

5-2023

Identifying Social Media Users that are Susceptible to Phishing Attacks

Zoe Metzger
William & Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorsthesis>



Part of the [Data Science Commons](#), [Other Computer Sciences Commons](#), and the [Social Media Commons](#)

Recommended Citation

Metzger, Zoe, "Identifying Social Media Users that are Susceptible to Phishing Attacks" (2023).
Undergraduate Honors Theses. William & Mary. Paper 2013.
<https://scholarworks.wm.edu/honorsthesis/2013>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

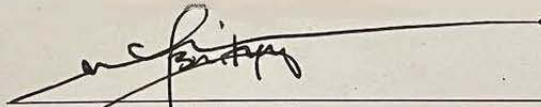
Identifying Social Media Users that are Susceptible to Phishing
Attacks

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Data Science from
William & Mary

by

Zoe Anne Metzger

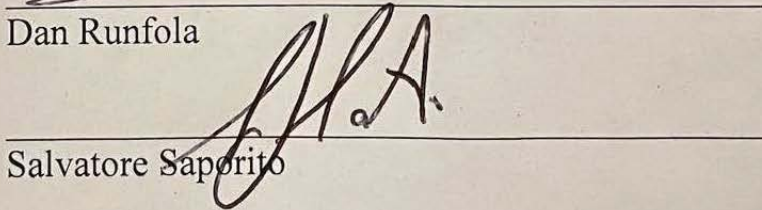
Accepted for Honors
(Honors, High Honors, Highest Honors)



Alexander C. Nwala



Dan Runfola



Salvatore Saporito



WILLIAM & MARY

CHARTERED 1693

THE COLLEGE OF WILLIAM & MARY

HONORS THESIS

**Identifying Social Media Users that are Susceptible to Phishing
Attacks**

Author:

Zoe A. METZGER

Advisor:

Alexander C. NWALA

*A thesis submitted in fulfillment of the requirements for
Honors in the degree of Bachelors of Science in the*

Data Science Program

Williamsburg, Virginia

April 21, 2023

THE COLLEGE OF WILLIAM & MARY

Abstract

Dr. Alexander C. Nwala
Data Science Program

Bachelors of Science

Identifying Social Media Users that are Susceptible to Phishing Attacks

by Zoe A. METZGER

Phishing scams are a billion-dollar problem. According to Threatpost, in 2020, business email compromise phishing attacks cost the US economy \$ 1.8 billion. Social media phishing scams are also on the rise with 74% of companies experiencing social media attacks in 2021 according to Proofpoint. Educating users about phishing scams is an effective strategy for reducing phishing attacks. Despite efforts to combat phishing, the number of attacks continues to rise, likely indicative of a reticence of users to change online behaviors. Existing research into predicting vulnerable social media users that are susceptible to phishing mostly focuses on content analysis of their posts or the users they interact with, and not their behaviors. In contrast, in this research, we study the online behaviors of social media users on Twitter to identify those that are susceptible to phishing attacks. Specifically, we analyzed the behaviors of social media users that succumb to phishing scams in comparison to a control group of users that did not, to identify behavioral patterns that distinguish them. Online actions encompass aspects such as liking and sharing habits, the nature of posts, duration of engagement in posting activities, among others. We classified control and susceptible users based on their page metrics with a KNN model (F1: 0.897) and also based on user behavioral metrics with a logistic regression model (F1 score: 0.903).

Contents

Abstract	i
1 Thesis	1
1 Introduction	1
2 Literature Review	2
3 Data	3
3.1 Political Dataset	4
3.2 Financial Dataset	5
3.3 Page Metrics	6
3.4 Posting Behaviors	7
4 Methods	8
4.1 Page Metrics	9
4.1.1 Visualizing Susceptible vs. Control Users	9
4.2 BLOC Metrics	13
5 Results	14
5.1 Page Metrics	14
5.1.1 Linear Regression	14
5.1.2 Classification Models on Page Metrics	14
5.2 BLOC	17
5.2.1 PCA Analysis	17
5.2.2 Classification Models on BLOC Metrics	19
6 Discussion	21
7 Conclusion	22
8 Acknowledgements	24
A Appendix	25
1 Control Tweet IDs	25
2 BLOC Alphabet	26

References

List of Figures

- 1.1 A screenshot of a simulated phishing scam posted by Bossetta, 2018 to lure potential susceptible users. The users clicked on this post and filled out the form in order to be labeled as susceptible (Bossetta, 2016) 5
- 1.2 A screenshot of a post from the New York Times used to collect control users for the political dataset. The post discusses the election in non-partisan verbiage and includes a link. 5
- 1.3 A screenshot of a post used to collect susceptible users. The users labeled as susceptible retweeted or liked this post. The posts included cashtags of both popular and obscure ticker symbols (Cresci et al., 2018b) 6
- 1.4 Screen grab of a post used to collect the control users. The post contains cashtags of only reputable ticker symbols. 7
- 1.5 Representation of different BLOC strings for different kinds of behaviors (Nwala, Flammini, and Menczer, 2022) 8
- 1.6 A scatter plot for the political dataset of users’ log following versus age of account. The control and susceptible users are labeled by color. 9
- 1.7 A scatter plot for the financial dataset of users’ log following versus age of account. The control and susceptible users are labeled by color. 10
- 1.8 CCDF and Histogram of Log Following for the political dataset . 10
- 1.9 CCDF and Histogram of Log Following for the financial dataset . 11
- 1.10 CCDF and Histogram of Log Followers for the political dataset . 11
- 1.11 CCDF and Histogram of Log Followers for the financial dataset . 11
- 1.12 CCDF and Histogram of Age for the political dataset 12
- 1.13 CCDF and Histogram of Age for the financial dataset 12

1.14 PCA plot for the political BLOC dataset with 74 BLOC features. . .	18
1.15 PCA plot for the financial BLOC dataset with 97 BLOC features. . .	18
1.16 PCA plot for the political and financial BLOC datasets with 73 BLOC features.	19

List of Tables

1.1	Most frequent BLOC words in the financial (Fin.) and political (Pol.) datasets split by control and susceptible users. Some notable differences across both susceptible and control users are in bold red text.	13
1.2	Table for the political dataset comparing classes for KNN	15
1.3	Table for the financial dataset comparing classes for KNN	16
1.4	Table for the political dataset comparing classes for Logistic Regression	16
1.5	Table for the financial dataset comparing classes for Logistic Regression	17
1.6	Table for the political BLOC word dataset comparing classes for KNN	20
1.7	Table for the financial BLOC word dataset comparing classes for KNN	20
1.8	Table for the political BLOC word dataset comparing classes for Logistic Regression	21
1.9	Table for the financial BLOC word dataset comparing classes for Logistic Regression	21

List of Abbreviations

API	Application Programming Interface
CCDF	Complementary Cumulative Distribution Function
KNN	K-Nearest Neighbors
PCA	Principal Component Analysis
SNS	Social Network Sites

Chapter 1

Thesis

1 Introduction

Online users continue to make security mistakes or give away sensitive information which costs individuals, companies, and organizations billions of dollars. On average, a single data breach costs a company over \$4 million (Henriquez, 2023). Succumbing to a phishing scam is easy and could be as simple as clicking a link in an email, text message, or even a twitter post. Cybersecurity specialists continue to struggle with completely preventing and blocking the multiple forms phishing scams. This includes phishing, spear phishing (the attacker poses as an acquaintance), whaling (attacker targets a high-ranking executive), vishing (attacks over the phone), smishing (attacks over sms messaging), and more. Education has been identified as one of the best ways to prevent successful attacks (Sheng et al., 2010). Identifying susceptible individuals is often based on recognizing specific personality traits and analyzing user content.

