THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UTILIZING FRIENDSHIP NETWORKS TO PREDICT USER SUSCEPTIBILITY TO FAKE NEWS

ROBERT COWELL
SPRING 2019

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Computer Science
with honors in Computer Science

Reviewed and approved* by the following:

Dongwon Lee
Associate Professor of Information Sciences and Technology
Thesis Supervisor

Danfeng Zhang
Assistant Professor of Computer Science and Engineering
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

While misinformation and disinformation have always existed in society, the prevalence of fake news on social media threatens to create divisions in society and erode trust in real news. Researchers have studied characteristics of fake news and developed accurate models to identify it on social media. However, understanding the human element of this societal problem is important. This thesis studies Twitter replies to fake news posts surrounding the shooting at Marjory Stoneman Douglas High School in Parkland, Florida and proposes a model for predicting a user's level of susceptibility to fake news by utilizing features derived from a user's friendship network. The features include user-based, clustering, centrality, degree, and psychology-based features. The final model, gradient-boosted trees (XGBoost) trained on a combination of 27 features from the aforementioned feature categories, achieved an AUC of 0.715. This model can be used in tandem with existing fake news detection models to create a sliding-scale intervention method based on predicted user susceptibility.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I want to thank Dr. Dongwon Lee, my thesis supervisor, for his guidance throughout this process. I am incredibly grateful to have been brought onto the user susceptibility project and allowed to contribute to and learn from the interesting work Dr. Lee oversees.

Thank you to Tracy Shen for her patience, guidance and willingness to integrate me into her research. I'm proud of the work that we have accomplished together.

Chapter 1

Introduction

1.1 Background

Disinformation, deliberately misleading or false information, has always existed in society. However, the invention of social media, connecting millions of people around the world, provided an unparalleled conduit for the spread of disinformation. It has been called rumors, hoaxes or propaganda in the past. Today, it is called fake news, a term popularized during the 2016 United States presidential election. In this thesis, fake news is defined as news that is verifiably false and intended to mislead readers [13]. As allegations have emerged about outside actors utilizing social media and fake news as tools to divide the American public, the importance of fully understanding this threat is critical.

Social media has become the primary news source for an increasing number of people, with 68% of Americans reporting that they obtain their news at least occasionally from social media websites [11]. News on social media is faster and easier to consume than traditional news sources, such as television or newspapers, leading many to rely exclusively on social media for news. Social media also allows news to spread to thousands of people within seconds. With 71% of Twitters users saying they get news from the website [11], it is important to minimize the likelihood of fake news gaining traction on Twitter.

The power of fake news on Twitter was displayed on April 23, 2013. The official Twitter account of the Associated Press tweeted, "Breaking: Two Explosions in the White House and Barack Obama is injured." The tweet reached the AP's two million followers instantaneously. Within three minutes, the Dow Jones Industry Average fell 150 points, wiping out an estimated $136 billion [5]. The stock market recovered shortly after AP employees revealed the account had been compromised by hackers. However, the damage had been done. Fake news has undeniably real consequences on our lives, influencing decisions and sowing discord in society.

1.2 Research Goal

Fake news is a societal problem, not just a technological one. However, there is a distinct lack of research into understanding the human element of this phenomenon—why people fall for fake news. Without the ability to identify those who are susceptible to fake news, it will be difficult to truly make an impact on its spread. In this thesis, susceptibility is defined as the likelihood of someone believing fake news on social media. Susceptibility is measured on a five-point scale, ranging from highly susceptible, slightly susceptible, neutral, not-quite susceptible to not-at-all susceptible. While susceptibility cannot be directly observed, a user's level of agreement with fake news tweets was used as an observable alternative. Then, a friendship network of users and their immediate friends was created. This thesis addresses the following question: Can the susceptibility level of users on Twitter be accurately predicted from features derived from their friendship network?

1.3 Related Work

Fake news detection has been studied extensively. Researchers have used a combination of features, including the content of the fake news, the characteristics of the user spreading the news and the network of those who have interacted with the news [13]. The networks can take many forms, such as interaction networks to model how fake news spreads and friendship networks to model how users form communities [13].

There is limited prior research on susceptibility to fake news in the computational domain. However, psychological studies have examined the characteristics of those who believe rumors and fake news. Factors such as overclaiming knowledge and low analytical thinking scores, correlated strongly with the likelihood of believing fake news [8]. A phenomenon in psychology called the illusory truth effect, the tendency to believe information after repeated exposure, has been found to apply to both rumors and fake news [1, 8]. There are two additional well-studied psychological phenomena that work in conjunction to affect the susceptibility of a user. One of these phenomena is the echo chamber effect, where users tend to form groups with other like-minded users, creates self-selecting communities that reduce the likelihood of users seeing information or opinions critical of their views [9]. The other psychological phenomenon is confirmation bias, where an individual tends toward information that confirms their existing views regardless of its veracity. These two effects combine to increase the likelihood of the user believing fake news that confirms their views [7].

