A detection model to sense Human vs Social bots on Twitter

Kumar Saurabh

M.Sc. in Computing
in Big Data Analytics
and Artificial
Intelligence

2021

lyit | Institiúid Teicneolaíochta Leitir Ceanainn | Letterkenny Institute of Technology

Computing Department, Letterkenny Institute of Technology, Port Road, Letterkenny, Co. Donegal, Ireland.

A detection model to sense Human vs Social bots on Twitter

Author: Kumar Saurabh

Supervised by: Dr Gary Cullen

A thesis submitted in partial fulfilment of the requirements for the

Master of Science in Computing in Big Data Analytics and Artificial Intelligence

Submitted to Quality and Qualifications Ireland (QQI)
Dearbhú Cáilíochta agus Cáilíochtaí Éireann

September 2021

Declaration

I hereby certify that the material, which l now submit for assessment on the programmes of study leading to the award of Master of Science in Computing in Big Data Analytics and Artificial Intelligence, is entirely my own work and has not been taken from the work of others except to the extent that such work has been cited and acknowledged within the text of my own work. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution. I understand that it is my responsibility to ensure that I have adhered to LYIT's rules and regulations.

I hereby certify that the material on which I have relied on for the purpose of my assessment is not deemed as personal data under the GDPR Regulations. Personal data is any data from living people that can be identified. Any personal data used for the purpose of my assessment has been pseudonymised and the data set and identifiers are not held by LYIT. Alternatively, personal data has been anonymised in line with the Data Protection Commissioners Guidelines on Anonymisation.

I consent that my work will be held for the purposes of education assistance to future students and will be shared on the LYIT Computing website (www.lyitcomputing.com) and Research THEA website (https://research.thea.ie/). I understand that documents once uploaded onto the website can be viewed throughout the world and not just in the Ireland. Consent can be withdrawn for the publishing of material online by emailing Thomas Dowling; Head of Department at thomas.dowling@lyit.ie to remove items from the LYIT Computing website and by email emailing Denise McCaul; Systems Librarian at denise.mccaul@lyit.ie to remove items from the Research THEA website. Material will continue to appear in printed formats once published and as websites are public medium, LYIT cannot guarantee that the material has not been saved or downloaded.

Signature of Candidate: Kumar Saurabh Date: 31/08/2021

Acknowledgements

I would like to thank the following people, without whom I would not have been able to complete this dissertation.

First of all, I would like to thank **Dr Kevin Meehan** who helped me for choosing this dissertation topic. Without his help I was finding it really hard to stick to any topic. Secondly, I would like to thank **Dr Shagufta Henna** and **Dr Edwina Sweeney** for being supportive throughout the taught module and bringing challenges during the Artificial Intelligence and Big Data Architecture modules which proved to be very beneficial in a way which helped in working on the dissertation.

Apart from the professors, I would like to say a special thanks to my supervisor **Dr Gary Cullen** whose insights and knowledge always steered me throughout the research and finally pushed me towards the completion of this dissertation. It was his hard work and knowledge which helped me write so many pages with ease.

Last but not the least, I would like to thank my family members in India for believing in me and being supportive throughout this corona pandemic and not to forget my seniors and friends from LYIT for being helpful and cooperative during these tough days. Their motivation helped me finish the taught module on a good note and finally completing this dissertation successfully.

Abstract

Twitter is a new web application that serves as both an online social network and a microblogging platform. Text-based posts are used by users to communicate with one another. Twitter's popularity and open structure has attracted a significant number of automated programs known as bots, which look to be a double-edged sword for the social media platform. Bots that deliver news and update feeds generate a vast number of innocuous tweets, whereas harmful bots transmit spam or malicious content. More intriguingly, a term called cyborg has evolved to describe either a bot-assisted human or a human-assisted bot in the middle of the human-bot divide. This study focuses on classification of human and bot accounts on Twitter to assist human users in identifying who they are communicating with. The research begins by collecting data from Twitter and implementing algorithms on more than 67 thousands accounts including humans and bots. In terms of tweeting behaviour, tweet content, and account attributes, there is a noticeable differences between humans and bots. This report talks about a classification system based on the bot data that contains machine learning based components. It determines whether the user is a bot or not using one of the feature taken from the combined data. The proposed classification system's efficacy is demonstrated by some experimental examinations.

Acronyms

Acronym	Definition	Page
OSN	Online Social Network	1
PC	Personal Computer	1
US	United States	1
et al.	And others	2
SMS	Short Message Service	4
DM	Direct Message	4
URL	Uniform Resource Locator	4
BC	Before Christ	5
IRC	Internet Relay Conversation	5
NLP	Natural Language Processing	8
PIN	Personal Identification Number	11
CEO	Chief Executive Officer	12
API	Application Programming Interface	12
ISP	Internet Service Provider	19
DNS	Domain Name System	19
ID	Identity/ Identification	23
ML	Machine Learning	26
CSV	Comma Separated Values	26
GB	Gigabyte	27
SSD	Solid State Drive	27
OS	Operating System	27
EMD	Entropy Minimization Discretization	27
ROC	Receiver Operating Characteristics	31
AUC	Area Under the Curve	31
XGB	Extreme Gradient Boosting	31
GUI	Graphical User Interface	31
Qt	Quasar Toolkit	31

GTK	General Image Manipulation Program Toolkit	31
TP	True Positive	32
TN	True Negative	32
FP	False Positive	32
FN	False Negative	32
NB	Naïve Bayes	39
DT	Decision Tree	39
LR	Linear Regression	39
RF	Random Forest	39
FPR	False Positive Rate	39
MSE	Mean Squared Error	41
R2	Coefficient of Determination	41
AWS	Amazon Web Services	55

Table of Contents

Table of Figures

Table of Tables

Table of Code Listings

1. Introduction

Social media is one of the important part of everyone's life. Today, people's daily life have been influenced by online social networks (OSNs) such as Facebook, Instagram, and Twitter. Indeed, they are the most popular web applications, with over 3 billion users worldwide. OSNs allow users to connect with one another to engage, consume, generate, and share content. People can easily communicate with others through social media from anywhere as long as they have cell phones, tablets, computers, and an internet connection. Some people may have multiple accounts on a single social networking platform.

Thanks to technological advancements, a social network is no longer limited to a single island, but instead encompasses the entire globe. People can now communicate with people outside of their local areas and exchange ideas on a global scale thanks to sites like Facebook and Twitter. Having such a big network, of course, has its drawbacks. Because people do not meet in person, it is difficult to tell if someone in their social network is a real person or a robot (also known as social bots) controlled by a malevolent actor who is attempting to harvest data from users through spam and fishing tactics. Organizations are undoubtedly under continual security threats, which not only cost billions of dollars in damage and recovery, but also have a negative impact on their brand. A botnet-assisted attack against these companies is a well-known danger. "Botnets cost over 9 billion in damages to US victims and over 110 billion globally", according to the US Federal Bureau of Investigation. Every year, about 500 million computers are attacked, resulting in 18 victims every second. Rustock, the most well-known assault, infected 1 million PCs and sent up to 30 billion spam emails every day. As a result, defending against botnet-assisted assaults is critical.

The goal of this study is to obtain a better knowledge of the most effective bot detecting techniques. In social networking analysis, several bot detection approaches have been deployed, and considering the prominence of social media in everyday life, this looked like an excellent subject to focus on. Many people's minds are on social media these days because it has frequently been used to spread disinformation.

1.1. Background

As the internet becomes more pervasive in everyday life, experts are increasingly interested in bot hunting. Bots can become highly prominent as a result of this greater connectedness, especially with the abundance of fake information. A botnet is a group of bots, or agents, infected hosts that are managed by botmasters via command and control (C2) channels. Botmaster, which might be deployed among multiple agents within or outside the network, is used by a hostile opponent to control the bots. Because bots are frequently used to promote the same message, one way of bot detection is to compare tweets. Time interval entropy was the content-based feature employed by (Perdana *et al.* 2015). In a nutshell, this is a technique for capturing the regulatory aspects of tweeting behaviour, which can be a strong indicator of automation. The authors utilized unigram matching to determine tweet similarity. Unigram matching tries to anticipate the next item in a sequence using probability. The authors used the number of matching words and the number of words in a tweet to produce a similarity score for each tweet. The author compared results based on time interval entropy, tweet similarity, and a combination of the two. They discovered that combining the two metrics yielded substantially more accurate results.

Although classic machine learning approaches have been employed to detect bots, academics have begun to favour a more unique method of detection known as graph analytics. In the physical, biological, and social sciences, graphs can be used to model a wide range of interactions. Because of their versatility, these graph structures are particularly suited to modelling social media networks, with the vertices representing people and the edges representing how they are connected to one another, such as whether they are

1.2. Research Question

The main objective of this paper is to try and build an effective model which identify bot accounts on twitter. This research paper will try to find the answers of the question:

- Can the determination of False Positive or True Negative ratios be improved using the regression and classifier models or Gradient Boost is required, while detecting bot accounts?

1.3. Report Outline

The remaining part of this research paper is divided in several sections which will focus on covering all aspects of bot detection techniques.

In Chapter 2, a deeper look into the concept of bots and how they are influencing the real world both positively and negatively is discussed. It will cover the techniques which has been implemented in past and which are currently being used for detection of social bots.

2. Literature Review

Twitter is a one of the most popular social networking website based in the United States that allows a person to send and receive messages which is termed as Tweets in the language of Twitter. Unauthorized or unregistered person can only see the tweets, while registered person can post, like, and retweet them. Until April 2010, the Twitter service can only be accessed using SMS while now, it can be accessed using website or mobile application. One of Twitter's most important resources is the Tweet. A Tweet, in its most basic form, can be up to 280 characters long and can be shared publicly or privately, depending on the account's settings. However, a range of items, such as media, a location, polls, and URLs, can be linked to a Tweet. Users can send private DMs as well as public messages on Twitter. Accounts can be set to protected, which limits this information (and all tweets) to approved followers.

Many accounts on social media platforms like Twitter are now managed by automated agents known as bot accounts. Typically, people use the power of these accounts to increase traffic to their websites, influence community opinion on a specific topic, recruit people to their organizations, which may or may not be illegal, manipulate people for stock market actions, propagate some fake news, and blackmail people to spread their personal information. As an outcome, a social bot detection mechanism becomes extremely important in order to protect people from malicious accounts.

Currently, research on social robot identification by domestic and international researchers suggests that studying social robot characteristics is more significant than improving classifiers. As a result, a number of academics have proposed new features for detecting and recognizing social robots. New social robots, on the other hand, can be adjusted to fit the meaning of new features, and proposing new features requires a lot of time and work. Few research explore the relationship between different aspects, and following dimension reduction, the combination of different features is frequently a simple combination.

2.1. Influence

Social media has a far longer history than we might think. Despite the fact that it appears to be a new trend, sites like Facebook are the natural result of centuries of social media evolution. The earliest method of communication was letters which was written messages delivered by hand of one person to another. The first kind of postal service dates back to 550 BC, and over the ages, this basic delivery system would grow more widespread and simplified. The telegraph was invented in 1792 (Hendrick 2013). This made it possible to send messages across great distances far faster than a horse and rider could. Though telegraph messages were brief, they were a revolutionary means of communicating news and information. The pneumatic post, invented in 1865, provided another option for letters to be transported swiftly between recipients, though it is no longer widely used outside of drive-through banking. In the last decade of the 1800s, two significant inventions were made: the telephone in 1890 and the radio in 1891 (Hendricks 2013). Both technologies are still in use today, although newer versions are far more advanced than their forerunners. People could converse instantly across enormous distances thanks to telephone lines and radio waves, something that mankind had never experienced before.

In the twentieth century, technology began to change at an incredible rate. Following the creation of the first supercomputers in the 1940s, scientists and engineers began to develop ways to connect such computers, eventually leading to the establishment of the Internet. In the 1960s, the first versions of the Internet, such as CompuServe, were created. During this time, primitive kinds of email were also developed. By the 1970s, networking technology had advanced to the point that UseNet, launched in 1979, allowed users to connect via a virtual newsletter. Home PCs or personal desktops were becoming more popular in the 1980s, and social media was getting more complex. IRCs, or Internet relay conversations, were introduced in 1988 and were popular long into the 1990s. Six Degrees, the first well-known social media site, was founded in 1997 (Hendricks 2013). It allowed users to create a profile and connect with other individuals.

Social media increased in popularity after the development of blogging. In the early 2000s, sites like MySpace and LinkedIn grew in popularity, while Photobucket and Flickr allowed

online photo sharing. In 2004, Google launched Orkut and became the leader in social networking industry for few years until it was upgraded to Google Plus and then finally it was closed. In 2005, YouTube launched, ushering in a whole new way for people to communicate and share across vast distances. Facebook and Twitter were both available to people all around the world by 2009. These sites are still among the most popular social networking sites on the web. Tumblr, Spotify, Foursquare, and Pinterest were among the first sites to emerge to fill unique social networking niches. There are a variety of social networking sites available today, and many of them may be linked to allow cross-posting. This offers an atmosphere in which users can communicate with the greatest number of individuals while maintaining the intimacy of one-on-one contact. One can only hypothesize about how social networking will evolve in the next decade, or perhaps 100 years (Hendrick 2013).

While Facebook, Instagram, Pinterest, etc remains famous among common society, Twitter is one of the social network platform which is widely famous and is being used by the corporate society. Famous celebrities, businesspersons and corporate people throughout the world used this platform to speak to the world in the form of tweets (a short message published on twitter platform).

According to Twitter's privacy policy, it collects personally identifiable information from its users and distributes it with third parties. If the firm changes hands, the provider maintains the right to sell this information as an asset. Advertisers can target users based on their history of tweets and may quote tweets in adverts addressed directly to them, despite the fact that Twitter does not display advertising.

Figure 1 A layout of Twitter homepage

With the growth of Twitter, some harmful elements were introduced named as social bots. They act like a human account but instead they were programmed in such a way that they will tweet multiple times using false words on some of the posts of any famous personality. They are basically deployed to create rumours on social media to divide public opinions.