In this research, we sought to find a new way to identify a user's potential susceptibility to phishing scams. We specifically investigated Twitter users. Instead of trying to identify a social media user's susceptibility through a content analysis of their posts and descriptions, we analyzed the behaviors of social media users who succumbed to phishing scams in comparison to a control group of users who did not, to identify behavioral patterns that distinguish them. This research seeks to answer: Can we identify a Twitter user's potential phishing susceptibility based on their page and behavioral metrics?

By utilizing data from prior research, the Twitter API, page metrics, and behavioral metrics, we were able to provide insight into the behaviors which distinguish susceptible users who succumbed to phishing scams in comparison to

a control group of users who did not. Page metrics are accessible metrics which appear on a users profile such as, followers and following. Behavioral metrics look at the kind of behaviors a user does on the twitter platform such as posting and retweeting.

This paper is organized as follows. In Section 2 we discuss the gap in literature. Section 3 we explain our data collection and cleaning. Section outlines our methodology followed by the analysis in Section 5. Finally, the paper is concluded with a discussion and conclusion in Sections 6 and 7, respectively.

2 Literature Review

Social engineering frequently claims victims both on and off the internet. Social engineering can be defined as the process in which a bad actor manipulates a victim into giving away sensitive and/or valuable information. A common form of social engineering that occurs through communications online or over the phone. Phishing can be defined as the process in which a bad actor disguises their attempt at stealing information through a seemingly trustworthy digital communication. For example, many phishing attempts occur over email when bad actors get their victims to click on links and/or provide personal or business information. With this stolen information in hand, bad actors might be able to break into additional online resources of their victims and steal sensitive information such as credit card and banking details. Researchers agree, education is the best way to prevent phishing susceptibility (Sheng et al., 2010). Wang et al. showed an increase in scam knowledge causes users to be less likely to fall for phishing (Wang et al., 2012).

Two main areas of study that investigate how to identify users that are susceptibility to phishing include psychological and content analysis. The content analysis research has addressed user Facebook or Twitter profiles. Analysis of the Big Five personality traits shows a correlation between having the traits of extroversion, agreeableness, openness, conscientiousness, and neuroticism and higher levels of susceptibility (Tornblad et al., 2021). Golbeck, Robles, and Turner, 2011 created a bridge for personality to social media susceptibility by determining Facebook users' personalities based on their profile information. Frauenstein and Flowerday went beyond only determining personalities

on social media, but also investigated the relationship between the perceived personality and their susceptibility to phishing. Specifically, they investigated the relationship between user personality and the kinds of posts they interacted with to understand what traits were more susceptible to phishing (Frauenstein and Flowerday, 2020). Accordingly, they concluded that extroversion and conscientiousness, may lead to phishing susceptibility. While research into personality traits that correlate with susceptibility to phishing are valuable, implementing such analyses is not always possible especially since quantifying personality traits online is challenging.

Social media manipulation which could be seen as a form of phishing, was frequently discussed in relation recent US elections(Ratkiewicz et al., 2021). The research has illustrated how to identify potential bots looking to manipulate users' political views (Ferrara et al., 2020). Even though social media manipulation is a form of social engineering, we focus specifically on identifying phishing susceptibility. By using a neural network model, Razaque et al., 2021 identified safe versus malicious clickbait content. Malicious clickbait often leads to obtaining information about the user or access to their computer. However, the research into susceptibility is limited and not based on behavioral metrics.

A major contribution of this effort is to investigate the online behaviors (vs. personality or content) of user that are susceptible to phishing. A benefit of this approach is that we do not have to make many assumptions on the nature of behaviors. Investigating metrics such as usage and interaction frequency is informed by research that discusses how habitual use and frequent exposure to content correlates with higher rates of susceptibility.

3 Data

In this section, we provide details as to the data used in this analysis. It can be summarized as a dataset focused on individuals who have interacted with financial or political information. We needed to extract the posts of both users who are susceptible to phishing (which we call susceptible users) and a control set to better understand if there are differences between them. Our dataset of susceptible users was provided by authors studying both political and financial social media manipulation. In the next two sections we describe these datasets

and the respective control datasets we generated. We then collected 11 control posts and the users who interacted with them. We moved on to try and understand which metrics we wanted to collect and what they would mean in terms of potential susceptibility. Finally, we collected each user's BLOC, a string representation of a user's posts, to compare the behaviors of each user.

3.1 Political Dataset

The political dataset includes accounts of users that succumbed to a phishing scheme with a political focus, while the control users includes those who interacted with trustworthy political content (Bossetta, 2018). It includes Twitter users who interacted with the spear phishing bots from *A Simulated Cyberattack on Twitter: Assessing Partisan Vulnerability to Spear Phishing and Disinformation ahead of the 2018 U.S. Midterm Elections* (Bossetta, 2018). This research sought to identify a potential partisan divide between Twitter users who succumbed to a simulated phishing scheme (see Figure 1.1) in October 2018 before the 2018 US midterm elections. It is worth noting that the simulated phishing scams included politically neutral headlines in order to avoid biasing the dataset. Even though the original political dataset includes 197 users, only 106 users were still active as at the time of this writing.

We generated a control dataset of users for the political dataset by collecting the users who interacted with legitimate neutral political content during October 2018. We extracted control users by collecting the accounts of users who interacted with four posts from The New York Times, The Wall Street Journal, and The Hill. These were chosen because they report on the election using non-partisan verbiage. In addition, all three accounts are verified news sources according to Media Bias Fact Check which assesses the political bias of news sources (Zandt, 2015). Media Bias Fact Check gives a score out of ten based on four categories. The categories are biased wording/headlines, factual/sourcing, story choices, and political affiliation. The Hill received a label of least biased. The Wall Street Journal received a right-center bias, . so the New York Times which has a left-center bias was selected to balance it out. Ultimately, this process resulted in the selection of a four posts (the Tweet IDs for these posts are



FIGURE 1.1: A screenshot of a simulated phishing scam posted by Bossetta, 2018 to lure potential susceptible users. The users clicked on this post and filled out the form in order to be labeled as susceptible (Bossetta, 2016)

in Appendix Section 1). An example of one of the control posts can be seen in Figure 1.2.



FIGURE 1.2: A screenshot of a post from the New York Times used to collect control users for the political dataset. The post discusses the election in non-partisan verbiage and includes a link.

3.2 Financial Dataset

Similar to the political dataset, the financial dataset was provided by Tardelli et al. who studied users that interacted with phishing posts as part of *Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter* (Cresci

et al., 2018a). Specifically, the data was created in 2017 by collecting users who interacted with stock-market based phishing posts that were amplified by bots (e.g., Figure 1.3). It included 1,311 active non-bot susceptible users.



