THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

COLLEGE OF INFORMATION SCIENCES & TECHNOLOGY

Using Data Mining to Detect Hate Speech on Twitter

JOSHUA CICCARELLI
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Data Sciences
with honors in Data Sciences

Reviewed and approved* by the following:

Anna Cinzia Squicciarini
Associate Professor of Information Sciences & Technology
Thesis Supervisor

John Yen
Professor in Charge of Data Sciences in IST
Honors Advisor

* Electronic approvals are on file.

ABSTRACT

The intention of this experimental study is to create a machine learning algorithm that can accurately classify tweets as malicious or not malicious. The dataset used for the first experiments is 80k tweets that were gathered using Twitter's API. Each row contains features such as text, follower count, reply text, reply count and more. This dataset was processed for text features and was used as training data to train a machine learning algorithm to automatically detect cyber hate speech. The results were measured by calculating the root mean squared error of the testing data and using this as a factor to compare models. The dataset used for the second experiment contains 24K tweets and has labels and counters for each tweet. Overall, the goal of this study is to further research regarding machine learning on social media and to help social media platforms detect vulgar content. There exists a large amount of malicious content on social media platforms, especially Twitter, and a well-trained machine learning model can detect this content so it can be removed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge three people who helped me conduct my research, write my thesis and succeed as a Schreyers scholar. To Professor Anna Squicciarini, my supervisor and my data science professor, I am thankful for your guidance and advice during my research and data science classes. I gained my most valuable experience with data science while working with you and learning from you. To Markus Huff, my classmate and research assistant, thank you for helping me throughout my research and for helping me succeed. To Dr. John Yen, thank you for being my honors advisor and giving me guidance with my classes and work. I would have not been able to accomplish what I have accomplished without the three of you and am grateful for your help.

Chapter 1

Introduction

Over the past decade, the rise of social media has led to the constant creation of material online for people to view and give their own feedback. However, this constant stream of content results in vulgar and malicious material that shouldn't be allowed to be posted on social media platforms. Recently, especially on platforms like Twitter, there has been a surge to create a better detection algorithm and safeguard measures to find and report this vulgar material. To put it into numbers, the Huffington Post has reported that 15,000 malicious tweets are sent every day and nearly 100,000 every week (Fitzgerald). These numbers go to show that cyber-aggression and cyber-bullying on Twitter is a relevant issue and will remain a relevant issue for the years to come.

The key variables that will be used to evaluate and form a machine learning algorithm are as follows: text, hashtags, retweet count, follower count, reply count and reply text. These are the most informational portions of a tweet and allow for identification of continued cyber-aggression. The central variable will be the text in each tweet because this is where the actual hate speech can be identified. The purpose of this study is to apply machine learning models that can properly detect whether a tweet contains malicious content, hate speech, vulgar language, or offensive content, and then sort these malicious tweets depending on if they are cases of cyber aggression or bullying. Cyber-bullying in its most basic sense is simply cyber-aggression that happens repeatedly between two users with a power imbalance.

This study will contribute to the field by creating a solution to a major social media problem that the world currently faces. Additionally, it will look into twitter datasets that use a reply count and evaluate reply text which has not been done on a large scale yet. Even if the study fails at creating an accurate algorithm, the publication of the process will help the field advance.

Chapter 2

Literature Review

In the article, “Deep Learning for Hate Speech Detection in Tweets,” the authors are experimentally researching hate speech detection by using several deep learning methods on a 16k annotated tweet dataset. The question that guided the study was whether the use of different deep learning methods could make a significant impact on hate speech detection and if so, which one is the best. The study tested this by using CNNs, LSTMs, and FastText on the same annotated dataset and seeing how they performed. The only independent variable used was the actual text from the tweet; however, the authors broke it down by using a bag of words classifier and dictionary so that they could extract features. In the end, the study found that CNNs performed the best all-around and that combining multiple deep learning methods led to poorer results rather than just using one method. The results were measured by the RSME accuracy of each algorithm.

This article provided valuable information because it explained that using more than one method is not a feasible solution. Additionally, it gives information that will direct my own study because we will begin by testing random forest classifiers and convolutional neural networks rather than LSTMs or FastText. The study was slightly shorter than I think it should have been because it should have provided more information regarding the procedure of the actual experimentation.

In the article, “Mean Birds: Detecting Aggression and Bullying on Twitter,” the authors discuss the differences between cyber-bullying and cyber-aggression before they perform an experiment using a twitter dataset and machine learning. They are looking into whether they can

create a better detection algorithm as well as if accounts they deem as aggressive have been taken down by twitter. The researchers only used the text as their independent variable; however, they used bag of words and other tools to extract features such as sentiment, word embedding, basics, power difference popularity and reciprocity. They performed tests on their dataset by using random forest classifier and tree-based approaches to create a detection algorithm and analyze it. The results of the study were that many accounts they flagged as aggressive were still active on twitter. Additionally, they identified many difficulties with differentiating between aggression and bullying. For instance, the authors explain how users displaying cyber-bullying "are not very active as per number of posts overall, [but] when they do become active, they post more frequently than typical users, and do so with more hashtags, URLs, etc" (Chatzakou).

