Project Report
Study of Abuse on Social Media
Data Mining - Spring 2018
Indiana University, Bloomington, IN, USA

Divya Rajendran
divrajen@iu.edu
Under Prof. Vincent Malic

Objectives and Significance:

With the increasing popularity of the online social media, every kind of feeling and emotion is being shared by individuals in the form of tweets or posts on the internet. While some of these tweets/posts tend to be positive while others tend to be negative, these can still further be classified into various human feelings, happiness, sadness, anxiety, depression, elation, euphoria and such. Can one try to use these tweets/posts in identifying abuse on social media?

By abuse we mean both abused and abusive versions on the online social media. The first form of expression is generally hard to identify. There are a few abused individuals who involuntarily express themselves, their hurt, humiliation and try to post/tweet in a passive tone, which is hard to pick up and is an urgent cry for help. These kinds of abused individuals sometimes require an initiative like trending hashtags, such as "#metoo", "#dontcoveritup", "#youknew", "#lonely", "#alone", "#WhyIStayed", "#bodyshamming", "#YouOKSis", "#EverydaySexism" and so on to initiate a response from the abused individuals and express themselves.

Some do express themselves and if there existed a system, basically a machine learning algorithm in place to identify such tweets/posts and direct a series of helpful articles and

links to reach out, then we would be able to change the lives of people who involuntarily seek out to express their hurt. We cannot simply ignore the abusive tweets as well, as an abusive person tends to be someone who was abused as a child and so seeks out to vent out their anger/frustration on someone like them.

We would like to try our best to find out how effectively we can classify a tweet/post as a cry to be heard or a tweet/post is an abusive one so that appropriate help measure will be directed toward the individuals posting such tweets. In this project, we would like to take advantage of the abundantly available data in the form of tweets from a micro blogging form of social media, Twitter, to identify and distinguish different varieties of tweets, to potentially use them to study the various kinds of abuse via social media tweets.

Background:

When we talk about abuse, we generally are inclined into thinking about a physical violence or a relationship gone wrong. So, what is Abuse? According to Merriam-Webster's dictionary [1], Abuse is defined as an improper use of an entity often for an improper benefit. We can rephrase that term improper as immoral as well, changing the definition of abuse to anything which is not used or done in conjunction with the basic human moral code.

There are various kinds of abuse which can be - physical, mental, verbal, psychological, emotional, sexual, financial, cultural or identity abuse [2]. These are explained as below:

1. Physical: where abuse relates to intentional physical harm on a person,
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 people's 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.

We are more concerned about the study of various kinds of the abuse on social media to study the various ways the social media is evolving to handle these tweets to identify people in abusive situations or who are abusive. When these people are recognized, how effectively we can try to passively direct them to relevant help pages or ads or links through which they can be empowered to stand up for themselves, no longer suffering the abuse in a firm and efficient manner is to be seen. Also, we can direct the abusive people with relevant links and material, to enlighten them and help them reform to be a better person, to live and let live, and to make the world a better place to live for the generations to come.

In this project we would like to work on the micro blogging site Twitter and we take the data from Twitter as a base and use sentiment analysis to analyze the entire tweet of 280 characters into classifying it as a positive or negative or neutral sentiments. There might be challenges while opting for sentiment analysis in conjunction with a classifying algorithm like a tweet expressing both a positive and a negative connotation.

Additionally, a tweet might contain emoticons as an expression, or tweets might contain words which would indicate abuse or be abusive if taken out of context. There might be links, memes or videos attached to a tweet and we would like to pre-process the data to remove these kinds of data to work mainly on textual data.

A tweet can be termed as a user status or post which describes their sentiments on various topics ranging from politics to movies to games and so on. A tweet is often re-tweeted meaning each tweet can be shared by a user with his friends in Twitter, who then have the access to see who originally tweeted something. These tweets are often accompanied by hash-tags, which are a series of words or phrases preceded by hashes which identify a message on a specific concept or topic.

Generally, on looking at a set of tweets, we can generalize them into a set of positive or negative or neutral tweets, and later we classify whether these sets of tweets are either an abused or abusive or a statement or fact kind of tweet. However, we need to find an effective combination of techniques from sentiment analysis, data mining, information retrieval to classify the tweets into the proper classifications of polarity.

Currently Twitter does try to remove offensive or abusive tweets by employing their anti-abuse filters [3]. These filters consider some abusive words, maps them with the account of origination, blocks the tweet if it finds that the account has not been verified with a phone number or an email account [3]. This filter is not enough to tackle abusive tweets, 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.