FIGURE 1.3: A screenshot of a post used to collect susceptible users. The users labeled as susceptible retweeted or liked this post. The posts included cashtags of both popular and obscure ticker symbols (Cresci et al., 2018b)

We generated a control set for the financial dataset by collected the accounts of users that interacted with one of the seven financial-related posts (e.g., 1.4) from a reputable source. The control posts contained stock commentary and listed their ticker symbols. The Twitter account, @Bespoke, was listed on Investopedia as one of the top 10 Twitter feeds investors should follow. In total, we extracted 386 control users who either liked, retweeted, or quoted a control post (Parker, 2022).

3.3 Page Metrics

Informed by previous research, we extracted page metrics for each susceptible and control user for both datasets. The page metrics represent an overview of a users behaviors on social media. These metrics include *followers*, *following*, *created at*, and *tweet count*. These metrics summarize a user's behaviors on the platform. We can understand how many people they interact with, how many posts they create, and how long they have used the account on this platform. Each metrics could provide useful signals for identifying susceptible users and could help approximate personality traits (Tornblad et al., 2021). For example, the



FIGURE 1.4: Screen grab of a post used to collect the control users. The post contains cashtags of only reputable ticker symbols.

number of followers and following can represent extroversion (Golbeck, Robles, and Turner, 2011 and Seidman, 2020). Extroversion has been identified more often in susceptible users (Tornblad et al., 2021). The variable *created at* which represents when the account was created can be used to calculate the age of an account. The *age* variable can be used to calculate duration of access to the Twitter platform. Heartfield et al. state users who access a platform for longer amounts of time were less susceptible to phishing (Heartfield, Loukas, and Gan, 2016). The more familiar a user is with a platform allows them to potentially be better at identifying when something is not right or a scam. Additionally, tweet count which represents the number of tweets a user has posted could help approximate excessive use or internet addiction which is also linked to higher susceptibility (Tornblad et al., 2021).

3.4 Posting Behaviors

Collecting and analyzing users posting behaviors allows us to better understand a pattern of usage for each user. Previous research has not investigated these

4.1 Page Metrics

We began by plotting various page metrics to see if any separated the control and susceptible groups.

4.1.1 Visualizing Susceptible vs. Control Users

Figures 1.6 and 1.7 plots the age vs. log of the number of followers for the political and financial datasets. At first glance, the scatter plots, as seen in Figures 1.6 and 1.7, do not show a significant difference between the susceptible and control groups. However, looking more closely at the scatter plots for log followers versus age, we can see there are a few susceptible users with lower numbers of log followers and age than the control group. We also see a potential cluster of control users with a higher number of log followers and older accounts. The control observation is more significant in the plot for the political dataset and the susceptible observation is more significant in the financial dataset plot.

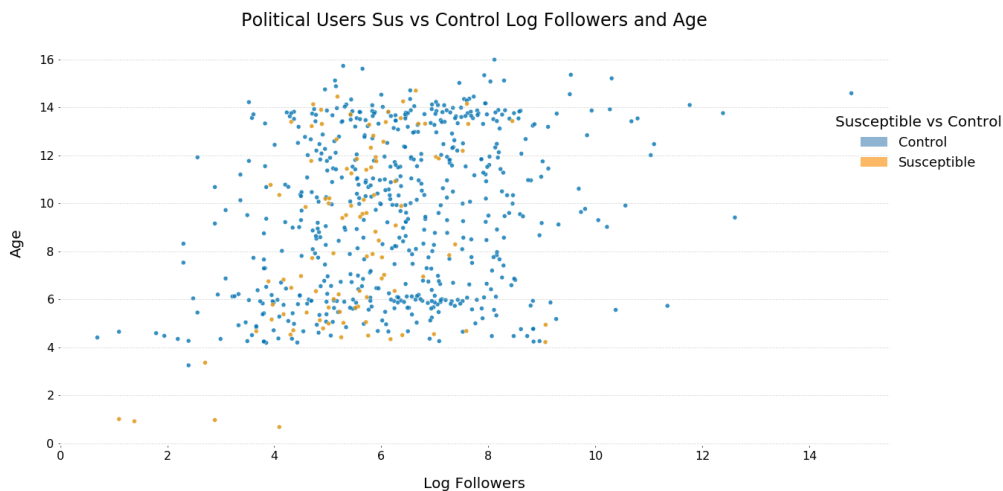


FIGURE 1.6: A scatter plot for the political dataset of users' log following versus age of account. The control and susceptible users are labeled by color.

CCDF displays the probability the users have the number of the metric or more. For example, if we look at Figure 1.8 we can see 100% of our users follow 1 account or more. The CCDF plots are used to expand upon the potential differences seen in the scatter plots. CCDF plots show the proportion of the users represented by the given value or less. Figure 1.8 and Figure 1.9 demonstrate

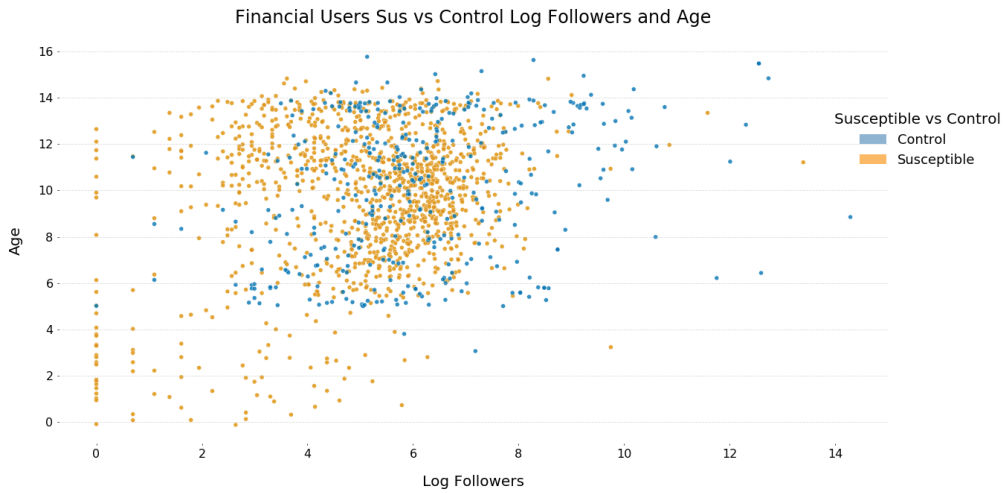


FIGURE 1.7: A scatter plot for the financial dataset of users' log following versus age of account. The control and susceptible users are labeled by color.

