

Prairie View A&M University

Digital Commons @PVAMU

---

All Theses

---

5-2020

## Sentiment Analysis on Social Media Via Machine Learning

Christina Hastings

*Prairie View A&M University*

Follow this and additional works at: <https://digitalcommons.pvamu.edu/pvamu-theses>

---

### Recommended Citation

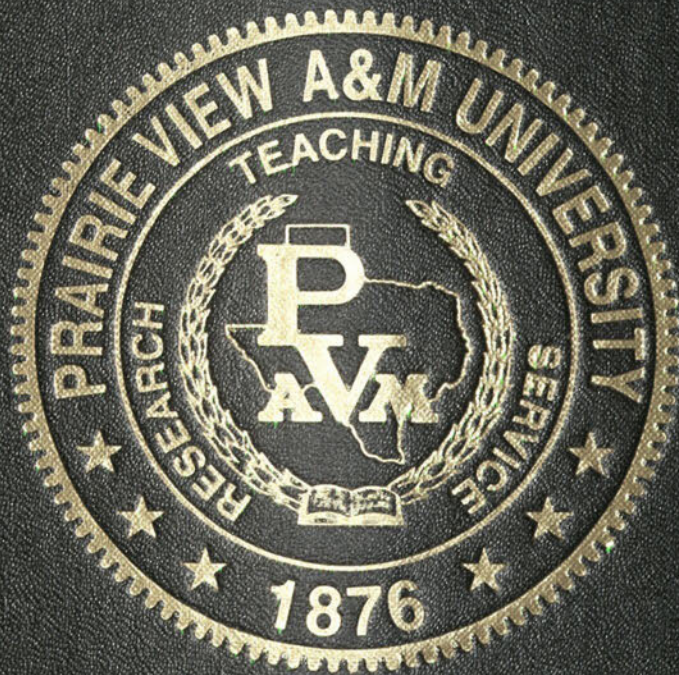
Hastings, C. (2020). Sentiment Analysis on Social Media Via Machine Learning. Retrieved from <https://digitalcommons.pvamu.edu/pvamu-theses/3>

This Undergraduate Thesis is brought to you for free and open access by Digital Commons @PVAMU. It has been accepted for inclusion in All Theses by an authorized administrator of Digital Commons @PVAMU. For more information, please contact [hvkoshy@pvamu.edu](mailto:hvkoshy@pvamu.edu).



# SENTIMENT ANALYSIS ON SOCIAL MEDIA VIA MACHINE LEARNING

CHRISTINA HASTINGS



MASTER OF SCIENCE  
ELECTRICAL ENGINEERING

ROY G. PERRY COLLEGE OF ENGINEERING  
PRAIRIE VIEW A&M UNIVERSITY

2020



# SENTIMENT ANALYSIS ON SOCIAL MEDIA VIA MACHINE LEARNING

A Thesis by

CHRISTINA HASTINGS

Submitted to the Office of Graduate Studies of  
Prairie View A&M University  
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2020

Major Subject: Electrical Engineering

# SENTIMENT ANALYSIS ON SOCIAL MEDIA VIA MACHINE LEARNING


A Thesis by


CHRISTINA HASTINGS

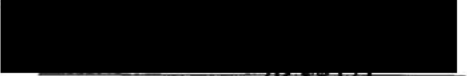
Submitted to the Office of Graduate Studies of  
Prairie View A&M University  
in partial fulfillment of the requirements for the degree of


MASTER OF SCIENCE

Approved as to style and content by:


  
Xishuang Dong  
Chair of Committee


  
John Fuller  
Committee Member


  
Kelvin Kirby  
Interim ECE Department Head

  
Pamela Obiomon  
Dean, College of Engineering

  
Lijun Qian  
Committee Member

  
Xiangfang Li  
Committee Member

  
Irvin W. Osborne-Lee  
Associate Dean, College of Engineering

  
Dorie Gilbert  
Dean, Graduate Studies

May 2020

Major Subject: Electrical Engineering



## **ABSTRACT**

Social media are shaping users' attitudes and behaviors through spreading information anytime and anywhere. Monitoring user opinions on social media is an effective solution to measure users' preferences towards brands or events. Currently, supervised machine learning-based methods dominate this area. However, as far as we know, there is no comprehensive comparison of performances of different models to figure out which model will be better for individual datasets. The focus of this thesis is to compare the performance of different supervised machine learning models. In detail, we built six classifiers, including support vector machine, random forest, neural network, Adaboost, decision tree, and Naive Bayes on two datasets and compare their performance. Furthermore, we introduced feature selection to remove unrelated attributes to preprocess the data and compare performance by building classifiers on the preprocessed data. Experimental results show that without feature selection, there is no significant difference in the performance. After feature selection, random forest outperformed other classifiers.

## ACKNOWLEDGMENTS

First, I thank my committee chair, Dr. Xishuang Dong, for his guidance and effort throughout my research process. I also send thanks to the other members of my committee, Dr. Lijun Qian, Dr. Xiangfang Li, and Dr. John Fuller, for their assistance.

This research work is supported in part by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

Finally, I sincerely thank my parents, Nicole and Reginald Blow, for their unwavering love and support throughout my academic career. Special thanks goes to my sister Ciarra for her loving support and for teaching me how to write again.

## TABLE OF CONTENTS

	Page
1 INTRODUCTION .....	1
1.1 Background and Motivation .....	3
1.2 Proposed Approach .....	4
1.3 Problem Formulation .....	5
1.4 Significance of this Research and Contributions .....	5
1.5 Structure of the Thesis .....	6
2 LITERATURE REVIEW .....	7
2.1 Sentiment Analysis .....	8
2.2 Twitter .....	13
2.3 Machine Learning .....	14
2.4 Discussion .....	14
2.5 Final Thoughts .....	14
3 APPLICATION TO SENTIMENT ANALYSIS .....	15
3.1 Algorithms .....	15
3.1.1 Random Forest Classifier .....	15
3.1.2 Support Vector Classification .....	15
3.1.3 Neural Network .....	17
3.1.4 Gaussian Naïve Bayesian .....	18
3.1.5 Decision Tree Classifier .....	18
3.1.6 Adaboost Classifier .....	19
3.2 Feature Selection .....	20
3.2.1 Select K Best .....	20
3.2.2 Extra Tree Classifier .....	20
3.2.3 F-Test .....	21
3.3 Data Sets .....	21
4 METHOD AND RESULTS .....	23
4.1 Research Strategy .....	23



	Page
4.2 Research Method .....	24
4.1.1 Pre-processing .....	24
4.1.2 Data splitting and vectorization .....	24
4.1.3 Evaluation Metrics .....	26
4.1.4 Feature selection implementation and repeat .....	27
4.2 Approach.....	27
4.3 Sample Selection .....	28
4.4 Research Process .....	28
4.5 Ethical Considerations.....	28
4.6 Research Limitations .....	28
4.7 Experimental Results .....	29
4.7.1 Experiment 1 .....	29
4.7.2 Experiment 2 .....	32
5 CONCLUSION AND FUTURE WORK .....	37
5.1 Conclusion .....	37
5.2 Future Work.....	38
REFERENCES .....	39
VITA .....	45

## LIST OF FIGURES

FIGURE	Page
3.1 Support vector classifier [25] .....	16
3.2 Neural network with three hidden layers [27]... ..	17
3.3 Example of a decision tree [34] .....	18
4.1 Coordinates for the array .....	25

## LIST OF TABLES

TABLE	Page
3.1 All Categories and the number of tweets associated with each one.. .....	22
4.1 Performance comparison on dataset 1 .....	29
4.2 Performance comparison on dataset 1 with K best feature selection .....	30
4.3 Performance comparison on dataset 1 with extra trees feature selection .....	31
4.4 Performance comparison on dataset 1 with F-Test feature selection .....	32
4.5 Performance comparison on dataset 2.....	33
4.6 Performance comparison on dataset 2 with K Best feature selection .....	34
4.7 Performance comparison on dataset 2 with extra trees feature selection .....	35
4.8 Performance comparison on dataset 2 with F-Test feature selection .....	35



## CHAPTER 1

### INTRODUCTION

In 2020, social media platforms continue to reign superior as a form of communication of information within everyday society. Since the turn of the century, the power of technology has become a global phenomenon. The speed by which data can travel has improved significantly. This power is now at the fingertips of the average citizen. Through the creation of programs such as chat rooms, blogs, and instant messaging, the art of face-to-face communication is experiencing a sharp decline. Social media has made the traditional newspaper obsolete because the smallest tidbit of information can become global news in a matter of seconds through a tweet.

