DETECTION OF ZIKA VIRUS MISINFORMATION ON TWITTER THROUGH CHAINED GRADIENT BOOSTING AND DESCRIPTIVE FEATURES

by

VICTOR JOSEPH RUBERTI JR.

(Under the Direction of Ismailcem Budak Arpinar)

ABSTRACT

As the most recent spread of the Zika virus continues, there is an increasing presence of Zika related information on social media platforms such as Twitter. The information found on Twitter provides a unique opportunity to obtain real-time news and firsthand accounts about a variety of subjects and events but there is potential for the spread of misinformation. During crisis events, it is important for users to be able to find accurate and timely information regarding safety precautions and potential threats. This research provides a valuable opportunity to detect misinformation related to the Zika virus on Twitter through a chained gradient boosted model using a unique set of descriptive features associated with Tweet formation, author credentials, subject matter, and author intention.

INDEX WORDS: Misinformation Detection, Zika Virus, Social Media

DETECTION OF ZIKA VIRUS MISINFORMATION ON TWITTER THROUGH CHAINED GRADIENT BOOSTING AND DESCRIPTIVE FEATURES

by

VICTOR JOSEPH RUBERTI JR.

BS, Georgia College & State University, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2016

DETECTION OF ZIKA VIRUS MISINFORMATION ON TWITTER THROUGH CHAINED GRADIENT BOOSTING AND DESCRIPTIVE FEATURES

by

VICTOR JOSEPH RUBERTI JR.

Major Professor:	Ismailcem Budak Arpinar
Committee:	John M. Drake
	Krzysztof J. Kochut

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2016

Acknowledgements

I would like to take this opportunity to thank Dr. Arpinar, Dr. Drake, and Dr. Kochut for their support in this research. Without Dr. Arpinar and Dr. Drake's continued direction and guidance, this project would not have been possible.

Table of Contents

List of Tables

Page

List of Figures

Page

Chapter 1

Introduction

Zika virus infections were reported as early as 1952 but the virus' most recent spread across the globe has caused concern and panic amongst the general public as complications associated with the virus are being discovered [49]. As a result of the current outbreak, social media platforms and news headlines are being flooded with information relating to the Zika virus. Twitter is a popular example of a social media platform where users strive for recognition by spreading information in the form of Tweets. Due to its rising popularity, a large number of users rely on Twitter as their primary news source [23].

Twitter provides the unique opportunity for users to discover firsthand accounts and real-time information about a wide variety of subjects, such as Zika, but since anyone can author a Tweet, there is no way to confirm if a Tweet includes valid information or not. Twitter users that strive for recognition and a larger following often post Tweets as quickly as possible and may end up posting misinformation regardless of the fact that they believe their Tweet is true. Furthermore, there are users who will intentionally spread misinformation to shock readers and gain recognition.

Since there is a potential for accidental and intentional misinformation on Twitter, it is important for users who are seeking information related to a crisis event, such as the

spread of the Zika virus, to be able to identify the validity of a Tweet. In regards to public health, it is important to identify and correct misinformation to avoid panic and allow the general public to equip themselves with proper knowledge on how to prevent or handle disease. Additionally, it is important to understand why people spread misinformation and what types of people spread misinformation. Figure 1 demonstrates an instance of how misinformation presents itself on Twitter. By exploring the attributes associated with a person who spreads this type of information, it is possible to understand why they spread the misinformation. Through further analysis of the spread misinformation, it is possible to derive more attributes that can potentially distinguish misinformation from true information. The purpose of this research is to utilize these attributes as features incorporated into a predictive model to determine whether or not a Tweet includes misinformation and gain a better understanding of who spreads misinformation and how misinformation presents itself on Twitter.

Figure 1 - A Tweet containing confirmed misinformation [52]

Chapter 2

Background Information

2.1 Zika Virus

Even though there is a common misconception that the Zika virus is a newly identified virus, it was first isolated in April 1947 from a rhesus monkey found in the Zika forest of Uganda where the virus got its name [20]. Scientists in Uganda and the United Republic of Tanzania were the first to identify the Zika virus in humans in 1952. The Zika virus belongs to the genus Flavivirus and is most commonly spread by infected mosquitoes belonging to the Aedes genus. In tropical regions, the Aedes aegypti is the main transmitter of the Zika virus and is the species most commonly associated with the virus. Aedes mosquitoes are primarily active during the day and are also responsible for spreading dengue and chikungunya viruses. The mosquitoes spread the virus by feeding on an individual previously infected with the Zika virus and then biting other humans [61].

In 2014, researchers confirmed that the virus can also be transferred from mother to infant via transplacental transmission or during delivery of the baby. It was not widely accepted that the Zika virus can be transmitted sexually until the United States reported a case of sexually transmitted Zika virus infection in Texas on February 2, 2016. Prior to the case in Texas, there were two other documented cases of sexually transmitted Zika virus

infection dating back to 2008 and 2013. It was not until the 2013 case that researchers were able to isolate the Zika virus from semen and provide concrete evidence that the Zika virus can be sexually transmitted [27]. During the course of this research, other Zika virus infection transmission mediums have been reported, including blood transfusion and laboratory exposure [61].

An individual infected with Zika may not present with any noticeable symptoms but if any occur, they are usually mild. Fever, rash, joint pain, and conjunctivitis are the most common symptoms associated with the Zika virus. If symptoms are present, they usually last in an infected individual for two to seven days. Since the symptoms associated with Zika are common among other infections, Zika virus is usually only a concern if the individual presenting symptoms visited an area infected with Zika. Blood or urine tests are used to definitively diagnose a Zika virus infection [59].

Prior to 2015, the virus was thought to only cause mild illness. It was not until Brazil confirmed its first case of Zika virus infection in May 2015 and began reporting associations between Zika virus infection and Guillain-Barré syndrome in July 2015 and microcephaly in October 2015 that links to other complications were investigated. Guillain-Barré syndrome is a disorder where the immune system of an individual begins to attack their peripheral nervous system. The disorder often presents with weakness and tingling in the legs but can increase in intensity until a person is paralyzed and in some rare cases, the condition can lead to death [35]. The CDC currently states that there is a strong correlation between Zika virus infection and Guillain-Barre syndrome [62]. The New England Journal of Medicine (NEJM) reports 1,474 cases of Guillain-Barre syndrome

potentially associated with Zika virus infection in Bahia, Brazil, Colombia, the Dominican Republic, El Salvador, Honduras, Suriname, and Venezuela. To emphasize this association, Venezuela has experienced an 877% increase in the incidence of Guillain-Barre when compared to a baseline before widespread Zika virus infection [36].

Microcephaly is a condition that usually occurs during fetal development where the head of an infant is underdeveloped and smaller than normal. The condition is usually caused by impaired growth but when the condition is associated with Zika, it is often caused by the death of tissues and cells leading to the brain. An individual with microcephaly may experience neurological problems such as seizures and problems with cognition, motor functions, speech, coordination, and balance [35]. At the time of this research, the consensus at the World Health Organization (WHO) is that Zika virus infection is a cause of microcephaly [63]. Additionally, the CDC has characterized the birth defects associated with Zika infection, including microcephaly, as Congenital Zika Syndrome. This syndrome accounts for microcephaly, decreased brain tissue, eye damage, limited range of motion, and muscle restrictions that could result from Zika virus infection during pregnancy [60].

At the present, there is no vaccine or definitive cure for Zika virus infection but since the symptoms are usually mild in nature, resting and drinking fluids are encouraged and additional medical attention is only required if symptoms worsen. According to the WHO, the best way to prevent infection is to cover as much of the body as possible, use screens for doors and windows, and use insect repellent [63].

As Zika virus infections continue to spread, it is important for the public to consume and spread accurate information. There are a large number of countries with active Zika virus transmission, including the United States, so it is vital for the people in these areas to learn proper prevention techniques and become familiar with the risks associated with Zika [58]. For this reason, the ability to determine the validity of information and stop the spread of misinformation associated with the virus is necessary for maintaining a healthy and educated global population.

2.2 Twitter

With 310 million active monthly users and 1 billion unique visits per month, the online social media platform and microblogging service known as Twitter allows users to broadcast and consume real-time information from all over the world [2]. This information is packaged into 140 character messages called Tweets, which any registered Twitter user is allowed to send. Tweets can be read by anyone who can access Twitter via its website, SMS, or mobile application [38]. Since its founding in 2006, Twitter has grown into a global platform where users post content ranging from prominent events to ordinary tasks. On average, 6000 Tweets are posted every second and approximately 500 million Tweets are posted per day [54]. A Twitter user can follow other registered users to see their content. When posting a Tweet, Tweets can be grouped by using a hashtag or reposted from another Twitter user and shared to the reposting user's followers [55].

Since this sharing and grouping process occurs in real-time, Twitter functions as an ideal platform for information consumption of time sensitive events

From the statistics listed above, it is apparent that Twitter is a popular information sharing platform and due to its popularity, Twitter data is frequently used for research relating to human behavior [4]. The opinions and concerns found in millions of Tweets posted per day are a valuable tool for gauging public knowledge and reactions on certain subjects. With its use in research, Twitter's popularity has caused some users to rely on Tweets as their primary source of news. Users turn to Twitter for breaking news on emergent situations because it is possible to read firsthand accounts of users experiencing certain events in real-time. According to a report from the Pew Research Center from May 2016, 59% of Twitter users use Twitter as a news source, which is a 7% increase from their report from 2013 [23]. With the rising use of Twitter as a new source, it is important to remember that any Twitter user can post information as a Tweet. These Tweets may provide a unique perspective on a variety of subjects and events but without any form of fact checking, there is a chance that a Tweet contains misinformation. Accordingly, it is important for users to be able to determine if a Tweet they are using as a news source contains misinformation.

Chapter 3

Related Works

3.1 Twitter as a News Source

Since Twitter is commonly used as some user's primary news source and any of Twitter's millions of users can instantly post Tweets, researchers have started to recognize information credibility as a valuable area of research to pursue on social media data. Kwak, Lee, Park, and Moon helped establish the notion of Twitter as a news source by examining 41.7 million user profiles and 106 million Tweets showing that 85% of Tweet topics were news related [28].

3.2 Features

3.2.1 User Features

After Twitter became known for its potential role as a news source, it became a popular source of information regarding emergent events. A study on Tweets relating to emergent events was published in 2011 by Castillo, Mendoza, and Poblete that found features associated with Tweets that identify credibility. An account's registration age, follower count, and number of Tweets posted were the features that best described a perceivably credible Tweet [11]. Castillo's research established the importance of features

related to the user and their credibility. A similar study published by Gupta and Kumaraguru in 2012 also found user-based features, especially number of followers, to have a strong correlation with the perceivable credibility of a Tweet [25]. In the same way, recent research composed by Liang, He, Xu, Chen, and Zeng focused on identifying rumors on the Chinese microblogging service, SinaWeibo, and found user-based features, such as registration age and number of followers, to be important indicators of posts containing rumors [29].

3.2.2 Tweet Features

The basic content of a Tweet has proven to be a useful indicator of a Tweet's credibility. For instance, the occurrence and number of hashtags in a Tweet have proven useful when distinguishing between credible and speculative Tweets. Kalyanam, Velupillai, Doan, Conway, and Lanckriet conducted a study in 2015 of the hashtags found in Tweets related to Ebola and observed that credible Tweets contained more hashtags than speculative Tweets [26].

In addition to hashtag occurrence and frequency, the occurrence of URLs has also been explored. Castillo's previously discussed credibility study observed that Tweets containing URLs were deemed more credible than those without URLs [11]. Gupta's and Kalyanam's credibility research also found similar results [25] [26]. A Retweet analysis performed by Zubiaga, Liakata, Procter, Hoi, and Tolmie in 2016 detailed the flaw with studying perceived credibility by showing how reputable users were starting more misinformation and backing it with evidence in the form of URLs to appear more

credible. Due to its disputed significance, URL presence is an important feature to explore in regards to misinformation identification on Twitter [64].

3.2.3 Propagation Features

Information relating to how a Tweet propagates is an important feature to consider due the potential of explaining a Tweet's popularity, who is spreading the Tweet, and why they are spreading the Tweet. In a different study published in 2010 by Mendoza, Poblete, and Castillo, they analyzed the Retweet network of Tweets containing misinformation and true information relating to the 2010 earthquake in Chile. Their analysis found that misinformation propagates differently than true information because users were questioning misinformation more often than the true information [34]. This is a valuable observation because it provides evidence that features related to propagation can distinguish between false and true information. In the following year, research conducted by Qazvinian, Rosengren, Radv, and Mei further backed the claim that propagation features are useful in identifying misinformation in Twitter with their study on identifying emergent rumors on Twitter. Their classifier yielded 84.8% accuracy when using network features relating to Retweet behavior to classify rumors [40].

An equally important use of propagation features is to describe the type of user spreading misinformation. Zubaiaga, Liakata, Procter, Hoi, and Tolmie's 2016 study analyzed the Retweet threads of Tweets containing rumors and found that reputable users with a high follower to following ratio, such as news agencies, were more likely to spread misinformation [64].

3.3 Classification Techniques

The previously discussed information credibility and misinformation classification studies employ supervised learning classification models. Castillo's 2011 information credibility study utilized a J48 decision tree classifier to predict a Tweet's perceivable credibility with an accuracy of 86% and precision and sensitivity between 70% and 80%, Precision denotes the number of false positives and sensitivity denotes the number of true positives [11]. Gupta's 2012 Tweet credibility study made use of Ranking SVM, which is an implementation of the support vector machine algorithm that uses pair-wise ranking functions to sort results based on credibility score [25]. Liang's 2015 rumor identification research tested five types of classification algorithms to find the best performing classifier. Logistic regression, SVM, Naive Bayes, decision tree, and K-nearest neighbors were tested. Among all of the precision and sensitivity results of the tested classifiers, the decision tree classifier performed the best with precision and sensitivity of approximately 85% [29].

Chapter 4

Approach

4.1 Problem Definition

Due to the rise in popularity and previously stated adoption of Twitter as a news source, it is important to detect and understand how and why misinformation spreads on Twitter [23]. This importance becomes emphasized when misinformation revolves around a public health concern such as the spread of the Zika virus. There are currently a significant number of countries with active Zika virus transmission, including the United States, and the ability to detect misinformation and understand why it spreads is a valuable tool to maintain public health [58]. This research aims to answer three main questions regarding misinformation on Twitter: (1) is it possible to accurately predict whether or not a Tweet related to Zika carries misinformation, (2) what indicators associated with a Tweet can distinguish its validity, and (3) what type of users spread misinformation.

Consider a collection of Zika related Tweets. This research's goal is to utilize optimal features to produce a set of Tweets with prediction labels indicating if a specific Zika related Tweet contains misinformation or not. In order to accomplish this, a chain of two gradient boosted models from R's Generalized Boosted Regression Model (GBM) package is employed. The purpose of chaining these models is to use the resulting

classifications from one model as features in the next model to generate a unique features set. The reasoning behind the use of this type of model is discussed in Section 4.2.4 and the chaining process is detailed in Section 4.2.5. The features utilized in this chained model approach are a combination of features effectively used in previous studies with a group of features attempting to describe the subject and intention of a Tweet and its author. For a full description of the features used in this research, refer to Chapter 6. In this chained approach, the first classification model determines if the Tweet is making a claim or not. This information is potentially valuable since misinformation usually presents itself as a claim or statement of fact. The claim classifications obtained from this model are then used as a feature in combination with 40 base features to accomplish misinformation classification. The resulting misinformation classifications are then used to answer the three main research questions stated previously.

