

Air Force Institute of Technology

**AFIT Scholar**

---

Theses and Dissertations

Student Graduate Works

---

3-2023

## Classification and Analysis of Twitter Bot and Troll Accounts

Callan P. McCormick

Follow this and additional works at: <https://scholar.afit.edu/etd>



Part of the [Computer Sciences Commons](#), and the [Operational Research Commons](#)

---

### Recommended Citation

McCormick, Callan P., "Classification and Analysis of Twitter Bot and Troll Accounts" (2023). *Theses and Dissertations*. 7457.

<https://scholar.afit.edu/etd/7457>

This Thesis is brought to you for free and open access by the Student Graduate Works at AFIT Scholar. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AFIT Scholar. For more information, please contact [AFIT.ENWL.Repository@us.af.mil](mailto:AFIT.ENWL.Repository@us.af.mil).



**CLASSIFICATION AND ANALYSIS OF  
TWITTER BOT AND TROLL ACCOUNTS**

THESIS

Callan McCormick, Second Lieutenant, USAF  
AFIT-ENS-MS-23-M-143

**DEPARTMENT OF THE AIR FORCE  
AIR UNIVERSITY**

***AIR FORCE INSTITUTE OF TECHNOLOGY***

**Wright-Patterson Air Force Base, Ohio**

DISTRIBUTION STATEMENT A  
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-MS-23-M-143

CLASSIFICATION AND ANALYSIS OF TWITTER BOT AND TROLL  
ACCOUNTS

THESIS

Presented to the Faculty  
Department of Operational Sciences  
Graduate School of Engineering and Management  
Air Force Institute of Technology  
Air University  
Air Education and Training Command  
in Partial Fulfillment of the Requirements for the  
Degree of Master of Science in Operations Research

Callan McCormick, B.A.  
Second Lieutenant, USAF

March 23, 2023

DISTRIBUTION STATEMENT A  
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-MS-23-M-143

CLASSIFICATION AND ANALYSIS OF TWITTER BOT AND TROLL  
ACCOUNTS

THESIS

Callan McCormick, B.A.  
Second Lieutenant, USAF

Committee Membership:

LTC Phillip M. LaCasse, Ph.D  
Chair

Maj Michael J. Garee, Ph.D  
Member

## Abstract

Social media bots imitate a human user, regularly posting content and interacting with other users, to point where to the common eye it can be hard to distinguish the difference between a bot account and a genuine user. This presents a challenge to any enterprise, military or civilian, that seeks to understand the attitudes, opinions, or motivations of a population of interest.

This research trains, tests, and analyzes bot and troll classification models using publicly available, open source datasets. Specifically, it applies decision tree, random forest, feed forward neural networks, and long-short term memory neural networks with hyperparameters tuned via designed experiment to five labeled bot datasets created between 2011 and 2020 and one dataset labeling state-sponsored disinformation accounts or trolls. The first three models utilize account profile features, while the last model applies natural language processing techniques, specifically GloVe embedding, to analyze a user's Tweet history. Results indicate that the random forest model outperforms the other three models with an average F1 score of approximately 0.879 for bot classification and 0.938 for troll classification. Additionally, this analysis explored the robustness of models trained on these open source corpora by training a model on each of the five datasets and testing on each of the four others. Overall, results are diminished, with an average F1 score of 0.601 on bot detection models and 0.462 for troll detection models. Lastly, the model was applied to unlabeled Twitter accounts to attempt to quantify the proportion of bots following prominent Twitter accounts.

# Table of Contents

	Page
Abstract .....	iv
List of Figures .....	vi
List of Tables .....	vii
I. Introduction .....	1
1.1 History of Twitter .....	3
1.2 Problem Statement .....	6
1.3 Research Objectives .....	7
1.4 Document Overview .....	8
II. Background and Literature Review .....	9
2.1 Bot Classification .....	9
2.2 State-Sponsored Disinformation Campaigns .....	13
2.2.1 Troll Classification .....	14
III. Methodology .....	19
3.1 Datasets .....	19
3.1.1 Natural Language Processing .....	21
3.2 Models .....	22
3.2.1 Decision Trees .....	22
3.2.2 Random Forest .....	23
3.2.3 Neural Networks .....	23
3.2.4 Recurrent Neural Networks .....	26
3.2.5 Long-Short Term Memory .....	26
3.3 Evaluation Metrics .....	27
3.4 Model Architecture .....	28
IV. Results and Analysis .....	30
4.1 Model Comparisons and Hyperparameter Tuning .....	30
4.2 Model Examination .....	33
4.3 Data Experimentation .....	34
4.3.1 Accounts associated with State-Sponsored Disinformation .....	34
4.3.2 Elon's Challenge .....	35
V. Conclusions .....	38
5.1 Future Work .....	39

## List of Figures

Figure		Page
1	Decision Trees . . . . .	23
2	Random Forest . . . . .	24
3	Neural Network TLU . . . . .	25
4	Deep Neural Network . . . . .	25
5	Recurrent Neuron . . . . .	26
6	LSTM cell (Géron, 2022) . . . . .	27
7	Dataset trained v. Dataset tested . . . . .	33
8	Analysis of 100 Twitter Followers . . . . .	36
9	Analysis of 50,000 Twitter Followers . . . . .	37
10	Analysis of Twitter Followers with Tweet history . . . . .	37



## List of Tables

Table		Page
1	Cresci-17 .....	20
2	User Features .....	21
3	Decision Tree .....	30
4	Random Forest DOE .....	31
5	Random Forest Results .....	31
6	Neural Network DOE .....	32
7	Neural Network Results .....	32
8	LSTM DOE .....	32
9	LSTM Results.....	33

# CLASSIFICATION AND ANALYSIS OF TWITTER BOT AND TROLL ACCOUNTS

## I. Introduction

Social media is an integral part of life in the 21st century. For many people, it connects them to family and friends and is also their source of news and information. With the inherent importance of social media comes potential dangers. Technology exists that imitates human behavior online in the form of bots. Bots have existed for almost as long as computers. Some common examples are chatbots or robocalls, which are algorithms made to hold a conversation. With the advent of social media has come another form of bots: social bots. These bots imitate a human user, regularly posting content and interacting with other users, to the point where to the ordinary eye, it can be hard to distinguish the difference between a bot account and a human user. One of the most concerning uses of social bots is for the purpose of state-sponsored information operations. The 2021 Interim National Security Strategic Guidance states, “Anti-democratic forces use misinformation, disinformation, and weaponized corruption to exploit perceived weaknesses and sow division within and among free nations” (Biden Jr, 2021).

Three elements are central to state-sponsored disinformation: medium, message, and audience (Nemr and Gangware, 2019). Thus, human behavior is intrinsic to understanding the proliferation of disinformation. Social media exploits humans’ need to belong, allowing people to find like-minded communities to which they may not have access in their local communities. While this can be positive in many ways, such as connecting to distant family and friends, it facilitates common cognitive biases

in interpreting information. Nemr and Gangware (2019) lay out several such biases: selective exposure, confirmation bias, and motivated reasoning, which respectively seeks information that solidifies, analyzes the data as consistent with, and applies higher scrutiny to information inconsistent with one’s beliefs . One recent example of this phenomenon is the debate surrounding the Covid-19 vaccine; after viewing online content on the vaccine’s adverse effects, someone hesitant about its efficacy might view this as confirmation of that belief (Liao, 2021).

The sheer magnitude of information spread via social media amplifies these cognitive biases, considering the fact that six thousand Tweets are sent every second (Nemr and Gangware, 2019). Thus, people are more likely to seek out pages and sources confirming their beliefs, leading to echo chambers. Cinelli et al. (2021) explore the echo chamber’s effect on social media, defining echo chambers as “environments in which the opinion, political leaning, or belief of users about a topic gets reinforced due to repeated interactions with peers or sources having similar tendencies and attitudes.” They quantify the political leaning of users and find that on Facebook and Twitter, a strong correlation exists between the leaning of a user and their nearest neighbors, indicating the presence of echo chambers on these sites.

Cognitive biases and echo chambers rely heavily upon a user’s emotions, making it challenging to correct factual misunderstandings. Simply fact-checking or encouraging critical thinking is not a sufficient solution to combat disinformation for several reasons. Often there is an unwillingness to change one’s mind even after seeing new, differing information, a phenomenon known as belief perseverance (Nemr and Gangware, 2019). Additionally, research shows that disinformation believers already perceive themselves as critical thinkers (Freelon, 2017). Fact-checking requires the repetition of false information, which could further confirm restated disinformation. Thus, Nemr and Gangware (2019) suggest “repeating facts, offering solid evidence,

preemptively warning about and debunking disinformation themes and encouraging openness” as potential methods to challenge disinformation on social media. However, these solutions are only possible by first identifying the existing disinformation themes. While all bots and trolls are not necessarily spreading disinformation, identifying these accounts can limit the scope of the search for these themes.

## **1.1 History of Twitter**

In 2006, Twitter was formed from a failed podcast startup, Odeo. An employee of Odeo, Mr. Jack Dorsey, floated around the idea of ‘status’ where people could text their status out to the public. This idea eventually won a hackathon at Odeo and became the company’s new focus. Over the next four years, Twitter began solidifying itself into the company that it is today with the first use of the hashtag, promoted Tweets for advertising, and actor Ashton Kutcher as the first user to reach 1 million followers (Meyer, 2020). However, at the same time, the first cases emerge of Twitter as a political and international tool.