The invention of the modern-day e-mail dates back to 1972, while American Online, AOL, messenger was only created in 1997. In the 1970s through the early 1990s, email was primarily used for work only topics and discussions [1]. Only in the late 1990s through the early 2000s was email and AOL messenger used as a form of communication for the average citizen as a means to communicate with family and friends outside of the working environment. This method of communication quickly transformed into MySpace for the newest generation. In 2009, MySpace was quickly taken over by Facebook and currently operates on a level playing field with Snap Chat, Twitter, and Instagram. These platforms aid in the production of mass quantities of information known as big data. The three major characteristics of big data are volume, variety, and velocity [2].

All the social media platforms listed are the primary means of communication between individuals who wish to communicate with the masses. The wide use of social media has increased the need for threat detection. This form of communication comes with a list of advantages and disadvantages. Some of these advantages are communication with friends and family who reside across the nation and in foreign countries. Social media has become the fastest and most efficient method of keeping people apprised of their everyday life.

With advantages also come the disadvantages of social media. One of these disadvantages is Internet trolls, defined as individuals who take pleasure from causing other people grief by starting fights or making outlandish accusations through Internet posts [3]. These trolls are also known for cyberbullying because they do not understand that there are actual people on the other end of the computer screen [3]. On a more serious note, cyberterrorism is an ever-evolving concept that is becoming more prominent with the development of social media platforms. These platforms do not closely monitor the nature of posts.

Researchers have been developing various approaches to examine the dark side of Twitter and the vast diversity of the users. Machine learning is the primary approach to analyzing these platforms. Machine learning will be discussed further in the literature review. Ultimately, the purpose of this research is to showcase the deceitful nature of various parts of the social media community. Not all tweets or other posts made on social media are created based on truth or verifiable facts. One of the data sets used in this paper is from the 2016 Presidential Election. Russian hackers attempted to influence voter's opinions by creating false tweets with the intent to incite rage and draw attention to the

unjust treatment of young Black Americans by local police departments. The Mueller Report was released to the public in April 2019, and it detailed all of the information collected in regards to the security breach [4]. The report begins with the Democratic National Committee and its cyber response team making a public statement about their systems being breached and expanded to include other evidence collected [4]. That is just one example among the countless issues within the United States that generated public outcry.

The evolution of social media has taken place over the course of many centuries. Each century brought about a multitude of opportunities to transform the way information is spread. For example, sending handwritten letters through the mail and the distribution of newspapers to every household was considered the earliest forms of social media. Over time these means of sharing information have become obsolete. With the new era, these same concepts have simply been improved. This was done to fit the current generations incessant need for information at the tips of their fingers. Social media provides them with that instant gratification and can continue to apply to life through future evolution.

### **1.1 Background and Motivation**

The researcher's motivation for bringing this potentially disastrous issue to the forefront is because "knowledge is power." Without being properly appraised of the situation, people do not have the ability to make an informed decision. This leads to the question: how will anybody know of the severity of the situation without the appropriate data? For many people, the Russian hacking scandal was a myth created by naysayer Democrats who were displeased with the outcomes of the 2016



Presidential Election, where Donald J. Trump was elected President. Since then, the 45<sup>th</sup> Presidential term has been plagued with scandal since the very day the President took office [5]. The alleged Russian hacking tainted the 45<sup>th</sup> President's term due to the division of the nation. Researchers are actively working to attain the ability to determine a real threat versus a fake threat on Twitter and various other social media platforms [5]. The ultimate goal of the research is to help protect people from imminent threats. The research gathered would merely be a stepping stone for this researcher to reach the said goal.

## **1.2 Proposed Approach**

Social media is constantly presented as an issue within discussions because of the relatively new threats that attest to the dangers of the Internet. The main problem is that various social media platforms are being used to influence impressionable young people who will actively adjust their behaviors and patterns based on what they see being posted. Within the past 7-10 years, a disturbing increase has occurred in the number of users who post content that would be deemed ominous. The post will eventually be taken down from the website, but nobody bothers to examine what the post truly means.

The researcher intends to explore these types of tweets and hopefully provide the readers with insight into the dangers of ignored threats. The literature behind this study spans several areas of interest, but few focus on threat detection. The full threat of social media has not been fully uncovered, but as time progresses, so does the level of awareness. The ability to detect threats and events through the analysis of a user's

unsettling messages posted to social media is still at an infancy level of exploration. Still, the full idea is slowly beginning to take form.

### **1.3 Problem Formulation**

During the 2016 Presidential Election, a group known as the *Internet Research Agency (IRA)* attempted to redirect public opinion by creating fake Twitter accounts. These accounts would then go on to discretely impact public opinion [5]. The Russian hackers' primary goal was to persuade the American people through false tweets that portrayed a more pessimistic view of the issues going on around the United States of America. The majority of the said issues had been completely fabricated to create a public uproar. At this moment in time, fabrication is considered one of the most dangerous misuses of the social media community.

### **1.4 Significance of This Research and Contribution**

The importance of this study was derived from the increase in violent actions being created on social media platforms globally. In today's energetic society, people find it difficult to go a month without hearing, seeing, or experiencing some form of brutal tragedy that could have potentially been avoided. The sheer amount of schools, bars, churches, and shopping centers that have been the ideal locations for mass shootings within the last year alone has reached staggering levels. More parents are considering home-schooling rather than sending their children to school. The chances of parents never seeing their child again after going to school has significantly increased in recent years.

This researcher believes that most of these attacks could have potentially been prevented had there been an algorithm designed to analyze potentially dangerous

word patterns. With the need for global safety against terroristic actions, the world needs to find a way to be more proactive instead of reactive. The algorithms currently in use by Twitter to obtain threatening messages should render Twitter accounts unusable while the threat is being assessed. This method has a variety of drawbacks because there are no efforts currently in place to prevent users from having access to the flagged accounts, even when the tweeters behavior has been deemed questionable. Removing a post from public view does not actively translate into action being taken to protect targeted individuals from potentially dangerous situations. Covering up the issue does not equate to addressing and potentially fixing the problem.

Many people in society today will likely have some form of social media. Ominous individuals will often display dangerous characteristics that can be deemed threatening or alarming over time. The time-variant is the missing element that most algorithms do not consider. Therefore, instead of spending time focusing on the outright threats against individuals, the true focus should be shifted towards implied acts of violence. These types of risks are harder to detect due to the language used and can be hard to discern, but that is why this avenue of threat detection research is necessary.

## **1.5 Structure of the Thesis**

This thesis begins with the literature review, which discusses the various elements that have been employed throughout the thesis, the techniques used by the researcher to complete this research, the analysis of the results from the experiments, and the conclusion and future work.



