Twitter Bot Detection

Dria Fabrizio
aufa6169@colorado.edu
Camden Funkhouser
cafu8688@colorado.edu
Tony Joo
Tony.Joo@colorado.edu

Abstract

Twitter bots have become a major talking point due to the influence they have in our society. In this paper we outline multiple examples of both academic work and private sector work in the effort to fight bots by detecting them. These examples were picked specifically to show a variety of effective ways of detecting Twitter bots using Artificial Intelligence. The results are quite interesting, both in the performance of these methods and in the variety. The key takeaways that we outline in this paper is that user account level data can be more accurately and easily predicted using machine learning methods, while tweet level data can be better predicted with deep learning models. The comparison of two companies we chose were not actually both companies, due to the nature of this field, instead we chose to focus on comparing Twitter's internal methods to an academic project called Botometer.

1. Introduction

The problem of social media bots has increasingly become a widespread problem for society as at large. Social media bots do more than just cause a nuisance, they can have huge societal consequences such as influencing elections. To begin, we will first discuss the nature of what bots are, what they do, and why they are a problem that needs to be addressed.

A simple definition of social media bots are automated programs used to engage in social media. They can mimic the behavior of human users through textual information such as tweets or actions or behaviors such as creating posts, sharing links, liking posts, following users, and commenting. There is a huge variety in the level of sophistication of twitter bots, and an equally huge variety in terms of purpose. While some are fully automated, others can have partial human control, and in some cases there are "Master of Puppets" scenarios where thousands of bots may be being controlled by one human. The level of sophistication of twitter bots can range from a simple account whose sole purpose is to follow and like posts to create nearly indistinguishable tweets which can fool people into believing they are human accounts.

The true problem of bots comes not from their individual capabilities but their power in large numbers. Huge amounts of bots acting in unison could artificially amplify the popularity of a person or movement. For example, it is not uncommon for people to purchase likes and followers on social media from gray market websites, which use bots to provide this service. Another danger is the deterioration of the democratic nature of the internet. During the Arab Spring movement in 2010-2012, governments used Twitter bots to deliberately overwhelm social media feeds to deliberately counter dissent online[1] which set a dangerous precedent in the use of social media manipulation by a governmental entity. But, this is also a problem here in the United States as well. During the infamous 2016 U.S. presidential election, research showed that as much as 20% of political discussion on social media was generated entirely by bots[3]. Perhaps the most prevalent form of bots most people are familiar with are spam bots. Spam bots are used for illicit advertising by spamming and promoting website links to commercial websites. There is also the category of financial bots, which are not the main focus of this paper, but are interesting in its own right. Financial bots scan and read twitter posts to do sentiment analysis in order to make trades on the stock market automatically. In 2013, a Syrian hacker group managed to gain control of the twitter account of the Associated Press[1], and tweeted fake news that a terrorist attack had occured on the White House and President Obama was injured, causing an instantaneous slide in the stock market. This shows just how powerful and dangerous twitter bots have become.

It is pretty clear that Twitter bots can be extremely harmful if left unchecked, but how can this be done? How are bots detected? Twitter claims that its internal bot detection system is working, however, they do not provide us with much information on how they are detecting them, and what information they give us about their methods seem relatively primitive. Luckily, academic researchers in this field can provide us with a peek behind the curtain, frequently making their code publicly available for transparency. In this paper, we will discuss some interesting methods in the field of twitter bot detection, which can be broken down into four distinct categories:

1. Human Classification
2. Statistical Methods
3. Machine Learning
4. Deep Learning

Now, let us look into these methods that researchers have created in the past few years to combat the abundance of bots on Twitter, how Twitter itself says it is mitigating the problem including recent developments, and uncertainties with the platform in light of its recent purchase by Elon Musk.

2. Research

2.1. Twitter Research

Twitter is taking some measures to detect bot accounts internally, in an effort to stop the flow of misinformation and spam. The company's focus has been on being more proactive in identifying problematic accounts and behavior instead of waiting until a report is filed and received. The firm's investments in this space are having a positive impact, however, with many improvements made to the machine learning algorithms over time.