4.2 Classification

It is difficult to conclusively verify information regarding emergent subjects, but it is possible to gauge credibility and predict the occurrence of misinformation. In order to find a classification technique that can predict misinformation with improved prediction accuracy over previous approaches, the performance of various classification techniques were researched and detailed in Section 3.3. From the precision and sensitivity results of the researched techniques, the decision tree classifier performed the best for identifying rumors with precision and sensitivity around 85% [29]. With this in mind, the next step is to understand why decision tree learning excelled in rumor identification and how can

the technique be further improved and utilized to accomplish misinformation detection. The following sections detail how decision trees can be improved using a combination of weaker decision trees to form an ensemble, how an ensemble can be used to further improve prediction accuracy through a technique called boosting, and how this research utilizes boosting to accomplish misinformation classification [14] [42].

4.2.1 Decision Trees

Decision tree classification works by utilizing a tree structure as a predictive model. In the tree structure, the first node encountered is called the root node, the nodes inside the tree with only one incoming edge and two or more outgoing edges are known as internal nodes, and the nodes with only one incoming edge and no outgoing edges are referred to as leaves. In the case of a decision tree, the leaves represent labels and the root and internal nodes represent decisions relating to attributes associated with a particular data instance. These decisions specify if a particular data instance satisfies a certain condition or not. After the decision process no longer encounters any internal nodes and reaches a leaf node, then the label associated with that leaf is assigned to that data instance [37]. Since the decisions within the tree structure inherently perform feature selection, the success of decision tree classification in the domain of information credibility may result from the ability to include a large number of features and discover meaningful patterns among them.

4.2.2 Ensemble Techniques

Although decision tree classifiers proved to be the most efficient performers when considering information credibility and identifying rumors, it is recognized that using a single decision tree as a predictive model can produce unstable results [15]. This notion originates from the idea that any changes to the training data can cause entirely different models to be generated. In order to address this variance, an ensemble of decision tree models can be employed in the place of one decision tree to produce more stable results [14].

There are several ensemble approaches used in machine learning to avoid overfitting and improve the stability of decision tree learning but the most popular techniques are stacking, bagging, and boosting. Stacking uses the outputs of different base models as features for a final model. Bagging, which stands for Bootstrap Aggregation, works by creating multiple models in a parallel fashion from different bootstrap samples where each sample is a random sample of the training dataset with replacement. Sampling with replacement allows for certain training examples to appear more than once or not at all in a particular bootstrap sample. The bagging process combines the multiple models by majority voting for classification and as a result, it is able to reduce variance but increases bias [57]. Boosting relies on converting weak models into a single powerful model through a process of incremental adjustments. Contrary to the parallel ensemble creation used in bagging, boosting generates its ensemble in a sequential fashion with models being added to strengthen areas of the ensemble that misclassified in previous iterations [9]. Compared to bagging, boosting has proven more

effective at improving prediction accuracy of decision tree classifiers and is capable of reducing bias and variance [42] [6].

4.2.3 Boosting Techniques

The first and most popular implementation of boosting is Freund and Schapire's 1995 Adaptive Boosting algorithm (AdaBoost) [17]. AdaBoost gets its name from how it adaptively adjusts weights on the weak models and training data to improve predictions. The algorithm starts with a model built from equally weighted training data then creates another model that aims to improve the prediction accuracy of the previous model by increasing weights given to predictions with high errors rates in the previous model. The weak models are sequentially added to help manage difficult sections of the data [9].

A statistician named Breiman created a framework known as ARCing algorithms in 1996, which stands for Adaptive Reweighting and Combining in an effort to better define boosting algorithms such as AdaBoost. Breiman's framework explained that AdaBoost and other boosting algorithms were ARCing algorithms because they perform a weighted minimization of the misclassification rate and recompute weights for all iterations [6].

In an effort to generalize AdaBoost and improve upon Breiman's framework, Friedman proposed a statistical framework called Gradient Boosting Machines in 2001, which is commonly referred to as gradient boosting. In Friedman's approach, boosting is thought of as a numerical optimization problem that follows a gradient descent approach that minimizes a loss function by sequentially adding weak models to an ensemble [18]. In gradient descent, the local minimum of a function is found by using weights to

calculate costs and taking the derivative of the calculated cost. The derivative is used to find the slope of the function used to calculate cost and with this slope, the direction of the function can be determined and the weights can be updated to minimize cost [10].

The generic structure of a gradient boosting algorithm is comprised of a loss function, weak models, and an additive model [8]. In the case of classification, the purpose of the loss function is to assign a numeric value to inaccurate predictions where minimizing this value results in more accurate predictions [31]. How the loss function is minimized is determined by the classification problem and since the gradient boosting framework is generic, any differentiable loss function can be applied. The additive model adds the weak models, which are regression trees in the case of gradient boosting, to the ensemble one at a time. In order to follow the previously detailed gradient descent procedure, the algorithm updates the weights of the new tree in a way that follows the gradient and minimizes classification error [8]. Trees are added until prediction accuracy on the training data is perfect or a predetermined number of trees are added to the ensemble [9].

4.2.4 Generalized Boosted Regression Models in R

R is a programming language primarily used for statistical computing and features several predictive modeling packages equipped with popular techniques for improved prediction accuracy [43]. Among these packages is the Generalized Boosted Regression Model (GBM) package that implements the gradient boosting technique previously described in section 4.2.3 [45]. Since decision tree learning proved successful in the

domain of information credibility and gradient boosting addresses the disadvantages commonly associated with decision tree learning, R's GBM package was selected for classification in this research. Since the classifications in this research only care about two unique values, whether or not a Tweet includes misinformation, the data follows a Bernoulli distribution so the Bernoulli loss function was selected to be used by GBM's gradient boosting algorithm. The final classifications of a GBM using gradient boosting with a Bernoulli loss function are continuous predictions ranging from 0 to 1. After rounding the predictions, a value of 1 indicates a positive classification and a 0 indicates a negative [44].

In order to avoid generating a model that is too complex and fails to generalize the data, GBM allows you to limit the number of trees added to the ensemble during creation [44]. The models were first fit using a large number of trees and the number of trees required to minimize the loss function was recorded. This recorded number was then used to limit the number of trees added to the ensemble during creation of the final model.

4.2.5 Chained Models

This research utilizes two GBMs in a unique chained approach to accomplish misinformation classification. The first model predicts whether or not a Tweet is making a claim or not and is described in detail in Chapter 7. The resulting binary classification from the claim model is then used as a feature in the second model, which performs the final misinformation classification and is detailed in Chapter 8. This chaining technique

aims to produce a more accurate misinformation classification by diversifying the final model's feature set.

4.3 Dataflow and Architecture

The following dataflow steps detail how data is utilized and manipulated throughout the classification process performed by the Claim Detection Model in Section 4.3.1 and the Misinformation Detection Model in Section 4.3.2. The architecture forming the main components of this research is detailed in Section 4.3.3.

4.3.1 Claim Detection Dataflow

1. Inputs: D1 (features of all Tweets), D2 (randomly selected subset of D1 with manual labels Y1 for "making a claim" or "not making a claim")
2. Numerical analysis of D1; define set of transformations T for variables in D1 as needed; refer to a data set D transformed by T as T(D)
3. Perform Principal Component Analysis (PCA) on T(D1); select the top 10 Principal Components (PC) as additional features; compute PC scores for all records in D; append these to other datasets yielding D2P
4. Randomly split D2p into D2aP (training) and D2bP (testing) while maintaining equivalent ratios of Y1 labels between D2aP and D2bP
5. Train model for "claims" using a Generalized Boosted Regression Model (GBM) to learn labels Y1 from features in D2aP yielding M1

6. Output: Apply model to M1(D1); append the resulting "claim" prediction labels to D1 as an additional feature yielding D3 (features of all Tweets including claim or not)

4.3.2 Misinformation Detection Dataflow

1. Inputs**:** D3 (features of all Tweets including claim or not and excluding the top 10 Principal Components from Claim Detection), D4 (subset of D3 a positive Y2 label indicates "confirmed misinformation" and a negative Y2 label "confirmed credible information")
2. Apply T to variables in D3 as needed yielding T(D3)
3. Perform PCA on T(D3); select the top 10 PCs as additional features; compute PC scores for all records in D; append these to other datasets yielding D4P
4. Randomly split D4P into D4aP (training) and D4bP (testing) while maintaining equivalent ratios of Y1 labels between D4aP and D4bP
5. Train model for "misinformation" using a GBM to learn labels Y2 from features in D4aP yielding M2
6. Output: Apply model to M2(D3); append the resulting "misinformation" prediction label to D3 yielding D5 (features of all Tweets including claim and misinformation)

4.3.3 Architecture

The architecture forming the main components of this research is illustrated in Figure 2. Each component of this architecture is described below.

1. Data Collection - Tweets related to Zika are collected from Twitter Streaming API.

2. Feature Set Generation - Features resulting from data analysis and previous studies are generated from the Zika Tweets.
3. Transformation - Features exhibiting log normal distribution are log transformed.
4. Principal Component Analysis (PCA) – PCA is separately conducted on the generated features for each models' feature sets, and the top 10 Principal Components (PCs) are included in both feature sets.
5. Claim Detection - A randomly selected subset of the collected Zika Tweets is manually annotated as being a claim or not. This subset is split into training and testing data. The training data is used to generate the Claim Detection model in R using the Generalized Boosted Regression Model (GBM) package. The resulting Claim Detection Model is applied on the entire collection of Zika Tweets. The resulting claim classifications are added to the Misinformation Detection Model's feature set.
6. Misinformation Detection – Credible information and misinformation are identified and categorized in the Zika Tweets using reputable sources. Using these categories, a selected subset of the Zika Tweets are annotated as "Confirmed Credible" or "Confirmed Misinformation" with equivalent ratios of each annotation. This subset is split into training and testing data with each split retaining equivalent ratios of each annotation. The training data is used to generate the Misinformation Detection GBM in R. The Misinformation Detection Model is applied on the entire collection of Zika Tweets to accomplish misinformation detection.

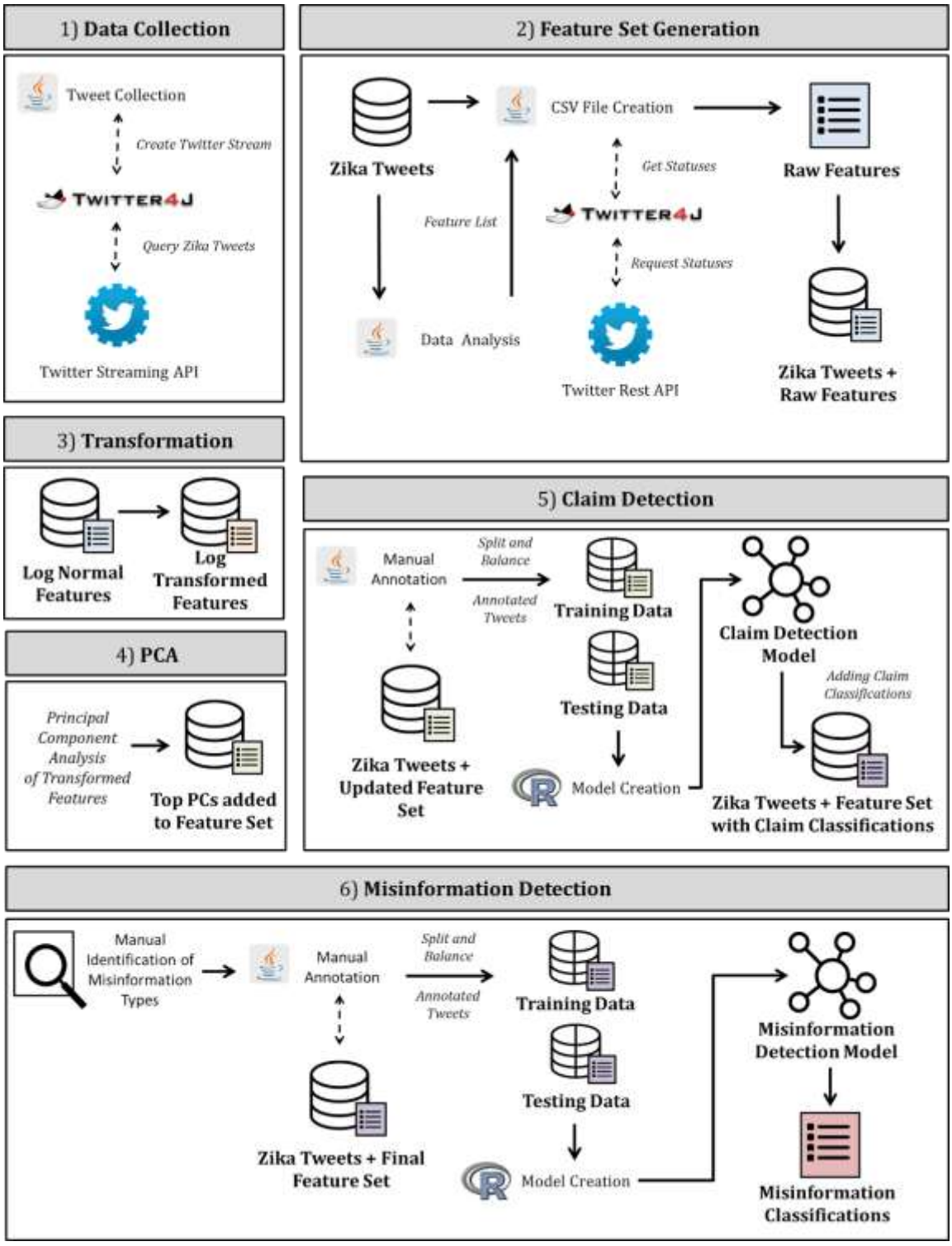

Figure 2 - Architecture Overview

Chapter 5

Dataset

5.1 Data Collection

In order to obtain Tweets related to the Zika virus, the Java library Twitter4J was used to access Twitter's Streaming API. Since Twitter rarely permits applications to use the entire stream of all incoming public Tweets, a customized public stream was employed to collect Tweets related to Zika in near real-time. The public streams accessible by the Streaming API are believed to allow access from 1% to 40% of all newly submitted Tweets [53].

A Tweet collected from a public stream is extracted as a component from a status with attributes detailed by the Twitter API. These attributes allow for the stream to be customized to filter out Tweets based on certain criteria. The criteria used in this research filtered Tweets containing the word "Zika" in the content and hashtags of the Tweet and only considered Tweets written in English. In order to account for future data processing, the numerical ID of the desired Tweet and Author were recorded along with the content of the Tweet and the location listed in the Author's profile. In the instance of a user with no location listed in their profile, a value of null was recorded for their location.

Data was collected over the course of 80 days from February 17, 2016 to May 7, 2016 and 996,443 Tweets relating to Zika were collected. On average, the stream collected

approximately 12,500 Tweets per day. The smallest amount of Tweets collected in a single day occurred on March 20, 2016 with 3,295 Tweets collected. The largest amount of Tweets collected on a single day occurred on April 13, 2016 with 55,343 Tweets collected. The large volume of Tweets collected on April 13, 2016 coincides with an official statement made by the CDC stating that the "Zika virus infection can cause microcephaly and other severe fetal brain defects" [12].

5.2 Data Analysis

In order to understand the contents of the collected data and gauge its potential use for misinformation detection, a word frequency analysis was performed on the dataset. Before processing the word frequencies, certain words known as stop words were filtered out of the dataset to help focus on the content of the Tweets. The stop word list consisted of contextually redundant words such as "Zika", single letters, and insignificant parts of speech such as articles and prepositions.

The resulting frequency list, shown in Table 1, allowed for the recognition of different subjects and intentions. For example, there are a significant number of words indicating a new report such as "first" occurring in 6% of Tweets, "new" occurring in 6% of Tweets, and "confirmed" occurring in 3% of Tweets. Tweets regarding a new report or evidence might include information that is not confirmed because the Tweet's author wanted to be the first to report the information.