The Mean Birds article is very informative because it is the most closely related paper to my study. The authors provide a very clear step by step process for data collection which was very helpful when my team was collecting our own dataset. Cyber aggression and cyber bullying are hard to differentiate from each other, but, hopefully, the use of reply text and reply count will help to ease this differentiation.

In the article, "What a B!tch!: Cyber Aggression Toward Women of Color," the authors want to provide a deeper understanding of how hate speech tweets are developed and how they spread in the social media community. The authors managed this by creating a 24k tweet dataset over the span of a week in 2018 and extracting features for their study. Again, the only independent variable used was the text from the tweet, but additional features can be extracted from the text. The actual method of annotation in this study was manual rather than through the use of deep learning. Therefore, the results of the study were the most accurate out of any of the

five studies because machine learning makes mistakes that manual annotation doesn't such as with sarcasm. The results of this study were that there is a large amount of hate speech tweets that is being allowed to remain on twitter. The authors used classifiers that identified if the tweet contained racism because that was their field of study; however, the useful part of this paper for my study is the data collection and statistics section, not their own results.

This is an interesting article because it takes a different approach and sheds some light on more social than technical aspects of my study. The article described the parameters they used and explained how they choose keywords to obtain the most relevant dataset that they could. Additionally, analyzing tweets by race may give more information regarding tweet differentiation of aggression vs bullying. The use of their labels and their features will provide some references during our own training and could possibly result in a higher accuracy if we choose to use their labels.

In the article, "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," the authors discuss the use of crowdsourcing to analyze tweets and also the various labels that can be used in twitter research. They sought out to create an accurate algorithm to detect numerous labels rather than hate vs normal and they wanted to see which labels are the most relevant. The authors used numerous filtering techniques on the dataset to clear out unwanted independent variables from the data. Additionally, they used a boosted sampling method on the data that allows for a higher amount of hate speech tweets to be learned from. The method used by the authors was crowdsourcing rather than machine learning methods. They did this to evaluate how useful crowdsourcing would be in the future and whether it could be used on the training data to give models a more accurate label column to learn from. The result of the

study was that crowdsourcing is a highly valuable method. Additionally, the researchers made a coherent and clear outline of their steps and methodology so that someone could repeat their experiment which was a clear goal of theirs.

This article provides clear methodologies that my team followed during data collection and gave us the idea to use crowdsourcing. Despite this, my team was able to find a partially annotated tweet dataset that we choose to base our classifiers off of because it was easier to use. However, this article gave us ideas regarding sample boosting and crowdsourcing that we plan to implement later on.

In the article, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," the authors explain how they labeled a 16k tweet dataset. Unlike the other papers, this study uses critical race theory to label their tweets so that their labels can be more accurate and have a clear explanation behind them. They use the text as their main independent variable; however, they extract features via indicators of the text. For instance, if the text has a racial slur it will have a racial slur indicator or if it mentions a minority then it will have a minority indicator and so on. The authors use a combination of critical race theory and logistic regression with 10-fold validation to create a model on the data. The actual model that they created doesn't seem to be noteworthy but the labeling method they used is very informative. The study found that using gender information improves accuracy of models; however, location and length data reduces the accuracy of the model.

The findings of this paper are extremely important because the study did a lot of work for my team already by limiting the amount of feature extraction that will be needed. Locational data is an attractive feature, but the researchers clearly state that they tested it and that

it is detrimental to model accuracy. This paper will help my team to narrow our focus regarding features and raises important questions regarding the inter-connectedness of demographics with hate speech.

Chapter 3

Cyber Aggression & Cyber Bullying

Cyber Aggression and Cyber Bullying have become more of a problem in the last decade because social media makes it easy for people to attack one another. Especially on platforms like Twitter, there is a lack of accountability and detection which enables malicious users to send aggressive messages to each other. The goal of this study is to create a classifier that can identify if there is aggressive material in a tweet and differentiate between cyber aggression and cyber bullying. According to Robin Henderson, Psy.D. and Chief Executive of Behavioral Health Services at Providence Oregon, “The difference between the aggression and bullying is that cyber-aggression can be a one-off situation or remark that the perpetrator doesn’t necessarily know is wrong, whereas cyberbullying is a malicious and targeted approach where the perpetrator has an intent to harm the other person” (Team). It is important to differentiate between cyber bullying and cyber aggression because cyber aggression can be an indicator that the account is becoming malicious while cyber bullying is a clear indicator that the account is malicious. Social media platforms that implement a detection system to identify cyber bullying and cyber aggression will be more equipped to deal with hate speech and malicious users.

Role	Description
Perpetrator	Sender of aggressive message
Reinforcer	Likes or favorites an aggressive message; Sends message in support of perpetrator
Victim	Target of cyber aggression (by self-admission or direct targeting)
Defender	Likes or favorites messages against perpetrator; Sends message in defense of victim
Bystander	Part of conversation that includes an aggressive message, but takes no role in perpetuating or alleviating aggression
Informer	Reports aggression to site administrator

Table 3.1: Cyber Aggression Participants

Felmlee, Diane. "The Social Networks of Cyberbullying on Twitter: Concepts, Methodologies, Tools, and Applications." ResearchGate, Jan. 2019, www.researchgate.net/publication/330128155_The_Social_Networks_of Cyberbullying_on_Twitter_Concepts_Methodologies_Tools_and _Applications.