## **CHAPTER 2**

### **LITERATURE REVIEW**

There was a time when online threats were viewed as angry anonymous thoughts typed on a social media platform. As time has passed, those same threats have started to become a reality. The persistence of online threats has increased as social media has grown. Researchers have increased their efforts to proactively create methods to predict when social media threats are most likely to transpire.

On September 11, 2001, the Twin Towers in New York City were destroyed during a terrorist attack. This attack was deemed the worst terroristic attack on United States soil [6]. While that remains true, that was not the only attack that has taken place. In 2019, there were two shootings in less than 24 hours that claimed innocent lives. The leftist Tweets about the mass shooting that took place on August 3, 2019, in El Paso, Texas, were liked and re-tweeted by the gunman of the mass shooting that occurred in Dayton, Ohio on August 4, 2019. Both leftist and rightest tweets can indicate violent views and tendencies. Leftist refers to the ideologies of Republicans; while rightest refers to the ideologies of Democrats. With the increase in threats, the need to have a method to filter out and evaluate the messages has become imperative.

All of these singular moments can create an illustration in the readers' minds of what is to come next. Simple actions such as bags left unattended at the airport or randomly placed in large, well-populated shopping locations can result in a widespread panic that can cause unintended harm to the public. For this reason, more

people live in fear and actively avoid going to events, centers, and generally crowded places. The aim of this literature review is to present the research conducted on social media and threat detection.

## **2.1 Sentiment Analysis**

Sentiment analysis, also known as opinion mining, takes messages that have been left on social media and examines the senders intent with the message. The analysis comes from analyzing the structure of the message, the words used within the message, and determining whether a positive or negative connotation is conveyed. This function is deployed through the use of word embedding. In layman's terms, that means providing weights to the words to determine if they are important or not. These weights can also be used for feature selection, which is presented later in this chapter.

In [7] the authors broke down the different levels of sentiment analysis from document level to entity/aspect level. The document analysis level task is to determine the overall opinion. The sentence-level analysis task is to assess the opinion of all the sentences, whether a positive, negative, or neutral idea is proposed. The entity/aspect level analysis of the tweeter's likes and dislikes can be found. This level of analysis allows for finer-grained exploration [7].

In [8], the authors' sourced their data from Facebook comments and applied sentiment analysis techniques. This article explained the different fields that sentiment analysis could be used for, such as e-trade, marketing, politics, and decision making [8]. The methods shown in the article served as a vehicle to determining how data should be gathered and the different short-hands used in messages to express thoughts [8].

The authors considered the relationship between the textual information of a tweet and sentiment dissemination to display the emotion conveyed in [9]. The results were obtained after studying the inversion of feelings in the message and finding some stimulating properties [9]. Random forest machine learning was used to determine the polarity of emotions expressed in the messages. The researchers are among the first to use sentiment dissemination models in an effort to improve the sentiment analysis of Twitter messages [9].

In [10], the authors' produced an in-depth study of sentiment analysis and the various techniques that can be used to implement this approach. Machine learning and lexicon-based methods were compared in this article [10]. The machine learning approach utilized three different algorithms to produce outputs. The lexicon-based process used a sentiment dictionary full of opinion words to match the data in an effort to measure the overall polarity of the tweet. The Corpus-based lexicon approach was also used, which provides dictionaries related to specific domains [10]. These types of dictionaries are generated by a set of initial opinion terms and grown through the search of related words. This was done by way of mathematical or semantic practices [10].

The authors of [11] utilized sentiment analysis and machine learning to construct a method of segregating opinions of tweets. These messages were separated into positive, negative, and neutral categories [11]. The machine learning algorithms used were the support vector machine, Naïve Bayes, and neural networks. When the deep bidirectional, BERT, model was applied, the results produced were able to be fine-tuned to provide better results [11].

In [12], hackers on social media were studied. This group studied their communities, how they shared knowledge, coordination, and recruitment efforts [12]. This study proposed a set of indicators to be used to analyze the communication patterns, which included technical discussions, positive and negative sentiments along with threats. Twitter was the platform used for this study, and it was found that there were different indicating factors for different types of hackers [12]. For instance, hackers with higher skill levels used more technical terminology in their tweets. The hackers that were motivated by profit and ideology expressed more recruitment type language as opposed to those that were stirred by revenge or prestige [12].

The authors' in [13] examined the use of Twitter and Facebook, by amassing large quantities of data that can be sourced for opinion mining or sentiment analysis. The sites used assisted in locating the sentiments of users on a specified subject or product. In that study, a structure was proposed, which collected data from social networking sites using the Twitter and Facebook API's [13]. From here, the challenge of big data was countered using Hadoop through the map-reduce framework. The entire data set is mapped and then reduced to a smaller size data set to lessen the workload. Finally, the content is analyzed, and the final results are presented a graph for comparison purposes that can be viewed in the paper [13].

In [14], the practice of not writing sentences with the correct grammar and spelling was considered. Social networking sites are common grounds for shorthanded spellings of words [14]. These spellings lead to a wide range of uncertainties such as lexical, syntactic, and semantic errors, which results in difficulty locating the actual data order. That study aimed to define how studies of social media have applied text mining

and text analytics strategies to determine how to categorize critical themes from the information [14]. The study focused on exploring text mining studies in connection with Facebook and Twitter, which are the two most prominent social media platforms used today [14].

In [15], the micro-blogging site Twitter was employed to deploy original ontology-based techniques to produce a more efficient sentiment analysis for Twitter messages. The uniqueness of this article stems from the use of a sentiment grade instead of a sentiment score. The grade will be given for each distinct concept expressed in the message [15]. The overall product resulted in a more thorough analysis of the opinion of the post regarding specific topics [15].

Reference [16] looked into the various ways people express themselves online through messages. This study used a wide variety of messages from Twitter and machine learning applications to classify text. The authors found that none of the classification models outperformed the other. However, different models can be combined to boost the benefits of all the models [16]. The studies also suggested that this work could be applied to other languages in the future. The difficulty of classifying sarcasm and negation expression were other topics for future work to accurately and efficiently classify these sentiments [16].

The authors' of [17] demonstrated the importance of emoticons in messages. The emotion these emojis are used to display is an essential part of all messages. The proposed method of their study was to manually create an emoticon sentiment lexicon to improve the lexicon-based sentiment classification method [17]. Over Two thousand Dutch tweets and messages from forums that all contained emoticons were manually

annotated for the sentiment. With this corpus, paragraph-level analysis of the sentiment improved the accuracy significantly [17].

Gaikwad [18] explored the different trends of sentiment analysis, along with the various approaches used to examine the sentiment. These trends were machine learning, ontology-based, and other unsupervised methods of exploration [18]. For machine learning, Naive Bayesian, maximum entropy classifier, and support vector machine were used to produce outputs for the accuracy. Ontology-based sentiment analysis can be used for an in-depth investigation of twitter posts. The author of this study ultimately concluded that the best method for the most accurate results would be to combine machine learning and ontology-based methods to classify the tweets [18].

A brief glimpse of the various methods that can be used to perform sentiment analysis was provided in [19]. Different grouping strategies were presented. Some of the groupings were SentiStrength, SentiWordNet, and Happiness Index [19]. These different groupings are used to express how different methods are dependent on the information presented [19].

In [20], sentiment analysis on Twitter data implementing SentiCircle, a lexicon-based approach was performed. Senticircle takes the co-occurrence patterning of words and the different contexts in which they appear to acquire their semantics [20]. Updates are allowed to be made to the pre-assigned strengths and polarities in sentiment lexicons in this format. These authors' approach allows for entity-level and tweet-level detection of the sentiment [20]. Three different data sets and three different sentiment lexicons were used during the course of their experiment. The proposed method outperformed the baselines in both accuracy and F-measure on entity-level subjectivity and polarity [20].