Bots, or automated accounts on social media sites, have gotten a lot of attention lately because of their potential to have a big impact on the social spaces where they operate. While bot accounts can be used for a variety of reasons, both good and bad, much of the existing research treats bots as a behaviourally monolithic category. Studies that only distinguish between human and bot accounts back this up. While such research is useful for recognizing bots in general, more research into dynamic descriptions of bots is needed to better understand the range of actions and consequently objectives for which bots are built. To a certain purpose, it is recommended that bot accounts' text be viewed as a record of an individual's information foraging activity in a linguistic resource environment in order to generate textual content. The language that emerges from this foraging activity could be exploratory or exploitative in nature. A strategy to text written over several days by a small sample of Twitter users, including suspected bots, has been applied in this work. It is discovered that, on average, the suspected bots behave less exploratively than the

suspected human accounts in the small sample. A brief about social bots is explained in next section.

2.2. Social Bots

Bots, short for robots or software robots have been around since the birth of computers. Chatbots, which Alan Turing envisioned in the 1950s as algorithms designed to hold a conversation with a human, are one intriguing example of bots. For decades, the goal of creating a computer algorithm that passes the Turing test has fuelled artificial intelligence research, as evidenced by projects like the Loebner Prize, awarding progress in natural language processing. Many things have changed since the early days of artificial intelligence, when bots like Joseph Weizenbaum's ELIZA, which imitated a Rogerian psychiatrist, were created as demonstrations or for entertainment. A social bot is a computer algorithm that creates content and interacts with people on social media in an attempt to emulate and maybe change their behaviour. For the past few years, social media platforms have been populated by social bots.

2.2.1. Useful Bots

Bots are initially designed to help the humans to achieve their goals with ease. They are programmed to perform a specific task. Bots, when deployed correctly, have proven to be beneficial in a variety of ways. If you've ever experienced a live chat support for customer service, you've probably had the uneasy feeling that the "person" you're conversing with is actually a robot. Chatbots have grown in popularity over the years, thanks to significant advances in artificial intelligence and other underlying technologies like natural language processing. Chatbots are becoming smarter, more responsive, and more useful, and humans can see even more of them in the upcoming years. Despite the fact that chatbot technology is distinct from natural language processing technology, the former can only progress at the same rate as the latter; without further advances in NLP, chatbots will be at the mercy of algorithms' current ability to detect subtle nuances in both written and spoken dialogue.

2.2.2. Harmful Bots

Bots as discussed earlier are automated software robots which performs specific tasks based on input from a user or act according to a set pattern. Some of the bots are peaceful and, in general, harmless or even helpful: bots that automatically combine content from numerous sources, such as simple news feeds, fall into this group. For customer service, brands and businesses are increasingly using automatic responders to questions. Although these bots are intended to provide a valuable service, they can occasionally be destructive, such as when they help spread false information or rumours. Analysis of Twitter messages in the aftermath of the Boston marathon bombing demonstrated that social media can help with early detection and characterisation of emergency situations. Every new technology brings with it the potential for abuse, and social media is not left far behind. A second group of social bots consists of hostile entities created with the intent of causing harm. With rumours, spam, viruses, misinformation, defamation, or even just noise, these bots deceive, exploit, and control social media discussions. The issue isn't just determining the authenticity of the content being promoted, this was a challenge long before the development of social bots, and it's one that computational approaches can't solve. Bots generate a fresh difficulty in that they might offer the misleading appearance that some piece of information, regardless of its truth, is extremely popular and backed by a large number of people, exerting an impact against which there’s still a need to build antibodies.

2.3. Twitter Bots

A Twitter bot is a computer software that automatically posts to Twitter, tweeting, retweeting, and following other accounts. According to a recent research, 20 million accounts on Twitter were fake in 2013, accounting for less than 5% of all accounts (D’Onfro 2021). Fake accounts are frequently used to swiftly establish big following populations for advertising, while others respond to tweets that contain a specific word or phrase. Writing tweets, retweeting, and like are some of the tasks that a Twitter bot can do on their own. Twitter does not object to the use of Twitter bot accounts as long as they do not violate Twitter's Terms of Service by sending spammy automated messages or tweeting deceptive links. Twitter bots, like botnets in general, can carry out a variety of jobs, from simple tasks

like following a user to more complex tasks like having conversations with other users. Social bots are a class of bot that interacts with users and creates information to support a specific point of view. The content's authenticity seems to have no influence on the social bot's detection. Social bots are believed to account for 15 to 18 percent of all Twitter accounts currently in use. This bot detection study aims to develop upgraded algorithms for detecting social bots that actively resist being detected by standard bot identification techniques. Bots exist in a variety of shapes and sizes on Twitter. One form of bot exists just to artificially inflate an account's number of followers. The number of Twitter followers impacts the account's power since the number of followers determines how broadly the account's message is transmitted and how much weight it receives. Users are more inclined to trust a Twitter handle with 1 million followers than one with 100 followers. Increasing one's fame and attracting more human followers by using bots to artificially inflate the number of followers on an account is a strategy to do so. When Twitter bot accounts are examined, it is evident that they appear in a variety of forms, some of which are quite basic and others which are so complicated that they are difficult to detect even by humans. They imitate human accounts to escape detection, develop techniques to friend or follow human accounts, and support one another as a huge network to gain trust. Further, a large group or network of these accounts can work together to manipulate Twitter's hot topics for harmful objectives. Twitterbots can turn public opinion on culture, products, and political agendas by automating the production of large numbers of tweets that mimic human conversation. The sheer volume of these bot accounts, as well as their rising intricacy, makes human detection of these accounts difficult. The new user registration process appears to be an appropriate location for detecting and preventing these sybil and bot accounts. Companies that facilitate this service generate fake Twitter accounts that follow a large number of people; some of these Twitter accounts may even submit illegitimate tweets to appear legitimate.

Users can buy followers, favorites, retweets, and comments on numerous websites that cater to increasing a user's image through the collection of followers, in addition to content-generating bots. Users' profiles acquire more attention as they gain more followers, increasing their popularity.

2.4. Privacy and Security

Twitter always remains in the eye of attackers and social spammers to promote hatred and jealousy all around the world. Many incidence were reported in the past which pointed out questions on Twitter. Nitesh Dhanjani and Rujith identified a security issue on April 7, 2007 (Gilbertson 2011). Because Twitter uses the sender's phone number as authentication, unscrupulous users might use SMS spoofing to change someone else's status page. If the spoofer knew the phone number associated with their victim's account, they may exploit the flaw. Twitter offered an optional personal identification number (PIN) that users could use to authenticate SMS-originating messages within a few weeks after this discovery. A dictionary attack on a Twitter administrator's password resulted in the compromising of 33 high-profile Twitter accounts on January 5, 2009. Some of the accounts that were hacked sent forged tweets, including messages about drugs (Wortham 2021). On June 11, 2009, Twitter released a beta version of their "Verified Accounts" feature, which allows users with public profiles to certify their account name. The status of these accounts is indicated by a badge on their home pages (Cashmore 2009). İnci Sözlük uncovered a problem in Twitter in May 2010 that might allow a person to force people to follow them without their consent or knowledge. Conan O'Brien's account, for example, was altered to get roughly 200 fraudulent subscriptions after it was configured to follow only one user(Mashable 2021). The US Department of Justice filed a subpoena on December 14, 2010 requesting Twitter to submit information regarding accounts registered to or linked with WikiLeaks. "It's should be the policy to notify users about law enforcement and governmental requests for their information, unless it’s forbidden by law from doing so" Twitter said in a statement. It was revealed in August 2012 that there existed a market for bogus Twitter followers, which were used to boost politicians' and celebrities' perceived popularity. Nearly every politically affiliated account from the White House to Congress to the 2016 campaign trail has been linked to the underground market for fake followers, sometimes known as bots(Carroll 2012). Campaign staff or associates of political politicians were among those involved in the creation of non-human followers, or bots. For $20, one website provided 1,000 bogus followers. The "bots" were frequently created by persons from Eastern Europe and Asia. Two Italian academics estimated that 10% of all Twitter accounts were "bots" in 2013, however other estimates have put the figure even higher (Samuelsohn 2021b). In April

2013, Twitter implemented a two-factor login verification as an added protection against hacking after a series of high-profile thefts of official accounts, including those of the Associated Press and The Guardian (Rodriguez 2021). After an outcry, including a petition with 100,000 signatures, over Tweets that included rape and death threats to historian Mary Beard, feminist campaigner Caroline Criado-Perez, and member of parliament Stella Creasy, Twitter announced plans to introduce a "report abuse" button for all versions of the site in August 2013 (Moyer 2021). In December 2014, Twitter introduced new reporting and blocking practices, including a blocking method designed by Randi Harper, a GamerGate target. CEO Dick Costolo claimed in February 2015 that he was "frankly ashamed" of how Twitter handled trolls and abuse, and that the company has lost users as a result (Mello Jr. 2014). On May 10, 2019, Twitter announced that it has already terminated 166,513 accounts for inciting terrorism between July and December 2018, claiming that the platform's "zero-tolerance policy enforcement" has resulted in a steady drop in terrorist organisations attempting to utilize it (Holt 2021). On July 21, 2020, Twitter banned 7,000 accounts and limited 150,000 more that had ties to QAnon. The bans and restrictions were implemented after QAnon-affiliated accounts began harassing other users through crowding or moderating tactics, which involved numerous accounts attacking the same person at the same time (Ben and Zadrozny 2021). Marking the start of Donald Trump supporters' protests across the United States in January 2021, Twitter suspended more than 70,000 accounts, claiming that they were "dedicated to the propagation of this conspiracy theory across the service" and shared "harmful QAnon-associated content" on a large scale (Porter 2021). Twitter's developer API is widely regarded as one of the most open and powerful of any major technological business. Developers' interest in Twitter grew quickly after its inception, pushing the business to expose its initial public API in September 2006. The API soon became well-known as a model for public REST APIs, and it is frequently quoted in programming lectures (Bynder 2021).

2.5. Detection Techniques for Bots

For all of the reasons mentioned above, the computing community is working on developing advanced ways to detect social bots or distinguish between humans and bots (Boshmaf *et*

al. 2013). Bot accounts come in several different forms on Twitter. Some are extremely rudimentary, while others have extremely intricate systems that mirror human behaviour. It gets more difficult to distinguish between synthetic and human actions. Detecting these bot accounts is a major study issue, particularly during the 2016 US presidential election. The tactics now deployed by social media platforms appear insufficient to counteract this issue, and research initiatives in this direction have only recently begun. Some of the methods used for social bots detection is listed below.

2.5.1. Graph Based

Various tasks for social bot detection have been framed in an adversarial setting. Multiple social bots (also known as sybils) can be controlled by an adversary to fake numerous identities and conduct an attack or infiltration. This characteristic is used to detect densely related sybil groups. To identify tightly linked social community bots, one standard approach is to use off-the-shelf community detection methods, however, the community detection algorithm chosen has been shown to have a significant impact on the detection algorithm's success. A smart attacker may use the controlled sybil accounts' connectivity to simulate the community structure of the portion of the social network occupied by legitimate accounts, making the attack invisible to approaches that depend simply on community detection. Connecting and talking with strangers is one of the key features of some of the sites like Twitter and Tumblr. In these situations, the innocent-by-association paradigm produces a high number of false-negative results. Some authors pointed out the limitations of the assumption of only finding groups of social bots or legitimate users: real platforms may contain many mixed groups of legitimate users who were targeted by some bots, and sophisticated bots may be successful in large-scale infiltrations, making it impossible to detect them solely based on network structure data.

2.5.2. Feature Based

According to (Wang *et al.* 2020), focusing on behavioural patterns has the advantage of being easily stored in features and used with machine learning techniques to understand the signature of human-like and bot-like activities. This enables for later classification of accounts based on their observed actions. Bot or Not?, for example, is an example of a feature-based system. It was the first publicly available social bot detection tool for Twitter, and it was released in 2014 to raise awareness about the presence of social bots. Like other

feature-based systems, it uses a detection algorithm based on highly predictive features to capture a variety of suspicious behaviours and effectively distinguish social bots from humans. Bots are always changing and evolving; analysing the highly predictive behaviours that feature-based algorithms may detect may reveal fascinating trends and provide unique possibilities to learn how to distinguish between bots and people. User meta-data is regarded as one of the most predictive and interpretable features. However, further research is needed to discover sophisticated techniques that combine human and social bot characteristics. Detecting bots or compromised accounts with feature-based solutions is currently very difficult.

2.5.3. Crowdsourcing

When (Wang *et al.* 2020) looked at the possibility of human detection, proposing that social bot detection be crowdsourced to legions of workers. They designed an Online Social Turing Test platform as a proof-of-concept. It was assumed that detecting bots is an easy task for humans, who have an unrivalled capacity to judge conversational nuances such as sarcasm or persuasive language, as well as to spot growing trends and abnormalities. The detection rate for hired workers declines over time, according to the authors, but it is still excellent enough to be employed in a majority voting technique, in which the identical profile is displayed to numerous workers and the majority's view determines the final conclusion. This technique has a low false positive rate, which is an important feature for a service provider.

2.6. Machine Learning

Machine learning is a type of data analysis that automates the creation of analytical models. It's a field of artificial intelligence based on the notion that computers can learn from data, recognize patterns, and make choices with little or no human input. Machine learning now is not the same as machine learning in the past, thanks to advances in computer technology. It was influenced by pattern recognition and the idea that computers may learn without being trained to do certain tasks; artificial intelligence researchers sought to test if computers could learn from data. The periodic component of machine learning is essential because models may adjust autonomously as they are exposed to fresh data. They use prior

computations to generate consistent, repeatable judgments and outcomes. It's a science that's not new, but it's gaining large popularity.