Table 3.1 contains the participants involved in cyber aggression. The experiments in this paper are attempting to identify perpetrators and reinforcers so that their accounts can be blocked, and their posts can be deleted. Identifying Victims and Defenders is not as important because they are not posting malicious material and they are merely receiving hateful content.

Chapter 4

Overview of Data Mining

Data mining is the process of analyzing raw data to extract information and identify patterns that the data contains. It is comprised of three disciplines: statistics, artificial intelligence, and machine learning. “Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed” (Expert.ai Team). The experiments that were conducted during the research for this thesis used machine learning classifiers to learn from training data so that the system could detect occurrences of malice in future tweets it processed.

Data mining is becoming increasingly important because the volume of data produced is doubling every two years. Technological advancements have increased the access to technology in the world and users are generating enormous amounts of data every day. Data has become very valuable and data mining allows for this data to be sifted through and made sense of. It allows data analysts to use the information to assess outcomes and make informed decisions and predictions.

Training/Testing

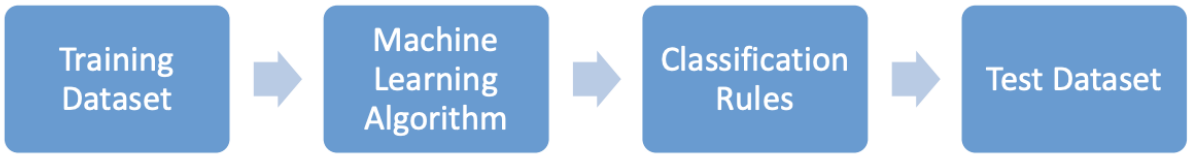

Figure 4.1: Machine Learning Processing

When training a machine learning algorithm, the first step is to divide the data into a training dataset and a testing dataset. The typical split is for 80 percent of the data to be used for training and twenty percent of the data to be used for testing. By splitting the data, the data analyst can validate the accuracy of the algorithm by feeding it the testing data after the algorithm is trained and calculating its accuracy. The sklearn library offers a method, train_test_split(), in python that makes this easy and allows for randomized splits. The analyst can extract different features from the data to make the model more accurate and can alter the parameters of the algorithm to fine tune its performance.

Overfitting

Overfitting is when a machine learning algorithm learns too much of the details and noise in the training and testing dataset which decreases its performance when used on different data. When the algorithm is used on new data, the model will fail to accurately classify the data because it will look for specific patterns that were only in the training dataset and the original testing dataset. Overfitting becomes more likely when more features are used because the algorithm is more likely to identify specific patterns across the features that are unique to the training dataset. Therefore, in these experiments, the goal was to identify the key features for identifying malicious tweets so that the model could be as accurate as possible for future tweets that the algorithm processed.

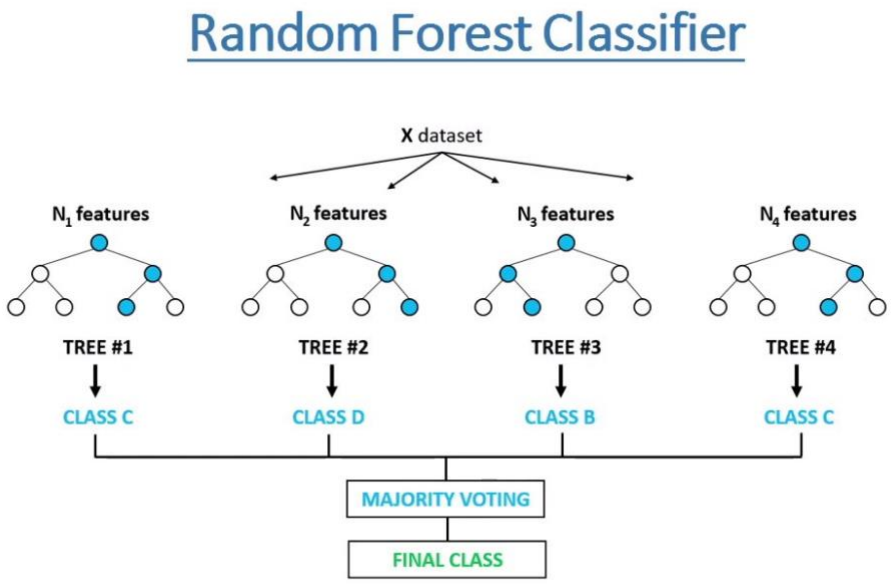

Figure 4.2: Random Forest Classifier Example

Random Forest Algorithm

Random forest algorithm is one of the most widely used machine learning algorithms because it produces high accuracy results without complex parameter tuning. It is a supervised learning algorithm, it requires labels, that creates an ensemble of decision trees to make classifications or predictions. “Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model” (Donges). In the experiments conducted during research, random forest classifier proved to be the most accurate out of all the machine learning algorithms used. It consistently made more accurate predictions for both of the datasets used and did better than other popular algorithms such as xgboost and adaboost classifier. Figure 4.2 is a diagram that shows how random forest classifiers use an ensemble of decision trees to create an algorithm

that can accurately classify data. Random forest combines numerous decision trees with a specified max depth to create a final class which can perform classification or regression.