Words hinting at disease-specific topics such as "CDC" occurring in 6% of Tweets, "spread" occurring in 3% of Tweets, and "outbreak" occurring in 3% of Tweets emphasize

the potential severity of a Tweet's content and may elicit fear in the reader causing them to Retweet and spread potential misinformation.

The presence of words, such as "may" occurring in 4% of Tweets, "could" occurring in 3% of Tweets, and "should" occurring in 1% of Tweets, demonstrate how people talk when they are uncertain about a topic. Additionally, if a user is unsure about the information they are posting and want to distance themselves from any potential blame, they will distance themselves from the Tweet and include words such as "says" occurring in 4% of Tweets, "study" occurring in 3% of Tweets, and "reports" occurring in 2% of Tweets. The insight gained from the aforementioned frequency analysis helped form the basis of the string variables included in the descriptive features of the Claim Detection Model and Misinformation Detection Model discussed in Section 6.4.

Table 1 - Word Frequencies of the 996,443 Zika related Tweets

Word Frequency (% of Tweets) Word

72287 (7%) health	34074 (3%) outbreak	21046 (2%) south
67242 (6%) case	33528 (3%) birth	20768 (2%) sexually
66190 (6%) first	32117 (3%) pope	19803 (2%) Puerto
62836 (6%) new	31993 (3%) study	18992 (1%) confirms
59527 (6%) who	31992 (3%) link	18758 (1%) should
59382 (6%) women	31698 (3%) could	18052 (1%) sexual
59216 (6%) cdc	27476 (2%) travel	17926 (1%) cause
53142 (5%) news	26909 (2%) risk	17770 (1%) evidence
52565 (5%) pregnant	26351 (2%) scientists	17577 (1%) combat
51661 (5%) brazil	25771 (2%) transmitted	17544 (1%) research
48677 (5%) cases	25131 (2%) vaccine	17184 (1%) officials
45485 (4%) microcephaly	24563 (2%) defects	16985 (1%) test
41810 (4%) fight	24556 (2%) transmission	16890 (1%) rio
40132 (4%) says	23992 (2%) world	16576 (1%) report
39994 (4%) may	23454 (2%) reports	15666 (1%) disease
39897 (4%) mosquitoes	23334 (2%) ebola	15516 (1%) linked
37633 (3%) spread	21580 (2%) contraception	15364 (1%) help
36195 (3%) confirmed	21306 (2%) brain	14912 (1%) infection
35242 (3%) mosquito	21276 (2%) olympics	14857 (1%) crisis
34922 (3%) will	21179 (2%) google	14350 (1%) house

Chapter 6

Features

Since this research performs misinformation classification using a chained model, two sets of features are used. The first set of features focuses on Tweet formation, author credentials, subject matter, and author intention. These 40 features are grouped into the following categories: User, Propagation, Tweet, and Descriptive Features and are detailed in the following sections. For the Claim Detection Model, Principal Component Analysis (PCA) is performed on the 40 features and the top 10 Principal Components (PCs) are used in the feature set. The second set of features used for the Misinformation Detection Model consists of the same 40 categorized features used for claim classification combined with the resulting label from the Claim Detection Model indicating if a Tweet is making a claim or not. PCA is performed on the 41 features and the top 10 PCs are used as features for the Misinformation Classification Model.

6.1 User Features

In past Twitter-based research, features pertaining to a user's Twitter account proved useful in determining the credibility of a Tweet. Specifically, Castillo's 2011 research pioneered the use of features such as an account's registration age, follower count, and number of Tweets posted for determining how credible a Tweet appears [11].

In addition to account age, number of followers, and number of Tweets posted by a particular Tweet's author, this research incorporates the number of users followed by the author, their follow ratio, which is the number of users following the user divided by the number of users the user is following, and whether or not the author has their location listed in their Twitter profile. In order to track how and where misinformation is spreading, the location of a user spreading misinformation needs to be known. The most obvious solution to this would be to obtain GPS coordinates or tagged location associated with a Tweet. However, this technique only yielded a few Zika Tweets per day. With an alternative approach, we extracted the location listed in the Tweet author's profile and then geocoded the result to determine if an actual location was listed. Geocoding the listed location is necessary because their profile's location can be set manually by the user. More information regarding the geocoding process can be found in Section 9.2. By incorporating proven user features and other features relating to a user's profile into classification, this research aims to define what type of user spreads rumors. The full list of User Features and their descriptions is shown in Table 2.

Table 2 - Description of User Features

User Features	
Name	*Description*
accountAge	Number of days since account creation
locationListed	Location listed in the user's profile
numFollowers	Number of users following the user
numFollowing	Number of users a the user is following
followRatio	numFollowers / numFollowing
numTweetsPosted	Number of Tweets posted by the user

6.2 Tweet Features

In a similar fashion to the user features listed above, previous research also established the importance of incorporating features based on a Tweet's content when determining its credibility [11]. By incorporating Tweet features such as contains hashtags, number of hashtags used in the Tweet, number of characters in the Tweet, contains a question mark, contains an exclamation point, and contains a URL, this research aims to discover potential flags of credibility within a Tweet's content. The full list of Tweet Features and their descriptions is shown in Table 3.

Table 3 - Description of Tweet Features

Tweet Features	
Name	*Description*
hasCountry	Tweet contains name of a country
hasExclamation	Tweet contains a "!"
hasHashtags	Tweet contains hashtags
hasMultimedia	Tweet contains multimedia (video or image)
hasNumber	Tweets contains a number
hasQuestionMark	Tweet contains a "?"
hasURL	Tweet contains a URL
numHashtags	Number of hashtags used in the Tweet
timespan	accountAge – time the tweet was posted
tweetLength	Number of characters in the Tweet

6.3 Propagation Features

The propagation features incorporated in this research are if the Tweet has been favorited, the number of times the Tweet has been favorited, if the Tweet is a Retweet, and the number of times the Tweet has been Retweeted. Previous research found that a Tweet with many Retweets is deemed more credible by the reader [11]. Since a reader is more likely to accept a Tweet with many Retweets, it is possible that a malicious Tweet author seeking to elicit a panic response from their followers may intentionally Tweet

misinformation to generate more Retweets. For this reason, it is important to examine these propagation features to determine what misinformation is popular and why it is spreading. The full list of Propagation Features and their descriptions is listed in Table 4.

Table 4 - Description of Propagation Features

Propagation Features	
Name	*Description*
isFavorited	Tweet has been favorited
numFavorites	Number of times the Tweet has been favorited
isRetweet	Tweet is a retweet
numRetweets	Number of times the Tweet has been retweeted

6.4 Descriptive Features

Unlike the traditional features discussed in the aforementioned studies, this research also considers the presence of certain string variables that may describe the subject and intention of a Tweet. By definition, a rumor is "a currently circulating story of uncertain or doubtful truth" [48]. With this definition in mind, we must consider what words could identify a Tweet as uncertain. When a user describes a topic they are uncertain about, they may speak about the possibility of a statement with words like "could" and "may". With time sensitive information, users might try to report on an event before they know the facts and they might try to distance themselves from an uncertain claim with phrases like "they are saying" and "study suggests". When regarding disease-

specific information, statements might be accompanied with buzzwords like “CDC” and “outbreak” to emphasize the seriousness of an event. Seeing that Zika has become a popular topic in the media, words like “microcephaly” and “mosquitoes” are prominent among advisories related to Zika. The full list of Descriptive Features and their descriptions is shown in Table 5. After the data analysis performed in Section 5.2, words were associated with each Descriptive Feature. The complete list of words associated with each Descriptive Feature is listed in Table 6.

Table 5 - Description of Descriptive Features

Descriptive Features	
Name	*Description*
isAssociation	Tweet contains an association string variable
numAssociation	Number of association string variables in the tweet
isCase	Tweet contains a case string variables
numCase	Number of case string variables in the tweet
isDiseaseSpecific	Tweet contains a disease-specific string variable
numDiseaseSpecific	Number of disease-specific string variables in the tweet
isNewReport	Tweet contains a new report string variable
numNewReport	Number of new report string variables in the tweet
isPolitical	Tweet contains a political string variable
numPolitical	Number of political string variables in the tweet
isPreventive	Tweet contains a preventive string variable
numPreventive	Number of preventive string variables in the tweet
isReinforced	Tweet contains a reinforced string variable
numReinforced	Number of reinforced string variables in the tweet
isTranmission	Tweet contains transmission string variable
numTransmission	Number of transmission string variables in the tweet
isUncertain	Tweet contains an uncertain string variable
numUncertain	Number of uncertain string variables in the tweet
isZikaSpecific	Tweet contains a Zika-specific string variable
numZikaSpecific	Number of Zika-specific string variables in the tweet

Table 6 - String variables associated with each Descriptive Feature type

String Variables for Descriptive Features	
Type	*String Variables*
Association	"associated", "cause(s,ing)", "disorder", "implicated", "indicates" "link(ed,ing)", "not zika", "tied", "zikalinked"
Case	"case(s)", "first", "found", "reports", "test positive"
Disease-Specific	"condition", "epidemic", "medication", "outbreak", "pandemic", "spread(ing)", "symptom", "vaccine"
New Report	"advise", "believe", "breaking", "claim", "confirm(s,ed), "detected", "evidence", "experts", "fear", "investigate", "new report", "reveal", "scientists"
Political	"congress", "fund(s,ing)", "government", "laws", "obama", "pope", "potus", "president"
Preventive	"avoid", "combat", "control", "detect", "develop(ed)", "experimental", "fda approves", "fight", "funding" "prevent(ion)", "protect", "screen(ed,ing)", "stop", "test", "vaccine"
Reinforced	"cdc", "experts", "fda", "health officials", "officials", "scientists", "who"
Transmission	"catch", "caught", "contract", "contracted", "sexually", "spread", "through sex", "transmission", "transmitted"
Uncertain	"appear to be", "could", "experts say", "good chance", "if true", "likely", "may", "maybe", "might", "perhaps", "possible", "potential", "some people think", "study finds", "study suggests", "they are saying", "this might be"
Zika-Specific	"aedes", "guillain", "microcephaly", "mosquito(es)", "pregnancy", "pregnant"

6.5 Data Transformation

Density plots were generated in R for the continuous numerical variables in the data set to visualize the distribution of individual features. These plots illustrate the probability that a feature instance will exhibit a certain value where the Y-axis indicates the probability and the X-axis indicates the value. After examining the density plots of values among continuous numerical variables in the feature sets, it was apparent that the distribution of the data was asymmetric and skewed with most of the density on the left. The density plot of the numerical variable numFollowers is shown in Figure 3 and illustrates this left-skewed distribution. Figure 3 shows that the values have a large range from 0 to 140,000 with most of the values occurring between 0 and 20,000. Since this research employs logistic regression for modeling, it is important to make sure that the numerical training features have as close to normal distribution as possible because it becomes more difficult for classifiers to interpret patterns from skewed distributions [41]. The left-skewed plot shown in Figure 3 illustrates that that the values follow an approximate lognormal distribution, which allows for the opportunity to log transform the data to obtain an approximate normal distribution, which is shown in Figure 4 [30]. The numerical features that exhibited lognormal distribution and were log transformed for the two models are listed in Table 7.

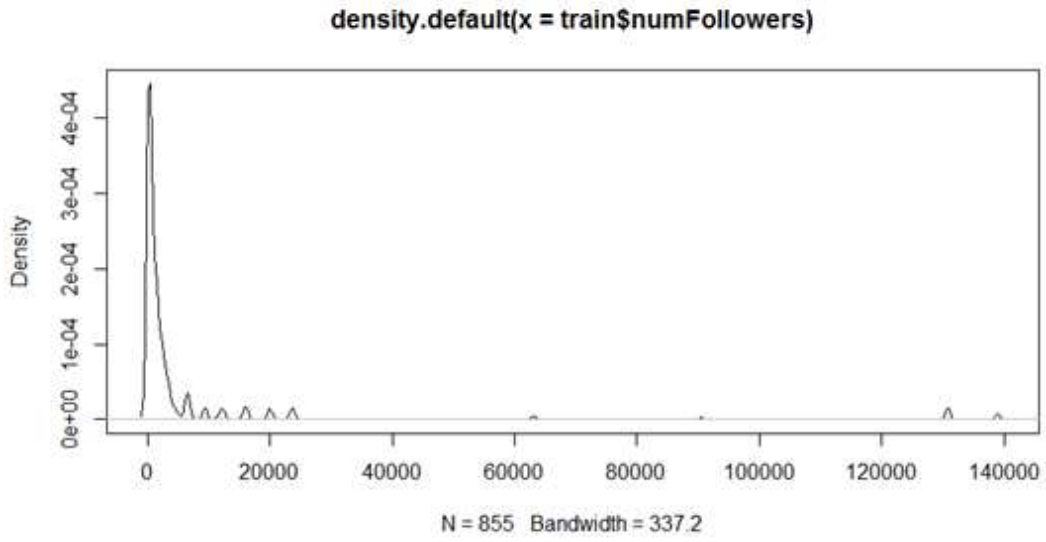

Figure 3 - Density plot of "numFollowers" before log transformation

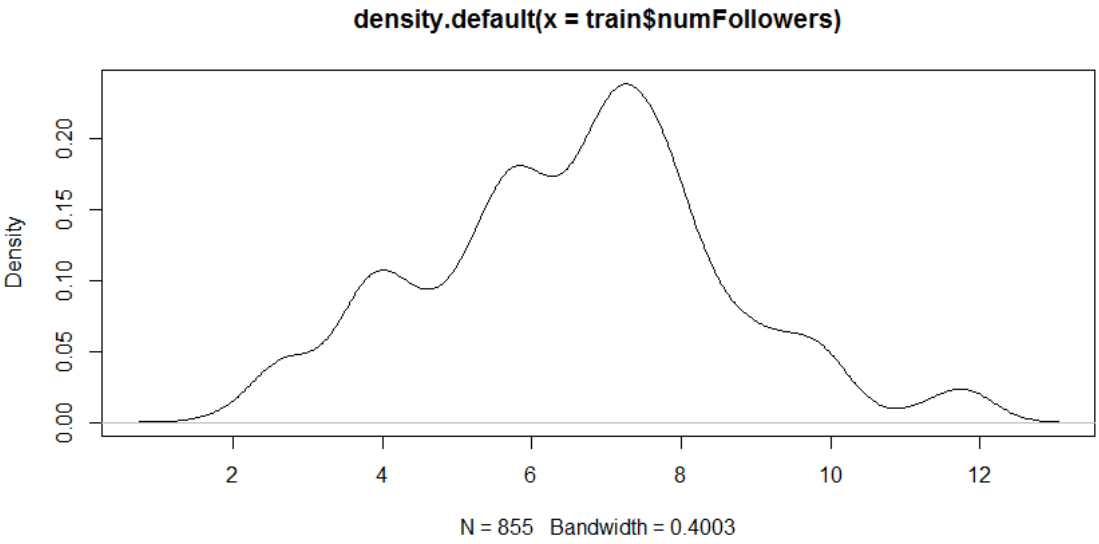

Figure 4 - Density plot of numFollowers after log transformation

Table 7 - List of features that exhibited lognormal distribution and were log transformed for claim and misinformation classification

Log Transformed Features	
Name	*Description*
numFollowers	Number of users following the user
numFollowing	Number of users a the user is following
followRatio	numFollowers / numFollowing
numTweetsPosted	Number of Tweets posted by the user
numFavorites	Number of times the Tweet has been favorited
numRetweets	Number of times the Tweet has been retweeted

6.6 Principal Component Analysis

In order to discover potential patterns amongst the feature set prior to classification, Principal Component Analysis (PCA) was performed on the two feature sets used in this research. The reasoning behind performing PCA is to extract new variables from the existing base features by scoring them based on how much they impact data variance. In order to extract these new score features, Principal Components (PC), which are linear combinations of the original feature values, must be generated from the existing features [46]. PCA generates these PCs by reducing the dimensionality of the original data and finding which direction in the new representation offers the most variance. The direction with the highest variance is the first PC and is the attribute that best summarizes the data. The second PC is the direction with the next highest variance

that is orthogonal to the first PC's direction [22]. When PCs are used in classification, each data instance can have up to the number of PCs as there are original features but the more PCs you incorporate accounts for decreasingly less variability [39].