The first occurrence of Twitter on the international political stage is what is known as ‘Moldova’s Twitter Revolution’ in April 2009. The protests occurred in reaction to a parliamentary election in Moldova where the Communist Party received 50% of the vote compared to the exit polls, which projected 35%. The following week saw protests grow by 10,000 to 30,000 protesters between the first and second days of protests (Mungiu-Pippidi and Munteanu, 2009). This growth is attributed to Twitter and other social media sites, as official media did not cover the protests. Ultimately, Moldova conducted a second election in July, where the opposition parties formed an alliance. Around the same time, in June 2009, similar post-election unrest was occurring in Iran. In this case, it was due to incumbent President Mahmoud Ahmadinejad winning the majority, with each opposition candidate claiming manipulation. Days

into the post-election protests, Twitter was scheduled for maintenance. However, Mr. Jared Cohen, a State Department policy planning staff member, directly requested that Dorsey delay the maintenance so that Twitter would remain up during the protests, to which he eagerly complied (Lichtenstein, 2010). Additionally, this story was released to the public, which initiated one of the first conversations about social media's role in international relations, specifically since the official policy in Iran at the time was nonintervention.

In 2011, a number of revolutionary movements occurred throughout the Middle East, coined the 'Arab Spring.' However, at the time, they were also referred to by the media as the "Twitter or Facebook revolution" (Alhindi et al., 2012). The revolutions were instigated in Tunisia in December of 2010 following Mohamed Bouazizi, a street vendor, who set himself on fire in protest. A month later, on January 25th (National Police Day), revolts broke out in Egypt in response to increasing police brutality.

A standard critique of Twitter and other social media's role in these revolutions is that they would have taken place either way due to the unrest in countries (Alhindi et al., 2012). Undoubtedly, a requirement for revolution is socioeconomic and political discontent. However, social media enables organizers to express and organize discontent to a global audience with a marginal cost of almost zero. In the case of Egypt, in June of the previous year, blogger Khaled Said was beaten by two police officers. His story went viral (a term for rapid, worldwide dissemination) leading to a Facebook group titled "We are all Khaled Said," where people expressed their frustrations with the government and later organized protests. A poll from protesters showed that 50% used Facebook, 16% used Twitter, and 50% heard about protests from an online source (Tufekci and Wilson, 2012). Perhaps the strongest argument for the role of social media in these protests was that Egyptian President Mubarak tried to shut down the Internet in the last week of January. However, Google and

Twitter teamed up with a ‘speak to Tweet’ feature by leaving a voicemail, resulting in around 10,000 Tweets daily during the blackout period (Liedtke, 2011). Eventually, 18 days after the initial protest President Mubarak stepped back. Thus, social media played a critical role in shaping this globally significant event.

Social media has also been a prominent force within social movements in the United States over the past few years. The Black Lives Matter (BLM) movement began in 2013 following the acquittal of Mr. George Zimmerman on charges of murder and manslaughter when he shot and killed African American teenager Trayvon Martin (Edrington and Lee, 2018). The movement grew following the shooting of another African-American teenager, Michael Brown, in 2014 and, most recently, the death of Mr. George Floyd in 2020. The hashtag, #BlackLivesMatter, allowed the widespread dissemination of themes and messages beyond simply a network of friends via the trending page on Twitter. Edrington and Lee (2018) analyzed the function of Tweets relating to BLM and found that the majority of Tweets were information focused, followed by action and community, respectively.

Similarly, Twitter was a significant force in the 2017 ‘Me Too’ movement. Following a New York Times article addressing sexual misconduct by Hollywood film producer Harvey Weinstein, actress Alyssa Milano encouraged others to respond ‘me too’ if they had experienced sexual harassment or assault (Brünker et al., 2020). In the next 24 hours, the phrase was used over 500,000 times on Twitter. While initiated in the US, the movement spread throughout the world. Thus, social media uniquely allows users to share information and experiences, good or bad.

Conversely, social media, especially Twitter, has been seen as a force to polarize people further. This phenomenon is best seen through President Donald Trump, who frequently used the platform to share his unfiltered opinions, so much so that Twitter decided to ban him following the January 6th riots at the U.S. Capitol. Specifically,

two Tweets were cited by Twitter as an encouragement to his followers not to support the election results and incite violence. However, this ban was later lifted by CEO Elon Musk following a Twitter poll soliciting Twitter users' preferences.

In April of 2022, Musk publicly expressed his desire to acquire Twitter. His stated motivation was to increase free speech on the platform. However, in the following months, he reversed his stance because he believed that the number of bots and trolls exceeded Twitter's projected amount of 5% (Wile, 2022). After a brief period of back-and-forth lawsuits, the acquisition officially took place on October 28, 2022. The question of user verification has been a major priority to address. In an effort to increase revenue and reduce harmful accounts, Twitter has altered the verification process of users, which previously was reserved for accounts "notable in government, news, entertainment, or another designed category" (O'Connell, 2022). Instead, Twitter Blue now charges users \$8 to become verified. After some initial challenges with users changing their names to impersonate notable people, causing a brief suspension of the service, Twitter Blue was renewed with pending visual blue check marks (and an increase in price for iPhone users). As of December 2022, the service provides users with the following features: bookmark folders, custom app icons, themes, custom navigation, top articles, reader, and undo Tweets.

## **1.2 Problem Statement**

Social media is inherently both a catalyst and a hindrance to the progress of democracy. Its open nature facilitates a wide range of views and opinions, spreading outlooks different from one's own. In many ways, this can be positive by spreading awareness of movements unrelated to a user, such as those in the Middle East, to Americans. Similarly, BLM spread to non-black Americans via social media. In general, social media has democratized speech.

In addition to the influence of the everyday user, disinformation also impacts the efforts of open-source intelligence (OSINT). The aim of OSINT is to use publicly available sources to extract information. OSINT is only as informative as the media from which it polls. Thus, trolls and bots pose a significant hindrance. Media infested with automated bots and foreign state-sponsored trolls could inevitably lead to inaccurate analysis.

Thus, when Tweets share misinformation and disinformation, consequentially, the everyday user is harmed. Truly positive communication is only possible between two genuine users. As shown above, with the advent of echo chambers, honest discourse is already complicated without the interference of non-authentic accounts.

There are two challenges, the first being the presence of misinformation and disinformation. However, combating this is out of this project's scope as it is a natural consequence of free speech. The second challenge is merely the existence of artificial accounts posing as human users. As described above, these accounts can contribute to a distorted view of human users, which is both relevant to military intelligence and the private sector. Thus, this research aims to distinguish bot accounts from human users.

### **1.3 Research Objectives**

The primary objective of this research is to ascertain the extent to which previous bot classification methods are genuinely successful in their goal. To this aim, a handful of the most commonly used datasets will be applied against some of the most favorable methods. Specifically, this research explores how an approach trained on one dataset performs on a different one. Secondly, research in bot classification heavily outperforms troll classification. Therefore, this research aims also to classify trolls and explore how successful bot classification methods perform with troll accounts.



Lastly, the methods will be applied to unlabeled Twitter to determine the amount of bots.

## **1.4 Document Overview**

This document is organized as follows. Chapter II provides an overview of relevant background information on bot and troll classification. Chapter III details the process of the various models considered: decision tree, random forest, neural network, and long-short term memory. Chapter IV presents the results of various models on a collection of datasets, explores its performance, and tests it on troll and unlabeled data. Finally, Chapter V discusses the conclusions drawn from the results.

## II. Background and Literature Review

This chapter reviews the differences between classifications of bots versus trolls. Specifically, this research focuses on Twitter data. Section 2.1 gives a survey of previous Twitter bot classification models and data used. Section 2.2 further explains the inherent differences between Twitter bots and trolls and how foreign states take advantage of US social media through state-sponsored disinformation campaigns. Finally, section 2.2.1 surveys means to classify trolls, a relatively recent area of interest motivated by allegations of foreign interference in the 2016 Presidential election.

### 2.1 Bot Classification

Lee et al. (2011) utilized a social honeypot for seven months in order to lure, classify, and filter content polluters on Twitter. They cite three essential advantages to the usage of social honeypots: “automatically collecting evidence on content polluters, no interference or intrusion on the activities of legitimate users in the system, and robustness of ongoing polluter identification and filtering” (Lee et al., 2011). They deployed 60 social honeypot accounts and monitored accounts that interacted with those initial 60 users. Once they had accounts to classify the users, they utilized 30 different classification algorithms using various features under four groups: User Demographics, User Friendship Networks, User Content, and User History. While all algorithms performed well, they specified that tree-based classifiers, specifically random forests, performed the best. Their study resulted in identifying 36,000 potential content polluters. Some exciting trends they found in bot accounts were that they only posted four times a day on average and had roughly 2,000 followers and following, whereas a real account usually had 100-1000.

Yang et al. (2020) explore data selection with the goal of labeling bots in real-time; such a method must use minimal user data. For classification, they also found that a random forest classifier with cross-validation tends to result in near-perfect area under the curve (AUC), and thus used this classifier with 100 trees. For their data, they conglomerated all publicly available data sets. The finding is that a subset of the training data had the best results on the evaluation metrics: “cross-validation accuracy on training data, generalization to unseen data, and consistency with more feature-rich classifier on unlabeled data” (Yang et al., 2020). This finding contradicts traditional thought, which suggests that using all training data in the classifier would result in the best model.

Kudugunta and Ferrara (2018) leverage a deep neural network using long short-term memory (LSTM); with synthetic minority oversampling to generate a large labeled data set. In terms of bot detection, they not only wanted to classify accounts but singular Tweets. Similarly to Yang et al., the authors chose to use a relatively small number of features from basic user metadata for the sake of model efficiency and interpretability. For accounts themselves, they do not find deep learning techniques necessary and instead find using an AdaBoost classifier successful, with an AUC of 99.81%. In order for Tweets to be appropriate for LSTMs, they transformed Tweets as embeddings using the pre-trained set GloVE. Their efforts proved very effective, being able to predict a bot or not from a single Tweet with 96%, as well as 99% accuracy given account data. At the time of publication, they were the first to attempt to label a single Tweet as a bot or not.

