Detection of Hate And Offensive Speech In Social Media

Sanket Sheth
Graduate Student in Computer Science
Rochester Institute of Technology
Rochester, NY 14623
sas6792@g.rit.edu

Abstract

Social media with over 2 Billion users worldwide has become one of the if not the greatest outcomes of this century and along with the positives come negatives and a recent scare that has surfaced to be a harsh reality is social media abuse and bullying and with the social network companies in dispute and confusion over the right to speech acts they end up having little moderation over the spoken text. The image and video-based moderation are quite ahead in comparison to written text which remains undetected. The idea is to use various classification algorithms and sentiment analysis to train on twitter and other social media data to develop a model that detects potential cases of cyberbullying or abuse, also going forward the model may consider user history, user interactions and other social media based features to improve efficiency. There is also a chance to further the project by introducing deep yearning for a more refined classification.A classification model was implemented with experiments and analysis carried out for different features and models with about 73% accuracy using random forest with a maximum depth of 60.

1 Introduction

Social media abuse and bullying are one of the harsh realities of this century. The methods proposed here are a way for detecting cyberbullying and hateful abuse. One of the key challenges here is to separate hate abuse from offensive language, as the young lingo often involve using abusive offensive words used in normal daily conversation which transcends to social media and the main task would be to detect posts that contain hate speech or bullying regardless of the offensive language used in it. This gives way to a classification solution which would classify posts into different classes like, hate, offensive, normal, etc. There are several ways the classification data can be taken further, one approach is to detect future tweets of such activity like hate or bullying, but one other approach is to classify the user itself that is based on the users timeline determine whether the user participates in such wrongful activities.

2 Past Related Work

2.1 Past Methods

There exist several methodologies explored for cyberbullying or abuse detection using different types of classification algorithm to achieve it, the primary method to achieve this is using a bag of words, this is a method that gives high recall but ends up with wrong results of classifying many offensive tweets as hateful and which enables the high numbers of false positives even with a high recall rates (Kwok and Irene and Wang and Yuzhou, 2013). Using syntactic features to identify the intensity of words that use parts of speech tagging to detect abuse and hate (Silva and Mondal and Correa and Benevenuto and weber, 2016), neural networks have also been explored but the context of hate seems too broad with the results achieved.

A graph-based approach is also used based on likes and comments to build bipartite graphs and identify negative behavior (Hosseinmardi and Mattson and Rafiq and Han and lv and Mishra, 2015). Deep learning that us using CNN and LSTM for classification of comment abuse is also explored. One such approach that intrigued me uses logistic regression for classification of text in classes like hate and offensive texts, the results obtained by that research did not turn out to be

promising (Davidson and Warmsley and Macy and Weber , 2017), while another approach uses the twitter streamer data and use the random forest to extract three different forms of features including network-based features like friend lists, user tweets, mentions etc. and achieves good results with it. The number of features used for is 30, but this is user based and classify users as being a bully, aggressor etc. (Chatzakou aand Kourtellis and Blackburn and Cristofaro and Stringhini and Vakali, 2017). All these techniques end up using classification with varied forms of features to either detect bullying or detect hate and abusive comments. The features used are also polarizing in comparison to each other. The line between offense, abuse, hate, and bullying is very thin and most of the times the results overlap few classes together.

2.2 Past Data Sets

There are several datasets available but most of them turn out not to be labeled, this creates a problem, one of the most comprehensive data set available is twitters streamer dataset which gives user information along with their timelines and amounts to 1.9 million tweets approximately, as this data set is not labeled it cannot be used still I tried to search for a labeled version of the set but to no avail concerning my needs. The best dataset at hand is a portion of crowd annotated dataset extracted from the streamer dataset using the vocabulary of the hatebase.org of hate words. This dataset has data labeled in three classes that is hate, offensive and neither. This right now is the primary source of data that I am planning to use for the project. Other options that I explored are like the data set used for is very detailed and comprehensive but is not available (Chatzakou aand Kourtellis and Blackburn and Cristofaro and Stringhini and Vakali, 2017). Other social network websites also have data like Wikipedia, for example, the Wikipedia detox project used in research where comment abuse was considered. Also, Kaggle has few datasets that include only sentiment analysis features labeled.