One of the ways that Twitter is combating the bot problem is to reduce the visibility of suspicious accounts in Tweet and account metrics. To achieve this the company has started implementing a process of “updating account metrics in near real time”[9]. This will help prove which accounts seemingly stumble upon hundreds and thousands of followers and likes overnight and which accounts have authentically generated interest. Initially the bot detection process incorrectly classified legitimate accounts as spam, with a sometimes lengthy appeals process, now a two-factor authentication method is employed to alleviate the appeal queue.

Another way that Twitter is fighting spam is by improving the company’s sign up process. “To make it harder to register spam accounts, we’re also going to require new accounts to confirm either an email address or phone number when they sign up to Twitter”[9]. Incorporating this two-factor authentication method in the sign up process certainly helps ensure that a real person is creating the account. Even though these accounts and numbers can be faked, implementing this process will slow down the quantity of fake accounts created.

More steps that are being taken in the fight include the audit of existing accounts for signs of automated sign-up. The business wants to take action to be proactive and “ensure that every account created, passes some easy, automatic security checks designed to prevent automated signups”[9]. By having an audit this will help Twitter determine a list of potential fraudulent and/or bot accounts created in sign up. This can potentially nip the problems in the bud in some cases.

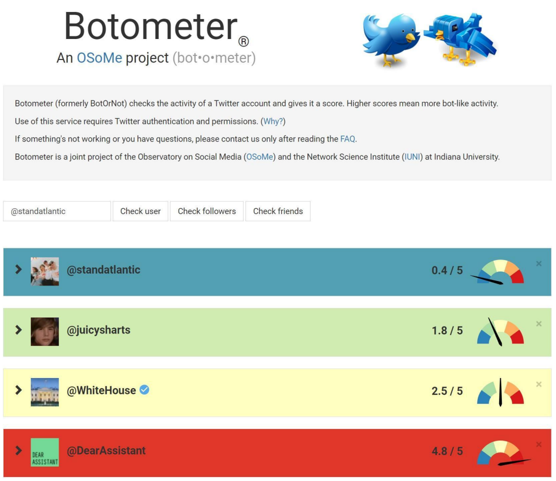
The last action that Twitter is taking to stop fraud is the expanding of the firm's malicious behavior detection systems. Some of the measurable actions include automating some processes where suspicious activity is spotted. Examples of suspicious activity include: “exceptionally high-volume tweeting with the same hashtag, or using the same @username without a reply from the account you’re mentioning”[9].

There are some steps that an individual user can take to improve their security on their own that Twitter recommends. These include “enabling two-factor authentication, regularly reviewing any third-party applications, not reusing passwords, U2F security keys”[9]. These are all actionable steps an individual can take to improve their privacy and security.

This problem is not going away and has actually evolved and become worse since this article was written. However Twitter had a plan for the future of their security involving bot detection. The firm said that, “going forward, Twitter is continuing to invest across the board in our approach to these issues, including leveraging machine learning technology and partnerships with third parties”[9]. Most of these actions were taken in response to the overwhelming quantity of bot and spam accounts that were prevalent on Twitter during the 2016 presidential election. This same style of mis information bot account spam campaign is being currently employed in the conflict in Ukraine and it will be interesting to see how Twitter addresses the problem going forward.

2.2. Botometer

Botometer is a machine learning based twitter bot detection project created by faculty from Indiana University Bloomington, USA. The paper entitled “Botometer 101: Social bot practicum for computational social scientists” by Emilio Ferrara[10] *et al.* details and outlines the machine learning methods utilized behind the scenes for their public user interface Botometer.com. The way Botometer works on the surface level for users of the site is that the user inputs the @username of someone they want to test whether it is a bot or not. The algorithm from Botometer then scrapes through the user account and posting metadata of that user and runs it through their algorithm. As can be seen in figure 1 below, the user interface shows blue colors and a low number close to zero when the algorithm believes it is a human controlled account, and red with a high number close to five for an automated account.



Botometer®

An OSoMe project (bot•o•meter)

Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score. Higher scores mean more bot-like activity.

Use of this service requires Twitter authentication and permissions. (Why?)

If something's not working or you have questions, please contact us only after reading the FAQ.

