5-2021

Detecting Bots Using a Hybrid Approach

Edmund Kofi Genfi
Montclair State University

Abstract

Artificial intelligence (AI) remains a crucial aspect for improving our modern lives but it also casts several social and ethical issues. One issue is of major concern, investigated in this research, is the amount of content users consume that is being generated by a form of AI known as bots (automated software programs). With the rise of social bots and the spread of fake news more research is required to understand how much content generated by bots is being consumed. This research investigates the amount of bot generated content relating to COVID-19. While research continues to uncover the extent to which our social media platforms are being used as a terrain to spread information and misinformation, there still remain issues when it comes to distinguishing between social bots and humans that spread misinformation. Since online platforms have become a center for spreading fake information that is often accelerated using bots this research examines the amount of bot generated COVID-19 content on Twitter. A hybrid approach is presented to detect bots using a Covid-19 dataset of 71,908 tweets collected between January 22nd, 2020 and April 2020, when the total reported cases of Covid-19 were below 600 globally. Three experiments were conducted using user account features, topic analysis, and sentiment features to detect bots and misinformation relating to the Covid-19 pandemic. Using Weka Machine Learning Tool, Experiment I investigates the optimal algorithms that can be used to detect bots on Twitter. We used 10-fold cross validation to test for prediction accuracy on two labelled datasets. Each dataset contains a different set (category 1 and category 2) of four features. Results from Experiment I show that category 1 features (favorite count, listed count, name length, and number of tweets) combined with random forest algorithm

produced the best prediction accuracy and performed better than features found in category 2 (follower count, following count, length of screen name and description length). The best feature was listed count followed by favorite count. It was also observed that using category 2 features for the two labelled datasets produced the same prediction accuracy (100%) when Tree based classifiers are used.

To further investigate the validity of the features used in the two labelled datasets, in Experiment II, each labelled dataset from Experiment I was used as a training sample to classify two different labelled datasets. Results show that Category 1 features generated a 94% prediction accuracy as compared to 60% accuracy generated by category 2 features using the Random Forest algorithm. Experiment III applies the results from Experiment I and II to classify 39,091 account that posted Coronavirus related content. Using the random forest algorithm and features identified Experiment I and II, our classification framework detected 5867 out of 39,091 (15%) account as bots and 33,224 (85%) accounts as humans.

Further analysis revealed that bot accounts generated 30% (1949/6446) of Coronavirus misinformation compared to 70% of misinformation created by human accounts. Closer examination showed that about 30% of misinformation created by humans were retweets of bot content. In addition, results suggest that bot accounts were involved in posting content on fewer topics compared to humans. Our results also show that bots generated more negative sentiments as compared to humans on Covid-19 related issues. Consequently, topic distribution and sentiment may further improve the ability to distinguish between bot and human accounts.

Keywords: Social Bots, Human, Misinformation, information, Detection Technique, Hybrid Approach, Social Networking Features, Sentiments Features

Montclair State University

Detecting Bots Using a Hybrid Approach

by

Edmund Kofi Genfi

A Master Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

May 2021

College of Science and Mathematics
Department of Computer Science

Thesis Committee:

Dr. Christopher Leberknight

Thesis Sponsor

[redacted]

Dr. Bharath Samanthula

Committee Member

[redacted]

Dr. Boxiang Dong

Committee Member

[redacted]

DETECTING BOTS USING A HYBRID APPROACH

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

by

Edmund Kofi Genfi

Montclair State University

Montclair, NJ

2021

Acknowledgments

I dedicate this work to God Almighty. I am very delighted for his complete protection and guidance throughout my Master thesis project. I also dedicate this work to my mother who has supported me tremendously throughout my university education. "We must consistently acknowledge the efforts of people who in one way or another contributed immensely to the turning point of our lives"- John Osten.

In the light of this statement, I will like to express my deepest and hearty appreciation to my supervisor – Prof. Christopher Leberknight, Sir, I am very grateful for your complete supervision and constructive criticism that helped tremendously in the completion of this work and contributed immensely to my understanding of Computer science research.

My gratitude goes to all lecturers of the department of Cybersecurity for the knowledge they have imparted in me. Special thanks go to Daniel Chege (Ms. Cybersecurity), Murad Hasan (Ms. Cybersecurity) and my fiancée, Tiffany Opoku-Antwi for their role in the research process.

Contents

LIST OF ILLUSTRATIONS

Figure Page

List of Tables

Table Page

Chapter 1 Introduction

1.1 Background and Motivation

There has always been the need to study how Bots or network of Bots (Sybils) affects social media and its impact on politics and national security. If you are an individual that searches for daily news on social media, like most people do, then you may be exposed to many types of fake and misleading content (Dunn et al., 2011). For example, hoaxes, rumors, fabricated stories, conspiracy theories, and click-bait are all forms of misleading content (Dunn et al., 2011). While malicious social bots often wage disinformation campaigns by targeting political or economic content, the volume of such campaigns render manual detection infeasible. Social media users are often unable to identify content created by social bots. Scrolling through your favorite social media page, it may not be obvious if you come across a bot account.

A malicious bot is a compromised computer under the direction of a human operator called "Botmaster" (Feily et al., 2009). The term "Bot" is derived from the word "Robot", and just like Robots, bots are created to perform a specific function in an automated manner (Feily et al., 2009). These bots are pieces of software programs that run on infected machines without the user knowing about their existence (Al-Hammadi & Aickelin, 2017).

Botnets or Sybils (network of compromised computers) have become a huge cybersecurity problem and have been used as a means to carry out most forms of cyber-attack (Eslahi et al., 2012).

The presence of these computerized agents has been observed in many sections of social media applications such as Twitter which has been the most affected (Shao et al., 2018). These social media bots create a platform for the spreading of several illegal activities such as launching DDOS attacks against specific targets (Feily et al., 2009). A publication on MIT Technology Review in 2020 reported that researchers observed that about half of some 200 million tweets on the novel COVID-19 likely came from bots, with many of them spreading false information, pushing conspiracy theories, and advocating for the United States to loosen restrictions in order to reopen America (*Nearly Half of Twitter Accounts Pushing to Reopen America May Be Bots | MIT Technology Review*, 2020). Figure 1.1 shows a typical Botnet architecture.

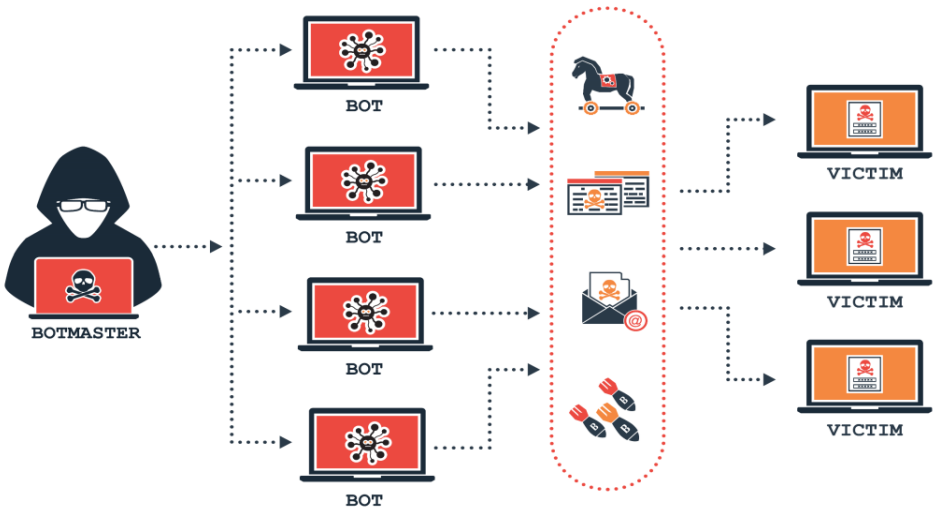


Figure 1.1: Botnet Architecture (Adapted from Depositphotos)

Though hard to verify, researchers have also put forward claims about how fake news can change how people think during a pandemic (Evanega et al., 2020; Chen et al., 2020). Yet we have seen many forms of demonstration of real harm in 2020 caused by the spread of

misinformation on social media relating to COVID-19 (P. Wang et al., 2018). The influence of fake news on Twitter during the 2016 US presidential elections is a crucial example that shows why much attention and research is needed to deal with malicious social media bots. Using a casual model , the authors used a dataset of about 171 million tweets to identify 30 million tweets spreading either false information or extremely biased news from about 2.2 million users (Bovet & Makse, 2019).

Another research study used a dataset with 3.6 million tweets and observed that about 23.6% of those tweets that were examined were spreading hate speech by dividing public views on issues concerning Brexit or the Catalan referendum (Rosso, 2019). While misleading content is not something new, many online information platforms do not have adequate safeguards to control and the spread of misinformation. It is now easy to use social media to influence public opinion due to the low cost of creating fake websites and the existence of several software-controlled social media profiles or pages (Dunn et al., 2011).

Internet users believe in social contracts (Dunn et al., 2011) and can be made to accept and spread content produced in a certain way (P. Wang et al., 2018). Moreover, the augmentation of misleading news through social bots overwhelms our fact-checking capacity because of our definite attention, as well as our propensities to consider what appears current and to believe information in a social environment. A well worked out strategy is required to fight against the spread of misinformation online (Dunn et al., 2011). People need education when it comes to the consumption of news on all internet platforms by the use of algorithms to widen the exposure to varied views and if malicious social bots

are the reason for the spread of misinformation, then there is the need to focus our attention on coming up with techniques to detect these malicious bots.

Mary Papenfuss from the HuffPost reported that there has been ongoing research about how social media bots are spreading misleading content about the novel Coronavirus (COVID-19) pandemic in May 2020. Researchers are yet to come up with a conclusion about the entities or organizations that may be primarily responsible for the bots. The primary objectives of this study are therefore to find out if the spread of misinformation during the Coronavirus (COVID-19) pandemic era was done by activities of Bots using a hybrid Bot Detection model.

False news, extensively disseminated over all internet platforms, can be considered as a form of computational propaganda (Howard et al., 2017). Social media have provided a platform for substantial volumes of fake, dramatic and other forms of junk news at delicate moments in our social setting, though most platforms disclose little about how much of this content there is or how it impacts those who use the platform (Howard et al., 2017).

The United States Department of Homeland Security reported in 2020 that the World Economic Forum has identified the spread of disinformation as one of the top 10 threats to society(*COVID-19 Exploited by Malicious Cyber Actors | CISA*, 2020). It has been reported that bots can jeopardize online information platforms as well as our society (Ferrara et al., 2016). Prior studies have done a sensational job trying to figure out the best malicious Bot detection technique to help slow down the spread of fake news on all online information platforms, however the bot strategies continue to evolve to evade detection. Today, some social bots have been used to penetrate political discourse, control the stock market and steal private information (Bovet & Makse, 2019). Prior to the 2020 United

States elections, social media sites especially Twitter was flooded with bots that could evade most bot detection techniques. In a new study, researchers at University of Southern California identified thousands of bot accounts on Twitter that were uploading information related to Donald Trump, President Biden and their political campaigns. Many of these automated accounts were spreading disinformation and far-right conspiracy theories such as "pizzagate" and QAnon (*Twitter Bots Poised to Spread Disinformation Before Election - The New York Times*, 2020). Although social media platforms such as Twitter and Facebook have worked effortlessly to control the impact of malicious social bots on their respective platforms, identifying these bots still remain a difficult task and warrant further research. The detection of social bots and the motive behind the spread of certain sensitive and malicious information continues to be a significant research endeavor (Ferrara et al., 2016).

1.2 Organization of the study

This paper is organized into five (5) chapters. Chapter 1 which is the introductory chapter includes the background and motivation of the study, organization of the study, literature review and limitation of the study. The review of literature is an attempt to study prior studies on social bots to help have a better understanding of the issue or the problem this research seeks to solve. Chapter two includes the objectives of the study, problem statement and hypothesis. Chapter 3 describes the methodology, the dataset used in this study as well as the description of the experiments conducted. Chapter 4 focuses on analysis of the data, experimental results, misinformation and topic analysis, entities responsible for the spread of Covid-19 misinformation and sentiment analysis. Chapter five is the summary of the research findings and recommendations for future research.

1.3 Literature Review

Today, the research computing community is still designing sophisticated methods that can automatically detect or prevent malicious social bots that spread misinformation on online platforms. Bot detection techniques can be broadly divided into three distinct groups: (1) Graph-Based Social Bot Detection, (2) Crowdsourcing Social Bot Detection and (3) Feature-Based Social Bot Detection (Ferrara et al., 2016).

1.3.1 Graph Based Social Bot Detection

Graph Based Social Bot Detection is an intuitive way of representing network communications using graphs. A strategy developed known as BotChase presents a two-phased graph-based bot detection system that controls both unsupervised and supervised Machine Leaning. The authors application of BotChase could detect several types of bots and showed toughness to zero-day attacks (Daya et al., 2020). The author also observed that the BotChase strategy that they implemented was suitable for large-scale data and different network topologies. The authors in (Chowdhury et al., 2017) also proposed a bot detection technique based on topological characteristics of nodes within a graph. The authors administered a self-organizing map clustering method that was applied to establish clusters of nodes in the network based on these characteristics.

Previous research has also proposed a method that can isolate nodes in clusters of small size while containing the majority of the normal node in the same big cluster (Daya et al., 2020). Furthermore, a Graph- based malware activity detection was introduced by this

technique which makes use of a sequence of DNS queries in order to achieve robustness against evasion techniques (Lee & Lee, 2014).

While Graph-Based detection can be applied without knowledge of a specific language a major challenge is the availability of information that captures the complete topology of the network. The best bot detection technique that applies Graph- Based Social Bot Detection uses a hybrid analysis of flow-based and Graph- based traffic behaviors (W. Wang et al., 2020). The authors argued that only using graph-based analysis would result in false negatives or false positives or can even be eluded by malicious bots (W. Wang et al., 2020). To address the limitation with graph-based analysis they proposed another model known as BotMark that uses a hybrid analysis of flow –based and graph-based network traffic behaviors (W. Wang et al., 2020). The authors technique was able to characterize the botnets actions thoroughly as compared to other techniques. (W. Wang et al., 2020) report that one limitation with BotMark is that Botnets can use a legitimate server as their C&C communication to avoid detection. Since this paper will not be investigating network communication patterns between nodes, this study will not adopt this technique to detect malicious Twitter accounts.

1.3.2 Crowdsourcing Social Bot Detection

Wang et.al (2020) looked at the possibility of bot detection by humans. The authors recommended crowdsourcing of social bot detection to multitudes of workers. An online Social Turing test platform was created to see if humans can easily detect bot through the evaluation of conversational nuances like sarcasm or perspective language or to look at developing patterns and irregularities. The abilities of individuals were tested using data

obtained from Facebook and Renren which is an online Chinese social networking platform. The authors observed that the detection accuracy of both "experts" and "turkers" under various conditions vary tremendously in their effectiveness with experts consistently producing near perfect results. Though a great technique, crowdsourcing bot detection method has its drawbacks and might not be cost effective to help achieve the listed objectives and answer the research questions that this paper seeks to address.

1.3.3 Feature-based Social Bot Detection

Feature-based Social Bot Detection focuses on behavioral patterns that can be easily encoded in features and adopted with machine learning strategies to observe the patterns of human-like and bot-like behaviors (Ferrara et al., 2016.). Feature-based Social Bots Detection makes it easier to categorize accounts based on their detected behaviors (Ferrara et al., 2016).

The first social bot detection interface for Twitter in 2014 was made public to educate individuals on online information platforms about the presence of malicious bot activities (Ferrara et al., 2016). The authors proposed a bot detection algorithm that uses predictive features that detect a variety of malicious behaviors to deduce if information was created by a bot or human. A collection of networks, linguistic and application-oriented variables are used as likely features that associate certain characteristics to humans or bots (Ferrara et al., 2016). The challenge with using Feature-based bot detection is finding ground truth dataset that can be used as a training set to classify an unlabeled dataset. Another challenge is that the characteristics of a bot is increasingly becoming more humanlike so relying on only user account features may lead to incorrect classifications.

Chapter 2 Objectives of the study

2.1 Overview

The main objective of this present study is to contribute to an understanding of how social media bots spread misleading information on online information platforms and to find out if the spread of misinformation during the Coronavirus (COVID-19) pandemic era was done by activities of Bots or other individuals/organizations using hybrid approach that incorporates sentiment features, national language processing and social networking features to detect bots. The two main objectives for this research are to:

1. Identify twitter features that provide high discrimination quality for detecting bots
2. Investigate the spread of misinformation by bots during the initial months of the COVID-19 pandemic

Prior research indicates that the user metadata and user content provide the greatest discrimination accuracy (Shin et al., 2012). However, details on the specific features within each category are not reported. To optimize the performance of the classifier, optimal features within each category will be identified. Also, there is a discrepancy in prior research regarding the quality of network features. While user meta-data and user content have shown to perform the best (Shin et al., 2012), other studies suggest network features provided the highest accuracy for detecting content polluters (Dhital & Gonen, 2019). We aim to use a hybrid approach that incorporates user account features and sentiment features to detect malicious bot in Twitter.