For this research, the top 10 PCs were incorporated as features in both classifiers to observe how it affects classification accuracy. For every data instance, 10 values were added as a feature indicating how much the particular data instance impacts the variation on a particular PC. These numerical values are referred to as loadings, which are a measurement of the correlation between the original features and the derived PCs [32]. After feature set generation for the Claim Detection Model and the log transformation of appropriate features is performed, PCA was performed on the 40 base features. In a similar fashion, PCA was performed on the 40 base features and the resulting claim classifications from the Claim Detection Model. The top 10 PCs were added to the Misinformation Detection Model's feature set. Incorporating PCs in conjunction with existing features could highlight patterns in the data that would have been overlooked because each PC offers a unique expression of the data set where a particular data instance receives a score based on how its feature values impact the data variance of the entire data set represented by the 10 PCs [22].

Chapter 7

Claim Detection

Before misinformation detection occurs, a model for predicting whether or not a Tweet is making a definitive claim or not is applied to the Zika dataset. This prediction is later used as a feature in the Misinformation Detection Model (MDM). The purpose of determining if a Tweet is making a claim or not derives from the idea that misinformation will present as a statement of fact. Tweets that question a topic, are unsure about a topic, or our providing an opinion on a topic are not considered as making a claim. For this research, a Tweet is making a claim if it presents a statement as the truth. The idea is that a Tweet making a definitive claim should have a higher chance of containing misinformation than a Tweet that is not making a definitive claim.

7.1 Manual Annotation

In order to create the training and testing data for the Claim Detection Model (CDM), a randomly selected subset of 1800 Tweets were selected from the 996,443 Zika related Tweets. Through a process of manual annotation, we labeled each of the 1800 Tweets as making a claim or not making a claim. The criteria for labeling a Tweet as making claim is if the author of the Tweet presented what they were Tweeting as a statement of fact. An example of a Tweet labeled as making a claim is shown in Figure 5.

Regardless of if this statement is true or not, the author of the Tweet is still making a claim about the mosquitoes that cause Zika. The 1800 manually annotated Tweets were split into training and testing data where 75% or 1350 Tweets were reserved for training and 15% or 450 were reserved for testing.

Figure 5 – Example Tweet manually labeled "isClaim" [1]

7.2 Feature Set

The feature set for the CDM contains 40 base features that are grouped as being user-related, Tweet-related, propagation-related, and descriptive. For a full description of

the 40 features and why they were used, see Chapter 6 of this thesis. After the relevant features were log-transformed, Principal Component Analysis (PCA) was performed on the 40 features and the top 10 Principal Components (PCs) were used to create 10 additional features. For a description on how and why PCA-based features were incorporated, refer to Section 6.6 of this paper. In total, 50 features were used to construct the Claim Detection Model's feature set; these features are listed in Table 8.

Table 8 - Complete list of features used in the Claim Detection Model's feature set

Claim Detection Model's Feature Set				
accountAge	hasNumber	isAssociation	isPreventive	PC1
locationListed	hasQuestionMark	numAssociation	numPreventive	PC2
numFollowers	hasURL	isCase	isReinforced	PC3
numFollowing	numHashtags	numCase	numReinforced	PC4
followRatio	timespan	isDiseaseSpecific	isTranmission	PC5
numTweetsPosted	tweetLength	numDiseaseSpecific	numTransmission	PC6
hasCountry	isFavorited	isNewReport	isUncertain	PC7
hasExclamation	numFavorites	numNewReport	numUncertain	PC8
hasHashtags	isRetweet	isPolitical	isZikaSpecific	PC9
hasMultimedia	numRetweets	numPolitical	numZikaSpecific	PC10

7.3 Model

Using R's GBM package, a Generalized Boosted Regression Model (GBM) was created to function as the Claim Detection Model using the training data derived from the manual annotation and feature generation described in the previous sections. The training data consists of the 50 features listed in Table 8 and the manually annotated feature signifying if a particular Tweet was making a claim or not denoted as "isClaim". The reasoning behind why the GBM package was chosen for the modeling in this research and how it functions is explained in Section 4.2.4 of this paper.

During creation of the model, the "isClaim" feature is used as the dependent variable and the maximum number of trees to add to the ensemble was initially set to 5000. After creation, the loss function was minimized after adding 846 trees. Accordingly, the model was recreated with the maximum number of trees set to 846. Since the data follows a Bernoulli distribution, the Bernoulli loss function was selected during the generation of the model. For more details on the loss function refer to Section 4.2.4 of this paper. The final classifications of the model are continuous predictions ranging from 0 to 1. These were rounded and a value of 1 indicates a positive "isClaim" label and a value of 0 indicates a negative "isClaim" label [44].

7.4 Evaluation

For the sake of optimizing predictions and understanding the importance of log transformation and the impact of including Principal Components (PCs) as features,

three different versions of the Claim Detection Model were utilized for evaluation. The first version referred to as Model A does not log transform any of the features exhibiting lognormal distribution and includes PCs as features. The second version referred to as Model B log transforms the appropriate features and includes PCs as features. The third version referred to as Model C log transforms the appropriate features but does not include the top 10 PCs as features. Models A and B are compared to evaluate the usefulness of log transformation. Models B and C are compared to evaluate the usefulness of incorporating PCs as features.

7.4.1 Metrics

For the Feature Relevancy analysis, the relative influence values generated from the GBM package are used as feature relevancy scores indicating the amount of influence a particular feature has on classification. For the Testing Set Validation, confusion matrices are created and bar graphs are used to display performance statistics. For Cross Fold Validation results, a line graph is used to display the individual area under the ROC curve (AUC) values of the different models.

The confusion matrices used in the evaluation show how well a particular model performed on the testing data. The matrices display four values that are used to compute the performance statistics shown in the bar graphs. The four values are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy of a particular model is determined by the following formula: $\frac{(TP+TN)}{n}$ where n is the size of

the data set. In a similar fashion, error rate is determined by $\frac{(FP+FN)}{n}$ and used to find out how often the model is incorrect. Sensitivity, which is how often the model predicts 1 when the value is actually 1, is determined by $\frac{TP}{num\ of\ actual\ 1s}$. Specificity is determined by $\frac{TN}{num\ of\ actual\ 0s}$ and is how often the model predicts 0 when the actual value is 0. Precision is determined by $\frac{TP}{num\ of\ predicted\ 1s}$ and indicates how often the model correctly predicts an actual value of 1 [50].

The line graphs used in the evaluation section display the Area Under the Curve (AUC) values for a particular model. AUC is calculated using an ROC curve that plots the Sensitivity, which is the True Positive Rate, against the False Positive Rate, which is how often the model predicts 1 when the actual value is 0 and determined by $\frac{FP}{num\ of\ actual\ 0s}$. These ratings range from 0 to 1. An AUC value is calculated by obtaining the percentage of the values under a ROC curve and is used to gauge the performance of a model [47].

7.4.2 Feature Relevancy

In order to gauge the influence of individual features in the three versions of the Claim Detection Model, a feature relevancy score was generated from the GBM package for every feature in the feature set of each model. The top 30 feature relevancy scores for the three versions of the model are shown in Table 9, 10, 11 respectively. The scores are

based on the relative influence values placed on the feature by the GBM package. A high relevancy score indicates that a particular feature had high influence on determining whether or not a Tweet is making a claim or not.

Recall Model A introduced in section 7.4 where the lognormal features are not log transformed and Principal Components (PCs) are included as features. In Table 9 it is apparent that the included PC features are the most influential features in Model A. Since the first PC has the highest relevancy score and represents the most variance in the data set, it demonstrates that the score relating to how a particular instance of a feature impacts the variance of the dataset is useful for discerning if a Tweet was making a claim or not. Since Model A is performing claim detection, it is understandable that the descriptive feature "isNewReport" would indicate a claim since a new report would present facts. Since Model A does not use log transformation, it is interesting that the features exhibiting lognormal distribution, such as numTweetsPosted, numFollowing, and followRatio, had a high influence on classification.

In contrast to Model A, Model B does log transform the lognormal variables. The purpose of Model B is to gauge the impact of log transformation since PCs are used as features in both. Table 10 illustrates the most influential features in Model B and shows that Model B follows a similar pattern as Model A with PCs being the most influential features. It is important to note that after log transformation was performed on the features, the transformed features became more influential when comparing Model A without log transformation against Model B with log transformation. For example, the "numTweetsPosted" feature received a relevancy score of 1.894 before log transformation

and 2.357 after log transformation. From observing the changes from Table 9 to Table 10 detailed above, it is observable that log transforming relevant features increases their influence.

In contrast to Model B, Model C does not include PCs as features. The purpose of Model C is to gauge the impact of incorporating PCs as features. Table 11 illustrates the most influential features in Model C. When comparing Model B to Model C using Table 10 and Table 11, the absence of PCs as features results in a higher distribution of influence amongst the features with descriptive features and user related features being the most influential.

Table 9 – Feature relevancy scores for Claim Detection Model A

Feature Relevancy Scores for Claim Detection Model A (Not Using Log Transformation and Including PCs)			
Feature Name	*Relevancy Score*	*Feature Name*	*Relevancy Score*
PC 1	30.656	**isDiseaseSpecific**	1.012
PC 5	21.516	**numFollows**	0.774
PC 6	14.723	**timespan**	0.668
PC 2	3.829	**isUncertain**	.636
PC 7	3.336	**accountAge**	0.561
PC 9	3.212	**numRetweets**	0.543
isNewReport	2.077	**isCase**	0.470
numTweetsPosted	1.894	**hasHastags**	0.443
PC 4	1.721	**numHashtags**	0.440
PC 3	1.707	**isAssociation**	0.351
PC 8	1.652	**isPrevention**	0.350
numFollowing	1.644	**isPolitical**	0.267
PC 10	1.622	**hasNumber**	0.211
tweetLength	1.550	**hasURL**	0.177
followRatio	1.045	**isFavorited**	0.166

Table 10 – Feature relevancy scores for Claim Detection Model B

Feature Relevancy Scores for Claim Detection Model B (Using Log Transformation and Including PCs)			
Feature Name	*Relevancy Score*	*Feature Name*	*Relevancy Score*
PC 1	25.403	**followRatio**	0.745
PC 7	24.879	**isTransmission**	0.719
PC 6	13.587	**isDiseaseSpecific**	0.693
PC 5	5.138	**isCase**	0.528
PC 3	4.778	**hasURL**	0.504
PC 10	4.734	**hasNumber**	0.475
numTweetsPosted	2.357	**isAssociation**	0.417
PC 8	2.221	**accountAge**	0.410
isNewReport	1.955	**timespan**	0.358
tweetLength	1.686	**timespan**	0.355
numFollowing	1.627	**isPreventive**	0.345
PC 9	1.625	**isFavorited**	0.266
PC 2	1.008	**hasHashtags**	0.226
PC 4	0.887	**isUncertain**	0.189
numFollowers	0.827	**numRetweets**	0.154

Table 11 – Feature relevancy scores for Claim Detection Model C

Feature Relevancy Scores for Claim Detection Model C (Using Log Transformation and Excluding PCs)			
Feature Name	*Relevancy Score*	*Feature Name*	*Relevancy Score*
numTweetsPosted	11.709	**isPolitical**	1.710
isCase	10.612	**isUncertain**	1.670
isAssociation	8.310	**isPreventive**	1.158
numFollowing	7.979	**hasQuestionMark**	1.018
followRatio	7.716	**numHashtags**	0.943
numFollowers	6.903	**isFavorited**	0.903
tweetLength	4.674	**isDiseaseSpecific**	0.773
accountAge	3.561	**hasMultimedia**	0.673
timespan	3.390	**hasNumber**	0.606
isTransmission	3.253	**hasURL**	0.588
isNewReport	2.662	**locationListed**	0.580
hasCountry	2.494	**hasHastags**	0.421
numRetweets	2.486	**isRetweet**	0.402
isZikaSpecific	2.189	**numPreventive**	0.384
numCase	2.025	**isReinforced**	0.159

7.4.3 Testing Set Validation

Recall the 1800 manually annotated Tweets that were split into training and testing data where 75% or 1350 Tweets were reserved for training and 15% or 450 were reserved for testing. In order to further evaluate the accuracy of the three versions of the Claim Detection Model on unknown data, the models were applied to the testing data set. As detailed in section 7.4.1, confusion matrices are used as a metric to evaluate how well a model performs on testing data. The four values provided by a confusion matrix are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For a description of each value and how it is derived, refer to section 7.4.1.

Table 12 shows the confusion matrices for Model A and Model B. The differences in these matrices highlight the performance impact of not using and using log transformation on relevant features. It is apparent that log transforming lowered the occurrence of FPs from 18 to 10, which is an important improvement for better predicting Tweets that are not making a claim. Additionally, log transformation increased the number of TPs, which indicates an improvement in the identification of a Tweet containing a claim. Since the TN and FN values remained virtually the same while the other rates improved, it is evident that log transforming relevant features increases prediction accuracy.

The accuracy, sensitivity, specificity, and precision statistics displayed in Figure 6 are derived from the confusion matrix values from Table 12. A description of these performance statistics and how they are derived is detailed in Section 7.4.1. From Figure 6, it is evident that Model B exhibits a lower sensitivity than Model A indicating that

Model B did not perform as well as Model A when comparing the number of correctly identified Tweets making a claim versus the number of Tweets that were actually making a claim. Despite the lower sensitivity, Model B exhibits higher accuracy, specificity, and precision than Model A. The higher specificity value indicates that log transformed features are better at predicting if a Tweet is not making a claim when compared to the actual number of Tweets not making a claim. The higher precision value indicates that log transformed features are better at predicting if a Tweet is making a claim when compared to the number of Tweets predicted as making a claim. This further backs the claim that using log transformation increases predictive performance.

Table 13 shows the confusion matrices for Model B and Model C. The differences in these matrices highlight the performance impact of including and excluding PCs as features. Recall Model B includes PCs and Model C does not. From Table 13, it is apparent that excluding the top 10 PCs drastically reduced performance in every category. Excluding the PCs lowered the occurrence of TNs and TPs and increased the occurrence of FNs and FPs. This type of performance impact highlights the potential of including PCs as features in classification. The performance statistics displayed in Figure 7 further highlight the benefit of incorporating PCs since Model B, which includes PCs as features, exhibits higher accuracy, sensitivity, specificity, and precision than Model C. which does not include PCs as features.