The tweet-level results were better than the normal SentiStrength overall, making this a beneficial approach to use in the future [20].

## 2.2 Twitter

Twitter is a social media platform that is used to communicate information to the masses quickly in 280 characters or less. These messages, known as tweets, have the ability to inform others about something as menial as what was eaten for dinner or something as important as a shooting at a church, school, or public gathering. These messages are important to the reader for different reasons. Messages have the ability to be informative, but they can also be destructive. These messages contain harmful and even false information that is being spread and potentially believed, by the readers.

The Pizzagate conspiracy theory was a rumor that began on Twitter in 2016 about a pizzeria in Washington DC that was said to be a front for a child abuse ring. Welch, a 28-year-old male, read about the conspiracy on Twitter and spent a few days researching before he drove from his home in Salisbury, North Carolina, to "self-investigate" the claims. He conducted research beforehand and attempted to enlist his friends to join him, but they refused.

According to an article written by Eric Lipton of *The New York Times* [21], he shot off a round from a rifle, the caliber was undisclosed, while he was inside and was arrested and afterward sentenced to four years in prison. This platform was used by [12] as a way to analyze hacking behaviors on social media and to figure out ways to detect these behaviors. Twitter is the social media platform that a large portion of researchers use.

## **2.3 Machine Learning**

Machine learning is when a set of methods can be applied to a collection of data to notice patterns the patterns found to assist in predicting outcomes from the collection of data [22]. There are two types of machine learning. Those are supervised and unsupervised learning. Supervised learning utilizes an input and an output to provide the machine information to learn patterns. For unsupervised learning, the provided input portion is used to locate the patterns [22]. The most commonly applied form of machine learning is supervised learning [22]. In most of the articles used, machine learning was the preferred method for sentiment analysis.

## **2.4 Discussion**

A detailed review of the literature disclosed that using more than trigger words is essential. Different information offered by the accounts is beneficial to review to provide context. Acquiring sub-types beyond threat and non-threat could also be helpful.

One of the limitations of the literature found was searching for threats of varying types--meaning all threat types were explored from personal threats to hackers to spammers. These different types of threats are essential and have a common thread with various ways to calculate the output.

## **2.5 Final Thoughts**

This literature review has provided an in-depth background of all of the reference material that was used throughout this study. The focus of this chapter is to explain the different components that were implemented, as well as examining the various works that were conducted surrounding the models and feature selections.

## CHAPTER 3

### APPLICATIONS TO SENTIMENT ANALYSIS

The strategy utilized was conducted by completing background research on the subject, locating an available data set, implementing a code that could analyze the data, and producing outputs from the six different algorithms used. From there, feature selection was applied and reran through the six algorithms to refine the results of the data. All conclusions were drawn from the final outputs and provided ideas for future uses. Different steps were taken to complete this study that will be explored and explained further.

#### 3.1 Algorithms

The algorithms used for this experiment are traditionally used for machine learning problems. Different algorithms are used to produce different outcomes. The more complex the question is, the more complex the algorithm. Six algorithms were used for the completion of this experiment. Those algorithms were random forest classifier, support vector classifier, neural network, Gaussian naive Bayesian, decision tree classifier, and Adaboost classifier. The algorithms were chosen to help to provide a wide array of potential outcomes to determine which model to use.

**3.1.1 Random forest classifier.** The Random Forest Classifier is an algorithm that creates a multitude of subset trees. These different trees then take the data set and "learn" what characteristics fit best with a given answer. From here, different trees implement majority-voting techniques to come to a final answer [23]. Random Forest Classifier was applied in [24] to classify the star reviews of a product. The investigation took place to help the creators enhance their brand [24].

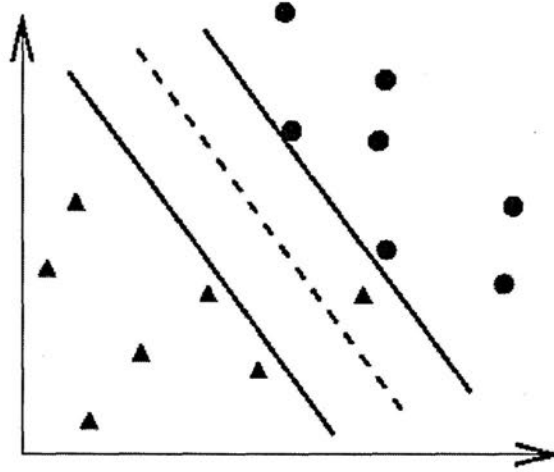
$$m_{M,n}(x, \theta_1, \dots, \theta_M) = \frac{1}{M} \sum_{j=1}^M \sum_{i=1}^n \frac{Y_i 1_{x_i \in A_n(x, \theta_j)}}{N_n(x, \theta_j)} \quad (3.1)$$

**3.1.2 Support vector classification.** The support vector classifier, SVC, is a subset of support vector machines, SVM, which is used for classification. SVC uses supervised machine learning to separate the information into different hyperplanes [23]. In laymen's terms, this means that two classes, class A and class B, are separated on either side of a graph. The line for the hyperplane can be drawn after determining the best fit concerning class A and class B in the diagram [23]. The authors [25] utilized SVM to present how different sets of features and weights on words can affect the output of SVM.

Figure 3.1 presents the support vector classifier.

**Figure 3.1**

*Support Vector Classifier [25]*



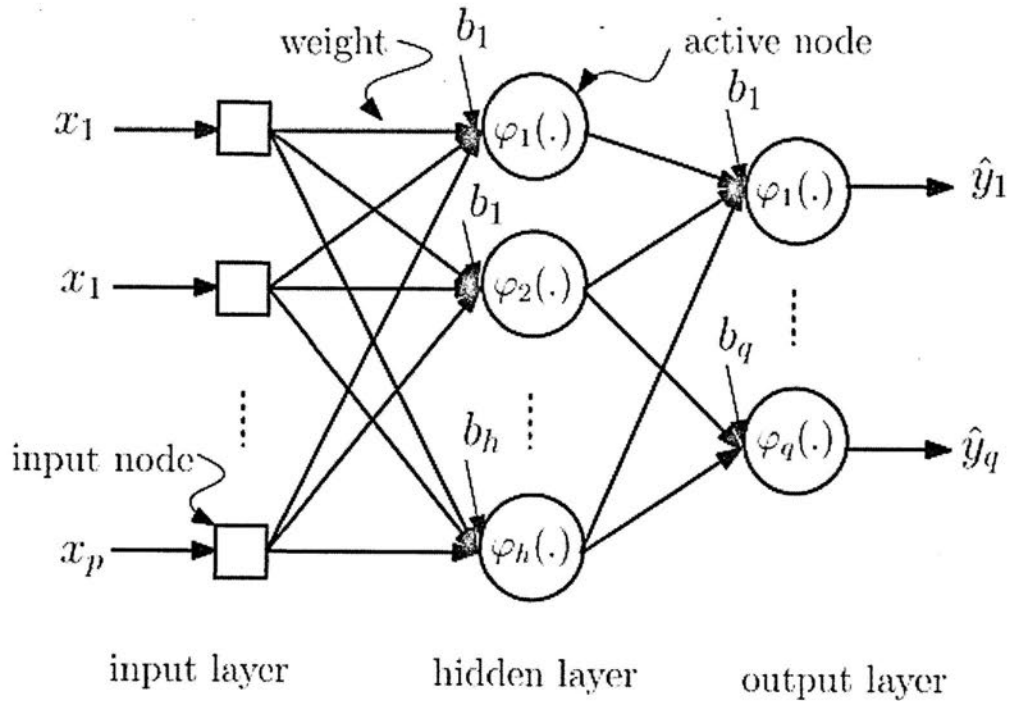
**3.1.3 Neural network.** A neural network is a computer-based "brain" that is loosely designed to mimic the way a human's brain functions. However, this has never

been achieved due to the fact that no one has ever being able to crack the complexity of the human brain [27]. Neural networks are a way of teaching a computer to perform by analyzing training data sets that have been hand-labeled. The neural network will then use the patterns that have been found and apply them to new information [27].