3. To identify the source (bot or human account) responsible for spreading misinformation during the Coronavirus (COVID-19) pandemic era.

2.2 Problem Statement

Many studies have been conducted on social media bots to examine how to detect them and how these bots spread misinformation in online information platforms. Prior research reviewed techniques that can be used to fabricate misinformation by combining social bots and fake news to spread misinformation in a social setting (Daya et al.,2019; Wang et. al in 2018). (Eslahi et.al in 2012), studied the characteristics of the malicious activities of Bots and Botnets and came up with various detection techniques as well the challenges that accompanied those techniques. (Shao et al., 2018) studied how social bots spreads fake news by analyzing 14 million messages that were spreading 400 thousand claims during the 2016 US presidential elections. The study concluded that social bots played a key role in the spread of fake news during that time. Another study used a dataset with 3.6 million tweets with a casual model and observed that about 23.6% of those tweets that were examined were spreading hate speech by dividing public views on issues concerning Brexit or the Catalan referendum (Rosso, 2019).

(Ferrara et al., 2016) studied the rise of social bots and its impact on several online information platforms. Every aspect of our society is impacted heavily by social media today as it allows users to interconnect and exchange content freely (P. Wang et al., 2018). (Shin et al., 2012) among others also used a technique known as EFFORT to efficiently and effectively detect Bot Malware. (Shin et al., 2012) report that EFFORT can detect all 15 real world bots related to their study.

Recently, as of August 2020, there has been ongoing research about how social media bots are spreading misleading content about the novel Coronavirus (COVID-19) pandemic. The authors in (Varol et al., 2017; Rosso, 2019; Shao et al., 2018; Daya et al., 2019; Kudugunta & Ferrara, 2018) among others have studied the impact of malicious social bots and ways malicious social bots can be detected or prevented. However, researchers are yet to come up with a conclusion about what interest/entity may be primarily responsible for the bots. Although several techniques to detect malicious social bots have been created, there still remain issues when it comes distinguishing between social bots and human bots that spread misinformation. Since manual bot detection is infeasible, this study will develop a novel automated method to identify bots. While many automated methods have been proposed they have mainly been driven by features available in Twitter and apply single method approaches based on application specific features (Wang et al., 2019; Shin et al., 2012; Dhital & Gonen, 2019). To accomplish this task a hybrid approach that combines a variety of factors to detect bots will be developed. Specifically, the proposed bot detection model will incorporate user account features, topic analysis and sentiment analysis. It is also our objective to test different user account features to see which feature or set of features produces the best classification accuracy. (Varol et al., 2017) for example achieved the best classification performance by using two user account features i.e., follower count and friend count while (Wijeratne et al., 2017) observed that favorite count, tweet count and friend count are top three features that produced the best classification accuracy in their research. We aim to use, test and rank all Twitter user account features available in our Covid-19 Twitter dataset to observe their prediction and classification accuracy. The context for the study of misinformation is Coronavirus (COVID-19) data on social media.

This research will develop a model to identify bots and provide insights for organizations or entities who have interest in controlling these bots that have spread of misinformation during the Coronavirus (COVID-19) pandemic era.

2.3 Hybrid Approach

This research proposes a hybrid method that integrates Twitter user account features, sentiments features and topic analysis, to detect malicious social bots. A hybrid approach is a way of combining multiple approaches to improve detection accuracy (Ferrara et al., 2016). Wang et. al in 2018 developed a practical system using a server-side clickstream technique that showed effectiveness and high detection accuracy in detecting fake identities. This present study will rely on a similar approach conducted in prior research to detect bots by analyzing topical content (Morstatter et al., 2016). In a prior study it was observed that the content posted by bots can be a solid indicator that can help detect them (Morstatter et al., 2016). The authors used Latent Dirichlet Allocation (LDA) to attain topic representation of each user. However, the issue with using content for bot detection is that the nature of the text features is sparse and have high dimensionality (Morstatter et al., 2016).

Based on the review of prior bot detection studies, while many bot detection methods have been proposed, the feature-based detection appears to be most promising method and is therefore the focus of this research. This research will investigate new features and features that have been underexplored in previous studies. Many studies have examined bot detection accuracy using specific Twitter user account features (Lee, Eoff, and Caverlee, 2011). We aim to use, test and rank all Twitter user account features and analyze their

prediction and classification accuracy compared to features from a previous study (Lee, Eoff, and Caverlee, 2011). In this paper, we investigate the following three features: (1) topic distribution, (2) listed count, and (3) favorite count. These three user account features have been rarely used in prior research. Topic distribution on Twitter has to do with the variety of sentiments expressed by users on any given issue. listed count is a curated group of Twitter accounts, and favorite count is the number of accounts a Twitter user has favorited.

Chapter 3 Methodology

This is an empirical based research that uses several datasets with three experiments to detect bots and bot generated content. We start by generating a Twitter dataset associated with the novel coronavirus COVID-19 in a three-month period between January 22, 2020 to April 23, 2020. The Twitter's search API is used to hydrate tweets from multiple countries in various languages that contained any word associated with COVID-19 (i.e., ncov19, corona, covid, covid-19, virus, coronavirus, ncov2019) that were used in (Lopez et al., 2020). In order to stick to Twitter's [Terms of service] (https://developer.twitter.com/en/developer-terms/agreement-and-policy), only the Tweet IDs of the Tweets collected are made available for non-commercial research use only.

The only keyword used hydrate tweets for the month of January was "Coronavirus" as there was less talk of the pandemic at that time. As news about the Coronavirus spread, additional keywords were added to the search list.

Month	**Keyword(s)**
January	Coronavirus, virus
February	Coronavirus, virus ncov19, ncov2019
March	coronavirus, virus, covid, ncov19, ncov2019
April	coronavirus, virus, covid, ncov19, ncov2019

Table 3.1: Shows the various months and the keyword used to hydrate the tweets.

The keywords, presented in Table 3.1, used for search tweets are: virus and coronavirus since 22 January, ncov19 and ncov2019 since 26 February, Coronavirus, virus, ncov19, ncov2019 since 7 March 2020 and all keywords were used to hydrate tweets for the month of April. A total of 71,908 tweets were sampled out of 115,000 tweets across the four-

month period that this paper focused on. Since there was a disproportionate amount of data collected in January compared to other months this data was excluded from the analysis. Moreover, twitter API can provide tweets up to 7 days so we ensured that there was a lag of 7 days in the dataset to make sure enough tweets were hydrated. It is worth noting that our dataset does not capture every tweet on twitter related to the Covid-19 keywords used for hydration due to Twitter's limits on how much tweets can be hydrated every 15 minutes.

However, it is also worth noting that there were some inconsistencies in our data collection process. For example, only tweets in English were hydrated from 22 January to 31 January, 2020, after this brief period we found an algorithm that could collect tweets in all languages. Our data collection technique could also track other keywords unrelated to Covid-19 which resulted in fewer tweets relating to Coronavirus in our dataset in the first few weeks.

3.1 Datasets

Obtaining a social bot dataset can be cumbersome due to the challenge in obtaining conclusive ground truth (Morstatter et al., 2016). Two labeled datasets are used for ground truth and serve as the training datasets: social honeypot and RTbust. The trained datasets are used to detect bots with three test datasets (1) Fame for sale, (2) BotWikiCelebrity, and (3) COVID-19. The classes for the Fame for sale and BotWikiCelebrity datasets are known and the trained datasets are used to evaluate classification accuracy against data where the classes are known. The COVID-19 data is unlabeled. While several ways to detect bots have been put forward, we use two approaches to label Twitter users as bots or humans: (1) social account features and (2) sentiment features. To train and test our model, we

selected five (5) datasets of verified human and bot account from Bot Repository (https://botometer.osome.iu.edu/bot-repository/datasets.html). We use Weka machine learning tool to test for prediction accuracy to help select the best labelled dataset that can be used as a training set for classification in this paper. The nature of the datasets and how we collected the five (5) datasets have been summarized below:

3.1.1Training Dataset

3.1.1.1 The Social Honeypot Dataset

We use the Social Honeypot dataset as a training set in this paper. We chose the social honeypot dataset because of its high prediction accuracy i.e., 99%. (Lee, Eoff, and Caverlee, 2011) created a honeypot that could attract content polluters in Twitter. (Lee, Eoff, and Caverlee, 2011) generated and deployed 60 social honeypot accounts in Twitter whose function was to act like Twitter users and report what accounts follow or otherwise communicate with them. (Lee, Eoff, and Caverlee, 2011) manipulated how frequent the honeypot account post and the sort of content that these accounts post on Twitter. The author's manipulation system ran from December 30, 2009 to August 2, 2010 and a total of 22,223 polluters and 19,276 legit users were detected from 5,643,297 tweets. (Lee, Eoff, and Caverlee, 2011) created a wide variety of user account features that were a part of one of four groups:

- **UD** screen name length, description length, account age
- **UFN** following count, follower count, the ratio of the number of following and the number of followers, bidirectional friend's percentage
- **UC** statuses count per day

- **UH** following change rate

(Lee, Eoff, and Caverlee, 2011) tested 30 classification algorithms using Weka machine learning toolkit on five user account features (i.e., screen name length, description length, followers count, following count, and statuses count) and found their results consistent with accuracy ranging between 98% to 95%. (Lee, Eoff, and Caverlee, 2011) used these five categories of features as these features produced the highest accuracy results in their experiment. Table 3.2 shows a breakdown of content polluters and legit users that was detected by the manipulation model built by (Lee, Eoff, and Caverlee, 2011).

Class	User Profiles	Tweets
Polluters	22,223	2,380,059
Legit Users	19,276	3,263,238

Table 3.2: Social Honeypot Dataset.

3.1.1.2 RTbust Dataset

We use the RTbust dataset as our second training set in this paper. With a prediction accuracy of 100%, (Mazza et al., 2019) had access to all Twitter metadata fields for each tweet, retweet and user in their dataset. To collect this dataset, the authors used Twitter Premium Search API to build a complete dataset using the following query parameters: lang: IT and is: retweet. The authors carried out a manual annotation of a small subset of the dataset to see the extent to which their technique was capable of correctly spotting bots and ended up with an almost balanced annotated dataset, comprising of 51% bots and 49% human accounts. The authors dataset consists of Italian tweets shared in a 2-week period specifically between 17 and 30 June, 2018. The authors dataset consisted of 9,989,819

retweets, shared by 1,446,250 different users. (Mazza et al., 2019) observed that on an average each user in their dataset retweeted about 7 times per day which was in line with current statistics that reported daily retweets between 2 to 50 for legitimate users. (Mazza et al., 2019) argue that although their dataset is mainly Italian, the analytical approach and the data collection process is strictly language independent. We use all 14,640,084 tweets from 1000 annotated accounts from the RTbust dataset in this paper.

3.2 Test Datasets

3.2.1 Fame for sale Dataset

The fame for sale: Efficient Detection of fake Twitter followers on twitter was used as a testing dataset in this paper. (Cresci et al., 2015) set up a project to recruit Twitter users to voluntarily join an academic study for discovering fake followers on Twitter. This initiative was set up by (Cresci et al., 2015) to create a dataset of verified human accounts on Twitter. (Cresci et al., 2015) launched a verification phase on the 574 human accounts and named this initiative as the "the fake project" dataset. The #elezioni2013 (E13) was also created by (Cresci et al., 2015) and it is made up of active Italian Twitter users, with different professional profiles and belong to assorted social classes.

To create their bot dataset, (Cresci et al., 2015) purchased 3000 fake accounts in April, 2013 from different Twitter online markets. To be specific, the authors purchased 1000 fake accounts from *http://fastfollowerz.com*, 1000 from *https://intertwitter.com* and 1000 fake accounts from *http://twittertechnology*.com. To create our legitimate user dataset, we sampled 235 out of 574 human accounts from "thefakeproject" (TFP) and 964 out of 1488 from the #elezioni2013(E13) verified human dataset. We created our bot dataset by

selecting all 1335 fake followers from the "intertwitter" (INT) dataset. Therefore, a total of 1199 legitimate accounts and 1335 fake accounts were used for the test dataset. The account details for the test dataset as well as the number of followers and friends are provided in Table 3.3

Dataset	Accounts	Followers	Friends
TFP (@TheFakeProject	235	183,166	152,664
E13 (#elezioni2013)	964	797,432	420,450
INT (intertwitter)	1335	22,518	517,100
Human Dataset **Bot Dataset**	1199 1335	980,598 22,518	573114 517,100
Testing Dataset	2534	1,003,116	1,090,214

Table 3.3: Shows statistics about total collected data for testing.

3.2.2 BotwikiCelebrity Dataset

The performance of the social honeypot dataset was not encouraging so we created the BotwikiCelebrity dataset as another test dataset to see if the classification framework from the social honeypot dataset can accurately distinguish between what is a human and what is a bot. We performed a cross-dataset analysis by using uploaded bot dataset on Bot Repository to create our final testing dataset. To create our final testing dataset, we merged the Self-identified bots (botwiki-verified) dataset from (Yang et al., 2020) and Celebrity account collected as authentic users (celebrity) dataset from (Onur et al., 2019) to create a new testing dataset. One way to analyze different labeled dataset is to look at the datasets in feature space (Yang et al., 2020). Visualizing the two datasets together was difficult as there were too many data points so instead we sampled 500 out of 699 verified bots from

the botwiki dataset and 500 out of 20,984 verified human accounts from the celebrity dataset to create a balanced dataset.

3.2.3 COVID-19 Dataset

Using Twitter's API, we hydrated tweets relating to Covid-19 from January to April to build the Covid-19 dataset. To quantify text and make sure certain characters are not counted, we removed characters such as (;, :, *. '', ,|,\, {,[, spaces etc.) from the text attribute. Using the user_attribute_string function we extracted the user attributes such as user_id, description, friend count, follower count etc. from the user column to create a total of 71,908 tweets out of 115,000 tweets that we collected from January to April.

Dataset	**Number of bots**	**Number of humans**	**Data points**	**Account Features**
Social Honeypot	22,223	19,276	41,499	5
RTbust	190	209	399	8
Botwiki	698	0	698	8
Celebrity	0	5971	5971	8
Covid-19	Unlabeled	Unlabeled	71,908	8

Table 3.4: Shows the datasets used for our experiment.

3.3 Hypotheses

To achieve the stated objectives in section 2.1, the following four hypotheses are investigated:

H_1: The spread of misinformation or disinformation by bots regarding content related to COVID-19 will be higher than the spread of misinformation or disinformation by humans. While it is known that bots spread low quality information on Twitter. (Shao et al., 2018),

we do not attempt to distinguish between misinformation and disinformation. It is therefore our objective to analyze the percentage of social media bots in our examined Covid-19.

H_2: The accuracy to detect misinformation by bots will be higher using twitter features such as favorite count, listed count, and topic distribution as compared to social honeypot features. In this paper, we propose three new features: (1) topic distribution, (2) listed count and (3) favorite count.

These three user account features have been rarely used in prior research. Topic distribution on Twitter has to do with the variety of sentiments expressed by users on any given issue. Listed count is a curated group of Twitter accounts, and favorite count is the number of accounts a Twitter user has favorited.

Most prior studies have relied on well-known Twitter user account features such as count, friend count, Tweet count, name length, account age and description length. These features are considered to be top features with the highest predictive power and few research studies have investigated features such as listed count, favorite count, and screen name length as they have a lower predictive power when it comes to distinguishing between a bot and a human (Varol et al., 2017; A. H. Wang, 2010; Cresci et al., 2015; Lee, Eoff, and Caverlee, 2011). Favorite count, listed count and screen name length will be used in our training and testing experiment to see how well they improve our classification algorithm.

H_3: The distribution of different topics will be greater for humans compared to bots. We expect humans to have a wider variety of topics expressed in Twitter as compared to bots.

Our reasoning is that we think bots are much more likely to have a target or an agenda that needs to be talked about to change the economic, political or social setting of an online platform. Hashtags will be extracted from tweets text for all tweets hydrated between 1st February, 2020 to 31st April 2020. Bot sentinel will be used to estimate hashtags with emerging popularity to help us test our hypothesis 3.

H_4: Detected bots will express more negative sentiments compared to humans.

The rationale behind this is that we believe bots are more likely to engage in creating negative inflammatory content compared to humans (Stella, Ferrara, & De Domenico, 2018).