Table 12 – Confusion Matrices comparing Model A against Model B on the testing data

Model A (Not using Log Transformation)			**Model B** (Using Log Transformation)		
n = 450	Predicted: 0	Predicted: 1	n = 450	Predicted: 0	Predicted: 1
Actual: 0	TN = 210	FP = 18	Actual: 0	TN = 209	FP = 10
Actual: 1	FN = 15	TP = 207	Actual: 1	FN = 16	TP = 215

Table 13 – Confusion Matrices comparing Model B against Model C on the testing data

Model B (Including PCs)			**Model C** (Excluding PCs)		
n = 450	Predicted: 0	Predicted: 1	n = 450	Predicted: 0	Predicted: 1
Actual: 0	TN = 209	FP = 10	Actual: 0	TN = 187	FP = 71
Actual: 1	FN = 16	TP = 215	Actual: 1	FN = 38	TP = 154

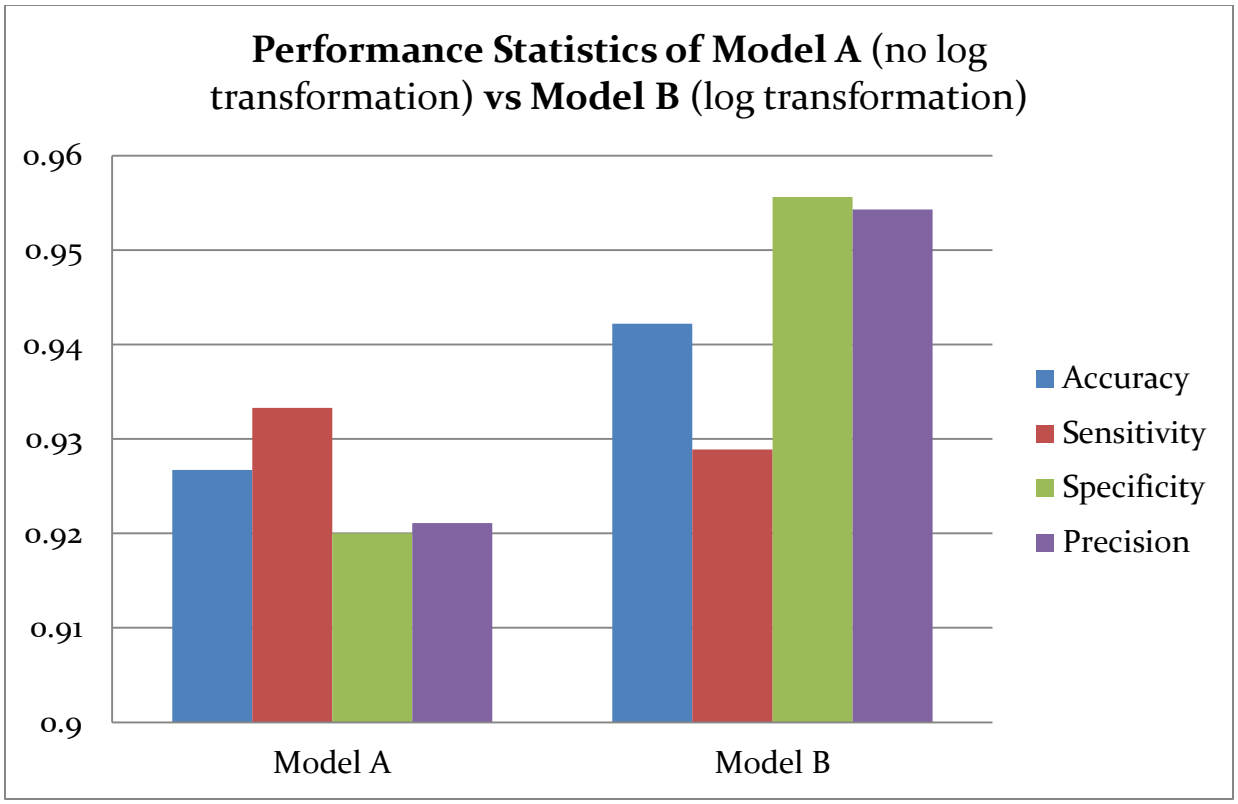

Figure 6 – Performance statistics comparing Model A (with no log transformation) against Model B (with log transformation) on the testing data

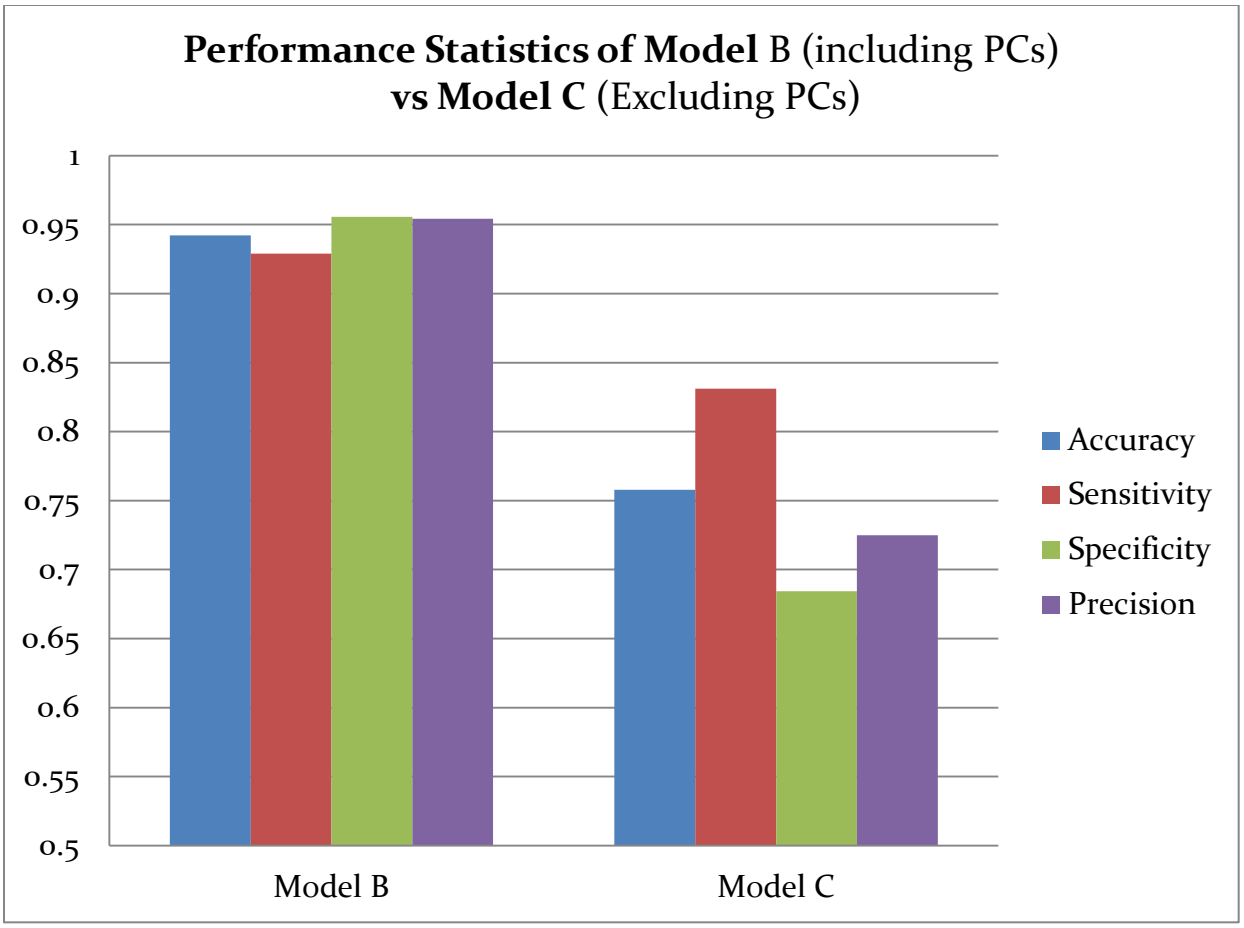

Figure 7 – Performance statistics comparing Model B (including PCs) against Model C (excluding PCs) when applied to testing data

After comparing the confusion matrices and performance statistics shown in the previous tables and figures, it is apparent that Model B, which used log transformation and included PCs, performed the best out of the three versions of the Claim Detection Model during Testing Validation. The success of Model B shows that using a feature set with approximately normal distribution among numerical features improves accuracy and performance. Additionally, Model B demonstrated the potential of incorporating features

from the existing base features by scoring them based on how much they impact data variance.

7.4.4 Cross Validation

In order to evaluate the previously explained three versions of the Claim Detection Model on more generalized data, a technique called K-Fold Cross Validation (KCV) was utilized. KCV analyzes the performance of the models by dividing the training dataset into k equal sized partitions. Every partition is separately used as testing data within the CV process. The process is repeated k times with a different partition being used to train a model during each iteration. The partitions that are not being used as testing data during an iteration are used as training data. The advantage of this procedure is that unlike the Testing Set Validation procedure utilized in the previous section, KCV allows for every piece of data to be used and evaluated as testing data [13]. For this research, 10 folds are used.

Individual AUC values for all 10 folds of Cross Validation performed on Model A and B are shown in Figure 8. This comparison highlights the performance impact of log transforming relevant features for claim detection. Excluding the virtually equal AUC value that occurred during fold 7, Model B outperformed Model A in every fold of Cross Validation. This indicates that log transforming relevant features increases the predictive performance across generalized unknown data.

Individual AUC values for all 10 folds of Cross Validation performed on Model B and C are shown in Figure 9. This comparison highlights the performance impact of

including and excluding PCs as features for claim detection. Model B drastically outperformed Model C in every fold of Cross Validation with an average AUC value of 0.9543 compared to Model C's average AUC value of 0.8105. This emphasizes the potential value of including PCs as features for claim detection.

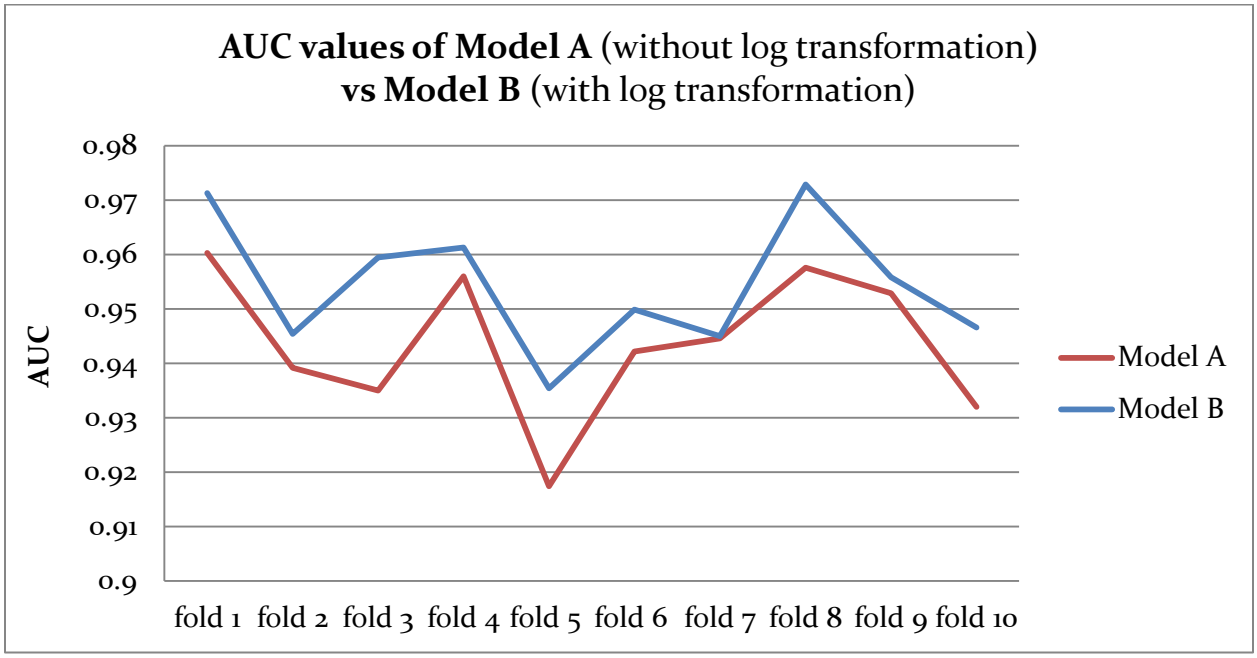

Figure 8 – Resulting AUC values from performing 10-Fold Cross Validation on Model A and Model B

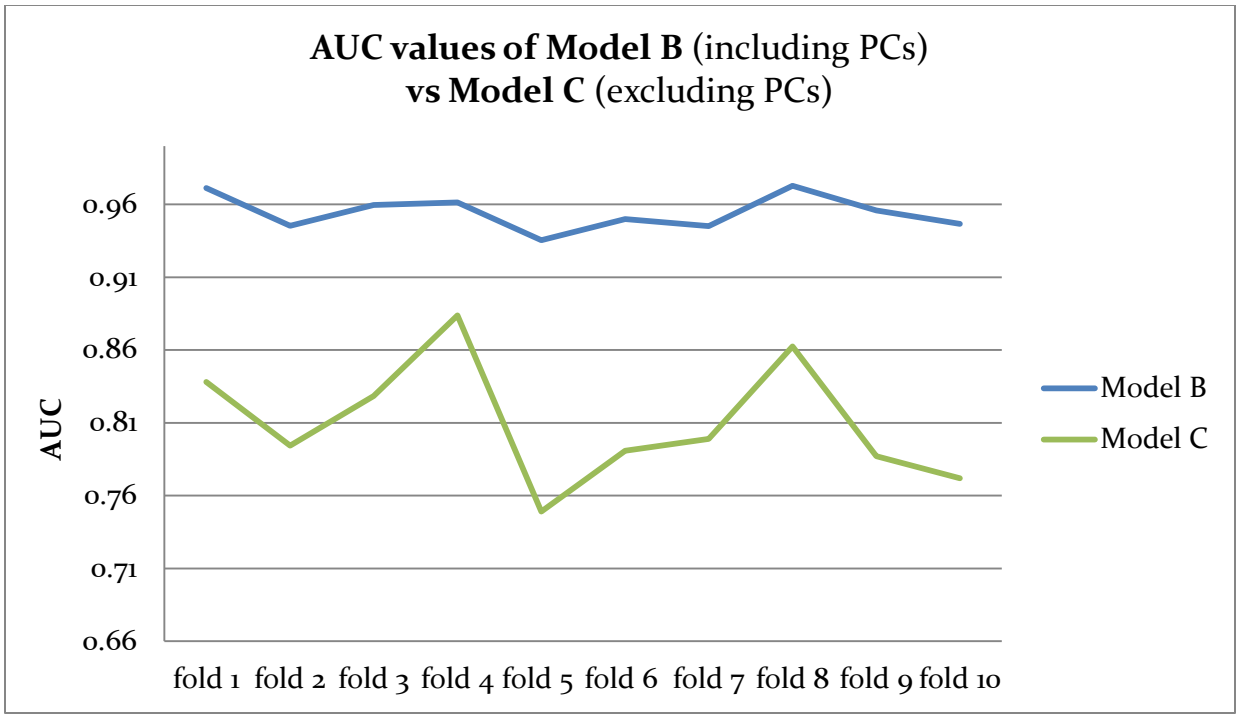

Figure 9 – Resulting AUC values from performing 10-Fold Cross Validation on Model A and Model B

7.5 Application

As a result of the performance shown by Model B during evaluation, the final version of the Claim Detection Model used log transformation on the appropriate features and included the top 10 PCs as features. This version of the Claim Detection Model had an accuracy of 0.9422 during Testing Validation and generated an average AUC value of 0.9543 during 10-fold Cross Validation. The selected model was applied to the entire dataset of 996,443 Zika related Tweets. The model predicted that 624,568 Tweets (62.68% of dataset) were making a claim and 371,875 Tweets (37.32% of dataset) were not making a claim. Examples of detected Tweets making a claim are presented in Table 14. Examples

of detected Tweets not making a claim are presented in Table 15. The resulting predictions were recorded and used as features in the Misinformation Detection Model detailed in Chapter 8.

