Twitter Bot Behavior: How Twitter Bots Interact With People

Alic Szecsei
University of Iowa
alic-szecsei@uiowa.edu

Willem DeJong
University of Iowa
willem-dejong@uiowa.edu

ABSTRACT

Twitter bots are often cited as affecting the political process by manipulating the trending topics data; similar behavior is also cited on other social platforms, such as Facebook or Instagram. We present our use of unsupervised machine learning, combined with Indiana University's *BotOrNot* service, to classify Twitter users as bots based on statistical analysis of their accounts, and then examine the ways in which they interact with other users. Determining how these bots interact with human users can help to focus bot-detection algorithms to target those bots that interact with human users in malicious ways.

1. INTRODUCTION

1.1 Background & Motivation

Social bots, also known as *sybil accounts*, are programs that automate interaction on social platforms. While some may simply be humorous or helpful accounts that don't attempt to hide their status as bots, others have more manipulative goals; they may flood a social network with spam, or attempt to more subtly influence the thoughts and behavior of the humans it interacts with. While social networks are extremely effective at causing social change and improving the quality of life of their users, they are also at risk of automated manipulation by bots.

Aral and Walker (2011) showed that social networks are highly effective at manipulating the public[1], and the automation of such behavior only increases this efficiency. In addition, Ratkiewicz (2011) showed that political bots actively manipulated the 2010 U.S. midterm elections[7].

1.2 Problem Statement

While there have been multiple approaches to bot detection[8, 10, 4], these have been restrained to simple detection. Very few have attempted to examine the ways that these fake accounts interact with real users. Our goal is to find a number of bot accounts and determine how they use social media to affect their target users. Determining which bots are attempting to manipulate social networks and which are providing services to human users is an import aspect of bot detection, and one we believe can be improved by examining how malicious bots interact with human users.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

1.3 Proposed Approach

In this paper, we use data from *BotOrNot*, a bot-detection service run by Indiana University, to determine whether or not users are bots. We then pull their latest tweets, as well as user data, and use the collected data in an unsupervised machine learning algorithm to cluster the users into 50 groups. We then take the data for each cluster and analyze common behavioral patterns.

1.4 Key Results

We found that bots tend towards extreme behaviors when interacting with humans: either they do not interact with other users through a specific vector, or they exhibit no moderation in doing so, while most human users tend towards a more moderate engagement across all aspects of the platform.

We also verified *BotOrNot*'s bot detection process, while presenting an unsupervised machine learning classification system to simplify behavioral analysis by grouping similar categories of users and bots together based on *BotOrNot*'s category subscores.

2. RELATED WORK

Davis (2016)[5] and Dickerson (2014)[6]. [TODO: Talk about cited papers, what their results were, how those results were relevant to our data] Lorem ipsum lorem ipsum ipsum lorem lorem ipsum.

Lorem ipsum lorem ipsum

Lorem ipsum lorem ipsum.

Lorem ipsum lorem ipsum.

Lorem ipsum lorem ipsum.

3. PROPOSED APPROACH

Many examinations of bot behavior on Twitter use ground truths created by verified accounts. However, the social behavior of verified Twitter accounts is wildly different from that of the general public. Verified users often have a celebrity status, and so are less likely to retweet other users, and usually will not have a small number of followers.

In addition, verified Twitter accounts occasionally belong to people who exhibit bot-like behavior, advertising their services without much variety between tweets, and consistently linking to their personal websites. While these users may be verified, they are not guaranteed to be run by real people, and are often linked to other services to simply tweet links.

Instead, we chose to start with Twitter accounts we knew or who followed our personal accounts, and then attempt to provide a more detailed classification system to account for these "verified bots."

To retrieve a list of human and bot Twitter accounts, we compiled an initial list of 113 users from the followers of our personal Twitter accounts and manually determined whether or not they were being automated. This initial list had an approximately even split between humans and bots. We then retrieved data for their followers, and their followers' followers, leaving us with 9,025 Twitter users, which we then classified.

3.1 Clustering

To analyze data for a large number of users, we sought to cluster these accounts into a relatively small number of users with similar characteristics. For example, organizational accounts that may have bot-like behavior but provide a valuable service to the social network and are therefore followed by a large number of users should be classified as a different type of bot than a typical spambot. To perform this clustering, we used data from Indiana University's *BotOrNot* service.