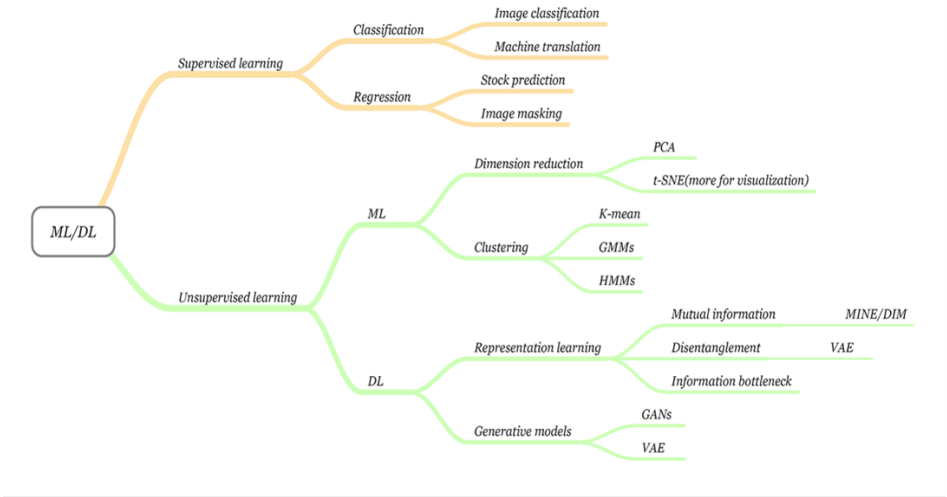

Figure 2 Machine Learning Algorithms Tree (Shukla et al. *2021)*

Machine Learning are mainly classified into two categories, however, three forms along with two of the widely known categories has been discussed below:

2.6.1. Supervised Learning

The most widely used machine learning paradigm is supervised learning. It is very easy to understand and apply. One may feed a learning algorithm data in the form of example-label pairs one by one, enabling the algorithm to predict the label for each case and providing feedback on whether it anticipated the correct answer or not. The algorithm will eventually learn to estimate the exact nature of the link between instances and labels. The supervised learning algorithm will be able to observe a new, never-before-seen sample and predict an appropriate label for it after it has been completely trained.

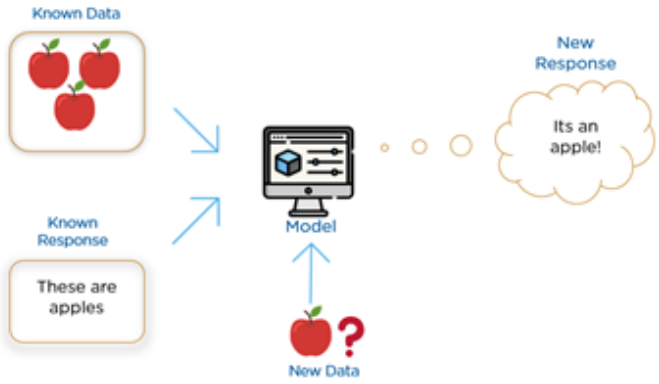

Figure 3 Supervised Learning (Heidenreich 2018)

Because of this, supervised learning is frequently defined as task-oriented. It is laser-focused on a single job, providing the algorithm more and more samples until it can reliably do that task.

2.6.2. Unsupervised Learning

The absolute opposite of supervised learning is unsupervised learning. There are no labels on it. Instead, the algorithm would be fed a large amount of data and given the tools necessary to comprehend the data's characteristics. It can then learn to group, cluster, and/or arrange the data in such a manner that a person (or another intelligent algorithm) can understand the newly structured data.

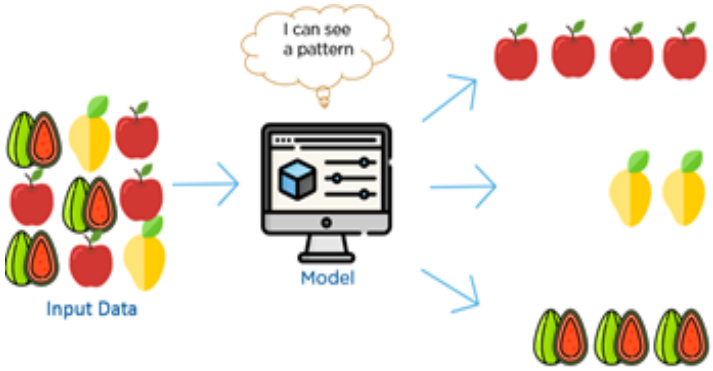

Figure 4 Unsupervised Learning (Heidenreich 2018)

The fact that the vast majority of data in the world is unlabelled is what makes unsupervised learning so exciting. For many businesses, having sophisticated algorithms that can make sense of terabytes of unlabelled data is a tremendous source of potential profit. That alone has the potential to increase productivity in a variety of sectors.

2.6.3. Reinforcement Learning

When compared to supervised and unsupervised learning, reinforcement learning is rather different. While the difference between supervised and unsupervised learning (the presence or absence of labels) is clear, the difference between reinforcement and non-reinforcement learning is less clear. Reinforcement learning is based on the actions of the learner. The areas of neuroscience and psychology have influenced it.

Figure 5 Reinforcement Learning (Heidenreich 2018)

The agent in Mario is learning algorithm, and the game represents the environment. The agent has a set of tasks to do. These will be the states of buttons. The updated state will be each game frame as time passes, and the change in score will be the reward signal. A reinforcement learning scenario will have to be put up to play Mario if all of these components are connected together.

2.7. History of Bot Detection

There are a lot of research which focuses on spam on social media sites. Spam and social bots have similar traits, but their detection mechanisms on social networks are vastly different. Spam is frequently intended, but social robots frequently publish information on social media that is extremely similar to that of normal people. Detecting whether a user is a social robot necessitates a review of the user's predicted tweets or a honeypot effort to reveal the person's attributes. According to (Mousavi *et al.* 2019), the random forest is an ensemble learning method mostly used for classification and regression procedures. During the training and output of the class, the researchers used decision-making approaches to create a large number of decision trees. Botnet approaches rely on the setup of a honeypot and the development of intrusion detection systems. Wang proposes a random forest approach for bot detection. The architecture implements support vector machine cross-

validation. For signature text data, critical phrases hashtags are employed, and structural patterns are used to gather user history. The support vector machine uses communication frequencies and opinion groups to execute correlation selection of the retrieved data and univariate selection with recursive feature elimination, which plays an important part in the bot identification phase. In this research system, (Sarabu *et al.* 2019) examined NetFlow data. The ISP network's flowing network traffic was monitored using the NetFlow & DNS data tool. The F-Test and data packet inspection are being employed in the study. Queries are used to extract data, which helps to overcome the difficulty of heavy computation and processing. (Lee *et al.* 2011) explained that Honeycup-based social robots can be detected using only a few ways. He developed a seven-month experiment in which 60 honeypots, which are also social robots that produce meaningless tweets, were deployed on social networks. Normal users, on the other hand, do not pay attention to these honeypots; only social users do. There are few traits that social robots can gain on the Internet. The qualities of social robots are the focus of the majority of machine learning-based social robot detection research. (Andriotis and Takasu 2018) mentioned that the number of fans, likes, and friends referred to as metadata, as well as the ratio, URL proportion, emotion vector, and topic vector gathered from tweets, were all listed. (Lingam *et al.* 2018) and others identify botnets based on the user's fan network; (Dickerson *et al.* 2014) proposed characteristics intersect in the emotion, topic, and timing of user's tweets, which is a rare study on mining the probable association between features.

2.8. Approach

We approach the social bot detection on Twitter as a supervised classification problem and use machine learning algorithms after data pre-processing and feature extraction operations. Large number of features are extracted by analysing tweets of Twitter user accounts, profile information and temporal behaviours such as changes in profile and tweet frequencies. In this report, the accounts that are tagged as bot on Twitter has been used to gather labelled data, assuming that the bulk of them are spam or bot accounts. For this report, Twitter Streaming API has been used to acquire data from Twitter between July 2021 and August 2021 in order to develop machine learning models and conduct tests. During the

above time period, the focus for collecting tweets and other data was on hot topics such as the Olympics in 2021, Covid, Vaccines, and so on.

The next section of this report is focused on the complete data preparation process before implementing machine learning algorithms.

3. Implementation

This chapter will examine all of the steps involved in putting the artefact into action. It will go over the processes that were taken to set up the environments that would be used to gather and process the data. A discussion of how the bot data was gathered and processed to obtain the required metrics, as well as how the model's other user-based metrics were obtained. Also, the training and testing of various machine learning algorithms will be investigated.

3.1. Dataset

When it comes to implementing machine learning algorithms, the data plays a big role in the success of a machine learning model. A genuine and real data becomes equally important when you are trying to make predictions. Because the data is the basic thing on which is required to implement any machine learning models. The train and test of model on the dataset needs to be done initially to see whether the algorithms are working good or not. The phase of creating a dataset or more than one datasets are divided in various phases. The first and foremost important part of dataset is the data collection. Beyond that comes the pre-processing and evaluation steps. These steps have been further discussed in this report. An overview of the tweet extraction from Twitter using Twitter API based on various parameters is shown in the figure below:

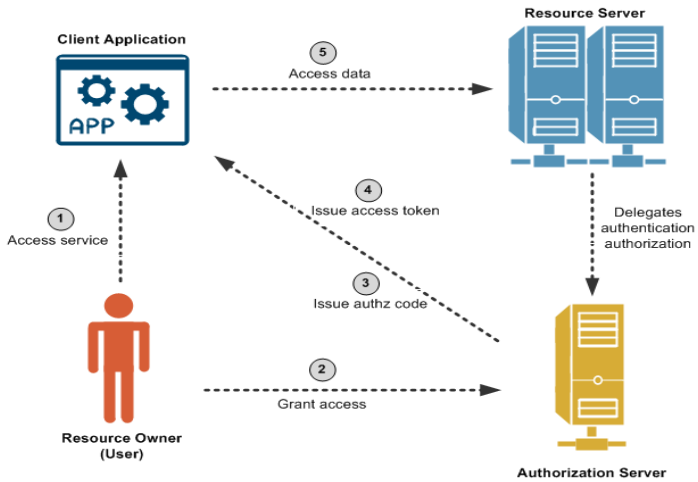

Figure 6 Tweets extraction using Twitter API (Owe 2020)

3.1.1. Data Collection

The data has been collected using the Twitter website itself after getting developer access from Twitter. There are basically two types of developer access on Twitter. One is the Standard developer access and another one is Student Research developer access.

A layout of Twitter API Developer Portal is shown in Figure 7. It is visible that there are two sections of access on the portal. One is the Standard and other one is the Academic Research. The Standard product track is suitable for both professional and inexperienced developers, from bot builders to hobbyists and even students. On the other hand the academic research access is for academics working on a non-commercial research project that necessitates or benefits from examining Twitter's conversational data. Once you apply for developer access, the developer team of Twitter API will communicate with you through the email which is registered with that particular twitter account and ask for some documents to verify your identity and know the reason for applying for the developer access. They may ask a few questions which needs to be answered and after verifying your identity and documents uploaded, they decide whether to allow permission for developer access.

For this report, both the access were applied however, only standard access was permitted. This research will move forward with the standard access of the developer portal of Twitter API. The maximum number of requests that may be made is determined by a time interval, or a set period of time. The fifteen-minute request limit interval is the most frequent. If an endpoint's rate restriction is 900 requests/15 minutes, then every 15-minute interval can have up to 900 requests.

With this Standard access, Twitter provides 1 project free of cost. The Tweet cap provided is 500,000/month for each project. However, there are certain endpoints which decides on how many Tweets can be pulled per month based on the filtered streams and recent searches.

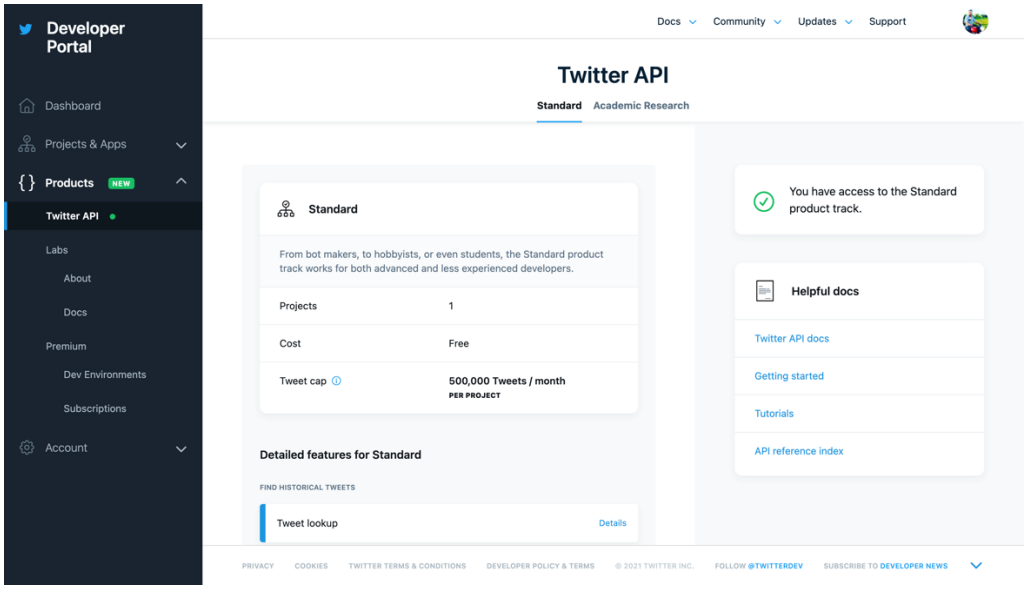

Figure 7 Twitter Developer Portal

Each project consists of its own Consumer Keys and Authentication Tokens. The keys and secrets is always unique for particular project and should be kept private and confidential.

There is a Monthly Tweet Cap Usage which keeps on track of the amount of tweets collected within a month and it resets after every month. So, if someone start extracting tweets on the first day of a month, the Tweet Cap will reset to 0 on the first day of next month. In the language of Twitter, you need to create an app to access Twitter data. The name of app used for this report is Bot Detection. Each app has its own unique app ID. The data can be accessed used the consumer key, consumer secret, access key and access secret which need to be generated while using the app for the first time. The keys and secrets can be used multiple number of times based on the requirements.

