from a Kaggle challenge on Toxic Comment Classification [1]. A filter can then be employed on the tweets classified, to dynamically filter out toxic comments and make the internet a safe place for people of all ages.

2 Background / Existing Work

Social media, Twitter, is chosen to classify data/tweets using a model trained on the train data obtained from the Kaggle challenge [1]. Each tweet can be of 280 characters capturing enough information and it can be replied to and/or shared across the platform.

Every tweet containing abusive information in verbal format would be affect the sentiments of several users. If such a tweet is shared across the platform the impact on users would be harmful. It would be beneficial to understand what abuse is to properly identify abusive comments on twitter.

Currently, Twitter tries to filter and remove abusive tweets by employing their anti-abuse filters [2]. These filters consider some abusive words, map them with the account of origination and block the tweet if the account is not verified with a phone number or an email account [2]. However, this filter is not enough as abuse can come from verified accounts as well. Additionally, there is no existing system to identify abused individuals based on the tweets they post.

2.1 Abuse / Kinds of Abuse

According to Merriam-Websters dictionary [3], Abuse is defined as an improper use of an entity often for an improper benefit. Rephrasing that term 'improper' as immoral, the definition of abuse changes to anything which is done in any media chosen to demean another person. There are several kinds of abuse, a few of which are explained below.

1) Physical: where abuse relates to intentional physical harm on a person, [4]
2) mental and psychological: where abuse relates to treating a person with contempt by repeatedly making them believe they are a nobody and good for nothing, or making a person depend more on themselves for self-gain,
3) verbal and emotional: where abuse relates to torture with words like swords digging deep into peoples hearts, hurting them every single step of their shared time,
4) financial: where abuse relates to stopping access to financial or misusing financial of a person, or even not allowing a person to earn to be more independent,
5) cultural or identity: where abuse relates to intentional isolation of a person based on their ethnicity or culture or their eating habits/dressing habits or their faith or a language.

2.2 Existing Work

There are a few research papers on abuse on twitter related to information abuse / tweet spamming [5], political abuse [6], offensive language detection [7], bullying detection [8], however none of these papers are aimed at classifying these tweets to identify levels of abuse.

In the paper [5], authors talk about detecting various kinds of spam, possibilities of phishing, identifying malicious tweets and uses clustering algorithms to identify these sections of spam, which gives an insight into applying similar kind of approaches for classifying tweets.

This paper [6], talks about social bots spreading fake news or mass creation of twitter accounts to support a political candidate. The paper talks about a type of abuse called astroturf meaning a substantial number of users are paid to post or say good things about a candidate and how such type of spamming activities have large consequences. This research is aimed at identifying spam.

The paper [7] is closely related to the effort of this project and talks about identifying offensive language and cyberbullying to detect and delete such offensive content from reaching adolescents.

The paper on [8] explores ways to detect individuals with depression, much earlier than the actual prognosis and in conclusion derived a 70% classification accuracy using a SVM model. Finally, this paper [**?**] talks about detecting bullying on social media by employing sentiment analysis and data mining algorithms, some useful insights can be obtained from this paper to identify abusive tweets.

The thesis work [9] is closely related to the plan for this project. The paper talks about getting lots of tweets with the words bully, bullied, bullying and extracted the features, built a bayes classifier to get the probability a tweet would be a case of bullying. This project is aligned closely in line with this work, with an exception of the approach and tweet classification.

3 Data Acquisition

The train data can be downloaded from the data page in the Kaggle competition [1]. For the test data, tweets related to a few hashtags like #metoo, #dontcoveritup, #youknew, #lonely, #alone, #WhyIStayed", #bodyshamming, #YouOK-Sis, #EverydaySexism has been collected using twitter API for test data.

The data has been obtained for the time line starting 01-Jan-2017 to 31-Mar-2018, 1000 tweets each day containing the nine hashtags. The extraction of tweets took a longer time than anticipated as Twitter has a time limit and a download limit. So the tweets for one month and two months at a time has to be downloaded before exhausting the limit.

These tweets are collected in language English and tweet attributes like the username, location, tweet date, tweet time, hashtags, quote status id, retweeted status, quote count, retweet count, reply count, tweet text have been mined. For this project we use only the tweet id and the tweet text.

4 Model / Formulation

A model is trained to identify and classify data into six different classes, assigning each class a probability, which tells us if the tweet indeed is abusive or not. This model is trained on pre-labelled train data taken from the Kaggle competition [1]. A series of steps followed is mentioned and shown in the flow chart below 1.

