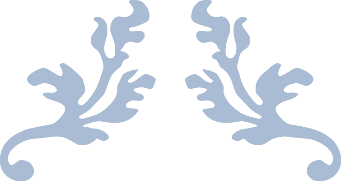
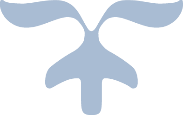
****

**MALICIOUS TWITTER BOT TYPE DETECTION**

**WEB ANALYTICS PROJECT**

****



**EXECUTIVE SUMMARY**

Bots on Twitter are accounts that are controlled by computer programs, automatically producing content, and interacting with other accounts. These programs are turned on and off without following a pattern, making them hard to identify. Some bots only exist for providing some interesting tweets to users daily, but some bots are intentionally spreading disturbing or misleading information, which negatively impacts users’ experiences. Those malicious bots indirectly hurt Twitter’s image and business revenue.

Although Twitter has already been able to identify most of the bot accounts, the company could have done better in Malicious Bots Classification. Therefore, in this project, we propose a novel bot type classification method by using TFIDF. With the supervised machine learning method, this study aims to detect malicious Twitter bot types based on bot behaviors and text analysis. To identify the important indicators for malicious bot type detection, we downloaded different malicious bots’ IDs from Bot Repository and used Python to crawl user’s information and tweets based on these IDs. The methodology of our study divides into two phases. In phase one, we used IBM SPSS Modeler to create testing models with 13 typical fields, because we can leverage the models to find the most important indicator of predicting a certain malicious bot type. In phase two, we created a dictionary for each type of malicious bots and calculated the TFIDF of keywords in the dictionary of each tweet. Finally, with three new fields, we got a better-performed model for malicious Twitter bot detection.

This study can help Twitter not just keep suspending all the bots-like accounts but identify malicious bot that should be banned in a more precise way. With a larger dataset from Twitter’s database, Twitter can optimize our method with a more comprehensive dictionary.

**I. BUSINESS GOAL ANALYSIS**

In recent years, Twitter, with its easy enrollment process and attractive user interface, has seen a proliferation of automated accounts or bots. While a few of these automated accounts engage in human conversation or provide community benefits, many are malicious. We define malicious bots as those that violate Twitter’s terms of service, including those that post spam content, adware and malware, as well as bots that are fake followers and part of sponsored campaigns to sway public opinion. These bots need to be controlled, or they will affect human users’ performance.

Twitter makes money through advertising, and if the number of human users reduces, it will affect the company’s revenue because advertising firms will not spend money buying promoted products from twitters. Furthermore, it also induces cost from government scrutiny. For instance, a subset of Twitter bots programmed to complete social tasks played an important role in the United States’ 2016 Presidential Election. Researchers estimated that pro-Trump bots generated four tweets for every pro-Clinton automated account and out-tweeted pro-Clinton bots 7:1 on relevant hashtags during the final debate. Deceiving Twitter bots fooled candidates and campaign staffers into retweeting misappropriated quotes and accounts affiliated with incendiary ideas. Concerns about political Twitter bots include the promulgation of malicious content, the increase of polarization, and the spreading of fake news.

How prevalent are bots and bot networks on Twitter? Estimates vary, with Twitter itself stating that less than 5% of its over 300 million active accounts are bots. Using a supervised machine learning approach with a manually curated set of Twitter bots, estimate that between 9% to 15% of active Twitter accounts are bots (both benign and malicious). Although Twitter has already been able to identify most of the bot accounts, the company could have done better in Malicious Bots Classification. Our goal is to build a better model for Twitter to identify different kinds of malicious bots, using typical features combined with our new features. We hope Twitter can suspend malicious bots better than before with the new model, and then offer a better community for other users.

**II. DATASET DESCRIPTION**

Once we figured out our business goal, we considered how many groups we should divide malicious bots. We referred several papers and defined malicious bots as Fake Follower, Scam Bot, and Spam Bot. The gold standard for confirming different types is accounts already recognized by other researchers1.

|  |  |
| --- | --- |
| **Malicious Bot** | **Description** |
| Fake Follower | Robot or inactive accounts that inflate number of followers of another account. |
| Scam Bot | Accounts that advertise scam sites. |
| Spam Bot | Accounts that spam different kinds of information by sending messages with the same content multiple times. |

Table 1. Malicious Bot Classification

We downloaded 900 IDs for each kind of malicious bots from Bot Repository2 and got 2700 IDs in total. After getting all the IDs, we thought of our features and decided to analyze from four perspectives, including tweet syntax, tweet semantics, temporal behavior, and user profile. We use 2700 IDs to crawl twitter information and tweets. Since Twitter has already banned some accounts and some accounts may not have any tweet, we finally got 2042 valid accounts and crawled the most recent 250 tweets for each account. Some accounts may not have 250 tweets in total, so we eventually crawled 138042 tweets for Fake Followers, 160245 tweets for Scam Bots, and 161956 tweets for Spam Bots.