The first step of the Tweet extraction is to assign the keys and secrets to a variable as shown in below code block.

```
!pip install tweepy
import tweepy
auth_key = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth_key.set_access_token(access_token, access_secret)
api = tweepy.API(auth_key)
...
```

Code Listing 1 Twitter API Implementation

The Tweepy library is installed on Google Chrome as it is not present by default. After installing the library, we will import the Tweepy library which will be required to pull out data from Twitter.

After this the most important step is to import the libraries without which no algorithm will work. Below are the list of all the libraries used in this report.

```
import tweepy
import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import tree, metrics
from sklearn.model_selection import train_test_split, KFold
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import roc_auc_score
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import scale, normalize
from sklearn.naive_bayes import GaussianNB
from xgboost import XGBClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.datasets import make_classification
from sklearn.metrics import plot_confusion_matrix
from sklearn.svm import SVC
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import ElasticNet
```

Figure 8 Implementation of Libraries

Next is the Authentication Handler. Tweepy supports both type of authentication. The first one is OAuth 1a which redirects user to authorize the application. However, the OAuth 2 is an authentication technique that allows an application to make API requests without knowing the user's context. In both the cases, the authentication is handled using the tweepy.AuthHandler class.

After these, the next step is to extract Tweets. There are various methods to pull out tweets but two of the most commonly used methods are ID and search.

In ID method, the Tweets are pulled out based on the unique ID of a particular user. ID is basically the user name of the Twitter profile. The second method is the search method. In this method, a popular hashtag (#) is used to pull out Tweets.

In this research, the second method is given preference over the first as to make Tweet extraction process easy and convenient. The basic intuition is that bot accounts seek trending topics in order to increase the exposure of their tweets. We believe that exposure is vital for bot accounts because one of their key qualities is to reach as many people as possible. Because it is read by a large number of people, trending subjects provide this possibility. As a result, the data is relied on tweets of popular subjects as the primary source of information. Using the search techniques, a large amount of user profile can be captured.

The search keywords used for this report are:

- Covid
- Olympics
- Vaccine
- Bot

The above three (Covid, Olympics and Vaccine) keywords were used to extract human account and last one was used to pull out Bot data from Twitter. However, the later (Bot) keyword was used to extract bot data from Twitter. It is assumed that the profile which uses bot as a profile name, location or in the tweets are bot accounts. So, the data from above 3 hashtags are considered to be posted by humans and the last one which uses bot in any form are generated by bot accounts.

The composition of various fields in the dataset used in this report is described below:

Table 1 Dataset Composition

Account Category	No of Accounts	Retweets Count	Friends Count	Followers Count	Listed Count	Favorites Count	Status Count
Human & Bot	68579	98844633	87182718	1050441947	5071152	2534315453	395707916
Bot	20000	31988024	11169494	29755260	347285	178687014	1197649167

3.1.2. Data Pre-processing

The data which is extracted from Twitter using Twitter API is in the form of CSV format. The User name itself is used as a feature for bot detection, so it is not required to convert its form. The rest of the fields such as the Tweets, ID, Language and Location is just to create a proper dataset. Other profile features which will be used for the machine learning algorithms are listed as:

- Friends Count → Number of friends in the user account
- Followers Count → Number of follower in the user account
- Listed Count → Number of people have added user their list
- Favourite Count → Number of favourites a user has in the account
- Status Count → Number of status in the user account
- Verified → Is the account verified or not?
- Extended Profile → Does the user has any extended profile?
- Bot → Is the account bot or human?

3.1.3. Design

The method proposed by (Kantepe and Ganiz 2017) for bot detection on Twitter, uses ML techniques following a lengthy process of data preparation and feature extraction. To acquire data, Twitter API and Apache Spark was used. They gathered data from 16000

Twitter accounts and identified 62 traits, which were divided into three categories; User features, Tweet features and Periodic features. For the training and test set, they used ratios of 60:40, 70:30 and 80:20 respectively. Logistic Regression, Naïve Bayes, Support Vector Machine (SVM) and Gradient Boost were the four classifiers used. For Gradient Boosted trees, the accuracy result was highest with 86% and the F1 score was 83%. (Ersahin *et al.* 2017) developed a classification technique for detecting fraudulent Twitter accounts by evaluating the dataset using EMD. They evaluated their data before and after discretization, and used the Naïve bayes classifier and F-measures to determine the system's prediction accuracy. Before applying the proposed approach, the results were 85.5%, but after employing the edge detection technique on chosen characteristics, the result was improved to 90.41%. The algorithm below shows the process involved in this report for the bot detection model:

Figure 9 Design Process of Bot Detection model

3.1.4. Requirements

The hardware used to perform machine learning algorithms locally is Apple's MacBook Air. The specifications of the local machine are 8GB Memory, 256GB SSD, new advance M1 chip which performs much better than i7 and i9 chips of Intel, and runs on latest macOS Big Sur version 11.5.1.

Figure 10 Local Machine requirements

The software requirements for this report consists of Google Colab. It is a cloud based platform from Google where we can perform machine learning tasks.

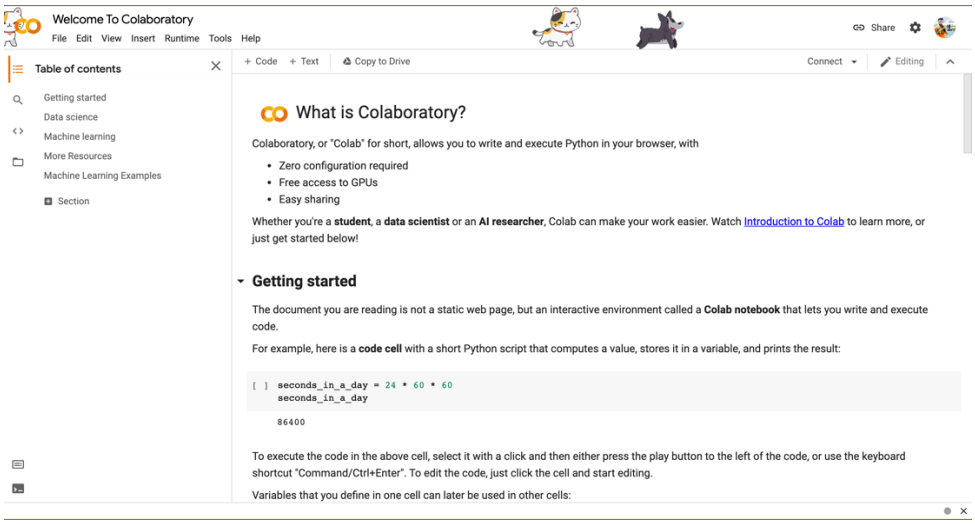

Figure 11 Google Colaboratory ('Google Colaboratory' 2021)

The Google Colab runs on Python 3 Google Compute Engine at the backend.

Others details of Google Colab is mentioned in figures below:

```
Filesystem      Size  Used Avail Use% Mounted on
overlay         108G   37G   71G  35% /
tmpfs            64M     0   64M   0% /dev
tmpfs           6.4G     0  6.4G   0% /sys/fs/cgroup
shm             5.9G     0  5.9G   0% /dev/shm
tmpfs           6.4G   28K  6.4G   1% /var/colab
/dev/sda1        76G   42G   35G  55% /etc/hosts
tmpfs           6.4G     0  6.4G   0% /proc/acpi
tmpfs           6.4G     0  6.4G   0% /proc/scsi
tmpfs           6.4G     0  6.4G   0% /sys/firmware
```

Figure 12 Disk Information (Saurabh 2021)

```
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 85
model name      : Intel(R) Xeon(R) CPU @ 2.00GHz
stepping        : 3
microcode       : 0x1
cpu MHz         : 2000.180
cache size      : 39424 KB
physical id     : 0
siblings        : 2
core id         : 0
cpu cores       : 1
apicid          : 0
initial apicid  : 0
fpu             : yes
fpu_exception   : yes
cpuid level     : 13
wp              : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 ss ht syscall
bugs            : cpu_meltdown spectre_v1 spectre_v2 spec_store_bypass l1tf mds swapgs taa
bogomips        : 4000.36
clflush size    : 64
cache_alignment : 64
address sizes   : 46 bits physical, 48 bits virtual
power management:

processor       : 1
vendor_id       : GenuineIntel
cpu family      : 6
model           : 85
model name      : Intel(R) Xeon(R) CPU @ 2.00GHz
stepping        : 3
microcode       : 0x1
cpu MHz         : 2000.180
cache size      : 39424 KB
physical id     : 0
siblings        : 2
core id         : 0
cpu cores       : 1
apicid          : 1
initial apicid  : 1
fpu             : yes
fpu_exception   : yes
cpuid level     : 13
wp              : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 ss ht syscall
bugs            : cpu_meltdown spectre_v1 spectre_v2 spec_store_bypass l1tf mds swapgs taa
bogomips        : 4000.36
clflush size    : 64
cache_alignment : 64
address sizes   : 46 bits physical, 48 bits virtual
power management:
```

Figure 13 CPU Specification (Saurabh 2021)

```
MemTotal:       13302928 kB
MemFree:        10603360 kB
MemAvailable:   12515200 kB
Buffers:          119780 kB
Cached:          1933332 kB
SwapCached:            0 kB
Active:           983228 kB
Inactive:        1472192 kB
Active(anon):     371636 kB
Inactive(anon):      432 kB
Active(file):     611592 kB
Inactive(file):  1471760 kB
Unevictable:           0 kB
Mlocked:               0 kB
SwapTotal:             0 kB
SwapFree:              0 kB
Dirty:             13240 kB
Writeback:             0 kB
AnonPages:        402260 kB
Mapped:           237408 kB
Shmem:              1140 kB
KReclaimable:     139928 kB
Slab:             183888 kB
SReclaimable:     139928 kB
SUnreclaim:        43960 kB
KernelStack:        4704 kB
PageTables:         5472 kB
NFS_Unstable:          0 kB
Bounce:                0 kB
WritebackTmp:          0 kB
CommitLimit:     6651464 kB
Committed_AS:    3012900 kB
VmallocTotal:   34359738367 kB
VmallocUsed:        7060 kB
VmallocChunk:          0 kB
Percpu:             1400 kB
AnonHugePages:         0 kB
ShmemHugePages:        0 kB
ShmemPmdMapped:        0 kB
FileHugePages:         0 kB
FilePmdMapped:         0 kB
CmaTotal:              0 kB
CmaFree:               0 kB
HugePages_Total:       0
HugePages_Free:        0
HugePages_Rsvd:        0
HugePages_Surp:        0
Hugepagesize:       2048 kB
Hugetlb:               0 kB
DirectMap4k:       88896 kB
DirectMap2M:     5150720 kB
DirectMap1G:    10485760 kB
```

Figure 14 Memory (Saurabh 2021)

Though the majority of tasks were performed on Google Colab, which is a cloud based platform from Google, Jupyter Notebook (‘Project Jupyter’ 2021)was used to perform some of the training of data to see if it is working efficiently on local environment as well as the cloud platform always have some limitations. A number of libraries will be used to achieve the research goal, including:

- **Pandas** → is a quick, powerful, versatile, and easy-to-use open source data analysis and manipulation tool build on top of Python programming language (Pandas 2021).
- **Numpy** → is a Python library that adds support for huge, multi-dimensional arrays and matrices, as well as a large number of high-level mathematical functions to operate on these arrays (Numpy 2021).

- **Random** → is a built in module which is used to generate pseudo-random variables. It can be used to achieve tasks like generate a random number, choose random elements from a list, shuffle elements randomly, and so on.
- **Matplotlib** → is a plotting library for Python with NumPy, the Python numerical mathematics extension. It provides an object-oriented API for embedding charts into applications utilizing GUI toolkits such as Qt, or GTK (Matplotlib 2021).
- **Sklearn** → is a free Python machine learning package. It includes support vector machines, random forests, gradient boosting, and k-means, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy (Scikit 2021).
- **Seaborn** → is a matplotlib-based Python data visualization package. It has a high-level interface for creating visually appealing and instructive statistics visuals (Waskom 2021).
- **XGBoost** → is a new library and since its introduction in 2014, it has been nicknames as the holy grail of machine learning hackathons and competitions. XGBoost has proven its mettle in terms of performance – and speed – on anything from forecasting ad click-through rates to identifying high-energy physics events. XGBoost is a distributed gradient boosting library that has been developed for efficiency, flexibility, and portability (Vidya 2021).

3.2. Evaluation

To evaluate the models, approaches similar to (Ersahin *et al.* 2017) will be used. In the past ROC curve were used along with F1 score, Accuracy and Confusion Matrix. However, in this report, F1 score, Accuracy, Precision and Confusion Matrix will be used along with AUC curve.

The above metrics are calculated based on True Positive, False Positive, True Negative and False Negative has been explained further in the table below:

Table 2 Parameters of Confusion Matrix

True Positive (TP)	**False Positive (FP)**	**True Negative (TN)**	**False Negative (FN)**
True Positive is the situation where actual values matches the predicted values.	False Positive is the situation where the predicted values were predicted false. This is referred as Type 1 error.	True Negative is a situation where predicted values matches the true or actual values.	False Negative is a situation similar to False Positive where predicted values were predicted false. This is referred as Type 2 error.
In this situation, the model predicts values as positive when the actual values were also positive.	In this situation, the model predicts the values as positive when the actual values were negative.	In this situation, the model predicted the values as negative when the actual values were also negative.	In this situation, the model predicted the values as negative when the values were positive.

3.2.1. Accuracy

The number of right predictions divided by the total number of guesses in the dataset gives the accuracy (ACC). The highest level of accuracy is 1.0, while the lowest level is 0.0. It can also be calculated by 1 – ERR.

Equation 1

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

Equation 2

$$Error\ (ERR) = \frac{FP + FN}{TP + TN + FP + FN} = \frac{FP + FN}{P + N}$$

3.2.2. Precision

Precision indicates how many of the instances that were accurately predicted turned out to be positive.