Wei and Nguyen (2019) proposition to use recurrent neural networks (bidirectional LSTM) with word embeddings. More specifically, they do not use any user metadata or user history to determine whether or not an account is a bot, a strategy which had been previously unexplored. They argue that the two main advantages of their

approach are that handcrafting features and prior knowledge are unnecessary to bot classification.

Miller et al. (2014) do not approach the problem as one of classification but instead as anomaly detection. Anomaly detection is generally most effective when there is mostly one class with outliers. However, the authors argue that stream mining is a natural technique due to speed and the vast amount of Twitter traffic. Stream mining, in regards to data, is defined as the process of learning information from a continuous stream of data from the internet (Nagwani, 2022). The data in this research included 3031 verified users and 208 spam accounts with feature emphasis on content, user information, and Tweet text. Specifically for Tweet text, the model introduced 95 one-gram features and uses two stream clustering algorithms to detect bots: StreamKM++ and DenStream. The former is an instance of k-means clustering, and the second is density-based clustering, which forms spherical and arbitrary clusters. DenStream achieved an accuracy score of 97.2%, while StreamKM++ received a score of 93%.

Cresci et al. (2016), is inspired by DNA techniques to classify bots and online behavior, modeling a user’s behavior with a character encoding that represents a user’s actions. Additionally, the authors emphasized four main steps of using digital DNA for classification: “acquisition of behavioral data, extraction of DNA sequences, comparisons of DNA sequences, and evaluations” (Cresci et al., 2016). They conducted two experiments with different encodings. The first experiment explored types of Tweets: “A for a simple Tweet, T for a reply, and C for a reTweet”(Cresci et al., 2016). The second experiment considered the content of the Tweets: “A for Tweets with URLs, T for Tweets with hashtags, C for Tweets with mentions, G for Tweets with media (pictures, videos), X for Tweets with a combination of previous entities, and N for Tweets with none of them” (Cresci et al., 2016). In order to actually classify

and determine similarities in bots vs. humans, they looked at the longest common substring among the sequences. In comparison to other prominent approaches, this approach performed better on multiple test sets.

Davis et al. (2016) provide potentially the most well-known bot classifier study and a publicly available web app, first released in May 2014. They group the features into six classes: network, user, friends, temporal, content, and sentiment. They initially tested their method on a data set of 15k bots and 16k humans with 5.6 million Tweets. This study also utilizes a random forest classifier, as well as ten-fold-cross-validation, on seven classifiers: one for each class mentioned above, and an overall score, resulting with an AUC of .95. Of all of the approaches mentioned in this review, this is one of the oldest and only one that provides easy accessibility to an average user.

Ali Alhosseini et al. (2019) propose that a social graph is also necessary in addition to a feature set. Thus, they create a graph neural network that incorporates both aspects. It considers the features of a user’s neighbors as their own. However, they consider a limited number of six features: age, favourites\_count, statuses\_count (number of Tweets), account name length, followers\_count, and friends\_count. This is likely partly due to the Twitter API limiting the number of requests. One alternative is to build the graph structure based on reTweets. The approach improved 8% on the area under curve accuracy compared to other leading bot classification research.

Most recently, Feng et al. (2021) explored gaps within current bot classification measures. The authors created a new data set, TwiBot-20, to test the current bot detection method’s performance. Three main issues occur in previous data sets used in bot detection: lack of user diversity, limited user information, and data scarcity. To overcome these drawbacks, their data set includes 229,573 users with 33,488,192 Tweets (the largest to date at the time of publication), including semantic, property, and neighborhood information on the user. Previously, all three had yet to be used

together. Lastly, to ensure user diversity, they employed breath-first search (BFS) via following relationships with seed users varying in topic and geographic location. They then test eight different bot classification methods on the data set in comparison to two other popular datasets, Cresci-17 and PAN-19.

## 2.2 State-Sponsored Disinformation Campaigns

One of the strongest actors in information warfare is Russia’s Internet Research Agency (IRA). It has been active since 2013, acting on behalf of the Russian government and businesses. It not only utilizes bots but also human employees who operate as trolls online. Their work especially came to light following the 2016 Presidential election. The Muller report found that accounts infiltrated all aspects of American society, including anti-immigration, tea party, BLM, LGBTQ, and religious groups (Beskow and Carley, 2020). Their goal is to seed disinformation and perpetuate divisiveness. They operate all over the world and within their own country. Thus, it is critical for US national security and the everyday Twitter user to identify the difference between troll accounts and legitimate users.

Beskow and Carley (2020) extend bot detection research and state-sponsored trolls on Twitter, characterizing and comparing Russian actions to Chinese actions in their respective information warfare campaign. However, this literature review focuses on Russian findings. The first characterization is to create a network based on user interaction on Twitter and highlight the most recent language used. The four most prominent languages were English, Russian, Arabic, and Spanish. Next, analysis of the account’s timeline reveals that Russian trolls are planned and purposeful, embedding themselves in whichever society/subculture before beginning a manipulation. Based on hashtag market share, it appears that the IRA infiltrates the American right moreso than the left, emphasizing the #MAGA. An additional concerning finding is

that multiple accounts appear as a news agency. Furthermore, in regard to news, the accounts amplify Russian-backed news agencies such as Russia Today (RT) and Sputnik. The research then explores how many trolls appear to be bots and finds that 9-15% exhibit automated behavior. Lastly, Beskow and Carley (2020) attempts to answer the question of how many similar actors are still active on Twitter; of his periphery data (accounts mentioned, replied to, or reTweeted by core accounts removed from Twitter), 85% of the accounts were active. Of these active accounts, roughly 10% exhibited potential influence in both American right and left politics. Their study is useful for this research because it demonstrates that simply using bot detection methods will likely only detect automated accounts, not trolls. However, the intense characterization of Russian troll activity can be utilized to generate potential features for classification purposes.

### **2.2.1 Troll Classification**

Following the 2016 US Presidential election, Twitter committed to more transparency regarding state-sponsored campaigns. They first released a dataset in 2018 of 3,841 IRA-related accounts. Thus, most research in troll classification was not possible until Twitter, and other researchers, identified trolls concerning Russian interference. Since 2018, Twitter has been regularly publishing datasets of accounts linked to state-sponsored campaigns from 17 different countries. Nonetheless, most research still focuses on Russia and the IRA's involvement in US politics.

One of the main challenges identified from previous work on bot classification is whether or not it is necessary to include Tweet data and, if so, how. This same difficulty applies to identifying trolls, especially since they are more complex than bots. Ghanem et al. (2019) attempt to tackle the identification of Russian trolls from a textual perspective. They utilize an IRA dataset released by Twitter, filtered only

to include the English Tweets of each user. Then, they queried the Twitter API from August 2016 until the end of 2016, filtered with politically related hashtags. These queries produce a dataset including 2% trolls and 98% users, which they claimed is indicative of the real world. However, the research does not explain how the status of accounts was confirmed. This challenge is common across troll classification, as the source for labeled trolls comes from Twitter releases.

Nonetheless, Ghanem et al. (2019) identify seven main themes across the dataset, using Latent Dirichlet Allocation (LDA): police shootings, Islam and War, Supporting Trump, Black People, Civil Rights, Attacking Hillary, and Crimes. The belief is that IRA trolls' behavior will 'flip-flop' between themes and thus consider the following thematic-based features independently of each other and within each theme mentioned above: emotions, sentiment, bad sexual cues, stance cues, bias cues, and morality. The researchers also attempt to profile IRA accounts utilizing native language identification (NLI) and stylistic tendencies. Finally, the research compares these features using logistic regression against baseline methods of random selection, majority class, Tweet2vec (character-based vector-space representations), and account features. All of the baseline methods perform poorly with .5 or lower accuracy. When considering the theme-based features, emotions led to the highest increase in accuracy. However, the NLI feature resulted in the best result of 91%, and with all features considered together, an accuracy of 94%. The authors consider the Tweets themselves to classify Russian trolls. However, the scope is limited to the 2016 US presidential election, thus prompting the question of whether it would be as effective on more generalized troll detection.

Another shortfall of troll classification research is the lack of diversification and fixation on Russian interference via the IRA. Alhazbi (2020) recognizes this shortcoming and instead focuses on classifying Saudi state-sponsored Twitter troll accounts via



behavior-based features (2020), theorizing that, since trolls are employees of the state and extrinsically motivated, their behavior will differ from that of a genuine user. At the time of this publication, most research into state-sponsored trolls focused on analyzing their behaviors and influence, not detection. It was found that trolls are likelier to Tweet during working hours and interact with others via reTweets and mentions. The author uses this finding to guide his feature selection: average number of Tweets per day, standard deviation, number of URLs, number of reTweets, number of replies, percentage of weekend Tweets, and time of Tweets. Using a set of the 500 most recent Tweets from Saudi trolls (released by Twitter in 2019) and verified Saudi users, he applied the following models: decision tree, random forest, AdaBoost, and gradient boosting machine. While all models performed well, Gradient boost outperformed the others with an accuracy of 94.4%. Additionally, the number of URLs was found to be the most crucial feature in each model, with the exception of Adaboost, where the time of Tweets was most important. Overall, Alhazbi (2020) has shown the importance of behavioral features when classifying state-sponsored trolls. Luceri et al. (2020) consider the behavior of troll accounts via inverse reinforcement learning (IRL). Again the focus is on the 2016 election using Tweets released during the Congressional investigation. Regular users who Tweeted about the election were also collected along with their non-political Tweets to reflect the true nature of their accounts. For both trolls and non-trolls, only accounts that shared at least ten posts and interacted with ten other posts were considered. The IRL relies on a Markov Decision Process (MDP) framework. In this case, the MDP comprises five states: Tweet, reTweet, reply, resharing of agent Tweets, and replying to agent Tweets. IRL uses max entropy and max entropy deep to return the reward vector, which explains the agent’s behavior between states. These reward vectors are then used as the features for machine learning models. The authors consider eight methods: k-neighbors,

SVC, Gaussian process, decision tree, MLP, AdaBoost, random forest, and naive Bayes. For max entropy, AdaBoost performed the best with 89.1% accuracy, and for max entropy deep, the gaussian process performed the best with 85.6% accuracy. Both of these results are at least 5% lower than Alhazbi’s behavior-based approach. However, research to represent the behavior of trolls vs. non-trolls is relatively new and it is reasonable to expect continued advancement in this domain.

