

University of Nevada, Reno

**Karma Chameleons: Data Collection Techniques and Account
Characterization for Bot Detection on Reddit**

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Computer Science and Engineering

by

Marissa Floam

Ms. Nancy LaTourrette, Advisor
May 2025



THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

Marissa Floam

entitled

**Karma Chameleons: Data Collection Techniques and
Account Characterization for Bot Detection on Reddit**

be accepted in partial fulfillment of the
requirements for the degree of

Master of Science

Nancy LaTourrette, M.S.
Advisor

Emily Hand, Ph.D.
Committee Member

Gi Woong Yun, Ph.D.
Graduate School Representative

Markus Kemmelmeier, Ph.D., Dean
Graduate School

May, 2025

Abstract

Malicious bots on social media platforms present an ever-evolving threat to the integrity of online communication. As platforms like Twitter/X, Meta, and Reddit continue to grow in popularity, so do opportunities for malicious actors to create bot accounts. These bots, which are often designed to spread deceptive material, spam, or unoriginal content, can significantly influence public opinion and suppress creativity. As a result, it is crucial that general users are able to easily identify these bots so they can be removed, protecting the authenticity of these online spaces.

This thesis addresses one of the main challenges in social media bot detection: the collection of bot or human account datasets with a clearly established ground truth that can be used to train machine learning models. Without these datasets, developing accurate models to distinguish between bot and human accounts becomes difficult. In contrast to Twitter/X, bot detection research targeting Reddit is scarce, even though it is an increasingly popular platform. This thesis outlines characteristics that differentiate bot accounts from human accounts on Reddit and proposes methods for creating reliable datasets for bot detection. Human and bot datasets are combined and tested in six decision tree models to determine the accuracy of each data collection method for use in bot detection. Accurately distinguishing between human and bot accounts is essential for dataset creation, and focusing on specific characteristics allows for clearer determinations between bot and human accounts. Ultimately, in order to maintain the integrity and trustworthiness of one of the most popular social media platforms, Reddit, this thesis addresses a critical gap in bot detection research by providing a reliable framework for data collection.

Dedication

To my mom.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Nancy LaTourrette, for her constant support and invaluable feedback throughout this process. She has been a mentor to me since the very start of my cybersecurity career and I could not have undertaken this journey without her.

To my sisters, thank you for always being there for me. Your love and encouragement have shaped me into the person I am today. I am endlessly lucky to have the three of you in my life.

To my friends, I am so grateful for all the late night gaming sessions that kept me sane and provided a distraction when I needed it most. Thank you for being my community and constantly reminding me that I am not alone.

Lastly, I would like to thank my cat, Mac 'n' Cheese, for her unconditional love and emotional support throughout my entire college career.

Table of Contents

1	Introduction	1
1.1	History of Social Bots	1
1.1.1	Impact on Public Opinion	2
1.1.2	Dead Internet Theory	4
1.2	Data Collection on Social Media	4
1.3	Terminology	6
1.3.1	Bot	6
1.3.2	Reddit	8
1.4	Contribution	10
2	Motivation	11
2.1	Bot Prevalence	11
2.2	Data Gathering and Testing	13
3	Literature Review	15
3.1	Bot Detection Techniques	15
3.1.1	Machine Learning Approach	16
3.1.2	Crowdsourcing Approach	17
3.1.3	Graph-based Approach	18
3.1.4	Anomaly-based Approach	19

3.2	Bot Detection Tools	19
3.2.1	Reddit	20
3.2.2	Twitter/X	21
3.3	Data Gathering	23
3.4	Current Research Gaps	24
4	Bots on Social Media	25
4.1	Twitter/X	25
4.2	Other Social Media Sites	26
4.2.1	Meta	26
4.2.2	Parler	28
4.2.3	Truth Social	28
5	Bots on Reddit	30
5.1	Reddit Bot Landscape	31
5.1.1	Types of Bots	32
5.1.2	Content	36
5.1.3	Account Attributes	37
6	Methods	39
6.1	Data Gathering	39
6.1.1	PRAW	40
6.1.2	Bot Detection within Reddit	40
6.1.3	Bot Datasets	43
6.1.4	Human Datasets	45
6.2	Machine Learning	49
6.3	Limitations	50

7	Analysis	52
7.1	Bot Dataset Analysis	52
7.1.1	Error Code Response	54
7.1.2	Account Profile Page	57
7.1.3	Bot Dataset Statistics	58
7.2	Decision Tree Analyses	59
7.2.1	Verified Humans Model	62
7.2.2	Location Subreddits Model	64
7.2.3	University Subreddits Model	65
7.2.4	Hobby Subreddits Model	67
7.2.5	Combined Humans Model	69
7.2.6	Refined Combined Humans Model	70
7.2.7	Decision Tree Comparison	72
7.3	Limitations	73
8	Conclusion and Future Work	74
	Bibliography	76
A	List of Human Subreddits	86
B	Decision Trees	88
B.1	Verified Humans Model	89
B.2	Location Subreddits Model	90
B.3	University Subreddits Model	91
B.4	Hobby Subreddits Model	92
B.5	Combined Humans Model	93
B.6	Refined Combined Humans Model	94

Chapter 1

Introduction

1.1 History of Social Bots

The emergence of social media as a means of communication has fundamentally changed the way that humans interact with each other. As platforms like Twitter/X, Meta, and Reddit continue to grow in popularity, they have become an avenue for malicious actors to create bot accounts to spread misinformation, harmful material, or unoriginal content. Malicious bots on social media, referred to as social bots, attempt to emulate human behavior such that they can evade detection and spread deceptive or low-quality content [1], [2]. Researchers and general users alike continuously attempt to detect these bots in order to preserve the platforms as a social space for humans to engage in discussion, creativity, and connection.

The detection of malicious bots on social media platforms is essential to maintaining a high-quality and authentic social space that is free from misinformation. However, as techniques evolve to detect them, so do the realism and complexity of the bots, thus making them harder to detect [3]. The cat and mouse game between

bots and bot detection tools ensures that the research surrounding bot detection must constantly advance.

Bots pose a significant threat due to their ability to influence public opinion. Two notable instances where bots were able to alter public opinion with misinformation include the COVID-19 pandemic and presidential elections.

1.1.1 Impact on Public Opinion

Bot accounts have historically been used during times of unrest to impact public opinion. Research on both the COVID-19 pandemic and the 2016 United States presidential election shows the use of bot accounts to influence users on social media platforms. Understanding the impact of bot accounts on public opinion is essential for recognizing the context in which they are used for malicious purposes.

COVID-19 Misinformation

During and following the COVID-19 pandemic, researchers have analyzed the connection between bots on social media and the spread of misinformation related to the pandemic and the vaccine for the SARS-CoV-2 virus. Ferrara et al. analyzed online conversation between the time of the first COVID-19 case reported on United States soil up until the government declared a state of national emergency, to characterize the prevalence of automated accounts that discuss the pandemic. Ultimately, this paper uses statistical characterization to determine that accounts with a "high bot score" engaged with political conspiracies and divisive COVID-19 content [4].

Moreover, Himelein-Wachowiak et al. described known Twitter/X bots and their engagement with COVID-19 and found that up to 66% of the bots analyzed were involved with discussion about COVID-19 [5]. The study also details the "human susceptibility to believing and sharing misinformation" [5], an essential point to make

when considering bot interactions on social media.

Election Interference

Social media is a common platform for the spread of political propaganda; current research indicates that social media manipulates and alters public opinion, sometimes facilitated through the use of bots [6]. Ferrara et al. focused on the 2017 French presidential election by analyzing millions of Twitter/X posts with machine learning, ultimately concluding that bot accounts are reused to spread political disinformation [7].

During the 2016 United States presidential election, Twitter/X identified automated bot accounts located in Russia or run by the Internet Research Agency (IRA). The IRA was a Russian company that often engaged in online propaganda on behalf of Russian interests [8] [9]. Over one thousand accounts were found to be associated with the IRA, while a total of 50,258 automated accounts were identified by the social media company as Russian-linked and "Tweeting election-related content during the election period" [10]. While these accounts were ultimately suspended for Terms of Service violations, it is important to illustrate the sheer impact that social media bots can have on public opinion, especially when considering election interference. The blog post by Twitter/X that discusses these accounts specifies that the information was shared with congressional investigators, demonstrating the seriousness of state-run botnets.

These examples present issues with bots in a real-world context, but similar situations will only continue to increase in frequency as bots become more sophisticated and popular on social media sites.

1.1.2 Dead Internet Theory

The population of bots on the internet continues to increase over time. A theory that has gained traction in online spaces called the "Dead Internet Theory" hypothesizes that a significant portion of internet traffic, posts, and users are populated by bots or AI-generated content [11] [12]. Social media sites may contribute to the basis of this theory for several reasons, such as boosting engagement through interactions between human and bot accounts [13]. The issue with this theory becoming reality results in an overall less authentic internet; this may lead to decreased creativity and diversity on social media as bots aim to blend in by engaging with trends and popular content [13]. As generative AI becomes more sophisticated and easier to implement for a general user, the frequency of bot or non-human accounts will only continue to grow as well. It is not just the frequency of bots online that poses a problem, but their anonymity as well, which prevents human users from knowing whether they are interacting with other humans or automated accounts. The detection of these bots therefore becomes essential to maintain the integrity of original content and use of social media: human connection and interaction. However, successful bot detection techniques remain difficult to achieve as researchers struggle to obtain meaningful and large datasets from social media sites.

1.2 Data Collection on Social Media

As bot detection remains a constantly changing field of research due to the advancement of bots, it becomes progressively more difficult to identify them. This is largely due to the fact that obtaining bot and human account datasets presents a major obstacle for researchers. Many bot detection techniques utilize machine learning, which requires a "ground truth" of accurate data for training purposes. This ground

truth in the context of bot detection requires a dataset of social media accounts that are unequivocally identified as bots or humans [2]. Determining a ground truth dataset typically involves manual annotation or labeling by researchers, which is a time-consuming task that introduces biases. Before labeling can occur, however, the actual obtaining of data—this is typically done through Application Programming Interface (API) calls—varies from platform to platform.

Two major social media platforms, Twitter/X and Reddit, implemented substantial API changes in 2023. Both social media platforms introduced a paid model; Twitter/X removed reading data for free altogether, and Reddit introduced limits on the number of requests made without charging a fee. This led to significant backlash from communities, developers, and researchers, as it resulted in accessibility concerns as well as forced third-party applications to shut down [14]. For example, the most popular third-party Reddit client for iOS devices, Apollo, shut down after the API changes would have resulted in the developer being charged \$20 million per year due to the volume of requests the app received [15], [16]. Additionally, third-party mobile clients for Reddit offered better accessibility for disabled users than the native application. For example, moderators for the r/Blind community who relied on third-party mobile applications announced that the API changes made it impossible for them to moderate [17]. Not only do these API changes affect general users of these platforms, but developers and researchers as well. Even before the changes, researchers expressed their frustration with the use of social media APIs to gather data. Magelinski et al. describe how deployment of certain classification algorithms is inhibited by API bottlenecks [18].

Furthermore, data collection techniques vary on different platforms, especially since the API changes. Many bot detection researchers focus on Twitter/X due to the sheer amount of datasets created and updated by other researchers, but this is

less feasible since 2023 as there is no longer free use of the API. Before the API changes, there were two main ways to gather data from Reddit: Python Reddit API Wrapper (PRAW) and Pushshift Reddit API. However, since the changes, Pushshift Reddit API usage has been limited to moderators only [19], leaving the only avenue for data collection to be with PRAW, which will be addressed in greater detail in Section 6.1.1. PRAW is a useful API tool that has limited free usage, particularly if a researcher needs large amounts of data. Unlike Twitter/X, however, Reddit has very few datasets accessible to researchers that are recent and large enough to analyze. While PRAW is available as a tool to pull data from Reddit, there still lies the issue of establishing a ground truth for bot detection. With the exception of a single list of Russian bots released by Reddit in 2017, publicly available datasets of confirmed bot accounts remain scarce.

This thesis focuses on data collection for bot detection specifically within the context of Reddit. To provide clarity, it is important to establish and define key terms related to both bots and the Reddit platform.

1.3 Terminology

1.3.1 Bot

In a very general sense, bots are automated software systems. In malicious bot detection research, there is a distinction between various types of bots. Javed et al. separates them into four different types: spam bots, scam bots, Sybils, and social bots. Spam bots are created to distribute unsolicited material or advertisements [1]. Scam bots are created for advertising scams, whether that be phishing or cryptocurrency related [20]. There are also Sybils, which are accounts with fake identities. Sybils are

named after Sybil attacks, in which an attacker creates false nodes on a network [21]. Our research focuses on the fourth category of bots outlined by Javed et al.: social bots.

Javed et al. separate social bots into three categories: political bots, misinformation bots, and cyborgs [20]. Political and misinformation bots exist for the sole purpose of engaging in political conversation or pushing content with misinformation respectively. Cyborgs are partially automated accounts that are further classified into bot-assisted human accounts or human-assisted bot accounts [20]. For example, bot accounts on Reddit can be initially created by humans and then automated by software in their interactions, like posting or commenting. Because of the vagueness of this definition, it is essential to distinguish between malicious bots and benign bots, particularly because research on bot detection focuses on detecting malicious bots. Benign bots are typically created by general users on Reddit and are not necessarily an official tool or account that is endorsed by Reddit staff.

A benign bot is one that provides a useful service to a user and does not otherwise spread harmful or unverified information. Examples of benign bots include various automated tools that general users can interact with, such as the **translate-into** bot which can translate a comment or post into another language, or the **RemindMeBot** bot which will send a message to a user to remind them about a post or comment. A benign bot that has been integrated into the site itself for moderators to use as a tool for content moderation is the **AutoModerator** bot, which makes it easier for subreddits to filter out unwanted content.

A malicious bot achieves the opposite: it is one that is created with the purpose of spreading disinformation, harmful material, or non-original content. Examples of malicious bots include a list of 944 suspicious accounts which Reddit discovered after an investigation into Russian attempts to exploit the site. These accounts are

suspected to be of IRA origin, similar to those described when discussing election interference on Twitter/X. Another type of malicious bot that general users find themselves encountering are "karma-farming" bot accounts, which will be described in Section 1.3.2 and 5.1.1.

For the purposes of this thesis, the use of the word *bot* will refer to this definition of malicious social bots.

1.3.2 Reddit

Because the focus of our research is data collection techniques for bot detection on Reddit, there is terminology specific to the platform that must be clearly defined.

Subreddit

Reddit hosts communities dedicated to a specific topic where users can interact. These communities are called **subreddits**, and are the core of the site's functionality. Subreddits can be identified by their unique prefix **r/** followed by the name of the subreddit. For example, a subreddit focused on cybersecurity would be referred to as **r/cybersecurity**. Similarly, user accounts have a unique prefix **u/** followed by the username. An example of this is the CEO of Reddit's account **u/spez**. A Reddit user may also be referred to as a **redditor**.

Karma Farming

Another key functionality of Reddit is its use of **upvotes** and **downvotes** on posts and comments. The difference between these is referred to as **karma**, which is essentially a representation of a user's reputation. One of the many goals of bot accounts is to appear legitimate to other users. This can be achieved with a higher karma

score, which implies that an account has more positive than negative engagement. To do this, accounts will engage in what is known as **karma farming**, where many low-effort, agreeable, or recycled content is posted on different subreddits with the sole purpose of gaining karma.

There are many reasons why bots exist on Reddit, similar to other social media platforms, but a key reason why bots are able to exist and thrive is due to content farming capturing normal user attention [13]. The ability for a bot account to content farm has evolved with the use of artificial intelligence as well, allowing bots to produce more specific, tailored posts that will increase engagement [13]. While bots on other platforms also seek to boost interactions on their posts and comments, karma farming is specific to Reddit as a reference to the voting point system.

Shadow Ban

Another term that is not unique to Reddit but rather online communities as a whole is **shadowbanning**. When an account is **shadowbanned**, their posts and comments will no longer be visible to others, but that is not apparent to the user themselves. A **shadow ban** can only be implemented by Reddit admins and is usually done to suspected bot accounts or users who continuously break rules. A key difference between an account on Reddit being shadowbanned and suspended is that an account can still post and comment when they are shadowbanned, but these interactions do not appear to anyone except themselves. On the other hand, when an account is suspended, the user is no longer able to post or comment on the site. A full and in-depth discussion of the difference between shadowbans and suspensions and how it applies to data collection and bot detection can be found in Section 7.1.