Chapter 5

Hypothesis & Method

Hypothesis

The five articles above, especially the articles written by Diane Felmlee, give information regarding the collection of twitter data, the sampling of the data, the feature extraction of the data and the analysis of the data. The articles summarize the thought-process behind using specific labels and how these labels came to be. Additionally, they explain how they selected their features, and which features to avoid during the training of the model. The mean bird's paper, especially, describes the differences between aggression and bullying and how this can be measured using only text data. Overall, these papers have many similarities; however, each one describes a different aspect of our research and gives helpful information on how to conduct our own research.

Hate speech detection on twitter can be accurately classified into cyber-aggression and cyber-bullying by extracting power imbalance, sentiment, and reply features that will result in a model that will hopefully further research on hate speech detection.

Method

Steps to conduct a study with the goal of creating a machine learning cyber-bullying detection algorithm for social media.

To begin this study, a thorough literature review of 5-10 current journals regarding the progression of automatic cyber-bullying and cyber-aggression detection algorithms on social media is required. This step resulted in information about using the Twitter API, the requirements of cyber-bullying posts, the relevancy of this topic, different machine learning

algorithms that can be used and tips on how to approach this type of study. Overall, the review of current work is important because it provides necessary information and guidance that can't be found elsewhere.

Next, a dataset needs to be compiled using the Twitter API so that it can be processed using the sklearn and pandas library, https://pandas.pydata.org/. The dataset should include text, datetime, twitter id, reply count, reply tweets, number of followers and number of retweets. The combination of these features with additional features extracted from the text is a hefty and informational set that can be used to train a model. These features can be gathered after a Twitter developer account is created and a script is run in python using libraries such as Tweepy, Twitter, and SearchTweets. Datetime can be broken down using the datetime library; and sklearn's count vectorizer can be used to extract text features. After this is done, a model can be trained which will accurately predict and classify future tweets based on their provided information.

The final step of the study is to investigate which classification algorithm should be used and to apply it. The most likely models that will be used is either a random forest classifier or xgboost gradient boosting because they are both highly accurate and versatile models that can learn from highly dimensional data. These models will be tested thoroughly with the data by changing parameters and features so that the highest accuracy can be achieved. These libraries can be reviewed online on the library website or on stack overflow. The results will be analyzed by measuring Root Mean Squared so that findings can be discussed, and full automation will be attempted if the study is accurate so that it can be sent to Twitter for possible implementation. The number of participants for this study is simply the number of twitter users that tweets are collected from which in this case is eighty thousand.

Chapter 6

Datasets

Experiment I Dataset

The dataset used for Experiment I, Dataset I, is a large dataset of malicious, normal and spam tweets that were pulled from Twitter's API in March of 2017 using a variety of keywords with the Search Tweets library, https://github.com/twitterdev/search-tweets-python. The dataset has 38,961 rows and 16 columns which include the following features: Datetime, Twitter_Id, Label, Hashtags, Text, Truncated, User_Mentions, Quote_Status, Retweet_Count, Retweet, User_Id, Location, Friends, Followers, List_Count, and Status_Count. Dataset I was selected to be used for the first experiment because of its length, variety of features and labels. Since it was already labeled, our research team was able to proceed into feature extraction and training without the need for crowdsourcing. Additionally, the Friends and Followers features made it possible to calculate the popularity of the user and classify the user into bins based on their popularity. Lastly, the inclusion of the Text made it possible to extract text features and the Retweet_Count feature made it easier to detect cyber aggression since cyber aggression can be as simple as retweeting a malicious tweet.

Label	Text	Hashtags
normal	I deserve a yoga retreat in Thailand with a yo...	[]
normal	if there is no faith? \nLet the Qutubas be use...	[{'text': 'Ø¹Ø§Ù…_Ø¹Ù„ÙŠ_Ù…Ù†Ø¹_Ø§Ù„Ù‚Ù†Ø§Ø',...
normal	Yes, but don't expect than other that peace an...	[]
abusive	Hate wasp and bee season, wasps keep flying in...	[]
spam	It's sooooooooo flattering to me when folks th...	[]
...	...	...
normal	if you look carefully on his forehead you can ...	[]
abusive	My mom: That girl in the car is so pretty, her...	[]
spam	I absolutely hate when grown ass adults act li...	[]
normal	Oh, the irony!! Kentucky Coal Museum installs ...	[]
normal	RT @JeremyPlattform: "You have to admire #HapA...	[{'text': 'HapAndLeonard', 'indices': [41, 55]}]

Figure 6.1: Experiment I Dataset Subset

Experiment II Dataset

The dataset used for Experiment II, Dataset II, is a dataset of hate speech, offensive and normal tweets that were pulled from Twitter's API at an unknown date using keywords. The dataset has 24,783 rows and 6 columns that include the following features: count, hate_speech, offensive_language, neither, class, and tweet. The important features for this dataset are hate_speech, offensive_language, class and tweet. The hate_speech and offensive_language columns are counters of the number of occurrences of hate speech or offensive language. The class column is a numerical classifier that labels the tweet as offensive, hate speech or neither. Lastly, the tweet column is the text of the tweet including mentions and hashtags.

tweet	Curse_Count	@	text_length
!!! RT @mayasolovely: As a woman you shouldn't...	0	1	25
!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	0	1	16
!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	3	1	21
!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	0	1	9
!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	2	1	26
...	...	...	...
you's a muthaf***in lie “@LifeAsKing: @2...	0	1	19
you've gone and broke the wrong heart baby, an...	0	0	13
young buck wanna eat!!.. dat nigguh like I ain...	1	0	13
youu got wild bitches tellin you lies	1	0	7
~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	0	0	18

Figure 6.2: Experiment II Dataset

Chapter 7

Experiment I

For this experiment, the goal was to create a machine learning classifier that could accurately identify if a tweet was spam, normal or malicious. If the classifier was able to accurately classify malicious tweets, the next step would be to classify the malicious tweets into subcategories of cyber aggression, cyber bullying, and neither.