In the context of studying social media users, susceptibility has often referred to a stage of compartmental epidemiological models. Given the similarities between information diffusion on social media and the spread of disease, researchers have used epidemiological models to study information diffusion on social media. One study used the traditional SIR (susceptible, infected, recovered) model [15]. Another used a variant on SIR, called SEIZ (susceptible, exposed, infected, skeptic), to model rumor propagation [4]. Both models defined a susceptible user as one who has not been infected. Infection was defined as a user spreading the information in question [4, 15]. However, user susceptibility is defined differently in this thesis, where susceptibility is the user's level of belief in a fake news post.

Another study that modelled viral information diffusion defined susceptibility as the level to which a user can be convinced to adopt items that have been introduced to the user. While not linked the the SIR/SEIZ epidemiological models, this concept of susceptibility was inseparable from the virality of the content and the user spreading the content, as opposed to a discrete concept as in this thesis [2].

Wagner et al. studied the susceptibility of social media users to social bots, which are commonly used to spread disinformation. Susceptibility was defined as any interaction with the social bot, or simply following the bot, regardless of the intent behind the interaction. The prediction model had two components: binary classification of susceptible/non-susceptible users as well as prediction of the level of susceptibility of a user that had already been determined to be susceptible user. While the binary classification task achieved an AUC of 0.71, the model predicting the level of susceptibility struggled to differentiate between highly susceptible and less susceptible users [14].

Rath et al. displayed the utility of user networks to understand their behavior. Specifically, the study used retweet networks to identify rumor spreaders on Twitter [10]. However, the study used interaction networks, as opposed to friendship networks. Another study employed friendship networks to classify rumors, indicating the utility of friendship networks in discerning user susceptibility to false information [6]. Both studies inspired application of network features to differentiate user susceptibility in this thesis.

While the literature in this realm addresses the identification of fake news and users who spread it, there has never been an attempt to predict user susceptibility to fake news. This thesis will contribute to the research field by presenting a model that accomplishes this important task.

Chapter 2

Data Collection

2.1 Susceptibility Dataset

The susceptibility dataset was comprised of replies to fake news tweets related to the shooting at Marjory Stoneman Douglas High School in Parkland, Florida (MSD shooting) on February 14, 2018. The dataset was built from seven tweets identified as fake news through verification with trusted third-party websites and news organizations (see Table 2.1). All replies from the post date to May 20, 2018 were collected, resulting in a total of 906 tweets. Using context of the tweet and a user's reply, annotators labeled a user's reaction to the fake news. The level of agreement that the user displays in their reply to the fake news was used as an observable analogue to a user's susceptibility. Each reply in the dataset was labeled on the scale: strong agreement, weak agreement, neutral, weak disagreement and strong disagreement. These classes correspond to highly susceptible, slightly susceptible, neutral, not quite susceptible, and not at all susceptible, respectively [12].

Table 2.1: Seven fake news tweets studied and their verification methods.

#	Fake News	Verified Fake By
1	MSD shooting survivor and gun control activist David Hogg is heavily coached on lines in interviews	USA Today
2	MSD shooting survivors and student activists have powerful backers	CBS News
3	#MarchForOurLives (MSD Shooting student activist led protest for gun control) is hiring people on Craigslist to participate in the march	snopes.com
4	MSD High School student Colton Haab calls out CNN Town Hall bias after they refused to approve his question	snopes.com
5	#MarchForOurLives attracted 850,000 people to participate	CBS News
6	CNN did not allow Colton Haab to ask a question at #StudentsStandUp because he was pro-gun rights	CNN
7	David Hogg is heavily coached by the FBI	New York Times

For example, a user replied to Fake News Tweet 1 writing, "Always knew these kids had ties to #Soros. The parents need to have a mental background check." The user's reply reveals that they strongly agree with the content of the fake news. Given that the user was in the strong agreement class, the user was categorized as highly susceptible to fake news.

Five human annotators, through Amazon Mechanical Turk, labeled each reply, with majority rule determining the final class label. Only 768 of the original 906 replies achieved a majority label [12]. The other 134 replies were not used in the model creation as annotators failed to come to a consensus.

Table 2.2: Distribution of susceptibility dataset.

Susceptibility	**Class**	**#**	**%**
Highly Susceptible	Strong Agreement	281	37
Slightly Susceptible	Weak Agreement	97	13
Neutral	Neutral	223	29
Not Quite Susceptible	Weak Disagreement	56	7
Not At All Susceptible	Strong Disagreement	111	14

The dataset also included basic user information scraped from Twitter (see Table 2.3). Some of the attributes are created by users as opposed to Twitter, such as the location attribute which resulted in entries ranging from “New York” to “None of Your Business, USA”, limiting its utility.

Table 2.3: Susceptibility dataset attributes.