Most bot or troll classification research focuses on classifying the account; however, Yilmaz et al. attempt to classify the Tweet. Using a dataset of 18,514 Tweets, half trolls and half non-trolls, they apply three-word embedding techniques: GloVe, ELMo, and BERT. Each word embedding is then fed to three deep neural network models: CNN, GRU, and transformer. BERT and ELMo performed better than GloVe, with an average accuracy of .84 across the different methods. In terms of the methods themselves, GRU performed the best. Therefore, Yilmaz et al. demonstrate that it is possible to detect a troll from the contents of a single Tweet.

Following the 2016 election, Clemson researchers classified the troll accounts themselves. Kim et al. attempted to create a model that would not classify between troll and non-troll but rather right troll, left troll, and news feed. The model classifies the Tweet itself using KNN and then uses majority voting to assign a classification to the account, based on 50 Tweets per account. KNN classifies a data point based on its distance from other points. The authors propose that a time-sensitive distance is necessary regarding Twitter, as specific phrases can take on new meanings at different times. The #MeToo is an example of a phrase taking on a significant change after the start of the movement in 2017. Thus, for their distance metric, they employ an exponential penalty for differences in time. Regarding the distance metric, the researcher uses the cosine similarity of a bag of words and minimum edit distance. It also proposes a new semantic edit distance (SED) that considers more similar words

as less of an edit. The similarity is computed via the cosine similarity between co-occurring words. The idea is that similar words are more likely to appear next to the same word. SED performs better than ED without the time penalty; however, when including the time penalty, ED performs better than the baseline SED. These results indicate that while both word similarity and time of Tweets are essential, temporal information is more critical.

Im et al. (2020) conduct research with similar aims as this research, specifically to classify Russian trolls. Their data set focused on Russian trolls using the English language and active during the 2016 election, with 2,286 troll accounts and 171,291 control accounts. Their classifier used five types of features: profile, behavioral, stop word usage, language distribution, and a bag of words. The model compared three different classifier methods logistic regression, decision tree with a max depth of 10, and Adaptive Boosted decision tree. Adaptive boost performed the best when using 10-fold cross-validation. The authors subsequently tested a sample of the data on Botometer to see if the accounts were valid. However, they found that characteristics associated with bots were not the dominant characteristics found in trolls. Similar to Beskow, their research has identified important characteristics of troll accounts on Twitter that can be used to classify them.

The literature has shown that Twitter trolls are an emerging study. However, the bot classification research is nonetheless valuable, as it shows various classifiers applied to Twitter data. Since bot classification research is more developed than that for trolls, more deep learning has been done, whereas troll classification utilizes more basic machine learning techniques.

### III. Methodology

This study analyzes five Twitter Bot datasets shared by *Bot Repository* and applies four different machine learning and deep learning models to each dataset. This section details the differences between the datasets and methods applied to them.

#### 3.1 Datasets

The first dataset is from Lee et al. (2011). To identify bot accounts, the authors deployed 60 social honeypot accounts. The idea of these accounts was to attract other polluter accounts. The initial 60 could post four kinds of Tweets: a standard Tweet either with or without a link, a Tweet on the top 10 trending topics, or a reply to another social honeypot. These social honeypot accounts ran from December 30, 2009, to August 2, 2010, resulting in 23,869 tempted users. Lee et al. (2011) found that Twitter later determined 23% of these accounts to be spam accounts and subsequently suspended said accounts, thus, displaying the authors' effective approach to identifying spam accounts. Ultimately, the final data set comprises 22,223 of the initially identified accounts, eliminating short-lived accounts. For legitimate accounts, they sampled 19,276 accounts and monitored them for three months to ensure the accounts were active and not suspended by Twitter. The dataset includes the metadata of users, their most recent 200 Tweets, and following/follower information.

Yang et al. (2020) research created a new dataset focused on political discussion during the 2018 US midterm elections. They manually identified genuine users based on their active participation in the discussion. Bot accounts were determined via correlations between the suspicious creation of an account and Tweet timestamps. The resulting dataset contains 42,446 bot accounts and 8,092 human accounts and again only included user metadata (Yang et al., 2020).

The third dataset analyzed is from Cresci et al. (2017) research on the paradigm-shift of social spambots. This dataset is an aggregation of nine datasets from 2009-2014. There is one dataset of genuine users, three of social spambots, four of traditional spambots, and one of fake followers, which are described in further detail in Table 1.

Table 1: Composition of Cresci-17 Dataset Cresci et al. (2017)

Group Name	Description	Accounts	Tweets	Year
Genuine Accounts	verified accounts that are human-operated	3,474	8,377,522	2011
Social spambots 1	reTweeters of an Italian political candidate	991	1,610,176	2012
Socail spambots 2	spammers of paid apps for mobile devices	3,457	428,542	2014
Socail spambots 3	spammers of products on sale at Amazon.com	464	1,418,626	2011
Traditional spambots 1	training set of spammers used by C. Yang, R. Harkreader, and G. Gu	1,000	145,094	2009
Traditional spambots 2	spammers of scam URLs	100	0	2014
Traditional spambots 3	automated accounts spamming job offers	433	0	2013
Traditional spambots 4	another group of automated accounts spamming job offers	1,128	0	2009
Fake Followers	simple accounts that inflate the number of followers of another account	3,351	196,027	2012

The fourth dataset is also provided by research from Cresci et al. in 2019 (Mazza et al., 2019). They collected all Italian reTweets in a two-week period in June 2018. The authors identified accounts that reTweeted the most as automated accounts through data filtration and manual annotation. However, in the final dataset, they omitted automated accounts that did not try to mimic a human either in metadata or Tweet activity. The resulting dataset includes 63,762 accounts, 51% of which are bots and 49% legitimate accounts (Mazza et al., 2019).

The fifth dataset, TwiBot-20, is from Feng et al. (2021) research in diversifying current bot datasets. Previous datasets have tended to focus on a specific topic. Instead, Feng et al. apply a breadth-first search from different seed users in the topics of politics, business, entertainment, and sports. The accounts were manually annotated through crowdsourcing, based on the following behaviors of bots: lack of originality of Tweets, automated activity, repeated Tweets, and irrelevant or phishing URLs. Four out of 5 annotators needed to agree on the classification of the account; otherwise, it went through further analysis by the research team. Thus, this dataset

is more robust than previous ones, with 5,237 human users and 6,589 bot accounts across multiple domains (Feng et al., 2021).

All of the datasets include user metadata information; however, not all include the same features. Ultimately, the following features were retained for each dataset: screen\_name, followers\_count, friends\_count, statuses\_count, and profile description. However, it was necessary to manipulate select features into a numerical format; for the screen name and description, the length was kept as a final feature. Then, additionally, the ratio between followers and friends was added as an additional feature. Finally, it was ensured there was no 'nan' or 'infinity' values by dropping rows that contained either. The final features used for modeling are described in Table 2.

### 3.1.1 Natural Language Processing

Models that consider the users' Tweets require natural language processing (NLP). Natural language processing is the branch of computer science and artificial intelligence that develops and implements algorithms to model and interpret human language. NLP tasks include but are not limited to translation, sentiment analysis, named entity recognition, part of speech recognition, and spell checking. In this case, NLP methods, specifically word embeddings, are applied to the Tweets in order to classify them as genuine or bot accounts. Word embeddings map real words to numeric-vector representation so that the computer can understand words using a

Table 2: User Features

<b>Feaure</b>	<b>Description</b>
Screen_length	length of user's screen name
followers_count	amount of followers of a user
friends_count	amount of accounts a user follows
ratio	ratio between follower_count and friend_count
statuses_count	number of Tweets of a user
description_length	length of a user's profile description

pre-trained word vector (Almeida and Xexéo, 2019). However, before embedding can take place, the text must be preprocessed. For each of the datasets with Tweets, all of a user’s Tweets are conglomerated. At this point, the Tweets are cleaned by lowercasing all words, removing the reTweet symbol ‘rt,’ substituting links with ‘https’, and removing numbers. The next step, tokenization, breaks text into individual words referred to as tokens. At this point, word embedding can take place; for this analysis, Stanford’s *GloVe: Global Vectors for Word Representation* is used for the pre-trained word vectors (Pennington et al., 2014). GloVe utilizes both local statistics and global statistics to create word vectors. Specifically, their pre-trained Twitter word vector with 2B Tweets, 27B tokens, 1.2M vocab with a dimension of 100 is used for this methodology. This pre-trained word vector is loaded in as a dictionary and the Tweet tokens are compared to the dictionary; if a match is found, the vector is copied into an embedding matrix to be used in the embedding layer of a model.

## 3.2 Models

This research performs bot classification using traditional machine learning and deep learning models, specifically decision trees, random forests, deep neural networks (DNN), and long-short term memory (LSTM) models.

### 3.2.1 Decision Trees

The first classification method is a decision tree, which dates back to the mid-20th century (Quinlan, 1996). Decision trees are made up of either leaf nodes, which label a class, or test nodes, which branch off into more subtrees. As an input, it takes in numerical feature vectors. At each occurrence of a test node, the best feature is selected to branch the current instance based on scoring criteria. In this case, the Gini Index is used, which calculates the class impurity, the probability of classifying

a feature incorrectly (Quinlan, 1996). This process repeats throughout subtrees until a leaf node and label classification is reached.