We would like to address this gap and apply existing data mining techniques and algorithms in conjunction with sentiment analysis to study the abused individuals based on their tweets so that an appropriate help is provided to them on time, thereby making a positive impact in our society. We would to like to focus on the trending tweets with the following hashtags, "bodyshamming", "YouOKSis", "EverydaySexism" to study the various ways in which these kinds of abuse occur in social circles.

There are a few research papers on abuse on twitter, which talk about abuse of information or spamming of tweets [4], political abuse [5], offensive language detection [6], bullying detection [7], however none of these papers are aimed at finding a distinction between abused and abusive versions of tweets.

If we take the paper *"An In-Depth Analysis of Abuse on Twitter by J. Oliver, P. Pajares, C. Ke, C. Chen and Y. Xiang"* [4], it talks about detecting various kinds of spam, possibilities of phishing, identifying malicious tweets and uses clustering algorithms to identify these sections of spam, which gives us an insight into applying similar kind of approaches for classifying our tweets into abused, abusive and neutral clusters.

And the paper on *"Detecting and Tracking Political Abuse on Social Media by J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, F. Menczer"* [5], 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. The research talks about ways to identify such spam.

The paper on *"Detecting Offensive Language in Social Media to Protect Adolescent On-*

line Safety by Y. Chen, S. Zhu, Y. Zhu, H. Xu" [6] is closely related to the effort of our current research and talks about identifying offensive language and cyber-bullying to detect and delete such offensive content from reaching adolescents.

The paper on *"Predicting Depression via Social Media by M. Choudhury, M. Gamon, S. Counts, E. Horvitz"* [7] 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, the paper on *"Twitter Bullying Detection by H. Sanchez and S. Kumar"* [8] talks about detecting bullying on social media by employing sentiment analysis and data mining algorithms, we draw some useful insights from this paper in identifying abusive tweets which was termed as an extremely difficult task in the paper.

The thesis work done by *Understanding and Fighting Bullying with Machine Learning by Junming Sui* [12] closely related to the plan we have 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. We are working closely in line with this work with an exception of the method used and that we are trying to identify abuse in 6 levels of toxicity.

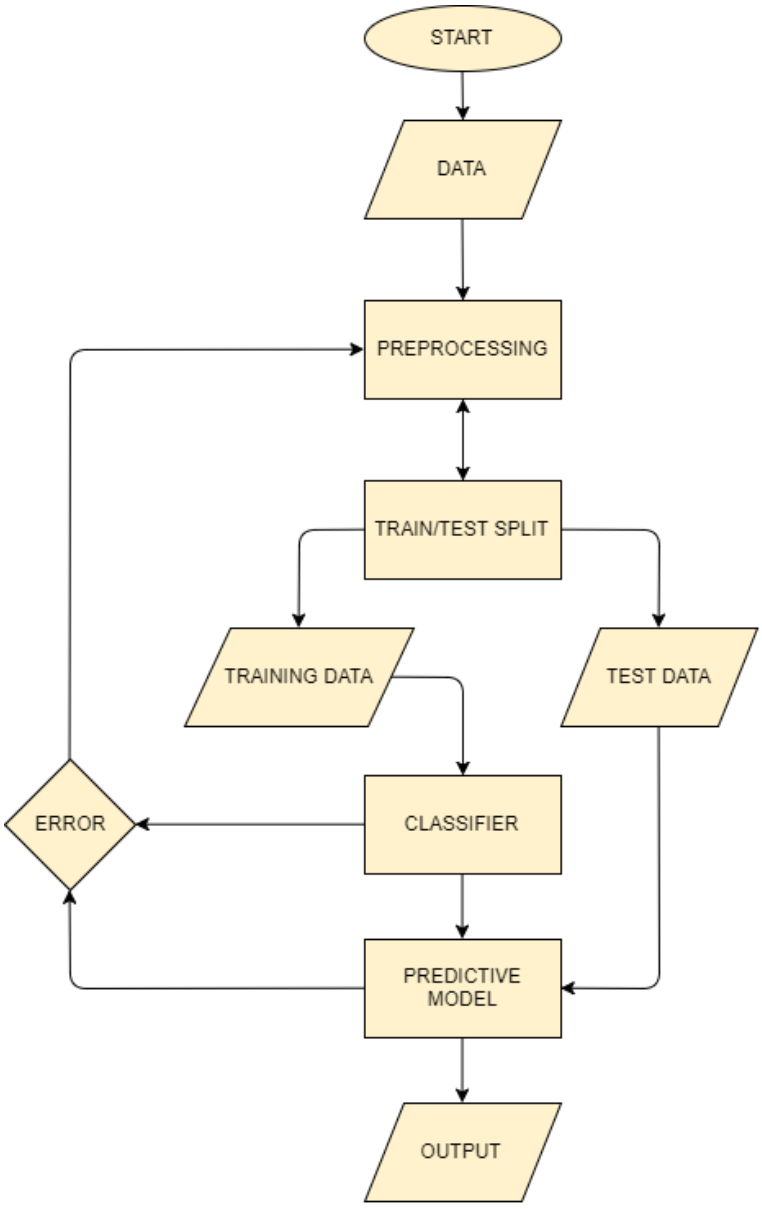
Methods:

As we have mentioned, Twitter is a micro blogging social media, which allows users to post statuses or opinions or comments or anything they want to share in real time and can be done so in the form of tweets. Each of these tweets can be retweeted, loved and commented upon, these can often contain hashtags. These tweets are short posts or messages of 280 characters in length, earlier this was 140 characters only and has been increased to 280 in September 2017.

For this project we collected data from Twitter API for all the users to get a dataset from a varied set of users, the tweets posted with the hashtag "#bodyshaming" in English. Each of these tweets has attributes like the username, location, tweet date, tweet time, hashtags, quote status id, retweeted status, quote count, retweet count, reply count and so on. We got the data for the timeline starting 01-Jan-2017 to 31-Mar-2018, 1000 tweets each day containing the hashtag we needed. We got every possible attribute for an initial analysis,

but ended up using only tweet id and text.

The extraction of tweets took a longer time than anticipated as Twitter has a time limit and download limit. So we had to take tweets for one month and two months at a time before we exhausted our limit. We extracted tweets making use of the twitter API using a search term "bodyshamming" for a period of one month each starting January 2017 and ending in March 2018.
The process of steps we followed is shown in the flow chart.



The steps in detail are below:

1. Data collection through Twitter API
2. Pre-process the obtained data – to remove lexicons, URLs, images, memes, GIFs and so on, conversion of upper to lower case characters.
3. Feature Extraction - done using a module "featureextraction" from existing library from scikit-Learn
4. Split the train and test data, fit the train and test data and transform the tweet text.
5. Use the train data to train the classifier - here we have used logistic regression.
6. test the model on the test data
7. Use cross validation techniques to get the ROC AUC score value for each label we have predicted for a tweet to see the accuracy of our model.

We wanted to label each tweet manually to either abusive or abused, however it was very tedious and an existing training data set from a *Kaggle challenge on identifying toxicity* [13] provided a solution to us. We used the train data set, which came pre-labelled into *toxic, severetoxic, obscene, threat, insult and identityhate.* We used this train data to train our classifier and used the tweet data we collected as a test set to test our model and find the accuracy of the fit.

We needed only the text of the tweet as a feature for our analysis, as we wanted to focus more on identifying abuse probability from a tweet directly than from the other features which are more suitable for finding out if a person is abusive, which we are not trying to identify in this project. Since we had the pre-labelled tweets with us, we did not find the need to work on sentiment analysis for now, we are planning to include that as a part of future research.
The tweets are pre-processed to remove numbers, special characters, newlines or tabs or extra spaces and then tokenize the words using the library "nltk" and words having more than 2 characters is appended back to make the comment. Then we used a feature extraction form from term frequency – inverse document frequency (td-idf) sub-module of scikit-learn to calculate the relevancy of a word from a tweet, which is stored in a sparse row matrix format which includes the probability of each word.

We chose a linear model, Logistic Regression as we have the labelled data in a binary format and we felt that this model would be very good in predicting the label of a tweet text. It calculates the probability of a tweet text to be in any of the 6 labels. The probabilities calculated for each of the tweet texts is saved into a predictions variable for every tweet text. This predictions data frame holds all tweets along with their corresponding labels and their predictions.

While we predict the label probabilities for each of these tweets, we even test our accuracy in the form of ROC AUC score values on 10 cross validations for each of the labels. We draw a correlation matrix to identify the probability of a tweet belong to any one of the above six labels. We draw a pie-chart to visually show the distribution of the tweets we have into the six categories. We even show a word cloud of the most frequent words which appear in each of the tweets based on the probability we obtained using the classifier.

Based on the probabilities we have achieved through our fit and comparing the probabilities of each of these tweets, we can identify which category a tweet is most likely to be classified under. Also, when we compared the ROC AUC scores for each of the labels we achieved performed much better than an average classifier of score 0.5.

Results:

We have obtained the ROC AUC scores for the classifier, by cross validating in 10 folds, we fit on our train data and tested on our tweets data set for each of the categories/labels. Our classifier performed better than an average classifier or a random classifier, which would turn out a score of 0.5. The results are as shown in Table 1.

We see that when we compare with an average classifier of score 0.5, our model performs a little better. However, model fit did not converge well for threat, insult and identity hate categories, so we cannot be sure of the predictions we achieved for these categories. We plan to obtain better data and find an efficient way to get them all labelled to understand our data more.