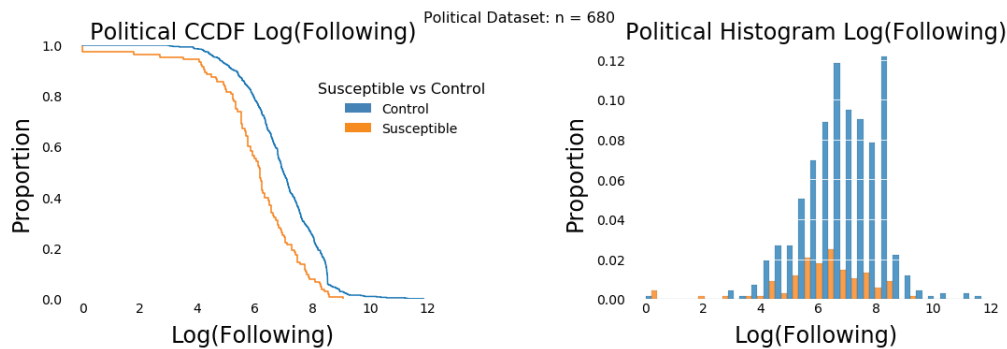


FIGURE 1.8: CCDF and Histogram of Log Following for the political dataset

that 50% of the users' number of log following is lower for susceptible users than the control users. This holds for both the financial and political datasets. Comparable results are shown for log followers in Figure 1.10 and 1.11. Plotting age demonstrates a trend for a subsection of the susceptible users which have younger accounts. Conversely, we see a trend for a subsection of the control groups which have older accounts (see Figures 1.12 and 1.13). The CCDF analysis shows different results than expected based on the literature about personality, susceptibility, and social media page metrics.