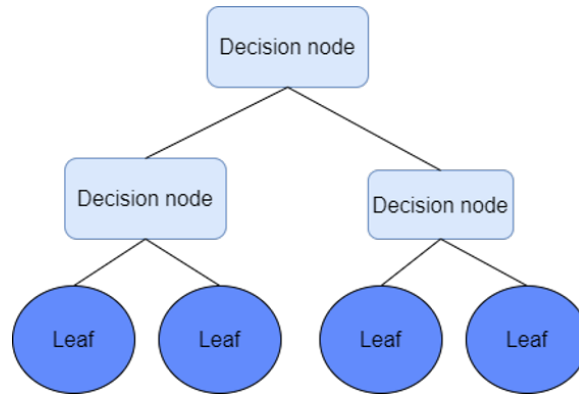


Figure 1: Decision Trees

### 3.2.2 Random Forest

The random forest classifier is an extension of decision trees. As indicated by its name, this model consists of a collection of decision trees, by which the assigned label comes from an aggregation of the labels from the individual trees. The main aim of random forest is to diversify the various decision trees, which is accomplished through random feature selection and bagging, as proposed by Leo Breiman in 2001. As opposed to decision trees, which create each partition based on scoring criteria, this model randomly chooses a feature to partition on. Bagging refers to how the training set is drawn from the original set with replacements for each tree. Breiman found that combining these two techniques increases accuracy. complex deep learning methods.

### 3.2.3 Neural Networks

In 1943, the first artificial neural network architecture was proposed by McCulloch and Pitts. They theorized that an artificial neuron could simulate a biological neuron.



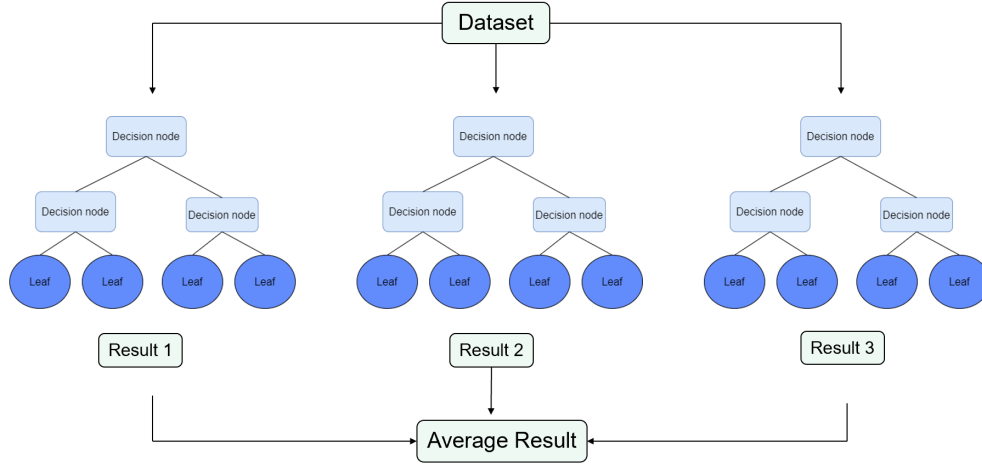


Figure 2: Random Forest

Using one or more binary inputs and one binary output, they demonstrated how a network could perform the following logical computations: identity, AND, OR, and NOT (McCulloch and Pitts, 1943). In 1957, Rosenblatt advanced artificial neurons with the Perceptron. A new kind of neuron, a threshold logic unit (TLU) which, takes in numerical inputs, each with an associated weight. The TLU then calculates output by taking the weighted sum of the inputs and performing the heaviside step function to this sum, as seen Figure 3. The Perceptron itself is made up of a single layer of TLUs connected to each of the inputs, which is an example of a dense layer as all of the neurons are connected to each neuron in the previous layer (Géron, 2022). These early artificial neural networks are the basis for more complex deep learning methods. Deep neural networks (DNN) resemble a neural network, with the addition of multiple TLUs known as hidden layers, as seen in Figure 4. DNNs were only successful in 1986 with the introduction of backward propagation. Forward propagation refers to the basic neural network structure of data flowing from the input layer to the output layer. Backward propagation works in the opposite direction to determine the error of the network. The chain rule is applied at each layer to determine the impact of the current layer on network error (Géron, 2022).

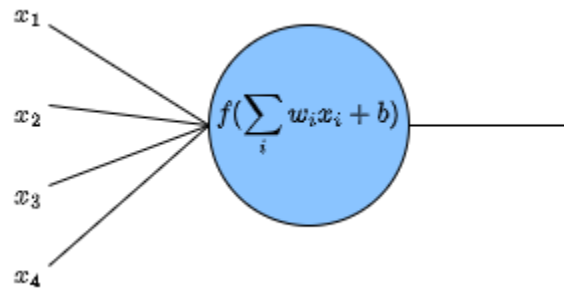


Figure 3: Neural Network TLU

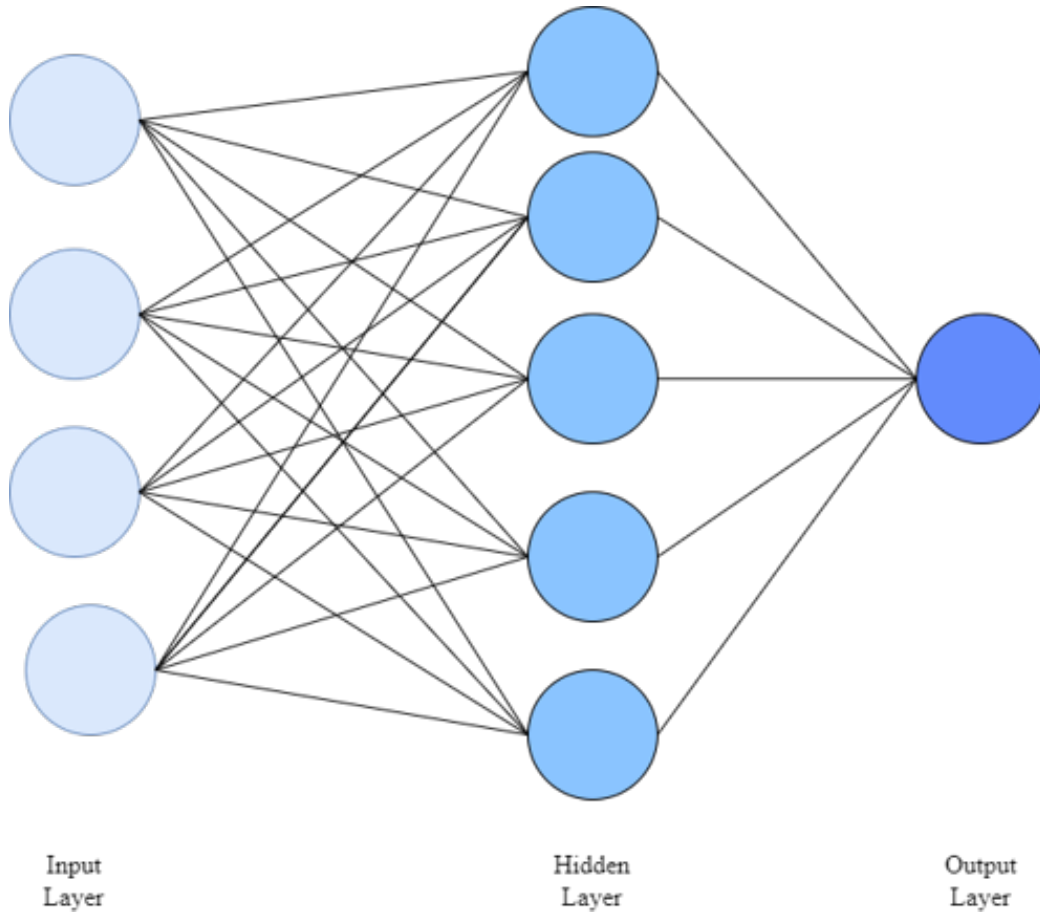


Figure 4: Deep Neural Network

### 3.2.4 Recurrent Neural Networks

The lack of memory is one of the limitations of traditional neural networks. Recurrent neural networks (RNN) aim to rectify the memory constraint by sending the previous output along with the following input. Thus, each recurrent neuron receives an input vector, the previous output vector, and corresponding weight vectors (Géron, 2022). Figure 5 demonstrates the process of a single recurrent neuron over time.

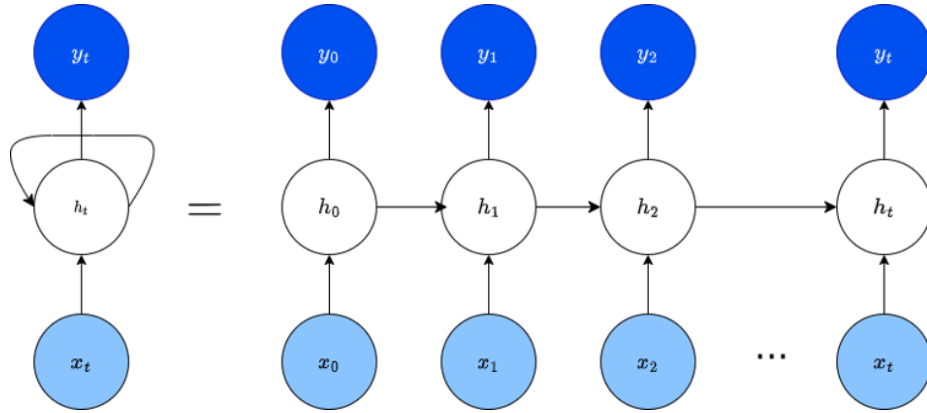


Figure 5: Recurrent Neuron

The nature of RNNs makes them most suitable for sequential data of arbitrary lengths, such as text or Tweets. However, over time, especially when processing long sequences, an RNN forgets its initial input as it is only capable of short-term memory.