Botometer is a joint project of the Observatory on Social Media (OSoMe) and the Network Science Institute (IUNI) at Indiana University.

@standatlantic | Check user | Check followers | Check friends

@standatlantic	0.4 / 5
@juicysharts	1.8 / 5
@WhiteHouse	2.5 / 5
@DearAssistant	4.8 / 5

Figure 1: The machine learning method that Botometer uses, as outlines in the publication

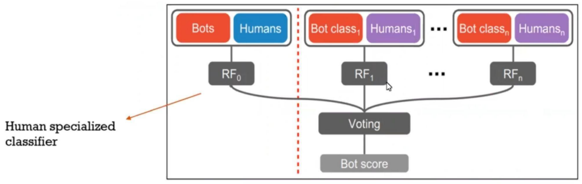


Figure 2: Ensemble of Random Forest from Botometer

Botometer was created by faculty from Indiana University and has the capability to detect if a user is a bot by inputting a username into their system. They use multiple random forests, each with 100 trees, and employ 1209 Features.It then outputs a bot score between 1 and 5, the higher the number, the more certain the model is that it is a bot. Using five-fold cross validation they were able to achieve an AUC of 0.99. Each random forest model gives a score between 0 and 1, 0 for human, 1 for bot; this is known as an Ensemble of Special Classifiers (ESC)

2.3. LSTM

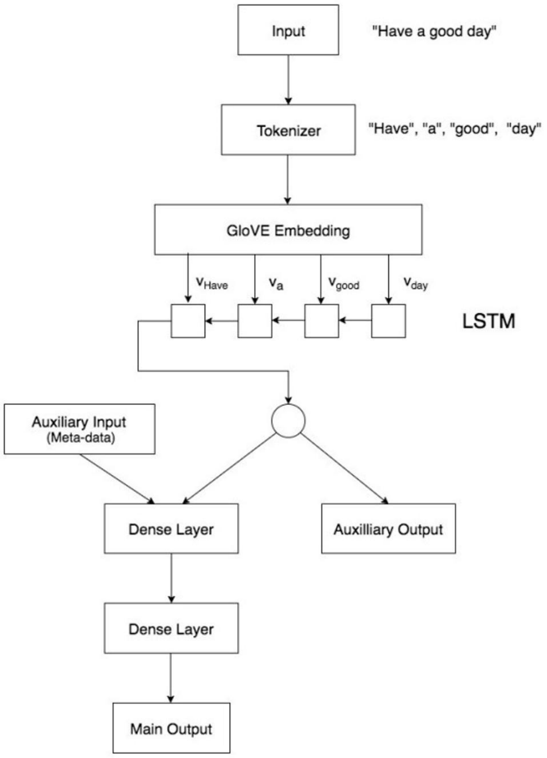
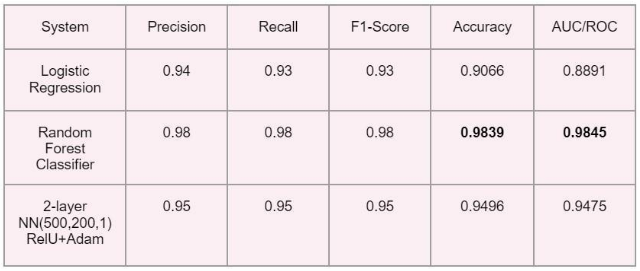
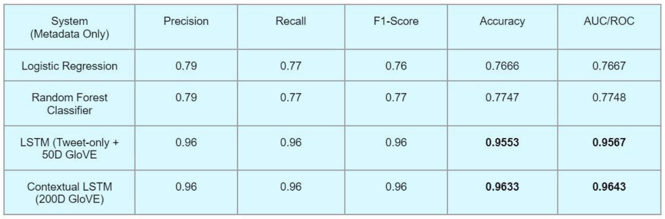


Figure 3: Flow Diagram of Contextual LSTM Model



System	Precision	Recall	F1-Score	Accuracy	AUC/ROC
Logistic Regression	0.94	0.93	0.93	0.9066	0.8891
Random Forest Classifier	0.98	0.98	0.98	**0.9839**	**0.9845**
2-layer NN(500,200,1) RelU+Adam	0.95	0.95	0.95	0.9496	0.9475

Figure 4: Model results at the account-level (user); data taken from[10], and formatted by the authors of this paper.