Also, researchers have extracted comments from YouTube videos and carried out analysis on them along with combining them with the officially released data set by Formspring.Another data set that was labeled was used for a research explored predictive symbols for hate speech detection concentrated on Twitter and also used the same grounds to explore the influence of an annotator on hate speech detection.(Waseem and Hovy, 2016) (Ross and Rist and Carbonell and Cabrera and Kurowsky and Wojatzki, 2017) (Waseem, Zeerak, 2016) The data here is labeled as racist or sexist or None. The tweets are denoted as numbers that are their streamer data ids which will involve preprocessing to mine them using any tweet crawler of good standards.

3 Data Set

Focusing more in the primary data, this data is a sub-part of the tweets extracted from the twitter streamer API using a dictionary of hateful words compiled by `hatebase.org` and from the tweets extracted entire timelines of the user is extracted. A random selection of 25K tweets is then extracted from the compilation of the user's timelines.This 25K tweets then underwent human annotation. The annotations by different humans is then compiled together to vote each tweet with a label of being Hateful, offensive and neither. These become the labels for the dataset. The tweets are documented in a csv file with 7 columns, with each column described below -

- Column 1: Tweet ID
- Column 2: Total number of human annotators
- Column 3: Total number of annotations for hate speech
- Column 4: Total number of annotations for offensive speech
- Column 5: Total number of annotations for neither
- Column 6: Final voted label attached to the tweet
- Column 7: Tweet text

Label	Example
Hate	0
Offensive	1
Neither	2

Table 1: Labels

The data set at hand is very unbalanced with offensive tweets way more than the other two and the number of hate speech tweets is extremely low. The split for the three labels is 1430,19190,4163 respectively. The ratio is about 6:77:17 for the three.

4 Approach

4.1 Methodology

The point is to detect the decision boundary between the three labels. Examples of the three labels is given below-

Hate Speech:

> "@MarkRoundtreeJr:LMFAOOOO I HATE "X" PEOPLE https://t.co/RNvD2nLCDR" This is why theres "X" people and "X" "

Offensive Speech:

> "@z0mbiedance: I made that 'Swear Word' lunch. @elizabethbatman" 'Swear Word' love lunch"

Neither:

> "@EdgarPixar: Overdosing on heavy drugs doesn't sound bad tonight." I do that 'Swear Word' shit every day."

As it can be observed from here, the boundary is very thin and detecting it will be a difficult task. Different classification algorithms are implemented and experimented to find out the best possible combination of features and model.

4.2 Features

Many language and text based features were considered. A list of all the features considered along with their relevance is mentioned in table 2.

Some features are very useful while some give very bad results as tested on the model that gave the best results. In almost all cases the offensive tweets were classified, but some features were specifically selected to create a model that can classify the hate and neither tweets as well. The features are extracted using various different methods like regular expressions, using scikit learn and some where provided with the data.

Features	Relevance
Presence of URLs	Negative
Number of Annotators	Neutral
Retweet	Neutral
Number of Words	Very Positive
POS Tags: Adjectives	Very Positive
Mentions	Very Positive
POS Tags: Cardinal Numbers	Positive
Length of Tweet(Characters)	Positive
Bag of Words(Count)	Very Negative

Table 2: Feature Relevance

4.3 Model

Several classification models were experimented with like logistic regression, svm , random forest etc. The model parameters were also meddled with to see which model aligns with the data at hand. In the end Random forest with a depth of 60 is selected. Logistic regression with l1 and l2 regularizer gave the worst results with l2 being slightly better than l1 while no effect was registered by changing the cost value. Svm gave moderately better results, with few classifications for labels other than offensive. The problem is with the data imbalance which is proved by normalizing the data to the number of samples as the smallest label that is hate. Also, a one on one comparison is carried out using random forest model for all three labels that is hate is compared with offensive , hate with neither and offensive with neither with drastically better results proving the effects of data imbalance. Also, there were changes carried out with the initial split of train and test ending with a 80:20 split for train:test. A list showing the rankings of the various models used is given below-