Figure 3.2 shows the neural network with three hidden layers.

**Figure 3.2**

*Neural Network With Three Hidden Layers [27]*



The authors [29] used trimmed data and applied this information to convolutional neural networks. These networks were based on the morphological pattern presented [29]. [30] showcased the power of neural networks to process large quantities of data promptly across different social media platforms.

**3.1.4 Gaussian Naïve Bayesian.** Gaussian Naive Bayesian is a subset of Bayes theorem, which is the fundamental belief that one event will occur because another event has occurred. Gaussian Naive Bayesian believes that continuous values are associated with each feature [31]. These features are perceived to be distributed according to Gaussian distribution. This function is also known as Normal distribution and produces a bell-shaped curve [31]. The equation necessary to solve for Gaussian Naive Bayesian is:

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \quad (3.2)$$

In [32], Gaussian Naive Bayesian was applied along with the CHI2 feature selection to optimize the outcomes from different review-based websites.

**3.1.5 Decision tree classifier.** Decision trees behave in a tree-like manner to solve the classification problem. Each "leaf" holds an attribute and must determine whether the information presented has this trait [33]. From here, the "leaf" has the option to send the information to the next "leaf" or send the information to the exit. This same procedure will continue until all of the information has been filtered [33].

Figure 3.3 shows an example of a decision tree [34].

**Figure 3.3**

*Example of a Decision Tree [34]*



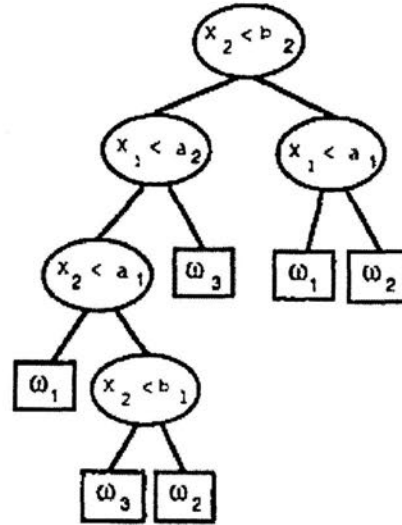


Figure 3.3 provides an example of how decision trees work. The authors [35] used the decision tree to perform sentiment analysis on Urdu news tweets using 150 positives and 150 negative words. Ferdin Joseph [36] worked with the decision tree to predict the outcome of the 2019 Indian general election. Joseph only used English tweets but deemed the work necessary to be spread to other languages [36].

**3.1.6 Adaboost Classifier.** Adaboost classifier is generally used to combine "weak learners" which are usually decision trees and random forest classifiers [23]. Adaboost begins by placing a classifier over the data set and creating multiple copies of the classifier on the same data set. However, the incorrectly weighted instances from the first classifier are changed to more complicated cases [23]. The authors [37] implemented adaboost to enhance their convolutional neural network for dealing with sentiment analysis. See equation (3.3).

$$H_{\text{final}}(x) = \text{sign}(\sum_k \alpha_k h_k(x)) \quad (3.3)$$

### 3.2 Feature Selection

Feature selection is an extra component that can be added to embedded wording as a way of extracting the words that have the highest weights as a way of refining results. Feature selection cuts down on computational costs and the time required to filter through all of the data [38]. Irrelevant feature inputs have the potential to produce overfitting. Overfitting is when a feature that defines a personal characteristic in the data is used to identify an overall trait [38]. Sifting out the features that have little to no effect on the output will contribute to keeping the model small and usable[38]. The authors [39] employed feature selection to compare feature selection evaluators on Twitter sentiment classification.

**3.2.1 Select K Best.** The Select K Best is a form of feature selection used in an attempt to improve the output results by obtaining the highest k scores [23]. The k is found by taking the best results from the data and using them to produce a better output. Any ties between features are broken in an undisclosed manner [23]. The author of this study was unable to locate literature on Select K Best.

**3.2.2 Extra tree classifier.** Extra Trees classifier is a step above Random Forest Classifier in that the same steps are taken, but the "tree" is assembled into a "forest." The same data set is used to implement the code [40]. However, at each "leaf" a random sampling of k features, features that have been deemed the best to implement, are analyzed [40]. As the "forest" progresses, multiple de-correlated decision trees will be left behind. The de-correlation of the trees assists in determining the final outcome for

each input [40]. The author of this study was unable to locate literature pertaining to Extra Tree Classifier.

**3.2.3 F-Test.** F-Test was the final feature selection used to produce an outcome. F-Test is used to compare between models and to check if there is a substantial difference between the models [41]. A hypothesis testing model is created with X being a constant and Y being the model created by the constant and a feature. The errors produced are then analyzed to determine whether these errors are considerable or were they introduced by chance [41]. F-Test is valuable for getting to know the importance of each of the features to improve the model [28]. Select K Best utilizing F-Test was applied in this research. The author of this study was unable to locate literature about F-Test.

### 3.3 Data Sets

The first data set was sourced from *FiveThirtyEight*, by an undergraduate researcher named Cody, on Russian troll accounts that were flagged during the 2016 presidential election. These accounts were flagged and collected in this particular data set because they were said to be attempting to influence voters [5]. The purpose of these tweets was to incite rage and hostility into the readers. Eight categories were made available by the website, as displayed in Table 3.1. However, only two were the primary focus of this research. The Right Troll tag represented the right-wing or Republican views, and the Left Troll tag served to signal the left-wing or Democratic opinions. Both sides were opposed to each other, and this was exploited in the hope of starting fights online that could then manifest off-line [5]. The main advantage of using a pre-published data set was that all of the information was sorted and organized in excel files. These were available on Github and pre-labeled with the unedited messages attached. These

files also contained the country of origin, the author, language, and the post type. The post type labels helped to discern whether the message was the original post or a retweet of a post. These labels helped to eliminate the redundancy of posts applied to the algorithms.

**Table 3.1**

*All Categories and the Number of Tweets Associated with Each One*

Category	Number of Tweets
Non-English	837,725
Right Troll	719,087
News Feed	599,294
Left Troll	427,811
Hashtag Gamer	241,827
Commercial	122,582
Unknown	13,905
Fearmonger	11,140

The second data set used was a simple sentiment analysis corpus. This corpus contained 1,578,627 classified tweets. All of the tweets were taken from the University of Michigan "Sentiment Analysis" competition listed on Kaggle and "Twitter Sentiment Corpus" made available by Niek Sanders [42]. The connotation of the tweet separated them. A "0" symbolized a negative connotation while a "1" symbolized positive connotation [42].

## **CHAPTER 4**

### **METHOD AND RESULTS**

This chapter will provide a brief overview to the steps taken throughout the completion of this research. The results from each experiment will also be provided and explained.

#### **4.1 Research Strategy**

This researcher reviewed the literature available on the subject matter and determined what findings were already published. After this step occurred, setting up the environments, choosing the algorithms, and determining which feature selections to implement followed. The researcher's advisor provided the algorithms and feature selections used throughout the study. The measuring parameters for each algorithm were prearranged. From here, the author implemented the program using Python programming language Version 3. Python is regarded as a high-level programming language that can be used by all major platforms and is freely distributable [43].