After crawling all the tweets, we used python to delete all stop words and calculated every term’s frequency for each group of malicious bots and picked those terms with a larger than 0.05% term frequency rate to create three dictionaries to calculate our new feature TFIDF. Since we had to add all TFIDFs up, we must normalize them to make them under the same weights and measurements. We used Python and Excel to calculate the normalized TFIDF of each tweet and the average normalized TFIDF of each account. The formula of the normalized TFIDF is shown below:

Normalized TFIDF =

We also used Word Cloud to check our term frequency, and the results of the Word Cloud were consistent with ours.

|  |  |
| --- | --- |
| **Category** | **Features Will Be Used** |
| Tweet Syntax | The average number of retweets of tweets for each account |
| Percentage of tweets containing URL or hyperlink for each account |
| Tweet Semantics | Keyword TFIDF |
| Temporal Behavior Features | Average number of tweets per day |
| User Profile Features | Number of followers of one account |
| Number of friends of one account |
| Number of tweets that one account has |
| Using default profile |
| Using default profile image |
| Using geography or location enabled |

Table 2. Features Used for Malicious Bot Type Detection

|  |  |
| --- | --- |
| **Data Description** | **Data Size** |
| Number of valid ID | 709 of Fake Followers, 641 of Scam Bot, 692 of Spam Bot |
| Number of tweets | 216173 |
| Total number of words | 138042 for Fake Follower, 160245 for Scam Bot, 161956 for Spam Bot |
| Number of keywords in a dictionary | 120 for Fake Follower, 126 for Scam Bot, 170 for Spam Bot |