Dataset Attributes	**Description**
ID	Unique numeric identifier assigned by Twitter
Name	User-provided display name
Screen Name	User-created handle
Reply	Contents of user’s reply to fake news
Label	Susceptibility score
Location	User-provided location
Protected	Whether the user’s account is private (binary)
Description	User-provided bio
Followers Count	Number of followers of the user
Friends Count	Number of friends (users who the user follows)
List Count	Number of user-created lists (groups of related users)
Created Time	Date when user created their account
Favorites Count	Number of tweets favorited by user
Timezone	User-provided timezone
Statuses Count	Number of tweets written by user
Language	Twitter interface language selected by user

2.2 Friendship Network

In order to extract network-based features, friendship networks were created from the fake news repliers and their immediate neighbors (hereafter called second-level users). Friendship on most social media platforms represent a mutual connection, where both users must agree to the friendship and the users will see each other’s posts. However, Twitter has two friendship types, follower and friend, which are asymmetrical friendships. If Alice follows Bob, Alice agreed to see Bob's posts without requiring Bob's consent, but Bob does not see Alice's posts. Followers of Alice are users who follow Alice, and therefore are the users who see Alice's posts. On the other hand, friends of Alice are users Alice follows, and Alice sees their posts.

The followers and friends of each replier were extracted using Twitter's API. There were 902 unique repliers of the 906 replies to the 7 fake news tweets. The users whose replies failed to get a majority label were still included in the friendship network. This resulted in 1,134,240 unique second-level users connected to the 902 repliers. Because of the exponential increase in users and rate limitations of Twitter's API, only users directly adjacent to the repliers were collected.

The raw user data was converted into a multi-edge directed graph. The direction of the edges in the graph was based on the nature of the relationship between the two users. For example, if a replier named Alice followed Bob (i.e. Bob was in Alice's friends list and Alice was in Bob's follower list), the corresponding edge in the graph would be: Bob $\rightarrow$ Alice. This direction resembles the information flow on Twitter, as only a user's followers will see their tweets and retweets.

There were two user-types of friendship networks and two community-types of friendship networks. The division of the collected friendship data into different friendship networks acted as lenses to give more or less detail to the structure of the network.

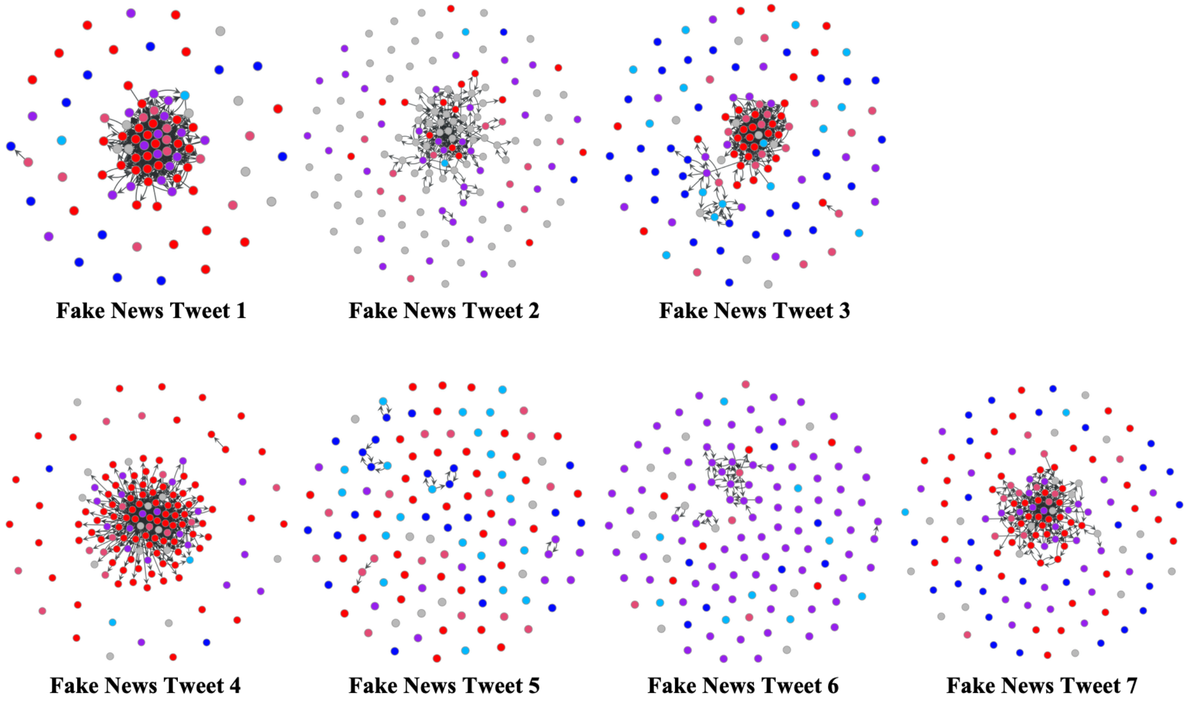
The user-types of networks were full networks and repliers-only networks. The repliers-only network represented the connections between repliers of the fake news tweets exclusively, and a full network contains both repliers and second-level users. The differentiation allowed the study of connections among repliers alone and the network as a whole.

The two community-types of networks were tweet-level networks and combined networks. Tweet-level networks were the communities based around each fake news tweet. A replier only appears in the tweet-level network if they interacted with the tweet in question. As a result, there are tweet-level networks for each fake news tweet. Each tweet-level network represents a nearly-disparate community, as there are only four repliers who appear in more than one tweet-level network. On the other hand, combined networks merge all users in the dataset into one network. The community-types allowed for features to be calculated in micro and macro contexts.

In total, 16 networks were created, with a repliers-only and full network for each of the seven tweet-level networks and one combined network.