Feature Extraction

Type	Feature
User (total: 10)	avg. # posts, # days since account creation, verified account # subscribed lists, posts' interarrival time, default profile image? statisticson sessions: total number, avg., median, and STD. of their size
Textual (total: 9)	avg. # hashtags, avg. # emoticons, avg. # upper cases, # URLs avg. sentiment score, avg. emotional scores, hate score avg. word embedding score, avg. curse score
Network (total: 11)	# friends, # followers, hubs, (d=#followers/#friends), authority avg. power diff. with mentioned users, clustering coefficient, reciprocity eigenvector centrality, closeness centrality, louvain modularity

Figure 7.1: Twitter Types and Features

For feature extraction, our research team used the features and text in Dataset I to create new features that could be used to train a random forest classifier algorithm. We imported the csv of tweets into python and used pandas to create new features for the classifier. We used Figure 8.1 as a reference during feature extraction to decide which features to extract and which features would be relevant to our research. Power imbalance is a difficult element of tweets to identify so instead, we used a friends/follower's ratio to measure the popularity of the user. Users who had a smaller value for this ratio were considered to be prominent users since their followers outnumbered their friends.

Feature	Description	Likelihood of Importance
Text	Cleaned text of the tweet	Likely
Follower Label	A follower classification that is ranked 1 -3	Unlikely
Curse Count	The count of curse words in the text	Very Likely
URL Presence	The presence of URL in the tweet	Unlikely
Retweet	Whether the tweet was retweeted	Likely
Uppercase Count	The count of uppercase characters in the tweet	Likely
Uppercase Presence	The presence of URLs in the tweet	Unlikely
Hashtag Count	The count of hashtags in the tweet	Likely
Hashtag Presence	The presence of hashtags in the tweet	Unlikely
Friends/Followers Ratio	The ratio of Friends/Followers for the sender	Very Likely

Table 7.2: Feature Extraction Table

Hashtag count was created by parsing the hashtag column with a for loop that counted the number of elements in the list of hashtags for each tweet. Hashtag presence was then created by parsing the hashtag count feature and creating a column with a binary value which represented the presence of hashtags. Curse Count was created by importing a list of social media curse words created by google and counting the number of matches in the text while parsing it with a for loop. The list included alternate spellings of curse words that are found on social media such as replacing the letter, I, with an exclamation point. The follower label was created by categorizing the number of followers into unpopular, average, and popular which were represented by numerical values. For uppercase count and uppercase presence, the text was parsed using a for loop and we used the NLTK library with a counter to create these features. Lastly, URL presence was created by parsing the text and friends/follower's ratio was created by

dividing the number of friends by the number of followers. Figure 8.2 shows the features we extracted, their description and the likelihood of importance for the random forest algorithm.

Results

The results of this experiment showed the difficulties of training a model to classify tweets into multiple categories. The model was unable to accurately classify the tweets using a random forest classifier, or adaboost classifier. The classifiers failed because they were not able to classify tweets as containing malicious content with a high enough accuracy to be helpful to social media developers or researchers. We tried to use bag of words to create more text features that the model could learn from, but the bag of words decreased the accuracy. When using bag of words, there was too many features because of the wide array of language used in tweets which caused the model to perform worse. I used k-fold cross validation to run experiments using the bag of words as a feature and the accuracy of the models ranged between 48.775 percent and 66.05 percent when using the bag of words. The tests using the bag of words were modeled with a Random Forest ensemble and I changed the “max_depth” parameter throughout the experiment. I found that the accuracy of the model was highest for depths 5-10 and then the accuracy decreased for depths higher than 10. The most important feature for detecting malice was Curse Count because a high number of curse words are used in most malicious tweets. The friends/followers ratio was not an important feature despite the information in the literature reviews that pointed to power imbalance being an important feature. The features that were extracted from the text such as hashtag, url, and uppercase offered little information gain for the Random Forest.