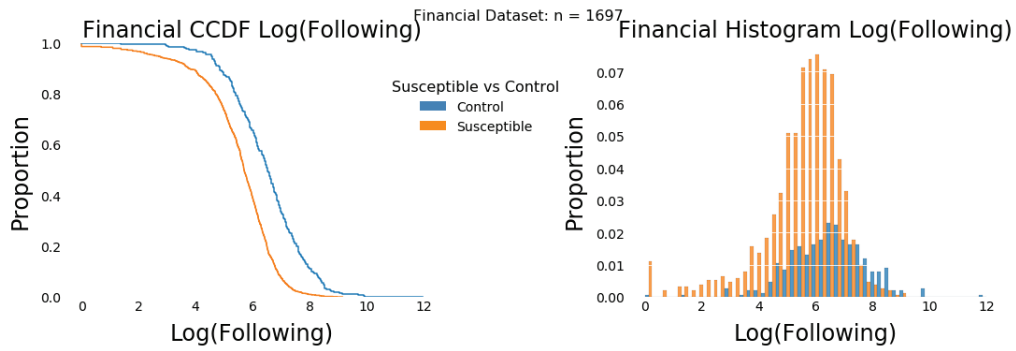


FIGURE 1.9: CCDF and Histogram of Log Following for the financial dataset

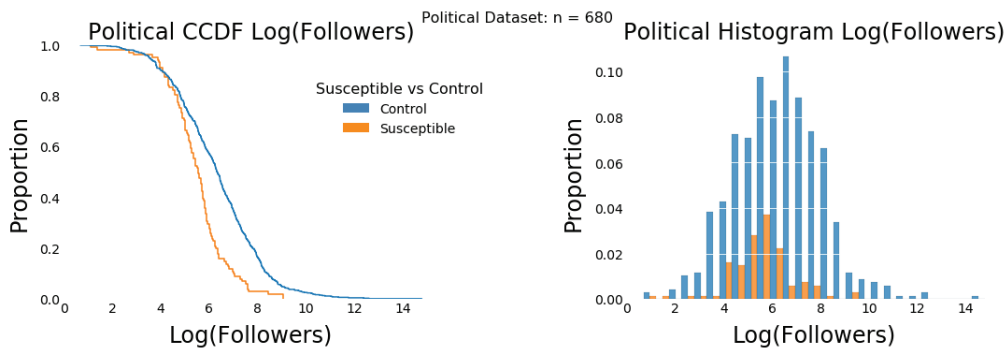


FIGURE 1.10: CCDF and Histogram of Log Followers for the political dataset

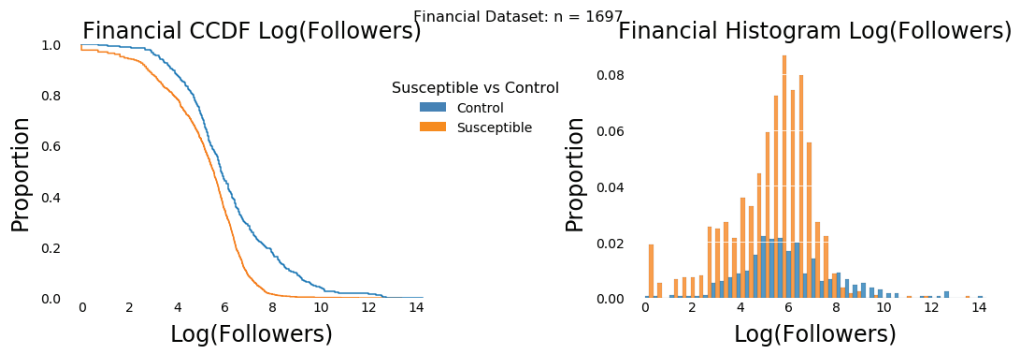


FIGURE 1.11: CCDF and Histogram of Log Followers for the financial dataset

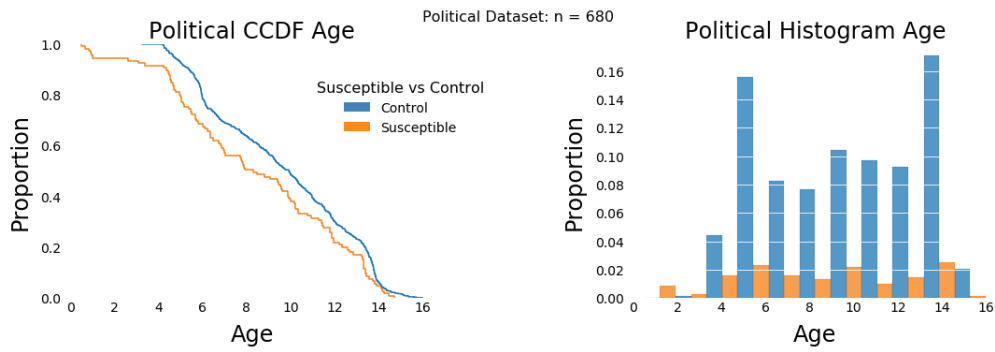


FIGURE 1.12: CCDF and Histogram of Age for the political dataset

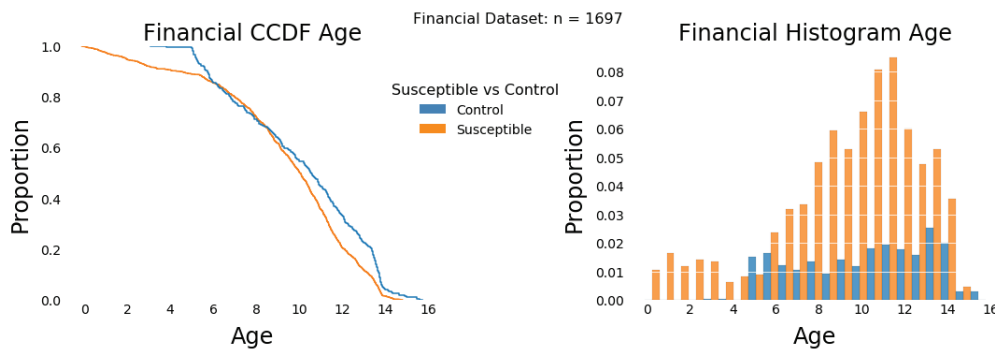


FIGURE 1.13: CCDF and Histogram of Age for the financial dataset

4.2 BLOC Metrics

Table 1.1 presents the top 15 BLOC words for each the political (Pol.) and financial (Fin.) susceptible and control users. Each BLOC word represents a different kind of action or content behavior. Table 1.1 (see Appendix Section 2) is a dictionary that explains the meaning of the symbols. The political dataset produced a vocabulary of 74 unique BLOC words and the financial dataset produced 97. We collected the BLOC words for tweets posted by users within the time frames the susceptible users were collected for both the political (Oct 1, 2018 to Oct 31, 2018) and the financial datasets (May 1, 2017 to Sep 31, 2017). We had to provide BLOC with the tweets collected for each user. Our original pull for data only required users to interact with phishing or control posts; they did not have to post themselves. As a result, we had users who had not posted within their respective time frames, so our datasets shrunk. For BLOC analysis we had 90 control and 312 susceptible users from the financial dataset and 86 control and 12 susceptible users from the political dataset.

Rank	Fin. Control	Fin. Susceptible	Pol. Control	Pol. Susceptible
1	T	T	T	T
2	t_w	mt	mt	t_d
3	mt	t_d	t_w	t_{min}
4	t_d	TT	t_d	mt
5	t_h	t_{min}	t_h	t_w
6	mUt	t_h	t	t_h
7	t	t_w	p	p
8	p	Emt	t_{min}	mmt
9	t_{min}	mUt	mmt	t
10	Emt	t	mUt	Hmt
11	TT	mqt	mqt	Ht
12	t_m	mmt	qt	mUt
13	Hmt	Emqt	Hmt	mmmt
14	mmt	Hmt	TT	mqt
15	mqt	EHmt	Emt	TT

TABLE 1.1: Most frequent BLOC words in the financial (Fin.) and political (Pol.) datasets split by control and susceptible users. Some notable differences across both susceptible and control users are in bold red text.

5 Results

Here we investigate the degree to if linear regression or machine-learning models trained on page and BLOC metrics could help distinguish susceptible users from control. These models help to explain behavioral metrics can be useful in determining a users' potential susceptibility.

5.1 Page Metrics

5.1.1 Linear Regression

We considered creating a linear regression model in order to have a model to predict if a user is likely to be susceptible. The linear regression model also provides a way to understand the potential significance of the different page metrics. Linear regression models give weight to each variable to create a model to predict the y variable. In our case, the y variable is the type, or whether the user is susceptible or not. The x variables provided were the page metric variables for each user. Linear regression makes a model only based on the data it is provided. We ran multiple linear regression models with different interaction variables. To create the most accurate linear regression models, we ran stepwise regression models with 2 variable interactions. The models produced for the political and financial data both resulted in R-squared values less than 0.5 which indicates the model does do a good job of approximating the data. The models created indicated that approximately 50% of the variance in the data could be explained by the metrics generated.

5.1.2 Classification Models on Page Metrics

We trained KNN and Logistic regression models on the page metrics features (i.e., *followers*, *following*, *created at*, and *tweet count*). We split the data using the `train_test_split` function provided by Sckit Learn. The function allows us to split the data into a training and testing group to be utilized in the models. we set the split to be equal for the training and testing groups. In order to promote more efficient optimization, all variables were normalized using a z-score normalization approach called standard scalar (Scikit Learn, [n.d.](#)). We applied 10-fold cross-validation to evaluate the performance of the classification models.

Before understanding and calculating the correctness of the models, we wanted to make sure the models were performing with little variance between the training and test data. The training data is given to the machine learning model to train it and the test data is used to check the models accuracy. Using 10-fold validation we calculated the mean training and test scores. The training score represents how well the model learned from the training data. The test score represents how the model works against data it has not seen before. If both of these scores are high and have low variance, the model has been tuned to be a best fit.

The KNN model for the political dataset received a training score of 85.72% and a test score of 85.29%. This low variance shows this model is a good fit for the data however this does not represent accuracy. To understand the correctness of the model, we looked at the precision, recall, and F1 scores. The KNN model for the political dataset produced a precision and recall score of 0.200 and 0.020 respectively. When we compare the classes for the models, we noticed the imbalance between the number of control and susceptible users. As a result, an F1 score would provide a better understanding of the model's performance. F1 is the harmonic mean between accuracy and recall. The F1 score represents the accuracy at detecting a true positive. In our case a true positive is correctly classifying a susceptible user. The F1 accuracy for the political dataset is 0.037 which is very poor. We also wanted to test the models against a baseline, so we calculated the F1 score if we assumed all the users were susceptible. The F1 score for the political dataset is 0.252. Our KNN model under performs this baseline test.

KNN Political Dataset	Actual Control	Actual Susceptible
Predicted Control	287	48
Predicted Susceptible	4	1

TABLE 1.2: Table for the political dataset comparing classes for KNN

The KNN model for the financial dataset produced a mean train score of 82.90% and a test score of 79.90%. The high scores and low variance represent the model is a good fit of the data. Knowing the model has been tuned well, we can look at the correctness of the model. The financial dataset produced a model with a precision and recall score of 0.849 and 0.910 respectively. The F1 score for

the financial dataset is 0.878. The KNN model for the financial dataset appears to be a correct model for detecting susceptible users. Once again, we calculated the F1 score given the assumption all users are identified as susceptible. The baseline model received an F1 score of 0.889. The KNN model therefore does not outperform the baseline model.

KNN Financial Dataset	Actual Control	Actual Susceptible
Predicted Control	60	61
Predicted Susceptible	110	618

TABLE 1.3: Table for the financial dataset comparing classes for KNN

We once again looked at the training and test scores for the logistic regression models before looking into their correctness at identifying susceptible users. The training and test scores for the political dataset were both 84.56%. The logistic regression model is therefore a good fit of the data. The model for the Logistic Regression algorithm for the political dataset was unable to identify any users as susceptible. Therefore the precision and recall scores are both 0. The F1 score for the political dataset is 0 which is again lower than the baseline model F1 score of 0.252.

LR Political Dataset	Actual Control	Actual Susceptible
Predicted Control	291	49
Predicted Susceptible	0	0

TABLE 1.4: Table for the political dataset comparing classes for Logistic Regression

The training score was 80.00% and the test score was 79.73% for the financial dataset. This model is also a good fit of the data. The financial dataset produced a model with a precision and recall score of 0.822 and 0.987 respectively. We once again looked at F1 to understand the model. The F1 score for the financial dataset is 0.897. The logistic regression model for the financial dataset appears to be the most representative model using page metrics to classify susceptible users. The logistic regression model performs better than the baseline model

with a 0.889 F1 score. Therefore, this model is more useful than just assuming all of the users are susceptible.