1.4 Contribution

The inherent issue with detecting bot accounts on Reddit in particular is a lack of datasets to train machine learning models. Since Reddit’s controversial API changes in April 2023, which introduced monetary charges for large amounts of API calls and resulted in a site-wide protest, it has become much more difficult to pull data from the site in large quantities without being throttled. Furthermore, there is no concrete way to establish a ground truth, that is, determine whether or not an account is a bot or a human, especially when Reddit does not publish lists of banned bot accounts. Outlining the characteristics that users agree are bot-like and compiling lists of accounts that adhere to these characteristics is essential to furthering bot detection technology. Similarly, collecting lists of accounts based on characteristics that are more human-like can achieve the same goal. Our research addresses the gap in data collection such that research surrounding bot detection can continue to improve on a popular social media site like Reddit.

Chapter 2

Motivation

2.1 Bot Prevalence

The prevalence of bots on social media sites is alarming, especially as they become more complex in the years following advancements in artificial intelligence technology.

As Ferrera et al. state:

"In recent years, Twitter bots have become increasingly sophisticated, making their detection more difficult. The boundary between human-like and bot-like behavior is now fuzzier." [3]

While this paper focuses on Twitter/X bots, the same holds true for other social media sites, including Reddit. According to a Securities and Exchange Commission (SEC) filing by Twitter/X in 2013, the site estimated that false or spam accounts represent less than 5% of their monthly active users [22]. This number, however, has grown significantly since then. While the company has not directly stated bot prevalence since this filing, some academic studies as recent as 2020 estimate that up to 15% of all accounts on Twitter/X are automated bot accounts [23]. This

is a significant increase considering current bot detection techniques lack datasets to properly identify these accounts, especially on social media sites outside of Twitter/X.

Reddit releases Transparency Reports biannually which offer insights and metrics into moderation efforts on the site, both by volunteer subreddit moderators and paid Reddit admins [24]. The reports indicate a separation between accounts banned for spam and content manipulation and other accounts banned or suspended for violating Reddit's Content Policy, which will be described in more detail in Section 5.1.1. Accounts that have been banned for spam or content manipulation are separated due to their disproportionate share of violations, making up 58.7% of all violations in the most recent Transparency Report [24]. While Reddit does not offer any sort of indication how many of these accounts are bots versus humans, the bot accounts that this thesis studies are in that 58.7% separation. In Reddit's most recent SEC Form 10-K filing, they state:

"We regularly deactivate false, spam, and bot accounts that violate our terms or policies, and exclude these users from the calculation of our DAUq [Daily Average Unique] metric; however, we will not succeed in identifying and removing all false, spam, and bot accounts, which means that our DAUq count could be overstated. We are continually seeking to improve our ability to estimate the total number of false, spam, and bot accounts and eliminate them from the calculation of our DAUq..." [25]

Reddit refrains from using the term "bot accounts" in any of their transparency reports, but it is mentioned here in the SEC filing as they describe the continuous effort to determine what accounts are false, spam, or bots. In order to maintain a useful and authentic Internet, bot detection is absolutely essential in social spaces, particularly on a site as widely used as Reddit. Many studies fail to consider Reddit as a means for bots to interact and spread misinformation, highlighting the need for

research. However, even with the importance of bot detection, there lies an issue with doing so: collecting meaningful datasets to test and make determinations on account validity.

2.2 Data Gathering and Testing

In general, it is difficult to achieve higher accuracy in machine learning outcomes without large and accurate datasets. The inherent issue with machine learning bot detection is a lack of datasets that are unquestionably populated with bot accounts or human accounts, which in turn means there is no way to train a model with complete certainty [2]. Outside of company-published lists, such as the 944 Reddit accounts that "were of suspected Russian Internet Research Agency origin" from Reddit's 2017 Transparency Report [24], these datasets are crafted manually and present many biases, especially as bots become more sophisticated and human-like.

There are communities on Reddit that focus on detecting bots using heuristic methods and scripts that they have developed over time, but they are not published such that bot creation and behavior cannot improve. Current researchers have also specified data gathering as a problem because Reddit as a company, as well as other social media sites, do not publish lists of bot accounts that are banned or suspended for being a bot. Many cite that building a robust bot detector for any social media platform is a difficult task due to the lack of large and accurate data sets, which are difficult to obtain or create because of the constant evolution of the bots themselves [2]. Furthermore, when using Reddit's Application Programming Interface (API) to gather a dataset of human accounts, there is still no way for the researcher to know with certainty that these accounts are in fact human. Some researchers compare bot accounts against "normal" accounts in their bot detection techniques, but this can

present some major issues if there exist any bot accounts in that "normal" dataset [26].

Ultimately, this fuels the necessity of research in this area; if there are no established techniques to gather both bot and human data from social media sites, there can never be an accurate classification of them. The collection of data is the first step to classifying whether or not an account is a bot, and techniques to do so intrinsically start with a manual and heuristic approach, but can be assisted with machine learning techniques. This thesis will further the conversation about data collection on social media sites when the label of whether an account is a human or a bot is unknown.

Chapter 3

Literature Review

To understand why collecting meaningful data on Reddit is so essential for the purposes of bot detection, a conceptualization of the current research regarding bot detection techniques, tools for detecting bots, and data gathering is needed. After a review of the current related literature, gaps in the research will be addressed in order to highlight the importance and motivation of this thesis.

3.1 Bot Detection Techniques

There are a few different techniques that are used to detect bots on social media sites. Over time, these techniques have evolved as they become more or less useful in the current bot landscape. Orabi et al. categorize these bot detection techniques into distinct types [2]. The most widely used approach for bot detection involves machine learning, but there also exists a graph-based approach, a crowdsourcing approach, and an anomaly-based approach. In general, a combination of some or all approaches could allow for the most efficient bot detection.

3.1.1 Machine Learning Approach

The most common type of social bot detection is a machine learning approach. Machine learning for bot detection is appealing due to its cost and time effectiveness. Several examples of supervised, semi-supervised, and unsupervised bot detection techniques are outlined by Orabi et al., like the behavior-based tool BotOrNot [2], [27].

One example of a machine learning approach is a study that analyzed a single social media post to determine whether or not an account is a bot. Mohammad et al. created a convolutional neural network (CNN) with Twitter/X data to make a determination on a single post; this analysis uses the content of the posts rather than profile features, which are typically used in a characteristic bot detection approach. The CNN was compared against an artificial neural network (ANN) as well as an approach that combined the two. Ultimately, the CNN model was found to be more accurate and had better performance than the ANN model, and while the combined model had better performance than the CNN model, it required more preprocessing [6].

Similarly, Magelinski et al. propose a CNN architecture with graph classification for detecting bot accounts on Twitter/X, illustrating that the non-linear nature of these models makes them significantly more challenging for botnets to circumvent [18]. Magelinski et al. collect bot accounts using many different currently existing bot detection tools to compare their efficacy. This study also cites difficulties in deployment due to social media API bottlenecks, a common issue in data collection. In general, there is a lack of datasets that are public, accurate, or large enough for proper machine learning processing. Many of the current studies use Twitter/X datasets, but do not focus on any other social media sites where bots may be prevalent, including Reddit, Facebook, or Instagram.

Proper bot detection on the Reddit platform suffers from a lack of large datasets.

More simply, Norlander utilizes simpler classification approaches with supervised machine learning due to the size of the datasets available [26]. Particularly with platforms that do not have large data sets like Twitter/X, using neural networks is far more challenging.

3.1.2 Crowdsourcing Approach

Another common type of social bot detection is crowdsourcing. This focuses on a more manual approach, using groups of humans to make a determination of whether or not an account is a bot. Because humans can more easily identify human-like behavior or nuances like sarcasm, it may seem like a more appealing approach than relying on machines for this task [3]. The idea of a crowdsourcing approach was first introduced in 2012 by Wang et al. who created a Social Turing Test. This study found that humans looking for social bot accounts achieved near-zero false positives, especially if the ones who were detecting bots were well-motivated experts or undergraduate students [28].

However, this approach also has its limitations. First, this approach was introduced in 2012, when social media platforms were far less popular than they are today. In 2017, Cresci et al. researched the accuracy of a crowdsourcing approach and found that while this approach succeeded in identifying traditional spam bots (such as email), it failed to detect social spam bots [1]. To implement this approach on a much larger scale, particularly as bots have become more sophisticated and harder to detect, could lead to more false positives. Because of the increased popularity of social media and the rise of social bots, it is not a cost or time-effective solution in detecting bots and there must be some sort of machine learning involved [3]. Due to the lack of bot versus human account datasets, particularly on Reddit, a combination of both a machine learning and crowdsourcing approach may be necessary for the

future of bot detection.

3.1.3 Graph-based Approach

A graph-based approach to bot detection utilizes the mathematical concept of a graph in order to model relationships between objects, in this case, a social network structure to classify accounts as bots or humans [2]. Several popularized graph-based approaches use machine learning in combination, such as the *Íntegro* system introduced by Boshmaf et al. [29]. *Íntegro* focuses on fake accounts as opposed to malicious social bots, providing general users with a ranking scheme to detect them. The system predicts fake accounts based on user-level activities and ranks these accounts as weights on the graph [29]. Other approaches may use a graph-based approach in order to visualize and represent the information, while machine learning methods are then used to detect the bots [2].

Hurtado et al. utilize the graph-based approach for bot detection on Reddit, focusing on political discussion with the subreddit `r/The_Donald`, a popular community of Donald Trump supporters in 2016. This study uses the maximum edge weight metric to identify abnormal users, which includes bots [30]. Maximum edge weight determines the strongest single connection an account has with another account. After an analysis of two other large subreddits compared to `r/The_Donald`, the researchers determined that maximum edge weight was an effective way to detect abnormal behavior because of the sheer difference in edge weights. This study was done in 2019, and there is a lack of more current examples that are specific to Reddit, likely due to the API changes and overall lack of datasets available.

3.1.4 Anomaly-based Approach

An anomaly-based approach to bot detection is based on the assumption that legitimate, human users on social media have no reason to act with abnormal behavior. This assumption then leads to the conclusion that if an account is behaving abnormally, it is more likely to be a bot account [2]. Orabi et al. separate the anomaly-based approach into two techniques: action-based and interaction-based.

Action-based detection focuses on the actions that accounts take. This mainly includes posting behavior, as outlined by Echeverria and Zhou in their analysis of a ‘Star Wars’ botnet on Twitter, where over 350,000 bots posted random quotes from the series [31]. This study outlines the fact that these accounts are posting unlike humans. Action-based detection typically relies on metrics such as posting frequency or content similarity to determine if an account is a bot or a human.

Interaction-based detection focuses on the interactions between accounts. Lee et al. provide an example of interaction-based detection with their creation of honeypots, or accounts created with the sole purpose of attracting ‘content polluters’, to conclude that accounts that engage with these honeypots are more likely to be malicious [32]. This shows that bots are more likely to interact with each other with positive engagement to increase their legitimacy on the platforms.

3.2 Bot Detection Tools

Many researchers and general social media users have created bot detection tools to curb the increase of misinformation and content pollution. These tools continue to evolve alongside bots. Particularly with the popularity of generative AI, it will become more difficult to create effective bot detection tools. Many bot detection tools focus on one social media site, primarily Twitter/X, but Ng et al. propose BotBuster, a

multi-platform bot detector with a focus on both Twitter/X and Reddit that uses supervised machine learning with manual human annotations [33]. The following subsections are separated into bot detection tools on Reddit and Twitter/X.

3.2.1 Reddit

In 2018, Norlander focused on detecting Russian bots on Reddit, using the declared list of 944 Russian bot accounts that was published by Reddit in their 2017 Transparency Report. This list of bots allowed for the researcher to classify and analyze the data, comparing normal users in the same time frame as the bots. Normal users are assumed to be humans, which emphasizes the need for better characterization of not only bot accounts, but human accounts as well. The tool, called Reddit Bot Classifier, utilizes supervised machine learning and decision trees to classify an account as a bot or normal user [26]. The study focuses on post title and subreddit, comment body and subreddit, and finally, some account characteristics. Each of these features were analyzed individually, rather than as a collective. The limitations with this paper are the age of the data and the accounts; bot behavior and tactics have increased in sophistication since 2017 and have thus become more human-like, ultimately leading to a more difficult detection.

Saeed et al. present another detection system, TrollMagnifier, trained on the same Russian bot data. The model uses a combination of a graph-based approach and supervised machine learning, utilizing four different classifiers for training. TrollMagnifier was able to identify 1,248 potential troll accounts, but faces the same issues with a lack of meaningful datasets and confirmation that their tool was successful in identifying bot accounts [34].

3.2.2 Twitter/X

There are many bot detection tools that exist specific to Twitter/X. As stated previously, Twitter/X underwent strict API changes in 2023 which removed the free tier altogether, making it more difficult for developers and researchers to access data from the site [35]. Unlike Reddit, however, there are many publicly available datasets that researchers have been able to use for bot detection purposes. Below is an analysis of four different bot detection systems for Twitter/X that utilize a variety of the techniques described previously.

BotOrNot

Davis et al. created BotOrNot in 2014, a publicly available tool that classified Twitter/X accounts into a scale of human or bot. Because this tool relied on the Twitter/X API, it is no longer functioning, but between 2014 and 2016, the tool served over one million requests and classified over nine hundred thousand unique user accounts [27]. A user provides the tool with a Twitter/X username, and it would obtain the recent history of that account as well as posts that mention it. The data then runs through their classification algorithm. BotOrNot’s classification algorithm, a Random Forest supervised machine learning model, reviews over 1,000 features which are grouped into 6 classes: network, user, friends, temporal, content, and sentiment [27]. The ultimate goal of this tool was to offer a free bot detection service for researchers, reporters, and enthusiasts [27] and this tool served as one of the first popularized, comprehensive bot detection tools for social media.

BotWalk

BotWalk, an adaptive bot detection algorithm for Twitter introduced by Minnich et al., offers a different approach to typical bot detectors in its use of unsupervised

learning to detect bots as they change and evolve [36]. Similar to the features analyzed in BotOrNot, BotWalk reviews metadata, content, temporal, and network-based features. This system uses a combination of anomaly-based detection and machine learning with an unsupervised approach. The dataset that BotWalk uses was acquired with the Twitter/X API, which is now less accessible for researchers and would require payment.

DeBot

DeBot analyzes the correlation between user accounts on Twitter, concluding that highly synchronous accounts are more likely to be bots. Chavoshi et al. use an anomaly-based approach in combination with an unsupervised learning model, similar to BotWalk [37], but with a focus on detecting bots faster than Twitter/X can suspend them. DeBot is also trained on all languages and not just English, which Chavoshi et al. describe as the reason their model is more accurate than BotOrNot [37].

RTbust

Similar to the previous two Twitter/X bot detection tools, RTbust utilizes an unsupervised feature extraction approach. Unlike the other tools described, RTbust implements clustering to analyze malicious retweeting patterns [38]. RTbust's data collection includes almost 10 million retweets gathered in a 2-week span, collected with Twitter/X's API. The model uses clustering to distinguish between normal tweeting behavior and anomalous or suspicious behavior, visualized with a technique deemed ReTweet-Tweet (RTT) [38].

3.3 Data Gathering

Orabi et al. describe in depth the issues with this research area: collecting a labeled dataset. However, there are ways to manually annotate datasets to assist in bot detection, which this thesis addresses, while still acknowledging the inherent issues and limitations with this approach.

"Even though the resulting labeled datasets are not perfectly accurate, they enable researchers build practical models or get acceptable performance measurements. These labels are usually produced by human decision on social media accounts, automated methods that have almost perfect precision, or odd behavior exposed by some accounts that is very unlikely to be originated by legitimate users." [2]

We are still able to attempt classification of accounts with manually labeled datasets, however, this process is slow and as it will be shown in future sections, does not produce large amounts of data [2]. In addition to this, Orabi et al. summarizes the collection of datasets for bot detection with a focus on Twitter, and as such, there still remains a gap in dataset collection for Reddit.

Much of the previous research in this area collected data with the Twitter/X or Reddit API. However, due to the 2023 API changes, the process of data collection for researchers and general developers has become more difficult [14], [39], [20]. Javed et al. describe how much of a financial burden these changes bring to researchers, particularly independent researchers who focused on Twitter/X due to the number of public figures who use the site [20]. Reddit is still compatible with the Python Reddit API Wrapper (PRAW), which our research utilizes; however, due to recent API changes, rate-limiting measures are in place that restrict the number of requests that can be made to the API within a given time. Twitter/X no longer offers a free tier for API calls.