System (Metadata Only)	Precision	Recall	F1-Score	Accuracy	AUC/ROC
Logistic Regression	0.79	0.77	0.76	0.7666	0.7667
Random Forest Classifier	0.79	0.77	0.77	0.7747	0.7748
LSTM (Tweet-only + 50D GloVE	0.96	0.96	0.96	**0.9553**	**0.9567**
Contextual LSTM (200D GloVE)	0.96	0.96	0.96	**0.9633**	**0.9643**

Figure 5: Model results at the tweet-level; data taken from[10], and formatted by the authors of this paper.

In their 2019 study, “Identifying Twitter Bots Using a Convolutional Neural Network”[2], Michael Farber, Agon Qurdina, and Lule Ahmedi took a different approach to bot detection than Kudugunta and Ferrara by taking advantage of the natural language processing capabilities of convolutional neural networks, experimenting with different architectures, until they were able to settle on a model that performed relatively well with a test accuracy of 90.34%. Where Kudugunta and Ferrara considered both the content of a user’s tweets, and metadata about the account[10], Farber et al. focused solely on the tweets themselves. Additionally, instead of looking at individual tweets from any given account, they combined all of a user’s tweets into a single input for their dataset.

When experimenting with the architecture of their CNN, Farber et al. discovered that reducing the amount of features by adding one dimensional MaxPooling layers after each convolutional layer improved model performance. They created two models using this architecture, one using a layer size of two, and the other using a layer size of four, the second of which had a higher validation accuracy than the first. The authors attributed this to having a relatively small dataset, making a highly complex model with millions of trainable weights lead to overfitting. They then further tweaked this smaller model by trying different types of word embeddings, and ultimately settled on a 300-dimensional word2vec embedding.

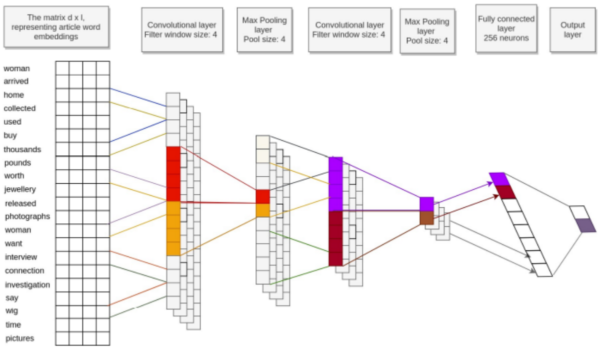

The matrix d x l, representing article word embeddings
Convolutional layer Filter window size: 4
Max Pooling layer Pool size: 4
Convolutional layer Filter window size: 4
Max Pooling layer Pool size: 4
Fully connected layer 256 neurons
Output layer
woman
arrived
home
collected
used
buy
thousands
pounds
worth
jewellery
released
photographs
woman
want
interview
connection
investigation
say
wig
time
pictures

Figure 6: The architecture of Farber et al’s CNN, courtesy of their paper[2].

3. Research on two Competitors

Currently, the two Twitter bot detecting services we have discussed (Botometer and Twitter’s internal system) both rely on machine learning and the behavior of accounts, rather than the content that these accounts produce. Meanwhile, researchers are developing well performing models for bot detection using deep learning techniques, which use natural language processing to analyze users’ tweets. This suggests that the current services available for detecting bots may be lacking, and more sophisticated methods may need to be put in place. This comes with a caveat, though, which is that there are not many of these services, nor is there much in the way of research into deep learning methods for this particular task as it is a relatively new field of study to address a relatively new issue with social media.

Additionally, since each of these models use different data, it is difficult to compare these models. Because of this, we cannot definitely conclude that there is one prevailing method for bot detection. However, based on the information we have, it appears that Kudugunta and Ferrara’s model, which takes a hybrid approach of looking at user activity and tweet content, has the potential to significantly reduce the amount of bots on Twitter.