### 3.2.5 Long-Short Term Memory

In 1997, Hochreiter and Schmidhuber introduced the concept of Long Short-Term Memory (LSTM) to solve RNNs memory problems (Hochreiter and Schmidhuber, 1997). An LSTM memory cell replaces the basic RNN memory cell, learning what to forget, what to remember, and what to output. Similarly to an RNN, it takes in the input vector  $x_{(t)}$  and previous memory. However, in this case the short-term memory  $h_{(t-1)}$  and long-term memory  $c_{(t-1)}$  are separated. Both the input vector and short-term memory layer are passed to four layers. The main layer acts similarly to

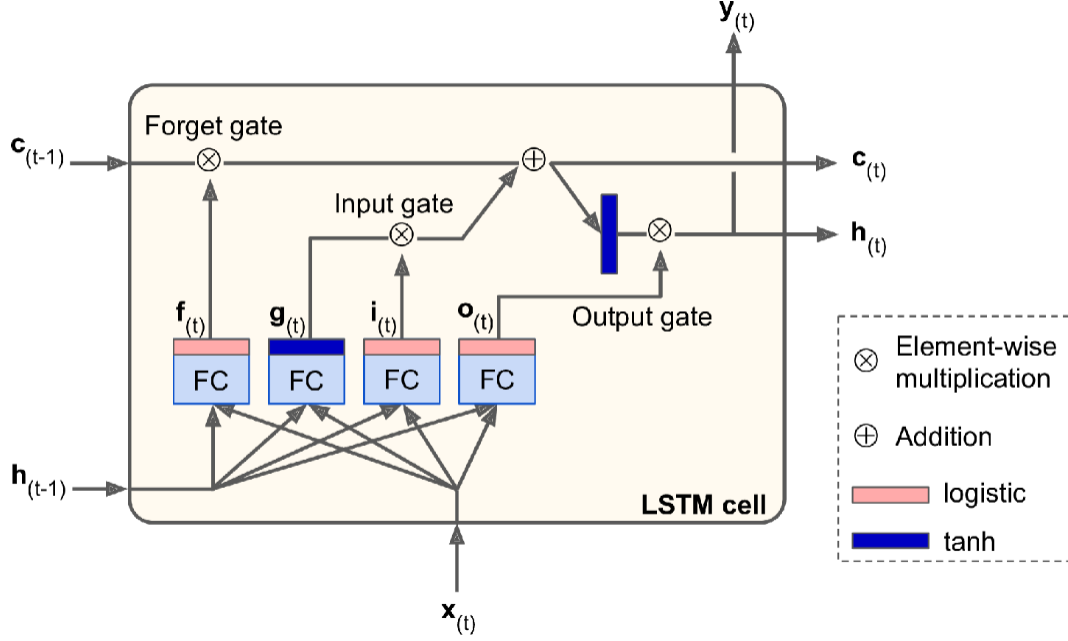


Figure 6: LSTM cell (Géron, 2022)

an RNN cell, analyzing the input vector and previous outputs. However, in this case, only what is deemed important is stored in long and short-term memory (Géron, 2022). The other three layers are gate controllers for the forget, input, and output gates. These layers use a sigmoid activation function; outputting a number between 0 and 1, determining how much information to pass on (Géron, 2022). Specifically, the forget and input gate control what is deleted and added to the long-term memory. In contrast, the output gate determines what should be outputted as short-term memory for the current time step.

### 3.3 Evaluation Metrics

The various models and subsequent experiments consider four metrics: accuracy, precision, recall and F1 score. Accuracy is defined as the percentage of bots and

genuine users correctly identified.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision is defined as the proportion of true positive results over the total predicted positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined as the proportion of true positive over total actual positive.

$$Recall = \frac{TP}{TP + FN}$$

Lastly, the F1 score incorporates both recall and precision.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

While all of these metrics are collected, the F1 score is used as the final evaluator, and when relevant, the averaged F1 score across multiple datasets is considered.

### 3.4 Model Architecture

This analysis considers four models: a decision tree, a random forest, a deep neural network, and an LSTM. All of the models were tested using Python 3.8 in Google Colab, which, when available, utilizes a GPU. The following packages were used throughout the analysis: Pandas (1.3.5), Numpy (1.21.6), Sklearn (1.0.2), Keras(2.11.0), and NLTK (3.7). All of the models use a train-test split of 70% to 30%, respectively. The baseline method, the decision tree, made use of the sklearn decision tree classifier and its default values. The random forest model also utilized sklearn and its default parameters, with the exception of the number of estimators and the maximum number

of features, which were tuned.

The next two models are deep learning models; thus, they mainly utilize keras. The neural network structure is a Dense layer 1(6 Neurons, Relu activation) and Output dense layer (1 neuron, sigmoid activation) using the Adam optimizer with a default learning rate of .001. Batch Normalization is also applied, and the number of epochs considered is 100 with early stopping based on validation accuracy with a patience of 10. For this model, two hyperparameters are considered for tuning the batch size and whether or not to add a regularizer.

The last model considered is an LSTM with the following structure: embedding layer, LSTM layer (32 units, Relu Activation), Dropout(.2), Dense layer 1 (128 neurons, Relu), Dropout(.2), Dense layer 2 (32 neurons, Relu activation), Output dense layer (1 neuron, sigmoid activation). Similarly to the neural network, the Adam optimizer with a learning rate of .001 is utilized. 20 epochs are considered with early stopping based on validation accuracy with patience of 10. This model also considers batch size as a hyperparameter, as well as the number of words to tokenize. For the first three models, all five datasets are trained using user features described in Table 2. The last model utilizes the content of a user’s Tweet history to classify. Thus only the three datasets with Tweet data are trained on this last model.

## IV. Results and Analysis

This section reports the results of the various models and dataset tests in efforts to answer the research objectives. Section 4.1 compares the four model on each of the datasets. Section 4.2 evaluates the best best model from Section 4.1 by training the model on one dataset and testing on another. Lastly, Section 4.3 experiments with troll datasets and unseen Twitter data.

### 4.1 Model Comparisons and Hyperparameter Tuning

The five datasets described in Section 3.1, Caverlee, Midterm, Cresci-17, Cresci-19, and TwiBot-20, were applied to the four models described in Section 3.3. The first method tested and used as a baseline is a Decision Tree, results for which are shown in Table 3. As the results indicate, the Midterm and Cresci-17 datasets perform exceptionally well on this model with high accuracy and F1 scores. However, TwiBot-20 potentially appears to be underfitting with high accuracy and a low F1 score. Overall, the average F1 score across the datasets is **0.760**.

Table 3: Decision Tree Results

<b>Dataset</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Caverlee	0.855	0.862	0.869	0.865
Midterm	0.939	0.963	0.962	0.962
Cresci-17	0.943	0.961	0.963	0.962
Cresci-19	0.676	0.707	0.654	0.680
Twibot-20	0.809	0.320	0.339	0.330
Average	0.844	0.763	0.757	0.760

The following three models each consider relevant hyperparameters via a two-factor factorial design of experiments (DOE). For random forest, the two factors considered are the number of estimators and maximum features, as described in

Table 4. The number of estimators determines the number of trees in the forest, the two factors considered were the sklearn default of 100 and an increased number of 500. Maximum features is described as the number of features to consider when looking for the best split. For this factor, the default of none was considered, as well as the square root of the number of features. The sklearn user guide indicates that in general 1 or none provide good results, but specified square root of the number of features for classification tasks.

Table 4: Random Forest DOE

Feature	High	Low
Number of Estimators	500	100
Maximum Features	sqrt	none

As shown in Table 5, the hyperparameter tuning is essentially negligible for the random forest model. The average of the F1 scores across each of the datasets for the different combinations resulted in an F1 score of **0.87**. Nonetheless, the combination of 100 estimators and square root of number of features resulted in the overall highest F1 score of **.8797**.

Table 5: Random Forest Results

		Maximum Features	
Number of Estimators		Low	High
	Low	0.8716	<b>0.8797</b>
	High	0.8711	0.8753

The next model considered is a Deep Neural Network. The two hyperparameters tuned for this model are regularizer and batch size, as shown in Table 6. A regularizer attempts to prevent overfitting by applying penalties on layer parameters. Batch size describes the number of samples considered before the model is updated. This experiment considers a batch size of 32 and 64.



Table 6: Neural Network DOE

Feature	High	Low
Regularizer	Yes	No
Batch Size	64	32

Table 7 shows the results of the hyperparameter tuning on the neural network model. In comparison to the random forest model, the hyperparameter tuning of the neural network indicates preferred parameters. Specifically, no regularizer and a batch size of 64 leads to the highest average F1 score of **0.8224**. However, this is nonetheless lower than the F1 score of each variation of the random forest model.

Table 7: Neural Network Results

		Batch Size	
Regularizer		Low	High
	Low	0.7744	<b>0.8224</b>
	High	0.7066	0.6900

The final model considered is an LSTM. This model differs from the previous three as it considers the Tweets themselves as opposed to the user metadata features. As explained in Section 3.1, only three datasets included Tweet information. The factors considered are again batch size with the same high and low values; as well as the number of words. The number of words describes the number of words to be tokenized and put into the dictionary created by the glove embedding and Tweets.

Table 8: LSTM DOE

Feature	High	Low
Number of Words	1500	5000
Batch Size	64	32

Similarly to the random forest model, the differences between the hyperparameters considered is not considerably significant, as seen in Table 9. However, the low factor

for the number of words contributed to a .02 rise in the average F1 score. Overall, the combination of 5000 number of words and a batch size of 32 resulted in the highest average F1 score of **.8355**.