Additionally, I trained the data using ADA Boost Classifier, which is a gradient boosting algorithm, but the performance was worse than Random Forest. I expected the ADA Boost Classifier to improve the accuracy of the model because it combines multiple classifiers to create a more accurate classifier. While using the same features, the RSME accuracy for ADA Boost Classifier ranged from 51.3 percent to 61.5 percent while the Random Forest accuracy was 66.05 percent. However, when using bag of words in addition to the same features, Random Forest's accuracy decreased to 22 percent. I varied the n estimators' parameter for ADA Boost Classifier between 100 and 200, and 100 n estimators resulted in the higher accuracy. Overall, Random Forest performed the best when it was trained with two to four text features; bag of words decreased the accuracy because it created too many features.

```
def train_estimators(n_estimators):
    clf = RandomForestClassifier(n_estimators = n_estimators, max_depth=5, criterion = 'gini', random_state=0)
    clf.fit(Xtrain, Ytrain)
    pred = clf.predict(Xtest)
    acc_list.append(accuracy_score(Ytest, pred))
```

Figure 7.3: Random Forest Classifier Method

```
def train():
    clf = AdaBoostClassifier(DecisionTreeClassifier(max_depth=5), n_estimators=100, random_state = 0)
#    clf = AdaBoostClassifier(n_estimators=100, base_estimator = LR, random_state=0)
#    clf = RandomForestClassifier( max_depth=5, criterion = 'gini', random_state=0, class_weight = 'balanced')
    clf.fit(Xtrain, Ytrain)
    pred = clf.predict(Xtest)
    acc_list.append(accuracy_score(Ytest, pred))
```

Figure 7.4: ADA Boost Classifier Method

Chapter 8

Experiment II

For this experiment, the focus was to create a machine learning model that could classify tweets with an accuracy higher than 70% which Experiment I was unable to accomplish. The dataset for this experiment came with labels that classified the tweets as hate speech, offensive or neither. The dataset I used can be found at the following url: https://data.world/thomasrdavidson/hate-speech-and-offensive-language

Feature Extraction

For this experiment, I used the insight of the last experiment and used a limited number of features. I choose not to use bag of words because it reduced the accuracy of the model in Experiment I. Using the Google list of social media curse words, I created a Curse_count feature similar to Experiment I because it was the primary indicator. The dataset included a feature, count, which was the total count of indicators that were counted by crowdsourcing. However, I choose to only use this feature in the control test because this feature would not be available in future experiments and it is subjective. Next, I used a for loop to calculate the number of elements once the tweet was split by spaces and then created a mention (@ sign) presence feature to include some power imbalance. If a tweet has mentions, then it is more likely to be directed at someone which could be an indicator of cyber bullying or cyber aggression. Although it could be un-related to malice, it was worthwhile including and seeing if this feature increased or decreased the accuracy of the model. Lastly, I choose to use only Random Forest Classifer for Experiment II because Experiment I showed that Random Forest outperformed the other classifiers (Logistic Regression, XG boost and ADA Boost Classifier).

Results

Features	Label	Classifier	Depth	Accuracy
Curse_count, count	class	random_forest	5	0.7635
Curse_count, text_length	class	random_forest	5	0.7624
Curse_count, text_length, at sign presence	class	random_forest	5	0.7649
Curse_count, count	class	random_forest	7	0.7635
Curse_count, text_length	class	random_forest	7	0.7645
Curse_count, text_length, at sign presence	class	random_forest	7	0.7661
Curse_count, count	class	random_forest	10	0.7635
Curse_count, text_length	class	random_forest	10	0.7643
Curse_count, text_length, at sign presence	class	random_forest	10	0.7635
Curse_count, count	class	random_forest	50	0.7635
Curse_count, text_length	class	random_forest	50	0.768
Curse_count, text_length, at sign presence	class	random_forest	50	0.768

Figure 8.1: Experiment II Log

The curse count was the most important indicator of malice in a tweet, so the curse count feature was used in all 12 tests of this experiment. Text length and at sign presence increased the accuracy of the model by a tenth of a percent. Although these features increased the accuracy of the model, the increase in accuracy was minimal and might have been specific to this dataset. For parameter tuning, the model became more accurate when its max_depth was increased. The accuracy increased from 76.49 percent to 76.80 percent when the max_depth was increased from a depth of 5 to a depth of 50. Although the accuracy increased, it was only an increase of .31 percent which could be specific to this dataset and not twitter datasets in general. I think that a max depth of 10 is the best depth to use for twitter malice detection because 10 was the best depth for Experiment I and the accuracy did not increase by much when using a depth of 50. Overall, the takeaways from this experiment are that Curse count is the most important indicator of malice, using fewer features results in a higher accuracy and a max depth of 10 is the best depth to use for Random Forest classifier.

```
clf = RandomForestClassifier(max_depth=50, random_state=0)
clf.fit(Xtrain, Ytrain)
pred = clf.predict(Xtest)
print(accuracy_score(Ytest, pred))
```

Figure 8.2: Random Forest Classifier Code

```
clf = AdaBoostClassifier(DecisionTreeClassifier(max_depth=5), n_estimators=100, random_state = 0)
clf.fit(Xtrain,Ytrain)
pred = clf.predict(Xtest)
print(accuracy_score(Ytest, pred))
```

Figure 8.3: ADA Boost Classifier Code

Chapter 9

Analysis

Comparison Analysis

The results of Experiment I and Experiment II showed that the main indicator of malice is a high count of curse words. This finding is intuitive and it Although other features increase the accuracy of the model, these features minimally increase the accuracy and are most likely specific to the dataset. Random Forest Classifier performed the best out of the classifiers and consistently produced the most accurate classifications. It outperformed ADA Boost Classifier, XG Boost, and Logistic Regression. The Random Forest Classifier performed best when the max depth parameter was set to 10. Although the performance was slightly better when using a max depth of 50 in Experiment II, I think that this was specific to this data and that 10 is the best depth. Power imbalance did not turn out to be a useful feature, but this study only used a friends/follower's ratio to measure power imbalance. It did not take into account the power of the user receiving the tweet which could improve the importance of the feature. The results of these experiments showed that a fewer number of features is better for accuracy because it is easier for the model to learn from, and it reduces overfitting.