We also hypothesize that the reason both Twitter and Botometer relies on machine learning algorithms such as Random Forests is that they rely more heavily on user based data. Twitter of course has access to the highest quality data available, and the ESC method of using many specialized classification random forests works incredibly well on this type of data as well. Although Neural Networks can work better on tweet level data, perhaps these companies are relying on the quality of user based data to achieve high results with the advantages in speed and computing power of less complex machine learning algorithms

4. Recommendation (Regulator role)

As a regulator this is a very challenging problem. While these types of algorithms seem great for society as a whole since they are fighting misrepresentation and bias, the algorithms themselves could be introducing misrepresentation and bias as well. The issue of regulating twitter bot detection is essentially part of the broader issue of transparency for social media sites such as twitter. There has been a growing public distrust of Twitter and the way they use their algorithm, some claiming a political bias and others claiming a problem with misinformation.

Recently the broader topic of Twitter’s algorithm has been thrust into the spotlight by Tesla’s buyout of Twitter. “Some of Twitter's technology is already open source, which means it's publicly available for anyone to view, rework, and use for other purposes. But what Musk was asking, essentially, was whether the rules that computers follow to determine what you see in your Twitter (TWTR) feed should be public, too. Over a million votes were tallied by the time the poll closed, with an overwhelming amount of "yes" votes (82.7%).”

The purpose of regulation is to ensure fairness, trust, and transparency for society, but the issue with simply exposing the algorithms is that only a handful of people could truly understand what the code and algorithms are doing. Even practitioners would have a tough time given the black box nature of deep learning methods. Therefore, we propose to go one step further and not only require transparency of the algorithms, but to also be required to show interpretable models for people outside of the company to actually understand what the algorithms are doing. By requiring companies whether it's bot detection companies or large platforms such as twitter to show interpretable machine learning models to the public this would give potential clients and consumers a better understanding of what the algorithms are actually doing.

The main argument in opposition to regulations of this type is that private companies should have the right to control their product in the way they see fit, whether for competitive advantages or simply to appeal to their own target demographics. As the regulations stand today, there is no requirement for transparency of algorithms from any company, and realistically it would be a huge challenge for a bill of this type to get passed in the real world.

However, for the sake of this paper, we are going to discuss the particular issues with transparency that we have seen among the models we have previously mentioned, and provide potential solutions. Across all of the models discussed in this paper, Botometer was the only one that had all of its data publicly available. As such, we cannot make assumptions about the quality of the data, or what types of bots the models were trained on. This is concerning because a model that seems to perform well may suffer if faced with bots that are more sophisticated in their ability to mimic human behavior than those that the model was trained on. Additionally, while we were excited by the performance of the LSTM model, we soon realized that they did not explicitly explain how they were able to combine contextual user data into the dense layer of their model with the tweet data. Without this information, it is difficult for others to replicate the model and confirm that it works as intended.

From our look into Twitter's internal bot detection system, we know that Twitter does not provide much information about the methods used, nor were we able to find any official summary statistics indicating how well the system works. We can intuit that machine learning was used based on the information we were able to find on the website that said user behavior was used for bot detection in the same way that Botometer's algorithm is. But that still does not tell us what specific sort of algorithm is being used, nor does it provide us with insight into the performance of the algorithm.

Moreover, while the LSTM, CNN, and Botometer model provided us with some summary statistics, as shown in figures 4 and 5, but without providing us a full confusion matrix, we are still left with questions as to where the model makes mistakes. If the model tends towards false positives, this means that legitimate, human operated accounts are being flagged as bots, which could be considered an infringement on free speech. While Twitter does have an appeals system for misclassified accounts, that process involves the flagged accounts being reviewed by humans, which could become a slow and frustrating process if too many accounts are misclassified. On the other hand, if the model tends towards false negatives, that means that bots are slipping through the cracks, which as we have mentioned earlier in this paper, can have negative effects that are detrimental to society.

Our recommendations to mitigate these issues are twofold: making systems more transparent, and making them more interpretable. For making systems more transparent, we recommend making all data and code publicly available, and providing confusion matrices as well as measures of performance. For interpretability, we recommend the implementation of Python packages such as ELI5 (Explain Like I'm 5) for machine learning models, or LIME (Local Interpretable Model-Agnostic Explanations) for deep learning models. Additionally, since the general public should not be expected to understand the code, we suggest that bot detecting services write brief summaries of what their models are doing in layman's terms.