Equation 3

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

A precision would tell whether our model is reliable or not. When FPs are more of a worry than FNs, precision is a helpful indicator. In music or video recommendation systems, e-commerce websites, and other applications, precision is critical. Customer turnover and company losses might arise from incorrect results (Vidya 2020).

3.2.3. Recall

Recall indicates how many of the actual positive cases our model was able to properly anticipate.

Equation 4

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

When False Negative beats False Positive, Recall is a valuable metric. In medical situations, recall is critical since it doesn't matter if we raise a false alert; the genuine positive cases should not go unnoticed! (Vidya 2020)

3.2.4. F1 Score

During the practice, when we work to increase our model's precision, the recall drops, and vice versa. In a single value, the F1 score captures both trends.

Equation 5

$$F1\ score = \frac{2}{\left(\frac{1}{Recall}\right) + \left(\frac{1}{Precision}\right)}$$

3.2.5. AUC Curve

The Area Under the Curve (AUC) is a summary of the ROC curve that measures a classifier's ability to discriminate between classes.

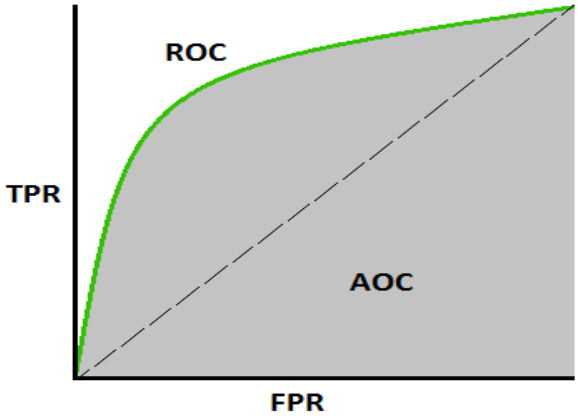

Figure 15 ROC-AUC Graph

3.2.6. Confusion Matrix

A Confusion Matrix is basically a (N x N) matrix which is used to evaluate the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values to the machine learning model's predictions. This provides us with a comprehensive picture of how well our classification model is doing and the types of errors it makes.

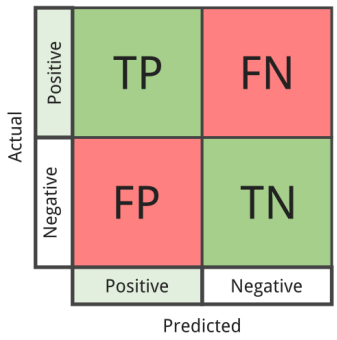

Figure 16 Confusion Matrix (MLeeDataScience 2021)

Above, a (2 X 2) Confusion Matrix has been shown with 4 values. In the above figure there are Actual values and Predicted values. Let's discuss the matrix further:

- There are two values **Positive** and **Negative** which represents the target variables
- The rows represents the True Label or Actual values of the target variables.
- The columns represents the Predicted Label of the target variables.

3.3. Feature Engineering

Researchers have progressively developed ways to detect these accounts as bots have multiplied and their usage has been widely debated in the media and political organizations. The same openness and simplicity with which social media APIs make it easy to create and

utilize automated accounts also makes it easy to collect data to detect them (Beskow and Carley 2019). Bot engineers modify and adapt as detection efforts increase in order to live and flourish in a dynamic environment. The primary focus is to enhance not just the models that identify bots, but also the labelled data that is used to train them are motivated by the need for better accuracy in the face of shifting signals.

3.4. Bot Classification

For bot detection, the approach of supervised learning is adapted as it will become very difficult to make any prediction if the method is unsupervised. The usage of labelled datasets is the key difference between the two methodologies. Simply put, supervised learning algorithms use labelled input and output data, whereas unsupervised learning algorithms do not. The algorithm learns from the training dataset by iteratively making predictions on the data and adjusting for the correct answer in supervised learning. While supervised learning models are more accurate than unsupervised learning models, they do necessitate human interaction to properly identify the data.

A supervised learning model, for example, can forecast the length of your commute based on the time of day, weather conditions, and other factors. But first, you'll have to teach it that driving in rainy weather takes longer.

Unsupervised learning models, on the other hand, function independently to uncover the structure of unlabelled data. It's important to remember that validating output variables still necessitates human intervention. An unsupervised learning model, for example, can detect that online buyers frequently purchase groups of products at the same time.

3.5. Modelling

When the data collection phase is complete, we can start the modelling process which is implementing machine learning algorithms on our dataset to see which model is performing best on the dataset. The models used in this report are:

- **Linear Regression :** Linear regression basically means finding a linear relationship between target and one or more variables.
- **Lasso Regression :** Lasso regression is a type of regularization. For a more precise prediction, it is preferred over regression techniques. Shrinkage is used in this model. Data values are shrunk towards a central point known as the mean in shrinkage.
- **Ridge Regression :** Ridge regression is a model tuning technique that may be used to analyse data with multicollinearity. When there is a problem with multicollinearity, least-squares are unbiased, and variances are significant, the projected values are distant from the actual values.
- **Elastic Net :** The penalties from both the lasso and ridge methods are used to regularize regression models in elastic net regression. The methodology combines the lasso and ridge regression methods by learning from their flaws to enhance statistical model regularization.
- **Decision Tree :** For classification and regression, Decision Tree (DT) is a non-parametric supervised learning approach. The objective is to learn basic decision rules from data characteristics to construct a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.
- **Random Forest :** A random forest is a machine learning approach for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving difficult problems by combining many classifiers.
- **K-Fold :** The K-Folds approach is popular and simple to comprehend, and it typically produces a less biased model than other methods. Because every observation from the original dataset has a probability of showing up in the training and test sets.
- **Naïve Bayes :** The Bayes Theorem is used to create a set of classification algorithms known as Naive Bayes classifiers. It is a family of algorithms that share a similar premise, namely that each pair of characteristics being categorized is independent of the others.
- **XG Boost :** XGBoost is a distributed gradient boosting toolkit that has been tuned for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to construct machine learning algorithms. XGBoost is a parallel tree boosting algorithm that solves a wide range of data science issues quickly and accurately. The best part

is that it runs on various other platforms such as Hadoop, SGE, MPI, etc. and solve a lot of problems with ease.

3.6. Conclusion

This chapter discusses on implantation of machine learning algorithms on the collected data in CSV format. The models were run on the datasets collected in the form of combined data frame and bot data frame. In the combined dataset, humans as well as bots accounts and data related to them are present while in bot dataset, only bot related data are present. In the next chapter, the results of various machine learning algorithms working of the data frames has been discussed.

4. Results and Analysis

This section describes the outcomes of the algorithms used in the preceding chapter and evaluates them using the Accuracy, Precision, Recall, F1 score, AUC curve, and FPR parameters. These measures will be used to compare the results of the classification algorithms such as Linear (LR), Lasso, Ridge, Elastic Net, Decision Tree (DT), Naïve Baise (NB), Random Forest (RF), and XGB to the results achieved. The Regression family is compared based on MSE, R2 Score and AUC Score however, the other models including the classifiers will be compared based on accuracy, precision, recall and F1 score.

4.1. Dataset Visualization

The summary of combined data frame which includes both bot and human accounts is shown in Figure 17.

```
The summary of dataset is :

Friends Count       8.718272e+07
Followers Count     1.050442e+09
Listed Count        5.071152e+06
Favorite Count      2.534315e+09
Status Count        3.975708e+09
Verified            1.368000e+03
Extended Profile    3.642800e+04
Bot                 2.000000e+04
dtype: float64
```

Figure 17 Summary of Dataset

The description of the data frame is shown in Figure 18.

	Friends Count	Followers Count	Listed Count	Favorite Count	Status Count	Bot
count	68001.000000	6.800100e+04	68001.000000	6.800100e+04	6.800100e+04	68001.000000
mean	1282.079940	1.544745e+04	74.574668	3.726880e+04	5.846543e+04	0.294113
std	5085.910043	4.173216e+05	1548.995422	8.243172e+04	1.284710e+05	0.455646
min	0.000000	0.000000e+00	0.000000	0.000000e+00	1.000000e+00	0.000000
25%	98.000000	5.400000e+01	0.000000	3.140000e+02	2.724000e+03	0.000000
50%	403.000000	2.720000e+02	1.000000	6.122000e+03	1.554500e+04	0.000000
75%	1180.000000	1.123000e+03	8.000000	3.605300e+04	5.763100e+04	1.000000
max	689319.000000	5.435129e+07	210372.000000	1.549803e+06	2.905164e+06	1.000000

Figure 18 Description of Dataset

The Pearson Correlation Matrix is shown in Figure 19.

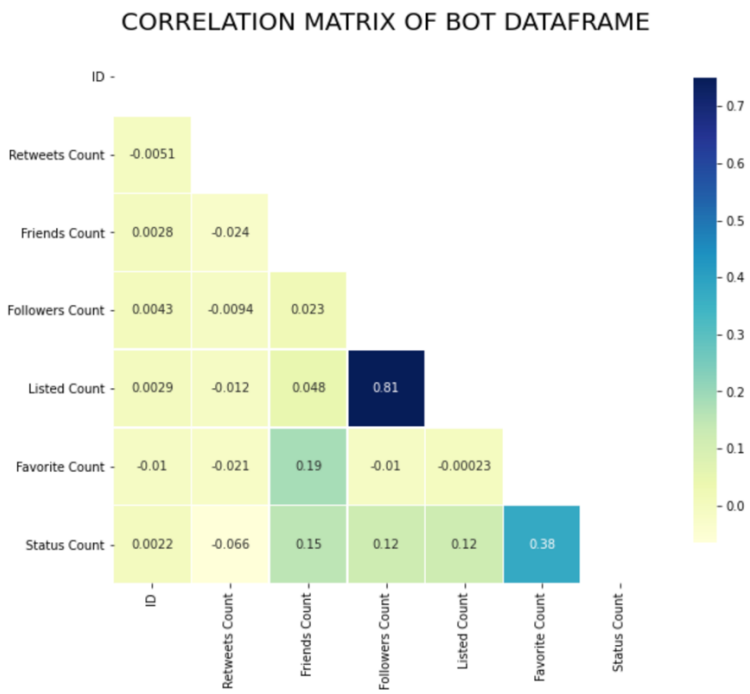

Figure 19 Correlation Matrix of Bot and Not Bot accounts

4.2. ML Algorithms Results

Overall, 10 different ML algorithms were implemented on the dataset in this report however, the parameters used for the regression models were different from the classification models. The results based on the parameters are discussed below:

4.2.1. Linear Regression

The Linear Regression is the simplest form of all of the machine learning algorithms. The Mean Square error, R2 and AUC is shown in Figure 20.

```
**LINEAR REGRESSION**

For Train Set :-
Mean Squared Error : 2.6458729202838594e-16
R2 Score : 1.0
ROC AUC Score : 1.0


For Test Set :-
Mean Squared Error : 2.4665186712775476e-16
R2 Score : 1.0
ROC AUC Score : 1.0
```

Figure 20 Classification Report Linear Regression

4.2.2. Lasso Regression

The Lasso Regression is an advance form of the Regression models and is preferred over other regression models in this category.

The Mean Square Error, R2 score and AUC Score is almost same as Linear Regression. MSE is best amongst all the model of this family. The results of Lasso model is shown in Figure 21.

```
**LASSO REGRESSION**

For Train Set :-
Mean Squared Error : 0.1097045549853881
R2 Score : 0.9420626494024391
ROC AUC Score:  1.0


For Test Set :-
Mean Squared Error : 0.10955240615137468
R2 Score : 0.9420621010369202
ROC AUC Score:  1.0
```

Figure 21 Classification Report Lasso Regression

4.2.3. Ridge Regression

The Ridge Regression is also from Regression family and is considered as a good model among Regression models. The MSE is higher than Lasso model however, it is better than the Linear Regression.

The R2 Score and AUC were almost same as the above two regression models. The result is shown in Figure 22.

```
**RIDGE REGRESSION**

For Train Set :-
Mean Squared Error : 2.1007969414580747e-06
R2 Score : 0.999999999978754
ROC AUC Score : 1.0


For Test Set :-
Mean Squared Error : 2.0995237049421526e-06
R2 Score : 0.9999999999787206
ROC AUC Score : 1.0
```

Figure 22 Classification Report Ridge Regression

4.2.4. Elastic Net

Elastic Net is the last member in this report from the Regression family. This model also performed good giving MSE lesser than Linear Regression, Lasso Regression and even Ridge Regression.

The R2 score and AUC was however close to or similar to the other three models of the family. The Figure 23 shows the exact numbers for the Elastic Net model.

```
**ELASTIC NET REGRESSION**

For Train Set :-
Mean Squared Error : 0.09791981743223889
R2 Score : 0.9538416260123513
ROC AUC Score : 1.0


For Test Set :-
Mean Squared Error : 0.09778401280634066
R2 Score : 0.9538411891325072
ROC AUC Score : 1.0
```

Figure 23 Classification Report Elastic Net

Elastic Net can also be considered as a good model as it is giving a very good MSE and R2 score and is the one of those model including Lasso which is giving a very good result on the bot and human accounts.

More prediction can be made after testing it on other datasets after applying more feature engineering for the bot detection.

4.2.5. Decision Tree Classifier

Decision Tree is considered as one of the most important classifier model of machine learning algorithms. The confusion matrix shows that the classifier was able to predict 77 False Negative and 23 True Positive out of the sample of 100. There is no False Positive or True Negative which is a very good thing. When it was tested on large scale putting the test size to 20% of the overall dataset, it was able to justify it's previous performance. Overall it can be considered as a good model seeing the accuracy in Figure 24. However, it cannot be used as a dependable model for bot detection. The precision, recall and F1 score is 100% and it has been tested multiple times by increasing the dataset but it is giving same accuracy. This can be a case of over fitting. To check whether the model has been overfitted, we need to check this with more datasets and more user features before making a judgement about its reliability.

```
**DECISION TREE CLASSIFIER**

Accuracy is 100.00%
Recall is 100.00%
Precision is 100.00%
F1 Score is 100.00
Area Under Curve is 1.0000

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     48001
         1.0       1.00      1.00      1.00     20000

    accuracy                           1.00     68001
   macro avg       1.00      1.00      1.00     68001
weighted avg       1.00      1.00      1.00     68001
```

Figure 24 Classification Report DT

The confusion matrix for Decision Tree classifier can be seen in Figure 25, where it has predicted all the True Positives and False Negatives very efficiently.