The previously detailed success of Model B provided the evidence needed to utilize log transformation in the final Claim Detection Model and Misinformation Detection Model by showing that using a feature set with approximately normal distribution among numerical features improved accuracy and performance. Additionally, the successful incorporation of Principal Components (PCs) provided evidence of the significance and potential of incorporating features from the existing base features by scoring them based on how much they impact data variance. Combined with the success of the descriptive and PC features previously detailed, the Claim Detection Model demonstrates that Tweets can be accurately classified as making a claim or not through the incorporation of unique features detailed in Section 7.2 and the use of the gradient boosting technique detailed in Section 4.2.4.

Table 14 – Examples of Tweets Making a Claim

TweetID	*Example Tweet*
709513194134437889	"scientists find first links between zika and temporary paralysis"
710081073237995520	"cuba reports first case of zika virus in the country"
715605652819673088	"colombia reports 32 cases of zikalinked birth defects"
719401704345591810	"zika virus is now linked to adem"
721790646625837057	"peru reports first case of sexually transmitted zika virus"
702387470911086593	"cdc reports 14 new sexually transmitted zika cases"
722257626432712704	"zika mosquito found in chile first time in decades"
700125308104667136	"zika virus 100% man-made in a lab !!"
702241927811436544	"#alaska the mosquito that started today's zika outbreak in brazil was killed off"
702242594949046273	"venezuela won't talk to colombia about zika"
722362623635955712	"doctors link monsanto pesticides not zika to birth defects"
704720964215554049	"first case of zika virus identified in philadelphia resident"
699985933894963200	"the first known sexual transmission of #zika virus in the u.s. was in 2008"
708435207431520256	"zika virus is now causing serious brain infections in adults"
700127753702633472	"microcephaly cases up 10% in brazil amid #zika scare"
702545017714290688	"zika virus is man-made because it's only mosquito designed to fly during daytime."

Table 15 – Examples of Tweets Not Making a Claim

TweetID	*Example Tweet*
717055300877074432	"will angolas yellow fever outbreak be another zika?"
720470642898898948	"meet the scientists hunting zikacarrying mosquitoes"
711640970115162112	"we should have a well funded response team in place to combat outbreaks like this cdc fundingresearch is too thin"
719654926121435136	"lets use lasers to fight the zika virus"
721373658636103680	"waiting until mosquito seasons over typical gop stupidity"
720301717925933056	"7 common places mosquitoes breed in your yard help avoid zika virus"
720398203837763587	"can we take bets on how long it'll take congress to fund zika prevention treatment?"
719654096794308608	"what do you mean we dont have enough money to research zika virus?"
708442011813044226	"12 mosquito bites what are the chances i have zika virus?"
704191180108636160	"zika virus: a little less speculation, a little more action"
709807009105772545	"spring break and zika how to protect yourself from mosquitoes"
722369614198353920	"eradication of the mosquito spreading the zika virus in brazil unlikely"
705407071869652992	"by putting out a public statement with the words i believe you do nothing but cause unwarranted mass hysteria"
711541995672621056	"will the zika outbreak ease abortion laws?"
705876200099414016	"scientists may have finally solved zikas scariest mystery "
700566613100113921	"can scientists prove zika virus is causing birth defects?"

Chapter 8

Misinformation Detection

Once the Claim Detection Model (CDM) is applied to the dataset, the resulting claim classifications are used to chain the CDM to the Misinformation Detection Model (MDM). This chained approach is accomplished by incorporating the claim classifications as features for the MDM. The purpose of the MDM is to detect misinformation by classifying a Tweet as either misinformation or not misinformation. The MDM is trained on manually annotated data comprised of Tweets that are either confirmed credible or confirmed misinformation. We manually annotated a subset of Tweets as either confirmed credible or confirmed misinformation based on verified by reports from reputable health agencies. The manual annotation process is detailed in Section 8.1 below.

8.1 Manual Annotation

In order to appropriately represent instances of Tweets containing misinformation, it is important to also consider Tweets that contain credible information. Therefore, the training and testing data for the Misinformation Detection Model consists of a subset of 3600 Tweets selected from the 996,443 Zika related Tweets containing confirmed credible information and confirmed misinformation. In order to minimize sample selection bias by over representing positive labels of confirmed misinformation in the dataset,

confirmed credible Tweets were also identified to function as a negative label. Sample selection bias was further reduced by proportionally stratifying the subset of 3600 Tweets by confirmed misinformation or confirmed credible information [56]. The resulting training and testing data consists of 1800 positive labels of misinformation representing confirmed misinformation and 1800 negative labels of misinformation representing confirmed credible information. The training data consists of 2700 Tweets where one half of the Tweets are positive labels of misinformation and the other half is negative labels of misinformation. In similar fashion, the testing data consists of 900 Tweets where one half of the Tweets are positive labels of misinformation and the other half is negative labels of misinformation.

8.1.1 Confirmed Misinformation

The 1800 positive labels of misinformation were selected by discovering different subjects of misinformation in the dataset through manual data analysis and verified by reports from reputable health agencies such as the Centers for Disease Control and Prevention (CDC). We performed manual data analysis on the contents of the Zika data set by searching for Tweets that included information contrary to verified information from reputable health agencies. Based on this analysis, certain categories of misinformation were apparent. Consider the following information publically available from various health agencies. Excluding basic symptoms, Zika virus infection is only known to cause Guillain-Barré Syndrome and birth defects, including microcephaly, decreased brain tissue, eye damage, limited range of motion, and muscle restrictions

according to the CDC [60]. Even though there are recorded deaths related to complications from Zika virus infection (New England Journal of Medicine, 2016), Zika virus infection rarely causes severe reactions or death [16] [3]. The CDC also reports that the Zika virus can be transmitted from human to human contact through sex, blood transfusion, or from mother to child during pregnancy [61].

Despite this publically available information, the data analysis revealed Tweets in the dataset claiming that Zika is extremely fatal, not transmittable from human to human contact, man-made by the government for population control, created by pharmaceutical companies to sell vaccines, and causes Autism. For a full list of the identified subjects of misinformation and an example of each, see Table 16. Using 115 phrases associated with the subjects in Table 16, 1800 positive labels of misinformation were selected from the dataset of 996,443 Zika Tweets. A threshold was used to prevent individual subjects from being overly represented during manual annotation.

Table 16 - Complete list of subjects used as criteria for manually annotating confirmed misinformation with example Tweets given for each subject

Misinformation Subjects

Subject	*Description*	*Example Tweet*
Autism	Zika virus infection causes autism	"The Zika virus linked to the development of autism, bipolar disorder and schizophrenia."
Fatal	Zika virus infection is extremely fatal	"It's being reported that Zika virus is being transmitted through sexual intercourse and more deadly than Ebola."
GMO	Zika virus caused and/or spread by genetically modified mosquitos	"Zika virus was caused by genetically modified mosquitoes and now labeled a global issue."
Man-made	Zika virus is a man-made virus	"There's no such thing as Zika, its a man made virus designed to cause panic amongst the most unfortunate"
Pesticide	Symptoms associated with Zika virus infection are actually caused by pesticide/larvicide/herbicide	"Larvicide Manufactured By Sumitomo, Not Zika Virus, True Cause Of Brazil's Microcephaly Outbreak"
Pharmaceutical	Pharmaceutical companies are covering up side effects using the Zika virus infection	"Tdap vaccinations causes microcephaly...BigMedia stuck on Zika virus as the cause due to relationship with BigPharma"
Pollution	Symptoms associated with Zika virus infection are actually caused by pollution	"Leaked report: petrochemical pollution causes microcephaly. Not Zika."
Transmission	Zika virus is not transmitted from human to human	"The Health Dept says the 1st case of Zika in SA poses no risks to residents as the virus is not transmitted from human to human @Faizie143"
Vaccine	Symptoms associated with Zika virus infection are actually caused by vaccines	"@AJEnglish VACCINES CAUSED THIS NOT THE ZIKA VIRUS."

8.1.2 Confirmed Credible Information

In a similar fashion to the positive labels of misinformation, the 1800 negative labels of misinformation were selected by finding examples of credible information in the dataset through manual data analysis and verified using public reports from the CDC, World Health Organization, and Dick, Kitchen, and Haddow's paper on the isolation of the Zika virus [59] [60] [61] [62] [63] [20]. The subjects include background information regarding the Zika virus, how it is transmitted, symptoms, health risks, and prevention techniques. For examples of each subject of credible information, see Table 17. Using 30 phrases associated with the subjects in Table 17, 1800 negative labels of misinformation were selected from the dataset of 996,443 Zika Tweets. A threshold was used to prevent individual subjects from being overly represented during manual annotation.

Table 17 - Complete list of subjects used as criteria for manually annotating confirmed credible information with example Tweets given for each subject

Credible Information Subjects

Subject	*Description*	*Example Tweet*
Health Risks	Health risks and effects of Zika Virus infection	"Zika virus confirmed to cause microcephaly"
History	Origin and history of the Zika virus	"Zika was first discovered in 1947"
Prevention Techniques	Techniques backed by health agencies to prevent exposure and infection of the Zika virus	"Best way to prevent Zika is to avoid mosquito bites: wear long clothing, use insect repellents, stay/sleep in screened rooms #ReutersZika"
Symptoms	Symptoms associated with Zika virus infections	"Symptoms of Zika virus: Most cases cause no symptoms.Those who do develop symptoms:mild headaches, fever, rash, conjunctivitis & joint pain."
Transmission	How the Zika virus is transmitted	"Zika can be transmitted sexually "

8.2 Feature Set

In similar fashion to the CDM, the MDM also incorporates the 40 user-related, Tweet-related, propagation-related, and descriptive base features and uses appropriate log transformation when necessary. Additionally, the resulting claim classifications generated from applying the CDM to the dataset were used as features in the MDM. The claim classifications indicate if a Tweet is making a claim or not and are stored in a feature called "isClaim". Principal Component Analysis (PCA) was performed on the 41 features and the top 10 Principal Components (PCs) were used to create 10 additional

features for the MDM. For a description on how and why PCA-based features were incorporated, refer to Section 6.6 of this thesis. In total, 51 features were used to construct the MDM's feature set; these features are listed in Table 18.

Table 18 - Complete list of features used in the Misinformation Detection Model's feature set

Misinformation Detection Model's Feature Set				
accountAge	hasNumber	isAssociation	isPreventive	isClaim
locationListed	hasQuestionMark	numAssociation	numPreventive	PC1
numFollowers	hasURL	isCase	isReinforced	PC2
numFollowing	numHashtags	numCase	numReinforced	PC3
followRatio	timespan	isDiseaseSpecific	isTranmission	PC4
numTweetsPosted	tweetLength	numDiseaseSpecific	numTransmission	PC5
hasCountry	isFavorited	isNewReport	isUncertain	PC6
hasExclamation	numFavorites	numNewReport	numUncertain	PC7
hasHashtags	isRetweet	isPolitical	isZikaSpecific	PC8
hasMultimedia	numRetweets	numPolitical	numZikaSpecific	PC9
				PC10

8.3 Model

The Misinformation Detection Model is a gradient boosted model created using R's GBM package, which is detailed in Section 4.2.4 of this thesis. The training and testing

data for the model is formed using the 3600 resulting Tweets from the manual annotation process detailed in Section 8.1. The training data set consists of 75% of the manually annotated Tweets with 1350 Tweets positively labeled as misinformation and 1350 Tweets negatively labeled as misinformation. The testing data set consists of the remaining 25% of the manually annotated Tweets with 450 Tweets positively labeled as misinformation and 450 Tweets negatively labeled as misinformation. The manually annotated labels of misinformation are denoted as "isMisinformation" and function as the dependent variable during model creation.

During creation of the model, the manually annotated labels of "isMisinformation" were used as the dependent variable and the maximum number of trees to add to the ensemble was initially set to 5000. After creation, the model was able to minimize the loss function using 697 trees so the model was recreated limiting the number of trees added to the ensemble to 697 trees. For the same reason discussed in Section 7.3, the model was generated with a Bernoulli loss function and the resulting continuous predictions were rounded where a value of 1 indicates a positive "isMisinformation" prediction and a value of 0 indicates a negative "isMisinformation" prediction [44].

8.4 Evaluation

Using the knowledge gained from the evaluation of the Claim Detection Model in Section 7.4, a single model was used for the evaluation of the Misinformation Detection Model. The model log transformed the features with lognormal distribution and included the top 10 PCs as features in the feature set.

8.4.1 Metrics

In a similar fashion as the evaluation of the Claim Detection Model, relative influence values generated from the GBM package were used as feature relevancy scores to measure the amount of influence each feature has on Misinformation Detection. In order to gauge the accuracy of detection on unknown data, a confusion matrix is created during Testing Set Validation and a bar graph is used to compare performance statistics against a baseline approach detailed in Section 8.4.3. In order to gauge the accuracy of detection on a more generalized set of unknown data, AUC values resulting from Cross Fold Validation are displayed in a line graph. For more detail on the purpose and meaning of these evaluation metrics, refer to Section 7.4.1.

8.4.2 Feature Relevancy

In order to gauge the influence of individual features on misinformation detection, feature relevancy scores were generated for every feature in the feature set. Table 19 shows the top 30 feature relevancy scores of the Misinformation Detection Model's feature set. The top 10 features account for 84% of the total relative influence in the feature set.

PC 1 is the feature with the highest relevancy score at 22.179%. This is reasonable due to the fact that the first PC accounts for the highest variance in the data set and consequently, is potentially the best feature to summarize the data [22]. The included 10 PCs account for approximately 45% of the entire feature set's influence. Since the PC features symbolize variance amongst the data set, it is difficult to perceive what patterns

in the data that the individual PCs actually represent. Regardless, the incorporation of the top 10 PCs as features greatly influences how the model detects misinformation.

The second highest relevancy score at 16.226% belongs to the descriptive feature "isReinforced" and shows that the incorporation of string variables such as "CDC", "experts", and "officials" are good indicators of whether or not a Tweet contains misinformation. The following descriptive features proved noteworthy as well with "hasCountry" at 8.643% relevancy, "numAssociation" at 7.521% relevancy, and "isNewReport" at 4.950% relevancy. The impact of these features shows that mentioning a country in a Tweet, frequently using words related to Zika associations, and Tweeting about new Zika-related reports are worthy indicators for detecting misinformation.

Unfortunately, the "isClaim" feature from the Claim Detection Model did not have a high enough relative influence to make the list of top 30 features relevancy scores. This shows that the features produced by the Claim Detection Model and used in the Misinformation Detection Model did not have a significant impact on classifying misinformation. Even though the Claim Detection Model and the Misinformation Detection Model were separately successful in their goals, the chained approach did not prove particularly useful when compared to the other features in the Misinformation Detection Model's feature set such as the Principal Components (PCs) and descriptive features. This finding could result from misinformation presenting as uncertainty, a question, or an opinion because the Claim Detection Model only classified a Tweet as making a claim if it was definitively stating a fact.

Table 19 – Feature relevancy scores for every feature in the Misinformation Detection Model's feature set

Feature Relevancy Scores of the Misinformation Detection Model			
Feature Name	*Relevancy Score*	*Feature Name*	*Relevancy Score*
PC 1	22.179	**PC 9**	0.976
isReinforced	16.226	**numRetweets**	0.927
hasCountry	8.643	**numTweetsPosted**	0.861
PC 4	7.911	**PC 7**	0.830
numAssociation	7.521	**PC 10**	0.784
tweetLength	5.507	**numFollowers**	0.745
PC 3	4.964	**numFollowers**	0.638
isNewReport	4.950	**timespan**	0.351
isAssociation	3.830	**accountAge**	0.347
PC 6	2.472	**isZikaSpecific**	0.286
PC 2	2.121	**isDiseaseSpecific**	0.278
isCase	2.001	**hasHastags**	0.215
PC 5	1.323	**numHashtags**	0.149
PC 8	1.317	**hasUrl**	0.091
followRatio	1.202	**locationListed**	0.072

8.4.3 Testing Set Validation

Recall the 3600 manually annotated Tweets that were split into training and testing data where 75% of the Tweets were reserved for training data and the remaining 25% were reserved for the testing data. The Tweets were stratified in way that ensured that each dataset contained 50% positive labels of misinformation and 50% negative labels of misinformation. In order to validate the accuracy of the Misinformation Detection Model on unknown data, the model was applied to the testing data set of 900 Tweets. Table 20 shows the confusion matrix of the resulting predictions. The confusion matrix indicates that the model produced 17 false negatives and 10 false positives out of the 900 classifications resulting with an overall error rate of 2.89% on the testing data. 421 Tweets were correctly predicted as not containing misinformation and 453 were correctly predicted as containing misinformation resulting in an overall accuracy of 97.11% on the testing data.