3.4 Experiments

We first aim to replicate the results found in (Cresci et al., 2015) and the merged BotwikiCelebrity dataset by using the classification framework we build using the Social Honeypot Dataset and the RTbust Dataset. Comparing the results obtained through the experiment and the ones reported in (Cresci et al., 2015), (Yang et al., 2020) and (Onur et al., 2019) will increase the level of confidence in the hybrid approach that this research will rely on. To achieve this objective, this paper will test ten features (see Table 3.5) as seen in (Cresci et al., 2015), (Yang et al., 2020) and (Onur et al., 2019).

Twitter Attributes	Description	(1)	(2)	(3)	(4)
Name	Twitter defines "name" as the name of the user	x		x	x
Listed count	Curated group of Twitter accounts	x		x	x
Favorite count	The number of accounts a user has favorited	x		x	x
Statuses count	The number Tweets (including retweet) issued by a specific user	x		x	x
Sentiment	Sentiments expressed on a given subject	x	x	x	x
User id	User features based on Twitter meta-data	x	x	x	x
Screen name	Handle or alias that a specific user identifies with	x	x	x	x
Follower count	The total number of Twitter users that follow a specific user	x	x	x	x
Friends count	Total number of Twitter users that follow a specific user	x	x	x	x
Description Length	Total character count of a description in a user profile	x	x	x	x

Table 3.5: Shows the features that will be used in this study. Features that are not seen in the Social Honeypot dataset are shaded in grey. New feature proposed shaded in light blue. Features that are seen in all the datasets used in this study are are not shaded. Datasets*:* (1) = **Fame for sale**, (2) = **Social Honeypot**, (3) = **Self-identified bots**, (4) = **RTbust**

These ten features are highly predictive and capture several suspicious behaviors which make it easier to differentiate between a bot and a human account using a bot detection algorithm (Ferrara et al., 2016). If the results demonstrate potential, this study will implement a new method to detect social bots by using a hybrid approach that incorporates sentiment features

and user account features. The new approach that this study seeks to implement will then be tested on the COVID-19 dataset of millions of tweets between February, 2020 and April, 2020. We explain how we conducted our experiments in subsections 3.4.1 – 3.4.4.

3.4.1 Experiment I

We use the Social Honeypot Dataset and the RTbust dataset as our baseline dataset in our first experiment. Using Weka machine learning tool (Witten et al., 2005), we followed the same classification framework used by the authors in (Lee, Eoff, and Caverlee, 2011) and (Mazza et al., 2019) to see what the dataset's prediction accuracy is. We tested 20 classification algorithms, such as random forest, naive Bayes, logistic regression and tree-based algorithm, all with default values for all parameters using 10-fold cross validation. 10-fold cross validation is a way of dividing the original data into 10 equally-sized sub-samples, and executing 10 training and validation procedures (Lee, Eoff, and Caverlee, 2011).

3.4.2 Experiment II

In experiment II, we take the best training dataset from Experiment I to classify the Fame for sale dataset. The social honeypot dataset or the RTbust dataset would be supplied as a training set for the Fame for sale dataset. We test for accuracy by replacing the class labels in the Fame for sale dataset with questions marks (?). Using Weka machine learning tool (Witten et al., 2005), we try to replicate the results found in the Fame for sale dataset by using our training datasets from experiment I. If the prediction results are not encouraging, we aim to do a cross data analysis on Bot Repository to create a final testing dataset.

3.4.3 Experiment III

Results from Experiment II will show which training dataset will be used for our unlabeled Covid-19 dataset in Experiment III. We aim to classify the unlabeled Covid-19 dataset with the best classification framework from Experiment II. An independent data analysis would be done by randomly comparing 7000 detected tweets from bots and 14,000 detected tweets from humans out of a total of 39,091 tweets to understand the sort of misinformation or information that were being amplified between February and April. We intend to achieve this by using an online sentiment tool known as Bot Sentinel. A fact checking tool known as Poynter would also be used to check for misinformation in Experiment III.

3.4.4 Experimental Steps

Initially, the plan for this research was to study and collect data on Twitter users that were actively posting Covid-19 tweets overtime but that was time consuming and very expensive to achieve. We therefore rely on a dataset that we collected from January to April to test the new approach that we propose for this paper. However, based on pilot studies we conducted it was determined that tweet length and sentiment expressed over time was the most accurate method for distinguishing a bot from a human. As a result, we relied on daily tweet length plus all other user account features to improve the accuracy of our results.

Additionally, there are other user account features such as location, verified, protected, default profile image among others were discarded as there was little to no information to collect for these features. To deliver accurate results, this study will try to duplicate the results found in (Cresci et al., 2015); (Onur et al., 2019) and (Yang et al., 2020) by using

behavioral patterns established in (Mazza et al., 2019) and (Lee, Eoff, and Caverlee, 2011) to see if the experiment can come up with the same results seen in (Cresci et al., 2015) and (Yang et al., 2020).

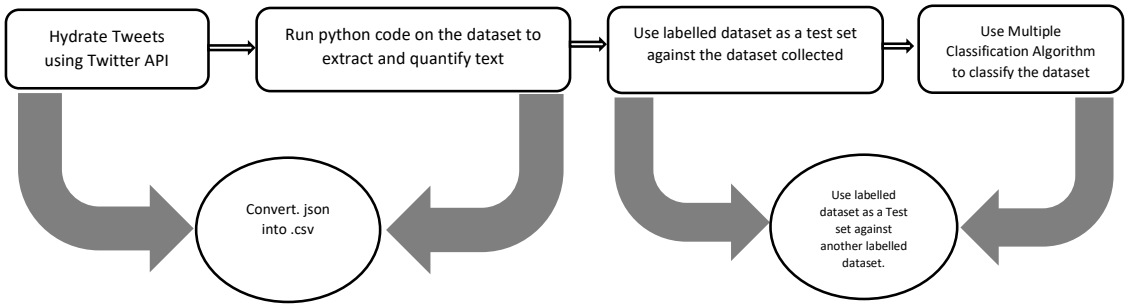


Figure 3.1: A graphic displaying our research plan.

Comparing the results obtained through the experiment and the ones reported in (Cresci et al., 2015); (Onur et al., 2019) and (Yang et al., 2020) will increase the level of confidence in the hybrid approach that this research will rely on. Figure 3.2 is a graphic display of the steps that will be taken to achieve the goals of this paper.

Chapter 4 Experimental Results

For the following experiments metadata associated with user accounts are used as the features to detect bots. Tweets and tweet content is not used in the subsequent analyses.

4.1 Experiment I

For the Social Honeypot dataset, we found the results from Weka consistent, with a prediction accuracy ranging from 100% to 87% across most classifiers (15 out 20 tested). For the other 5 out of 20 tested, accuracy ranges between 84% to as low as 53%. Since the Social Honeypot dataset has fewer features as compared to the RTbust dataset, we first tested the same features seen in the social honeypot dataset for the RTbust dataset. Prediction accuracy for the RTbust dataset ranges between 100% to 63% across most classifiers (15 out 20 tested) and for the other 5 out of 20 tested, accuracy ranges between 63% to as low as 52%.

We observed that the strength of the classification lies primarily in the preference of features used. Tree-based classifiers generated the best accuracy results. We also observed an increase in prediction accuracy from 5% to 20% across most classifiers when all features are used for the RTbust dataset. While it is clear that Tree-based classifiers produced the best accuracy results, we observed that classification accuracy significantly drops after Random Forest classifier in Table 4.1.

To understand why that is the case, we examined the nature of the social honeypot and the RTbust dataset to come up with possible explanation as to why that happened. It is worth

Classifier	Social Honeypot	RTbust (Same Features as Social Honeypot)	RTbust (all features)
Random Tree	100%	100%	100%
Kstar	100%	100%	100%
IBk	100%	100%	100%
Random Forest	99.9%	100%	100%
REPTree	93.5%	82.2%	79.1%
J48	93%	81%	85.2%
LMT	92.5%	79%	81.7%
Decision Table	92.5%	69.6%	79.9%
JRip	92.7%	78.6%	82.9%
PART	92.4%	72.6%	82.2%
Multilayer Perception	91.7%	62.4%	81.9%
BayesNet	89.6%	69.6%	79,6%
SGD	89.4%	63.1%	75.9%
SimpleLogistic	87.1%	63.6%	76.1%
Logistic	87.1%	63.6%	75.6%
SMO	84.9%	59.6%	72.4%
OneR	81.1%	77.9%	81.4%
NaiveBayes	72.7%	52.6%	56.2%
NaiveBayesMultinominal	56.8%	58.1%	62.9%
ZeroR	53.5%	52.3%	52.3%
Average	88%	74%	80%

Table 4.1: shows the prediction accuracy of our two baseline datasets.

noting that some classifiers work well with smaller dataset while others do well with large datasets. NaiveBayes, Logistic Regression, ZeroR etc. works well when the dataset is small

as these classifiers has enough room to construct the decision boundary (*Text Classification with Extremely Small Datasets | by Anirudh Shenoy | Towards Data Science*, 2019.).

On the other hand, Tree-based classifiers and Random Forest work well with large datasets as they require little data preparation and can handle both numerical and categorical data (*7 Types of Classification Algorithms - Analytics India Magazine*, 2020). (Kirubavathi Venkatesh & Anitha Nadarajan, 2012), (Lee, Eoff, and Caverlee, 2011), (Ji et al., 2016), (Cresci et al., 2015), (Yang et al., 2020.), (Onur et al., 2019), (Davis et al., 2016) among others have shown that Random Forest classifier is the best classifier when it comes classifying a large Twitter dataset. For the purpose of this research, Random Forest has been used to examine the results of this study.

To see which additional user account feature improved the prediction power for the RTbust dataset, using Weka machine learning tool we compare the performance of the features that are not used in the social honeypot dataset to the features used in the social honeypot dataset. We grouped the user account into two categories with Category 1 being the features used in the social honeypot dataset and Category 2 representing the features that are not seen in the social honeypot dataset. We used Tree-based classifiers only to test and compare these categories of features. The categories and the results are shown below:

Category 1: (nonsocial honeypot features): listed count, favorite count, length of name and number of tweets

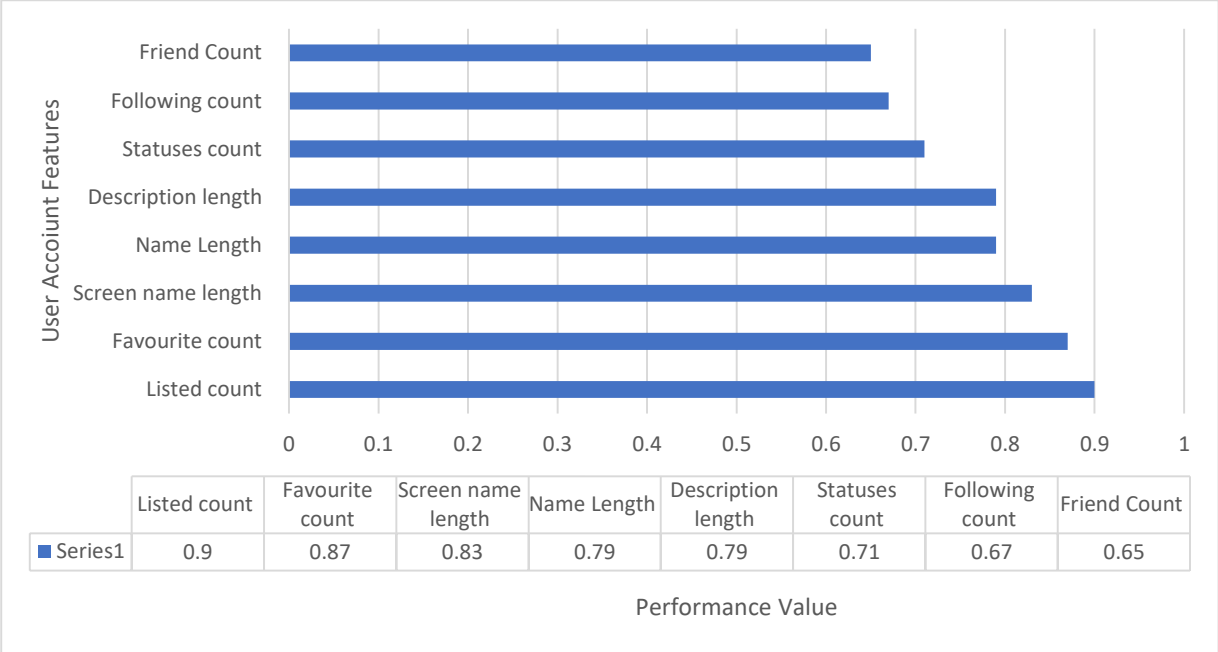
Category 2: (social honeypot features): follower count, following count, length of screen name and description length

Classifier	RTbust (all features)	Category 1 (social honeypot)	Category 2 (not social honeypot)
Random Tree	100%	100%	100%
Random Forest	100%	100%	100%
J48	85.2%	81%	80%
LMT	81.7%	79%	81%
REPTree	79.1%	82%	83%
Average	89%	88%	89%

Table 4.1.1: Prediction accuracy using all features, Category 1 and Category 2 .

We observed that features found in Category 1 performed better than those found in Category 2. In general, there was a 1% to 2% prediction accuracy increase across most Tree-based classifiers that were used. Figure 4.1.1 also shows that the RTbust dataset performs better when all user account features are used. Figure 4.1.1 and Figure 4.1.2 shows user account features from the RTbust and Social Honeypot dataset and their order of importance. The x-axis shows the user account features and their respective values.

Figure 4.1.1 shows that listed count, favorite count, screen name length, name length and description length improve the prediction accuracy of the RTbust classification framework as compared to statuses count, following count and friend count. Figure 4.1.2 on the other hand shows consistent performances from the features used for the social honeypot dataset.



	Listed count	Favourite count	Screen name length	Name Length	Description length	Statuses count	Following count	Friend Count
Series1	0.9	0.87	0.83	0.79	0.79	0.71	0.67	0.65

Figure 4.1.1: Shows RTbust user account features and their performance values.

To understand why performance values are not consistent across the features used in the RTbust dataset, we examined the nature of the RTbust dataset and noticed that there was not much difference statistically when we compare the following count, friend count and statuses count of bots to that of humans. There was however a major statistical difference when we compare bots to humans using listed count, favorite count, screen name length and name length. It is worth noting that Figure 4.1.1 and 4.1.2 shows the order of importance of features for both dataset and does not necessarily mean that features with low performance values are not good for making predictions. RTbust dataset for example shows that prediction accuracy is high when all features are used compared to when few are used.

However, the social honeypot dataset performed poorly when used as a training dataset for the fame for sale dataset because the dataset does not have 3 out of the top 4 performing features which are better for classifying unlabeled data and this explains why the RTbust dataset is the best for classifying the unlabeled Covid-19 dataset.

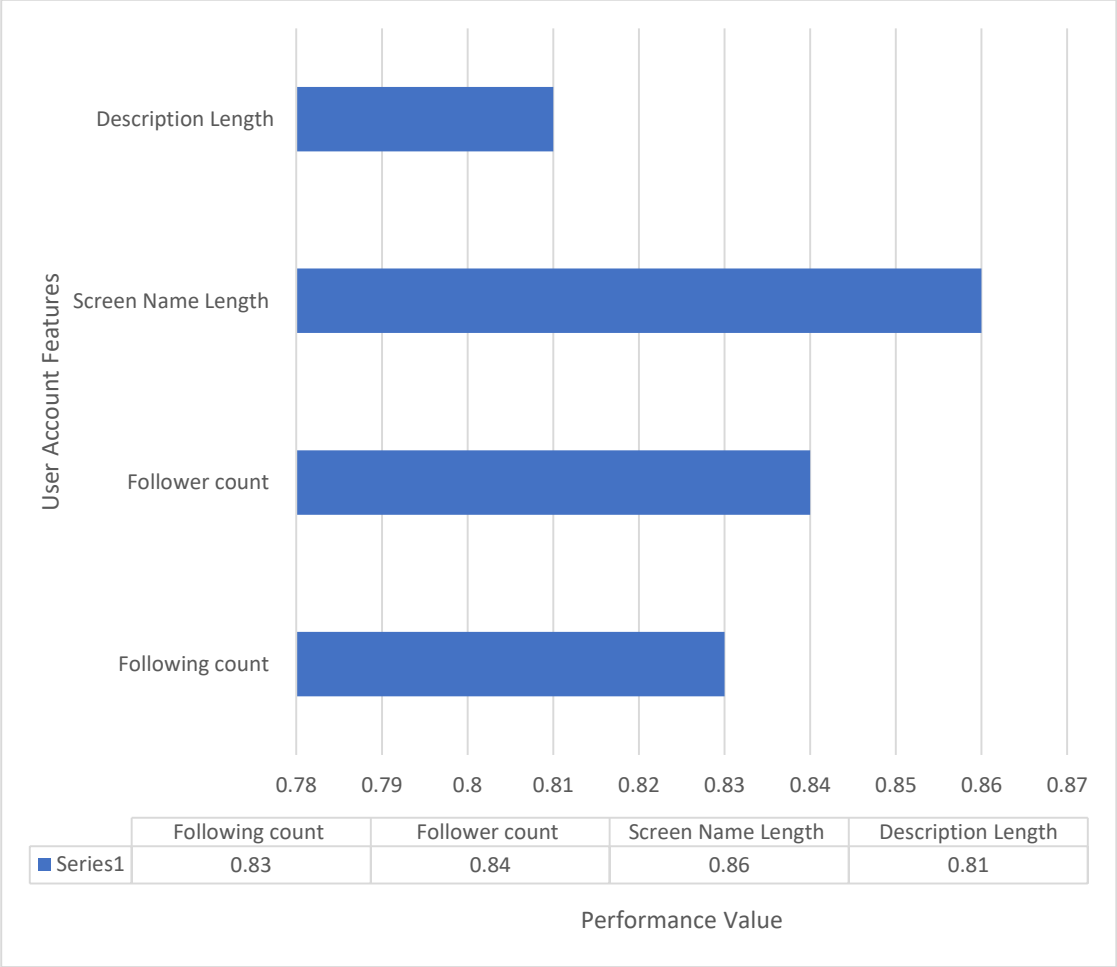


Figure 4.1.2: Shows Social Honeypot user account features and their performance values.