For the test data set we have now, the results we achieved imply that the tweets gath-

Label/Category	ROC AUC score
toxic	0.7852
severe toxic	0.9059
obscene	0.8163
threat	0.8744
insult	0.8184
identity hate	0.7699

Table 1: ROC AUC scores for each of the categories

ered had more abuse and we plan to perform more analysis on these tweets, which would help us identify if the user is abused or is an abuser or in fact an observer who reported this on twitter.

The relative counts of the tweets classified into each of these categories is shown in Table 2. From the counts we observe in the table, we observe that the more number of tweets fall under the category *toxic* giving us more insight into our tweets content and how we can make our model better in identifying toxic texts immediately.

Label/Category	count
toxic	3222.59
severe toxic	314.12
obscene	1834.83
threat	77.88
insult	1535.86
identity hate	299.84

Table 2: relative counts for each of the categories

We want to show a sample of tweets the model has categorized in the table 3. We see the quality of our predictions based on the text, however since we pre-processed tweets to remove hashtags as well some of these tweets had lot their actual meaning and turned out to be categorized under abuse categories.

We see that the tweets contained words (please excuse the choice of words), *fuck, victim, slut, skinny, lizard* and so they are classified with higher predictions toward toxic and next on obscenity. Our next step here would be to perform a LDA topic modelling by labelling at least 10K tweets and train them on a classifier and test it out on the test data to label the tweets into different topics based on which we can further modify the classifier to identify different levels of abuse from a tweet text.

Text	**toxic**	**severe toxic**	**obscene**
stop thinner girls fuck you	0.82331	0.471533	0.757814
stop you lizard with followers	0.823193	0.342628	0.755836
our idol not yours you not fan you bodyshame h...	0.824557	0.452731	0.782472
stop chirag you fuckboy	0.83385	0.49115	0.772364
was victim own mother you are built just like ...	0.640645	0.119456	0.473602
slutshaming and bitch you want maim you	0.696313	0.295018	0.596264
today your skinny friends hang out with your f...	0.603834	0.107051	0.563073

Table 3: Sample tweets classified

Below is a pie diagram in Figure 1, indicating the distribution of tweets into six of the categories which we try to assign our test data into. We find that the tweets are categorized more into the categories *obscene, insult and toxic* and lesser into *sever toxic, identity hate and threat*. We would see in a correlation matrix below that the tweets which are classified into toxic category are marked obscene or as an insult as well, which would account to the larger pies occupied by these categories.

Even though our model performs well, we are not entirely satisfied with this result as we did not find a better way to label so much of data, we relied on the labelled data obtained from kaggle data set to train our classifier. Our next steps are to obtain more data and write an automatic labelling algorithm to identify certain words or topics and label the tweets into the six categories, without manual intervention. This would help have a clear command on our categories and the words which are most common under them, this would help us better analyze the tweets