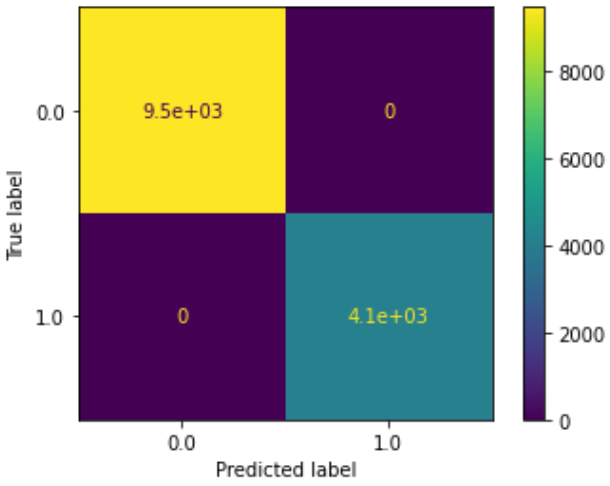

Figure 25 Confusion Matrix DT

The AUC Graph based on ROC_AUC score can be seen in Figure 26 which is exactly 1. As we can clearly see the value of True Positive and False Positive varies from 0 to 1, where 1 is considered as best.

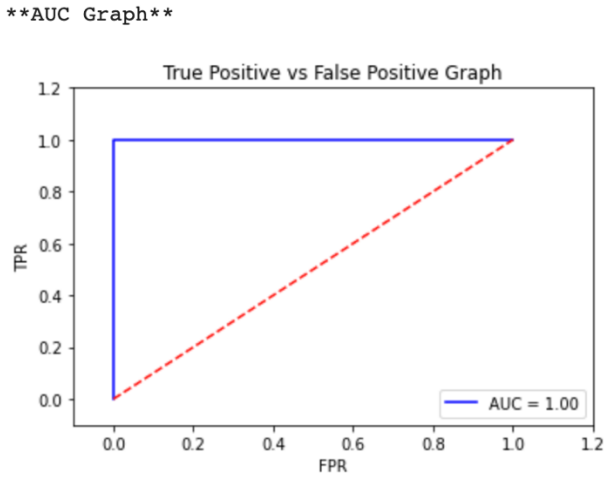

Figure 26 AUC Graph DT

4.2.6. Logistic Regression

Logistic Regression is one of the most important and advance regression model. In this report, the regression model performed very well. The accuracy shows around 80% with a precision of 82% can be trusted and hence we can do a little improvement for better results.

```
**LOGISTIC REGRESSION**

Accuracy is 80.41%
Precision is 82.38%
Recall is 42.14%
F1 Score is 0.56
Area Under Curve is 0.69

Classification Report:
               precision    recall  f1-score   support

         0.0       0.80      0.96      0.87     12020
         1.0       0.82      0.42      0.56      4981

    accuracy                           0.80     17001
   macro avg       0.81      0.69      0.72     17001
weighted avg       0.81      0.80      0.78     17001
```

Figure 27 Classification Report Logistic Regression

Figure 28 shows the confusion matrix of Logistic Regression where it was able to predict 35000 True Positives and 6300 False Negatives which can be considered as very good result. The number of outliers were 10100. So, this model has a lot of scope of improvement.

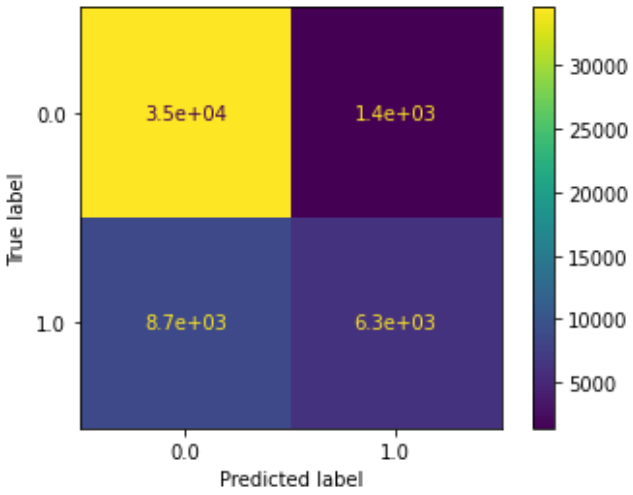

Figure 28 Confusion Matrix Logistic Regression

The AUC Graph of Logistic Regression in Figure 29 shows that the AUC score is almost 70% which is a healthy score based on testing it on dataset close to 70000 accounts.

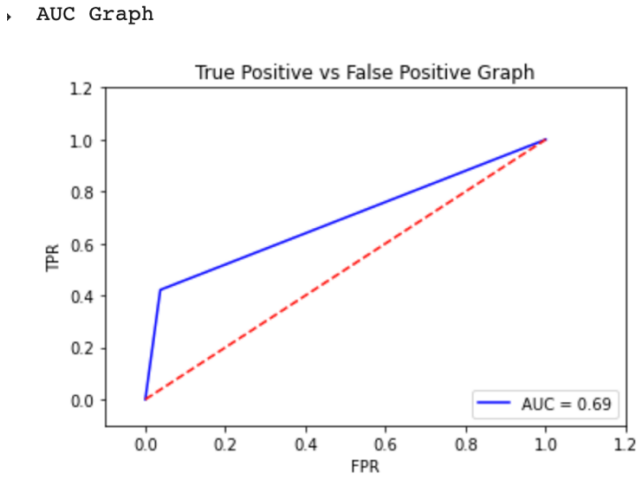

Figure 29 AUC Graph Logistic Regression

4.2.7. K-Fold Splits

In K-Fold, 10 splits is counted as ideal so, in this report, the number of splits assigned to K-Fold was 10. The different results for every split has been discussed in Table 3.

Table 3 K-Fold Results Each Split

Parameters & No. of Splits	**Accuracy**	**Precision**	**Recall**
N = 1	96.66%	0%	0%
N = 2	58.94%	95.84%	41.10%
N = 3	86.19%	80.67%	43.25%
N = 4	71.90%	90.11%	45.27%
N = 5	82.28%	82%	42.30%
N = 6	88.21%	64.91%	43.10%
N = 7	81.03%	80.60%	46.75%
N = 8	83.69%	70.01%	36.36%
N = 9	90.63%	46.80%	45.17%
N = 10	62.82%	94.23%	39.20%

In Table 3, we can see that for most of the splits, the accuracy is above 80%. For split 1 & 9, the accuracy is highest and going beyond 90%. The precision is highest for split 2 and recall is usually among the 40s.

The confusion matrix of K-Fold was combined with Logistic Regression and it has performed superbly well by leaving only 3 False Positive and 1 True Negative.

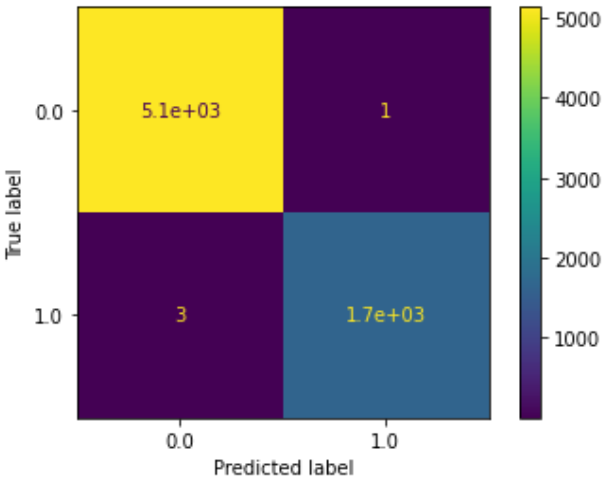

Figure 30 Confusion Matrix K-Fold with Logistic Regression

4.2.8. Naïve Bayes

Naïve Bayes performed well as compared to logistic regression. The accuracy was 86% which is pretty good with a precision of 95%. The recall value was 54% and F1 score was 0.69 (69%). The AUC score was 0.77. The classification report of the algorithm is shown Figure 31.

```
**NAIVE BAISE**

Accuracy is 85.56%
Precision is 94.77%
Recall is 53.86%
F1 Score is 0.69
Area Under Curve is 0.76
Classification Report:
              precision    recall  f1-score   support

         0.0       0.84      0.99      0.91     48001
         1.0       0.95      0.54      0.69     20000

    accuracy                           0.86     68001
   macro avg       0.89      0.76      0.80     68001
weighted avg       0.87      0.86      0.84     68001
```

Figure 31 Classification Report Naive Bayes

The confusion matrix of Naïve Bayes is shown in Figure 32 and it is clear this model has performed similar to Logistic Regression in terms of taking out True Negative and False Positives which overall makes a figure of 9790. It has predicted 47000 False Negative and 11000 True Positives.

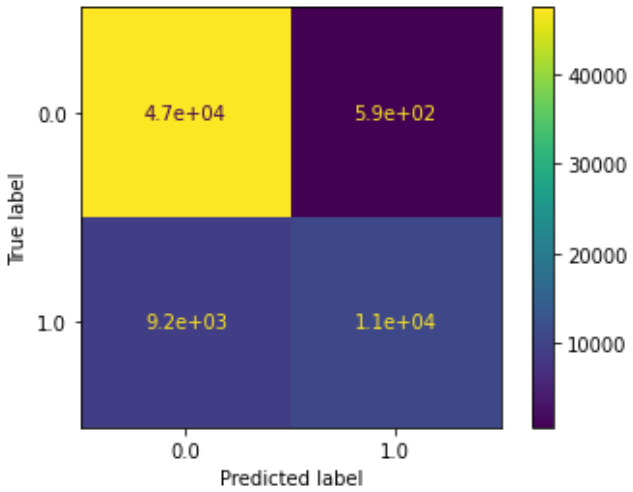

Figure 32 Confusion Matrix Naive Bayes

The AUC Graph based on the ROC_AUC score of Naïve Bayes model which is 0.76 is shown in Figure 33.

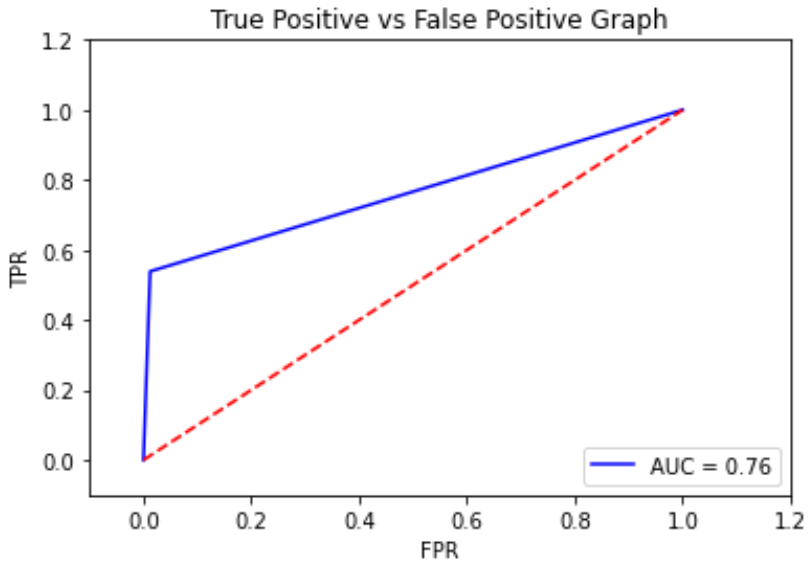

Figure 33 AUC Graph Naive Bayes

4.2.9. XG Boost

A lot was expected from XG Boost model which is one of the advance form of machine learning algorithms. It uses Gradient Boosting technique for fast and high performance tree model. While implementing it on the dataset, it can be seen that the model works good with the training set however, the result with the test dataset is not satisfactory.

The value came out to be 0.82 for the training set and 0.55 for the test set. This model was used on a different set of data frame however, the result was almost the same. The performance of this model may be poorer on the basis of the dataset it has been tested upon however, there are other models used in this report which has performed better on the same set of dataset. It delivers a good accuracy of 82% on the training set of data frame however, this model can be report based on implementation in this report.

The confusion matrix of XG Boost is shown in Figure 34. It has successfully predicted 55000 combinations of True Positives and False Negatives and threw 12700 outliers which is a mix of True Negatives and False Positives. This model can be relied on but to a certain extend.

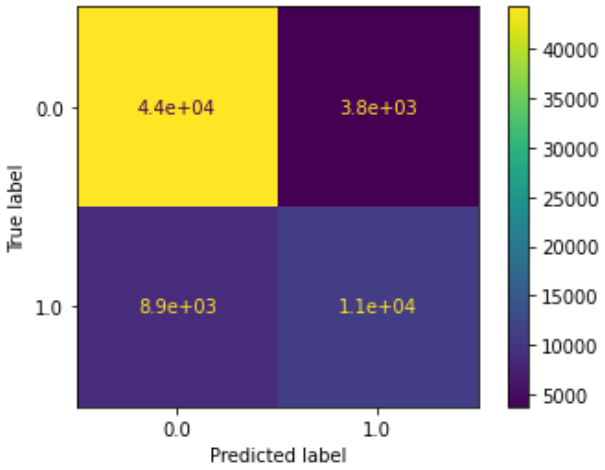

Figure 34 Confusion Matrix XG Boost

4.2.10. Random Forest Classifier

Random Forest proved to be very accurate in finding only the bot accounts out of the dataset and it left all the human accounts as it is. The accuracy is 99.7% which is extremely satisfying. This model is a good performer in the bot detection. The classification report of Random Forest can be seen in Figure 35.

```
Confusion Matrix:

[[2094    0]
 [   6    0]]

Accuracy: 0.9971428571428571
Classification Report:
              precision    recall  f1-score   support

       False       1.00      1.00      1.00      2094
        True       0.00      0.00      0.00         6

    accuracy                           1.00      2100
   macro avg       0.50      0.50      0.50      2100
weighted avg       0.99      1.00      1.00      2100
```

Figure 35 Classification Report Random Forest

4.3. Analysis

In this section of the chapter, we analyse various machine learning models used in this report for bot detection. To start with the regression models, it is clearly visible that they performed better than the classifier models. The R2 Score, ROC_AUC Score is usually same for all the four regression model (Linear, Lasso, Ridge and Elastic Net) however, there is minor difference in the R2 Score of Elastic Net Regression which looks more genuine than the other three models. The MSE of Lasso is 0.09, slightly better than 0.1 of Lasso Regression and remains the best performer among the Regression family.