3.4 Current Research Gaps

Researchers collectively agree that the inability to acquire large and meaningful datasets hinders the development of bot detection [2], [20]. This is, in part, a result of API changes on large social media sites like Twitter/X and Reddit, which make it more difficult and expensive for developers and researchers to access large amounts of data [40]. Additionally, because there is no ground truth established in this area, research can be very limited. Social media companies do not typically publish lists of known bot accounts after they are suspended and the majority of datasets that contain bots are compiled manually by users [33]. The few studies that focus on Reddit rely on the same set of Russian bot accounts published by the site in 2017 because there is not a more recent ground truth for bots. Many studies make assumptions in their data in order to create detectors or analyze detection techniques, such as assuming normal users, that is, randomly acquired users who interact with the site, are human. This assumption paired with machine learning to determine if an account is a bot will naturally introduce error into the models [26]. An outline of what characteristics make an account appear more bot-like versus more human-like is necessary, but these characteristics would need to be constantly evolving alongside the bots. Many Reddit tools created for the purpose of bot detection keep their criteria hidden so that bots cannot use them to improve, however, this hinders research in this area, resulting in a Catch-22.

Finally, much of the research surrounding bot detection on social media focuses on Twitter/X, and does not consider a popular social media site like Reddit. Bot detection tools that are specific to Twitter/X may not work on Reddit due to the specific features that are employed [40]. Given that Reddit functions differently from other social media platforms, it is essential to establish an overview of typical bot behavior on major platforms as a basis for comparison.

Chapter 4

Bots on Social Media

Bot behavior varies across different social media platforms. This chapter focuses on two key questions related to bots: how they are used on a specific platform, and what are the specific characteristics that can identify them. First, it is important to review the platform where most bot detection research already exists: Twitter/X. Next, there will be a short overview of other social media platforms with unique use cases. Lastly, an in-depth analysis of bots on Reddit will lead into the focal point of this thesis.

4.1 Twitter/X

Twitter/X has been the primary focus of bot detection research because its API made bot-related datasets widely accessible. Although this task has become more difficult since the 2023 API changes, it is still important to discuss the differences in bot behavior and characteristics from platform to platform. Twitter/X functions differently than other social media sites with a focus on shorter posts compared to Facebook or Reddit. The Trending tab, which is accompanied by keywords and

hashtags that are being posted with high frequency, allows for bots to quickly boost agreeable content that can be seen by a wide audience. This includes a vast amount of misinformation, as indicated by previous studies on the topic [41], [42], [43].

In general, Javed et al. describe five key features that are used to identify harmful Twitter/X bots. This includes content, sentiment, account information, account usage, and social network [20]. The vast majority of bot detection relies on content as a feature, which includes the posts that users make, as well as the social network, that is, how users interact with each other. However, our research focuses on the account information and account usage features, which are attributes specific to the account, such as creation date or number of posts over a specific period of time.

Given the analysis of bots on Twitter/X, it is also important to present some unique use cases on other social media platforms like Meta, Parler, and Truth Social.

4.2 Other Social Media Sites

Although recent bot detection research focuses primarily on Twitter/X, other social media platforms offer noteworthy examples of bots that contribute to a broader understanding of their behavior, interactions, and overall impact. This includes platforms like Meta alongside more niche communities such as Parler or Truth Social.

4.2.1 Meta

Meta, a technology conglomerate, owns and operates two major social media platforms: Facebook and Instagram. Much of the research focused on bot detection excludes Meta due to the difficulty of data extraction. However, there still exist bots on these sites, so their behavior must be addressed.

There are few studies that focus on general bot detection on Facebook, even

fewer that focus on social bots in particular. Obadimu et al. conducted a sentiment analysis on Facebook bots, which is a method of using natural language processing to determine whether the content of a message expresses a positive, negative, or neutral attitude [44]. In a comparison between Facebook bot entities and Twitter/X bot entities, Obadimu et al. find that the term "Facebook bot" had a much higher positive sentiment than "Twitter bot" [44], arguing that this may be a result of both a perception of misinformation on Twitter/X as well as an affiliation of political figures, which would indicate higher levels of polarization [45], [44]. In general, bots on Facebook focus more on general engagement and conversation than those of political Twitter/X bots. To detect these bots on Facebook, one study suggests that normal users are likely to make posts with a richer vocabulary, more content, and an overall increased readability compared to bot accounts [46].

Instagram began as a platform to share photos and videos, but this has changed over time to include a wide usage by social media influencers for marketing. In turn, many bot accounts on Instagram focus on increasing engagement and follower count to gain a wider reach for advertising products [47]. One way to detect bot accounts on Instagram is to consider the follower-to-like ratio, that is, the number of followers an account has compared to the number of likes they receive on posts. An account that has many followers but receives a small amount of likes on their posts is indicative of bot-like behavior [47]. Additionally, an account with a large amount of posts in a short amount of time may also signal bot-like behavior; in order to appear more human-like and realistic, many posts are made on account creation to seem like a genuine profile [47].

4.2.2 Parler

Parler describes itself as an "uncensored social media for free speech and authentic engagement" [48]. Alipour et al. note that users who prefer communities with less moderation will engage with platforms like Parler, but that also allows malicious users to more easily deploy bot accounts [40]. There is little published research done on Parler, but Alipour et al. argue that while bot detection tools more easily identify bots on Parler than other social media platforms, the behavior of human users often mirrors that of bots, increasing the likelihood of misclassification.

Parler is a unique use case for bots on social media due to its advertisement of being an "uncensored social media". This is vastly different from other popular platforms like Twitter/X, Meta, and even Reddit. Another similar case to Parler is the Truth Social platform.

4.2.3 Truth Social

Truth Social is a highly political platform founded and released by President Donald Trump in 2022. Truth Social states that it is a platform which encourages an "open, free, and honest global conversation without discriminating on the basis of political ideology" [49].

Galicia analyzed Truth Social's algorithm and concluded that it often recommends bot posts to normal users, which results in a high level of engagement with the bots. "Presumed bots on Truth Social have a disproportionate power to set conversations on the platform which allows bad actors room to easily manipulate the algorithm's flaw" [50]. Galicia also noted that the majority of the content posted by presumed bot accounts were pictures or meme-like content, which differs from bot behavior on other platforms. Kolk makes a similar distinction, noting the use of memes as well as a lack of creativity or originality in bot posts [51].

These traits of bot behavior are not limited to the previously mentioned platforms and can also be observed on Reddit.

Chapter 5

Bots on Reddit

In order to understand the context behind bots on Reddit, a discussion on the platform's structure is necessary. Reddit is a social network based in the United States that revolves around community forums and three central user actions: posting, commenting, and voting. Reddit is a popular platform, with an average of 101.7 million daily unique users as of December 2024 [25], and is the third most visited social media website in the world, just behind Meta and Twitter/X. While Reddit differs from Meta and Twitter/X in that it is more anonymous and forum-based, there is no denying its popularity. The popularity of the site results in millions of accounts and pieces of content being removed for spam, vote manipulation, or other violations of Reddit's Content Policy [24]. Because of its popularity and differences from other social media platforms, Reddit presents a unique challenge for bot detection. This chapter will provide a comprehensive discussion of the current bot landscape on Reddit.

5.1 Reddit Bot Landscape

While many of the factors driving bot activity are common across all social media, Reddit's platform-specific features create further opportunities for bot presence. This primarily includes karma farming for the purpose of selling accounts, which is more easily done on a site like Reddit compared to Twitter/X, since Reddit is an anonymous forum. Reddit user u/Blackfeathr notes that "Karma lends legitimacy to an account on Reddit. It makes a user seem more "trustworthy" which is obviously the goal, especially if you're trying to sell or make fake reviews for a product or service" [52]. The abundance of bots on Reddit largely exists because they can be sold. However, there has been limited research on Reddit bot detection due to the challenges associated with obtaining datasets. Many of the characteristics described in this section were gathered from bot-hunting communities and individual users who create bot detection tools, which will be described further in Section 6.1.2. It is important to note that even if bot hunters can agree that bots express specific characteristics, it is possible for human accounts to share the same characteristics as well. In a similar vein, there may be characteristics of bots that are so human-like that it may be impossible to distinguish them. As u/Blackfeathr describes:

"The [bots] I'm talking about are the ones that try to blend in with everyone else. They try to trick you into thinking they're real people. They are the most insidious of all, because when they are done with their first task, gaining karma, they move on to more nefarious tasks after being sold to whoever is willing to buy. These activities range from spreading misinformation/disinformation, propaganda, promoting a product, or outright scamming people with bootleg dropship merch. There is a large market for buying high karma accounts, and businesses, governments, and other entities will pay big bucks to have that kind of influence" [52]

The purpose of this subsection is to describe the different types of bots that currently exist on Reddit as well as provide a description of known characteristics that bot hunters agree upon. These characteristics will be used to develop data collection methods for bot and human accounts in Chapter 6, but as bots become more advanced, these traits may change, requiring constant monitoring and further research.

5.1.1 Types of Bots

There are multiple different types of bots that exist on Reddit. Similar to the types of bots distinguished in Section 1.3.1, bots on Reddit can be categorized even further by the actions they take. As described previously, Reddit releases Transparency Reports biannually which describe moderation efforts on the site. The most recent report, outlined for January to June 2024, provides some helpful statistics and analysis into the type of content or accounts that are removed due to Content Policy violations, under which bot activity is categorized. Reddit describes their Content Policy as "a set of principles-based rules that apply to the entire Reddit platform, including our users and content on Reddit" [24]. The main rule in the Content Policy that applies in this context is Rule 2. The rule is as follows:

"Abide by community rules. Post authentic content into communities where you have a personal interest, and do not cheat or engage in content manipulation (including spamming, vote manipulation, ban evasion, or subscriber fraud) or otherwise interfere with or disrupt Reddit communities" [24].

Reddit also provides helpful insight into what content manipulation actually is. This includes vote manipulation, defined as "creating and employing multiple accounts, voting services, or any software to manipulate vote counts" [53], but perhaps more

important is Reddit's definition of spamming. Reddit provides multiple examples of behavior that may be considered spam [54], included but not limited to:

- Repeatedly posting unrelated/off-topic/link-farmed content.
- Repeatedly posting the same or similar comments in a thread, subreddit or across subreddits.
- Programming bots that harm/break Reddit, including bots intended to promote content/products/services.

Bot hunters and Reddit users alike agree that bot behavior on Reddit fits into these criteria. Content manipulation is the umbrella that the bots analyzed in this thesis fall under, and the majority of "bot hunting" guides created by general users focus specifically on content manipulation bots. Karma farming bots are responsible for significant amounts of content manipulation on the platform, as reported by both bot hunters and everyday users. Both in the lack of posting authentic content and engagement in vote manipulation, these bots present a very Reddit-specific issue that human users face when interacting with the platform every day. Furthermore, Reddit states that

"...the vast majority of admin removals were for content manipulation, which includes spam (accounting for 66.5% of these removals), and issues like vote manipulation (attempts to interfere with Reddit's upvote/downvote tallies) and other attempts to artificially promote content (collectively making up 1.8% of these removals)" [24].

This indicates that the majority of the content removed from the site by admins is spam. It is important to note that this does not necessarily mean that most or all content removed for spam was posted by bots, only that they are under the same

umbrella. Both the type of content removed by admins as well as a full breakdown of the content removed by Content Policy violations during the time frame of this Transparency Report can be seen below.

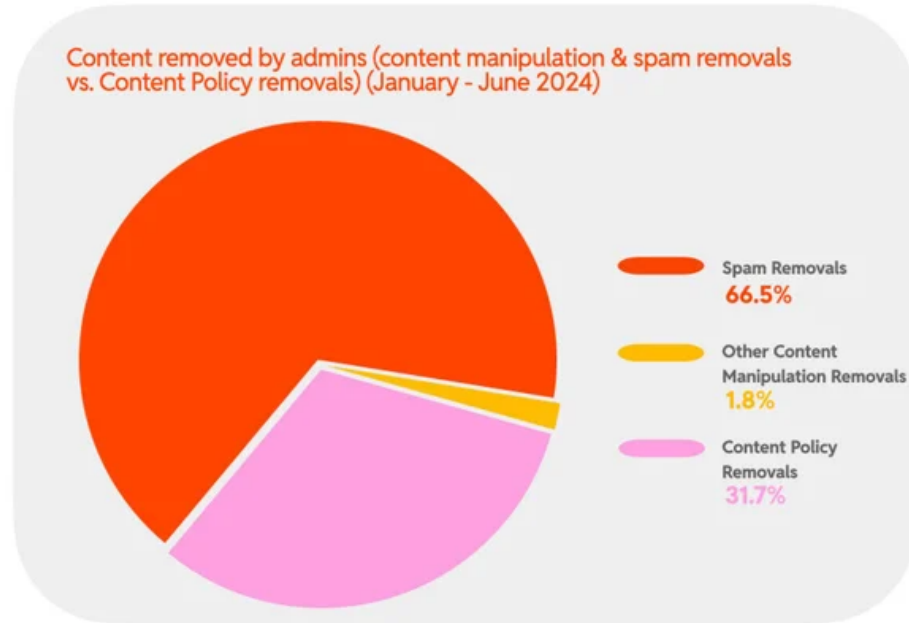


Figure 5.1: Content removed by Reddit admins from January to June 2024 [24].

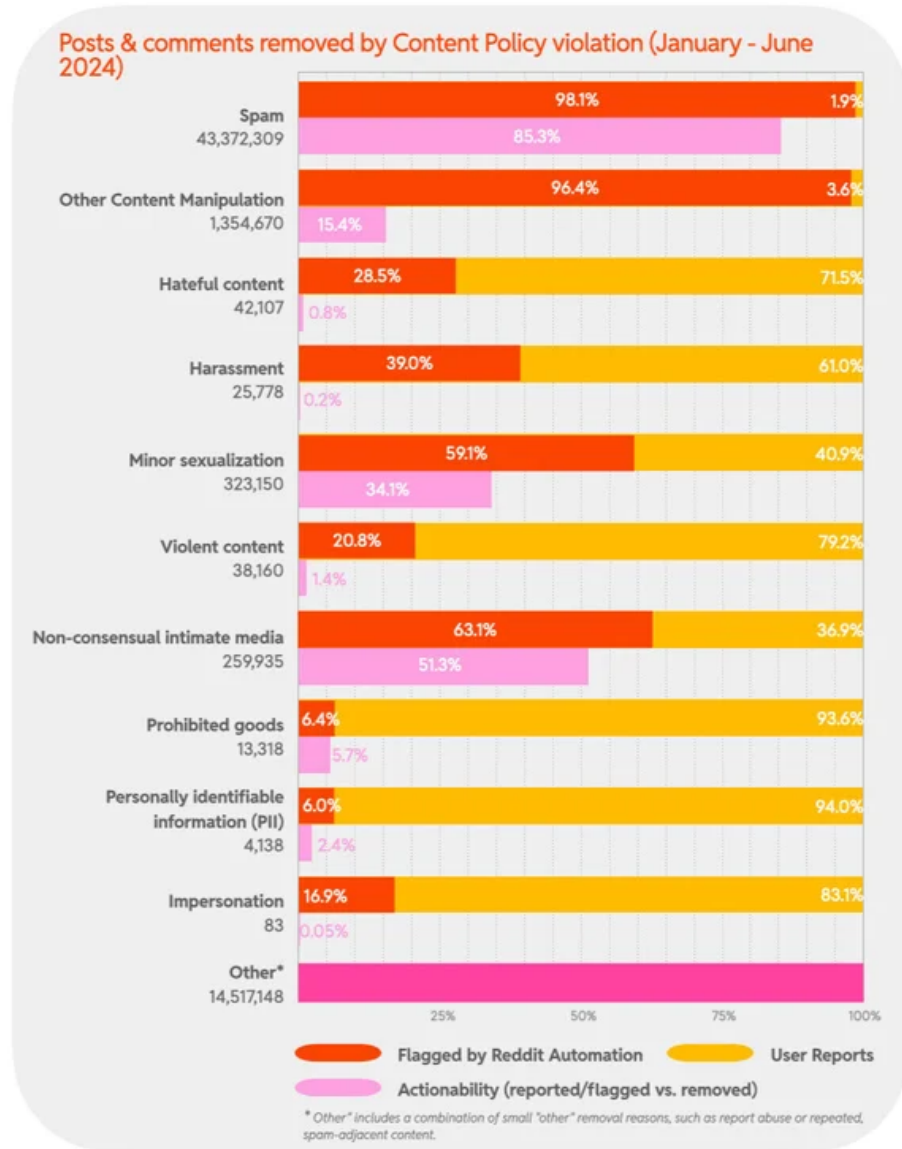


Figure 5.2: Content removed by Reddit for Content Policy violations from January to June 2024 [24].

As indicated by the figures, the majority of content removed for Content Policy violations falls under the same umbrella as bot activity: spam and other content manipulation.

The following two subsections describe the main characteristics used to determine if an account on Reddit is a bot or not: the content that they post and the attributes of the account.