Table 20 – Confusion matrix resulting from applying the Misinformation Detection Model to the testing data

Misinformation Detection Model (Applied on the Testing Set)		
n = 900	Predicted: 0	Predicted: 1
Actual: 0	TN = 421	FP = 9
Actual: 1	FN = 17	TP = 453

In order to compare the resulting performance statistics to another classifier, the prediction statistics detailed in Table 20 were compared against a classifier that used a non-boosted decision tree approach to identify rumors. Recall Liang's recent study on identifying rumors on the Chinese microblogging service known as Sina Weibo [29]. Like Twitter, Sina Weibo allows its users to use 140 characters for their posts and incorporates the basic concepts of following, reposting, hashtags, and multimedia functionality [19]. Liang's study was recently conducted and focused on identifying rumors, thus it was chosen as a base approach for comparison in this research. Specifically, the decision tree model tested in their research was used as the base approach because it produced their best results [29]. Since Liang's study only provides precision, sensitivity, and F1 score for their approach, the compared performance statistics are precision, sensitivity, and F1 score. F1 score is an accuracy score that considers precision (P) and sensitivity (S) by calculating with the formula $2\frac{PS}{P+S}$ [33].

The sensitivity of the Misinformation Detection Model signifying how often the model detected misinformation in a Tweet when there actually was misinformation in the Tweet was 0.9612. The specificity of the model signifying how often the model detected no misinformation when there actually was no misinformation was 0.9805. The precision of the model signifying how often the model correctly detected misinformation in a Tweet was 0.9791 [50]. The F1 score was 0.9700 and represents the accuracy of the model by considering the weighted average of precision and sensitivity [33]. The performance comparison of the Misinformation Detection Model from this research against the baseline approach is shown in Figure 10. From the graph, it is apparent that the

Misinformation Detection Model in this research performs more accurately than the baseline approach.

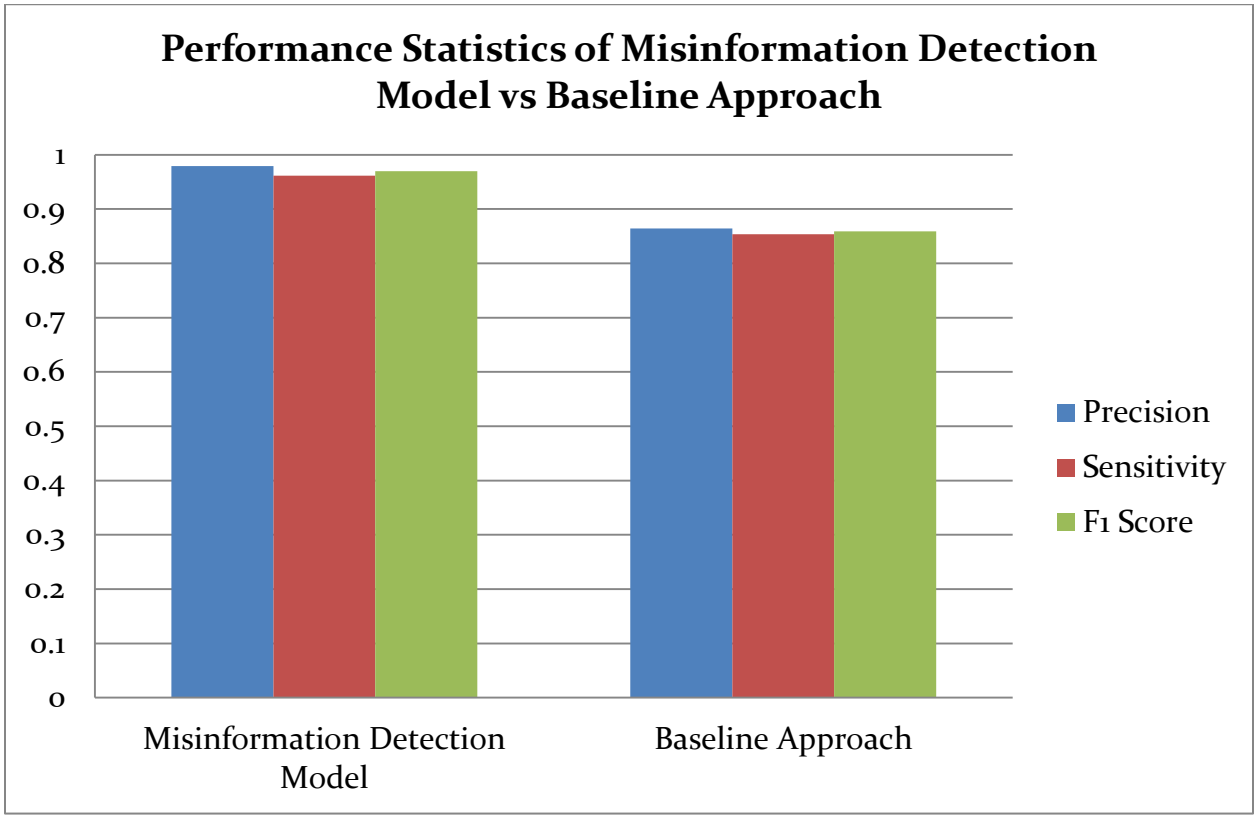

Figure 10 – Performance statistics of the Misinformation Detection Model applied on the training data against the baseline approach

8.4.4 Cross Validation

For the same reasons discussed in Section 7.4.4, 10-Fold Cross Validation was utilized to evaluate the Misinformation Detection Model on generalized data. This technique evaluates generalized data because 10-Fold Cross Validation allows for every piece of data to be used and evaluated as testing data instead of relying on a single testing and training set comparable to the Testing Set Validation technique used in Section 8.4.3.

The resulting AUC values resulting from each fold of 10-Fold Cross Validation on the Misinformation Detection Model are shown in Figure 11. The average AUC value over the 10 folds was 0.9940. It is apparent from the AUC results that the Misinformation Detection Model is able to handle generalized unknown data with a stable and high level of accuracy. This is an important observation because it hints that this model can be applied to a variety of Zika-related data sets.

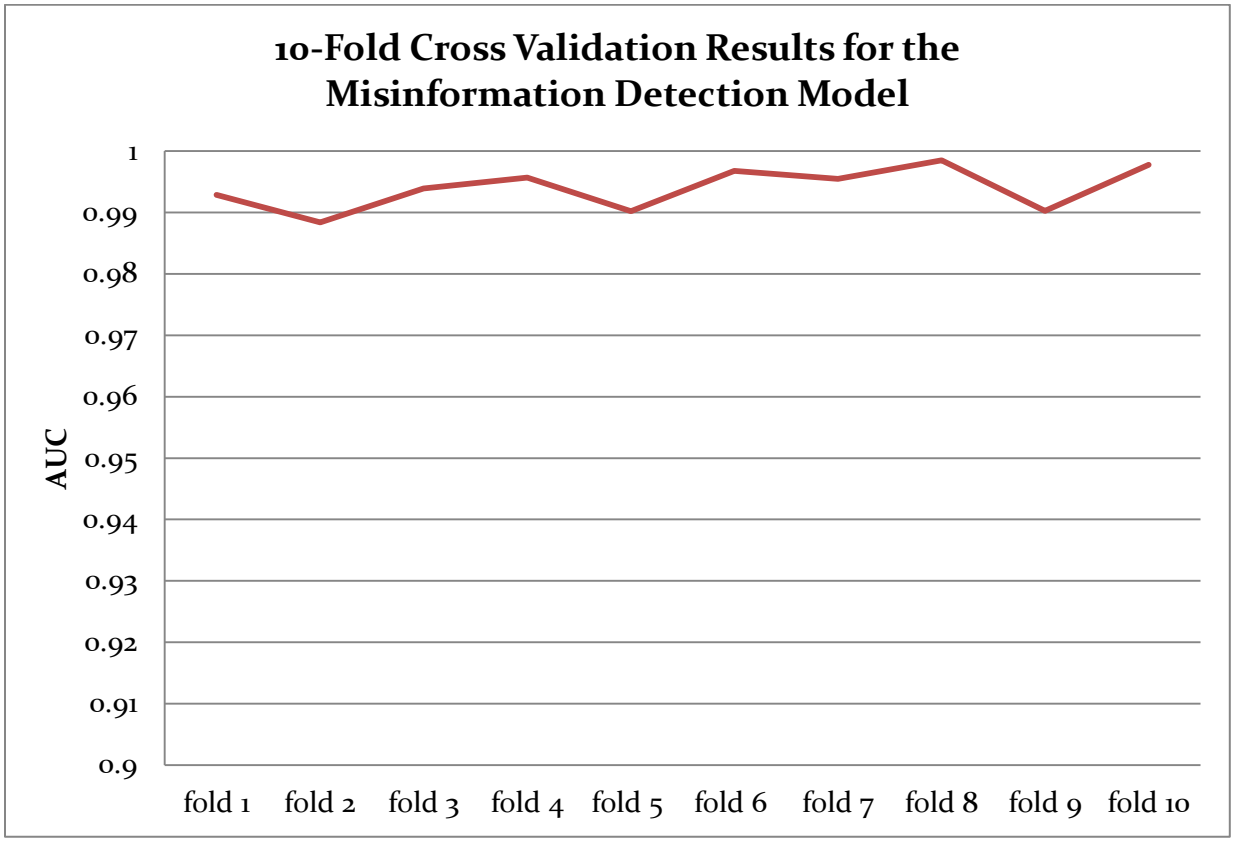

Figure 11 – Resulting AUC values for each fold from performing 10-Fold Cross Validation on the Misinformation Detection Model

8.5 Application

After evaluation, the Misinformation Detection Model was applied to the entire dataset of 996,443 Zika related Tweets. The model detected that 253,421 Tweets (25.43% of dataset) contained misinformation and 743,022 Tweets (74.57% of dataset) did not contain misinformation. The resulting predictions were analyzed and their results are presented in Chapter 9.

Chapter 9

Results

The 253,421 Zika-related Tweets detected as containing misinformation were analyzed and mapped to discover what type of Zika-related misinformation is being spread, where it is spreading from, and who is spreading the misinformation. This dataset will be referred to as detected misinformation.

9.1 Prevalent Misinformation

In order to gauge what kind of misinformation was being spread, the detected misinformation was analyzed to find which previously identified subjects of misinformation appeared the most. Table 21 shows that information relating to the transmission of the Zika virus, Zika virus being extremely deadly, and symptoms associated with Zika being attributed to pesticides were the most prevalent subjects of misinformation in the detected dataset. These previously identified subjects account for approximately 14% of the detected dataset.

Upon further data analysis of the detected misinformation, new subjects of misinformation were identified. These subjects are listed in Table 22 and account for approximately half of the detected dataset. Misleading accounts regarding Zika outbreaks were among the most commonly identified Tweets in the detected dataset at 51,557

Tweets. Tweets exaggerating or underestimating the health effects and risks associated with Zika were also noteworthy with 47,379 Tweets.

Table 21 – Prevalence frequencies and percentages of previously identified subjects of misinformation

Prevalence of Previously Identified Subjects of Misinformation			
Name	*Description*	*% of Tweets*	*# of Tweets*
Transmission	Zika virus is not transmitted from human to human	6.39	16,196
Fatal	Zika virus infection is extremely fatal	3.25	8,236
Pesticide	Symptoms associated with Zika virus infection are actually caused by pesticide	2.05	5,195
Man-made	Zika virus is a man-made virus	0.71	1,799
GMO	Zika virus caused and/or spread by genetically modified mosquitos	0.54	1,368
Pollution	Symptoms associated with Zika virus infection are actually caused by pollution	0.37	937
Autism	Zika virus infection causes autism	0.37	935
Pharmaceutical	Pharmaceutical companies are covering up vaccine related symptoms using the Zika virus infection	0.21	532

Table 22 – Prevalence frequencies and percentages of newly identified subjects of misinformation

Prevalence of Newly Identified Subjects of Misinformation			
Name	*Description*	*% of Tweets*	*# of Tweets*
Outbreaks	Misleading accounts of Zika virus spread	20.34	51,557
Health Risks	Misleading effects and risks of infection (Zika is not harmful or causing cancer etc)	18.7	47,379
Government Conspiracies	Government (hiding information, creating Zika etc)	4.5	11,405
Cure/Vaccine Exists	A cure/vaccine for Zika infection exists	1.27	3,231

9.2 Mapping Misinformation

In order to gauge where the Zika-related misinformation is spreading from, locations were extracted from the profiles of users with positive labels for the "locationListed" feature in the detected dataset. Since user locations can be manually entered by the user, the locations were geocoded using Bing Maps API to determine if the location was valid [5]. Of the 253,421 Tweets in the detected dataset, 159, 905 Tweet authors listed a valid location in their profile. After these locations were geocoded, a word frequency count was performed to distinguish what areas of the world were spreading the

most Zika-related misinformation. Since Tweets in English were exclusively considered during data collection, the Tweets including misinformation from the listed regions were only in English. Table 23 shows the amount of times a region is linked to Tweeting misinformation. The United States of America, United Kingdom, and India were found to distribute the most misinformation.

Table 23 – Frequency percentages of countries based on the number of Tweets linked to detected misinformation

Top 10 Countries Associated with Zika-related Misinformation	
Name	*Occurrence %*
United States of America	18.9
United Kingdom	15.1
India	6.6
Canada	3.6
Nigeria	3.3
Brazil	2.3
Australia	2.1
Venezuela	1.9
Spain	1.9

In order to visualize the origin and spread of Zika-related misinformation, the GPS coordinates obtained from geocoding the detected misinformation were plotted using the Google Maps API [21]. The resulting map is interactive and allows you to hover over a location to read the Tweet associated with it. An example of this behavior is shown in Figure 12 with a screenshot of the system. 25,000 Tweets are displayed in the screenshot. Including all of the detected misinformation at once would obstruct interaction with the location points on the map.

Figure 12 – Map of a subset of 25,000 Tweets selected from the detected misinformation

9.3 Indicators of Misinformation

In order to discover indicators of misinformation, the feature relevancy scores generated in Section 8.4.2 were used as a guide to group relevant features for analysis on the Tweets with detected misinformation. After analyzing the features with highest relevancy, Tweets containing misinformation included the name of a country in 46.64% of their Tweets compared to 8.86% of Tweets not containing misinformation. This may derive from Tweets with misleading Zika outbreak locations. The average Tweet length of a Tweet containing misinformation used 107.23 characters proved to be larger than the 103.14 characters used on average in Tweets not containing misinformation. Additionally, the detected misinformation was Retweeted more with an average of 6.2 Retweets compared to the average of 2.6 Retweets of the Tweets without misinformation.