Table 2.4: A comparison of different friendship networks.

Fake News Tweet #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Combined**
Repliers Network Nodes	90	148	123	147	120	130	149	902
Full Network Nodes	292,770	184,773	321,814	532,103	124,441	174,443	232,864	1,135,142
Repliers Network Edges	395	180	253	448	16	39	209	8849
Full Network Edges	994,662	648,800	938,454	1,224,057	258,810	400,314	722,605	5,156,735



Fake News Tweet 1

Fake News Tweet 2

Fake News Tweet 3

Fake News Tweet 4

Fake News Tweet 5

Fake News Tweet 6

Fake News Tweet 7

Figure 2.1: Repliers-only tweet-level friendship networks with susceptibility labels. (Red = highly susceptible, pink = slightly susceptible, purple = neutral, light blue = not quite susceptible, dark blue = not at all susceptible)

In addition to the connections between the repliers and second-level users, basic user information was collected on the second-level users, mirroring the attributes contained in the susceptibility dataset for future analysis (Table 2.3).

Chapter 3

Feature Extraction

There were five categories of features extracted from the friendship networks. The categories were user features, degree features, clustering features, centrality features and psychology-based features.

3.1 User Features

User attributes are the basic information and metrics that exist for every user on Twitter. Many of the attributes in the susceptibility dataset (see Table 2.3) are user attributes.

There are two types of user attributes: Twitter-created and user-created. For user-created attributes, users are given a textbox to enter information (e.g. bio and location), resulting in an unlimited range of entries as well as potential errors in the entries. In addition, there is no way to verify the user was truthful when entering these attributes. Twitter-created attributes, on the other hand, are maintained by Twitter, such as the number of statuses posted by the user. Twitter-created attributes are consistent in format and more likely to be accurate than user-created attributes. As a result, user features only consisted of Twitter-created attributes.

Table 3.1: User features and their descriptions.

Name of Feature	**Description**
Statuses Count	Number of Tweets written by user
Favorites Count	Number of tweets favorited by user
List Count	Number of user-created lists (groups of related users)

Within the context of the friendship networks, the selected user features identify the user's level of engagement on Twitter.

3.2 Degree Features

The connections between a user and their friends and followers is an important trait of a social media user. The connections indicate a user's relative importance and influence within their community. Using the friendship network model mimicking information flow described in Section 2.2, the in-degree of a user was defined to be their friends count and out-degree of a user was their follower count.

Table 3.2: In/Out degree features and their descriptions.

Name of Feature	**Description**
In-Degree	Number of friends of the user (i.e. friends count)
Out-Degree	Number of followers of user (i.e. followers count)
Repliers-Followed Count	Number of friends of user who also replied to fake news tweet
Influence	$\frac{O(u)}{I(u)}$, where $O(u)$ is out-degree of user u and $I(u)$ is in-degree
Weighted Influence	$\frac{O(u)}{I(u)} * O(u)$, influence weighted by out-degree
Second-Level Influence	$\sum_v \frac{O(v)}{I(v)}$, where v is second-level users of replier u

From the edges in the friendship network, repliers-followed count was created, which represents the number of repliers to the same fake news tweet that another replier follows (i.e. in-edges to the replier in question), to understand the relationships between repliers.

The influence features (see Table 3.2) are basic measures of a user's influence over others on Twitter. From the asymmetrical relationship of Twitter friendships, tweets of a user who has more followers are seen by a larger audience than users with fewer followers, increasing the likelihood of influencing a follower with their tweets.

However, some social media users perform a technique known as "follow-for-follow" on Twitter and other social media platforms to increase their follower count. These users follow large numbers of users hoping the target users reciprocate and follow them back. But users who perform follow-for-follow now have equal, if not more, friends compared to followers. As a result, influence cannot be determined based on follower count alone. Alternatively, celebrities are examples of truly influential users. Celebrity social media accounts often have tens or hundreds of friends, but thousands or millions of followers.

A measure for influence must be based on both follower count and friend count to control for follow-for-follow. As a result, the influence feature is the ratio between a users followers count and their friends count. A typical celebrity will have an influence score in the thousands whereas a typical follow-for-follow user will have an influence score of less than one.

However, this influence feature can fail to capture an accurate influence score for some users. For example, a user who has 1,000 followers and 100 friends has the same influence value as a user with 10,000 followers and 1,000 friends when the standard influence feature is used. Intuitively, the user with more followers would be more influential on Twitter, because more users see their posts. Therefore, another measure of influence was required to correct for ties in the standard influence feature. Weighted influence scales the influence score by the user's followers count, resulting in a larger value for those who have the same standard influence score but more followers.

Lastly, both influence scores do not consider the users who follow the repliers. A user who has influential followers increases the user's influence. To calculate this third influence measure, called second-level influence, the standard influence values for all followers of the user in question were summed. A user with a high second-level influence does not necessarily have large values for the other influence features.

3.3 Clustering Features

Clustering features fall into two categories, graph clustering and traditional clustering. Graph clustering identifies groups of highly connected users. Traditional clustering, however, utilizes a user's characteristics to group similar users, irrespective of their location in the network. Clustering coefficient, triangle count, clique count and cycle are the graph clustering features (Table 3.3). Cluster, common neighbor, average neighbor and majority neighbor are features representing more traditional clustering techniques (Table 3.3). The features were calculated on each type of friendship network described in Section 2.2.