5. Conclusion

In this paper we have discussed what bots are, what they are capable of, and why they need to be actively stopped. Furthermore, we introduced the four general categories of detecting bots: humans, statistics, machine learning, deep learning. We discussed four specific research sources, Twitter's internal bot detection methods, Botometer, Contextual LSTM, and CNN.

The key takeaways are firstly that power in numbers of twitter bots have huge implications for society, and need to be actively stopped. But, even after analyzing many different successful methods, there is no one size fits all solution. For example, the results from the paper *Deep Neural Networks for Bot Detection* showed that these simpler machine learning methods can actually perform better than neural networks when it comes to using user account level data. But, when looking at contextual text data from the tweets themselves, neural networks have a huge advantage.

This does not mean that neural networks are the clear winner, as some of the downsides of neural networks are that they take more computer power and time, and given the nature of the huge amounts of data generated by bots this can become a real hindrance. It is also important to note that when comparing Accuracy, AUC/ROC, and other performance metrics, different models cannot at all be compared to each other since the datasets they used are completely different.

In our comparison between Twitter's internal methods and Botometer, it was noted that both these organizations use (relatively) simple machine learning algorithms. Given what we learned about how machine learning algorithms work on account level data, we hypothesize that given Twitter themselves have the most amount of high quality metadata of a user, they might be relying on these simpler machine learning methods due to the effectiveness and speed when using the high quality data that they have. It is of course possible that they are employing deep learning as well, but it is unclear whether this is the case as there is no mention of it published for the public.

In conclusion, the cat and mouse battle between may never end, due to the nature of the ever increasing sophistication of AI powered Twitter bots. As detection models get more sophisticated and further developed through research, these methods could also be applied towards the creation of bots as well. The problem of bots is not limited to the Twitter platform, of course. All social media platforms have their own types of bots as well. But, given all of the methods that we have discussed, we are optimistic that at the very least the low level unsophisticated bots will perhaps see a decline in the future, and there are many methods and angles to effectively tackle the problem of bot detection.

6. References

[1] Yarno Ritzen, Aljazeera, *It exists to demobilize opposition: How Twitter fails Arabs:* https://www.aljazeera.com/news/2019/7/16/it-exists-to-demobilise-opposition-how-twitter-fails-arabs

[2] Michael Farber, Agon Qurdina, Lule Ahmedi, University of Prishtina, Kosovo, *Identifying Twitter bots using a Convolutional Neural Network, 2019:* http://ceur-ws.org/Vol-2380/paper_227.pdf

[3] Cloudflare, *What is a social media bot?,* https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/

[4] Emilio Ferrara, *The Rise of Social Bots, 2016:*https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext

[5] Stefan Wojcik, *5 things to know about bots on twitter* https://www.pewresearch.org/fact-tank/2018/04/09/5-things-to-know-about-bots-on-twitter/

[6] Jeanna Smialek, Time Inc., *Twitter bots may have boosted Donald Trump's Votes by 3.23%, Researchers Say, 2018:* https://time.com/5286013/twitter-bots-donald-trump-votes/

[7] Adam Rowe, *Bots are Impersonating Twitter Users for Paypal and Venmo Scams, 2021:* https://tech.co/news/bots-are-impersonating-twitter-users-for-paypal-and-venmo-scams

[8] Maria Temming, How Twitter bots get people to spread fake news, 2018: https://www.sciencenews.org/article/twitter-bots-fake-news-2016-election

[9] Del Harvey, Yoel Roth, *How Twitter is fighting spam and malicious automation,* 2018: https://arxiv.org/pdf/1802.04289.pdf

[10] Sneha Kudugunta, Emilio Ferrar, *Deep Neural Networks for Bot detection, 2018:* https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation

[11] Kai-Cheng Yang, Emilia Ferrara, Filippo Menczer, Indiana University, *Botometer 101: Social Bot practicum for computational social scientists,* 5 Jan 2022: https://arxiv.org/pdf/2201.01608.pdf