Table 3. Data Description and Data Size



Figure 1. Word Cloud of Fake Followers’ Tweets



Figure 2. Word Cloud of Scam Bots’ Tweets

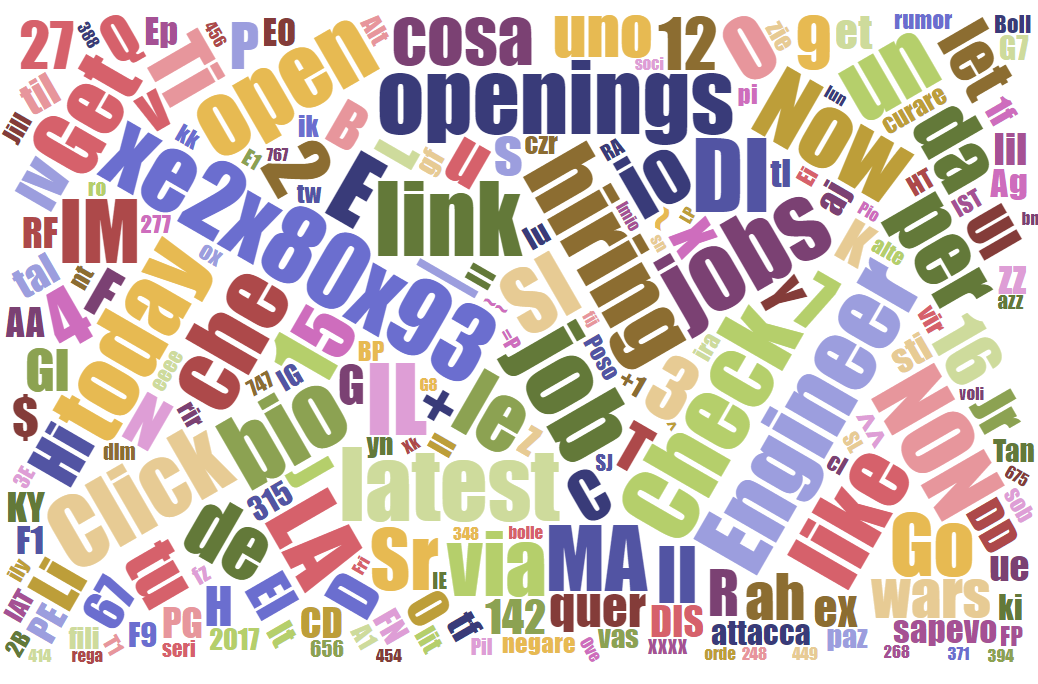


Figure 3. Word Cloud of Spam Bots’ Tweets

**III. SYSTEM DESIGN**

This project includes two phases as shown below. First, we studied Twitter’s business background, then we set our goal and downloaded the raw data. We used Python and Excel to get more features and processed the data. In phase one, we used typical features of tweet syntax, temporal behavior and user profile to build a model to identify different types of malicious bots. In phase two, we calculated every term’s frequency for each group of malicious bots and picked those terms with a larger than 0.05% term frequency rate to create three dictionaries to calculate our new feature TFIDF, and then we used this new feature of tweet semantics to build a new model to see whether we can build a better model with this new feature.



Figure 4. Research Flowchart

**IV. SYSTEM IMPLEMENTATION**

**i. Phase One: Random Forest Classification with Typical Features**

Random Forest methodology, also known as Random Decision Forest, is a classification algorithm consisting of many decision trees. In phase one, we used IBM SPSS with Random Forest to build an uncorrelated forest of trees to predict the type of malicious bots based on 13 identified fields of user behaviors, which are denoted in the table below.

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| id | User ID for each Twitter account |
| Average\_of\_retweet\_count | Average number of retweets of tweets for each account |
| Average\_of\_favorite\_count | Average number of the times tweets favorited by Twitter users for each account |
| bot\_type | 1. Fake Follower Bots 2. Scam Bots 3. Social Spam Bots #1: retweeters of an Italian political candidate 4. Social Spam Bots #2: spammers of paid apps for mobile devices 5. Social Spam Bots #3: spammers of products on sale at Amazon.com 6. Traditional Spam Bots #1: Training set of spammers used by Yang et al 7. Traditional Spam Bots #2: Automated accounts spamming job offers 8. Traditional Spam Bots #3: Another group of automated accounts spamming job offers |
| bot\_group | 1. Fake Followers  2. Scam Bots  3. Spam Bots.  These groups are derived from 8 types of bots. |
| num\_of\_followers | Number of followers for each account |
| num\_of\_friends | Number of friends for each account |
| status\_num | Number of tweets for each account |
| default\_profile | Using default profile |
| default\_profile\_image | Using default profile image |
| geo\_enabled | Turning on geo-location |
| average\_tweet\_per\_day | Average number of tweets an account posts daily |
| percentage\_of\_url\_tweet | Percentage of tweets containing URL or hyperlink for each account |

Table 4. Features Used in Phase One

The target for the Random Forest model is bot\_group, which containing three nominal values of Fake Followers, Scam Bots, and Spam Bots. Spam Bots contains all the bot types of 3,4,5,6,7, and 8. We set id and bot\_type as typeless measurements since these fields have no significant impact on our analysis. The dataset was partitioned into 70% of the training set and 30% of the testing set. The Random Forest test classified malicious bots with a high level of accuracy, 99.43% correctly on training data, and 91.05% on testing data. Since the training set may be overfitting, we focus on the testing set. Regarding the testing set, the model performed best at predicting Spam Bots, with accuracy of 96.226%, following by Fake Followers and Scam Bots, with accuracy of 89.862% and 87.019%, respectively. Overall, the Random Forests can be applied to classify malicious bot types with the features above at a high level of accuracy.

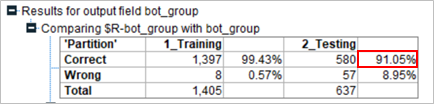


Figure 5. Random Forest Accuracy Analysis

The Random Forests Classification applied nine predictor inputs to train 1405 records, which is 70% of the dataset. The result indicated that some features matter more than others. Pertaining malicious bot behaviors classification, the model identified the most important predictor as average of retweet count, following by number of followers, average daily tweet, number of tweets, number of friends, percentage of tweets containing URL or hyperlink per account. Using the default profile, the average number of favorites and turning on geo-location do not demonstrate significant results for classification.