The Decision Tree in classifier models throws a 100% accuracy in all the four parameters namely accuracy, precision, recall and F1 score respectively. However, it shows a lot of True Negative and False Positive while plotting the confusion matrix. Logistic Regression and Naïve Bayes gave an accuracy of 80% and 85% respectively and performed better than XG Boost which was expected to perform better than others. Random Forest gave an accuracy of 99.7% in a sample of 2100 accounts which outperformed than all the other machine learning models.

4.4. Conclusion

To conclude this chapter, we can say that the models which were expected to performed better than basic machine learning algorithms, performed at par or below par in different scenarios. The regression models performed very good which was not expected. Decision Tree Classifier, outperformed as compared to other models. There is still a scope of improvement as they have not been tested on other datasets available in the market however, with the dataset used with this report, some of the algorithms performed exceptionally well. In the next chapter, we will discuss further about the limitation and the future aspect of our model for bot detection.

5. Conclusion

The goal of the study is to emphasize the necessity of using machine learning to detect bots on social media platforms and to address the resource constraints that different models for limiting the spread of hatred and abuse encounter. The research aim to apply machine learning techniques in order to improve existing methods and address challenges with accurate bot detection. Our research question was focused on determination of False Positive and True Negative and this research somehow was able to justify that if we see the results of some of the algorithms which gave more than 90% of accuracy. However, the question is partially answered as we can see that other models struggle to give expected results including the XG Boost. While XG Boost was not quite effective on the testing data, it was good with the training dataset and gave more than 80% accuracy which can be counted as a par model based on the dataset. All the models except few were running promptly and were giving large number of True Positives and False Negatives however, the number of outliers were also in thousands. This happened only in running on accounts which counts in thousands. There is still scope of improvement in some of models among which Naïve Bayes and Logistic Regression are the best example. The next section will discuss about the limitations of models used in this research paper.

5.1. Limitations

The recommended system's limitation is analysing the actual volume of the dataset required for improving bot identification quality and raising the speed with which user requests are processed. Due to the dynamic nature of the environment, it is obvious that manual interpretation is prone to inaccuracy. Though there are models in this report which outperformed from some of the expected models for e.g. Decision Tree and K-Fold performed better than XG Boost which was supposed to give the best result. However, there is still some limitation as only one of the feature was used in this report to perform the training and testing of models on the data frame. If the data is large in millions and there is not an option to categorize the data based on 0 and 1 for bots, the above models could be less effective and may not give better results. The data which is gathered from Twitter doesn’t require much cleaning like in almost 60000 accounts, 8000 were the

accounts which has less information of has nan values. So, this could also be one of the reason of the models being so good. This can be considered as a limitation of the models which is specific to a particular data frame. One more limitation which was seen during the implementation of algorithms was the speed. When the models were trained after increasing or adding more data to the initial dataset, the speed of the model was slow and it took around 1-2 minutes to plot confusion matrix in case of DT model. This could be more time consuming when the models would run of dataset which contains millions of tweets. This could be fixed when running it on Big Data platforms. One of the major problems which may arise in the models testes is that they have been tested on only one parameter and the authenticity of all the models can only be relied upon after testing it on more user features which includes tweets, emotion and choice of words.

5.2. Future Work

The use of modern data mining techniques for comparing numerous databases and sources while generating the dataset is the research's future focus, since it improves the accuracy of detecting bots from large datasets. In this report, various machine learning algorithms were applied on the data collected from Twitter in the form of humans and bots. We have seen some of the models have outperformed and gave 100% accuracy. This might be a case of over-fitting however, the data was increased and yet again the model was very accurate. The XG Boost classifier was expected to be the best among all the models used in this report however, Decision Tree and Random Forest proved to be better than XG Boost. The K-Fold also performed good in giving accuracy of more than 80% is almost every split which is a good thing. There is still a lot which can be improved in bot detection model where models can be run on amount of datasets to check the accuracy of models. There are few more bot detection techniques those of which have proved to be very effective and it can be implemented by taking help of Big Data platform to work on datasets which are in GBs. Cloud based resources can be used like AWS (‘Free Cloud Computing Services - AWS’ 2021) as the working environment. In future, we can work upon more user features of bot detection which has been discussed in the previous chapter. After working on more user

features, the ML model can easily be judged looking at the performance of models on different parameters.

References

Abril, D. (2021) Twitter's User-Reported Violations Jumped 19%—but the Number of Accounts Punished Dropped [online], *Fortune*, available: https://fortune.com/2019/05/10/twitter-transparency-report-abuse/ [accessed 23 Aug 2021].

Alothali, E., Zaki, N., Mohamed, E.A., Alashwal, H. (2018) 'Detecting Social Bots on Twitter: A Literature Review', in *2018 International Conference on Innovations in Information Technology (IIT)*, Presented at the 2018 International Conference on Innovations in Information Technology (IIT), IEEE: Al Ain, 175–180, available: https://ieeexplore.ieee.org/document/8605995/ [accessed 21 Aug 2021].

Andriotis, P., Takasu, A. (2018) 'Emotional Bots: Content-based Spammer Detection on Social Media', in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Presented at the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–8.

AYDIN, İ., SEVİ, M., SALUR, M.U. (2018) 'Detection of Fake Twitter Accounts with Machine Learning Algorithms', in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Presented at the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 1–4.

'Basic evaluation measures from the confusion matrix' (2015) *Classifier evaluation with imbalanced datasets*, available: https://classeval.wordpress.com/introduction/basic-evaluation-measures/ [accessed 30 Aug 2021].

Ben, C., Zadrozny, B. (2021) Twitter Bans 7,000 QAnon Accounts, Limits 150,000 Others as Part of Broad Crackdown [online], *NBC News*, available: https://www.nbcnews.com/tech/tech-news/twitter-bans-7-000-qanon-accounts-limits-150-000-others-n1234541 [accessed 23 Aug 2021].

Beskow, D., Carley, K. (2019) 'Its All in a Name: Detecting and Labeling Bots by Their Name', *Computational and Mathematical Organization Theory*, 25.

Bhandari, A. (2020) 'AUC-ROC Curve in Machine Learning Clearly Explained', *Analytics Vidhya*, available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/ [accessed 28 Aug 2021].

Blesson, D. (2021) Twitter Tweets Extracting Using Tweepy [online], available: https://kaggle.com/blessondensil294/twitter-tweets-extracting-using-tweepy [accessed 21 Aug 2021].

Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M. (2013) 'Design and analysis of a social botnet', *Computer Networks*, Botnet Activity: Analysis, Detection and Shutdown, 57(2), 556–578.

Bynder (2021) Top 8 Web APIs Bridging Today's Technology [online], *Bynder*, available: https://www.bynder.com/en/blog/8-apis-bridging-todays-technology/ [accessed 23 Aug 2021].

Caballero, J., Grier, C., Kreibich, C., Paxson, V. (n.d.) 'Measuring Pay-per-Install: The Commoditization of Malware Distribution', 16.

Carroll, R. (2012) Fake Twitter Accounts May Be Driving up Mitt Romney's Follower Number [online], *the Guardian*, available: http://www.theguardian.com/world/2012/aug/09/fake-twitter-accounts-mitt-romney [accessed 21 Aug 2021].

Cashmore, P. (2009) Twitter Launches Verified Accounts [online], *Mashable*, available: https://mashable.com/archive/twitter-verified-accounts-2 [accessed 30 Aug 2021].

Chen, Z., Subramanian, D. (2018) 'An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter', *arXiv:1804.05232 [cs]*, available: http://arxiv.org/abs/1804.05232 [accessed 21 Aug 2021].

Corcoran, I. (2021) 'Measuring the Influence of Bots in Twitter: Detecting Bots and Measuring Their Influence'.

Czakon, J. (2019) F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? [online], *neptune.ai*, available: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc [accessed 28 Aug 2021].

Daya, A.A., Salahuddin, M.A., Limam, N., Boutaba, R. (2020) 'BotChase: Graph-Based Bot Detection Using Machine Learning', *IEEE Transactions on Network and Service Management*, 17(1), 15–29.

Delua, J. (2021) Supervised vs. Unsupervised Learning: What's the Difference? [online], available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning [accessed 21 Aug 2021].

Dickerson, J.P., Kagan, V., Subrahmanian, V.S. (2014) 'Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?', in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, Presented at the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), 620–627.

D'Onfro, J. (2021) Twitter Admits 5% Of Its 'Users' Are Fake [online], *Business Insider*, available: https://www.businessinsider.in/twitter-admits-5-of-its-users-are-fake/articleshow/23479699.cms [accessed 30 Aug 2021].

Doshi, S. (2021) Building a Safer Twitter [online], available: https://blog.twitter.com/en_us/a/2014/building-a-safer-twitter [accessed 21 Aug 2021].

Eiman, A., Nazar, Z., Mohammed, E.A., Hany, A. (2021) Detecting Social Bots on Twitter: A Literature Review [online], available: https://ieeexplore.ieee.org/document/8605995 [accessed 21 Aug 2021].

Ersahin, B., Aktaş, Ö., Kilinç, D., Akyol, C. (2017) 'Twitter fake account detection', 388–392.

Erşahin, B., Aktaş, Ö., Kılınç, D., Akyol, C. (2017) 'Twitter fake account detection', in *2017 International Conference on Computer Science and Engineering (UBMK)*, Presented at the 2017 International Conference on Computer Science and Engineering (UBMK), 388–392.

Facebook, Twitter, options, S. more sharing, Facebook, Twitter, LinkedIn, Email, URLCopied!, C.L., Print (2013) Twitter Adds Two-Step Verification Option to Help Fend off Hackers [online], *Los Angeles Times*, available: https://www.latimes.com/business/technology/la-fi-tn-twitter-two-step-verification-hackers-20130523-story.html [accessed 21 Aug 2021].

Facebook, Twitter, options, S. more sharing, Facebook, Twitter, LinkedIn, Email, URLCopied!, C.L., Print (2014) Op-Ed: Mining Twitter Gold, at Five Bucks a Pop [online], *Los Angeles Times*, available: https://www.latimes.com/opinion/op-ed/la-oe-0601-lotan-buying-followers-20140601-story.html [accessed 21 Aug 2021].

Feiner, E.J., Lauren (2020) Twitter Knocks down Bernie Sanders' Suggestion That Russian Trolls Are behind Online Attacks from His Supporters [online], *CNBC*, available:

https://www.cnbc.com/2020/02/20/twitter-knocks-down-sanders-suggestion-russian-trolls-behind-supporters.html [accessed 21 Aug 2021].
Feng, Y., Li, J., Jiao, L., Wu, X. (2019) 'BotFlowMon: Learning-based, Content-Agnostic Identification of Social Bot Traffic Flows', in *2019 IEEE Conference on Communications and Network Security (CNS)*, Presented at the 2019 IEEE Conference on Communications and Network Security (CNS), 169–177.
Free Cloud Computing Services - AWS [online] (2021) *Amazon Web Services, Inc.*, available: https://aws.amazon.com/free/ [accessed 30 Aug 2021].
Gannarapu, S., Dawoud, A., Ali, R.S., Alwan, A. (2020) 'Bot Detection Using Machine Learning Algorithms on Social Media Platforms', in *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, Presented at the 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), 1–8.
Gilbertson, S. (2011) Twitter Vulnerability: Spoof Caller ID To Take Over Any Account | Webmonkey | Wired.Com [online], available: https://web.archive.org/web/20110721211341/http://www.webmonkey.com/2007/04/twitter_vulnerability_spoof_caller_id_to_take_over_any_account/ [accessed 23 Aug 2021].
Google Colaboratory [online] (2021) available: https://colab.research.google.com/notebooks/intro.ipynb [accessed 26 Aug 2021].
Heidari, M., Jones, J.H.J., Uzuner, O. (2021) 'An Empirical Study of Machine learning Algorithms for Social Media Bot Detection', in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Presented at the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 1–5.
Heidenreich, H. (2018) What Are the Types of Machine Learning? [online], *Medium*, available: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f [accessed 21 Aug 2021].
Hendrick, D. (2013) 'Complete History of Social Media: Then And Now', *Small Business Trends*, available: https://smallbiztrends.com/2013/05/the-complete-history-of-social-media-infographic.html [accessed 22 Aug 2021].
Holt, K. (2021) Twitter Suspensions for Promoting Terrorism Drop yet Again [online], *Engadget*, available: https://www.engadget.com/2019-05-09-twitter-transparency-report-terrorism-child-exploitation.html [accessed 23 Aug 2021].
Home / Twitter [online] (2021) *Twitter*, available: https://twitter.com/home [accessed 21 Aug 2021].
Honan, M. (2021) Killing the Fail Whale With Twitter's Christopher Fry | WIRED [online], available: https://www.wired.com/2013/11/qa-with-chris-fry/ [accessed 21 Aug 2021].
Johnson, S. (n.d.) 'Twitter Inc (TWTR) Could Use Gamergate Autoblocker Model To Block Millions of Fake Accounts?', available: https://www.techinsider.net/twitter-inc-twtr-could-use-gamergate-autoblocker-model-to-block-millions-of-fake-accounts/1120221.html [accessed 21 Aug 2021].
Kantepe, M., Ganiz, M.C. (2017) 'Preprocessing framework for Twitter bot detection', in *2017 International Conference on Computer Science and Engineering (UBMK)*, Presented at the 2017 International Conference on Computer Science and Engineering (UBMK), 630–634.