5.1.2 Content

Much of the research surrounding bot detection focuses on the content posted by accounts. Whether it be text in the form of a post or comment, or images and memes, analyzing the actual content is a common way to determine if an account is a bot. Saeed et al. discuss the previously mentioned Russian bot or troll accounts, stating that

"...troll accounts are more likely to reply to each other or to make submissions with the same title. We identify several features that characterize troll accounts (e.g., the fraction of comments made on submissions by troll accounts or the fraction of submissions with the same title as a troll account's submission) and use them to train classifiers and identify additional troll accounts in the wild" [34]

One recurring feature that is mentioned throughout bot-hunting subreddits is repeated reposted content. Because it is the easiest way to farm karma on Reddit, many bots will repost very popular posts from months or years ago in hopes of getting the same engagement as the original post did.

Additionally, low-effort, generic post titles and content are also common. Reddit user u/tyrannosnorlax details the easiest way to determine if an account is a bot is through the contents of their comments. They describe how bots will most likely engage with the most popular comments on newer, rising posts [55]. These comments are also typically short, agreeable, and common; examples included in their post about identifying bots are phrases such as "I agree dude" or "Well said", but u/tyrannosnorlax also compares bot speech to that of Yoda from Star Wars, or object-subject-verb word order, such as "Well spoken, you are" or "Perfectly said this is" [55]. However, a very common technique that these repost bots utilize is not only reposting the original post, but also reposting comments from other users on that

original post as well [55], [52], [56]. In addition, these bot comments are primarily posted on popular subreddits such as r/AskReddit, which is described by multiple users who have created bot-hunting guides on the platform [52].

5.1.3 Account Attributes

Since most research tends to prioritize bot account data collection while overlooking methods for gathering human account data, it is critical to examine both types of datasets to ensure a machine learning model’s accuracy in predicting bot accounts. The main features that this thesis focuses on are account attributes of both bots and humans. This includes username, post and comment count, karma scores, and other data-focused metrics of an account.

There are general account attributes that may indicate bot behavior. Account creation date is a prominent characteristic that many bot hunters use because bot accounts get suspended frequently; many suspicious accounts average 90 days in age, but typically remain under a year old [55]. Furthermore, u/blackfeathr notes that bot accounts typically do not have a profile picture, or have a randomized Reddit avatar [52]. Bot hunters also consider an account’s post-to-comment ratio. An account with many posts but no comments is suspicious to bot hunters, as that is not how a human would typically interact with the platform [52]. Karma score is another attribute considered by bot hunters, though on its own it is insufficient to reliably distinguish between a bot and a human, as bot accounts seek to increase their karma as quickly as possible.

Another common way that bot hunters identify bot accounts on Reddit is by their username. Bot hunting guides describe common username conventions of bot accounts, separated into three categories. First are random strings of numbers and letters, with examples being UcGhz6NmE, BuYtlpHEq, or GHJKxse7y. Second are

generic first and last names that sound human, typically female. These names often add on a string of letters at the end, which u/tyrannosnorlax notes is usually 's'. The examples provided are MariaJamesss, MeganAnthonysss, and OliverWilliamsss. Last, there are the default Reddit usernames, which follow a convention of Word-Word####, like Wild-Laundry5628 for example. These three categories are outlined by u/tyrannosnorlax and u/blackfeathr in their respective bot hunting guides [55], [52].

It is important to note that relying on only one or two of the outlined characteristics to identify a bot will result in false positives, because humans may exhibit similar attributes. The above bot classifications are mainly heuristic methods that bot hunters agree upon based on their experiences with bots on the platform. With the comprehensive understanding of why bots exist on Reddit and what they might look like, the following two chapters will present our methods and analysis of data collection techniques for both bot and human accounts.

Chapter 6

Methods

This chapter outlines the methods used to gather datasets of both bot and human accounts on Reddit and also discusses the methods of machine learning used to compare the datasets. Building on previously discussed attributes associated with bot behavior and introducing new strategies for compiling human account datasets, these methods will support future bot detection research in establishing a ground truth for training and evaluating machine learning models.

6.1 Data Gathering

This section details the data collection methods employed in this thesis, beginning with an overview of how the Python Reddit API Wrapper (PRAW) is used to extract data from Reddit, followed by a discussion of the specific bot and human datasets compiled.

6.1.1 PRAW

The Python Reddit API Wrapper (PRAW) is an open-source Python library that enables users to interact with Reddit’s API. It allows researchers and developers to efficiently retrieve data from Reddit without manually handling individual API requests. To use PRAW, it is necessary to create an application on the Reddit preferences page. The data collection for our research required only a script-type application, which is the simplest configuration to implement with PRAW. Once this setup is complete, data can be accessed by creating an instance of the Reddit class to establish a client connection. PRAW’s comprehensive documentation offers resources for retrieving various types of data from the platform. Several attributes of a Redditor, that is, the PRAW class representing Reddit users, are used in the data collection techniques described in this thesis, including `username`, `link_karma`, `comment_karma`, `submissions`, and `comments`. With a foundational understanding of how PRAW is used to collect data from Reddit, the following sections will describe the methods for assembling bot and human account datasets.

6.1.2 Bot Detection within Reddit

To contextualize how bot account datasets are currently compiled on Reddit, we first outline the current state of bot detection within the platform itself. There exist multiple subreddits and benign bot accounts on Reddit that function for the sole purpose of detecting malicious bots.

Examples of bot detection accounts are as follows:

1. `u/bot-sleuth-bot`
 - `u/bot-sleuth-bot` is a bot created by a Reddit user that can be prompted by other users. The bot runs on a point system that performs various

checks on the account and counts how many "suspicion points" accrue. It then divides these points by the total number of possible points, and responds to user prompts with a "suspicion quotient" value from 0 to 1 for the likelihood that the account is a bot. The majority of the specific criteria that this bot uses to make a bot determination are not public, and the author of the bot specifies this is because unethical bot makers might use the knowledge to improve their bots [57].

2. u/HelpfulJanitor

- u/HelpfulJanitor is a bot created by a Reddit user that can be added to a subreddit for moderation purposes. This bot will scan the subreddit for new posts and check the author's post history, and if it meets certain criteria, it will remove all posts by the user and then ban them from the subreddit [58]. Similar to u/bot-sleuth-bot, this bot has its own list of criteria to determine if the account is a malicious bot, which is not public.

3. u/bot-bouncer

- u/bot-bouncer is a moderation bot whose purpose is to protect subreddits against harmful or disruptive bots, similar to u/HelpfulJanitor. This bot has its own dedicated subreddit, r/BotBouncer, where suspected bot accounts are posted and then classified by both automated and human classification [59]. The majority of the suspected bot accounts collected for this thesis are sourced from r/BotBouncer, which is described in further detail in Section 7.1.

Examples of bot detection subreddits are as follows:

1. r/RedditBotHunters

- r/RedditBotHunters is a community that discusses bots and their patterns on Reddit, such that they can be identified. This subreddit is full of resources put together by a community of users who seek to identify bots, including guides, tools, and actions a user can take once they detect a bot [60].

2. r/TheseFuckingAccounts

- This subreddit is run by the creator of the u/HelpfulJanitor bot, and serves as a hub for users to submit and track accounts that are suspicious [61]. It is updated multiple times a day with potential bot accounts detected by u/HelpfulJanitor as well as other users in the community.

3. r/botwatch

- Similar to r/RedditBotHunters, r/botwatch is a community that is "dedicated to the continued interest, observation, discussion, and study of the bots that dwell in Reddit" [62]. Posts on this subreddit are less frequent than the other two subreddits mentioned above, but still contain resources for users to identify bots on Reddit.

These communities and tools created by Reddit users are essential to furthering the research of bot detection on the site. Much of the contribution provided by this thesis could not have been possible without the bot-hunting communities within Reddit, who share the same goal of identifying, reporting, and ultimately removing malicious bot accounts altogether. With an understanding of where current bot-hunting communities on Reddit reside, we can continue with a detailed discussion of the bot and human datasets themselves.

6.1.3 Bot Datasets

Bot account datasets are a central focus in most bot detection research, particularly because obtaining them is a well-documented challenge. Current research indicates that the majority of bot account datasets are either collected manually or acquired with previous bot detection tools, like ones described in Section 3.2. However, both approaches introduce inherent biases. Manual collection often results in selection bias, as human annotators tend to focus on accounts that display overtly suspicious behavior, potentially overlooking more subtle cases. Tool-based collection relies on models trained with this selection bias, inheriting errors from them. The goal is to minimize the amount of error introduced by collecting accurate datasets for use in machine learning models. Additionally, because Reddit differs from other social media platforms in structure and user behavior, tools designed for platforms like Twitter/X may not be directly applicable. As a result, alternative methods for collecting bot account data on Reddit are necessary.

Many researchers focusing on Reddit bot detection rely on the only official list of suspicious accounts published by the platform. As previously discussed, Reddit CEO Steve Huffman (u/spez) announced a list of 944 suspicious accounts following an investigation into Russian attempts to exploit Reddit. The accounts are kept visible for transparency purposes, but the announcement states that they will be removed in the future [63]. Notably, the announcement also stated:

"...our investigation did not find any election-related advertisements of the nature found on other platforms, through either our self-serve or managed advertisements. I also want to be very clear that none of the 944 users placed any ads on Reddit. We also did not detect any effective use of these accounts to engage in vote manipulation" [63].

This raises concerns, as these accounts did not engage in any of the commonly es-

tablished behaviors associated with bots on Reddit. Yet, researchers continue to use this list as the ground truth for identifying bot accounts on the platform, which risks undermining the accuracy and reliability of detection models trained on this data. It is also important to note that this list was published in 2018. In the seven years since, bots have evolved significantly, and differentiating them from human accounts has become a far more complex task.

The bot account data collection techniques introduced in our research have their basis in the bot-hunting communities within Reddit itself. Redditors have the necessary experience and knowledge surrounding typical human interactions and those of an anomalous user, which can be used to outline bot-like characteristics and automate detection. As the goal of data collection is to minimize the amount of error introduced in machine learning models, we focus on Reddit-specific characteristics that bot hunters collectively agree on, all previously established in Section 5.1. Using this information in collaboration with moderators from the largest bot-hunting subreddits and developers of Reddit bot detection tools described in Section 6.1.2, we compiled a list of 4,844 suspected bot accounts. This list includes accounts with recent activity and exhibits all the bot-like characteristics previously outlined. While it is still possible that some human accounts are included, the list represents the combined efforts of popular bot hunters on the platform, many of whom do not publish their criteria. Additionally, many of the accounts on this list are suspended or shadowbanned daily; a more detailed analysis of this fact will be provided in Chapter 7. It is generally inadvisable to rely on one or two bot-hunting techniques. For example, the u/HelpfulJanitor bot posts lists of suspected bot accounts multiple times per day. However, depending on this one bot hunter will introduce risk in the future, as its continued operation is contingent on ongoing support from its developer.

The list of suspected bot account usernames was compiled collaboratively by mod-

erators and users of bot-hunting subreddits (mainly r/BotBouncer and r/RedditBotHunters) as well as the developers of the most popular bot-hunting bots. Before an account is added to the list, it is reviewed by a team of bot-hunters from these subreddits. The bot criteria that these hunters use have been developed for years by the entire bot-hunting community and are constantly evolving alongside the bots themselves. This list of 4,844 account usernames was provided by an anonymous bot-hunting developer for our research purposes. We used PRAW on the list to create a bot account dataset with the following account characteristics: Username, Account Creation Date, Post Karma, Comment Karma, Total Karma, Post Count, and Comment Count. Two other attributes were manually added to this list for easier classification: Post Comment Ratio, which is the total number of posts divided by the total number of comments, and Bot Label. The Bot Label attribute is a boolean value assigned to accounts based on the list classification; in this case, a suspected bot account will have the label 1.

Because, to our knowledge, this is the largest and most up-to-date dataset of suspected bot accounts on Reddit, it is used for training with the various human account data gathering techniques in Chapter 7. Ongoing collaboration between researchers and bot-hunting experts on the platform is essential, particularly given the limited academic research on Reddit compared to platforms like Twitter/X. Currently, collaboration remains one of the most effective ways to collect bot account data for this type of analysis. However, there is an even larger gap in collecting human account datasets on Reddit, which will be addressed in the next section.

6.1.4 Human Datasets

Separate from bot account datasets, another inherent issue with machine learning bot detection techniques is establishing a ground truth of human accounts. Typi-

cally, researchers collect datasets with randomly gathered posts or accounts, which are not necessarily free of bots, introducing error into the machine learning model. These randomly gathered accounts are sometimes referred to as "normal" accounts. Therefore, it is essential that researchers can selectively compile the human account dataset in the same way they curate the bot account datasets. We propose methods to collect human datasets on Reddit by making a determination of which accounts are more likely to be human, as well as in which subreddits they are more likely to participate.

There are two sets of accounts that are all but guaranteed to have a human running them: public figures and Reddit administrators. First, when considering public figures, Reddit differs from other social media platforms in that they are not necessarily verified accounts, however, one of the most popular subreddits, r/IAmA, offers a space for public figures to answer questions from general users of Reddit. r/IAmA hosts interviews which are referred to as an "AMA" or "Ask Me Anything", with interviewees ranging from Barack Obama and Bill Gates to other non-famous people. The interviewees are required to submit proof that they are who they say they are, typically in the form of a photo of the interviewee holding up a sign with their Reddit username or the date and time of the AMA. To ensure the highest certainty of human accounts, we collected accounts that are associated only with public figures and featured as some of the most popular AMAs of all time.

Second, Reddit administrators are paid employees of Reddit Inc. who work directly on the platform, whether that be through content moderation or development. A list of current Reddit administrators is available on the r/reddit.com moderators page. Administrators can be verified with their Reddit Admin badge displayed on their profile, shown in Figure 6.1.

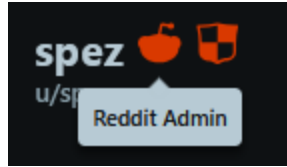


Figure 6.1: Reddit Admin badge on the CEO of Reddit’s profile [64].

However, this is not a complete list as it only includes administrators who engage publicly. Both this list of Reddit administrators and public figures from the r/IAmA subreddit have the highest degree of certainty regarding whether an account is operated by a human, because they must be verified in some way.

Additionally, we accumulated a list of communities where humans are more likely to interact. This technique of data gathering is supported by Ng and Carley and their BotBuster tool: "For Reddit humans, we collected users from 5 subreddits that generally require conscious writing and thought and manually verified the users are likely to be humans" [33]. One basis for human account data collection is that humans will participate in subreddits that require users to write and think consciously. Another basis for human account data collection is that because bots are more likely to post on larger, more generic subreddits, there is likely to be more human activity on smaller, more niche subreddits [52]. We selected smaller, niche subreddits where there is likely to be human activity and separated them into three different categories: hobby, location, and university. We chose these categories of subreddits because they are places where humans are more likely to interact, as suggested by Ng and Carley [33]; these communities are smaller and focused on a very specific topic or location, and it has been established that bot accounts tend to post on larger, more generic subreddits [52].

The list of hobby-based subreddits was collected manually by searching through Reddit’s Explore page, which separates subreddits into different categories. Additional entries were added to this list with an informal survey of confirmed Redditors

and their subreddit participation. A global list of local subreddits exists on the r/LocationReddits Wiki, of which the capital or largest city of each state in the United States was compiled into a list for our research purposes. Similarly, a list of university subreddits exists on the r/college Wiki, of which the most popular were compiled into a list. Of course, even categorizing these subreddits is not perfect, and one cannot assume that gathering data from them guarantees 100% human accounts, only that there is a higher likelihood for the subreddits to be bot-free.

We compiled approximately fifty of each type of subreddit with a full list found in Appendix A. In order to get account data from posters in these subreddits, we used PRAW to pull the authors of the top 5 posts from each subreddit over the past 24 hours. This resulted in a list of approximately 200 likely human accounts per subreddit category. A separate implementation with PRAW pulled account attributes. The attributes collected from each account which serves as the full human account dataset are as follows: Username, Account Creation Date, Post Karma, Comment Karma, Total Karma, Post Count, and Comment Count. Two other attributes were manually added to these lists for easier classification: Post Comment Ratio, which is the total number of posts divided by the total number of comments, and Bot Label, which is a boolean value assigned to accounts based on the list classification. Bot accounts were labeled 1, and human accounts were labeled 0.

It is incredibly important for bot detection research to focus not only on the bot account data collection but also on human account data collection. This is because in order to create an accurate machine learning model that can properly distinguish between bots and humans, the datasets must be as accurate as possible in establishing their ground truth. This requires ensuring that the bot account dataset contains as few human accounts as possible, and the human account dataset contains as few bot accounts as possible.

The following section provides a brief overview of the machine learning configuration used in the analysis. A full comparison of these human datasets with a supervised machine learning model is presented in Chapter 7.