Group 1: Fake Followers

Fake Followers tend to be inactive users. Based on our dataset, on average, a Fake Follower account has 55 followers and only post roughly three tweets every day. Overall, Fake Followers do not tweet or retweet frequently, have a low number of followers, and tend to use the default profile. These accounts are not real and only be there to inflate the number of followers of other accounts.

Group 2: Scam Bots

We classified Scam Bot behavior based on the predictor importance indicators and our dataset by utilizing graph analysis. According to the pie charts demonstrating portions of followers and friends that each type of malicious bot group has, on average, a Scam Bot has the lowest number of both followers of 32 and friends of 88. Besides, on average, a Scam Bot also performs at the lowest tweet frequency. Scam Bots tend not to frequently tweet every day and has the lowest number of total tweets among the three groups of malicious bots. However, a Scam Bot has the largest retweet counts of 2,072 among the three malicious bot groups. Overall, Scam Bots are inactive users with low tweet frequency, the least friends and followers but retweet the most among the three groups of malicious bots.

Group 3: Spam Bots

On the other side, Spam Bots tend to be active users with the highest frequency of activities on Twitter. Based on our analysis, among the three malicious bot types, Spam Bots has the most followers and friends. On average, a Spam Bot account has 8206 followers, while a Scam Bot and a Fake Follower only have 32 and 55 followers, respectively. A Spam Bot can daily post ten tweets, which align with its purpose of spamming other users by continuously posting tweets. Furthermore, tweets with hyperlinks are 21.76% of all the tweets that they posted. Overall, Spam Bots, though rarely retweet, are active Twitter robotic accounts, which have an extremely high number of followers and friends and tweets highly favorited by others, continuously tweet every day, and have the highest number of tweets per account compared to the other malicious bots.