4.2 Experiment II

In Experiment II, we first used each of our baseline datasets as a training set for our testing dataset. The social honeypot dataset was supplied as a training set for the Fame for sale dataset. The social honeypot dataset with a 99% prediction accuracy correctly classified all accounts that were verified as bots in the Fame for sale: efficient detection of fake Twitter follower's dataset but misclassified 1284 human accounts as bots. The social honeypot dataset could detect only 196 out of a total of 1199 verified human accounts as humans. Table 4.3 shows that the social honeypot dataset performed poorly when used as a training set for the Fame for sale dataset.

Using random forest classifier, the social honeypot dataset achieved a 47.6% precision and 51% accuracy. With so many incorrect classifications, we think the reason is due to the fact that the social honeypot dataset is an old dataset which does not have other features like name length, listed count, favorite count, reply count etc. that can be relied on to improve detection accuracy. Additional factors that contributed to the poor results from our first experiment will be discussed further below. The confusion matrix for the Social Honeypot and RTbust data is presented in Table 4.2.1 and Table 4.2.2.

Results	True Positive	True Negative
Predicted Positive	1335 (TP)	1003 (FP)
Predicted Negative	0 (FN)	196 (TN)

Measure	Value	Derivations
Sensitivity	1.0000	TPR = TP / (TP + FN)
Specificity	0.1635	SPC = TN / (FP + TN)
Precision	0.5710	PPV = TP / (TP + FP)
Negative Predictive Value	1.000	NPV = TN / (TN + FN)
False Positive Rate	0.8365	FPR = FP / (FP + TN)
False Discovery Rate	0.4290	FDR = FP / (FP + TP)
False Negative Rate	0.0000	FNR = FN / (FN + TP)
Accuracy	0.6042	ACC = (TP + TN) / (P + N)
F1 Score	0.7269	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.3055	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.2.1: Confusion matrix for the result from our Social Honeypot testing dataset. The metrics used were:screen_nameLength, description_length, following_count, friend_count,and statuses_count.

Results	True Positive	True Negative
Predicted Positive	1308 (TP)	112 (FP)
Predicted Negative	27 (FN)	1087 (TN)
Measure	Value	Derivations
Sensitivity	0.9798	TPR = TP / (TP + FN)
Specificity	0.9066	SPC = TN / (FP + TN)
Precision	0.9211	PPV = TP / (TP + FP)
Negative Predictive Value	0.9758	NPV = TN / (TN + FN)
False Positive Rate	0.0934	FPR = FP / (FP + TN)
False Discovery Rate	0.0789	FDR = FP / (FP + TP)
False Negative Rate	0.0202	FNR = FN / (FN + TP)
Accuracy	0.9451	ACC = (TP + TN) / (P + N)
F1 Score	0.9495	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.8916	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.2.2: Confusion matrix for the results of our second testing dataset. Metrics used were: favorite count, listed count, name length, and number of tweets.

With precision and accuracy at 72% and 80% respectively, the RTbust dataset performed better than the social honeypot dataset.

Dataset	Accuracy	F1
Social Honeypot	60%	0.73
RTbust	94%	0.95

Table 4.2.3: Comparison of accuracy and F1 for classifying Fame for Sale with the following features from social honeypot: tLengthofName, tLengthofScreenName, tNumberofListedCount, tNumberoffavoriteCount and tNumberofStatusesCount.

In our next RTbust classification experiment, we used the features that generated the best predictive accuracy (i.e., listed count, favorite count, length of name, length of screen name and statuses count) to see if the classification results would improve. The confusion matrix for the classification framework is shown in Table 4.2.4. With precision and accuracy at 92% and 95% respectively, we decided to add more features to see if the accuracy of the results improves. Since the strength of the classifier is dependent on the selective power of the metric used (Lee, Eoff, and Caverlee, 2011), in our next classification test, we randomly added description length. Results are presented in Table 4.2.5.

Results	True Positive	True Negative
Predicted Positive	968 (TP)	369 (FP)
Predicted Negative	168 (FN)	1308 (TN)

Measure	Value	Derivations
Sensitivity	0.8521	TPR = TP / (TP + FN)
Specificity	0.7800	SPC = TN / (FP + TN)
Precision	0.7240	PPV = TP / (TP + FP)
Negative Predictive Value	0.8862	NPV = TN / (TN + FN)
False Positive Rate	0.2200	FPR = FP / (FP + TN)
False Discovery Rate	0.2760	FDR = FP / (FP + TP)
False Negative Rate	0.1479	FNR = FN / (FN + TP)
Accuracy	0.8091	ACC = (TP + TN) / (P + N)
F1 Score	0.7829	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.6210	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.2.2: Confusion matrix for the results of our RTbust testing dataset. Metrics used were: follower count, following count, length of screen name and description length

Results	True Positive	True Negative
Predicted Positive	1275 (TP)	356 (FP)
Predicted Negative	61 (FN)	841 (TN)

Measure	Value	Derivations
Sensitivity	0.9551	TPR = TP / (TP + FN)
Specificity	0.7026	SPC = TN / (FP + TN)
Precision	0.7817	PPV = TP / (TP + FP)
Negative Predictive Value	0.9334	NPV = TN / (TN + FN)
False Positive Rate	0.2183	FPR = FP / (FP + TN)
False Discovery Rate	0.0449	FDR = FP / (FP + TP)
False Negative Rate	0.0202	FNR = FN / (FN + TP)
Accuracy	0.8357	ACC = (TP + TN) / (P + N)
F1 Score	0.8597	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.6858	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.2.4: Confusion matrix for the results of our second testing dataset. Metrics used were: tLengthofName, tLengthofScreenName, tNumberofListedCount, tLengthofdescription, tNumberoffavoriteCount and tNumberofStatusesCount.

We observed that precision and accuracy dropped from 92% to 78.2% and 95% to 84% respectively when description_length was introduced as a new feature. We decided to randomly add more features to see how the results changes. To do this, we added follower count, and following count to the set of features that we have already tested to observe the changes in precision and recall. The confusion matrix for the classification framework is presented in Table 4.2.6.

Results	True Positive	True Negative
Predicted Positive	150 (TP)	134(FP)
Predicted Negative	1185 (FN)	990(TN)

Measure	Value	Derivations
Sensitivity	0.1124	TPR = TP / (TP + FN)
Specificity	0.8808	SPC = TN / (FP + TN)
Precision	0.5282	PPV = TP / (TP + FP)
Negative Predictive Value	0.4552	NPV = TN / (TN + FN)
False Positive Rate	0.1192	FPR = FP / (FP + TN)
False Discovery Rate	0.4718	FDR = FP / (FP + TP)
False Negative Rate	0.8876	FNR = FN / (FN + TP)
Accuracy	0.4636	ACC = (TP + TN) / (P + N)
F1 Score	0.1853	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	-0.0107	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.2 5: Confusion matrix for the results of our second testing dataset. Metrics used were: tLengthofName, tLengthofScreenName, tNumberofListedCount, tNumberoffavorite, tTheNumberofFollowers, tTheNumberofFollowing and tNumberofStatusesCount.

We observed that the more features added to the classification framework, the less accurate our model becomes.

4.3 Experiment III

Poor results obtained from the social honeypot experiment led us to creating a final testing dataset. To understand why the social honeypot dataset performed poorly when it comes to detecting human accounts, we performed a cross-dataset analysis by using uploaded bot dataset on Bot Repository to create a new labeled dataset. We merged the Self-identified bots (botwiki-verified) dataset and Celebrity account collected as authentic users (celebrity) dataset to create a new testing dataset. One way to analyze different training dataset is to look at the datasets in feature space (Yang et al., 2020).

Visualizing the two datasets together was difficult as there were too many data points so instead we sampled 500 out of 699 verified bots from the botwiki dataset and 500 out of 20,984 verified human accounts from the celebrity dataset to create a balanced dataset. To achieve consistency, we supplied the social honeypot dataset as a training set to test the merged dataset. Using random forest classifier, the social honeypot dataset achieved a 58.9% precision and 65% accuracy (see Table 4.3.1) and this level of performance shows that no dataset can generalize well on all other datasets (Yang et al., 2020).

We observed few factors that contributed to the poor results from the social honeypot experiment. First, the datasets used had inconsistent classes. The social honeypot dataset had few features as compared to the botwiki and celebrity dataset. The few features that the social honeypot dataset has can only capture a tiny sector of a user account's characteristics. Third, the datasets used were annotated by different people with different standards using variety of methods (Yang et al., 2020).

Results	True Positive	True Negative
Predicted Positive	500 (TP)	350 (FP)
Predicted Negative	0 (FN)	150(TN)

Measure	Value	Derivations
Sensitivity	1.0000	TPR = TP / (TP + FN)
Specificity	0.3000	SPC = TN / (FP + TN)
Precision	0.5882	PPV = TP / (TP + FP)
Negative Predictive Value	1.0000	NPV = TN / (TN + FN)
False Positive Rate	0.7000	FPR = FP / (FP + TN)
False Discovery Rate	0.4118	FDR = FP / (FP + TP)
False Negative Rate	0.0000	FNR = FN / (FN + TP)
Accuracy	0.6500	ACC = (TP + TN) / (P + N)
F1 Score	0.7407	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.4201	TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN) *(TN+FP) *(TN+FN))

Table 4.3.1: Confusion matrix for the results of our second testing dataset. Metrics used were: tLengthofName, tLengthofScreenName, tNumberofListedCount, tNumberoffavorite, tTheNumberofFollowers, tTheNumberofFollowing and tNumberofStatusesCount.

To contrast results with RTbust data to detect bots the social honeypot dataset was used to classify the unlabeled COVID-19 data. We classified our COVID-19 dataset with the classification framework from the social honeypot dataset to see how many bots the classification algorithm detects in our COVID-19 dataset.

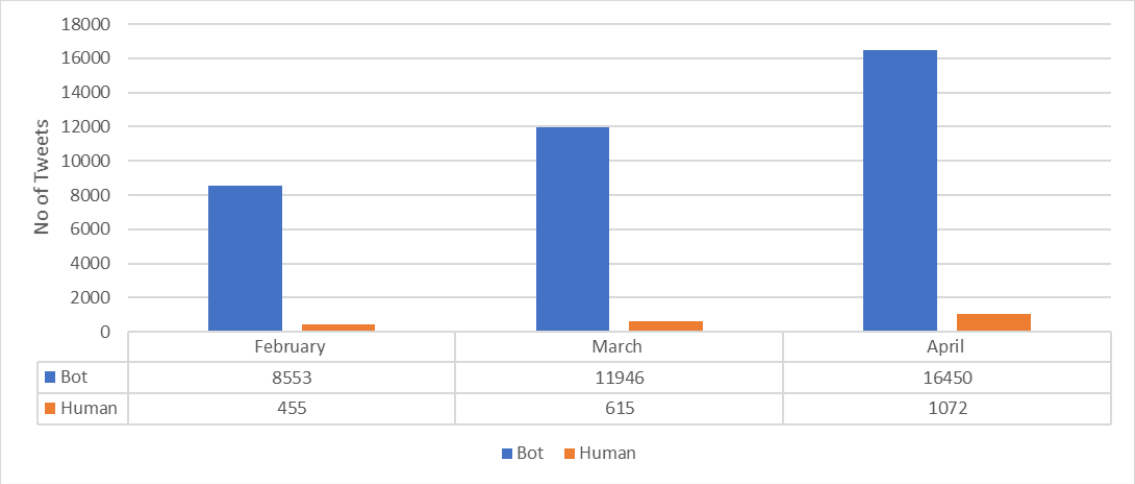


Figure 4.3.1: Covid-19 Trend analysis generated by using the social honeypot dataset as a training set. The metrics used were userID, screen_namelength, description_length following_count, friend_count, statuses_count.

Figure 4.3.1 shows that out of 39,091 tweets sampled from February to April, the model we built using the social honeypot dataset classified 36,949 user accounts as bots and 2,142 user accounts as humans. We also generated a trend analysis Trend analysis by using RTbust: Exploiting temporal patterns for botnet detection on twitter dataset to see how the model we built classifies the COVID-19 dataset. Results are inverted compared to the results obtained with the social honeypot data. Classifying the COVID-19 data with the social honeypot shows more content was created by bots compared to humans while the RTbust data suggests more human content was created compared to bots. This highlights the impact of using the wrong training data. Consequently, based on the results obtained

with RTbust from experiments I and II, it is our conjecture that RTbust provides a more accurate representation of bot generated content.

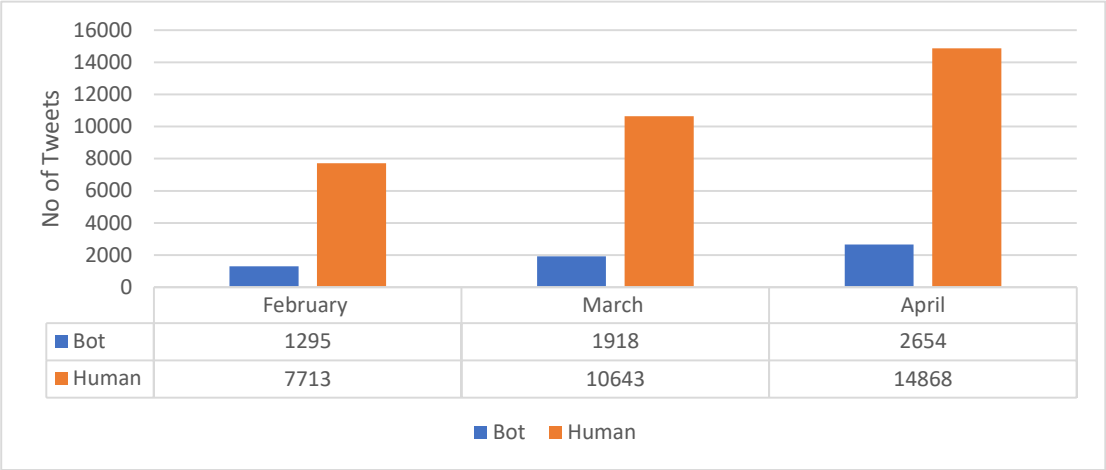


Figure 4.3.2: COVID-19 Trend analysis generated by the RTbust training set. The metrics used were userID, namelength, screen_namelength, listed count, and favorites count.

We started by using the user account features (i.e., listed count, favorite count, length of name, length of screen name and statuses count) that generated the highest precision and accuracy (i.e., 92% and 95% respectively) from the training and testing dataset experiment. Figure 4.3.2 shows that out of 39,091 tweets from February to April, the model we built using the RTbust dataset detected 5,867 user accounts as bots as compared to 36,949 bots detected by the model that we built using the social honeypot dataset. The RTbust classification model also detected 33,224 accounts humans as compared to 2,142 detected human accounts by the social honeypot classification framework

Based on the results in Table 4.3.2, it can be observed that there was a 43% increase in bot generated content with the RTbust data compared to only a 39% increase in bot generated content based on the social honeypot data. Even though the social honeypot data shows a

Month	# bots Social Honeypot (ACU < 0.5)	# bots RTbust (ACU > 0.94)
February	8553	1295
March	11946	1918
April	16450	2654
Average % increase (Feb-April)	39%	43%

Table 4.3.2: Shows the monthly classification of bots from the model we built using our training datasets.

greater number of bots each month compared to humans the percentage increase is actually lower. Also, we know that the social honeypot is not accurate but even still we can show that the average increase in number of bots is greater.