Table 3.3: Clustering features and their descriptions.

Feature Name	**Description**
Clustering Coefficient	$\frac{2T(u)}{deg(u)(deg(u)-1)}$, where $T(u)$ is triangles through u and $deg(u)$ is degree
Triangle Count	Number of triangles that include the user (i.e. $T(u)$)
Clique Count	Number of cliques the user is in
Cycle	Whether a cycle exists in the network through the user (binary)
Cluster	Five clusters from k-means clustering on adjacency matrix
Common Neighbor (Susceptible)	Whether the user follows the most frequent neighbor of susceptible users (binary)
Common Neighbor (Non-Susceptible)	Whether the user follows the most frequent neighbor of non-susceptible users (binary)
Average Neighbor	Average susceptibility score of user's neighbors
Majority Neighbor	Susceptibility score of the majority of user's neighbors

3.4 Centrality Features

Centrality features describe the importance of the user within the network. Centrality features were calculated on each type of friendship network.

Table 3.4: Centrality features and their descriptions.

Feature Name	**Description**
Degree Centrality	Fraction of nodes v connected to node u
Closeness Centrality	$C(u) = \frac{n-1}{\sum_v dist(v,u)}$
PageRank	Values assigned to user from Google's PageRank algorithm

Degree centrality is the fraction of users connected to a replier in the network. Closeness centrality is based on the distance between the replier and all other users in the network.

While large values for degree centrality and closeness centrality indicate a more important user in the friendship network, large values from the PageRank algorithm indicate a lack of importance. PageRank uses in-edges to calculate its importance value, resulting in a low value for users with exclusively out-edges. However, as explained in Section 3.2, a user with many followers (out-edges) is more influential than one with few followers. Thus, PageRank calculates a score for how unimportant a user is.

3.5 Psychology-Based Features

Using the findings of psychological studies related to fake news and rumors, features were based on characteristics correlated with fake news susceptibility. While unable to survey the users directly, these characteristics can be approximated from the behavior of the user and their neighbors.

Table 3.5: Psychology-based features and their descriptions.

Name of Feature	**Description**
Creator Follower	Whether the user follows the fake news tweet creator (binary)
Familiarity Score	Number of friends who retweeted fake news tweet
Scaled Familiarity	Familiarity score scaled by number of friends of the user

Confirmation bias and the echo chamber effect combine to increase the likelihood of a user agreeing with their friends' views [13]. Therefore, users who follow the fake news creator are more likely to believe their tweets as a result of the two psychological phenomena. In order to capture their effects, a binary feature was created to indicate if the user followed the originator of the fake news.

Pennycook et al. observed prior familiarity with fake news headlines was positively correlated with a one's belief in its veracity, known more generally as the illusory truth effect [8]. Prior exposure can be approximated by determining the number of times the tweet appears on a user's Twitter timeline. Twitter's timeline is a stream of tweets and retweets from a user's friends. A Twitter user can see a tweet in three ways: a user's friend posts a tweet, a user's friend retweets another user's tweet, or a user's friend replies to a tweet written by another friend. Given reply tweets on Twitter's timeline do not show the original tweet, the two pertinent ways for a user to see the fake news tweet is following the creator of the fake news or a friend retweets the fake news tweet. While a feature already existed to determine if the user follows the creator of the tweet, retweets were the factors not already captured. The familiarity score of a user is the number of user's friends who retweeted the fake news tweet. It is impossible to confirm that the replier saw any of these retweets, given Twitter curates a user's timeline to show the "best" tweets and a user may not have viewed their timeline to see all (or any) of the retweets.

Chapter 4

Results

4.1 Testing Methodology

The evaluation of each model utilized the following procedure. The prediction task was converted from a five-class classification to five one-vs-rest binary classifications. Then, 5-fold stratified cross validation was used to test each model. Stratified cross validation was used in order to preserve the imbalance of the classes (see Table 2.2). The evaluation metric of each model was the area under the receiver operating characteristic (ROC) curve averaged over the five folds.

First, a wide array of machine learning models were run on all features to compare their relative performance at prediction. Then, each feature category was compared to each other. Finally, features were trimmed to create a robust final model.

The aforementioned testing methodology provided the ability to determine the strengths and weaknesses of the models at predicting each susceptibility level. In addition, area under the ROC curve (AUC) was chosen because it is a popular metric for both model construction and model comparisons [3].

4.2 Model Comparisons

A wide range of machine learning models were tested to determine the highest-performing option for the final model. K-nearest neighbors (KNN), naive Bayes and support vector machines (SVM) were the simpler models tested. Random forests, gradient-boosted trees (XGBoost) and multilayer perceptrons (MLP) were the more complex models tested. The six models were chosen to cover the spectrum of techniques and complexity. Each model was tested with all features included, using the method described in Section 4.1.

Table 4.1: Comparison of mean 5-fold AUC for each model. A bolded cell indicates the highest mean AUC for the class.