Data Analysis

In the future, there are several areas that should be explored to further research regarding machine learning on social media. Firstly, the correlation of features needs to be analyzed to better understand which features can be combined to accurately measure a power imbalance. Currently, the understanding of a power imbalance between users during cyber-bullying relies mostly on number of followers and replies; however, word choice and use of hashtags could

implicate a power imbalance as well. A power imbalance is a key indicator of whether the hate speech is an instance of bullying or aggression so this would be an important topic to research.

Additionally, research should be done to transfer the knowledge learned from twitter machine learning efforts to other social media platforms. Although there has been some research on other platforms, it has not been as thorough or widespread as twitter research. If more is done to create algorithms on Facebook or Instagram, then new information might be discovered that could benefit the topic.

The goal of this study was to be able to classify tweets as normal, cyber aggression or cyber bullying, but the study failed to accomplish this. For future research with this same goal, I would recommend creating a dataset that is able to measure power imbalance by creating a graph dataset that connects tweets to its senders and receivers. Also, I recommend having a feature that is a counter for how many tweets the user has sent that contain malicious content. The combination of power imbalance, a tweet counter and curse word count may be informational enough to classify tweets as cyber aggression and cyber bullying.

Chapter 10

Discussion

Limitations regarding construct validity could be an issue for this study because all of the features will hold some sort of relationship; however, not all of them will be helpful. Since each feature is derived from the same tweet, correlations will automatically occur. Therefore, researchers need to use tools regarding feature importance to analyze which features are actually the most important. However, the tools currently in place are often not enough to differentiate between feature importance so some trial and error is required. In the future, new tools will most likely be configured and this trial-and-error step can be avoided.

Additionally, data gathered from Twitter's API have to be gathered during certain periods of time and have to be searched via hashtag or topic. However, this slightly biases the dataset and can cause for accuracy within the model to not transfer over to a real-life scenario. Therefore, in the future, researchers could work on combining Twitter datasets into a million-count tweet dataset so that more reasonable models can be generated.

Bibliography

Badjatiya, Pinkesh, et al. “Deep Learning for Hate Speech Detection in Tweets.” Proceedings of the 26th International Conference on World Wide Web Companion WWW 17 Companion, 2017, doi:10.1145/3041021.3054223.

Chatzakou, Despoina, et al. “Mean Birds.” Proceedings of the 2017 ACM on Web Science Conference - WebSci 17, 2017, doi:10.1145/3091478.3091487.

David, Davis. “Random Forest Classifier Tutorial: How to Use Tree-Based Algorithms for Machine Learning.” FreeCodeCamp.org, FreeCodeCamp.org, 13 Aug. 2020, www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/.

Donges, N. (2019, June 16). A complete guide to the random forest algorithm. Retrieved March 03, 2021, from https://builtin.com/data-science/random-forest-algorithm

Expert.ai Team. (2021, February 11). What is machine learning? A definition - expert system. Retrieved March 02, 2021, from https://www.expert.ai/blog/machine-learning-definition/

Felmlee, Diane, et al. “Chapter 6 What a B!Tch!: Cyber Aggression Toward Women of Color.” Advances in Gender Research Gender and the Media: Women’s Places, 2018, pp. 105–123., doi:10.1108/s1529-212620180000026008.

Felmlee, Diane. “The Social Networks of Cyberbullying on Twitter: Concepts, Methodologies, Tools, and Applications.” ResearchGate, Jan. 2019, www.researchgate.net/publication/330128155_The_Social_Networks_of Cyberbullying_on_Twitter_Concepts_Methodologies_Tools_and _Applications.

Fitzgerald, B. (2012, August 2). HOW MANY Mean Tweets Are Posted Every Day? Retrieved from https://www.huffpost.com/entry/bullying-on-twitter_n_1732952

Founta, Antigoni Maria, et al. “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior.” Proceedings of the Twelfth International AAAI Conference on Web and Social Media .

Team, P., Cyber aggression vs. cyberbullying and how to keep your child safe. Retrieved March 01, 2021, from https://blog.providence.org/archive/cyber-aggression-vs-cyberbullying-and-how-to-keep-your-child-safe

Waseem, Zeerak, and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” Proceedings of the NAACL Student Research Workshop, 2016, doi:10.18653/v1/n16-2013.

Appendix

A.

Waseem, Zeerak, and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” Proceedings of the NAACL Student Research Workshop, 2016, doi:10.18653/v1/n16-2013.