- 1: Random Forest with depth greater than 40
- 2: Random Forest with depth above 20 and below 40
- 3: Support Vector Machines
- 4: Random Forest with depth greater than 1 and below 20
- 5: Logistic Regression with l2 regularizer
- 6: Logistic regression with l1 regularizer
- 7: Logistic regression

5 Results and Discussions

Past results suggest that finding the decision boundary in itself is very difficult when it comes to offense and hate, there have been research with extremely bad results with some getting results above average. One biggest reason for this is also the bias created by humans during the process of human annotations. The classification rate achieved is 73% using random forest with maximum depth of 60. Also, after analyzing the various results achieved advantages and disadvantages of specific features and models is documented. Also, as we do not have optimal data use of a general baseline which turns out t be 33.33% is not advisable also using a prior work as baseline is pointless as with most systems the data used is different and although the task performed is in the same domain the labels and outputs are different. Also, the results obtained are not particularly impressive when it comes to this domain of research.

5.1 Optimal Results

The confusion matrix for the optimal results is tabulated below with the rows tabulating the actual labels and the columns tabulating the predictions. The true positives are the diagonal values.

Label	Hate	Offense	Neither
Hate	9	254	24
Offense	81	3475	301
Neither	38	644	131

Table 3: Confusion Matrix(Random Forest MaxDepth =60)

As can be observed here the model works very well for the offensive tweets but contrary to other models this also classifies hate and neither labels to a certain extent. The classification rate is given as 72.9%. The f1 score with macro average is 36.84%, which seems bad on surface but data imbalance is the root cause for this.

5.2 Comparative Results

Many models and features were used along with a method to nullify the data imbalance by normalizing the data set to the lowest count for a label. Firstly, logistic regression was used with l2 regularizer which gave bad results but a okay accuracy which is due to the high number of offensive tweets but the classification of other labels was nominal which made this model state one of the worst results.Thus accuracy can not be a good measure to judge the classifiers.Then support vector machines were used for the same task giving moderate results but took the most time to run computationally. The comparative results of the accuracy and f1 score is tabulated below for the various algorithms.

Logistic Regression (L2)	Measure
Accuracy	77.7%
F1 Measure(Macro)	29.6%

Table 4: Measure Logistic Regression

Support vector Machine	Measure
Accuracy	78.1%
F1 Measure(Macro)	32.4%

Table 5: Measure Support Vector Machine

Random Forest(MaxDepth=60)	Measure
Accuracy	72.9%
F1 Measure(Macro)	36.84%

Table 6: Measure Random Forest

Then we can look into the results obtained by skewing the data sets, by comparing labels one o one, we get interesting findings as when hate is compared with offense we get moderate classifications for both suggesting that offensive tweets had a negative involvement on the other labels.Also, when compared with the offensive the results for neither and hate although bad were better than most results we achieved, showcasing the bad effects of using a multi-class classification problem. A sample of one of the confusion matrix is displayed below when hate is compared with neither directly.The accuracy achieved was 67% and the F1 score was 52%.

Now, we further look into the results obtained by normalizing the skewed data by limiting the larger label data to the smaller label data which is hate data.When we analyze the results obtained here we see the clear effects of data imbalance as when tested hate against offensive we get a situation where the number of hate tweets predicted are more than offensive which also jsutifies our model which has special features to detect hate more than offensive that is it is biased towards hate a bit. The

Label	Hate	Neither
Hate	64	223
Neither	139	674

Table 7: Confusion Matrix for Hate vs Neither

accuracy achieved when we compare hate with offensive in a normalized skewed condition is 51.9% and the F1 score obtained is 51.8%. Below you can find the confusion matrix justifying these results.