	KNN	**Naive Bayes**	**SVM**	**Random Forest**	**XGBoost**	**MLP**
Highly Susceptible	0.533	0.610	0.499	0.729	**0.735**	0.518
Slightly Susceptible	0.437	0.509	0.503	**0.528**	0.480	0.464
Neutral	0.551	0.545	0.503	**0.696**	0.691	0.467
Not Quite Susceptible	0.591	0.614	0.506	0.658	**0.730**	0.596
Not At All Susceptible	0.605	0.711	0.501	0.856	**0.861**	0.557
Average	0.544	0.598	0.502	0.693	**0.699**	0.520

The strongest performing models were random forests and gradient-boosted trees, with an AUC averaged across all classes of 0.693 and 0.699 respectively. The random forest model traded stronger performance in the slightly susceptible and neutral classes for slightly worse performance in the rest of the classes. Despite the model's simplicity compared to other models tested, naive Bayes was the third highest performing with an average AUC of 0.598.

Gradient-boosted trees were selected for future work. With random forests and gradient-boosted trees having the two highest AUC values for every category, they emerged as the two contenders for selection. Gradient-boosted trees excelled in the highly susceptible, not-at-all susceptible categories as well as the overall average AUC, which indicated its utility for integration into social media for identifying those most vulnerable who require intervention or skeptics who require little attention.

4.3 Feature Comparisons

The relative prediction ability of the five feature categories were compared to understand their importance in the final model. All feature categories achieved an average AUC above random chance (0.5), indicating they all capture some information about user susceptibility.

Table 4.2: Comparison of mean 5-fold AUC for each feature category. A bolded cell indicates the highest mean AUC for the class.

	User	**Degree**	**Clustering**	**Centrality**	**Psychology-Based**
Highly Susceptible	0.539	0.583	0.641	**0.734**	0.699
Slightly Susceptible	0.491	0.559	0.480	0.502	**0.580**
Neutral	0.516	0.545	0.585	**0.702**	0.538
Not Quite Susceptible	0.603	0.596	0.600	0.671	**0.697**
Not At All Susceptible	0.545	0.664	0.768	0.828	**0.839**
Average	0.539	0.589	0.615	**0.687**	0.671

Centrality and psychology-based feature categories excelled with 0.687 and 0.671 AUCs, respectively. The centrality features achieved high AUCs for overall highly susceptible, neutral classes and the overall average. Psychology-based features won the slightly susceptible, not-quite susceptible and not-at-all susceptible classes, but did not get the best average AUC, because of the 16.4% difference in the neutral class AUC compared to centrality features.

User features achieved the lowest average AUC, with a significantly lower AUC for not-at-all susceptible class compared to the other features. However, there were only three user features, the fewest features per category.

The results of the feature comparison indicate that centrality and psychology-based features are important aspects of the final model.

4.4 Feature Selection

To select the best features, the model was trained with 5-fold cross validation. The information gain of each feature was averaged over the folds to calculate an aggregate information gain. Then, backward elimination was performed, where the feature with the lowest information gain was removed. Then, 5-fold cross validation was performed using the remaining features. To determine the best subset of features, the average AUC calculated across the folds was used. This technique prevents overfitting, because the model was not trained on the full dataset at any stage of feature selection.

Feature selection reduced the number of features from 36 to the 27 features used in the final model. The final model consisted of at least one feature from every feature category. All user features, degree features and psychology-based features were used in the final model.

Table 4.3: Final model features.

Feature Categories	**Features**
User Features	Statuses Count, Favorites Count, List Count
Degree Features	In-Degree, Out-Degree, Influence, Weighted Influence, Second-Level Influence, Repliers-Followed Count
Clustering Features	Cluster, Triangles, Clustering Coefficient, Replier's Graph Clustering Coefficient, Tweet-Level Clustering Coefficient, Tweet-Level Replier's Graph Clustering Coefficient
Centrality Features	Degree Centrality, Closeness Centrality, Tweet-Level Degree Centrality, Replier's Graph Closeness Centrality, Tweet-Level Replier's Graph Closeness Centrality, Replier's Graph PageRank, Tweet-Level PageRank, Tweet-Level Replier's Graph PageRank
Psychology-Based Features	Familiarity, Scaled Familiarity, Creator Follower, Average Neighbor

Table 4.4: Final model AUC.

Class	**AUC**
Highly Susceptible	0.751
Slightly Susceptible	0.519
Neutral	0.697
Not Quite Susceptible	0.748
Not At All Susceptible	0.862
Average	0.715