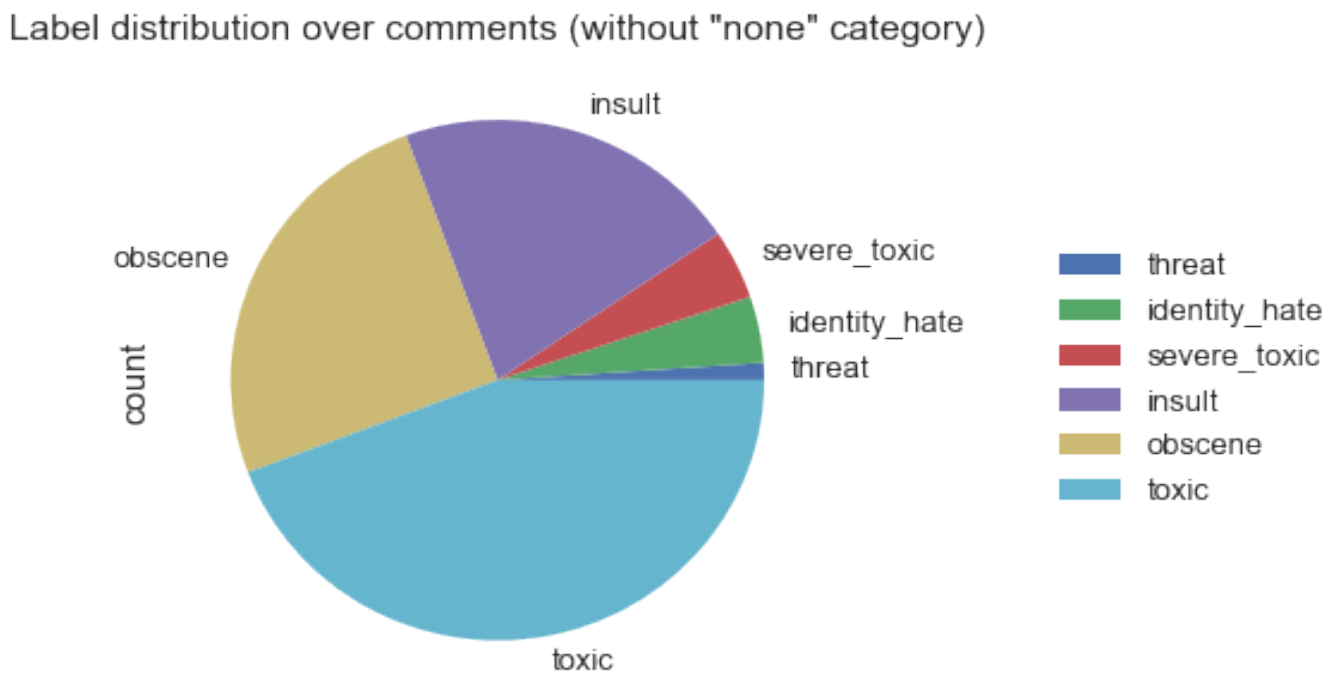


Figure 1: Tweet Distribution Pie Diagram

We have plotted a correlation matrix between each of these six categories and how each of them are related to each other, or the probability that a tweet of one category existing in other category as well. The correlation matrix for the tweets distributed into the six categories is shown in the figure 2.

We see from the matrix that the tweets categorized into toxic, had higher probability of being classified into obscene, severe toxic, insult and identity hate as well. The tweets classified into severe toxic can also be identified with insult, obscene and toxic categories. The tweets classified into obscene were closely correlated with tweets classified as insult, implying a higher probability of a tweet being classified into insult if its classified as obscene. Threat and identity hate have lesser correlations between the other categories in contrast to toxic, insult and severe toxic.

A word cloud built on train data depicting the different words with highest probabilties in toxic category is shown in figure 3.