BotOrNot analyzes a large amount of data retrieved from each user, including sentiment analysis and a temporal analysis to determine when users are likely to tweet. These different analyses are separated into six different category subscores: *Friend*, *User*, *Content*, *Temporal*, *Sentiment*, and *Network*[5]. Using machine learning classifiers, it assigns a score to a user, with higher scores indicating a larger amount of bot-like behavior.

BotOrNot separates its data into these six categories and performs machine learning classification on each category to determine its subscore. It also analyzes all of its data in a separate process to determine the overall score. Because these are separated, an account's subscores may not agree with its overall score; as an example, an account with all of its category subscores lower than another account may actually have a higher overall *BotOrNot* score.

To ensure that using *BotOrNot* would provide legitimate analysis of Twitter users, we used a small sample to validate its results. This lead us to discover a number of inconsistencies with *BotOrNot*'s overall score assignments. For example, a known Twitter bot was given a lower score than the personal account that the bot was attempting to imitate. Organizational accounts, such as the one belonging to the President of the United States, were often given a high *BotOrNot* score, which is a limitation that the official *BotOrNot* website discloses. One poorly-categorized account, the son of another user, had only made 3 tweets and had a *BotOrNot* score of over 90%.

After discovering these issues, we determined that more information was required for automated analysis of Twitter accounts. Using unsupervised machine learning to cluster accounts enabled us to successfully organize a large number of accounts into separate categories, which were then manually classified and verified. Following the recommendations of Bessi and Ferrera[2], we retrieved the most important descriptors of bots: whether they're using the default Twitter avatar and header image, and their retweet-to-tweet ratio, in addition to the *BotOrNot* score and category scores. While the number of retweets a user has is not readily available, it can be determined whether individual tweets on the user's timeline are retweets or original content. By retrieving the latest tweets of a user and calculating how many retweets were present, we can approximate that user's retweet-to-tweet ratio. Our best clustering results, however, were found when simply clustering based on a combination of the overall *BotOrNot* score and category subscores.

To classify each cluster, we set up a basic Python script using Selenium that displayed a sample set of Twitter feeds, and allowed a human classifier to submit a category for that Twitter user. Based on the overall classification associated with the cluster, we could determine what type of Twitter account a user who also belonged to that cluster was likely to be. We manually categorized 1,000 of these accounts, split evenly among each cluster, to verify both our clustering and *BotOrNot*'s overall scoring. This manual classification allowed us to verify whether or not the averaged *BotOrNot* score for each cluster was an accurate descriptor for the entire cluster.

3.2 Tweet Analysis

We examined the possible ways that bots can engage with human users on the Twitter platform, determining that these vectors consisted primarily of:

- Mentioning a user in a tweet

- Retweeting a user
- Using a popular hashtag or phrase
- Following a user
- Favoriting another user's tweet

We collected data about how many accounts each account was following, how many accounts followed them, and how many tweets each account had favorited. This user data, averaged across a cluster, gave insight into how each cluster of users tended to use the Twitter platform.

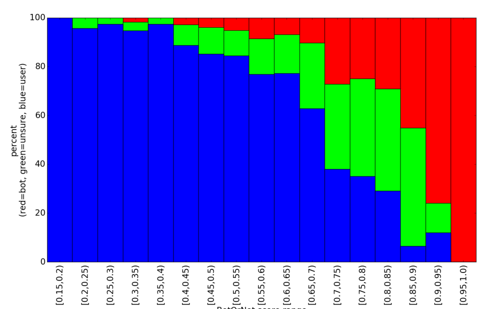
In addition to retrieving this user data, we also obtained the latest tweets made by each account, with a total of 625,053 tweets. We could then determine whether the tweet had been retweeted from another user, contained links, mentioned other users, and which hashtags were used.

Using these social behaviors, we were able to determine how each cluster and category of user tended to interact with other users. Focusing bot detection on these interactions could result in improved efficiency for spam removal services or other bot-related studies.

4. RESULTS & DISCUSSION

Examining the correctness of *BotOrNot* scoring was the first part of the analysis performed. We manually determined whether a number of accounts were bots, humans, or indeterminate. For the majority of accounts with a *BotOrNot* score over 50%, either we were unable to conclusively determine if the account was automated, or we determined that the account was definitely a bot. In addition, accounts with a *BotOrNot* score less than 50% were almost entirely found to be human users, as seen in Figure 1.

Figure 1: Manual identification of accounts



We next examined how an account's overall *BotOrNot* score aligned with the average of its corresponding category subscores, confirming that, with some deviations, the *BotOrNot* score was a better predictor of whether or not a user's account was automated. From Figure 2, we see that classifying the accounts based on the overall *BotOrNot* score is more effective than using the averaged category subscores.