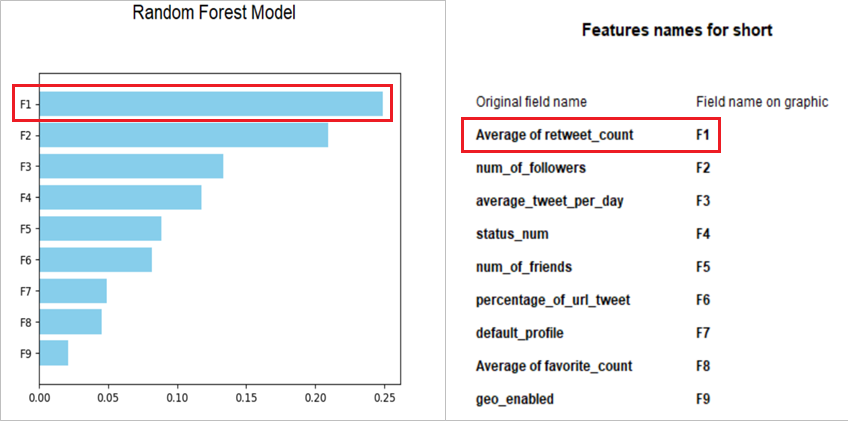


Figure 6. Random Forest Importance Indicator

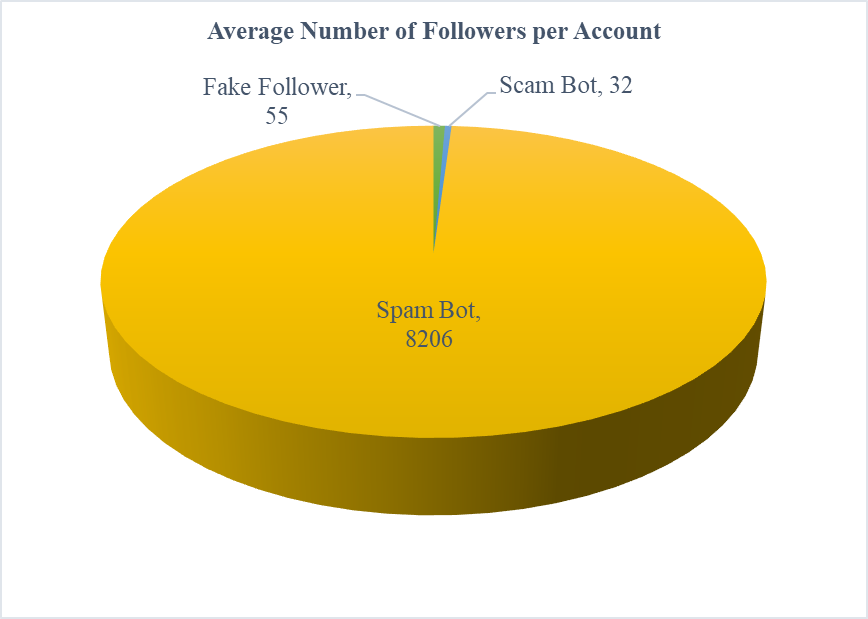


Figure 7. Average Number of Followers per Account

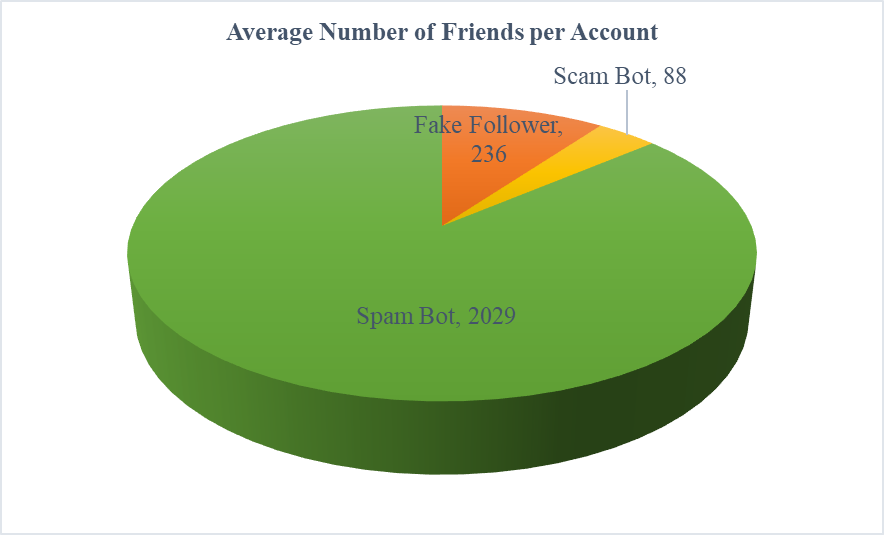


Figure 8. Average Number of Friends per Account

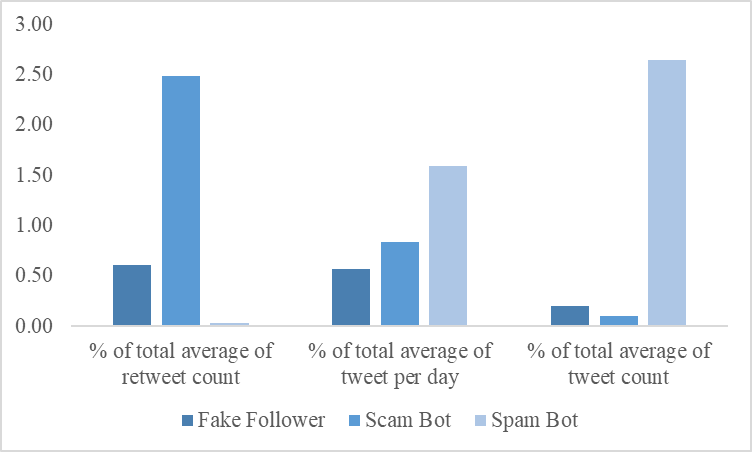


Figure 9. Tweet and Retweet Frequency

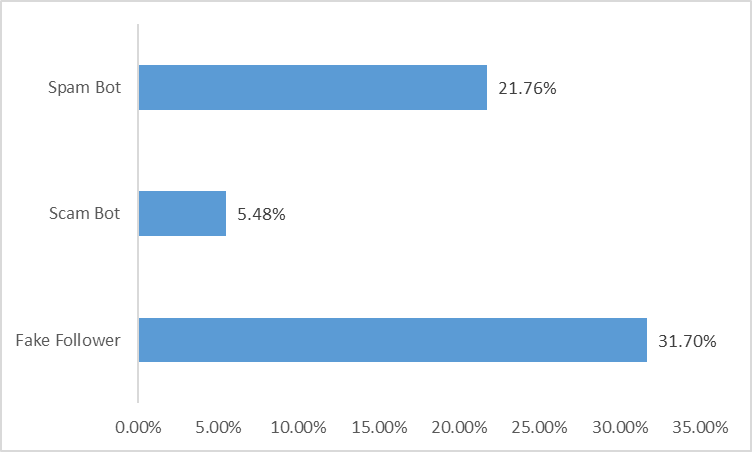


Figure 10. Percentage of Tweet with URL on Average per Account

**ii. Phase Two: Random Forest Classification with TFIDF**

We used Python package NLTK to get the keywords dictionary and frequency of each keyword. We input all 216173 tweets we crawled, removed all stop words, created a list of all words, and calculated the frequency of each of them. We used the words with more than 0.05% frequency rate to build a detection dictionary for each group: dict\_fake, dict\_scam, dict\_spam.

To calculate the fitting degree of each tweet with these three dictionaries, we used Python and Excel to calculate the keywords' TFIDF in each tweet to see whether these dictionaries contribute to our prediction. TF is normalized. In case the tweet has no keywords, or all the terms in the tweet are keywords, we also normalized the IDF.

Normalized TFIDF =

TF =

Keyword frequency: The times of keywords appear in that specific tweet

IDF =

We computed the TFIDF for each tweet based on all three dictionaries. The line chart below shows the average TFIDF of each group of bots. Fake Followers have the lowest TFIDF score in all three dictionaries and a little better match in Fake Follower dictionary, probably due to their small number of tweets. Scam Bots are in the middle, and Spam Bots have high TFIDF in all three dictionaries and especially in the spam dictionary. Spam Bots do post tons of tweets, and the result indicates that there are some words in the spam dictionary that have discriminative power to distinct Spam Bots from other bots.