1) Data collection through Twitter API
2) Pre-process the obtained data to remove lexicons, URLs, images, memes, GIFs and so on - using NLTK, twitter pre-processor, word_tokenize, stop-words
3) Feature extraction and transformation using TfidfVectorizer module from sklearn library - based on word and n-grams char analyzer
4) cross validation scores obtained on train data using Logistic regression model.
5) Use cross validation scores for each label to identify the accuracy of the model.
6) apply the regression model on test data, i.e., the tweets

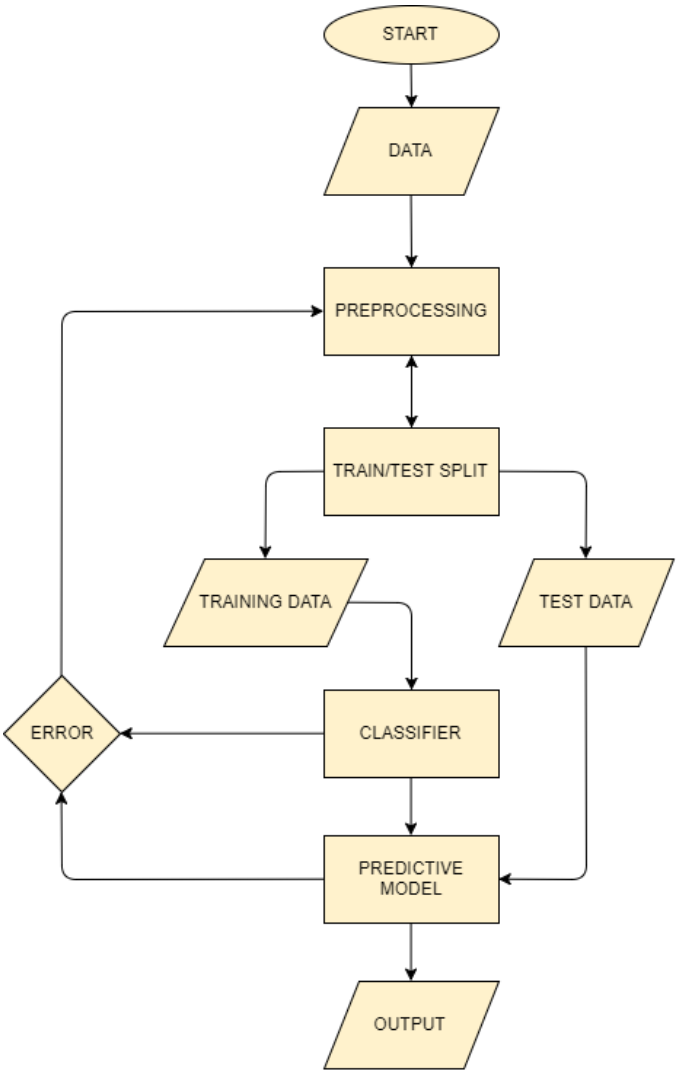

Fig. 1. Flow Diagram

5 Results

The logistic regression model achieved an accuracy of 98.6% on the train data.

The predicted probabilities for each of the class labels for the tweets is shown in the pie chart below 2. More than 50% of tweets have been predicted in toxic class label and the next highest number of tweets have been classified into obscene class.

The same information can be seen in the sum of tweets distribution into the six classes as shown in the figure 3.

The predicted classes correlation is shown in the figure 4. It looks like identity hate and threat are closely correlated to the other four classes.

And, the cross validation score for each of the class labels is mentioned in the table 1.

The word clouds of most frequent words in the prediction probabilities greater than an average is shown in the figure 5. Here, certain words stand out in each of these word clouds and some are common between the labels.