The model we built using social honeypot dataset (ACU < 0.5) misclassified most user accounts as bots while the RTbust model generated the results that we expected to see with an accuracy at 95%. We observed that between February and March, the number of bots detected increased by 32.4% from 1,295 to 1,918 and by 27.8% between the month of March to April. Figure 4.3.2 also shows an upward trend of legitimate users that were tweeting about the Coronavirus pandemic. A likely cause for this upward trend in human generated content could be due to several factors such as, high unemployment rates across all states, lockdowns and school shut downs. Twitter for example, has gained 14 million additional users from the end of 2019 to the start of 2020 which is 24% higher than from the end of 2018 to the start of 2019 (The Washington Post, 2020).

We also observed a greater percentage increase in the number of bots detected with RTbust training set (43%) as compared to Social Honeypot training set (39%) even though the social honeypot shows a greater number of bots detected each month compared to humans, the percentage increase is actually lower. Also, we know that the social honeypot is not accurate but even still we can show that the average increase in the number of bots is greater.

4.4 Misinformation and Topic Analysis

We performed an independent data analysis by randomly selecting 7000 detected tweets for bots and 14,000 detected tweets for humans to see the sort of information or misinformation that was been disseminated between January and April. Table 4.4.1 shows the number of detected bots and humans that we randomly selected for topic analysis and misinformation from our COVID-19 dataset.

Topic Analysis	**Misinformation Analysis**
Humans (N=1000)	Humans (N=14,000)
Bots (N=1000)	Bots (N=7000)

Table 4.4.1: shows the sample size for topic analysis and misinformation analysis

4.4.1 Bots

Bots were identified using the optimal features discussed in section 3 of this paper. Using Bot Sentinel, we matched some of the most used hashtags from the user ids like #coronavirus, #Covid-19, #Trump2020, #MAGA, #WWG1WGA, #TheGreatAwakening,

and #DarkToLight. Bot Sentinel is a free platform created to spot and track trollbots and malicious and untrustworthy Twitter accounts. Bot sentinel makes use of machine learning and artificial intelligence to observe Twitter accounts and classify those accounts being studied as social bots or not. Bot Sentinel stores these detected accounts in a database so that developers can extract these accounts for further studies. Bot Sentinel also acts a disinformation and misinformation tool by tracking, identifying and tagging malicious accounts that may be spreading false information (*Bot Sentinel* , 2019).

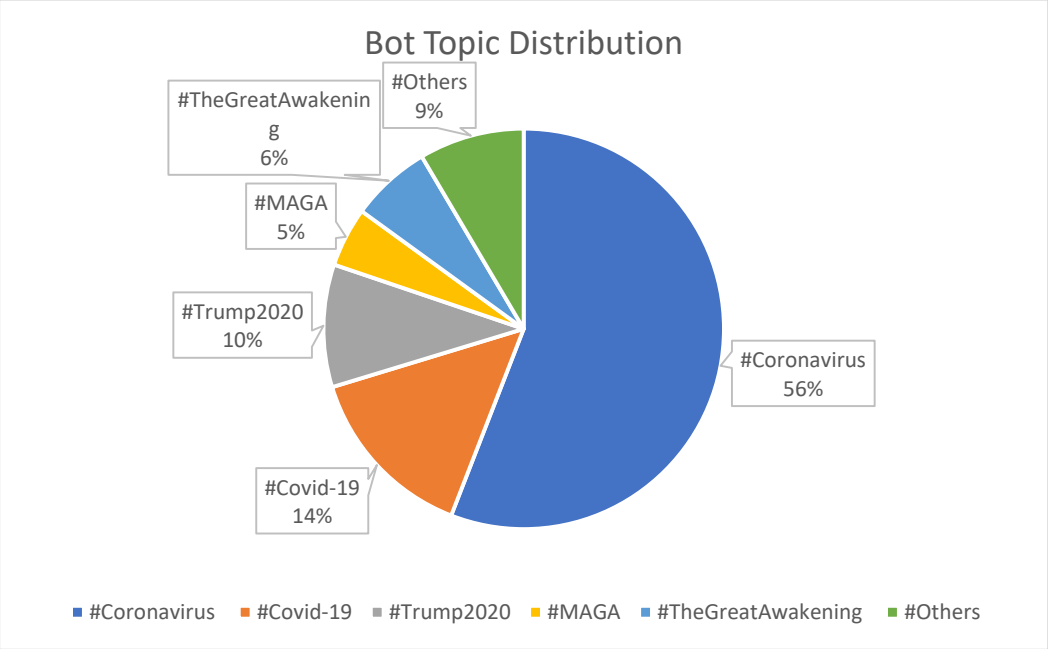


Figure 4.4.1:Shows the most used hashtags by bots in our COVID-19 dataset from February 2020 – April 2020.

Figure 4.4.2 shows that out of the 1000 bots we selected, 56% of the bots that were detected by our classification model were engaged in some form of conspiracies and political propaganda when we looked at some of the tweets that were posted. 14% were engaged in the discussion of the American public health. About 10% of the bots detected were engaged in Trump and 5G conspiracies. The rest of the bots detected (#others) were engaged in the

spread of other misinformation like “COVID is a hoax”, “bleach is a COVID cure”, “wearing a mask increases your chances of getting COVID” etc. Online fact checking tools like Poynter (http://www.poynter.com) and Bot Sentinel (*Bot Sentinel, 2019*) were used to detect misinformation generated by bots. The spread of conspiracies on online social media platforms is a well-established issue (E. Ferrara, 2020). It is worth noting that the actual number of coronavirus related bot tweets are probably higher, as Bot Sentinel only identifies hashtag terms (such as #Trump2020) and ignores “Trump2020” or “COVID-19”.

We also matched 1000 detected human tweets to see what sort of information or misinformation that was been disseminated. COVID-19 has had a significant impact on the quality of life in the US and around the world. During the months from February 2020 thru April 2020 (period for our dataset), there were more than 60,000 deaths in the US, unemployment level at 40 million, lockdowns and state of emergencies in all 50 states. Consequently, as the pandemic became more widespread more online information was being generated. Based on the analysis it can also be observed that there was an upward trend of misinformation from the month of February to April in Figure 4.4.3.

To estimate the amount of misinformation during the months of February through April of 2020, tweets generated by humans is also analyzed.

4.4.2 Humans

Figure 4.4.3 shows that about 25% of detected human tweets were engaged in the discussion of general health and self-care issues (#CoronavirusIsTheTruth, #TheGreatAwakening, #MAGA, #WWG1WGA, #quarantineandchill, #toiletpapercrisis, #workfromhome, #Fauci), 22% were engaged in the discussion of the American public

health, about 17% were engaged in the spread of conspiracies and political propaganda, 15% were engaged in WHO, Wuhan, vaccine and Trump issues, 11% were engaged in 5G and Covid-19 conspiracies, and the rest of the human tweets detected were engaged in the discussion of variety of topics (#others).

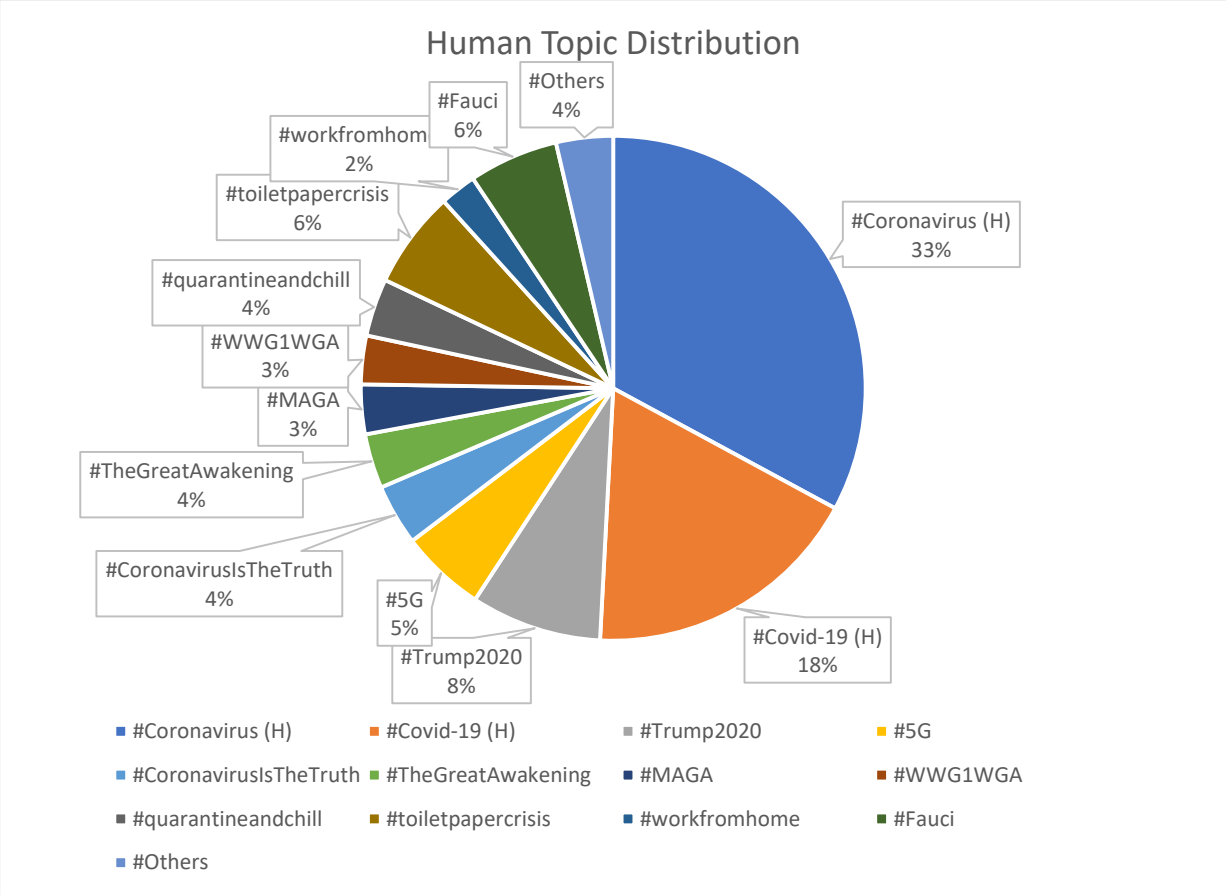


Figure 4.4 2:Shows the most used hashtags by bots in our Covid-19 dataset.

Figure 4.4.3 also shows that humans were engaged in a wide variety of topics as compared to bots. One notable distinction between bots and human tweets from Figures 4.4.2 and Figure 4.4.3 is that bots tend to be more narrowly focused on a small number of hashtags compared to humans.

This suggests bots may apply a more targeted or localized approach for spreading misinformation. In contrast the topic distribution for human tweets during the 3-month period analyzed consists of greater diversity of topics. The difference in topic distribution could be due the imbalance of data used for the analysis. Fewer tweets were analyzed for bots compared to humans. To address this issue, our data was normalized in the following analysis to more accurately measure the differences observed in topic distributions (see Figure 4.4.2.1).

Bot Sentinel (*Bot Sentinel, 2019*) and Poynter (http://www.poynter.com) were used to check for disinformation and misinformation that were disseminated by humans and bots on Twitter between February and April. 14,000 tweets for humans and 7,000 tweets for bots were randomly sampled from February 1, 2020 to April 30, 2020. A total of 21,000 tweets were used for misinformation analysis. We observed that #Coronavirus, and #Covid-19 are the most used hashtags with the most misinformation. The #other category is made up of other hashtags that were infrequently used by humans such as #Wuhan, #Virus, #Fauci etc.

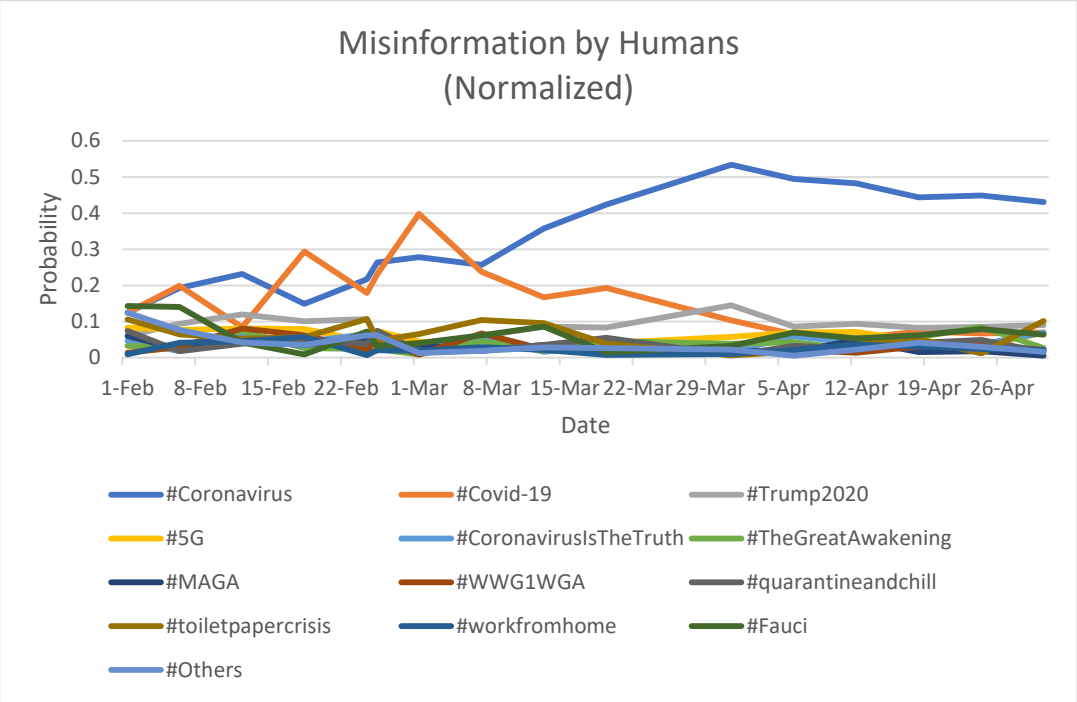


Figure 4.4.3: Misinformation by humans over time (1000 tweets)

4.4.2.1 Reasons for misinformation disseminated by humans

Possible reasons for misinformation disseminated by humans are grouped into four categories which have been explained below:

1. Ignorance

Ignorance was one crucial reason for the dissemination of wrong information in Twitter between Feb 1st to Feb 12th, 2020. We observed that at that time Twitter users did not really understand the nature of the pandemic, how the pandemic came about, how the Coronavirus spreads, and what to do and what not to do. We observed a lot of disinformation as compared misinformation between the first two weeks of February. Wrong information ranged from the spread of conspiracies like "the virus is a man-made weapon", "Lysol can cure Coronavirus", "the use of rubbing alcohol is enough to prevent Coronavirus" etc. We also observed wrongful claims when it came to the number of people

that had died from the virus or have been infected by the virus beginning February. For example, we observed claims that more than 100,000 people have died from the virus between February 1, 2020 to February 12, 2020. To solve the issue of disinformation and misinformation, the WHO launched a pilot program (EPI-WIN) in early January that extended to February to make sure that correct information are disseminated on various social media platforms. This action by the WHO was laudable but did little to bring down the issue of disinformation at that time. We observed that false information was retweeted later on in the month of February.

2. Retweeting of Bot Tweets

About 30% of the misinformation or disinformation that we detected through Bot Sentinel and Poynter came from retweets of Bot contents by humans. We observed that there were political agendas behind these fake coronavirus tweets by Bot that were retweeted by humans. For example, we observed that some Twitter users that oppose certain decisions made by China tend to retweet anything that is politically against China to create misunderstandings and make the people believe less in the Chinese authorities. Some of the retweets that we observed were tweets that targeted the American health system and leaders who are trying to manage the spread of the Coronavirus. We think that the purpose of these tweets was to undermine, destroy or disrupt the American health system.

3. Illiteracy

Illiteracy was also one of the main reasons that led to the spread of misinformation and disinformation among humans.in Twitter. For example, the WHO through its EPI – WIN project had to exposed the falseness or hollowness of the belief that sesame oil and breathing in the smoke can get rid of Coronavirus. We also observed that a lot of Americans

did not clearly understand most of the messages that were relayed by Dr. Fauci and the American Health System and this reflected in the tweets that the posted on their Twitter homepages.

4. Conspiracies and political propaganda

Conspiracy theories and false information about Coronavirus became a problem, as the pandemic spread across the globe. We detected a lot of tweets from February to April that were making it cumbersome for online social media users to spot trustworthy sources of information as these tweets were spreading conspiracies and political propaganda. The growing number of people getting infected and the enforcement of social distancing protocols led to widespread online discourse about the pandemic on various social media outlets with an increasing number of conspiracies and misinformation (Sharma et al., 2020). For example, Figure 4.4.5 shows some of the tweets that has been flagged as spreading conspiracies from the tweets that we analyzed.