The language can be accessed through Anaconda Navigator, which is an environment manager that assists in accessing work environments without having to open a command prompt and hunt for the conda environments [44]. The platform chosen for this experiment was Spyder, which was released on October 13, 2014. This scientific environment allows for building, editing, and debugging of codes written in Python for Python [45]. Scikit learn is a callable function for Machine Learning that is a simple and effective mechanism that can be used for predictive analysis data. These packages are easily accessible by anyone and can be used in various contexts [23].

## 4.2 Research Method

These messages were gathered by a third-party source and used to complete this research. These third-parties were discussed previously in section 3.3. The methods deployed were used to hone the findings from these data sets. The messages in the first data set were removed from Twitter, and researchers gained permission to distribute them due to the nature of the tweets.

**4.2.1 Pre-processing.** The data was cleaned to remove the parts that the code could not read to process the information. Cleaning entailed removing all punctuation, emojis, numbers, capitalization, and all other special characters. The components that remained in the messages were the various letters and words used to compose the messages. Right Troll and Left Troll labels were used for this study because the categories chosen had the highest number of English tweets available. During the cleaning process, a "hole check" had to be completed. Holes appeared after running these functions, and special characters removed, leaving behind a blank cell. This was done to guarantee there were no empty spaces because the next line of code could not be run with these holes. After cleaning, the file was saved to a new location for use in the next step.

**4.2.2 Data splitting and vectorization.** After data cleaning and saving the information to a new file, the data had to be split. The purpose of splitting the data was to separate the file into two sections. These sections were to train the models and test the model's outputs from the training. Training the model means "teaching" the model to read the message and determine whether the tweet is a Right Troll or a Left Troll. A ratio of 70% and 30% were used for training and testing, respectively.



These two groupings were then split up into `x_train`, `x_test`, `y_train`, and `y_test` to differentiate which file trained the models and which tested the models. Separation ensured that both of the test and train files remained isolated from each other, guaranteeing that contamination of the results had not occurred. Once the files were split Term Frequency-Inverse Document Frequency, TF-IDF, was then applied to the `x_train` and `x_test` files to transform the words into files that were read by the machine learning models to output the results. TF-IDF is a numerical statistic that is used to imitate the importance of a word in reference to a document that is then collected in a corpus [46].

Figure 4.1 is an example of the coordinates. These transformations produced NumPy Arrays, which are n-dimensional arrays that described the collection of weighted variables that represented every word provided in the document [47]. The shape of the arrays were (6958, 11994) for x-train, and (2982, 11994) for x-test in data set 1. For the `x_train` file, the `v.fit_transform` function shaped the data while `v.transform` on the `x_test` file allowed the file to be transformed into an array without manipulating the data to fit the `x_train` file. This allowed for factual findings at the end. The arrays were then utilized by the models for testing and training these files to provide the final results in tandem with the y files.

### **Figure 4.1**

*Coordinates for the Array*

0	(0, 11627) 0.5332848331...
1	(0, 19594) 0.5144134404...
2	(0, 10540) 0.6198550780...
3	(0, 15495) 0.2901795458...
4	(0, 7964) 0.2664737579...
5	(0, 8519) 0.6903166331...
6	(0, 14134) 0.7071067811...
7	(0, 18092) 0.5370028065...
8	(0, 10778) 0.2419753313...
9	(0, 12328) 0.5040562329...
10	(0, 20227) 0.4053634893...
11	(0, 13622) 0.5324795886...

**4.2.3 Evaluation Metrics.** The files were then applied to six established algorithms through the scikit learn packages. Each model was trained and tested with the performance measures being precision, recall, accuracy, and f1 score.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4.1)$$

The precision equation asks out of the total predicted positive results by the model, what is the actual percentage found [48].

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (4.2)$$

The recall equation focuses on the total number of predicted positives concerning the total number of actual positive results and determines what the number of positives, 1, found to be true versus the total number of positives predicted. Precision and recall focus on calculating the positive, 1, outcomes [48].

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.3)$$

The f1 score equation produces an average for precision and recall. The closer the answer is to 1 from 0, the better the model operates [48].

**4.2.3 Feature selection implementation and repeat.** For the three feature selections implemented, 100 features were chosen from the training file and applied to the test files. The features selected were then applied to each model, and the same four measuring variables were applied

These same steps were applied to data set 2. The shape of the arrays were (18200, 20456) for the `x_train` and (7800, 20456) for the `x_test`.

### 4.3 Approach

Once the data set was obtained, all conclusions were withheld until after completing the code. Final thoughts on the data could not be drawn in the beginning due to the range of the messages and the different information provided in each.

#### **4.4 Sample Selection**

After data set 1 was cleaned, over 1,000,000 usable tweets remained. From these tweets, 10,000 were randomly selected and were used during the first half of this study. A ratio of 70% and 30% were used for training and testing, respectively, for the first data set.

A second set of data was used to provide more evidence to the outcomes from the first data set. From this set, 26,000 tweets were applied from the original file. A ratio of 70% and 30% were used for training and testing, respectively, for the second data set.

#### **4.5 Research Process**

During the process of research, several obstacles had to be overcome. These obstacles included:

- The addition of a second data set
- Coding Setbacks

#### **4.6 Ethical Considerations**

All of the data used was sourced from Twitter accounts that were open to the public on this platform. Ethical considerations for this data set are unknown due to the connotation of the tweets.

#### **4.7 Research Limitations**

Both data sets were obtained from third-party sources. The author was not proficient in Python 3 before writing about this study.

## 4.8 Experimental Results

The results found for both datasets and the addition of each feature selection are presented.

**4.8.1 Experiment 1.** For the data set located in Table 4.1 without any feature selection applications, the models were found to produce pleasing results. The precision of the Gaussian Naive Bayesian was the highest at 80%. At the same time, the lowest value produced was for the random forest classifier at 63%. The complete opposite was the case for the recall with the random forest classifier, having an outcome of 100%. In contrast, the Gaussian Naive Bayesian had the lowest overall score of 54%. The Gaussian Naive Bayesian performed the worst overall with a final f1 score of 64% and an accuracy of 62%. The neural network performed the most consistent across all scores taken.

**Table 4.1**

*Performance Comparison on Dataset 1*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.633	0.633	1.0	0.775
SVC	0.748	0.778	0.843	0.809
Neural Net	0.728	0.785	0.785	0.785
GaussianNB	0.622	0.801	0.536	0.642
Decision Tree	0.679	0.740	0.760	0.750
Adaboost	0.686	0.686	0.928	0.789

Table 4.2 presents with the Select K Best feature selection added into the program. This feature selection produced consistent results with all six algorithms having a precision score of 63%. The recall scores ranged from 96% to 99%, with the f1 scores ranging from 76% to 77% for all of the models.

**Table 4.2**

*Performance Comparison on Dataset 1 with K Select Best Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.631	0.632	0.997	0.774
SVC	0.626	0.631	0.983	0.769
Neural Net	0.623	0.633	0.965	0.764
GaussianNB	0.623	0.633	0.966	0.764
Decision Tree	0.623	0.633	0.963	0.764
Adaboost	0.622	0.631	0.969	0.765

Table 4.3 shows the Extra Tree feature selection portion added to the program. As with the previous feature selection, the results were consistent across all six models. The precision score was found to be 63% for the models after rounding the third decimal. The recall portion reached 98% and above. The F1 score was found to be 77% for all of the models as well. The accuracy rounded to 63% for all, making this feature the most stable of the four Tables produced for data set 1.