LR Financial Dataset	Actual Control	Actual Susceptible
Predicted Control	25	9
Predicted Susceptible	145	670

TABLE 1.5: Table for the financial dataset comparing classes for Logistic Regression

5.2 BLOC

Interestingly, Table 1.1 suggests that susceptible users from the financial dataset do not engage in conversations (p) as often as their corresponding control counterparts since the BLOC p word is not part of the ranked list. Also, these users are more likely to posts tweets in short bursts (TT - Rank 4). Moreover, the control users also pause longer between posts compared to the susceptible users as seen by the ranking of t_w (pause for days).

Unlike the financial dataset, the top 15 BLOC words of the political dataset susceptible and control users are more identical. However, similar to the financial dataset, the control users are more likely to have longer pauses between their tweets (e.g., t_w - Rank 3 vs. t_{min} - Rank 3)

5.2.1 PCA Analysis

Table 1.1 outlines only the top 15 BLOC words of the susceptible and control users for both datasets. Here, we visualize the entire vocabulary generated by BLOC for both datasets by applying PCA to reduce the dimension of the BLOC vectors of users. Figures 1.14 and 1.15 show the PCA plots for the political and financial datasets, respectively. The result of projecting our data into 2 dimensions does not provide obvious clusters of users but instead demonstrates significant overlapping as well as several outliers. The explained variance was only 17% for both of these plots.

Figure 1.16 is a PCA plot which includes both datasets. Due to the shrinkage of the data, we wanted to see if applying PCA to more data would prove to be

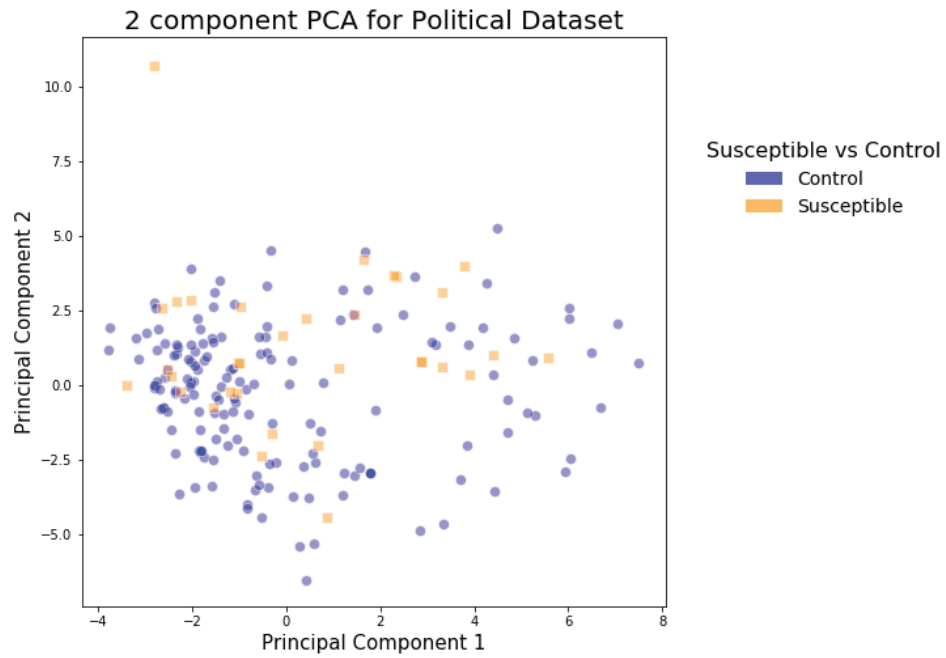


FIGURE 1.14: PCA plot for the political BLOC dataset with 74 BLOC features.

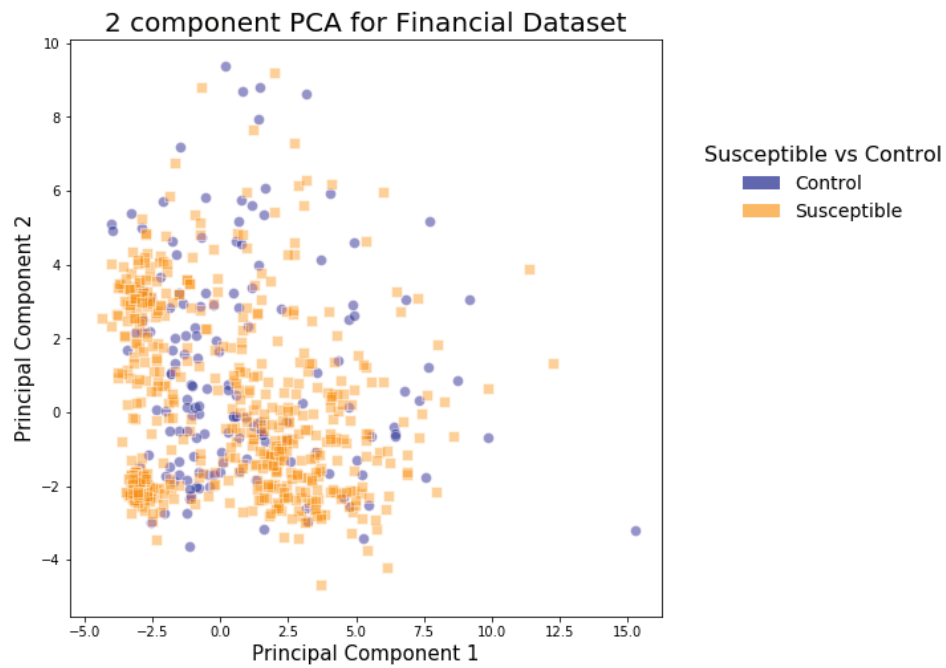


FIGURE 1.15: PCA plot for the financial BLOC dataset with 97 BLOC features.

more telling. In this PCA plot we begin to see clustering of control users and two more distinct clusters for susceptible users. The explained variance is 19% which is still low, but larger than the datasets plotted separately. This leads us to believe that with the addition of more data and using PCA or other clustering techniques, it may be possible to better model susceptibility with BLOC words.