6.2 Machine Learning

As discussed in Section 3.1.1, most bot detection techniques rely on machine learning models for classification. Although the biases within these models have already been discussed at length, our research employs decision trees not only for classification but to compare the effectiveness of different data collection techniques. A supervised machine learning model was chosen as the method of comparison for the following reasons. For classification purposes, a decision tree is a simple and easily interpretable model that works well on smaller datasets [65]. Due to the limited amount of data collected in the presented techniques, a decision tree is the ideal model to offer a basic, comparative baseline between the datasets.

In order to set up the supervised machine learning model, we utilize scikit-learn's **DecisionTreeClassifier**. The limitations of this model will be outlined in the next section, but in general, decision trees are a model that requires little data preparation and can be easily visualized. We propose six different decision tree models, where each model can be distinguished by the human dataset that is combined with the one bot dataset. The full bot dataset, as described in 6.1.3, is populated with 4,844 accounts from which we sample a matching number of bots for each human group. For instance, the verified humans dataset contains 86 accounts, so its corresponding model uses 86 bots and 86 humans. The combined dataset is used to train the model and test its accuracy. We utilize 80% of the total combined dataset for training the model, and 20% for testing. A detailed analysis of outcomes for these decision trees

will be discussed in Chapter 7.

To provide necessary context for the analysis in the following chapter, it is essential to first outline the limitations inherent in the data collection and machine learning methods discussed previously.

6.3 Limitations

There are many limitations when it comes to data gathering methods in the context of this research. Primarily, there still lacks an element of automation when it comes to acquiring both bot and human accounts. Compiling the list of over four thousand up-to-date bot accounts on Reddit required significant communication and collaboration with various bot-hunting experts on the platform, which is not a time-effective method for future researchers. Additionally, gathering the human account datasets was also a time-consuming task; compiling the lists of subreddits under each of the categories required considerable effort, and the number of accounts collected was limited in size. Most importantly, these methods of data collection are still not absolute. There may be cross-contamination; the list of bot accounts may still contain human accounts, and the lists of human accounts may still contain bot accounts. However, this is an ongoing challenge in bot detection research, and the goal is to minimize error whenever possible.

In addition to the limitations inherent with data collection are the recent API changes on both Twitter/X and Reddit. While the challenges that API changes have introduced have been discussed in depth throughout this thesis, it is still important to note the limitation it presents for future research. Much of the previous research in the field of bot detection relied on the use of APIs to gather data, which is now hindered completely on Twitter/X with costs, and frequent throttling on Reddit.

Lastly, when considering decision trees for use in comparing datasets, the main limitation of this approach is the risk of overfitting, particularly given the small datasets we are comparing. With limited data, decision trees can become too complex by fitting closely to the specific data points in the training set, which can reduce the model's ability to make accurate predictions when applied to new data. This can be addressed by adjusting the hyperparameters of the tree [66], and will be discussed in more detail in Chapter 7.

With the data collection methods, machine learning techniques, and their associated limitations established, the following chapter presents the analysis of these collected datasets.

Chapter 7

Analysis

This chapter evaluates the effectiveness of the data collection methods outlined previously by applying a supervised machine learning model to compare the lists of human accounts against the list of suspected bot accounts. The bot dataset is first separated by account status in order to assess its reliability for use in a machine learning model. Following this, we present a detailed comparison of the human account datasets to determine how they perform relative to one another. Lastly, it is important to discuss the limitations associated with these analyses and how they may impact the interpretation of the results.

7.1 Bot Dataset Analysis

Because only one complete dataset of bot accounts was acquired and tested against the various human account datasets, it is imperative that we analyze this dataset to ensure it is as accurate as possible. Over the course of sixty days, we looked at the same dataset of bot accounts in order to more accurately categorize the accounts into one of four areas: active, shadowbanned, suspended, or deleted.

The list of 4,844 bot accounts was run through a PRAW script to pull the various account attributes; however, accounts from this list are often suspended or shadowbanned by Reddit. It is unknown when the full list was compiled by the bot hunters, however, we obtained the full list on February 10th, 2025. The list of accounts was run through the PRAW script within a week. After attempting to use PRAW to collect all the suspected bot account attributes, only the information for 3,048 of the accounts was still active and available to pull with the Reddit API. By April 16th, 2025, the number of active accounts went down to 1,749. This means that since the list was compiled, over 60% of the accounts have been suspended, deleted, or shadowbanned.

The distinction between these three is important to note. If an account has been deleted, the user themselves must initiate that, and there is no administrative intervention. Site-wide suspensions and shadowbans, on the other hand, are given out only by Reddit admins. A suspension is a site-wide ban from the entirety of Reddit. Suspensions can be appealed, but a user will usually have to create an entirely new account once theirs is suspended. This may also be referred to as a "permaban" and is different from a shadowban. As described in Section 1.3.2, a shadowban renders an account invisible to all other users. This means that the account can still post, comment, upvote, or otherwise participate, but none of these actions will actually be applied. The goal is that the user does not know they are shadowbanned and will therefore not create another account. u/llamageddon01 created a helpful post describing the difference between suspensions and shadowbans and notes that shadowbans are typically used for bots and spammers [67]. Suspensions are given out to rule-breaking accounts, but if admins presume a user will continue to make accounts in order to break rules, a shadowban is a better alternative. Reddit as a company, however, is not entirely open about its use of shadowbans. When searching the Reddit

Help page, there are two posts that reference the term, only one of which names it directly. The "Reddiquette" page, that is, a list of values that redditors should abide by, states the following in regard to new submissions on the site:

"[Please don't] flood Reddit with a lot of stories in a short span of time. By doing this you flood the new queue. Be warned, your future submissions may be automatically blocked by the spam filter. Shadow banning (you can see your posts and votes, but no one else can) can, and will, take place in more severe cases" [68].

The only other reference to shadowbanning can be found on a help page related to a user's account status, which states "If your posts, comments, messages, and profile page aren't showing up as expected, your account might've been flagged for spam or inauthentic activity" [69]. These two articles from Reddit itself show that shadowbanning is a severe punishment for spamming or inauthentic activity, which is consistent with typical bot behavior. With all of this in mind, we can further analyze the dataset of 4,844 bot accounts and determine the percentage that are active, deleted, suspended, or shadowbanned.

Currently, there are two ways to tell if an account has been fully suspended or if it is only shadowbanned: the error code response after querying the account and the account's profile page.

7.1.1 Error Code Response

One way to determine if an account is deleted, suspended, or shadowbanned is by analyzing the error code response when attempting to query account information with PRAW. PRAW's `Redditor` class has an `is_suspended` attribute that will only return if the queried account is suspended. To determine whether an account is still

active, the account will not have the **is_suspended** attribute and PRAW will not return any errors. Moreover, accounts that are deleted or shadowbanned do *not* have the **is_suspended** attribute and will return a 404 error.

In order to distinguish between accounts that are deleted versus shadowbanned, we can simply check if the username is available. If a user deletes their account, their username becomes available. But, if an account is shadowbanned, the username will *not* be available, as the account technically still exists and can participate on the platform, albeit invisibly to other users.

To summarize, we are able to distinguish between accounts that are still active versus ones that have been deleted, suspended, or shadowbanned, as shown in Figure 7.1. Active accounts will not return any errors. Suspended accounts will have an **is_suspended** attribute. Deleted accounts will return a 404 error but their usernames are available for use. Shadowbanned accounts will return a 404 error but their usernames are *not* available for use.

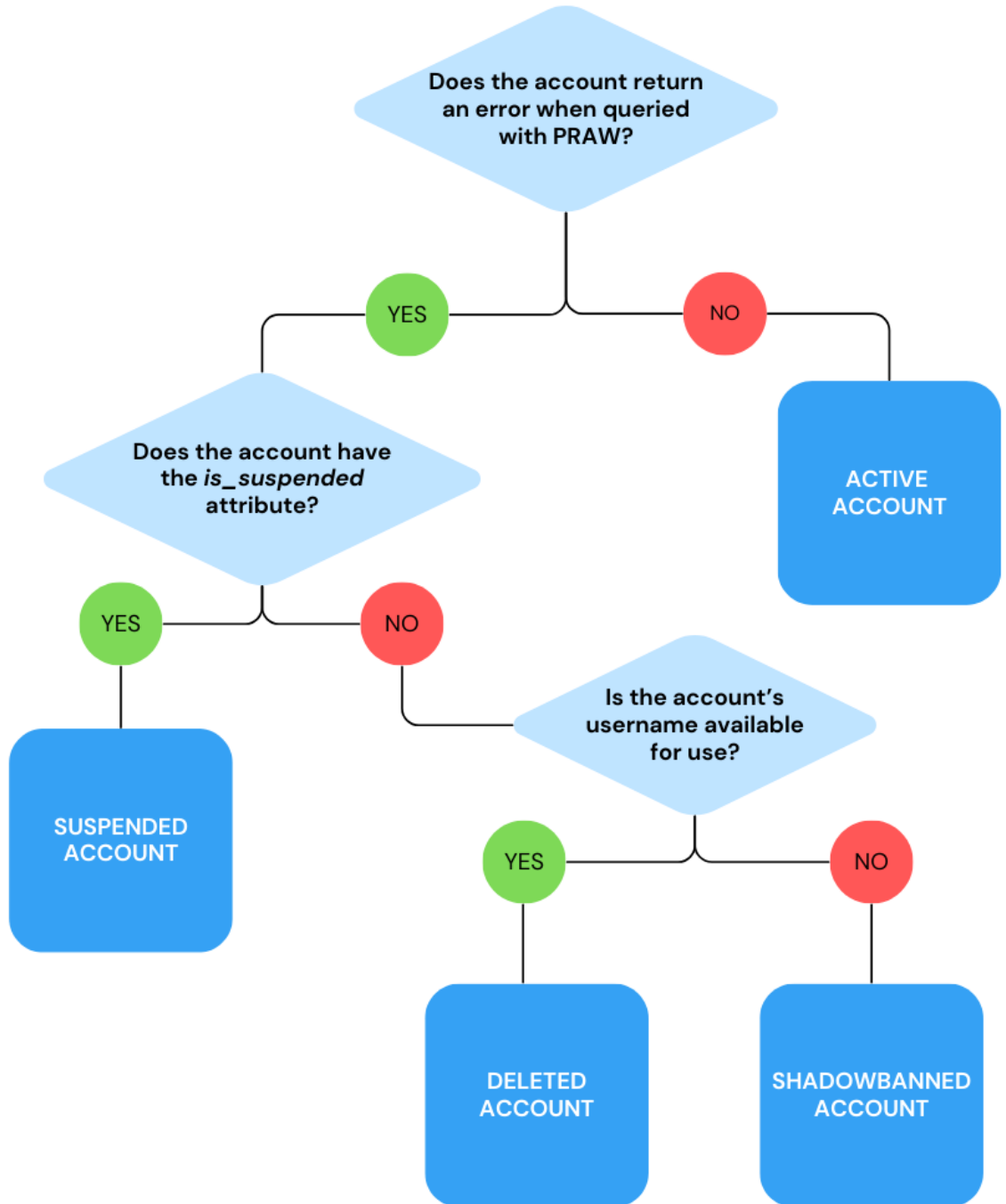


Figure 7.1: A flow chart to visualize the distinctions between active and non-active accounts on Reddit.

7.1.2 Account Profile Page

Another way to determine if an account is suspended or shadowbanned is by visually examining the account's profile page. This is only possible with the old Reddit format, as the new Reddit format shows the same message on an account's profile page for either status.

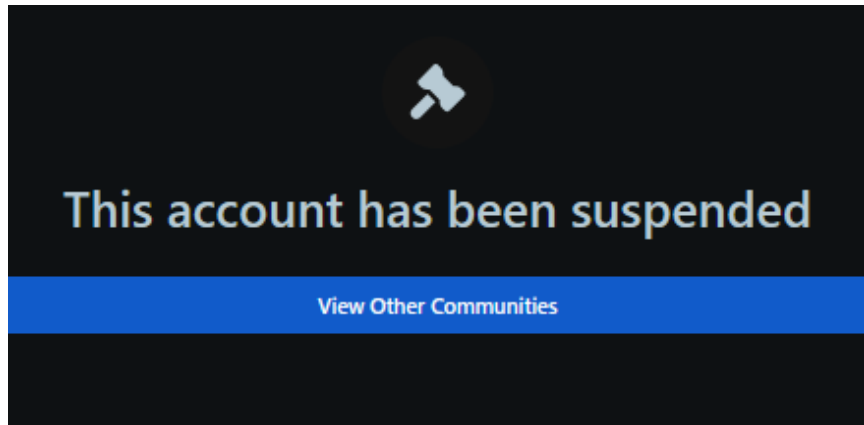


Figure 7.2: An example of the new Reddit account profile page for either a suspended or shadowbanned account.

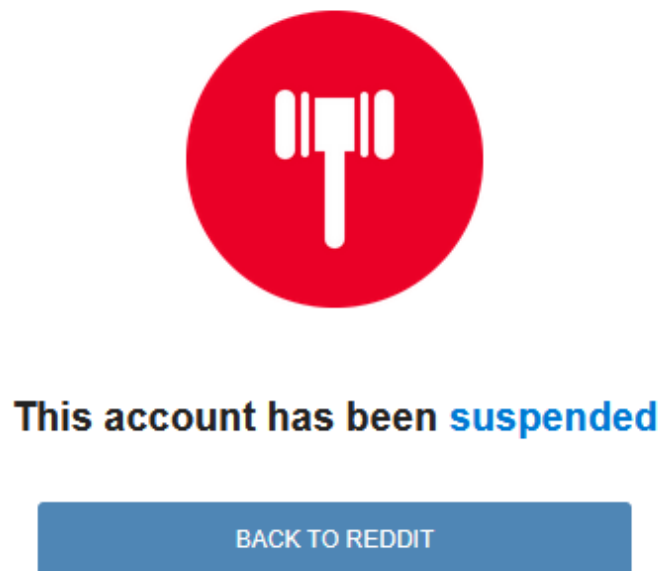


Figure 7.3: An example of a suspended account on the old Reddit account profile page.



Figure 7.4: An example of a shadowbanned account on the old Reddit account profile page.

As shown in Figures 7.2, 7.3, and 7.4, the old Reddit format used to distinguish between suspended and shadowbanned accounts, but this is no longer the case with the new Reddit format.

7.1.3 Bot Dataset Statistics

Using the above information, we were able to ascertain how many accounts in the full list of 4,844 suspected bot accounts are currently active from those that are non-active (deleted, suspended, or shadowbanned). For the initial data collection, there was no distinction between account statuses, only how many were active versus how many were not. For each attempt afterwards, we are able to see how the number of accounts which are suspended or shadowbanned changes over time.

Table 7.1: Active and Non-Active Accounts on Day 0

Active	Non-Active
3,048	1,796

Table 7.2: Active and Non-Active Accounts on Day 30

Active	Shadowbanned	Suspended	Deleted
1,850	2,592	402	0

Table 7.3: Active and Non-Active Accounts on Day 60

Active	Shadowbanned	Suspended	Deleted
1,771	2,662	411	0

Due to the fact that every single account which returned an error did not have an available username, we can conclude that none of the accounts were deleted. The largest portion of this bot dataset were shadowbanned, followed by the amount that are still active, with the smallest portion having been suspended by Reddit. Almost 55% of the total dataset have been shadowbanned since it was compiled, and that percentage continuously increases. We can therefore conclude the likelihood of this dataset containing mostly bot accounts is high, as accounts continue to be shadowbanned every day. This does not, however, mean that all of the accounts are definitively bots, as human users can be suspended or shadowbanned as well. A large portion of the accounts in this dataset are still active as well. With the overview of the bot account dataset and our analysis of the accuracy, we analyzed the supervised machine learning model used to compare all of the datasets.

7.2 Decision Tree Analyses

A decision tree model is used to compare the different types of human account datasets against the bot dataset. This analysis is divided into five subsections. The first

dataset includes Reddit staff accounts and public figures, which, as previously noted, is the dataset most likely to consist entirely of human accounts. Additionally there are accounts collected from location subreddits, followed by those collected from university and hobby-related subreddits. Finally, all five human datasets are combined and evaluated together.

Decision trees are an effective model in order to compare the different datasets. In the future, larger datasets may entail using an unsupervised machine learning method, but for general testing purposes scikit-learn’s **DecisionTreeClassifier** was utilized. The main limitation with this approach is the risk of overfitting, which is addressed with hyperparameter tuning [66]. We adjust `max_depth`, `min_samples_split`, and `min_samples_leaf`, depending on the size of the dataset. Smaller datasets should have smaller values in these variables, but in general, the decision tree should not split too soon on smaller datasets, resulting in overfit branches. The larger datasets can afford smaller leaves and more splits because of increased data to support decisions.