9.4 Users that Spread Misinformation

In a similar fashion to the technique used in Section 9.3, the relevant features from this dataset pertaining to the users were analyzed for the Tweets containing misinformation. Users who were found to not post misinformation utilized reinforced language, which included phrases such as the names of reputable health agencies, more often (30.35%) than those who were detected to post misinformation (3.07%). This is possibly due to a correlation between factual links from the CDC and other health agencies being posted by users spreading true information. Average account age for users that were shown to spread misinformation was smaller (1250 days) than those who did

not spread misinformation (1384 days). A younger account may indicate that users may use the account just for the spread of misinformation. Furthermore, users that were found to use misinformation listed their location in their profile less (60%) than the users that did not use misinformation (70%). People posting misinformation would more than likely be concerned with their privacy.

Chapter 10

Limitations and Future Work

The “isClaim” feature from the Claim Detection Model (CDM) yielded a low relative influence compared to the list of the top 30 features relevancy scores of the Misinformation Detection Model (MDM). The CDM and the MDM were separately successful in their goals, but the chained approach did not prove particularly useful when compared to the other features. The low relative influence of the “isClaim” feature could result from how misinformation presents in the data set. For example, a Tweet including misinformation could present as an uncertainty, a question, or an opinion. Since the CDM only classifies a Tweet as making a claim if it definitively states a fact, a tweet with uncertainty, a question, or an opinion may include misinformation but will not be classified as making a claim.

Even though the “isClaim” feature provided by the Claim Detection Model proved not to be relevant for misinformation classification, the other descriptive features utilized in this research future accounted for approximately 35% of the relative influence on the model. The manual data analysis required to generate the descriptive phrases used in these features was time consuming and delayed the other parts of the research. A system for automatically describing data via descriptive phrasing analysis could enhance future works. In a similar fashion, the limitation of manually annotating large amounts of data

to establish ground truth is tedious and could be improved through comparable efforts. In addition, the manual annotation process could be improved by employing more than one annotator of domain experts to avoid speculative data.

Since the descriptive features in this research rely on certain words to define a Tweet, they are limited to the selection of words. The words associated with each descriptive feature are shown in Table 6 and were derived from the data analysis performed in Section 5.2. The scope of the descriptive features could be improved by adding more words to the list of words associated with a certain descriptive feature.

Since Tweets are being posted in real-time to microblogging services such as Twitter, future work related to this research could apply a model similar to the one used in this research for real-time misinformation detection during crucial events. Since the model is geared towards understanding the concepts and reactions towards viral infection, it could easily be adapted for use during other medical emergencies. To accomplish real-time classification and learning, a system could establish training rules to automatically extract public health reports to use as ground truths and avoid manual annotation altogether.

In a recent news report, the New York Times Editorial Board detailed the problem of fake news stories on the social media platform Facebook. The report details how fake news can easily spread on social media to millions of users and there is no platform in place to block this type of information spread. The hoaxes and misinformation included in the regarded fake news has proven to be more popular than real news on Facebook and is generating the posters significant revenue. Due to its large audience, the fake news

being spread on Facebook has the potential to influence readers in a negative way. By applying descriptive features relating to the instances of fake news on Facebook, the Misinformation Detection Model created in this research could be retooled for the Facebook platform to help block content containing fake news [51].

Chapter 11

Conclusion

Accurate misinformation detection on the medium of Zika-related Tweets is possible through gradient boosting and is improved by incorporating features derived from PCA. When validated against unknown testing data, the Misinformation Detection Model generated in this research classified misinformation with an accuracy of 0.9711 and maintained an average AUC value of 0.9940 during 10-Fold Cross Validation, which was an improvement from the previous approaches discussed in this study.

Recall the three main questions of this research: is it possible to accurately predict whether or not a Tweet related to Zika is misinformation, what indicators associated with a Tweet can distinguish its validity, and what type of user spreads misinformation. Based on the performance results of the classifier, it is possible to accurately predict whether or not a Tweet related to Zika is misinformation. Based on an analysis of the features that were most relevant during classification, it was possible to identify attributes to distinguish the validity of a Tweet and describe user spreading the Tweet. For example, users who were found to post misinformation utilized reinforced language denoted by the "isReinforced" feature in 3.07% of observed cases of misinformation while the Tweets not labeled as misinformation included reinforced language in 30.35% of their Tweets.

The incorporation of features that described the subject and intention of a Tweet proved useful for detection with approximately 35% of the relative influence on the model deriving from descriptive features. Even though the chained approach of including the "isClaim" feature generated from the Claim Detection Model had minimal influence on misinformation detection, this research was able to accurately detect misinformation using unique features and raise awareness of the concerns related to Zika virus infection while also answering the three main research questions.

References

[1] *ABC Gold Coast.* (2016). *Twitter.com*. Retrieved 1 November 2016, from https://twitter.com/abcgoldcoast/status/702785492526346240

[2] *About Twitter*.(2016). About.twitter.com. Retrieved 7 May 2016, from https://about.twitter.com/company

[3] Arzuza-Ortega, L., Polo, A., Pérez-Tatis, G., López-García, H., Parra, E., Pardo-Herrera, L., & Rodríguez-Morales, A. (2016). Fatal sickle cell disease and Zika virus infection in girl from Colombia. *Emerging infectious diseases*, *22*(5), 925.

*[4] Baker, B. (2015). "Twitter data for research: from understanding relationships to spotting the Aurora Borealis".*Blog.twitter.com. *Retrieved 20 May 2016, from https://blog.twitter.com/2015/twitter-data-research*

[5] *Bing Maps REST Services*. (2016). *Msdn.microsoft.com*. Retrieved 20 July 2016, from https://msdn.microsoft.com/en-us/library/ff701713.aspx

[6] Breiman, L. (1996). Bias, variance, and arcing classifiers.

[7] Brownlee, J. (2016). *How to Evaluate Machine Learning Algorithms with R - Machine Learning Mastery*. *Machine Learning Mastery*. Retrieved 15 October 2016, from http://machinelearningmastery.com/evaluate-machine-learning-algorithms-with-r/

[8] Brownlee, J. (2016). "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning". *Machine Learning Mastery*. Retrieved 20 July 2016.

http://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[9] Brownlee, J. (2016). "Boosting and AdaBoost for Machine Learning". *Machine Learning Mastery*. Retrieved 12 July 2016. http://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/.

[10] Brownlee, Jason. (2016). "Gradient Descent For Machine Learning". *Machine Learning Mastery*. Retrieved July 20 2016. http://machinelearningmastery.com/gradient-descent-for-machine-learning/.

[11] Castillo, C., Mendoza, M., & Poblete, B. (2011, March).Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 675-684). ACM.

[12] *CDC Press Releases*. (2016). *CDC*. Retrieved June 14, from http://www.cdc.gov/media/releases/2016/s0413-zika-microcephaly.html

[13] *Cross Validation*. (2016). *Cs.cmu.edu*. Retrieved 20 October 2016, from http://www.cs.cmu.edu/~schneide/tut5/node42.html

[14] *Decision Trees*.(2016). *Scikit-learn.org*. Retrieved June 10 2016, from http://scikit-learn.org/stable/modules/tree.html

[15] Dwyer, K., & Holte, R. (2007, September). Decision tree instability and active learning. In *European Conference on Machine Learning* (pp. 128-139).Springer Berlin Heidelberg.

[16] *Fatal Zika Virus Infection with Secondary Nonsexual Transmission*. (2016). New England Journal of Medicine. Retrieved 3 November 2016, from http://www.nejm.org/doi/full/10.1056/NEJMc1610613?query=featured_home

[17] Freund, Y., & Schapire, R. E. (1995, March). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23-37).Springer Berlin Heidelberg.

[18] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[19] *Functional differences between Twitter and Sina Weibo*. (2015). Reach 7. Retrieved 30 October 2016, from http://blog.reach7.com/reach7s-china-guide-part-3-functional-differences-between-twitter-and-sina-weibo/

[20] G. W. A. Dick, S. F. Kitchen, & A. J. Haddow. (1952) Zika Virus (I).Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 46(5), 509-520. doi:10.1016/0035-9203(52)90042-4

[21] *Google Maps Geocoding API*. (2016). *Google Developers*. Retrieved 16 October 2016, from https://developers.google.com/maps/documentation/geocoding/start

[22] Gorakala, S. (2016). *Principal Component Analysis using R. R-bloggers*. Retrieved 16 August 2016, from https://www.r-bloggers.com/principal-component-analysis-using-r/

[23] Gottfried, J. & Elisa S. (2016). "News Use Across Social Media Platforms 2016". Pew Research Center's Journalism Project. *Retrieved 2 June 2016.*

http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

[24] Guillain-Barré Syndrome Fact Sheet.(2016). Ninds.nih.gov. Retrieved 2 July 2016, from http://www.ninds.nih.gov/disorders/gbs/detail_gbs.htm

[25] Gupta, A., & Kumaraguru, P. (2012, April).Credibility ranking of Tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (p. 2).ACM.

[26] Kalyanam, J., Velupillai, S., Doan, S., Conway, M., & Lanckriet, G. (2015). Facts and Fabrications about Ebola: A Twitter Based Study. *arXiv preprint arXiv:1508.02079*.

[27] Kindhauser, MK., Allen, T., Frank, V., Santhana, R., & Dye, C. (2016). Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization*, doi: http://dx.doi.org/10.2471/BLT.16.171082

[28] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World Wide Web* (pp. 591-600). ACM.

[29] Liang, G., He, W., Xu, C., Chen, L., & Zeng, J. (2015). Rumor Identification in Microblogging Systems Based on Users' Behavior. *IEEE Transactions on Computational Social Systems*, *2*(3), 99-108.

[30] *Lognormal Distribution*. (2016). *Mathworks.com*. Retrieved 3 August 2016, from http://www.mathworks.com/help/stats/lognormal-distribution.html?requestedDomain=in.mathworks.com

[31] *Loss function*. (2016). *Dictionary.com*. Retrieved June 20 2016, from http://www.dictionary.com/browse/loss-function

[32] Martins, T. (2013).*Computing and visualizing PCA in R*. Retrieved 16 August 2016, from https://tgmstat.wordpress.com/2013/11/28/computing-and-visualizing-pca-in-r/

[33] *Mean F Score*. (2016). *Kaggle.com*. Retrieved 30 October 2016, from https://www.kaggle.com/wiki/MeanFScore

[34] Mendoza, M., Poblete, B., & Castillo, C. (2010, July). Twitter under crisis: can we trust what we RT?. In *Proceedings of the first workshop on social media analytics* (pp. 71-79).ACM.

[35] *Microcephaly Information Page: National Institute of Neurological Disorders and Stroke (NINDS)*. (2016). Ninds.nih.gov. Retrieved 10 July 2016, from http://www.ninds.nih.gov/disorders/microcephaly/microcephaly.htm

[36] Oliveira, W., Coelho, G., Badaró, R., Cortez, J., Ospina, M., Pimentel, R., Masis, R., Hernandez, F., Lara, B., Montoya, R., Jubithana, B., Melchor, A., Alvarez, A., Aldighieri, S., Dye, C., & Espinal, M. (2016).Zika Virus and the Guillain–BarréSyndrome — Case Series from Seven Countries. *New England Journal of Medicine*, 375(16), 1598-1601.doi: 10.1056/NEJMc1609015

[37] Pang-Ning, T., Steinbach, M., & Kumar, V. (2006).Introduction to data mining. In *Library of congress* (Vol. 74).

[38] *Posting a Tweet*.(2016). *Twitter Help Center*. Retrieved 7 May 2016, from https://support.twitter.com/articles/15367

[39] *Principal Component Analysis*. (2016). *Nonlinear.com*. Retrieved 16 August 2016, from http://www.nonlinear.com/support/progenesis/lc-ms/faq/v4.1/pca.aspx

[40] Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011, July). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589-1599).Association for Computational Linguistics.

[41] *Quick-R: Density plots*. (2016). *Statmethods.net*. Retrieved 3 August 2016, from http://www.statmethods.net/graphs/density.html

[42] Quinlan, J. R. (1996, August). Bagging, boosting, and C4. 5. In *AAAI/IAAI, Vol. 1* (pp. 725-730).

[43] R: What is R?.(2016). R-project.org. *Retrieved July 1 2016, from https://www.r-project.org/about.html*

[44] Ridgeway, G. (2007). Generalized Boosted Models: A guide to the GBM package. *Update*, *1*(1), 2007.

[45] Ridgeway, Greg. (2015). *Cran.r-project.org*. Retrieved August 10 2016, from https://cran.r-project.org/web/packages/gbm/gbm.pdf

[46] Ringner, M. (2008). What is Principal Component Analysis?. *Nature Biotechnology*, *26*(3), 303. Retrieved from http://www.nature.com/nbt/journal/v26/n3/full/nbt0308-303.html

[47] *ROC curves and Area Under the Curve explained (video)*. (2014). *Data School*. Retrieved 30 October 2016, from http://www.dataschool.io/roc-curves-and-auc-explained/

[48] *Rumor*. (2016). *Oxford Dictionaries*. Retrieved 10 August 2016, from https://en.oxforddictionaries.com/definition/rumour

[49] Sikka, V., Chattu, VK.,Popli, RK., Galwankar, SC., Kelkar, D., Sawicki, SG., Stawicki, SP., &Papadimos, TJ. (2016). The emergence of zika virus as a global health security threat: A review and a consensus statement of the INDUSEM Joint working Group (JWG). *Journal Of Global Infectious Diseases*, *8*(1), 3. doi:10.4103/0974-777x.176140

[50] *Simple guide to confusion matrix terminology*. (2014). *Data School*. Retrieved 30 October 2016, from http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[51] The New York Times Editorial Board. (2016). *Facebook and the Digital Virus Called Fake News*. *Nytimes.com*. Retrieved 20 November 2016, from http://www.nytimes.com/2016/11/20/opinion/sunday/facebook-and-the-digital-virus-called-fake-news.html

[52] *The Syndicate on Twitter*. (2016). *Twitter*. Retrieved 1 November 2016, from https://twitter.com/boston_daryl/status/720246005850103808

[53] *Twitter Firehose vs. Twitter API: What's the difference and why should you care?.(2013). BrightPlanet*. Retrieved August 10 2016, from https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/

[54] *Twitter Usage Statistics - Internet Live Stats*. (2011). *Internetlivestats.com*. Retrieved 7 May 2016, from http://www.internetlivestats.com/twitter-statistics/

[55] *Using hashtags on Twitter*. (2016). *Twitter Help Center*. Retrieved 7 May 2016, from https://support.twitter.com/articles/49309

[56] *What are the advantages and disadvantages of stratified random sampling?*. (2015). *Investopedia*. Retrieved 5 August 2016, from http://www.investopedia.com/ask/answers/041615/what-are-advantages-and-disadvantages-stratified-random-sampling.asp

[57] Zhou, Z. H. (2015). Ensemble learning. *Encyclopedia of Biometrics*, 411-416.

[58] *Zika Case Counts in U.S.* (2016). *CDC*. Retrieved 12 May 2016, from http://www.cdc.gov/zika/geo/united-states.html

[59] *Zika Symptoms, Testing and Treatment*.(2016). CDC. Retrieved 14 May 2016, from https://www.cdc.gov/zika/symptoms/

[60] *Zika Virus – Microcephaly & Other Birth Defects*. (2016). *CDC*. Retrieved 3 November 2016, from https://www.cdc.gov/zika/healtheffects/birth_defects.html

[61] *Zika Virus - Transmission and Risks*. (2016). CDC. Retrieved 12 May 2016, from http://www.cdc.gov/zika/transmission/

[62] *Zika Virus- Health Effects & Risks*. (2016). CDC. Retrieved 2 July 2016, from http://www.cdc.gov/zika/healtheffects/index.html

[63] *Zika virus.*(2016). World Health Organization. Retrieved 12 July 2016, from http://www.who.int/mediacentre/factsheets/zika/en/

[64] Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analyzing how people orient to and spread rumors in social media by looking at conversational threads. *PloS one*, *11*(3), e0150989.