Table 9: LSTM Results

		Batch Size	
Number of Words		Low	High
	Low	<b>0.8355</b>	.8348
	High	0.8174	0.8135

After comparing the four models: decision tree, random forest, neural network, and LSTM, and their various hyperparamters; the random forest model with 100 estimators and maximum features as square root proved to be the highest performing in terms of F1 score.

## 4.2 Model Examination

To test the robustness of the labeled datasets, five random forest models were created, one trained on each of the five datasets. Model validation consisted of testing each model on each of the other four labeled corpora. Figure 7 contains F1 scores for each of the test sets and their average. Diagonals in Figure 7 are not calculated because each of the models were trained using the full dataset.

	Caverlee	Midterm	Cresci-17	Cresci-19	Twibot-20	Average
Caverlee	NA	0.510934	0.848488	0.225806	0.665404	0.562658
Midterm	0.660806	NA	0.92431	0.248175	0.420768	0.5635148
Cresci-17	0.831995	0.936401	NA	0.142494	0.514813	0.6064258
Cresci-19	0.325537	0.634965	0.782729	NA	0.356768	0.5249998
Twibot-20	0.712819	0.759474	0.855765	0.668132	NA	0.7490475

Figure 7: Dataset trained v. Dataset tested

On average, Caverlee, Midterm, and Cresci-17 perform extremely poorly when another dataset is tested on its model. However, the Twi-Bot20 model performs relatively well on each dataset with an average of **.7490**.

### 4.3 Data Experimentation

This section will test new data on the random forest models to explore how robust the model is. Specifically, it will test accounts that are associated with state-sponsored disinformation campaigns and the followers of the @Twitter account.

#### 4.3.1 Accounts associated with State-Sponsored Disinformation

From October 2018 to August 2022, Twitter periodically released datasets of accounts, their Tweets, and associated media that were connected to state-sponsored disinformation campaigns. Overall, they released 37 datasets originating from 17 countries, totaling in over 200 million Tweets.

Specifically, this analysis uses a dataset from May 2020 that includes 1,152 accounts engaging in state-backed political propaganda within Russia. According to Twitter, these accounts were suspended “for violations of our platform manipulation policy, specifically cross-posting and amplifying content in an inauthentic, coordinated manner for political ends.” These accounts were combined with the legitimate accounts from the Caverlee dataset to create a fully labeled dataset. This dataset was then used to create a random forest model, achieving an excellent F1 score of **.931**. Next, this dataset was used as the test set for the models trained by bot datasets as shown in Table 10.

In every case, models trained on bot-specific datasets were unable to discern between a state-sponsored troll and a human, as evidenced by their low F1 scores.

Table 10: Bot Classifiers tested on Troll accounts

<b>Dataset</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Midterm	0.526	0.319	0.645	0.427
Cresci-17	0.773	0.992	0.168	0.287
Cresci-19	0.753	0.568	0.396	0.467
Twibot-20	0.579	0.361	0.703	0.477
Average	0.722	0.670	0.407	0.428

#### 4.3.2 Elon’s Challenge

In May of 2022, Elon Musk temporarily suspended his purchase of Twitter due to the presence of bot/spam accounts. At the time, Twitter estimated spam and bot accounts made up less than 5% of users. On May 13, 2022, Musk declared in a series of Tweets his intention to randomly sample 100 followers of @twitter to see how many are bots, and invited others to repeat the experiment. He further specified that he chose 100 as the sample size, as that is the number Twitter uses to calculate its estimate. Thus, this section replicates this analysis.

A sample of 1000 followers of @twitter were pulled using the Twitter API. Accounts without a Tweet history were eliminated from the pool of account options. Then, a random sample of 100 accounts were chosen via the Python random module. These accounts were classified using each of the random forest models trained on the five different datasets. Figure 10 shows the results for the five datasets separately and combined. Combined was determined majority voting by adding the results of each dataset together, if the resulting value for an account was three or higher, it was assigned as a bot; otherwise, it was considered a human account. In other words, at least three of the models had to assign an account to be a bot. The results demonstrate that the random forest model trained on the different datasets largely agrees that roughly 80% of these 100 accounts are bots.

Since a sample size of 100 is relatively small, 50,000 followers of @twitter was also

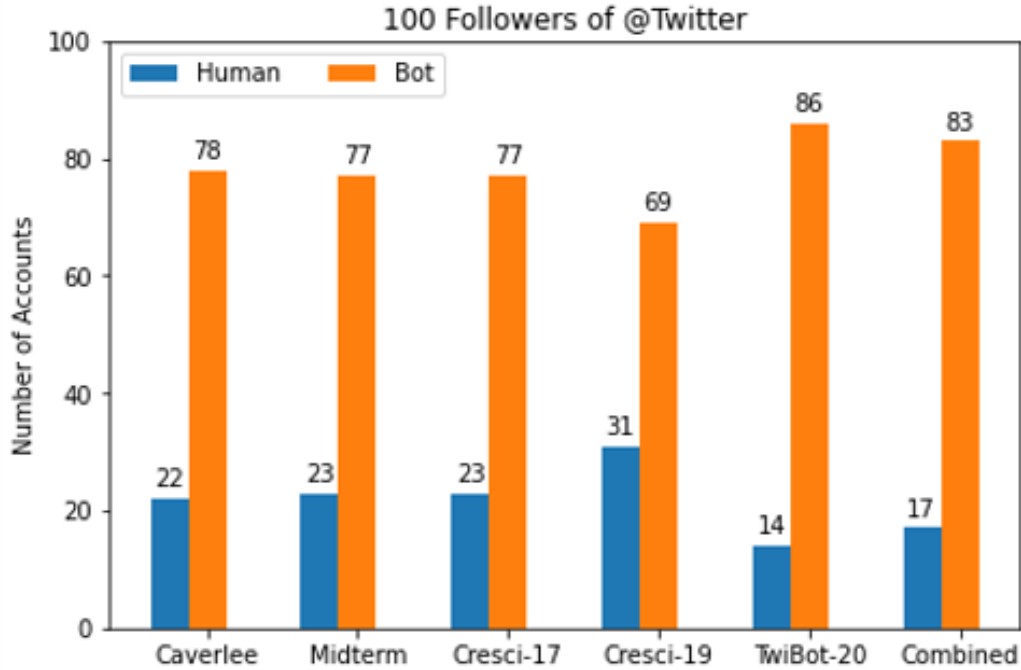


Figure 8: Analysis of 100 Twitter Followers

considered. With the exception of the Cresci-19 model, these results indicate a strong presence of bots following the @twitter account as shown in Figure 9.

Lastly, since the original 100 followers only looked at accounts with tweet history, accounts without a tweet history are removed from the 50,000 resulting in roughly 20,000 accounts. These results are consistent with the previous two tests indicating a strong presence of bots in the followers of the @twitter account.