We also analyze the confusion matrices, classification reports, and importance values of all five datasets. Firstly, the overall accuracy that is reported before each of these tables indicates the percentage of total predictions that were correct. The confusion matrices in this context provide the number of true positives (actual human, predicted human), false negatives (actual human, predicted bot), false positives (actual bot, predicted human), and true negatives (actual bot, predicted bot). A confusion matrix template can be seen below.

Table 7.4: Confusion Matrix Template [70]

	Actual: True	Actual: False
Predicted: True	True Positive (TP)	False Positive (FP)
Predicted: False	False Negative (FN)	True Negative (TN)

This is followed by the model’s classification report, that is, a summary of each

metric used for classification. The report will show the precision, recall, F1-score, and support for each class (human or bot). Each of these values can be calculated with the above confusion matrix definitions. The precision value measures the accuracy of positive predictions made by the model. A high precision value would indicate a low false positive rate. The formula used for precision can be seen below [70].

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

Recall measures how well the model identifies true positives. A high recall value would indicate a low false negative rate. The formula used for recall can be seen below [70].

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

The F1-score determines how balanced the model is with regard to its precision and recall. It is also known as the harmonic mean of precision and recall, and will always be a high value when precision and recall are also high. The F1-score is good for unbalanced data, and the formula used can be seen below [70].

$$F1score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Finally, the support value is the number of occurrences of each class in the dataset. We then evaluate the feature importance values. The feature importance values tell us how much each account attribute contributes to the model's decision-making process. It measures the relative importance of each feature when the tree splits; the more a feature is used to split the data and reduce impurity, the higher its importance score. These values are normalized to add up to 1.

The following sections will address each decision tree and its confusion matrices,

classification reports, and importance values in order to evaluate the accuracy of the collected datasets.

7.2.1 Verified Humans Model

We initiated the classifications with the smallest, most accurate dataset of 86 accounts. As discussed previously, the human account dataset containing account information for Reddit administrators and public figures is likely to have the least amount of bots because they require some level of verification that they are human.

Table 7.5: Verified Humans Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	16	1
actual bot (1)	4	13

Table 7.6: Verified Humans Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.80	0.94	0.86	17
1 (bot)	0.93	0.76	0.84	17
macro avg	0.86	0.85	0.85	34
weighted avg	0.86	0.85	0.85	34

Table 7.7: Verified Humans Model Feature Importance

Feature	Importance Value
Comment Karma	0.840371
Total Karma	0.071030
Comment Count	0.037297
Account Creation Date	0.035844
Post Count	0.015458
Post Karma	0.000000
Post Comment Ratio	0.000000

Accuracy: 0.853

After training a decision tree with an equal number of bot accounts, we obtained a total accuracy of 0.853. Of all accounts predicted as human, 80% of them were from the human account dataset, and of all human accounts, 94% were correctly predicted as such by the model. Of all accounts predicted as bots, 93% were from the bot account dataset, and of all bot accounts, 76% were correctly predicted as such by the model. The recall for bot accounts is much lower than that of human accounts. The F1-score of both is similar, which indicates good balance. Interestingly, this decision tree did not use Post Karma or Post Comment Ratio to classify accounts, so both of these features had an importance value of 0.

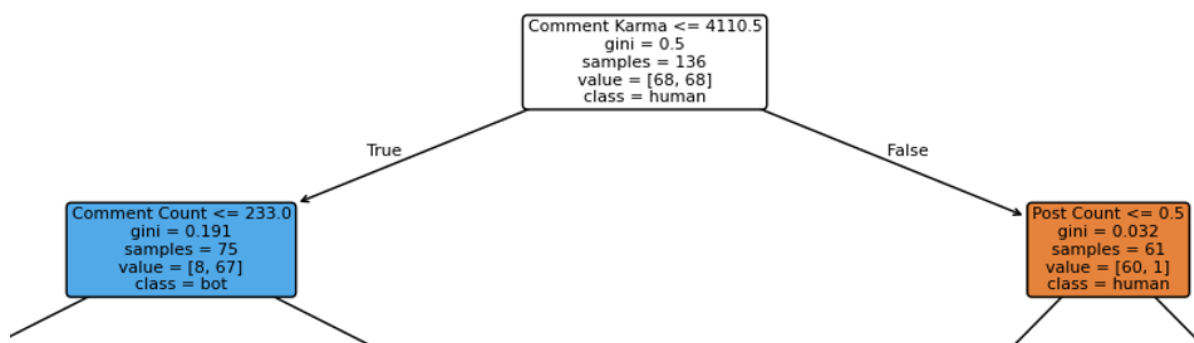


Figure 7.5: The Reddit staff and public figures decision tree initial split.

As shown in Figure 7.5, the most important feature for classification was Comment Karma, with a value of 0.84. This is misleading, however, due to the source of the public figure accounts. These accounts were collected from the r/IAmA subreddit, a community where comments are heavily favored. Because the public figures that hold these interviews are responding to questions in the comments, their comment karma will be significantly higher than a typical human user, thus impacting the importance value for that feature. This is also likely affecting the recall scores for bots and humans, as the model can easily identify humans based on their Comment Karma, since it is such a significant outlier. The full decision tree can be found in

Appendix B.

7.2.2 Location Subreddits Model

Next, we analyzed the accounts collected from location-based subreddits. This is the second smallest dataset with 186 accounts compiled.

Table 7.8: Location Subreddits Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	32	5
actual bot (1)	4	33

Table 7.9: Location Subreddits Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.89	0.86	0.88	37
1 (bot)	0.87	0.89	0.88	37
macro avg	0.88	0.88	0.88	74
weighted avg	0.88	0.88	0.88	74

Table 7.10: Location Subreddits Model Feature Importance

Feature	Importance Value
Post Karma	0.349626
Comment Karma	0.340710
Account Creation Date	0.282965
Total Karma	0.022579
Post Comment Ratio	0.004120
Post Count	0.000000
Comment Count	0.000000

Accuracy: 0.878

After training a decision tree with an equal number of bot accounts, we obtained a total accuracy of 0.878. Of all accounts predicted as human, 89% of them were from the human account dataset, and of all human accounts, 86% were correctly predicted

as such by the model. Of all accounts predicted as bots, 87% were from the bot account dataset, and of all bot accounts, 89% were correctly predicted as such by the model. Bot and human accounts had similar precision and recall for this dataset. The F1-score of both is also similar, which indicates good balance. This decision tree did not use Total Karma or Comment Karma to classify accounts, so both of these features had an importance value of 0.

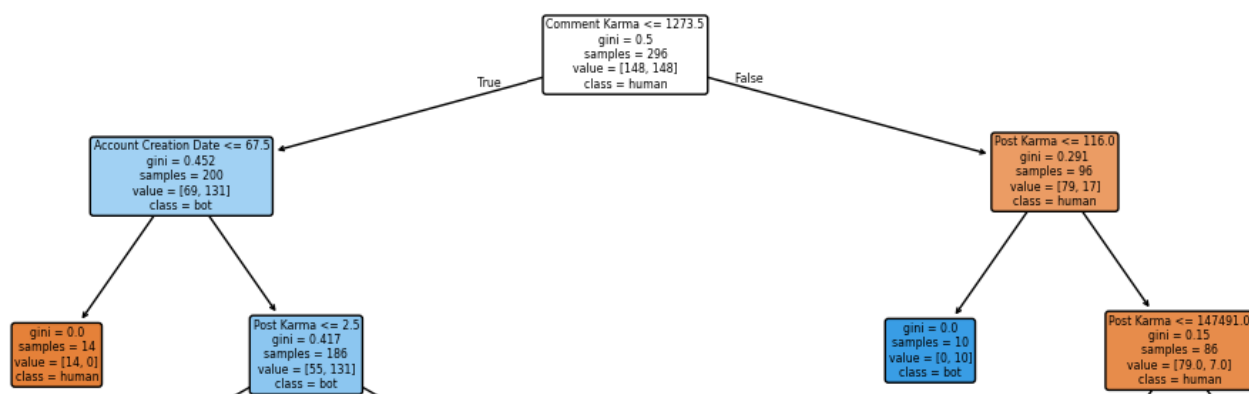


Figure 7.6: The location subreddit accounts decision tree two most important features.

As shown in Figure 7.6, the two most important features for classification were Post Karma, with a value of 0.35, followed by Comment Karma, which had a value of 0.34. The full decision tree can be found in Appendix B.

7.2.3 University Subreddits Model

Next, we analyzed the accounts collected from university subreddits. This is the second-largest dataset of the four with 219 accounts compiled.

Table 7.11: University Subreddits Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	37	7
actual bot (1)	11	33

Table 7.12: University Subreddits Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.77	0.84	0.80	44
1 (bot)	0.82	0.75	0.79	44
macro avg	0.80	0.80	0.80	88
weighted avg	0.80	0.80	0.80	88

Table 7.13: University Subreddits Model Feature Importance

Feature	Importance Value
Account Creation Date	0.396097
Post Comment Ratio	0.361372
Comment Karma	0.094543
Post Karma	0.086078
Post Count	0.039710
Comment Count	0.022201
Total Karma	0.000000

Accuracy: 0.795

After training a decision tree with an equal number of bot accounts, we obtained a total accuracy of 0.795. Of all accounts predicted as human, 77% of them were from the human account dataset, and of all human accounts, 84% were correctly predicted as such by the model. Of all accounts predicted as bots, 82% were from the bot account dataset, and of all bot accounts, 75% were correctly predicted as such by the model. This decision tree did not use the Total Karma feature for classification, so it had an importance value of 0.

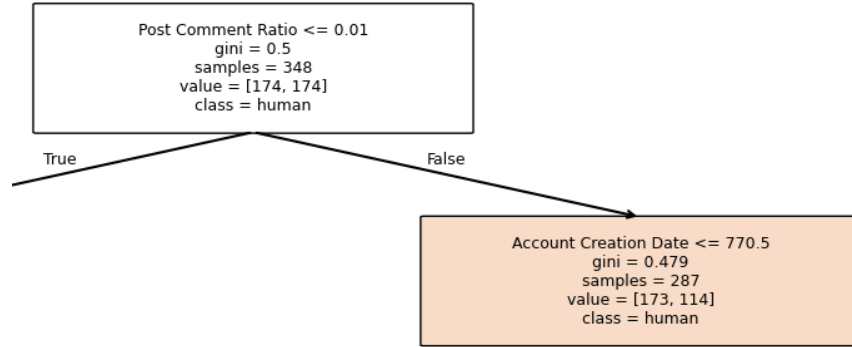


Figure 7.7: The university subreddit accounts decision tree initial split.

As shown in Figure 7.7, the two most important features for classification were Account Creation Date, with a value of 0.40, followed by Post Comment Ratio, which had a value of 0.36. The full decision tree can be found in Appendix B.

7.2.4 Hobby Subreddits Model

Next, we analyzed the accounts collected from the hobby-related subreddits. This is the largest dataset of the four with 241 accounts compiled.

Table 7.14: Hobby Subreddits Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	41	7
actual bot (1)	6	42

Table 7.15: Hobby Subreddits Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.91	0.85	0.88	48
1 (bot)	0.86	0.92	0.89	48
macro avg	0.89	0.89	0.89	96
weighted avg	0.89	0.89	0.89	96

Table 7.16: Hobby Subreddits Model Feature Importance

Feature	Importance Value
Post Karma	0.486442
Post Comment Ratio	0.217836
Account Creation Date	0.170587
Post Count	0.120470
Comment Count	0.004664
Total Karma	0.000000
Comment Karma	0.000000

Accuracy: 0.885

After training a decision tree with an equal number of bot accounts, we obtained a total accuracy of 0.885. Of all accounts predicted as human, 91% of them were from the human account dataset, and of all human accounts, 85% were correctly predicted as such by the model. Of all accounts predicted as bots, 86% were from the bot account dataset, and of all bot accounts, 92% were correctly predicted as such by the model. The F1-score of both is similar, which indicates good balance. This decision tree did not use Total Karma or Comment Karma to classify accounts, so both of these features had an importance value of 0.

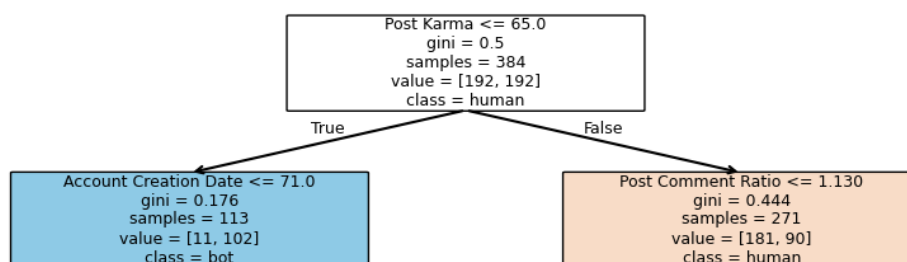


Figure 7.8: The hobby subreddit accounts decision tree initial split.

As shown in Figure 7.8, the most important feature for classification was Post Karma, with a value of 0.49. The full decision tree can be found in Appendix B.

7.2.5 Combined Humans Model

Lastly, we combine all previous datasets. This resulted in a dataset of 729 predicted human accounts.

Table 7.17: Combined Humans Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	131	15
actual bot (1)	42	104

Table 7.18: Combined Humans Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.75	0.90	0.82	146
1 (bot)	0.87	0.71	0.78	146
macro avg	0.81	0.80	0.80	292
weighted avg	0.81	0.80	0.80	292

Table 7.19: Combined Humans Model Feature Importance

Feature	Importance Value
Post Comment Ratio	0.368611
Account Creation Date	0.326625
Comment Karma	0.259414
Post Karma	0.022870
Post Count	0.009404
Total Karma	0.007466
Comment Count	0.005609

Accuracy: 0.801

After training a decision tree with an equal number of bot accounts, we obtained a total accuracy of 0.801. Of all accounts predicted as human, 75% of them were from the human account dataset, and of all human accounts, 90% were correctly predicted as such by the model. Of all accounts predicted as bots, 87% were from the bot account dataset, and of all bot accounts, 71% were correctly predicted as

such by the model. The verified humans dataset introduces error into this model, as verified humans do not have the same posting habits as other typical human users. Comment Karma is still heavily favored in this model due to the inclusion of these human accounts.

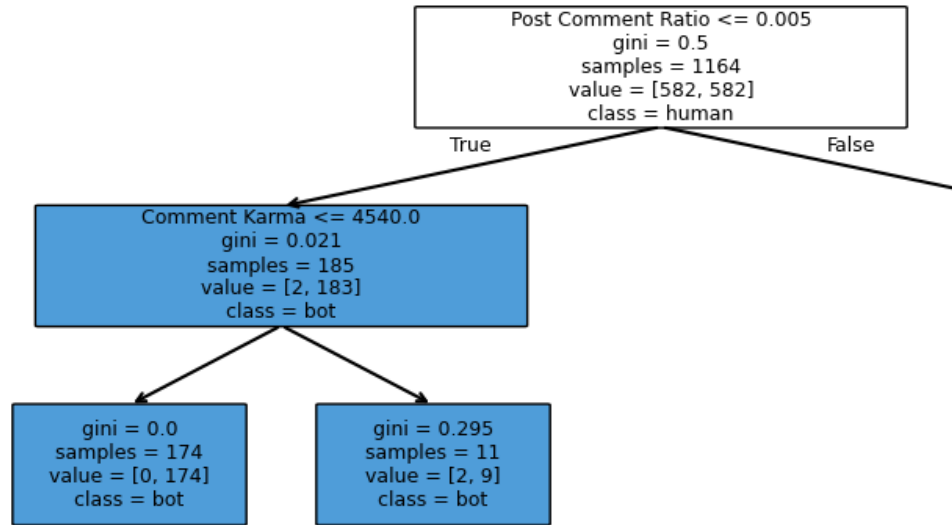


Figure 7.9: The combined human accounts decision tree initial split.

This decision tree utilized all the features provided to it for training and testing. The most important feature, Post Comment Ratio, is shown in 7.9. The full decision tree can be found in Appendix B.

7.2.6 Refined Combined Humans Model

In order to address the outliers presented by the verified humans model, we combined each human dataset except the one containing verified humans. We trained a model to check the accuracy without this dataset included.