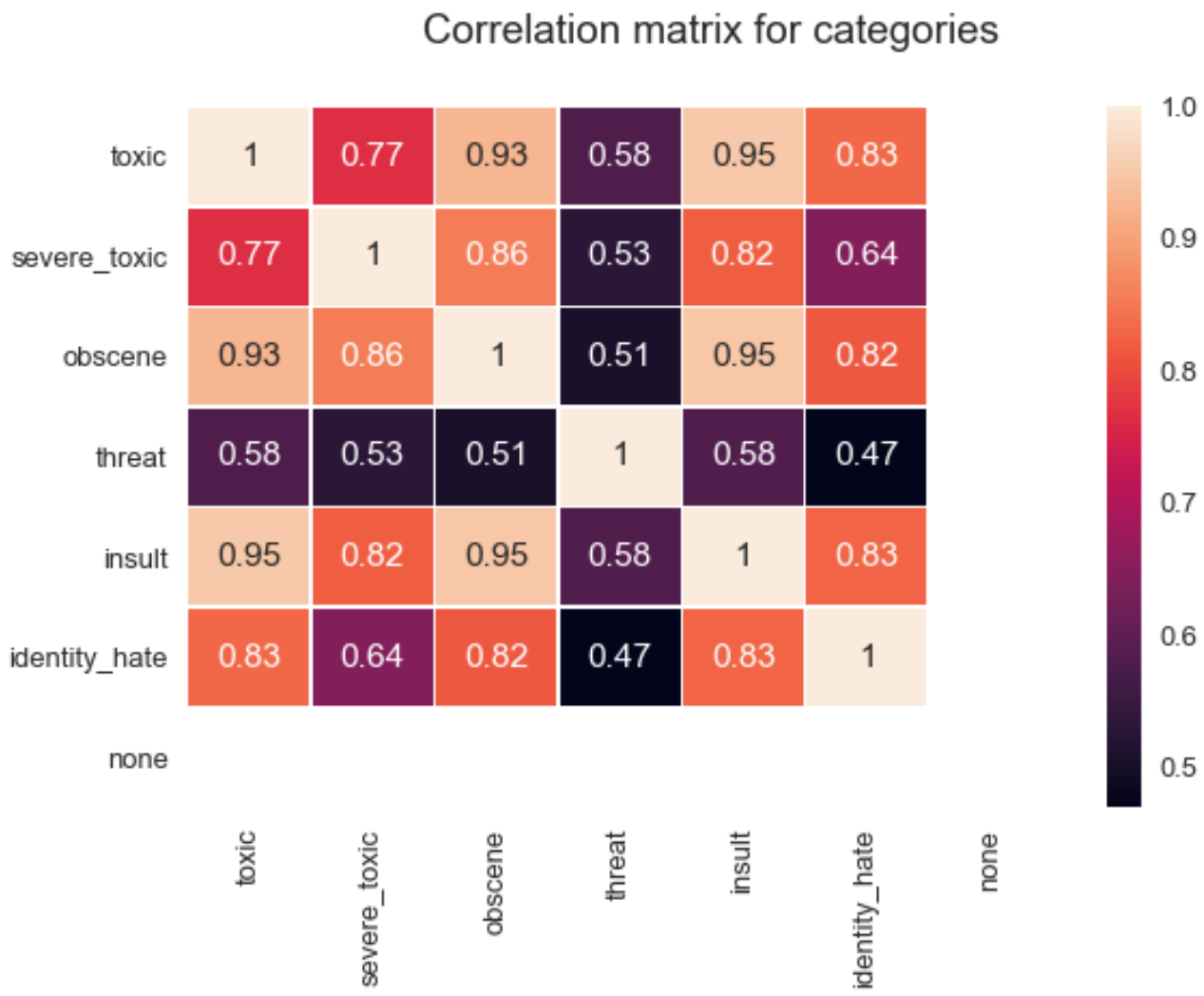


Figure 2: Correlation Matrix

A wordcloud built on test data depicting the different words with highest probabilty of being categorized into one of the categories, "toxic" is shown in the Figure 4.

Also, we have constructed a tf-idf vectoriser to calculate the frequency of words which fell under each of the six categories of toxicness of a tweet to better know the distribution of words in the train data. We had enough data to analyze our tweets. The figure is shown in Figure 5.