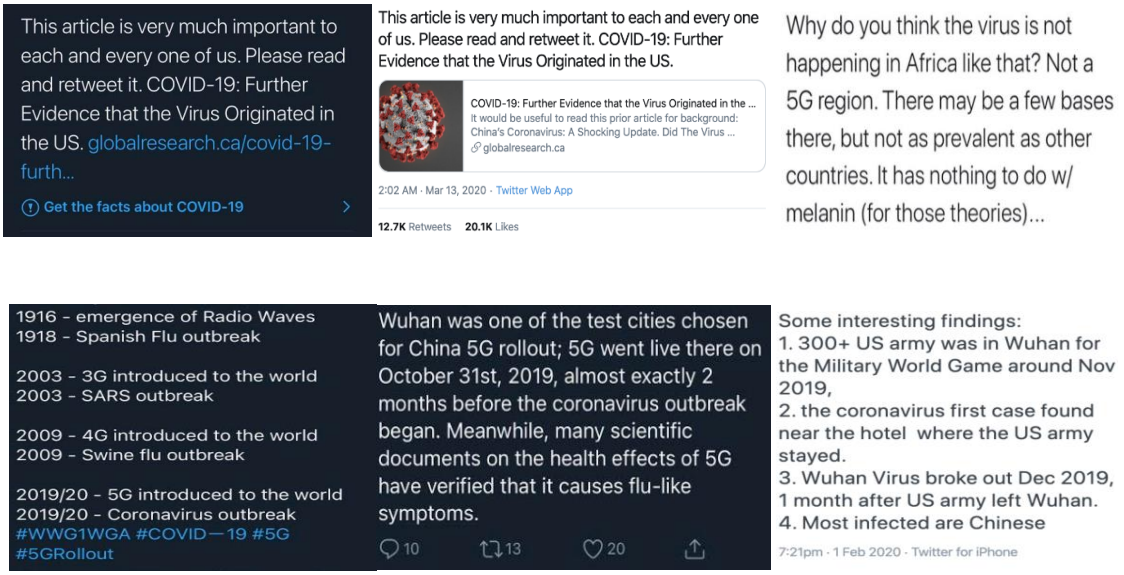


Figure 4.4 4: Shows examples of conspiracy tweets about 5G and Covid-19.

4.4.2.2 Reasons for misinformation disseminated by humans

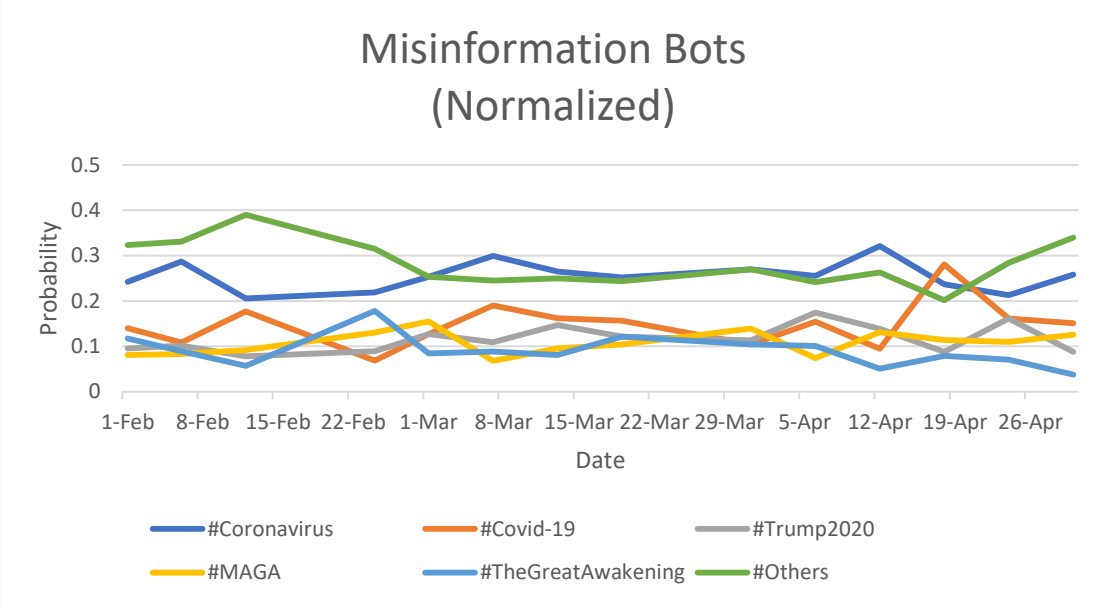


Figure 4.4.5: Shows Bot Misinformation and Disinformation Trend Analysis (N=500)

Figure 4.4.6 shows the most used hashtags and the amount of disinformation that was disseminated by bots from February 1st, 2020 to April 26th, 2020. At the time of this writing (March 2021), most of the bots especially QAnon and Pro Trump bots detected by our classification framework had already been taken down by Twitter so we relied mostly on the dataset content that we hydrated using Twitter's API to check for false information by using Poynter's FactChat. We observed a gradual rise in the level of misinformation disseminated by Bots in Twitter. We categorize the entities responsible for the spread of misinformation bots below:

1. Pro Trump Bots

We analyzed the bots detected by our classification framework and observed that in the month of March, when the pandemic was becoming an issue in the United States and all over the continent there were bots whose main agenda were to disseminate political

conspiracies, and indorse the conspiracy theory that the Coronavirus was a virus or a bioweapon created by China to destroy the United States. While analyzing tweet contents, we found a lot of Pro Trump bot accounts that retweeted Covid-19 related issues in a synchronized manner. We observed that, there were tweets from Pro Trump bots kept posting virus conspiracy theories over a period of time. These tweets were retweeted, liked many times by some Twitter users and had lots of impressions. At the time of this writing, Twitter had already suspended over 5000 Pro Trump bot accounts and others that were associated with it for amplifying certain political messages and spreading false information. (Hunt, 2019). The outcome of this Pro Trump was the augmentation of misinformation or disinformation by hardcore Trump supporters.

2. QAnon Bots

QAnon is a far right- wing, loosely organized association of supporters who accept a range of unproven beliefs. The Storm and the Great Awakening are two major things that QAnon followers are waiting for. The Storm has to do with the mass arrest of individuals in high official positions while the Great Awakening has to do with a single event that would show everyone that the QAnon beliefs were accurate the whole time (What Is QAnon? What We Know About the Conspiracy-Theory Group – WSJ, 2021). During the pandemic, QAnon followers added to their unproven belief that individuals that would take the Covid-19 vaccine increases the likelihood of them being classified as either homosexual or transgender in the future. While most of the QAnon bots that our classification framework detected were created by Researchers, we observed that most of the tweets by QAnon bots were liked and retweeted by bots and supporters of QAnon. We also observed Russian accounts that were backing these QAnon accounts.

3. Republican Bots

Republican bots are bots that were trying to deceive social media users in the United States and control the 2020 United States elections in favor of Donald Trump. While we did not find any connections between republican bots that our model detected and Russian operatives, (Chen et al., 2020) and (E. Ferrara, 2020) reported that these bots were created and operated by Russians. According to these authors, Russian operatives created these bots to make people support and vote for Donald Trump in the United States 2020 elections.

4. Human-Like Bots

In the past, Bots used to have simple tactics that were not difficult to spot but today, artificial intelligence (AI) tools that creates human-like language have made it cumbersome to detect certain malicious social media bots. This is due to the fact that, these human-like bots behave in the same way as humans which makes it difficult to tell what is real and what is not. Researchers have observed that these bots survive longer on social media platforms and can create a network of bots which are synchronized to act in a certain manner (E. Ferrara, 2020). Our detection model failed to detect any human- like bots but we were able to detect botnets that were working together to disseminate false information on Twitter using Bot Sentinel. Using Bot Sentinel, we examined user account features such as follower count, account age, tweet sentiment score, friend count etc. to tell if tweets from the account were coming from a human or a bot.

5. 5G Conspiracies

5G Conspiracy theories picked up steam in 2020 when the Russian government's news outlet issued a warning that 5G can kill (Evanega et al., 2020). The "5G can kill" warning was picked up by a French conspiracy website known as *Les moutons enrages*, which proposed a direct relation between Covid-19 and the installation of 5G towers in Wuhan, China. The unproven idea that there was a correlation between 5G and the novel Coronavirus started to spread on Twitter and broke into mainstream media coverage on April 5 with extensive reporting of destruction of 5G towers in the United Kingdom and other countries(Evanega et al., 2020). 5G conspiracy tweets was one of the common misinformation or disinformation tweets that we observed in our COVID-19 dataset. The fact-checking feedback we got from Poynter and Bot Sentinel shows how misinformed or disinformed individuals on Twitter have been during the early stages of the pandemic.

4.4.2.3 Bot Vs Human Misinformation Analysis

We focused on the most used hashtags for the detected bots and humans to see if bots have a higher likelihood to spread misinformation as compared to humans. Figure 4.9.2 shows that humans have a higher probability (0.24) to spread misinformation as compared to bots (0.20) from our Covid-19 dataset. We mentioned earlier that we observed that about 30% of tweets from detected humans that were spreading misinformation came from retweets of bot content so that explains why we are seeing a higher likelihood to spread misinformation by humans as compared to bots. When we account for humans retweeting bot content, humans actually may not be spreading misinformation intentionally. Figure 4.4.7 shows that bots on Twitter indirectly spread misleading content through humans by leveraging some human's inability to detect false information.

We also observed that after March 7th, 2020, there was a big separation between the #Coronavirus and the # Covid-19 hashtags for humans. The big separation seen in Figure 4.4.8 for humans has to do with Twitter's effort to crackdown Coronavirus related misinformation between March and April. Twitter put in place policies aimed at suspending tweets in all hashtags categories from user accounts that were disseminating misinformation about the Coronavirus between March and April.

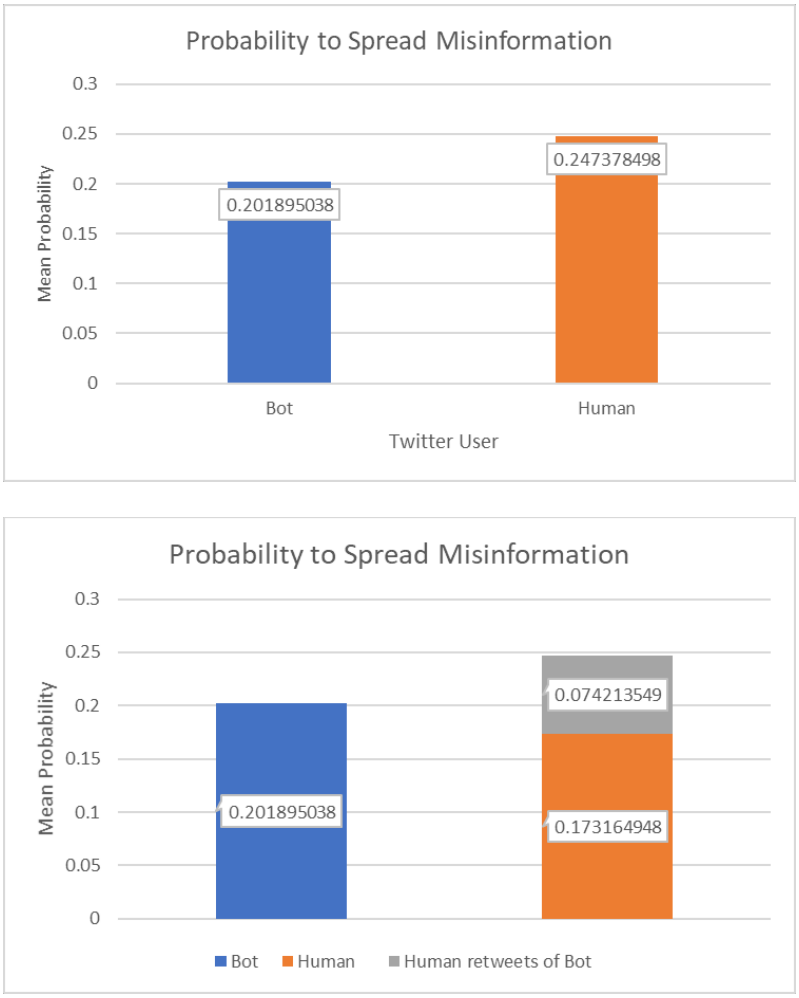


Figure 4.4.6: Shows the mean probability to spread misinformation (Bots vs Humans)

Since the #Coronavirus and #Covid-19 categories were the most used hashtags by bots and humans, they were the most suspended as compared to the other categories.

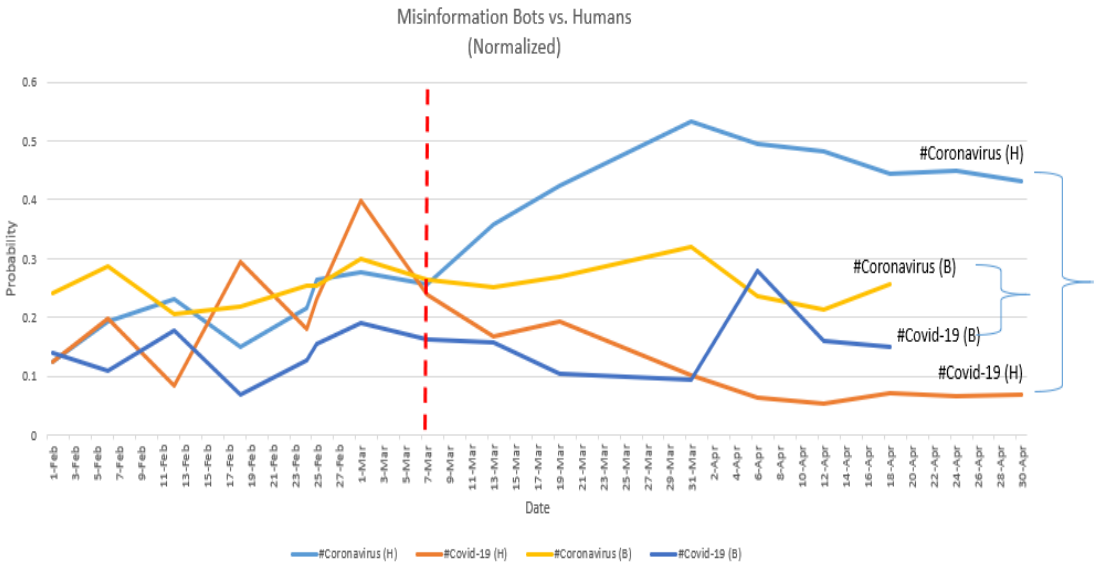


Figure 4.4.7: Shows the probability for misinformation (#Coronavirus, #Covid-19): bots vs humans.

4.3 Sentiment Analysis (Bots vs. Humans)

We analyze 100 tweets each for bots and humans detected in each of the #Coronovirus and #Covid-19 category for every two weeks from February to April from our Covid-19 dataset to see the sort of sentiments that were been expressed by the detected bots and humans As a result, we sampled a total of 3,200 tweets from February to April. Sentiments were extracted from the detected human and bot tweets to study their perception towards the coronavirus outbreak. We use (Hutto & Gilbert, 2014.) lexical sentiment extraction to generate the valence (positive or negative) of a Twitter user's tweet. We also relied on Bot Sentinel as an overall sentiment score generator to give every detected user account from the two most used hashtags by bots and humans (#Coronavirus and #Covid-19) a sentiment score and a sentiment rating.

As discussed earlier, most of the prominent issues that were discussed on Twitter between February and April centered around prevention measures such as the usage of hand sanitizers and Lysol, frequent hand washing and the wearing of mask, travel restrictions, global outbreaks (Italy, China, Germany, Iran etc.), symptoms and infections, global death rates, government response etc. Figure 4.4.9 and Figure 4.4.10 shows the weekly average sentiment score for detected humans and bots from the Covid-19 dataset.

From Figure 4.4.9 and Figure 4.4.10, we can see that the bots that our classification model detected were expressing more negative sentiments as compared to humans. We also observed that after 7th March, 2020 the level of negative sentiments expressed on the pandemic dropped. As discussed earlier, the reason why we are seeing a dip in sentiments expressed on Twitter with regards to the pandemic in the month of March has to do with Twitter's effort to crackdown misinformation when it comes to the pandemic.