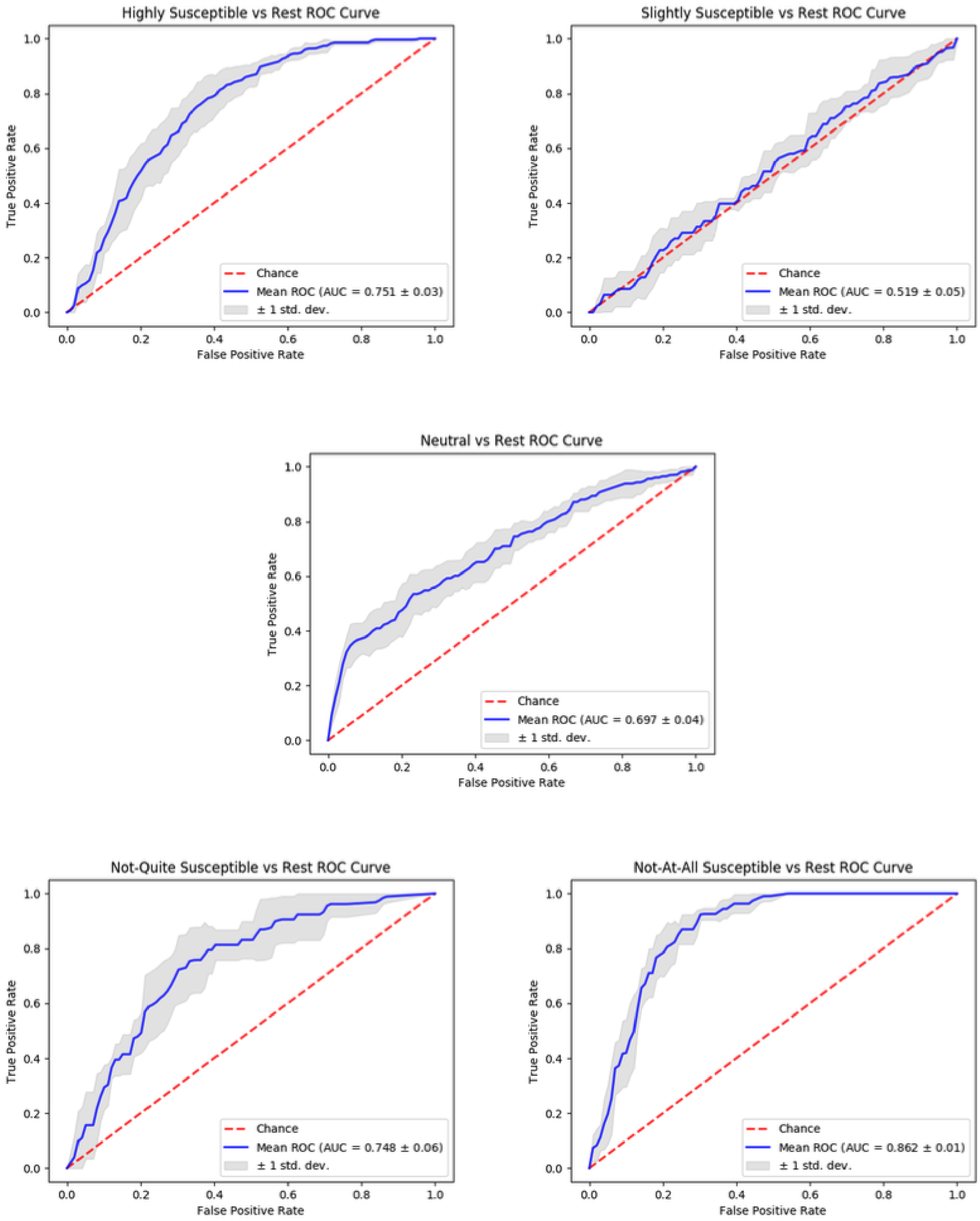


Figure 4.1: Final model ROC curves over the five folds.

The classifier struggled to successfully differentiate slightly susceptible users, only achieving 1.9% above random chance. The average AUC without slightly susceptible class was 0.764, showing the model's stronger performance on the other classes. However, most of the machine learning models in Section 4.2 had lower AUC values for the slightly susceptible class compared to other classes, with many models having performance worse than random chance. The only model that performed significantly better on the slightly susceptible class was the model trained with only psychology-based features, which achieved an AUC of 0.580, but had lower AUCs on the other classes compared to the final model.

The relative performance of each classifier is a metric of how different the average user from each class is. The final model had the highest AUC value (0.860) when classifying the not-at-all susceptible users. This indicates that there exists a difference in the friendship network between a not-at-all susceptible user and the other users. However, slightly susceptible users cannot be differentiated as confidently.

Chapter 5

Conclusion

5.1 Limitations and Future Work

There is inherent human error involved in the class labelling of the susceptibility dataset. Each annotator may have a different opinion of what constitutes a strong or weak agreement in a reply. Annotators may not have understood the context of the reply in the Twitter conversation. Also, annotators could have missed sarcasm in the tweet or lacked the knowledge to understand in-group references that could alter their label choice.

As a result of only 748 replies being assigned a clear majority label, the training dataset is small. Furthermore, the proposed model may not generalize well when tested with more data. In addition, the utility of the features selected in the final model may change with more data. The proposed model should be validated on a larger dataset to ensure its ability to generalize.

The friendship network is also restrained due to Twitter API limitations. Twitter imposes rate limits on the frequency of API calls. As a result, only immediate neighbors of repliers were extracted. A more expansive friendship network may uncover information that cannot be identified with only immediate neighbors, such as detecting communities within the network. A larger friendship network should be created in order to extract a better understanding of the networks that form around a fake news tweet, potentially improving the performance of the model.