Figure 3: Train data Toxicity Word cloud

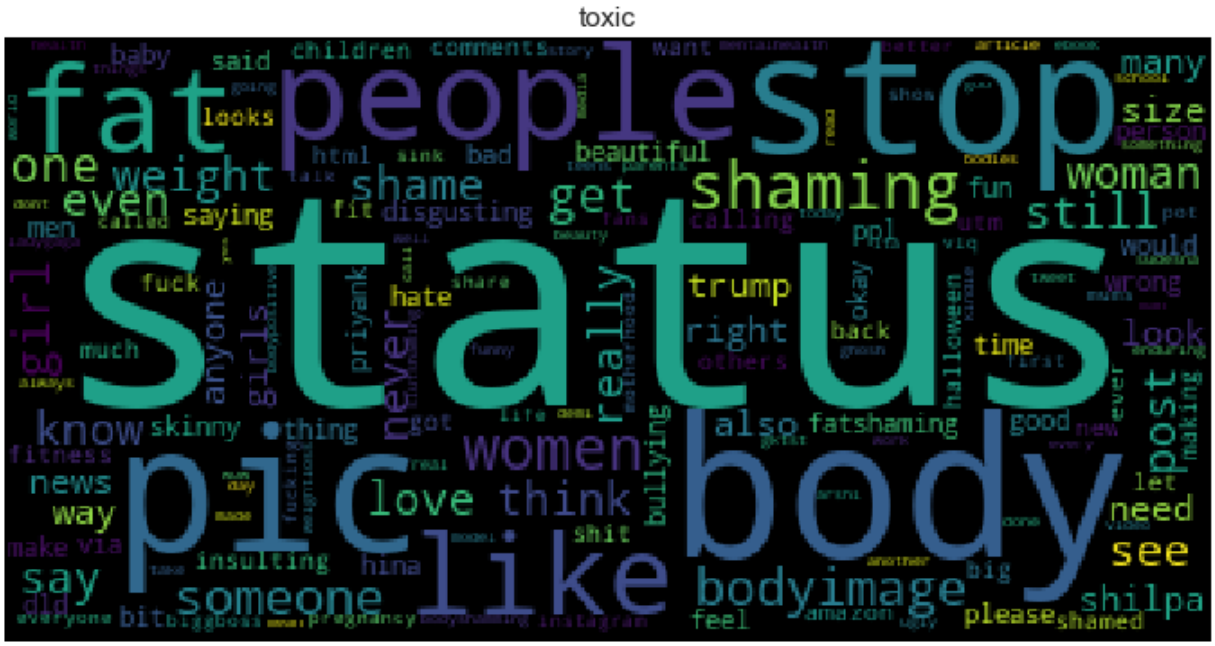


Figure 4: Test data Toxicity Word Cloud

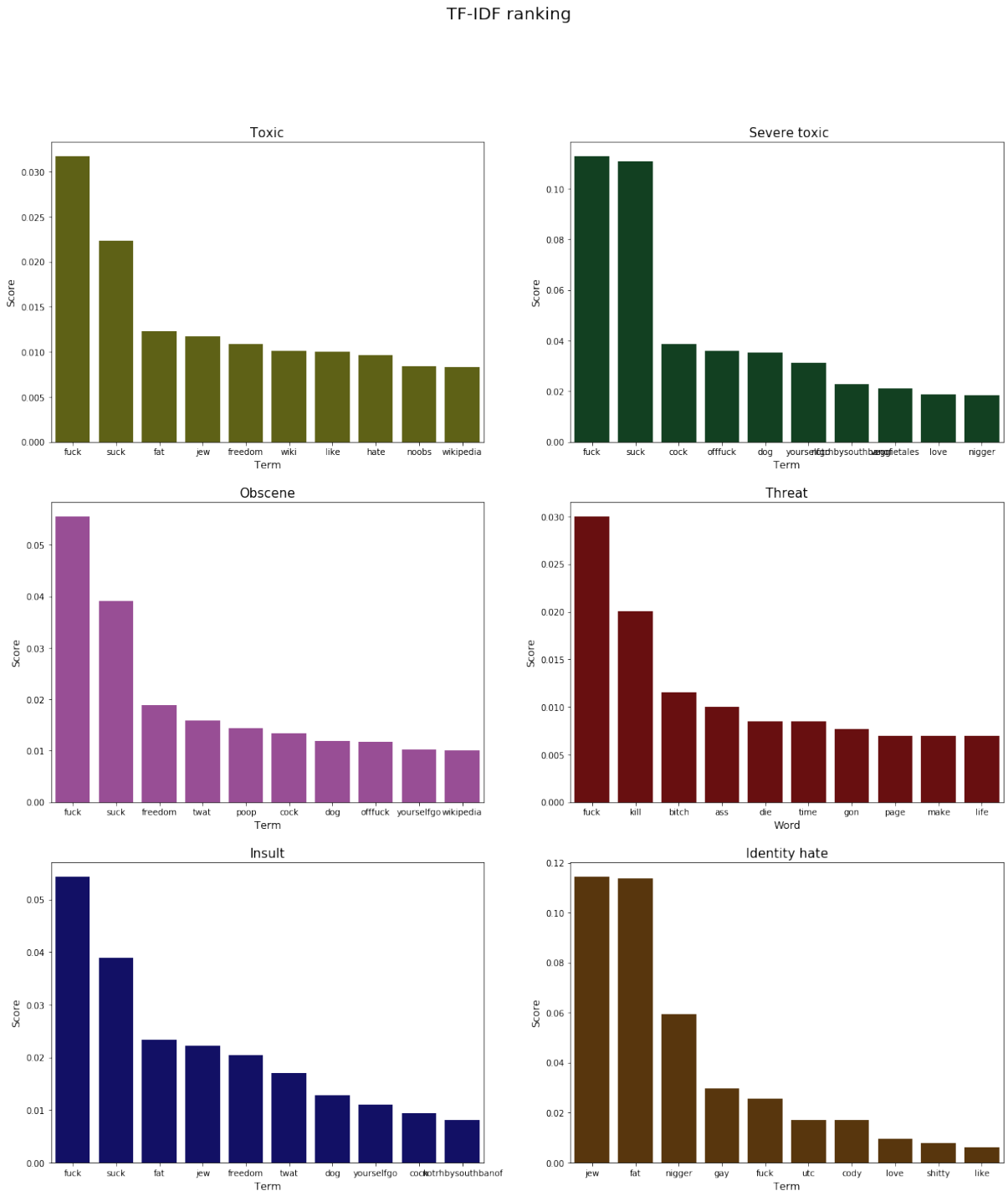


Figure 5: IF-IDF Ranking

Conclusions:

Working on this project, we realized that identifying abuse from tweets is a difficult process if we do not have labelled tweets readily available or if we cannot label the tweets manually.

Once we got the labelled train data, our analysis became much easier and we could categorize the tweets into the six categories such as *toxic, sever toxic, threat, insult, identity hate and obscene*. We had enough train data to train our classifier better, which performed better than a random classifier. I have achieved an accuracy of over 0.7 for each of the categories and I think this result is acceptable for my dataset.

However, I am not completely satisfied with the results as the predictions are a little haphazard due to overly cleaning the tweet text. I lost the actual meaning of a few text due to this and I should have been more diverse in my tweets selections to be able to identify the differences in the words in each of the tweets. I would consider my classifier does classify properly but, the way I built the pre-processing text removed certain words without which identification of a tweet into abused or abusive or observer category is not possible for verification, post classification.