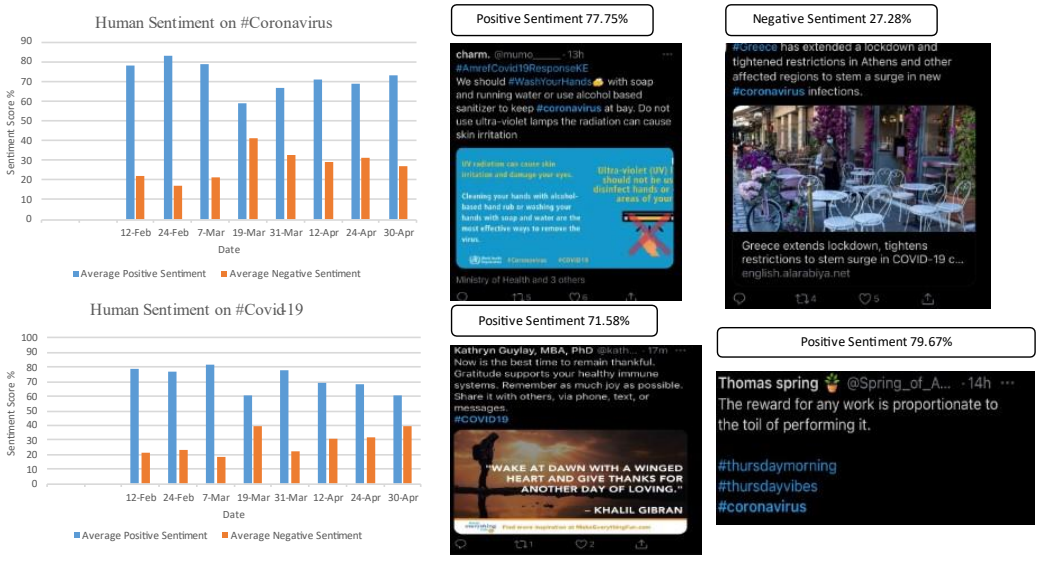


Figure 4.4.8: shows Human sentiment score on #Coronavirus and #Covid-19 (Left) and sample of tweets that show how tweets are rated (Right).

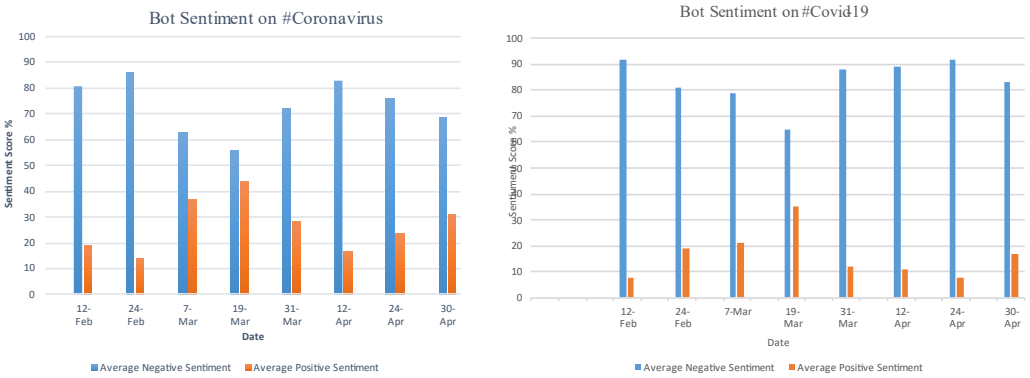


Figure 4.4.9: Shows Bot sentiment score on #Coronavirus #Covid-19

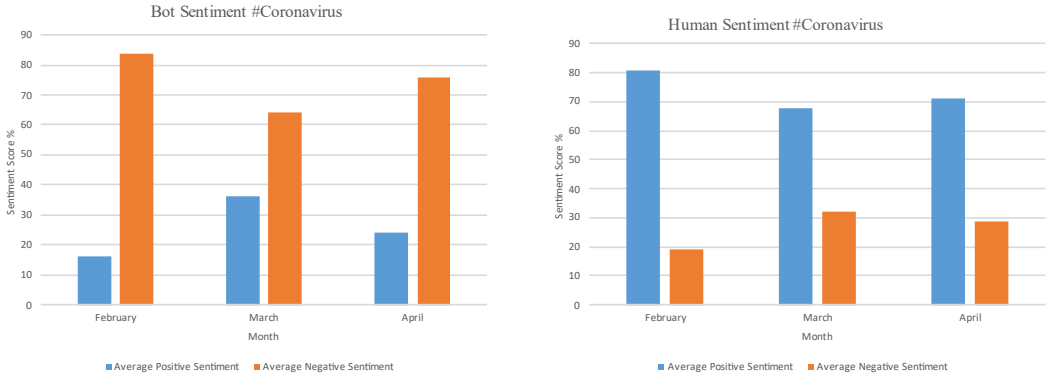


Figure 4.4 10: Shows the average sentiment (Bot vs Human) on #Coronavirus from February to March.

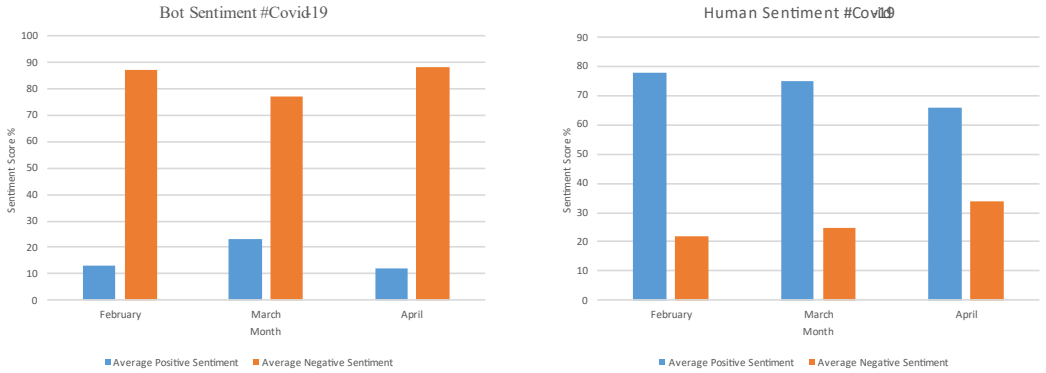


Figure 4.4 11: Shows the average sentiment (Bot vs Human) on #Covid-19 from February to March.

We also analyze the overall sentiment of a user's account by evaluating his or her account on Bot Sentinel for a general score. We use Bot Sentinel to rate the detected user accounts in the #Coronavirus and #Covid-19 category with a score from 0-100. The higher the score, the higher the likelihood that the account engages in the spread of false information and other malicious activities such as harassment, trolling etc. Bot Sentinel analysis several tweets per a Twitter account and the more a Twitter user engages in an act that is consistent with disruptive or problematic accounts, the higher their Bot Sentinel score is. A total of 900 unique user accounts (450 detected bots and 450 detected human accounts) from February to April were evaluated on Bot Sentinel for a general score. Figure 4.4.13 shows how Bot Sentinel rates a user's account based on his or her overall sentiments or tweets posted on Twitter with regards to Covid-19 and any other issue.

The purpose of rating and scoring the 900 unique Twitter accounts is to observe how many detected bots and humans fall into the Normal, Satisfactory Disruptive and Problematic categories on Bot Sentinel. Figure 4.4.14 shows that 189 out of the 450 bot accounts detected by our classification framework were flagged as accounts exhibiting disruptive behaviors. 89 and 158 detected bots exhibited Normal and Satisfactory Tweeting activities. 14 detected bot accounts produced no results which means that those accounts have been suspended temporarily or permanently by Twitter.

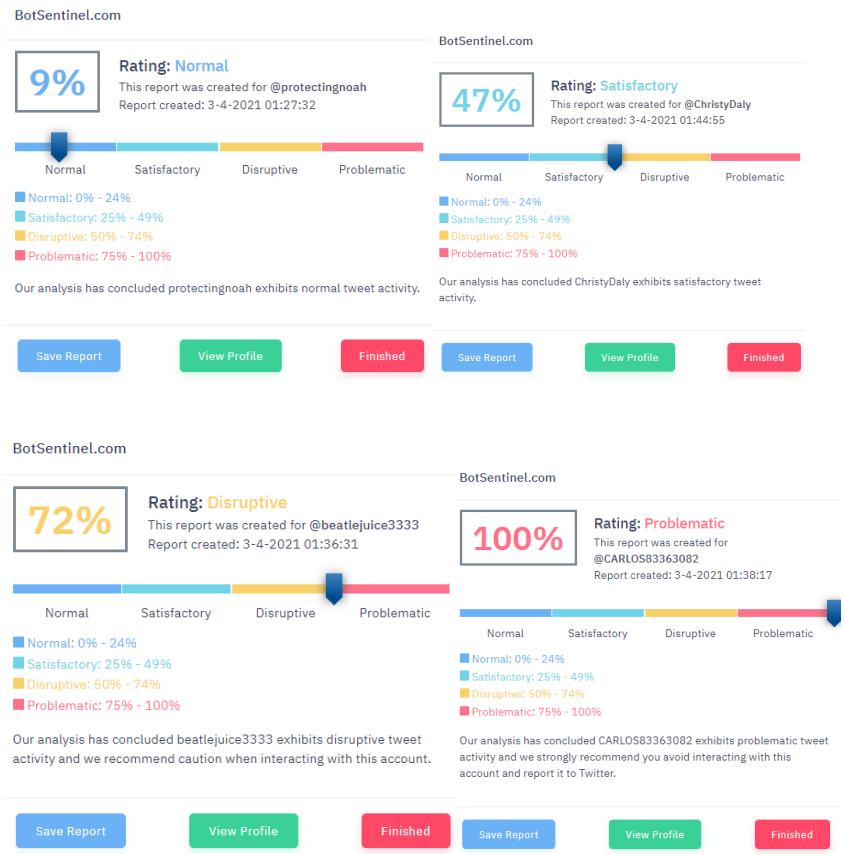


Figure 4.4.12: Shows examples of how Bot Sentinel rates a Twitter user Account.

The results obtained from Bot Sentinel shows how difficult it is to tell if an account belongs to a bot or human.

Today, we have bots exhibiting human behaviors and normal tweeting activities on social media platforms so the results obtained from Bot Sentinel is not surprising. Figure 4.4.14 also shows that out of the 450 selected human accounts that were evaluated on Bot Sentinel, 195 and 175 accounts exhibited Normal and Satisfactory tweeting activities. 63 out of the 450 human accounts were flagged as accounts exhibiting disruptive behaviors on Twitter. 17 detected human accounts were suspended temporarily or permanently which means that there were more suspended human accounts as compared to bots. Bot Sentinel does not show the specific reasons why those human and bot accounts were suspended. We believe that those accounts were suspended due to violations of Twitter policies. None of the accounts we evaluated on Bot Sentinel fell into the Problematic category.

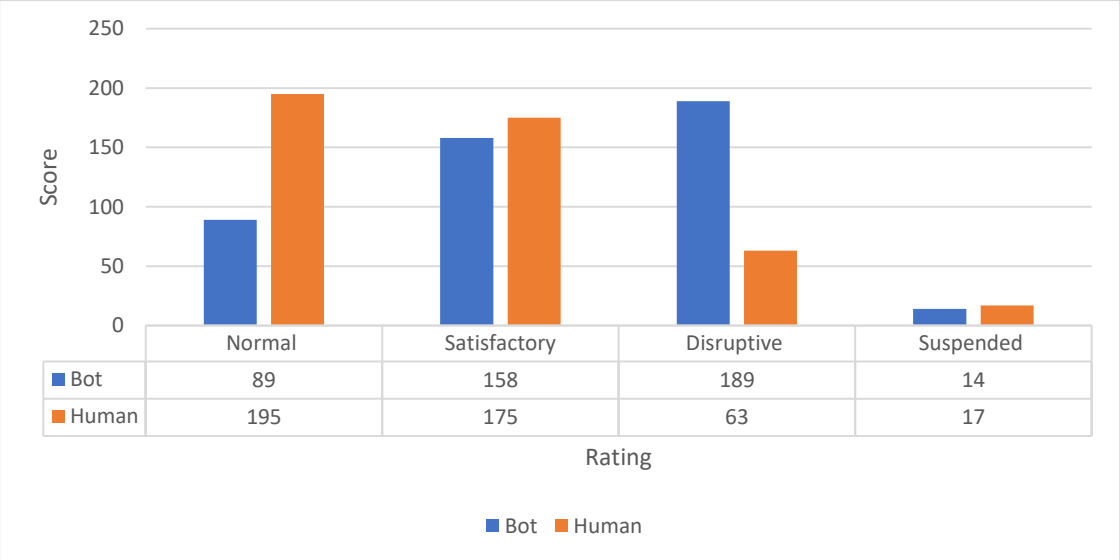


Figure 4.4.13: Shows Bot Sentinel rating and score for 900 unique Twitter accounts.

Chapter 5 Conclusions

SUMMARY OF MAJOR FINDINGS, CONCLUSIONS AND RECCOMENDATION

5.1 Introduction

We present the summary of major outcomes in chapter three and four in this chapter. Our conclusions and recommendations which we draw from the outcomes of the research are examined with respect to the objectives of the study which was to use a hybrid approach that incorporates user account features, topic analysis and sentiment analysis to detect bots on a large-scale Twitter dataset. We report the summary of key findings of this paper in Section 5.2. The concluding remarks and recommendations of the research outcomes have been presented in Section 5.3 and 5.4 respectively.

5.2 Summary of Major Findings

We proposed a hybrid approach that integrates Twitter user account features, topic analysis and sentiment analysis to detect malicious social bots. To achieve the objective of the study, we used the newly developed Twitter COVID-19 endpoint to access COVID-19 and coronavirus-related tweets across languages that provided a dataset of millions of tweets between February 1st, 2020 to April 30th 2020.

The Twitter's search API was used to hydrate tweets from multiple countries in various languages that contained any word associated with COVID-19 (i.e., ncov19, covid, covid-19, coronavirus, ncov2019) that were used in (Lopez et al., 2020). We sampled a total of 39,084 tweets out of 71,908 tweets across the three-month period that this paper focused on. To differentiate a bot from a human, we adopted some of the features used by

(Morstatter et al., 2016; Dickerson et al., 2014 Ferrara et al., 2016). As a result, we tested eight (10) highly predictive user account features which captures several suspicious behaviors to enable us to detect malicious social media bots. We relied on Bot Repository to create a training and testing dataset of already labelled dataset for our experiment. For our training dataset, we used the Social Honeypot Dataset as our first training dataset and the RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter as our second training dataset.

Using Weka machine learning tool, we followed the same classification framework used by the authors in (Lee, Eoff, and Caverlee, 2011) to see what the dataset's prediction accuracy is. In the first experiment, we tested 20 classification algorithms, such as, random forest, naive Bayes, logistic regression and tree-based algorithm, all with default values for all parameters using 10-fold cross validation. We found the results from Weka consistent, with a prediction accuracy ranging from 99% to 91% across most classifiers (15 out 20 tested) for our first training dataset. For the other 5 out of 20 tested, accuracy ranged between 90% to as low as 89%. We created our second training dataset by using 254 human accounts and 144 bot accounts from the RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter Dataset. We tested for prediction accuracy by using Random classifier and observed a 100% prediction accuracy.

In the second experiment, we used the fame for sale: Efficient Detection of fake Twitter followers on twitter as a testing dataset in this paper. To create our legitimate user dataset from the fame for sale dataset, we sampled 235 out of 574 human accounts from "thefakeproject" (TFP) and 964 out of 1488 from the #elezioni2013(E13) verified human

dataset. We created our bot dataset by selecting all fake followers from the "intertwitter" (INT) dataset.

We observed that the social honeypot dataset correctly classified all accounts that were verified as bots in the Fame for sale: efficient detection of fake Twitter follower's dataset but misclassified 1284 human accounts as bots. The social honeypot dataset could detect only 196 out of a total of 1199 verified human accounts as humans. Our second baseline dataset was also used as a training set for Fame for sale: efficient detection of fake Twitter follower dataset to see if the results are better than what the Social Honeypot dataset produced. With precision and accuracy at 92% and 95% respectively, the results from our second experiment with the same features used in the social honeypot dataset were better than the results produced by the Social Honeypot training set. We observed that our category 1 features (favorite count, listed count, name length, and number of tweets) performed better than those found in category 2 (follower count, following count, length of screen name and description length). In general, there was a 1% to 2% prediction accuracy increase across most Tree-based classifiers that were used. However, we observed that prediction accuracy of the RTbust classification framework dropped as we added social honeypot features (see section 4).

In our third experiment, we created a final testing dataset due to poor results obtained from our first testing dataset experiment. Using random forest classifier, the social honeypot dataset achieved a 58.9% precision and 65% accuracy. We observed few factors that contributed to the poor results from the social honeypot experiment. First, the datasets used had inconsistent classes. The social honeypot dataset had few features as compared to the botwiki and celebrity dataset. Second, the few features that the social honeypot dataset has

can only capture a tiny sector of a user account's characteristics. Lastly, the datasets used were annotated by different people with different standards using variety of methods.