5.2 Conclusion

Fake news undermines the trust in institutions with great importance to our democracy, news organizations and the government, while creating divisions in society. Social media has become the primary medium for fake news, and dissemination of fake news has been allowed to flourish on its platforms. While plenty of research exists related to the detection of fake news, little emphasis has been placed on understanding and predicting the users who fall victim.

This thesis found the following results:

1. It is possible to predict user susceptibility to fake news utilizing features derived from a user's friendship network.

2. Psychology-based features and centrality features play important roles in predicting user susceptibility.

3. The final prediction model, gradient-boosted trees trained on 27 features, achieved an average of 0.715 AUC.

Using fake news tweets related the Marjory Stoneman Douglas school shooting, a gradient boosting model was constructed to classify the level of susceptibility of a social media user, utilizing network-based features. The proposed model achieved moderate performance with an average AUC of 0.715, 21.5% higher than random chance.

In conjunction with traditional fake news detection models, the proposed model can be used to as part of an intervention strategy to reduce the spread and effects of fake news. The model can identify users that are at-risk or safe from fake news in social media. Then, social media companies can determine how to intervene for each level of susceptibility. Furthermore, this model can be integrated into an intervention protocol that social media giants can utilize. Because the model allows for variations of susceptibility to be gleaned for each user, an intervention strategy for this model can incorporate a sliding scale component. For example, the intervention could display warnings to a slightly susceptible user and hide a fake news tweet from a highly susceptible user. In a perfect world, this intervention could work to reach millions of social media users and protect them from falling into the traps of fake news outlets, thus protecting our democracy and institutions.

Bibliography

[1] Floyd H Allport and Milton Lepkin. 1945. Wartime rumors of waste and special privilege: why some people believe them. *The Journal of Abnormal and Social Psychology* 40, 1 (1945), 3.

[2] Tuan-Anh Hoang and Ee-Peng Lim. 2012. Virality and susceptibility in information diffusions. In *Sixth international AAAI conference on weblogs and social media*.

[3] Mohammad Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.

[4] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 8.

[5] Tero Karppi and Kate Crawford. 2016. Social media, financial algorithms and the hack crash. *Theory, Culture & Society* 33, 1 (2016), 73–92.

[6] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1103–1108.

[7] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[8] Gordon Pennycook and David G Rand. 2018. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. (2018).

[9] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. (2016).

[10] Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. 2017. From retweet to believability: Utilizing trust to identify rumor spreaders on Twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 179–186.

[11] Elisa Shearer. 2018. Social media outpaces print newspapers in the U.S. as a news source. `www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source`

[12] Tracy Jia Shen, Robert Cowell, Aditi Gupta, Yadav Amulya Le, Thai, and Dongwon Lee. 2019. How Gullible Are You? Predicting Susceptibility to Fake News. In *Proceedings of the 11th ACM Conference on Web Science*. ACM.

[13] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[14] Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)* 2, 4 (2012), 1951–1959.

[15] Laijun Zhao, Hongxin Cui, Xiaoyan Qiu, Xiaoli Wang, and Jiajia Wang. 2013. SIR rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications* 392, 4 (2013), 995–1003.

Robert Cowell

rh.cowell@gmail.com

Education

The Pennsylvania State University, Schreyer Honors College **May 2019**
Bachelor of Science in Computer Science University Park, PA
Relevant Coursework: Artificial Intelligence, Machine Learning, Data Mining

Professional Experience

Capital One **June 2018 – August 2018**
Software Engineering Intern McLean, VA

- Optimized natural language generation in transition from prototype to production for Eno intelligent assistant
- Developed data parser, logical validator and editor interface in Python for data compatibility and integration between intelligent assistant response-designer tool and natural language generation
- Extended functionality of intelligent assistant end-to-end testing tool UI using Angular for developer debugging
- Converted NLG's embedded database from Pandas DataFrame to SQLite in-memory database and refactored response content-resolving code to dynamically call only relevant handlers, increasing response speed by 400%

Optum **June 2017 – August 2017**
Software Engineering Intern Basking Ridge, NJ

- Developed natural language search engine for enterprise application analytics platform to provide insightful results without requiring technical knowledge, using Python, Rasa NLU and Elasticsearch
- Implemented natural language understanding and machine learning functionality using Rasa NLU to provide robust intent recognition, keyword extraction, and easy extensibility for future development
- Reduced the gap between complicated big data platform and business users and support for queries too complex or abstract for previous search engine

Research Experience

PIKE Research Group **April 2017 – Present**
SysFake Project

- Developed applications to create machine learning datasets to model a Twitter user's susceptibility to fake news
- Created social graph extractor to collect relationships of users who interacted with fake news for graph-based features using Python and Twitter API, and deployed application to AWS EC2 Instance
- Designed concurrent Twitter scraper using Python and Selenium to collect user information and tweet history

Hackathon

JPMorgan Chase & Co. Code for Good **October 2016**

- Developed a digital data collection system for fieldworkers in regions with limited access to technology to aid Free the Slaves, a nonprofit dedicated to eradicating modern-day slavery in Africa and Asia
- Created an automated SMS survey, utilizing Python and Twilio, to provide instantaneous data aggregation of fieldworkers' reports in areas with only basic cellular service and areas without smartphones

Skills

Python | Natural Language Processing | Machine Learning