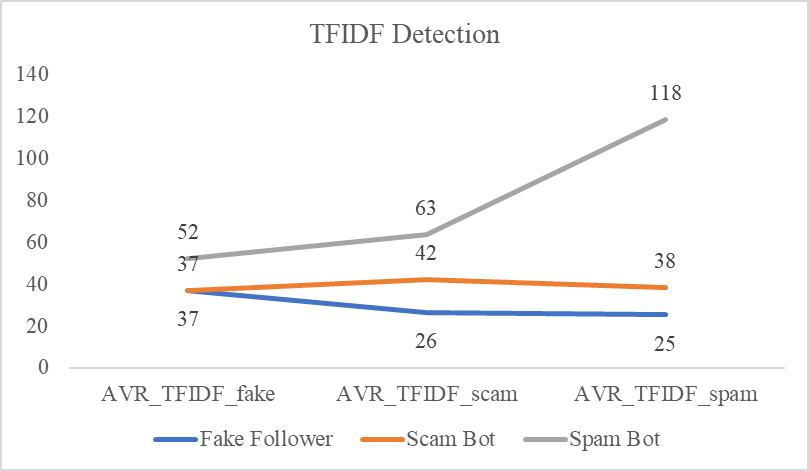


Figure 11. Average TFIDF of Each Group of Bots

We added these three average TFIDF as our new features in our detection model and built a new model to see whether these three new features can contribute to our overall accuracy.

The Random Forest Classification applied 12 inputs as predicting factors to train 1405 records again. The dataset was partitioned into 70% of the training set and 30% of the testing set. Adding three new inputs of average TFIDF scores of Fake Followers, Scam Bots, and Spam Bots, the model produced a new list of important indicators. The top three indicators now are average of retweet count for each tweet, the number of followers, and average TFIDF score of the spam bot. According to the result, the TFIDF score plays a significant role in classifying types of malicious bots. This new model predicted malicious bots correctly of 91.68% on the testing set.

A screenshot of a cell phone

Description automatically generated

Figure 12. Random Forest Importance Indicator with TFIDF Score

**V. EVALUATION**

In general, the model of phase two performed slightly better than the model of phase one based on the comparison of accuracy. We can’t tell whether it’s due to the new features or the random error. However, one thing we are sure about is that the TFIDF certainly contributes to analyze the bot types, because when we only used these three new features to detect different kinds of malicious bots, the model could predict correctly of 81.16% on the testing data.