Label	Hate	Offensive
Hate	157	130
Offensive	146	141

Table 8: Confusion Matrix for Hate vs Neither

Thus the comparative results proved very useful to justify the limitations stated in the methodology regarding the imbalanced data.

6 Future Work

For stretch goals if possible I would like to explore the effect of deep learning on the data for classification that is using the CNN and LSTM model mentioned in this paper [5]. Also, here the classification process is tweet based it will be interesting to explore the same for user based that is including features that are network based like number of followers, number of tweets, time since first joined, number of blocks etc. will bring an entire new dynamic in the model and might also be able to recognize a pattern in bullies and abusers. This might help achieve better the results of the normal method being explored in this project.

Another, step in terms of future work that can be considered is implementation of a mobile capable version of the system and to broaden the training set and include all social networking websites and chat messenger to alert the user when something that they are typing might be considered as offensive and out right hateful making the risk of causing even indirect bullying and offense low.

7 Conclusion

In conclusion, we can safely say a fairly good foundation has been created with which further work on this topic can be based, also one glaring problem of data imbalance has been addressed along with even more dangerous problem of human bias which the annotators brought in and which needs to be countered in order to dramatically alter the results achieved here for good. Human bias is something impossible to counter and maybe with the help of research in the field of annotator bias one can get much better results then what we achieved here.One more thing we can take away from this research is the potential of separate machine learning models when it comes to dealing with thin decision boundary with random forest giving the best results. Apart from that, it cannot be stressed enough but there and always will exist a fine line between hate and offense and modern day social media must be capable to differentiate it in the near future.

Also through this project we were able to corelate many different attributes to haters like the number of hashtags or the use of URL's or mentions used. Further research can open many doors in this domain of research.

The digital age has brought many boons connecting the world and bringing it closer than ever before, but cyber abuse and hate are one of the biggest problems of the day and there have been many cases of mental harassment that is seen all around the world due to it. It is also evident that social media companies are not regulating the content enough for many reasons including freedom of speech but lack of accuracy in detecting these posts should not be one of the reason. We also can positively say there is common pattern among the abusers and bullies which helps us create these models that help us learn their behavior and improving on the previous system as this regulation and detection is the immediate need of modern life.

Acknowledgments

I would like to thank Professor Cecilia Ovesdotter Al (Associate Professor at College of Liberal Arts, Rochester institute of Technology) for helping me in understanding this intriguing subject of natural language processing and also for constant guidance and motivation.

References

Kwok, Irene and Wang, Yuzhou, 2013. *Locate the Hate: Detecting Tweets against Blacks.*, AAAI

Leandro Araújo Silva and Mainack Mondal and Denzil Correa and Fabrício Benevenuto and Ingmar Weber,

2016. *Analyzing the Targets of Hate in Online Social Media*. Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.

Homa Hosseinmardi and Sabrina Arredondo Mattson and Rahat Ibn Rafiq and Richard Han and Qin Lv and Shivakant Mishra. 2015. *Detection of Cyberbullying Incidents on the Instagram Social Network*, CoRR

Thomas Davidson and Dana Warmsley and Michael W. Macy and Ingmar Weber, 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR*,

Despoina Chatzakou and Nicolas Kourtellis and Jeremy Blackburn and Emiliano De Cristofaro and Gianluca Stringhini and Athena Vakali, 2017. *Mean Birds: Detecting Aggression and Bullying on Twitter*. CoRR

Dinakar, Karthik and Jones, Birago and Havasi, Catherine and Lieberman, Henry and Picard, Rosalind, 2012. *Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying*. ACM Trans. Interact. Intell. Syst.

Zeerak Waseem and Dirk Hovy, 2016. *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*. SRW@HLT-NAACL

Waseem, Zeerak, 2016. *Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter*.

Björn Ross and Michael Rist and Guillermo Carbonell and Benjamin Cabrera and Nils Kurowsky and Michael Wojatzki, 2017. *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. CoRR