**Table 4.3***Performance Comparison on Dataset 1 with Extra Trees Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.632	0.633	0.998	0.775
SVC	0.632	0.634	0.991	0.773
Neural Net	0.629	0.633	0.986	0.771
GaussianNB	0.630	0.633	0.987	0.772
Decision Tree	0.627	0.632	0.984	0.769
Adaboost	0.628	0.633	0.984	0.770

Table 4.4 shows the F-Test feature selection applied to the program. Once again, the feature selection produced a consistent output for each model used. All of the outcomes for precision were 63%. The recall score ranged from 96% to 100%. The f1 score produced results of 76% and above. The accuracy was found to be 62% for the neural network, Gaussian Naive Bayesian, decision tree, and adaboost. The random forest and SVC had 63% accuracy.

**Table 4.4***Performance Comparison on Dataset 1 with F-Test Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.633	0.633	1.0	0.775
SVC	0.633	0.634	0.994	0.774
Neural Net	0.621	0.631	0.965	0.763
GaussianNB	0.621	0.631	0.966	0.763
Decision Tree	0.621	0.631	0.966	0.763
Adaboost	0.623	0.631	0.972	0.766

**4.8.2 Experiment 2.** Table 4.5 presents the following data set before feature selection. The neural network performed the worst with 0 outputs for precision, recall, and F1 scores. However, the accuracy of 50% was found for the neural network-- meaning the model was able to determine half of the messages were correct. This does not explain why no output showed for the other three measures. The random forest classifier produced the highest precision of 79.9% with an adaboost classifier having a better recall score of 84.9%. The randomization of the messages procured for this data set seems to have made determining the general emotion challenging to pinpoint. The range for accuracy was found to be between 50% and 72%. The range of outcomes for Table 4.5 was too broad.

**Table 4.5*****Performance Comparison on Dataset 2***

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.618	0.799	0.309	0.446
SVC	0.727	0.722	0.731	0.727
Neural Net	0.503	0.0	0.0	0.0
GaussianNB	0.554	0.627	0.255	0.363
Decision Tree	0.668	0.663	0.671	0.667
Adaboost	0.669	0.622	0.849	0.718

Table 4.6 shows the Select K Best included in the code. This feature selection was able to apply consistency to the numbers that Table 4.5 lacked. The precision for the neural network was better in Table 4.5 with an outcome of 40.7%, but the recall was less than 1%. The problem of having a recall of less than 1% also affected the random forest and decision tree classifier. The other three classifiers had recall outcomes of 98% and higher. The recall seems to have produced quite a few false negatives resulting in this score. Due to the recall and precision being on the lower side, the f1 scores for the decision tree, neural network, and random forest were all less than or equal to 0.01 or 1%. In Chapter 3.2.3, the researcher stated that the model is performing properly, the closer its value is to 1. Bearing this in mind, the three models, listed previously, performed horribly with this feature selection added.

**Table 4.6***Performance Comparison on Dataset 2 with K Best Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.503	0.333	0.0005	0.001
SVC	0.498	0.497	0.986	0.661
Neural Net	0.502	0.407	0.006	0.012
GaussianNB	0.498	0.497	0.99	0.662
Decision Tree	0.504	0.571	0.007	0.014
Adaboost	0.497	0.497	0.992	0.662

Table 4.7 had the Extra Tree feature selection applied. The results were less than pleasing. The f1 score for four of the models was found to be less than 2%. As explained in the paragraph above, this is not a good result. The problem once again falls to the recall scores, which are equal to or less than 1%. The outcomes for Gaussian Naive Bayesian and support vector classifier were 99%.

**Table 4.7***Performance Comparison on Dataset 2 with Extra Trees Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.505	0.75	0.005	0.009
SVC	0.497	0.497	0.989	0.662
Neural Net	0.504	0.596	0.007	0.014
GaussianNB	0.497	0.497	0.997	0.663
Decision Tree	0.505	0.673	0.01	0.019
Adaboost	0.505	0.667	0.006	0.012

Table 4.8 shows the F-Test feature selection applied to the program, and this one performed the best. Minus the random forest classifier that performed poorly with an f1 score of less than 1%, the other models had f1 scores of 66%. The other five models had the same precision score of 49.6% and recall scores above 98%.

**Table 4.8***Performance Comparison on Dataset 2 with F-Test Feature Selection*

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.503	0.364	0.001	0.002
SVC	0.495	0.496	0.987	0.660
Neural Net	0.495	0.496	0.987	0.660
GaussianNB	0.496	0.496	0.99	0.661
Decision Tree	0.495	0.496	0.987	0.66
Adaboost	0.496	0.496	0.989	0.661

The addition of the feature selection stabilized the outcomes produced and resulted in better overall performance from the models. For data set 1, the Extra Tree classifier performed the best overall of the four outputs. While certain values for each model was higher in other Tables, the outcomes for Table 4.3 were the best when taking all factors into consideration. Outputs of 63% for accuracy and precision as well as above 98% for recall were the most consistent. The f1 score was 77% for all of Table 4.3, which is why the researcher has deemed this Table to have the best results for data set 1.

The f1 scores produced by data set 2 were troubling due to the extreme falls that took place for recall scores. The precision was within range, but the recall kept falling short.

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

This chapter provides the final findings from this study and explains research areas the researcher wishes to explore in the future.

#### 5.1 Conclusion

The goal of this research was to establish whether or not a tweet could accurately convey the mindset of an individual who is making borderline violent threats on social media. Based on the results collected from the classifiers that were analyzed, the researcher can state that the potential to catch these messages is correct. This inference means that there is a method to filter through tweets that hint at darker intentions that aim to harm the general public. Refining the terms being analyzed and the placement pattern of these words provided the researcher with the ability to predict the intended act.

This topic was researched thoroughly and executed through the procedures, showed that the results produced various new insights upon completion. Although the data set was cleaned and run, the question of whether the data set was cleaned enough proved to be an issue due to several words that appeared to be stuck together. The over-usage of hashtags caused a multitude of problems, but eventually, these problems were rectified. This made the researcher to question what made data clean enough to input into an algorithm that was comprehensible to the program. Some of the limitations faced in the course of conducting the research were the use of established data sets that were based around fake news attempting to influence young minds and a randomly composed corpus. These messages can be interpreted as a form of threat detection, referencing data set 1, that the researcher ultimately aspires to present.

The findings from this research can be applied to future efforts as a means to determine whether a threat is plausible. Words spewed from anger and spite have the potential to be dangerous and possibly lethal. The contribution is that this research can seek out a threat being imposed through social media actively. The literature review provided insight into the limited research being conducted on social media networks to seek out domestic threats. Most of the literature addressed isolated studies that have been performed all around the world. The current question in the discussion is whether using social media to determine a person's state of mind is applicable. If so, can a researcher determine an individual's ill-intentions towards another individual or even a group of individuals? The problem with this mentality is that some people will publish posts that allude to them having a desire to harm others. These individuals appear to be prone to acts of violence and strive to cause more chaos.

## **5.2 Future Work**

This researcher or others can expand upon this work by obtaining different messages and expanding the number of features selected to filter out threats.