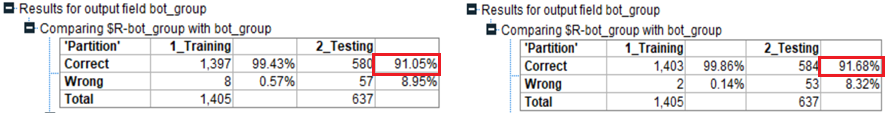


Figure 13. Phase One & Phase Two Model Evaluation

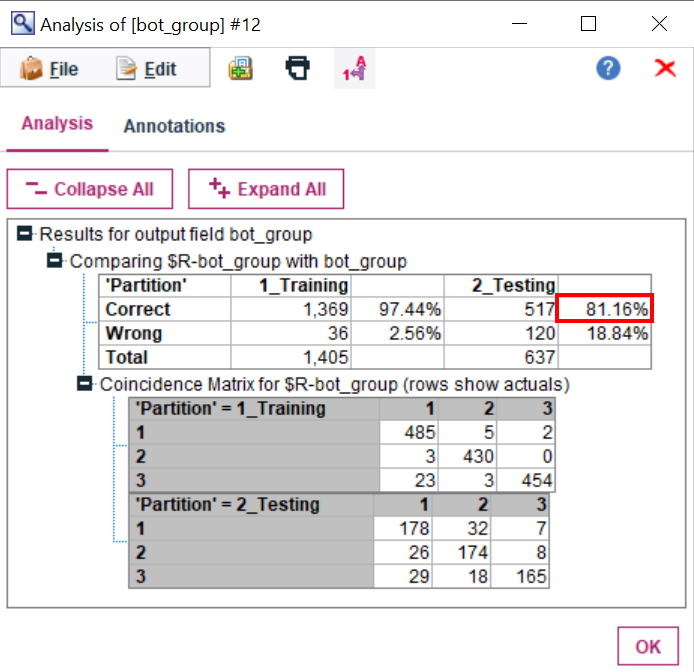


Figure 14. Random Forest Accuracy Analysis with Only TFIDF

**VI. CONCLUSION AND FUTURE DIRECTION**

**i. Conclusion**

In this project, we study the problem of classifying the different types of malicious bots on Twitter. First, we reviewed related work and classified the malicious bots as Fake Follower, Scam Bot, and Spam Bot. Although the existing methods these papers presented high accuracy in detection, most of them studied the problem in bot identification, and even though some of them also studied in malicious bots identification, none of them use keyword detection.

In our research, we target studying malicious accounts’ tweets as a means to build keyword dictionaries. Bot detection tools are more and more advanced, but bots are also becoming more and more human-like, which makes them more complicated to detect. We focus on finding features related to natural language. By running our script to get all the terms of each tweet and generating the report, we can identify the type of malicious bots.

In phase two, we train a Random Forest classifier using the features of the average TFIDF score of keywords in each type dictionary. We are able to create a multiclass classifier that performs with 91.68% accuracy when classifying into one of 3 classes.

Overall, Fake Followers are inactive accounts with the highest score of TFIDF\_fake, Scam Bots are also inactive accounts but with high retweet counts and the highest score of TFIDF\_scam, and Spam Bots are active accounts with the highest score of TFIDF\_spam. Besides, we can predict Spam Bots the best using the TFIDF score since TFIDF\_spam is in the top three important indicators of the model.

**ii. Future Direction**

In this project, we only focused on malicious bot type detection without combining bot detection, because we thought Twitter has already done well in bot detection. Therefore, some more work may need to be done to consolidate our model with other bot detection models. What’s more, in this project, we only use unigram to build the dictionary. However, bots are becoming more and more complicated, so we may need to use n-gram to build more advanced dictionaries in the future.

**REFERENCES:**

[1] Stefano Cresci et al. The paradigm-shift of social spambots. arXiv:1701.03017v1 [cs.SI] 11 Jan 2017.

Available at: https://arxiv.org/abs/1701.03017

[2] Cresci-2017 and Pronbots-2019 : https://botometer.iuni.iu.edu/bot-repository/datasets.html

Accessed: 2019-11-26.

[3] Lulwah Ahmad AlKulaib. Twitter Bots Multiclass Classification Using Bot-Like Behavior Features. B.S. in Computer Science, Gulf University for Science and Technology, June 2011

Available: https://scholarspace.library.gwu.edu/downloads/47429940p?

[4] Efthimion, Phillip George; Payne, Scott; and Proferes, Nicholas (2018) "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots," SMU Data Science Review: Vol. 1: No. 2, Article 5.

Available at: https://scholar.smu.edu/datasciencereview/vol1/iss2/5