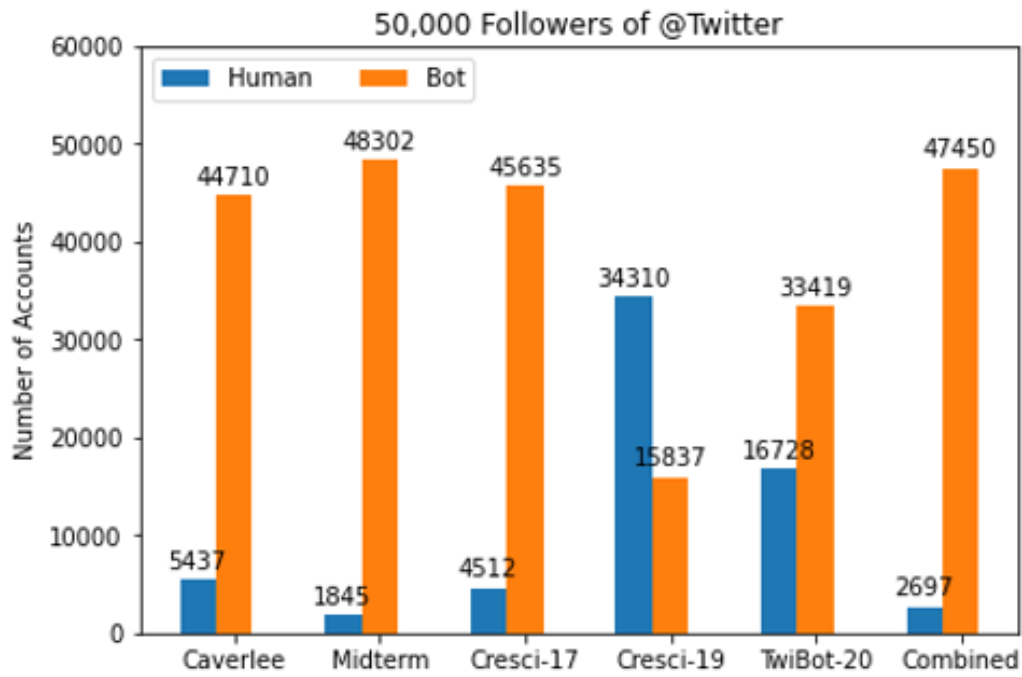


Figure 9: Analysis of 50,000 Twitter Followers

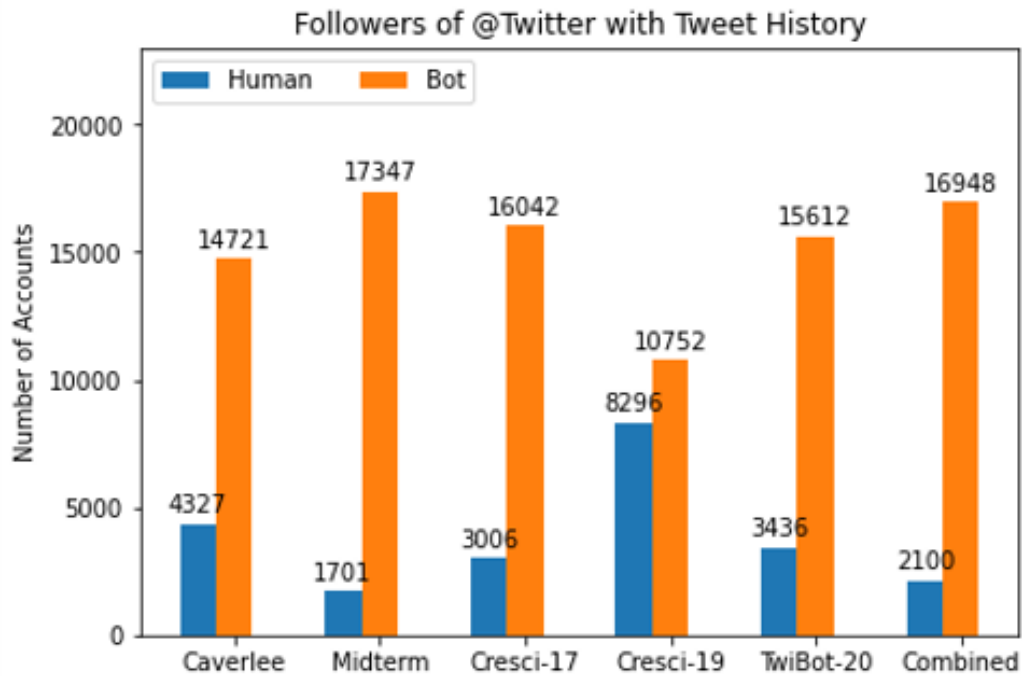


Figure 10: Analysis of Twitter Followers with Tweet history

## V. Conclusions

The results of this study demonstrate that the random forest model using user metadata as features is the best bot classifier considered, performing 4% better than LSTM, 5% better than NN, and 12% better than the decision tree in the F1 metric. However, since user features were not considered for the LSTM model, this in part indicates that user features are more useful than Tweet history in indicating the nature of an account. The same model, when trained and tested on troll accounts, is able to discern between trolls and humans. However, when trained on the bot datasets the model was unable to discern between trolls and humans. These results signify that bots', trolls', and humans' feature details all differ from one another. Similarly, with the exception of the TwiBot-20 dataset, the models trained on one dataset and tested on another are, on average, unable to identify other bots. These results suggest the current corpus of account types in datasets is not diverse enough and that bot accounts may be becoming more complex over time. Furthermore, previous bot classification models trained on these open-source datasets are not robust to validation on datasets created at different times and in different contexts. Lastly, the @twitter followers analysis indicates the presence of many bots on Twitter. Specifically, Twitter's corporate account is inflated by bot accounts. Considering the results of the first four datasets individually does not imply much. However, since the TwiBot-20 model performed well on unseen data these results are more trustworthy, especially as the average scores indicate that the classifiers mostly agreed with one another in their classification. Overall, this research has demonstrated that the presence of bots on Twitter is a real concern, and the need for more diverse datasets to increase the robustness of models.

## 5.1 Future Work

This analysis could be expanded by the types of models considered and a further deep dive into fresh data. Since the LSTM model only considered a user’s Tweet history, it would be interesting to see how the deep learning model performed on user metadata or a combination of metadata and Tweet history. There is also emerging research utilizing graph neural networks, however, this is somewhat limited by the scope of currently labeled datasets, mostly only including user metadata. For the data experimentation, more of the Twitter-released state-sponsored disinformation datasets could be applied. However, these datasets are limited in the sense there are no corresponding human accounts. Lastly, the Twitter follower experiment could easily be repeated among other highly followed users.



## Bibliography

- Alhazbi, S. (2020). Behavior-based machine learning approaches to identify state-sponsored trolls on twitter. *IEEE Access*, 8.
- Alhindi, W., Talha, M., and Sulong, G. (2012). The role of modern technology in arab spring. *Archives des sciences 1661-464X*, 65:1661–464.
- Ali Alhosseini, S., Bin Tareaf, R., Najafi, P., and Meinel, C. (2019). Detect me if you can: Spam bot detection using inductive representation learning. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW '19*, page 148–153, New York, NY, USA. Association for Computing Machinery.
- Almeida, F. and Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.
- Beskow, D. M. and Carley, K. M. (2020). Characterization and comparison of russian and chinese disinformation campaigns. *Disinformation, misinformation, and fake news in social media: emerging research challenges and opportunities*, pages 63–81.
- Biden Jr, J. R. (2021). Interim national security strategic guidance. Technical report, Executive Office of the President Washington DC.
- Brünker, F., Wischnewski, M., Mirbabaie, M., and Meinert, J. (2020). The role of social media during social movements—observations from the# metoo debate on twitter.
- Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

- Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2016). Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64.
- Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972.
- Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274.
- Edrington, C. L. and Lee, N. (2018). Tweeting a social movement: Black lives matter and its use of twitter to share information, build community, and promote action. *The Journal of Public Interest Communications*, 2(2):289–289.
- Feng, S., Wan, H., Wang, N., Li, J., and Luo, M. (2021). Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4485–4494.
- Freelon, D. (2017). Personalized information environments and their potential consequences for disinformation. In *Understanding and Addressing the Disinformation Ecosystem*, pages 38–44.
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. ” O’Reilly Media, Inc.”.
- Ghanem, B., Buscaldi, D., and Rosso, P. (2019). Textrolls: identifying russian trolls on twitter from a textual perspective. *arXiv preprint arXiv:1910.01340*.

- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Im, J., Chandrasekharan, E., Sargent, J., Lighthammer, P., Denby, T., Bhargava, A., Hemphill, L., Jurgens, D., and Gilbert, E. (2020). Still out there: Modeling and identifying russian troll accounts on twitter. In *12th ACM conference on web Science*, pages 1–10.
- Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467:312–322.
- Lee, K., Eoff, B., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 185–192.
- Liao, J. (2021). Undecided on getting a covid-19 vaccine? beware of these two cognitive biases.
- Lichtenstein, J. (2010). Digital diplomacy.
- Liedtke, M. (2011). In one short weekend, speak2tweet was born.
- Luceri, L., Giordano, S., and Ferrara, E. (2020). Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *International Conference on Web and Social Media*.
- Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., and Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

- Meyer, J. (2020). History of twitter: Jack dorsey and the social media giant.
- Miller, Z., Dickinson, B., Deitrick, W., Hu, W., and Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73.
- Mungiu-Pippidi, A. and Munteanu, I. (2009). Moldova’s ”twitter revolution”. *Journal of Democracy*, 20:136 – 142.
- Nagwani, N. K. (2022). *Stream Mining: Introduction, Tools & Techniques and Applications*, chapter 4, pages 99–124. John Wiley Sons, Ltd.
- Nemr, C. and Gangware, W. (2019). *Weapons of mass distraction: Foreign state-sponsored disinformation in the digital age*. Park Advisors.
- O’Connell, C. (2022). ’may not be notable’: Elon musk changes twitter verification again.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72.
- Tufekci, Z. and Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of Communication*, 62:363–379.
- Wei, F. and Nguyen, U. T. (2019). Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE*

*International conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pages 101–109. IEEE.

Wile, R. (2022). A timeline of elon musk’s takeover of twitter.

Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103.

<b>REPORT DOCUMENTATION PAGE</b>					<i>Form Approved</i> <i>OMB No. 0704-0188</i>	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. <b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b>						
<b>1. REPORT DATE</b> (DD-MM-YYYY) 23-03-2023		<b>2. REPORT TYPE</b> Master's Thesis			<b>3. DATES COVERED</b> (From — To) September 2021 — March 2023	
<b>4. TITLE AND SUBTITLE</b>  Classification and Analysis of Twitter Bot and Troll Accounts				<b>5a. CONTRACT NUMBER</b>		
				<b>5b. GRANT NUMBER</b>		
				<b>5c. PROGRAM ELEMENT NUMBER</b>		
<b>6. AUTHOR(S)</b>  Callan McCormick, 2nd Lt, USAF				<b>5d. PROJECT NUMBER</b>		
				<b>5e. TASK NUMBER</b>		
				<b>5f. WORK UNIT NUMBER</b>		
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB OH 45433-7765					<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>  AFIT-ENS-MS-23-M-143	
<b>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>  Intentionally Left Blank					<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>	
					<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>	
<b>12. DISTRIBUTION / AVAILABILITY STATEMENT</b>  DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.						
<b>13. SUPPLEMENTARY NOTES</b>  This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.						
<b>14. ABSTRACT</b>  This research trains, tests, and analyzes bot and troll classification models using publicly available, open source datasets. Specifically, it applies decision tree, random forest, feed forward neural networks, and long-short term memory neural networks with hyperparameters tuned via designed experiment to five labeled bot datasets created between 2011 and 2020 and one dataset labeling state-sponsored disinformation accounts or trolls. The first three models utilize account profile features, while the last model applies natural language processing techniques, specifically GloVe embedding, to analyze a user's Tweet history. Results indicate that the random forest model outperforms the other three models with an average F1 score of approximately 0.879 for bot classification and 0.938 for troll classification.						
<b>15. SUBJECT TERMS</b>  machine learning, artificial neural network (ANN), long short-term memory (LSTM), twitter, bot, troll						
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT</b>	<b>18. NUMBER OF PAGES</b>	<b>19a. NAME OF RESPONSIBLE PERSON</b>	
<b>a. REPORT</b>	<b>b. ABSTRACT</b>	<b>c. THIS PAGE</b>			LTC Phillip LaCasse, AFIT/ENS	
U	U	U	UU	53	<b>19b. TELEPHONE NUMBER</b> (include area code) (262) 470-7549; phillip.lacasse@afit.edu	