Kastrenakes, J. (2020) Twitter Notifies Users That It's Now Sharing More Data with Advertisers [online], *The Verge*, available: https://www.theverge.com/2020/4/8/21213593/twitter-data-sharing-pop-up-mobile-app-advertising-settings [accessed 21 Aug 2021].

Kudugunta, S., Ferrara, E. (2021) Deep Neural Networks for Bot Detection - ScienceDirect [online], available: https://www.sciencedirect.com/science/article/abs/pii/S0020025518306248?via%3Dihub [accessed 21 Aug 2021].

Lakshmanan, R. (2019) Twitter Bug Accidentally Shared Location Data of Some IOS Users [online], *TNW | Security*, available: https://thenextweb.com/news/twitter-bug-accidentally-shared-location-data-of-some-ios-users [accessed 23 Aug 2021].

Lee, K., Eoff, B., Caverlee, J. (2011) 'Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter'.

Leyden, J. (2021) Twitter SMS Spoofing Still Undead [online], available: https://www.theregister.com/2009/03/06/twitter_sms_spoofing_risk/ [accessed 21 Aug 2021].

Li, Y.-F., Guo, L.-Z., Zhou, Z.-H. (2021) 'Towards Safe Weakly Supervised Learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), 334–346.

Lingam, G., Rout, R.R., Somayajulu, D.V.L.N. (2018) 'Detection of Social Botnet using a Trust Model based on Spam Content in Twitter Network', in *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*, Presented at the 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), 280–285.

Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., Mata-Sánchez, J.I. (2019) 'Contrast Pattern-Based Classification for Bot Detection on Twitter', *IEEE Access*, 7, 45800–45817.

Lyons, K. (2020) Twitter Is Fighting Election Chaos by Urging Users to Quote Tweet Instead of Retweet [online], *The Verge*, available: https://www.theverge.com/2020/10/9/21509439/twitter-election-trump-quote-tweet-labels-rules-election [accessed 21 Aug 2021].

Machine Learning: What It Is and Why It Matters | SAS India [online] (2021) available: https://www.sas.com/en_in/insights/analytics/machine-learning.html [accessed 21 Aug 2021].

Machkovech, S. (2021) Twitter Announces Sweeping Update to Reporting, Blocking Tools | Ars Technica [online], available: https://arstechnica.com/information-technology/2014/12/twitter-announces-sweeping-update-to-reporting-blocking-tools/ [accessed 23 Aug 2021].

Mashable (2021) Twitter Bug Lets You Control Who Follows You [online], available: https://mashable.com/archive/twitter-follow-bug [accessed 23 Aug 2021].

Matplotlib (2021) Matplotlib: Python Plotting — Matplotlib 3.4.3 Documentation [online], available: https://matplotlib.org/ [accessed 28 Aug 2021].

Mello Jr., J.P. (2014) 'Twitter Gives Harassed Users a Little Ammo', *TechNewsWorld*, available: https://www.technewsworld.com/story/twitter-gives-harassed-users-a-little-ammo-81442.html [accessed 23 Aug 2021].

MLeeDataScience (2021) Visual Guide to the Confusion Matrix [online], *Medium*, available: https://towardsdatascience.com/visual-guide-to-the-confusion-matrix-bb63730c8eba [accessed 28 Aug 2021].

Mousavi, S.H., Khansari, M., Rahmani, R. (2019) 'A Fully Scalable Big Data Framework for Botnet Detection Based on Network Traffic Analysis', *Information Sciences*, 512.
Moyer, E. (2021) Twitter Updates Its Rules for Users after Uproar over Rape, Bomb Threats [online], *CNET*, available: https://www.cnet.com/tech/services-and-software/twitter-updates-its-rules-for-users-after-uproar-over-rape-bomb-threats/ [accessed 21 Aug 2021].
Narkhede, S. (2021) Understanding AUC - ROC Curve [online], *Medium*, available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [accessed 28 Aug 2021].
Nield, D. (2021) How to Hide Your Followers & Who You Are Following on Twitter [online], *It Still Works*, available: https://itstillworks.com/12757927/how-to-hide-your-followers-who-you-are-following-on-twitter [accessed 21 Aug 2021].
Numpy (2021) NumPy [online], available: https://numpy.org/ [accessed 28 Aug 2021].
Owe, P.C. (2020) Mining Twitter Data [online], *Medium*, available: https://towardsdatascience.com/mining-twitter-data-ba4e44e6aecc [accessed 27 Aug 2021].
Pandas (2021) Pandas - Python Data Analysis Library [online], available: https://pandas.pydata.org/ [accessed 28 Aug 2021].
Paul, M. (2021) Ireland to Become Privacy Regulator for 300m Twitter Users [online], *The Irish Times*, available: https://www.irishtimes.com/business/technology/ireland-to-become-privacy-regulator-for-300m-twitter-users-1.2180137 [accessed 21 Aug 2021].
Perdana, R., Muliawati, T., Harianto, R. (2015) 'BOT SPAMMER DETECTION IN TWITTER USING TWEET SIMILARITY AND TIME INTERVAL ENTROPY', *Jurnal Ilmu Komputer dan Informasi*, 8, 20–26.
Perez, S. (n.d.) 'Twitter bug disclosed some users' location data to an unnamed partner', *TechCrunch*, available: https://social.techcrunch.com/2019/05/13/twitter-bug-disclosed-some-users-location-data-to-an-unnamed-partner/ [accessed 23 Aug 2021].
Porter, J. (2021) Twitter Bans 70,000 QAnon Accounts as Conservatives Report Lost Followers - The Verge [online], available: https://www.theverge.com/2021/1/12/22226503/twitter-qanon-account-suspension-70000-capitol-riots [accessed 23 Aug 2021].
Project Jupyter [online] (2021) available: https://www.jupyter.org [accessed 26 Aug 2021].
Rawnsley, A., Stein, S. (2021) 'No Evidence' for Bernie Sanders' Russian Bot Claim, Experts Say [online], available: https://www.thedailybeast.com/experts-call-bs-on-bernies-russian-bot-theory [accessed 21 Aug 2021].
Rodriguez, S. (2021) Twitter Adds Two-Step Verification Option to Help Fend off Hackers - Los Angeles Times [online], available: https://www.latimes.com/business/technology/la-fi-tn-twitter-two-step-verification-hackers-20130523-story.html [accessed 23 Aug 2021].
RohanBhirangi (2021) *Detection of Twitter Bots Using Machine Learning Classifiers* [online], available: https://github.com/RohanBhirangi/Twitter-Bot-Detection [accessed 21 Aug 2021].
Royal Pingdom » Twitter Growing Pains Cause Lots of Downtime in 2007 [online] (2021) available:

https://web.archive.org/web/20101229114042/http://royal.pingdom.com/2007/12/19/twitter-growing-pains-cause-lots-of-downtime-in-2007/ [accessed 21 Aug 2021].
Samuelsohn, D. (2021a) Twitter Trend: Fakes [online], *POLITICO*, available: https://www.politico.com/story/2014/06/twitter-politicians-107672 [accessed 21 Aug 2021].
Samuelsohn, D. (2021b) Twitter Trend: Fakes - POLITICOPOLITICOSearchSearchCloseBack ButtonSearch IconFilter Icon [online], available: https://www.politico.com/story/2014/06/twitter-politicians-107672 [accessed 30 Aug 2021].
Sarabu, H., Ahlin, K., Hu, A.-P. (2019) 'Graph-Based Cooperative Robot Path Planning in Agricultural Environments', in *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, Presented at the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 519–525.
Saurabh, K. (2021) Google Colaboratory [online], available: https://colab.research.google.com/drive/1JBz86kOnizO0usbsN2oMbrsDHOYrEi3V#scrollTo=BmdwdBMVXBCr [accessed 27 Aug 2021].
Scikit (2021) Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.24.2 Documentation [online], available: https://scikit-learn.org/stable/ [accessed 28 Aug 2021].
Shukla, P., Iriondo, R., Chen, S. (2021) Machine Learning Algorithms For Beginners with Code Examples in Python [online], *Medium*, available: https://pub.towardsai.net/machine-learning-algorithms-for-beginners-with-python-code-examples-ml-19c6afd60daa [accessed 30 Aug 2021].
Simple Guide to Confusion Matrix Terminology [online] (2014) *Data School*, available: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ [accessed 30 Aug 2021].
Soni, J. (2021) *Detecting Bots on Twitter Using Machine Learning* [online], available: https://github.com/jubins/MachineLearning-Detecting-Twitter-Bots [accessed 21 Aug 2021].
Stone, B. (2021b) Monday Morning Madness [online], available: https://blog.twitter.com/en_us/a/2009/monday-morning-madness [accessed 23 Aug 2021].
Stone, B. (2021a) Introducing the Twitter API [online], available: https://blog.twitter.com/en_us/a/2006/introducing-the-twitter-api [accessed 23 Aug 2021].
Twitter Account: Twitter Suspended over 1.6 Lakh Terror-Promoting Accounts in Six Months [online] (2019) available: https://web.archive.org/web/20190531194121/https://economictimes.indiatimes.com/magazines/panache/twitter-suspended-over-1-6-lakh-accounts-for-promoting-terrorism/articleshow/69268206.cms#close [accessed 23 Aug 2021].
Twitter Developers [online] (2021) available: https://developer.twitter.com/en/portal/dashboard [accessed 21 Aug 2021].
Twitter Has Suspended More than 166,000 Accounts Related to Promotion of Terrorism-Technology News, Firstpost [online] (20:20:36 +05:30) *Tech2*, available: https://www.firstpost.com/tech/news-analysis/twitter-has-suspended-more-than-166000-accounts-related-to-promotion-of-terrorism-6611591.html [accessed 23 Aug 2021].

Twitter Is Trying to Block Images of James Foley's Death [online] (2021) available: https://finance.yahoo.com/news/twitter-trying-to-block-images-of-james-foleys-death-95278352899.html?guce_referrer=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnLw&guce_referrer_sig=AQAAAJBOvRpjRKXUzyfhsfKnDlfXgypzhyIG75NyeWNniZvF3_SgZbgUcuzdrTXkrUEAUu4m63YNGAv5Kx7Vl40En0bY3YhYeiPSh948cKuSuoAj4UBN5BazZD1kbO1rLXKr7vGpfJITEQhVBSAkP3QtL44qyQ_YWydrhHLxHetblWw2 [accessed 21 Aug 2021].

'Twitter removes French anti-Semitic tweets' (2012) *BBC News*, 19 Oct, available: https://www.bbc.com/news/technology-20004671 [accessed 21 Aug 2021].

Twitter Reports Fall in Extreme Content [online] (2021) *SBS News*, available: https://www.sbs.com.au/news/twitter-reports-fall-in-extreme-content/e0c40ed0-f6e9-4a59-ab0f-508d5eedf3e9 [accessed 23 Aug 2021].

'Twitter's Tony Wang issues apology to abuse victims' (2013) *BBC News*, 3 Aug, available: https://www.bbc.com/news/uk-23559605 [accessed 21 Aug 2021].

Tyagi, A. (2021) *Bot-Detection* [online], available: https://github.com/AayushTyagi1/Bot-Detection [accessed 21 Aug 2021].

Vidya, A. (2020) 'Confusion Matrix for Machine Learning', *Analytics Vidhya*, available: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/ [accessed 28 Aug 2021].

Vidya, A. (2021) XGBoost Algorithm | XGBoost In Machine Learning [online], available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ [accessed 22 Aug 2021].

Wang, A.H. (2010) 'Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach', in Foresti, S. and Jajodia, S., eds., *Data and Applications Security and Privacy XXIV*, Springer Berlin Heidelberg: Berlin, Heidelberg, 335–342.

Wang, W., Meichan, Z., Zhenzhen, G., Guangquan, X., Hequn, X., Yuanyuan, L., Xiangliang, Z. (2021) Constructing Features for Detecting Android Malicious Applications: Issues, Taxonomy and Directions [online], available: https://ieeexplore.ieee.org/document/8720030/ [accessed 21 Aug 2021].

Wang, W., Shang, Y., He, Y., Li, Y., Liu, J. (2020) 'BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors', *Information Sciences*, 511, 284–296.

Waskom, M. (2021) 'seaborn: statistical data visualization', *Journal of Open Source Software*, 6(60), 3021.

Why Fake Twitter Accounts Are a Political Problem [online] (2021) available: https://www.newstatesman.com/sci-tech/2014/05/why-fake-twitter-accounts-are-political-problem [accessed 21 Aug 2021].

Wofford, T. (2014) One Woman's New Tool to Stop Gamergate Harassment on Twitter [online], *Newsweek*, available: https://www.newsweek.com/one-womans-new-tool-stop-gamergate-harassment-twitter-288008 [accessed 21 Aug 2021].

Woolley, E.M. (2021) Blocked on Twitter: Software's Limits in the Fight against Online Hate - The Globe and Mail [online], available: https://www.theglobeandmail.com/technology/digital-culture/blocked-on-twitter-softwares-limits-in-the-fight-against-online-hate/article21920082/ [accessed 23 Aug 2021].

Wortham, J. (2021) Twitter-Savvy Hackers Tweak the Twitterati - The New York Times [online], available: https://bits.blogs.nytimes.com/2009/01/05/twitter-hit-by-hacker-phishers/ [accessed 23 Aug 2021].

XGBoost Documentation — Xgboost 1.5.0-Dev Documentation [online] (2021) available: https://xgboost.readthedocs.io/en/latest/ [accessed 28 Aug 2021].

Yu, H., Gibbons, P.B., Kaminsky, M., Xiao, F. (2010) 'SybilLimit: A Near-Optimal Social Network Defense Against Sybil Attacks', *IEEE/ACM Transactions on Networking*, 18(3), 885–898.

Zhang, C., Wu, B. (2020) 'Social Bot Detection Using "Features Fusion"', in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Presented at the 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 626–629.

Appendices

Appendix A: Code Listing

The GitHub Repository link with the code is listed below:

https://github.com/srb7600/Twitter-Bot-Detection