Table 7.20: Refined Combined Humans Model Confusion Matrix

	predicted human (0)	predicted bot (1)
actual human (0)	110	19
actual bot (1)	37	92

Table 7.21: Refined Combined Humans Model Classification Report

	precision	recall	f1-score	support
0 (human)	0.74	0.85	0.79	129
1 (bot)	0.83	0.71	0.76	129
macro avg	0.79	0.78	0.78	258
weighted avg	0.79	0.78	0.78	258

Table 7.22: Refined Combined Humans Model Feature Importance

Feature	Importance Value
Account Creation Date	0.398579
Post Karma	0.396057
Comment Karma	0.099881
Post Comment Ratio	0.057328
Comment Count	0.026131
Total Karma	0.022024
Post Count	0.000000

Accuracy: 0.779

After training a decision tree with an equal number of bots, we obtained a total accuracy of 0.779. This is the lowest accuracy of all six models. The recall for humans is higher than the recall for bots, with 85% of human accounts correctly predicted as humans and only 71% of bot accounts correctly predicted as bots. This highlights the fact that without significant outliers in human accounts, like Comment Karma, bots blend in very well and it is difficult to distinguish between the accounts just based on account attributes. This model did, however, remove Comment Karma as a significant factor in the model's importance, such that Post Karma and Account Creation Date became the most important features that impacted the model's decision-making process. The full decision tree can be found in Appendix B.

7.2.7 Decision Tree Comparison

The model that produced the most accurate overall results used the human accounts collected from hobby-based subreddits. This is different than the initial assumption that the dataset collected from Reddit staff and public figures would be the most accurate, although the accuracy value of that one as well as the location subreddit dataset are all extremely close, ranging from 0.853 to 0.885. The lower accuracy may also be a result of using such a small dataset. The least accurate models were the university subreddit account dataset and the combined datasets, likely due to conflicting feature patterns among the groups. For example, Reddit staff may have fundamentally different posting behaviors or karma scores than hobbyists.

When considering precision and recall, we find that the precision for predicting bots was often higher than the recall. This means that when the model labels an account as a bot, it is typically correct, but it often fails to identify all bots present. This suggests that false negatives are a consistent issue.

The feature with the highest importance across the board was Comment Karma, followed by Account Creation Date. Both Post Karma and Post Comment Ratio also had high importance. The least important features were Total Karma and Comment Count. Post Karma and Account Creation Date were valuable features across multiple datasets.

The results of the decision trees indicate that these data collection methods have a relatively high accuracy and are a useful way to compile datasets for the purposes of bot detection on Reddit. However, there are limitations to this analysis that must be addressed.

7.3 Limitations

The limitations of our analysis will be addressed in this section, including those of the supervised machine learning model used as well as the potential limitations of the data utilized.

As discussed previously, decision trees present a few challenges, the main one being overfitting. Before adding and adjusting the hyperparameters that limit the trees, such as `max_depth`, `min_samples_split`, and `min_samples_leaf`, accuracy results were extremely high due to overfitting. After adjusting these values, we were able to obtain more accurate results, allowing for better comparison of the trees. In addition, decision trees are not complex classification methods and are not ideal for nuanced conclusions. Comparing a decision tree model to other machine learning models would likely provide a more meaningful evaluation.

The comparison of these datasets results in the same limitation that many bot detection researchers outline: it is unlikely to absolutely determine if an account is a bot or a human. Therefore, if there are any human accounts in the bot datasets or bot accounts in the human datasets, this will introduce error into the models and the accuracy will suffer. Another critical limitation is the size of the datasets. Ideally, in this context, a machine learning model would have thousands of accounts on which to train and test. However, due to the manual extraction of some accounts and API limits, it is difficult to obtain datasets large enough for an ideal analysis.

Ultimately, while these challenges highlight areas for improvement in both the dataset collection and the comparison with decision trees, they also highlight the need for continued research in this space. The next chapter will summarize the key findings of our research, reflect on the implications of the results, and propose areas for future work in Reddit-specific bot detection research.

Chapter 8

Conclusion and Future Work

In order to address the evolving bot landscape on social media platforms, proper data collection techniques are necessary to ensure the authenticity of these online spaces. Bot detection researchers struggle to obtain large and meaningful datasets of bot and human accounts, and this research is especially limited on Reddit. This thesis has addressed these challenges by providing multiple techniques for compiling bot and human accounts on the social media platform Reddit to better establish a ground truth. These datasets were used to train and test a supervised machine learning model in order to evaluate how the different methods compare. Overall, the most accurate datasets contained human accounts collected from hobby-based subreddits, location-based subreddits, and the manually collected dataset of Reddit staff and public figures. The features that were most important in this analysis were Comment Karma, Account Creation Date, Post Karma, and Post-to-Comment ratio. The data collection techniques outlined in this thesis allow for researchers, developers, and general users to more accurately identify bots on Reddit. We also outline bot-hunting communities on Reddit so that researchers can seek them out in the future. Our research further helps compile accounts that are more likely to be human based

on subreddits that have high human interaction. However, several limitations exist and are important to address when considering future research in bot detection. The size and accuracy of the datasets, the time effectiveness of collaborating with bot-hunting communities on Reddit, and the inherent issues with decision trees as a model all present challenges.

The limitations of this thesis present opportunities for future research. The outlined features of bot accounts can be expanded, resulting in larger and potentially more accurate and meaningful datasets. In addition, the comparison process of human and bot accounts could be improved by testing these methods on various machine learning models. Automating the data collection process using the described methods would also assist researchers and developers to more easily identify bots and investigate applying these methods to other social media platforms. As bots continue to evolve and become more difficult to detect, data collection techniques that can evolve with bots are essential. Furthermore, the use of artificial intelligence in both bot creation and detection presents more implications for the future of bot detection research; while AI might make bot accounts easier than ever to implement, it can, in turn, be used to detect them. Ultimately, this research provides a foundation for future work in improving bot detection accuracy and maintaining the integrity of online communication across social media platforms.

Bibliography

- [1] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “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*, ser. WWW ’17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 963–972. [Online]. Available: <https://doi.org/10.1145/3041021.3055135>
- [2] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, “Detection of bots in social media: A systematic review,” *Information Processing & Management*, vol. 57, no. 4, p. 102250, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306457319313937>
- [3] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Commun. ACM*, vol. 59, no. 7, p. 96–104, Jun. 2016. [Online]. Available: <https://doi.org/10.1145/2818717>
- [4] E. Ferrara, “What types of covid-19 conspiracies are populated by twitter bots?” *First Monday*, May 2020. [Online]. Available: <http://dx.doi.org/10.5210/fm.v25i6.10633>
- [5] M. Himelein-Wachowiak, S. Giorgi, A. Devoto, M. Rahman, L. Ungar, H. A. Schwartz, D. H. Epstein, L. Leggio, and B. Curtis, “Bots and misinformation

- spread on social media: implications for covid-19,” *Journal of medical Internet research*, vol. 23, no. 5, p. e26933, 2021.
- [6] S. Mohammad, M. U. Khan, M. Ali, L. Liu, M. Shardlow, and R. Nawaz, “Bot detection using a single post on social media,” in *2019 Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*, 2019, pp. 215–220.
- [7] E. Ferrara, “Measuring social spam and the effect of bots on information diffusion in social media,” *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pp. 229–255, 2018.
- [8] J. Prier, “Commanding the trend: Social media as information warfare,” *Strategic Studies Quarterly*, vol. 11, no. 4, pp. 50–85, 2017. [Online]. Available: <http://www.jstor.org/stable/26271634>
- [9] A. Badawy, E. Ferrara, and K. Lerman, “Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign,” in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 258–265.
- [10] Twitter, “Update on twitter’s review of the 2016 us election,” 2018. [Online]. Available: https://blog.x.com/en_us/topics/company/2018/2016-election-update
- [11] Y. Walter, “Artificial influencers and the dead internet theory,” *AI & SOCIETY*, pp. 1–2, 2024.
- [12] D. D. Placido, “The dead internet theory, explained,” Jul 2024. [Online]. Available: <https://www.forbes.com/sites/danidiplacido/2024/01/16/the-dead-internet-theory-explained/>

- [13] P. Muzumdar, S. Cheemalapati, S. R. RamiReddy, K. Singh, G. Kurian, and A. Muley, “The dead internet theory: A survey on artificial interactions and the future of social media,” *Asian Journal of Research in Computer Science*, vol. 18, no. 1, p. 67–73, Jan. 2025. [Online]. Available: <http://dx.doi.org/10.9734/ajrcos/2025/v18i1549>
- [14] J. C. Wright, *Stakeholder management in change initiatives: Reddit changes its API pricing*. SAGE Publications: SAGE Business Cases Originals, 2024.
- [15] E. Roth, “A developer says reddit could charge him \$20 million a year to keep his app working,” May 2023. [Online]. Available: <https://www.theverge.com/2023/5/31/23743993/reddit-apollo-client-api-cost>
- [16] C. Selig, 2023. [Online]. Available: https://www.reddit.com/r/apolloapp/comments/13ws4w3/had_a_call_with_reddit_to_discuss_pricing_bad/
- [17] J. Peters, “r/blind mods are very unhappy with the state of reddit’s mobile mod tools right now.” Jul. 2023. [Online]. Available: <https://www.theverge.com/2023/7/1/23781394/r-blind-mods-are-very-unhappy-with-the-state-of-reddits-mobile-mod-tools-right-now>
- [18] T. Magelinski, D. Beskow, and K. M. Carley, “Graph-hist: Graph classification from latent feature histograms with application to bot detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5134–5141, Apr. 2020. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/5956>
- [19] u/lift_ticket83, “Reddit data api update: Changes to pushshift access,” 2023. [Online]. Available: https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/

- [20] D. Javed, N. Jhanjhi, N. A. Khan, S. K. Ray, A. A. Mazroa, F. Ashfaq, and S. R. Das, "Towards the future of bot detection: A comprehensive taxonomical review and challenges on twitter/x," *Computer Networks*, vol. 254, p. 110808, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389128624006406>
- [21] A. Hankins, T. Das, S. Sengupta, and D. Feil-Seifer, "Eyes on the road: A survey on cyber attacks and defense solutions for vehicular ad-hoc networks," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023, pp. 0585–0592.
- [22] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The darpa twitter bot challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.
- [23] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on twitter," *Computers & Security*, vol. 91, p. 101715, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404820300031>
- [24] Reddit Inc., "Reddit inc. transparency page." [Online]. Available: <https://redditinc.com/transparency>
- [25] —, "Form 10-k," 2024, accessed 2025-02-18. [Online]. Available: <https://www.sec.gov/ix?doc=/Archives/edgar/data/1713445/000171344525000018/rddt-20241231.htm>
- [26] B. Norlander, 2018. [Online]. Available: https://briannorlander.com/img/Reddit_Bot_Classifier.pdf

- [27] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, “Botornot: A system to evaluate social bots,” in *Proceedings of the 25th International Conference Companion on World Wide Web*, ser. WWW ’16 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, p. 273–274. [Online]. Available: <https://doi.org/10.1145/2872518.2889302>
- [28] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao, “Social turing tests: Crowdsourcing sybil detection,” 2012. [Online]. Available: <https://arxiv.org/abs/1205.3856>
- [29] Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, K. Beznosov, and H. Halawa, “Íntegro: Leveraging victim prediction for robust fake account detection in large scale osns,” *Computers & Security*, vol. 61, pp. 142–168, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404816300633>
- [30] S. Hurtado, P. Ray, and R. Marculescu, “Bot detection in reddit political discussion,” in *Proceedings of the Fourth International Workshop on Social Sensing*, ser. SocialSense’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 30–35. [Online]. Available: <https://doi.org/10.1145/3313294.3313386>
- [31] J. Echeverria and S. Zhou, “Discovery, retrieval, and analysis of the ’star wars’ botnet in twitter,” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ser. ASONAM ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1–8. [Online]. Available: <https://doi.org/10.1145/3110025.3110074>

- [32] K. Lee, B. Eoff, and J. Caverlee, “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*, vol. 5, no. 1, 2011, pp. 185–192.
- [33] L. H. X. Ng and K. M. Carley, “Botbuster: Multi-platform bot detection using a mixture of experts,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, p. 686–697, Jun. 2023. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/22179>
- [34] M. H. Saeed, S. Ali, J. Blackburn, E. D. Cristofaro, S. Zannettou, and G. Stringhini, “Trollmagnifier: Detecting state-sponsored troll accounts on reddit,” in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 2161–2175.
- [35] K. Bell, “Twitter shut off its free api and it’s breaking a lot of apps,” Apr. 2023. [Online]. Available: <https://www.engadget.com/twitter-shut-off-its-free-api-and-its-breaking-a-lot-of-apps-222011637.html>
- [36] A. Minnich, N. Chavoshi, D. Koutra, and A. Mueen, “Botwalk: Efficient adaptive exploration of twitter bot networks,” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ser. ASONAM ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 467–474. [Online]. Available: <https://doi.org/10.1145/3110025.3110163>
- [37] N. Chavoshi, H. Hamooni, and A. Mueen, “ DeBot: Twitter Bot Detection via Warped Correlation ,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2016, pp. 817–822. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/ICDM.2016.0096>

- [38] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, “Rtbust: Exploiting temporal patterns for botnet detection on twitter,” in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 183–192. [Online]. Available: <https://doi.org/10.1145/3292522.3326015>
- [39] N. Jess and H. G. Bayhan, “Visualizing the evolution of twitter (x. com) conversations: A comprehensive methodology applied to ai training discussions on chatgpt,” *arXiv preprint arXiv:2407.03484*, 2024.
- [40] S. A. Alipour, R. Orji, and N. Zincir-Heywood, “Behaviour and bot analysis on online social networks: twitter, parler, and reddit,” *International Journal of Technology and Human Interaction (IJTHI)*, vol. 19, no. 1, pp. 1–19, 2023.
- [41] M. M. Chiu, C. H. Park, H. Lee, Y. W. Oh, and J.-N. Kim, “Election fraud and misinformation on twitter: Author, cluster, and message antecedents,” *Media and Communication*, vol. 10, no. 2, pp. 66–80, 2022.
- [42] A. Dhiman and D. Toshniwal, “An unsupervised misinformation detection framework to analyze the users using covid-19 twitter data,” in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 679–688.
- [43] E. Alothali, N. Zaki, E. A. Mohamed, and H. Alashwal, “Detecting social bots on twitter: A literature review,” in *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 175–180.
- [44] A. Obadimu, E. Mead, S. Al-Khateeb, and N. Agarwal, “A comparative analysis of facebook and twitter bots,” 2019.

- [45] N. Agarwal, S. Al-Khateeb, R. Galeano, and R. Goolsby, “Examining the use of botnets and their evolution in propaganda dissemination,” *Defence Strategic Communications*, vol. 2, no. 1, pp. 87–112, 2017.
- [46] M. Ikram, L. Onwuzurike, S. Farooqi, E. D. Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, “Measuring, characterizing, and detecting facebook like farms,” *ACM Transactions on Privacy and Security (TOPS)*, vol. 20, no. 4, pp. 1–28, 2017.
- [47] Ü. Tunç, E. Atalar, M. S. Gargı, and Z. E. Aydın, “Classification of fake, bot, and real accounts on instagram using machine learning,” *Politeknik Dergisi*, vol. 27, no. 2, pp. 479–488, 2022.
- [48] [Online]. Available: <https://www.parler.com/>
- [49] Trump Media & Technology Group, “Truth Social,” <https://truthsocial.com/>.
- [50] B. Galicia, “A laboratory of revolution: Unravelling the deep story of truth social and its offline security implications,” 2024.
- [51] B. Kolk, “An incoherent truth: Truth social and democracy in our populist age,” 2024. [Online]. Available: <http://lup.lub.lu.se/student-papers/record/9151533>
- [52] u/blackfeathr, “How to identify bots on reddit,” 2023. [Online]. Available: https://www.reddit.com/r/LearnUselessTalents/comments/15tzjkb/how_to_identify_bots_on_reddit/
- [53] Reddit Inc., “What constitutes vote cheating or vote manipulation?” [Online]. Available: <https://support.reddithelp.com/hc/en-us/articles/360043066412-What-constitutes-vote-cheating-or-vote-manipulation>

- [54] —, “What constitutes spam? am i a spammer?” [Online]. Available: <https://support.reddithelp.com/hc/en-us/articles/360043504051-What-constitutes-spam-Am-I-a-spammer>
- [55] u/tyrannosnorlax, “Bots. how to identify them, and why do they exist on reddit?” 2022. [Online]. Available: https://www.reddit.com/user/tyrannosnorlax/comments/t0h466/bots_how_to_identify_them_and_why_do_they_exist/
- [56] u/WildFlemima, “General bot information megathread: How do bots act? what kinds of things do they post? how do we identify them, and why do they exist?” 2024. [Online]. Available: https://www.reddit.com/r/RedditBotHunters/comments/1f2xkii/general_bot_information_megathread_how_do_bots/
- [57] u/bot-sleuth-bot. [Online]. Available: <https://www.reddit.com/user/bot-sleuth-bot/>
- [58] u/HelpfulJanitor. [Online]. Available: <https://www.reddit.com/user/HelpfulJanitor/>
- [59] u/bot bouncer. [Online]. Available: <https://www.reddit.com/r/BotBouncer/wiki/index/>
- [60] “r/redditbothunters.” [Online]. Available: <https://www.reddit.com/r/RedditBotHunters/>
- [61] “r/thesefuckingaccounts.” [Online]. Available: <https://www.reddit.com/r/TheseFuckingAccounts/>
- [62] “r/botwatch.” [Online]. Available: <https://www.reddit.com/r/botwatch/>

- [63] u/spez, “Reddit’s 2017 transparency report and suspect account findings,” 2018. [Online]. Available: https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/
- [64] ——. [Online]. Available: <https://www.reddit.com/user/spez/>
- [65] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random forests and decision trees,” *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.
- [66] M. Bramer, “Avoiding overfitting of decision trees,” *Principles of data mining*, pp. 119–134, 2007.
- [67] u/llamagddon01, “What’s that wednesday? - bans, shadowbans and suspensions,” 2023. [Online]. Available: https://www.reddit.com/r/NewToReddit/comments/136hno1/whats_that_wednesday_bans_shadowbans_and/
- [68] Reddit Inc., “Reddiquette.” [Online]. Available: <https://support.reddithelp.com/hc/en-us/articles/205926439-Reddiquette>
- [69] —, “My account was flagged for spam or inauthentic activity.” [Online]. Available: <https://support.reddithelp.com/hc/en-us/articles/360045309012-My-account-was-flagged-for-spam-or-inauthentic-activity>
- [70] A. Burkov, *The Hundred-Page Machine Learning Book*. Polen: Andriy Burkov, Jan. 2019.