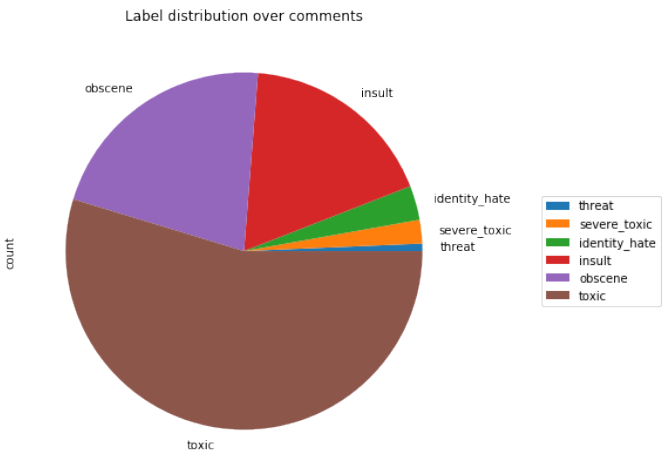

Fig. 2. Pie chart of class distribution on tweets

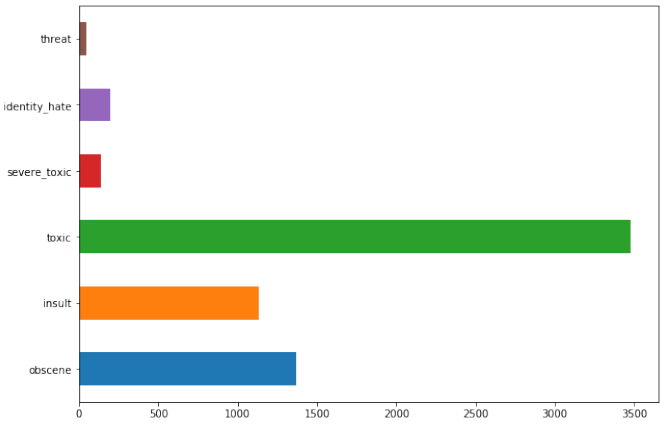

Fig. 3. Tweets predicted into classes

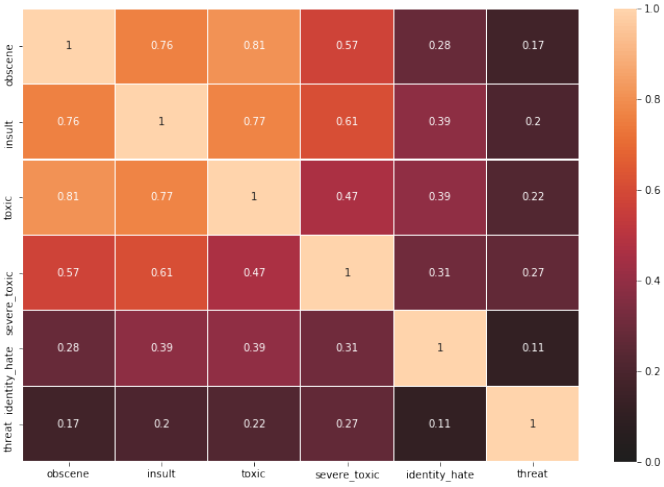

Fig. 4. Tweets predicted into classes

Fig. 5. Word Clouds for different classes

TABLE 1
CV Scores - classes

Class	CV score
obscene	0.9906158
insult	0.9830780
toxic	0.9789945
severe_toxic	0.9889034
identity_hate	0.9832962
threat	0.9888706

6 Discussion

Identifying abuse from tweets is a difficult process as readily labelled tweets are not available and the manual labelling is a tiresome and time consuming process.

The regression model performed well and it would have been great if the test data had some pre-labelled tweets to get the accuracy of predictions. It would be better to to include feature extraction based on other details of the tweet like origin, time of the tweet and employ different models for better model to filter abusive tweets.

Currently, the word clouds show words like people, say, look, stop, say and call as well into the classes. More test data should have been mined of twitter and the model tested on them.

7 Conclusion / Future work

The accuracy of the model is great and the model can be tested on live data only after testing it on a labelled test data.

For Future work, get more test data and try to employ any automatic pre-labelling tools and then test the model. Test data collected would be distributed across the diverse hashtags and mentions, so as to test the model on different settings.

It would be great if we can try to employ a text-CNN model or test different models, check their accuracy and pick out the best performing model.

References

[1] Conversation AI, "Toxic comment classification challenge." https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[2] S. Gibbs, "What can be done about abuse on social media?." https://www.theguardian.com/media/2017/dec/13/what-can-be-done-about-abuse-on-social-media, 2017.

[3] Merriam Webster, "Abuse." https://www.merriam-webster.com/dictionary/abuse/.

[4] REACH TEAM, "6 different types of abuse." https://reachma.org/6-different-types-abuse/, 2017.

[5] C. K. C. C. Jonathan Oliver, Paul Pajares and Y. Xiang, "An in-depth analysis of abuse on twitter," tech. rep., Trend Micro, September 2014.

[6] M. M. B. G. A. F. F. M. J. Ratkiewicz, M. D. Conover, "Detecting and tracking political abuse in social media," Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2011.

[7] Y. Z. H. X. Ying Chen, Sencun Zhu, "Detecting offensive language in social media to protect adolescent online safety," (Pittsburgh, USA), Symposium on Usable Privacy and Security, 2011.

[8] S. C. E. H. Munmun De Choudhury, Michael Gamon, "Predicting depression via social media," (Boston, MA), Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 2013.

[9] J. Sui, *Understanding and Fighting Bullying with Machine Learning*. PhD thesis, University Of Wisconsin, Madision, 2015.