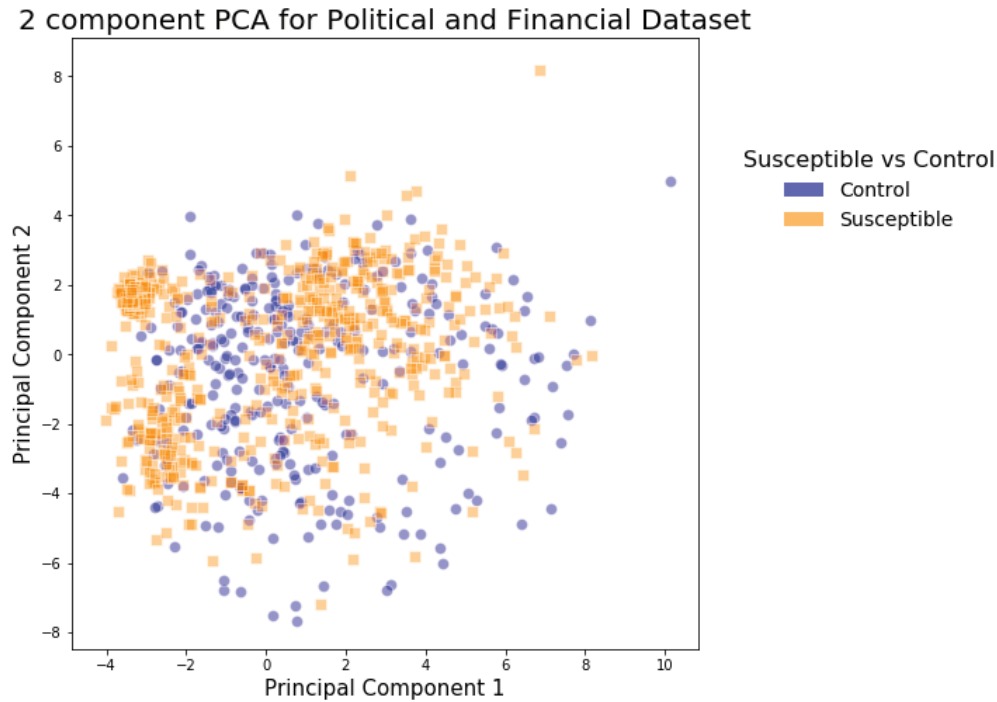


FIGURE 1.16: PCA plot for the political and financial BLOC datasets with 73 BLOC features.

5.2.2 Classification Models on BLOC Metrics

To test the additional value that BLOC metrics may have in promoting better classifications, we provided them to both the KNN and logistic regression classifiers. First we checked the fit and tuning of our models by looking at the training and test scores. The training score for the political dataset using KNN was 84.84% and the test score was 82.32%. This model represents a good fit of the data. The KNN model for the political dataset produced a precision and recall scores of 0. The F1 score for the political dataset is also 0. This model, therefore, does not correctly identify susceptible users. The F1 score for the political

dataset with BLOC baseline was 0.218. Once again, the baseline is created by calculating the F1 assuming all users are labeled as susceptible.

KNN Political BLOC Dataset	Actual Control	Actual Susceptible
Predicted Control	83	12
Predicted Susceptible	3	0

TABLE 1.6: Table for the political BLOC word dataset comparing classes for KNN

For the financial dataset KNN model we recieved a training score of 86.32% and a test score of 84.22%. This model is a also a good fit of the data provided. The financial dataset produced a model with a precision and recall score of 0.802 and 0.997 respectively. The F1 score for the financial dataset is 0.889. The KNN model for the financial dataset appears to be most accurate at detecting susceptible users based on their behavioral activities. The KNN model has a higher F1 score than our baseline assumption model which received a F1 score of 0.874.

KNN Financial BLOC Dataset	Actual Control	Actual Susceptible
Predicted Control	13	1
Predicted Susceptible	77	311

TABLE 1.7: Table for the financial BLOC word dataset comparing classes for KNN

We also used the logistic regression algorithm to create models for the political and financial datasets. The financial dataset produced a training score of 95.66% and a test score of 84.29%. This model has been tuned to be a good fit of the data. The model for the Logistic Regression algorithm for the political dataset resulted in precision and recall scores of 0.261 and 0.500 respectively. We once again looked at F1 to understand the model. The F1 score for the political dataset is 0.343 which is higher than our baseline model with a F1 score of 0.218. This is the most correct model for identifying susceptible users using the political dataset.

The logistic regression model for the financial dataset produced a training score of 93.26% and a test score of 87.82%. This model is also a good fit of the

LR Political BLOC Dataset	Actual Control	Actual Susceptible
Predicted Control	67	6
Predicted Susceptible	17	6

TABLE 1.8: Table for the political BLOC word dataset comparing classes for Logistic Regression

provided data. The financial dataset produced a model with a precision and recall score of 0.898 and 0.907 respectively. The F1 score for the financial dataset is 0.903 which is higher than the baseline F1 of 0.874. Therefore the model is more useful at predicting a users' susceptibility than assuming they are susceptible. The logistic regression model on the financial dataset proves to be the best model for identifying control and susceptible users when we take into consideration a holistic representation of both the page and BLOC metrics. Using the users' behavioral metrics through BLOC, allowed us to produce the models with the highest F1 scores for both datasets.

LR Financial BLOC Dataset	Actual Control	Actual Susceptible
Predicted Control	58	29
Predicted Susceptible	32	283

TABLE 1.9: Table for the financial BLOC word dataset comparing classes for Logistic Regression

6 Discussion

The findings did not appear to demonstrate significant accuracy in distinguishing between the Twitter behaviors of susceptible users and control users. However, there are some interesting findings as a result of this research. We were able to identify potential differences between the page metrics of users through the CCDF plots of log following, log followers, and age. These differences were most noticeable for subgroups of the datasets. The subgroups are represented by the tails of the histograms for the log following, log follower, and age histograms. The KNN and logistic regression models proved to not tell the whole story. For future research, it would be important and interesting to look at the

proportions of followers, following, and tweet count in terms of age. By normalizing these variables, would we be able to see clearer results and differences between the control and susceptible users?

The BLOC metrics demonstrated a few different behavioral patterns between the susceptible and control users when looking at both the political and financial dataset combined. The KNN and logistic regression models were unable to recognize susceptible users from control users. The PCA plot demonstrates some potential classification of different users who may become more apparent with a large set of user data. The most promising finding was using logistic regression to assess the actions of users from the financial dataset. This model was able to determine identifiable differences between the twitter actions between the control and susceptible users with an accuracy of 87.82% shown by the mean test score. Using BLOC sparked more questions about how patterns of usage can better be investigated and utilized. How would separating the types of BLOC words impact predictive modeling? For example, what would a model with only time between posts look like?

Everyone is at risk of falling for a phishing scam. Whether it be a scam pushed out by an organization to help educate us or a simple text which claims your credit card has been stolen, we all see and are susceptible to clicking on these links. Does this research demonstrate we are all equally susceptible? We do not think that is the final conclusion of this research question. We do believe we are contributing to further investigation into identifying and educating susceptible users. This research demonstrates there are data and mechanisms available to further investigate social media behaviors in relation to phishing susceptibility. Our research helps to move forward such research, reducing the gap in the literature on how to define and determine a social media users' susceptibility to phishing.

7 Conclusion

We were able to identify differences exist in behaviors between susceptible and control users. However, due to the significant overlap between the control and susceptible users, further analysis needs to be conducted with this research question. We were able to introduce the idea of using social media behavioral metrics

to predict if a user has the potential to fall susceptible to phishing scams. One of the biggest challenges of this research was data collection. It proved to be very time-consuming and challenging to find data that corresponds to susceptible users and phishing tweets. As a result, we ran analysis on two datasets containing a total of 2,380 users. Analysis on a larger variety of data (more users, more tweets, and more behavioral details) would be useful to expand and further develop this research and findings. The limitation of data most negatively impacting the classification models as these models overfit to the stronger class of the given dataset, missing the identification of the other class.