Appendix A

List of Human Subreddits

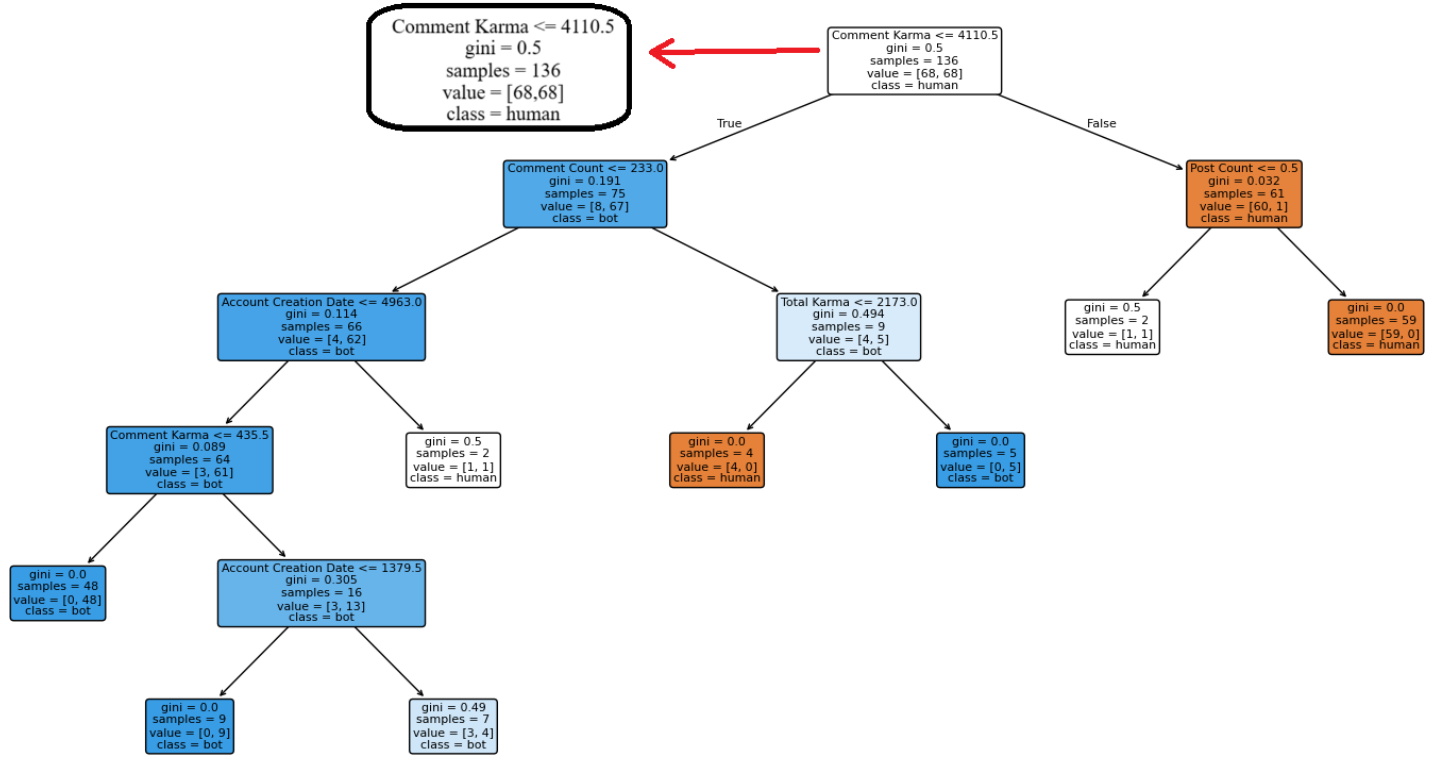
The following subreddits were included in the human account data collection for this thesis, separated by categories:

Appendix B

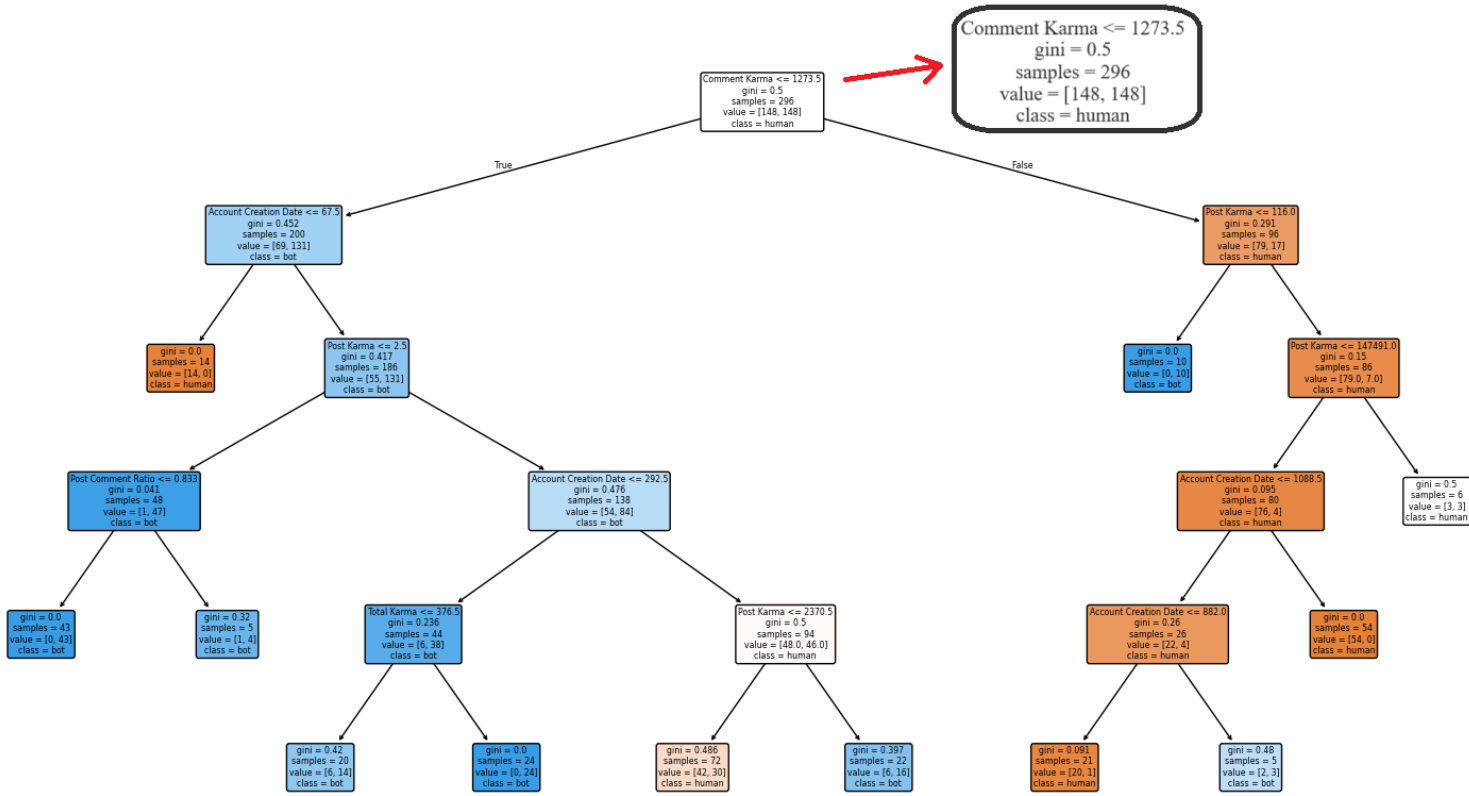
Decision Trees

The following decision tree visualizations are provided in this appendix.

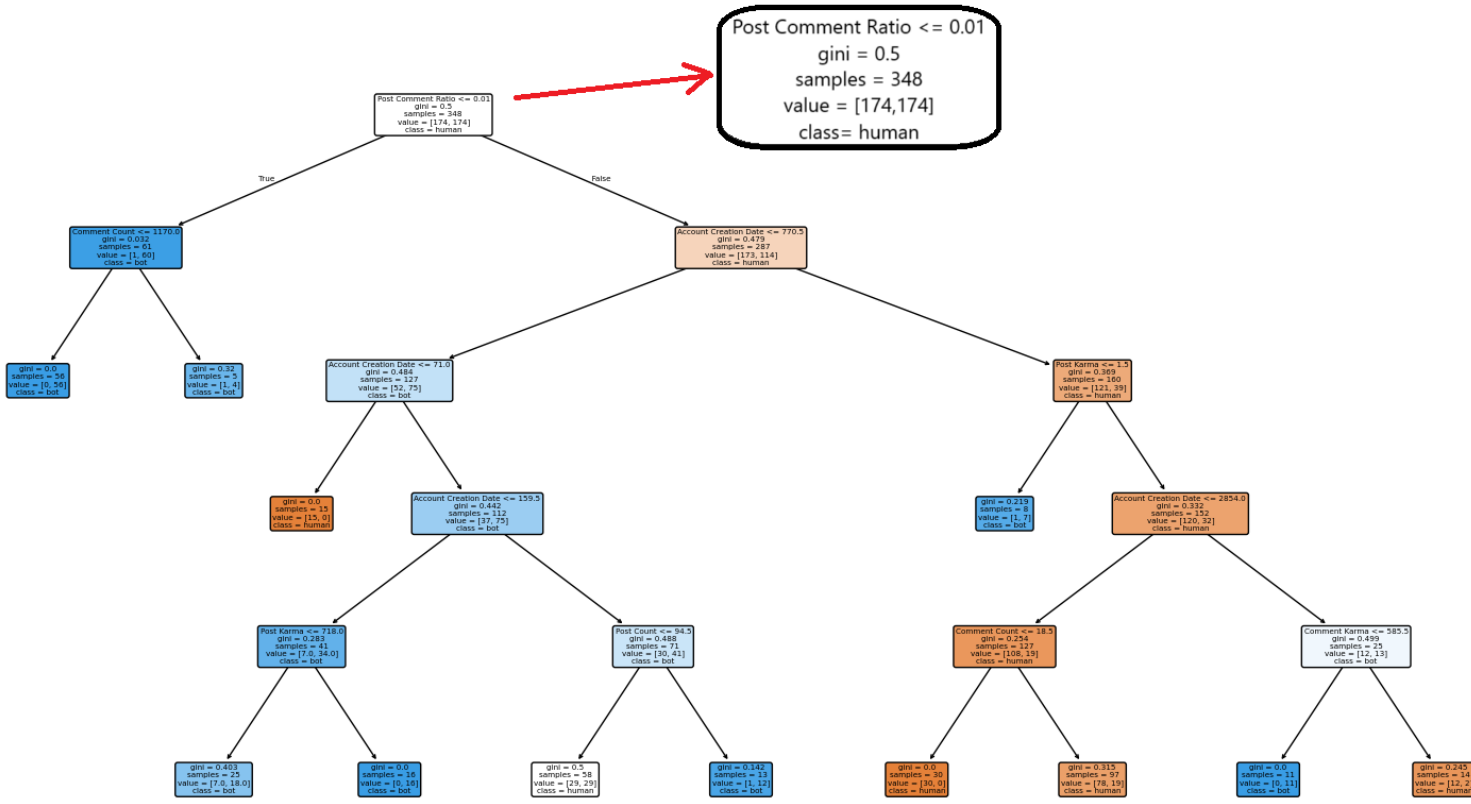
B.1 Verified Humans Model



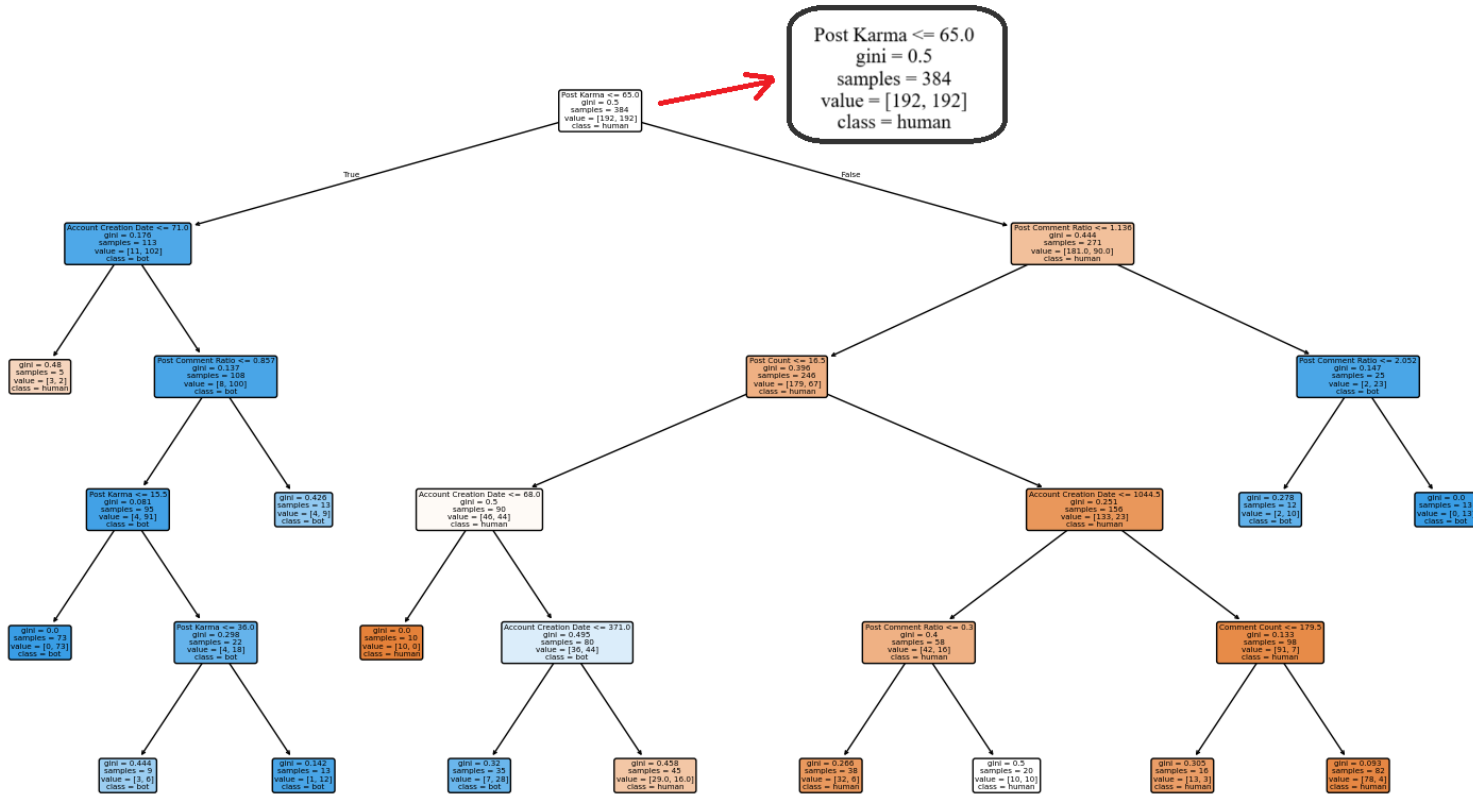
B.2 Location Subreddits Model



B.3 University Subreddits Model



B.4 Hobby Subreddits Model



B.6 Refined Combined Humans Model