Although the classifier does perform better than a random classifier, it could be better if I could get properly labelled data from twitter text for specific hashtags. For now I could get the classifier classify the text into the six categories, post which I want to break them down further into abused or abusive based on even very minor non trivial semantics like the tone of a sentence. I think I overdid my pre-processing and lost the context of tweets, I should tune down the steps to include words even of lengths 2 and consider the smileys mentioned in the tweet text's too to identify the tone of the text.

I assume now the semantics of a text for any of the hashtags is a little different and needs to be tackled in a precise manner to accurately categorize the tweets into abuse or abused. So, if I could get a data set with diverse topics or context, I would have a bag of different words and the probabilities of a word belonging to the abuse or abused versions would be easier to classify.

As next steps, I would like to get tweets from various kinds of hashtags, apart from #bodyshaming, like #mentalhealth, #abuse, #youknew, #youweresilent and so on which are closely related to bodyshaming to get the diverse range of words and accurately train the classification model to get better predictions on my test data. Also, I want to prepare and execute a machine learning algorithm which can identify the label of a tweet, instead of me manually labelling them. I would like to get the algorithm to label the tweets based on a set of words

which could be identified by Afinn Score and build my own sentiment vocabulary to identify the semantic orientation of the tweet. This would help me in labelling my train tweets and I can be sure of the results of my classifier.

Individual Tasks:

I worked alone on this project under the guidance of Prof. Vincent Malic. I collected Twitter data for the hashtag bodyshaming for a period of one year and three months, starting January 2017 and ending 31 March 2018. To get the labelled data, I used a train data set from Kaggle data sets [13] for pre-labelled texts which could be used in training my classifier. As planned earlier I calculated a polarity score for the tweets which told me whether a tweet has a positive or negative connotation. However, I didn't use this score in my analysis and instead I fit a model which would predict the categories each of the tweets would fall into and analyzed if we can truly identify abuse through twitter.

I worked on a similar project in Social Media Mining course I have taken this semester on a project called *"Study of political abuse on Twitter during U.S. House Elections 2017"* under Professor Vincent Malic. The project is based on identifying the misuse of social media to propagate fake news, its popularity among the voting community and its corresponding effect on the recent U.S. House election results. Even though it looks similar, the methods employed to classify tweets text is different. In Data Mining project, I tried to analyze different categories a tweet could be categorized into which would help in identifying versions of abuse in tweets and to identify if twitter data is enough to identify the instances of bullying.

References

[1] Abuse Definition
`https://www.merriam-webster.com/dictionary/abuse`

[2] Types of Abuse
`https://reachma.org/6-different-types-abuse/`

[3] Abuse on social media
`https://www.theguardian.com/media/2017/dec/13/what-can-be-done-about-abuse-on-social`

[4] An In-Depth Analysis of Abuse on Twitter by J. Oliver, P. Pajares, C. Ke, C. Chen and Y. Xiang, Trend Micro 225 E. John Carpenter Freeway Suite 1500 Irving Texas 75062 U.S.A. Tech. Rep., September 2014

[5] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A. and Menczer, F., (2011), ICWSM. [Google Scholar]

[6] Detecting Offensive Language in Social Media to Protect Adolescent Online Safety by Y. Chen, S. Zhu, Y. Zhu, H. Xu: Symposium on Usable Privacy and Security, Pittsburgh, USA (2011)

[7] Predicting Depression via Social Media by M. Choudhury, M. Gamon, S. Counts, E. Horvitz, Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, Boston, MA, July 8-11.

[8] Twitter Bullying Detection by H. Sanchez and S. Kumar, ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 15–15.

[9] Sentiment Analysis of Twitter Data: A Survey of Techniques by V. Kharde S. Sonawane (2016)

[10] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media, pp. 30–38 (2011)

[11] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011)

[12] Understanding and Fighting Bullying with Machine Learning by Junming Sui
`http://pages.cs.wisc.edu/~jerryzhu/pub/junming-thesis.pdf`

[13] Kaggle Data set
`https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data`

[14] Natural Language Processing:
`https://en.wikipedia.org/wiki/Natural_language_processing`

[15] Natural Language Toolkit:
`http://www.nltk.org/`

[16] Preprocessing using NLP:
https://nlp.stanford.edu/IR-book/html/htmledition/
dropping-common-terms-stop-words-1.html

[17] Scikit Learn Documentation:
http://scikit-learn.org/stable/documentation.html

[18] NumPy Documentation:
http://www.numpy.org/

[19] Pandas Documentation:
https://pandas.pydata.org/

[20] Logistic Regression:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.
LogisticRegression.html

[21] Plotly Documentation
https://plot.ly/python/

[22] The Book
Hands-On Machine Learning with Scikit-Learn and TensorFlow

[23] GetOldTweets-python
https://github.com/Jefferson-Henrique/GetOldTweets-python