Using other machine learning techniques could have also provided some different results and insights. Our analysis provided a summary of the potential differences between susceptible and the control users. A deeper dive into those specific differences could provide the behavioral differences between potentially susceptible and less susceptible users. By noticing differences at the extreme ends of the page metrics, we could potentially identify the extreme outlier cases of susceptible and less susceptible users.

At the time of publication, the BLOC tool has been applied to determining if Twitter users are bots or cyborgs. We were able to take this tool and apply it in a new way. The tool allowed us to contribute new analysis to the studies of phishing susceptibility. BLOC provided the behavioral metrics which is the unique feature of this research. Implementing more features of the BLOC tool could provide further and more comprehensive analysis.

There was a lack of literature on which social media behaviors could provide insight into a user's potential susceptibility. Our predictions based on the research into personality traits proved to either be the opposite or not impactful. Hopefully, with more research into phishing susceptibility on social media, more guidance could be implemented in answering our main research questions. When dealing with phishing susceptibility education is the best way to prevent successful attacks. This research provides one more step towards educating people about phishing and the drastic impact falling for an attack can have.

8 Acknowledgements

We would like to thank the committee Dr. Alexander C. Nwala (chair), Dr. Dan Runfola, and Dr. Salvatore Saporito. We acknowledge William & Mary Data Science Department for giving us the opportunity to conduct this work. A special thanks to Dr. Nwala for his help and advising; his tool, BLOC, proved to be instrumental in the analysis for this research.

Appendix A

Appendix

1 Control Tweet IDs

Political Control Tweet IDs:

- Wall Street Journal
 - 1047531845162864640
- New York Times
 - 1049099314155458562
- The Hill
 - 1057395754371899394
 - 1055838993940733953

Financial Control Tweet IDs:

- Bespokeinvest
 - 45680892923654145
 - 946870172597485568
- CNBC
 - 946853064274776066
 - 946829413047554048
 - 946891431079284737

- Stockwits
 - 944215945500700672
 - 946400637302034434

2 BLOC Alphabet

This is the list of BLOC words used in our analysis:

- t: Text
- H: Hashtag
- M: Mention of friend
- m: Mention of non-friend
- q: Quote of other's post
- E: Media object (e.g., image/video)
- U: link (URL)
- t_{min} : Pause less than a minute
- t_h : Pause less than an hour
- t_d : Pause less than a day
- t_w : Pause less than a week
- t_m : Pause less than a month
- T: Post message
- P: Reply to friend

Bibliography

- Bossetta, Michael (2016). *A Simulated Cyberattack on Twitter: Assessing Partisan Vulnerability to Spear Phishing and Disinformation ahead of the 2018 U.S. Midterm Elections*. DOI: <https://doi.org/10.5210/fm.v23i12.9540>. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/9540/7651>.
- (2018). “A simulated cyberattack on Twitter: Assessing partisan vulnerability to spear phishing and disinformation ahead of the 2018 U.S. midterm elections”. In: *First Monday*. DOI: <https://doi.org/10.5210/fm.v23i12.9540>. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/9540>.
- Cresci, Stefano et al. (2018a). “A simulated cyberattack on Twitter: Assessing partisan vulnerability to spear phishing and disinformation ahead of the 2018 U.S. midterm elections”. In: *ACM Transactions on the Web*. DOI: <https://doi.org/10.1145/3313184>. URL: <https://arxiv.org/abs/1804.04406>.
- (2018b). *Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter*. DOI: <https://doi.org/10.48550/arXiv.1804.04406>. URL: <https://zenodo.org/record/2686862#.ZD670HbMK3B>.
- Ferrara, Emilio et al. (2020). “Characterizing social media manipulation in the 2020 U.S. presidential election”. In: *First Monday*. DOI: <https://doi.org/10.5210/fm.v25i11.11431>. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/11431>.
- Frauenstein, Edwin Donald and Stephen Flowerday (2020). “Susceptibility to phishing on social network sites: A personality information processing model”. In: DOI: <https://doi.org/10.1016/j.cose.2020.101862>. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252086/>.
- Golbeck, Jennifer, Critina Robles, and Karen Turner (2011). “Predicting Personality with Social Media”. In: *Playing Well With Others*, pp. 253–262. DOI: <https://doi.org/10.1145/1979742.1979614>. URL: <https://dl.acm.org/doi/pdf/10.1145/1979742.1979614>.

- Heartfield, Lucas, George Loukas, and Diane Gan (2016). "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks". In: *IEE*, 6910—6928. DOI: <https://doi.org/10.1109/ACCESS.2016.2616285>.
- Henriquez, Maria (2023). *\$4.35 million — The average cost of a data breach*. Ed. by Security. URL: <https://www.securitymagazine.com/articles/98486-435-million-the-average-cost-of-a-data-breach#:~:text=The%20global%20average%20cost%20of,of%20a%20Data%20Breach%20Report.%E2%80%9D>.
- Nwala, Alexander C., Alessandro Flammini, and Filippo Menczer (2022). "A General Language for Modeling Social Media Account Behavior". In: URL: <https://arxiv.org/pdf/2211.00639.pdf>.
- Parker, Tim (2022). "10 Twitter Feeds Investors Should Follow". In: ed. by Investopedia. URL: <https://www.investopedia.com/financial-edge/0712/10-twitter-feeds-investors-should-follow.aspx>.
- Ratkiewicz, Jacob et al. (2021). "Detecting and tracking political abuse in social media". In: *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 297–304. DOI: <https://doi.org/10.1609/icwsm.v5i1.14127>.
- Razaque, Abdul et al. (2021). "Blockchain-Enabled Deep Recurrent Neural Network Model for Clickbait Detection". In: *IEEE*, pp. 3144–3163. DOI: <https://doi.org/10.1109/ACCESS.2021.3137078>. URL: <https://ieeexplore.ieee.org/abstract/document/9656746>.
- Scikit Learn, ed. (n.d.). *sklearn.preprocessing.StandardScaler*. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.
- Seidman, Gwendolyn (2020). *What Can We Learn About People From Their Social Media?* Ed. by Psychology Today. URL: <https://www.psychologytoday.com/us/blog/close-encounters/202009/what-can-we-learn-about-people-their-social-media#:~:text=The%20content%20on%20social%20media%20predicts%20personality.&text=Bachrach%20and%20colleagues%20found%20they, is%20related%20to%20their%20personality..>
- Sheng, Steve et al. (2010). "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 373–382.

DOI: <https://doi.org/10.1145/1753326.1753383>. URL: <https://dl.acm.org/doi/abs/10.1145/1753326.1753383>.

Tornblad, McKenna K et al. (2021). "Characteristics that Predict Phishing Susceptibility: A Review". In: pp. 938–942. DOI: <https://doi.org/10.1177/1071181321651330>. URL: <https://journals.sagepub.com/doi/pdf/10.1177/1071181321651330>.

Wang, Jingguo et al. (2012). "Phishing susceptibility: An investigation into the processing of a targeted spear phishing email". In: *IEEE*. DOI: <https://doi.org/10.1109/TPC.2012.2208392>. URL: <https://ieeexplore.ieee.org/document/6289402>.

Zandt, Dave Van (2015). *Media Bias Fact Check*. URL: <https://mediabiasfactcheck.com/>.