Online aggression represents a serious, and regularly occurring, social problem. In this piece the authors consider derogatory, harmful messages on the social media platform, Twitter, that target one of three groups of women, Asians, Blacks, and Latinx. The research focuses on messages that include one of the most common female slurs, “b!tch.” The findings of this chapter reveal that aggressive messages oriented toward women of color can be vicious and easily accessible (located in fewer than 30 seconds). Using an intersectional approach, the authors note the distinctive experiences of online harassment for women of color. The findings highlight the manner in which detrimental stereotypes are reinforced, including that of the “eroticized and obedient Asian woman,” the “angry Black woman,” and the “poor Latinx woman.” In some exceptions, women use the term “b!tch” in a positive and empowering manner, likely in an attempt to “reclaim” one of the common words used to attack females. Applying a social network perspective, we illustrate the tendency of typically hostile tweets to develop into interactive network conversations, where the original message spreads beyond the victim, and in the case of public individuals, quite widely. This research contributes to a deeper understanding of the processes that lead to online harassment, including the fortification of typical norms and social dominance. Finally, the authors find that messages that use the word “b!tch” to insult

Asian, Black, and Latinx women are particularly damaging in that they reinforce traditional stereotypes of women and ethno-racial minorities, and these messages possess the ability to extend to wider audiences.

B.

Founta, Antigoni Maria, et al. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." Proceedings of the Twelfth International AAAI Conference on Web and Social Media.

In recent years online social networks have suffered an increase in sexism, racism, and other types of aggressive and cyberbullying behavior, often manifesting itself through offensive, abusive, or hateful language. Past scientific work focused on studying these forms of abusive activity in popular online social networks, such as Facebook and Twitter. Building on such work, we present an eight-month study of the various forms of abusive behavior on Twitter, in a holistic fashion. Departing from past work, we examine a wide variety of labeling schemes, which cover different forms of abusive behavior. We propose an incremental and iterative methodology that leverages the power of crowdsourcing to annotate a large collection of tweets with a set of abuse-related labels. By applying our methodology and performing statistical analysis for label merging or elimination, we identify a reduced but robust set of labels to characterize abuse-related tweets. Finally, we offer a characterization of our annotated dataset of 80 thousand tweets, which we make publicly available for further scientific exploration

C..

Waseem, Zeerak, and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” Proceedings of the NAACL Student Research Workshop, 2016, doi:10.18653/v1/n16-2013.

Hate speech in the form of racist and sexist remarks are a common occurrence on social media. For that reason, many social media services address the problem of identifying hate speech, but the definition of hate speech varies markedly and is largely a manual effort (BBC, 2015; Lomas, 2015). We provide a list of criteria founded in critical race theory and use them to annotate a publicly available corpus of more than 16k tweets. We analyze the impact of various extra-linguistic features in conjunction with character n-grams for hatespeech detection. We also present a dictionary based the most indicative words in our data

Joshua Ciccarelli

jpc45@psu.edu | https://www.linkedin.com/in/joshua-ciccarelli-78454317a/
Website: JoshuaCiccarelli.com

Education

The Pennsylvania State University Expected Graduation Date: May 2021
Schreyer Honors Scholar
Bachelor's in Applied Data Science 4 Semesters Dean's List

Project Experience

On-Going Twitter API Research (Research Assistant) Spring 2020 - Present

- Creating a Random Forest Classifier using Twitter API with Python code that can detect possible cyber-bullying and cyber-aggression occurrences with a target accuracy of 85% or higher
- Participating in weekly meetings with a 4-person research team to report progress and integrate code

Stock Market analysis via Twitter Pipeline Spring 2018

- Gathered 1,500 - 2,000 tweets using Twitter API and Python libraries including Tweepy and Pandas to generate a dataset of tweets relevant to the target company
- Used logistic regression to classify new tweets and create a ratio of positive-to-negative sentiment with data visualization displaying whether the stock was trending up or down

Work Experience

Accenture (Technology Summer Analyst) Summer 2020

- Worked as a Technology Analyst on the PTP 2021 SAP Upgrade for BMS
- Networked with over 25 Accenture employees including 16 MD's across 6 of Accenture's divisions
- Used Excel, PowerPoint, SAP Business Client, Visio & more to create numerous SAP deliverables, spreadsheets, comparison analysis, and reports to assist the PMO leads on the project

Ciccarelli Law Office (Project Specialist): Summer 2014 - Summer 2017

- Assisted with Ciccarelli.com's award winning Search Engine Optimization by analyzing Key-Word search engine patterns and creating new pages as the firm branched out to new regions and types of law
- Handled client facing activities including scheduling appointments, answering phones and drafting client documents

Volunteer Work

Emergency Medical Services Fall 2016 - Summer 2019

- Emergency Medical Technician: National and Pennsylvania Certification

Extracurricular Activities

TAMID Group (Director of Technological Fund) Spring 2020 - Present

- Leading a 5-person team to create a data science stock market algorithm that will be used to compete against the investment fund portion of the club.

Nittany Data Labs (Presenter) Spring 2019 - Present

- Presenting multiple 40 minute in-person presentations to a club of 50 people on concepts like multi-variable linear regression for the purpose of explaining data science to business students and increasing public speaking skills

Relevant Coursework: Intro to Data Science, Data Management-Data Sci, Intro Programming Tech, Prog & Comp II, Org Data, Lang Log Disc Math, Elem Probability, Stat Data Science, Mach Learn & Data Analytics, Introductory Finance, Financial and Managerial Accounting

Software Experience: Python, SQL, R, Microsoft Excel, Microsoft Word