We classified our COVID-19 dataset by using the features that generated the highest precision and accuracy (i.e., 92% and 95% respectively) from the training and testing dataset experiment. Our classification framework shows that out of 39,091 tweets sampled, the model we built using the RTbust dataset classified 5,867 user accounts as bots as compared to 36,949 bots detected by the classification model built using the social honeypot dataset.

To do a topic and trend analysis between bots and humans, we used Bot Sentinel to match some of the hashtags from our Covid-19 tweet dataset like #coronavirus, #Covid-19, #Trump2020, #MAGA, #WWG1WGA, #TheGreatAwakening, and #DarkToLight etc. We observed that humans have a wide variety of topics expressed on Twitter as compared to bots. We also observed that the sort of information disseminated by bots are much more targeted as compared to humans. The most used hashtags for bots and humans from our topic analysis were #Coronavirus and #Covid-19.

User	**# of tweets for all hashtags**	**Fraction of Misinformation (human vs Bot) for all #hashtags**
Human	14,000	4,497
Bot	7,000	1,949
Total	21,000	6,446

Table 5.2 1 Total number of detected human tweets used for sentiment analysis (#Coronavirus and #Covid-19 only).

The dataset in Table 5.2.1 used for misinformation analysis contains twice as many tweets posted by humans compared to bots. We used Bot Sentinel, Poynter and other fact checking tools to check for misinformation among the detected bots and humans. 32% of

human posts were classified as misinformation, while 30% of posts created by bots were classified as misinformation.

Initial analysis suggested bots spread less misinformation compared to humans however, it was observed that that about 30% of tweets from detected humans that were spreading misinformation came from retweets of bot content. This result validates prior research suggesting humans frequently re-tweet bot generated content (Shao, et al., 2018). This may explain why we saw a higher likelihood to spread misinformation by humans as compared to bots. When we account for humans retweeting bot content, humans actually may not be spreading misinformation intentionally. We observed that bots on Twitter indirectly spread misleading content through humans by leveraging some human's inability to detect false information. We categorize the entities responsible for the spread of misinformation bots into: Pro Trump bots, QAnon bots, Republican bots, 5G conspiracies and Human-like bots. We also categorize possible reasons for misinformation disseminated by humans into: ignorance, illiteracy, retweeting of bot content and the spread of conspiracies and political propaganda. We focused on the most used hashtags for the detected bots and humans to see if bots have a higher likelihood to spread misinformation as compared to humans.

Moreover, we analyze 100 tweets each for bots and humans detected in each of the #Coronovirus and #Covid-19 category for every two weeks from February to April from our Covid-19 dataset to see the sort of sentiments that were been expressed by the detected bots and humans. As a result, a total of 3,200 tweets were used for sentiment analysis. Details for the number of tweets used for sentiment analysis is provided in tables 5.2.2-5.2.4

#Hashtag	Human	Bot	#Hashtag Total
#Coronavirus	800 tweets	800 tweets	1,600
#Covid-19	800 tweets	800 tweets	1,600
Total	1,600 tweets	1,600 tweets	3,200

Table 5.2 2 Total number of tweets used for sentiment analysis (#Coronavirus and #Covid-19 only**)**

#Hashtag	#Coronavirus	#Covid-19	#Hashtag Total
February	200	200	400
March	300	300	600
April	300	300	600
Monthly Total	800	800	1,600

Table 5.2 3 Total number of detected human tweets used for sentiment analysis (#Coronavirus and #Covid-19 only).

#Hashtag	#Coronavirus	#Covid-19	#Hashtag Total
February	200	200	400
March	300	300	600
April	300	300	600
Monthly Total	800	800	1,600

Table 5.2 4 Total number of detected bots tweets used for sentiment analysis (#Coronavirus and #Covid-19 only).

It was observed that the bots that our classification model detected were expressing more negative sentiments as compared to humans. We also we analyzed the overall sentiment score of a Twitter user's account by evaluating his or her account on Bot Sentinel. We evaluated 900 unique Twitter accounts (450 each for bots and humans) in our Covid-19 dataset and observed that 189 out of the 450 bot accounts detected by our classification framework were flagged as accounts exhibiting disruptive behaviors. 89 bot accounts were flagged as accounts exhibiting Normal behaviors. 158 bot accounts exhibited Satisfactory

Tweeting activities on Bot Sentinel. 14 detected bot accounts produced no results which means that those accounts have been suspended temporarily or permanently by Twitter. For humans, out of the 450 selected human accounts that were evaluated on Bot Sentinel, 195 detected human accounts exhibited Normal tweeting activities, 175 detected human account were flagged as accounts exhibiting satisfactory tweeting activities. 63 out of the 450 human accounts were flagged as accounts exhibiting disruptive behaviors on Twitter. 17 detected human accounts were suspended temporarily or permanently which means that there were more suspended human accounts as compared to bots. The sentiment results obtained from Bot Sentinel are provided in Table 5.2.5 and Table 5.2.6.

Hashtag	**Negative Sentiment**	**Positive Sentiment**	**N**
#Coronavirus	27.6%	72.4%	800
#Covid-19	28.9%	72.4%	800

Table 5.2 5 Fraction of negative and positive sentiment generated by humans on #Coronavirus and #Covid-19 from February to April

Hashtag	**Negative Sentiment**	**Positive Sentiment**	**N**
#Coronavirus	83.0%	29.5%	800
#Covid-19	83.6%	21.0%	800

Table 5.2 6 Fraction of negative and positive sentiment generated by bots on #Coronavirus and #Covid-19 from February to April

Comparing the results from both tables 5.2.4 and 5.2.5 it can be observed that bots generated more posts of negative sentiment compared to humans and humans created more posts with positive sentiment compared to bots. This result aligns with previous research that suggest bot strategies are often focused on increasing human exposure to negative and inflammatory narratives to exacerbating social conflict online (Stella, Ferrara, & De Domenico, 2018).

Today, we have bots exhibiting human behaviors and normal tweeting activities on social media platforms and identifying features and methods to detect them is becoming increasingly important. Results from this research provide insight into features and algorithms that can help detect bots. Specifically, we found the random forest algorithm provides the highest accuracy with twitter features such as favorite count and listed count compared to results obtained in prior research. In addition, sentiment and topic distributions are other key factors that may help to discriminate between bot and human social media behavior. Bots typically align with fewer topics compared to humans which suggest bots have a narrower and targeted approach. Also, bots tend to create more negative sentiment posts compared to human posts. A summary of the hypotheses and results for this research are summarized in Table 5.2.7.

Hypothesis	Description	Result
H_1	The spread of misinformation or disinformation by bots regarding content related to COVID-19 will be higher than the spread of misinformation or disinformation by humans.	**Supported:** Results from Experiment III (section 4) indicate bots spread more disinformation compared to humans
H_2	The accuracy to detect misinformation by bots will be higher using twitter features such as favorite count, and listed count, as compared to social honeypot features.	**Supported:** Results from Experiment I shows that, favorite count and listed count improves the accuracy to detect mis/disinformation as compared to social honeypot features (section 5.2).
H_3	The distribution of different topics will be greater for humans compared to bots. We expect humans to have a wider variety of topics expressed in Twitter as compared to bots.	**Supported:** Results from section 4 shows that the topic distribution for human tweets during the 3-month period analyzed consists of greater diversity of topics.
H_4	Detected bots will express more negative sentiments on Covid-19 related issues as compared to humans.	**Supported:** Results from section 4 shows that detected bots expressed more negative sentiments as compared to humans on Covid-19 related issues.

Table 5.2 7: Summary of results

5.3 Concluding Remarks

This research explores social media bots, Botnets, detection of malicious bots, the motive and entities behind the spread of misinformation by malicious bots during the Coronavirus (COVID-19) pandemic era between February 1st, 2020 and April 30th, 2020. Using a hybrid approach that incorporates Twitter user account features, topic analysis and sentiment features to detect bots on a large-scale Twitter dataset, we were able to detect malicious social media bots.

Our findings show that there were automated accounts that were used in a malicious manner to spread misinformation and unhealthy propaganda campaigns about the COVID-19 pandemic.

5.4 Recommendations

As of the time of writing this paper (mid-March, 2020), there was not enough studies that researched into social media kinetics in the context of COVID-19. Today, a lot of studies have observed the spread of misinformation and questionable content that relates to COVID-19 pandemic, (Lopez et al., 2020 ; Chen et al., 2020; E. Ferrara, 2020; Evanega et al., 2020 etc). Most of these studies have provided an incomplete outlook of online discussion and problems revolving around COVID-19, (Chen et al., 2020). There is a need for more research, as the landscape of information keeps evolving and more scientific knowledge are unveiled on how the spread of misinformation corrupts the online eco system, and also to help people understand what qualifies as a rumor, or misinformation.

Bibliography

[1] A. M. Dunn, O. S. Hofmann, B. Waters, and E. Witchel, “The spread of fake news by social bots,” *Proceedings of the 20th USENIX Security Symposium*. pp. 395–410, 2011.

[2] M. Feily, A. Shahrestani, and S. Ramadass, “A survey of botnet and botnet detection,” *Proc. - 2009 3rd Int. Conf. Emerg. Secur. Information, Syst. Technol. Secur. 2009*, pp. 268–273, 2009, doi: 10.1109/SECURWARE.2009.48.

[3] Y. Al-Hammadi and U. Aickelin, “Detecting Bots Based on Keylogging Activities,” *SSRN Electron. J.*, 2017, doi: 10.2139/ssrn.2830397.

[4] M. Eslahi, R. Salleh, and N. B. Anuar, “Bots and botnets: An overview of characteristics, detection and challenges,” *Proc. - 2012 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2012*, no. April 2014, pp. 349–354, 2012, doi: 10.1109/ICCSCE.2012.6487169.

[5] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–16, 2018, doi: 10.1038/s41467-018-06930-7.

[6] “Nearly half of Twitter accounts pushing to reopen America may be bots | MIT TechnologyReview.” https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/ (accessed Feb. 03, 2021).

[7] S. Evanega, M. Lynas, J. Adams, and K. Smolenyak, “Coronavirus misinformation: quantifying sources and themes in the COVID-19’infodemic’,” pp. 1–13, 2020, [Online]. Available: https://int.nyt.com/data/documenttools/evanega-et-al-coronavirus-misinformation-submitted-07-23-20-1/080839ac0c22bca8/full.pdf%0Ahttps://allianceforscience.cornell.edu/wp-content/uploads/2020/09/Evanega-et-al-Coronavirus-misinformationFINAL.pdf.

[8] E. Chen, K. Lerman, and E. Ferrara, “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set,” *arXiv*, vol. 6, no. 2, 2020, doi: 10.2196/19273.

[9] P. Wang, R. Angarita, and I. Renna, “Is this the Era of Misinformation yet? Combining Social Bots and Fake News to Deceive the Masses,” *Companion Web Conf. 2018 Web*, pp. 1557–1561, 2018, doi: 10.1145/3184558.3191610.

[10] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nat. Commun.*, vol. 10, no. 1, p. 7, 2019, doi: 10.1038/s41467-018-07761-2.

[11] P. Rosso, *Profiling Bots, Fake News Spreaders and Haters*. 2019, p. 2020.

[12] P. N. Howard, G. Bolsover, and S. Bradshaw, “Junk News and Bots during the U.S. Election: What Were Michigan Voters Sharing Over Twitter?,” no. March, pp. 1–5, 2017.

[13] “COVID-19 Exploited by Malicious Cyber Actors | CISA,” *National Cyber Awareness System*, 2020. https://us-cert.cisa.gov/ncas/alerts/aa20-099a (accessed Jul. 23, 2020).

[14] B. Y. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The Rise of Social Bots,” 2016.

[15] “Twitter Bots Poised to Spread Disinformation Before Election - The New York Times.”

[16] A. A. Daya, M. A. Salahuddin, N. Limam, and R. Boutaba, “BotChase: Graph-Based Bot Detection Using Machine Learning,” *IEEE Trans. Netw. Serv. Manag.*, vol. 17, no. 1, pp. 15–29, 2020, doi: 10.1109/TNSM.2020.2972405.

[17] S. Chowdhury *et al.*, “Botnet detection using graph-based feature clustering,” *J. Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0074-7.

[18] J. Lee and H. Lee, “GMAD: Graph-based malware activity detection by DNS traffic analysis,” *Comput. Commun.*, vol. 49, pp. 33–47, 2014, doi: 10.1016/j.comcom.2014.04.013.

[19] W. Wang, Y. Shang, Y. He, Y. Li, and J. Liu, “BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors,” *Inf. Sci. (Ny).*, vol. 511, pp. 284–296, 2020, doi: https://doi.org/10.1016/j.ins.2019.09.024.

[20] S. Shin, Z. Xu, and G. Gu, “EFFORT: Efficient and effective bot malware detection,” *Proc. - IEEE INFOCOM*, no. i, pp. 2846–2850, 2012, doi: 10.1109/INFCOM.2012.6195713.

[21] S. Kudugunta and E. Ferrara, “Deep neural networks for bot detection,” *Inf. Sci. (Ny).*, vol. 467, pp. 312–322, 2018, doi: 10.1016/j.ins.2018.08.019.

[22] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online Human-Bot Interactions : Detection , Estimation , and Characterization,” no. Icwsm, pp. 280–289, 2017.

[23] S. Wijeratne, A. Sheth, S. Bhatt, and L. Balasuriya, “Chapter 1 Feature Engineering for,” no. October, pp. 3–28, 2017, [Online]. Available: https://dev.twitter.com/streaming/overview.

[24] J. M. Piffaretti, “Senile ectropion,” *Orbit*, vol. 7, no. 4, pp. 261–266, 2010, doi: 10.3109/01676838809052830.

[25] C. E. Lopez, M. Vasu, and C. Gallemore, “Understanding the Perception of Covid-19 Policies By Mining a Multilanguage Twitter Dataset,” *arXiv*, pp. 1–4, 2020.

[26] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, “RTbust: Exploiting temporal patterns for botnet detection on twitter,” *WebSci 2019 - Proc. 11th ACM Conf. Web Sci.*, pp. 183–192, 2019, doi: 10.1145/3292522.3326015.

[27] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “Fame for sale: efficient detection of fake Twitter followers ☆,” pp. 1–34, 2015.

[28] K. Y. Onur, V. Clayton, A. D. Emilio, A. Flammini, and F. Menczer, “Arming the public with artificial intelligence to counter social bots,” no. December 2018, pp. 48–61, 2019, doi: 10.1002/hbe2.115.

[29] K. Yang, O. Varol, P. Hui, and F. Menczer, “Scalable and Generalizable Social Bot Detection through Data Selection.”

[30] “Text Classification with Extremely Small Datasets | by Anirudh Shenoy | Towards

Data Science." https://towardsdatascience.com/text-classification-with-extremely-small-datasets-333d322caee2 (accessed Jan. 18, 2021).

[31] "7 Types of Classification Algorithms - Analytics India Magazine." https://analyticsindiamag.com/7-types-classification-algorithms/ (accessed Jan. 18, 2021).

[32] G. Kirubavathi Venkatesh and R. Anitha Nadarajan, "HTTP botnet detection using adaptive learning rate multilayer feed-forward neural network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7322 LNCS, pp. 38–48, 2012, doi: 10.1007/978-3-642-30955-7_5.

[33] Y. Ji, Y. He, X. Jiang, J. Cao, and Q. Li, "Combating the evasion mechanisms of social bots," *Comput. Secur.*, vol. 58, pp. 230–249, 2016, doi: 10.1016/j.cose.2016.01.007.

[34] C. A. Davis, E. Ferrara, F. Menczer, and A. Flammini, "BotOrNot : A System to Evaluate Social Bots," pp. 273–274, 2016.

[35] Washington Post, "Twitter sees record number of users during pandemic, but advertising sales slow - The Washington Post." https://www.washingtonpost.com/business/economy/twitter-sees-record-number-of-users-during-pandemic-but-advertising-sales-slow/2020/04/30/747ef0fe-8ad8-11ea-9dfd-990f9dcc71fc_story.html (accessed Dec. 21, 2020).

[36] "Bot Sentinel | 2021." https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search/items/bot-sentinel.html (accessed Feb. 06, 2021).

[37] E. Ferrara, “What Types of Covid-19 Conspiracies Are Populated By Twitter Bots?,” *arXiv*, 2020, doi: 10.5210/fm.v25i6.10633.

[38] K. Sharma, S. Seo, C. Meng, S. Rambhatla, Y. Liu, and L. Angeles, “Covid-19 s m : a m t c,” vol. 2019, pp. 1–13, 2020.

[39] C. J. Hutto and E. Gilbert, “VADER : A Rule-based Model for Sentiment Analysis of Social Media Text,” pp. 216–225.

[40] Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435-12440.