## REFERENCES

- [1] Lichtenstein, S. "Knowledge development and creation in email." *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the. IEEE*, 2004.
- [2] Chen, M., Y. Liu and S. Mao, "Big data: A survey," *Mobile Networks and Applications*, pp. 171–209, Jan 2014.
- [3] Mihaylov, T., G. Georgiev and P. Nakov. "Finding opinion manipulation trolls in news community forums." *Proceedings of the nineteenth conference on computational natural language learning*. 2015.
- [4] Mueller, R. S. *Report on the investigation into Russian interference in the 2016 presidential election*. Vol. 1. Washington, DC: US Department of Justice, 2019.
- [5] Roeder, O. *Why We're Sharing 3 Million Russian Troll Tweets*. FiveThirtyEight. <https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/>. Jul 2018.
- [6] Watson, R. P. "The politics and history of terror." *America's War on Terror*. Routledge, 2016. 15-28.
- [7] Mittal, A. and S. Patidar. "Sentiment Analysis on Twitter Data: A Survey." *Proceedings of the 2019 7th International Conference on Computer and Communications Management*. 2019.
- [8] Kaur, R., H. Singh and G. Gupta. "A Review on Sentimental Analysis on Facebook Comments by using Data Mining Technique." (2019).
- [9] Pagar, N. and B. S. Satpue. "Survey Paper on Hybrid Approach for Twitter Sentiment Analysis using Supervised Machine Learning Algorithms."

- [10] Kharde, V. and P. Sonawane. "Sentiment analysis of twitter data: a survey of techniques." *arXiv preprint arXiv:1601.06971* (2016).
- [11] Kale, K. and P. M. Chawan. "Sentiment Analysis to Segregate Attributes using Machine Learning Techniques: A Survey." (2019).
- [12] Babko-Malaya, O., R. Cathey, D. Maimon, S. Hinton and T. Gladkova. "Detection of hacking behaviors and communication patterns on social media." *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017.
- [13] Gupta, S., A. Pandey and K. K. Paliwal. "Sentiment Analysis of Twitter and Facebook Data Using Map-Reduce."
- [14] Salloum, S. A., M. Al-Emran, A. A. Monem and K. Shaalan. "A survey of text mining in social media: facebook and twitter perspectives." *Adv. Sci. Technol. Eng. Syst. J* 2.1 (2017): 127-133.
- [15] Kontopoulos, E., C. Berberidis, T. Dergiades and N. Bassiliades. "Ontology-based sentiment analysis of twitter posts." *Expert systems with applications* 40.10 (2013): 4065-4074.
- [16] Patil, A., V. Magar, K. Kulkarni and A. Manwar. "A Survey on Classification of Sentiments from Twitter." *International Journal of Engineering Research & Technology (IJERT)*. 2015.
- [17] Hogenboom, A., D. Bal, F. Frasincar, M. Bal, F. Jong and U. Kaymak. "Exploiting emoticons in sentiment analysis." *Proceedings of the 28th annual ACM symposium on applied computing*. 2013.
- [18] Gaikwad, A. S. "Twitter Sentiment Analysis Approaches: A Survey." *International Journal of Engineering Research & Technology (IJERT)*, vol. 06, no. 01. 2019.

- [19] Bhardwaj, S. and J. Pant. "A Survey of Approaches for Sentiment Analysis on Social Media." *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2019.
- [20] Saif, H., Y. He, M. Fernandez and H. Alani. "Contextual semantics for sentiment analysis of Twitter." *Information Processing & Management* 52.1 (2016): 5-19.
- [21] Lipton, E.. "Man Motivated by 'Pizzagate' Conspiracy Theory Arrested in Washington Gunfire," *New York Times*.  
<https://www.nytimes.com/2016/12/05/us/pizzagate-comet-pingpong-edgar-maddison-welch.html>. 2020.
- [22] Murphy, K.. "Machine Learning: A Probabilistic Perspective." *The MIT Press*. 2012.
- [23] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2825-2830, 2011.
- [24] Karthika, P., R. Murugeswari and R. Manoranjithem. "Sentiment Analysis of Social Media Network Using Random Forest Algorithm." *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. IEEE, 2019.
- [25] Naz, S., A. Sharan and N. Malik. "sentiment classification on twitter data using support vector machine." *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018.

- [26] Hsu, C., C. Chang and C. Lin. "A practical guide to support vector classification." (2003): 1396-1400.
- [27] Demuth, H. B., M. H. Beale, O. D. Jess and M. T. Hagan. "Neural Network Design." *Martin Hagan*. 2014.
- [28] Ojha, V. K., A. Abraham and V. Snášel. "Metaheuristic design of feedforward neural networks: A review of two decades of research." *Engineering Applications of Artificial Intelligence* 60 (2017): 97-116.
- [29] Dhar, S., S. Pednekar, K. Borad and A. Save. "Sentiment Analysis Using Neural Networks: A New Approach." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018.
- [30] Paliwal, S., S. K. Khatri and M. Sharma. "Sentiment Analysis and Prediction Using Neural Networks." *International Conference on Advanced Informatics for Computing Research*. Springer, Singapore, 2018.
- [31] Rish, I.. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. 2001.
- [32] Wijayanto, U. W. and R. Sarno. "An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naive Bayes." *2018 International Seminar on Application for Technology of Information and Communication*. IEEE, 2018.
- [33] Kohavi, R. and J. Ross Quinlan. "Data mining tasks and methods: Classification: decision-tree discovery." *Handbook of data mining and knowledge discovery*. 2002. 267-276.

- [34] Safavian, S. R. and D. Landgrebe. "A survey of decision tree classifier methodology." *IEEE transactions on systems, man, and cybernetics* 21.3 (1991): 660-674.
- [35] Bibi, R., U. Qamar, M. Ansar and A. Shaheen. "Sentiment Analysis for Urdu News Tweets Using Decision Tree." *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2019.
- [36] Joseph, F. J. J.. "Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree." *2019 4th International Conference on Information Technology (InCIT)*. IEEE, 2019.
- [37] Gao, Y., W. Rong, Y. Shen and Z. Xiong. "Convolutional neural network based sentiment analysis using Adaboost combination." *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016.
- [38] Deng, K.. "Omega: On-Line Memory-Based General Purpose System C." *Carnegie Mellon University*. <https://www.cs.cmu.edu/~kdeng/thesis/feature.pdf>. Nov 1998.
- [39] Suchetha, N. K., A. Nikhil and P. Hrudya. "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification." *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2019.
- [40] AlindGuptaCheck . "ML: Extra Tree Classifier for Feature Selection." *GeeksforGeeks*. <https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/>. 2019.

- [41] Asaithambi, S.. “Why, How and When to apply Feature Selection.” *Towards Data Science*. <https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2>. 2018.
- [42] Naji. “Twitter Sentiment Analysis Training Corpus (Dataset).” *Thinknook*. <http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/>. 2018.
- [43] “What is Python? Executive Summary.” <https://www.python.org/doc/essays/blurb/>. 2019.
- [44] “Anaconda Navigator.” *Anaconda Navigator - Anaconda documentation*. <https://docs.anaconda.com/anaconda/navigator/>.
- [45] Team, S.. “Spyder: The Scientific Python Development Environment.” *Spyder Website*. <https://www.spyder-ide.org/>.
- [46] Manning, C. D., P. Raghavan and H. Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.
- [47] “Array objects.” *Array objects - NumPy v1.18 Manual*. <https://docs.scipy.org/doc/numpy/reference/arrays.html>.
- [48] Sammut, C. and G. I. Webb, eds. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

## VITA

Christina Hastings

Prairie View A&M University

Department of Electrical and Computer Engineering

### Education

M.S. Electrical and Computer Engineering, Prairie View A&M University, 2020.

B.Sc. Electrical and Computer Engineering, Prairie View A&M University, 2017.

### Employment

Prairie View A&M University, Graduate Research Assistant, 2017-2020.

### Internships

Air Force Research Lab (AFRL), Summer 2019.

Air Force Research Lab (AFRL), Summer 2018.

### Skills & Competencies

Have a solid background in electrical engineering as well as effective learning and communication abilities

Proficient in Python