Our unsupervised machine learning algorithms were able to similarly separate bot users and human users, as seen in Figure 3. Clusters are visible on the same horizontal line, and these clusters tend to be either mainly bots or mainly humans.

Figure 2: *BotOrNot* Scores vs Category Subscores

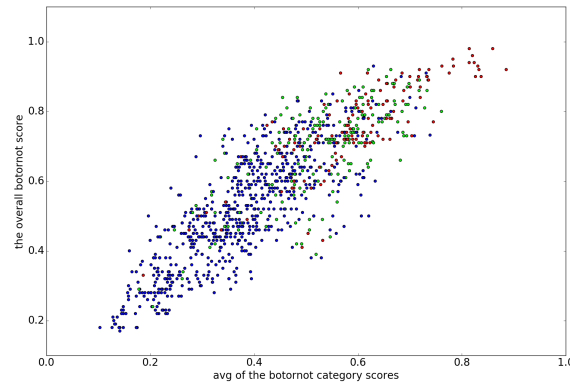
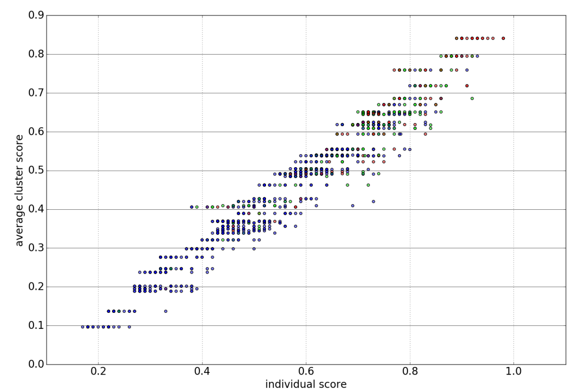


Figure 3: *BotOrNot* Clustering vs Average Scores



4.1 Mentions

As seen in Figure 4, we found that Twitter bots tended to not mention specific users, with most bot clusters having fewer than 17 mentions per user. Twitter's guidelines for bots are particularly explicit about this aspect of the service:

> If your application creates or facilitates automated reply messages or mentions to many users, the recipients must request or otherwise indicate an intent to be contacted in advance.[9]

In Figure 5, we examine the amount that accounts in clusters that mention users actually do so. If bots tended towards a similar rate of mentioning users as humans did, we would expect to see the data remain similar across clusters from Figure 4 to Figure 5. However, since bots tend to mention at a higher frequency when they do mention other users, clusters containing bots have comparatively higher mentioning rates in Figure 5. In clusters of humans, the mentioning characteristics of those users is more uniform. In contract, in the bot clusters, users tend to have either a medium amount of mentions, or none at all.

4.2 Retweets

Category	[0, 20%)	[20%, 40%)	[40%, 60%)	[60%, 80%)	[80%, 100%]
Human	10	150	277	204	21
Bot	0	1	17	70	53
Indeterminate	0	5	37	114	41

Table 1: Number of manually classified accounts within *BotOrNot* score ranges

Figure 4: Mentions Per Cluster

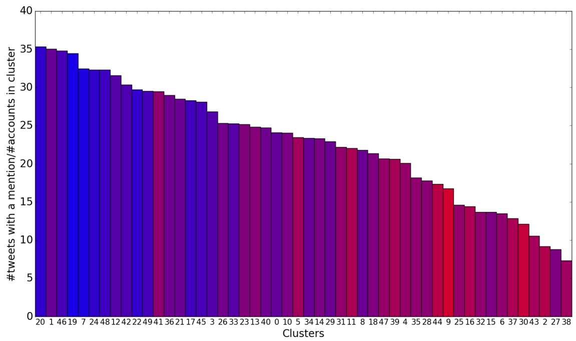


Figure 6: RTs Per Cluster

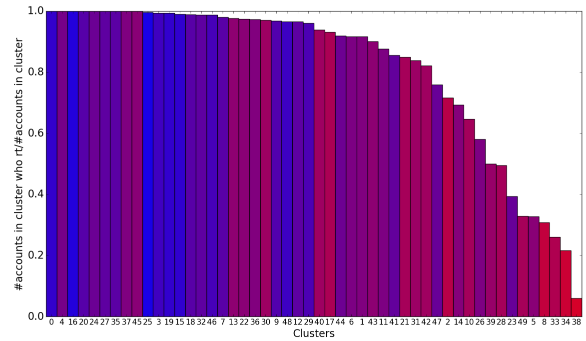
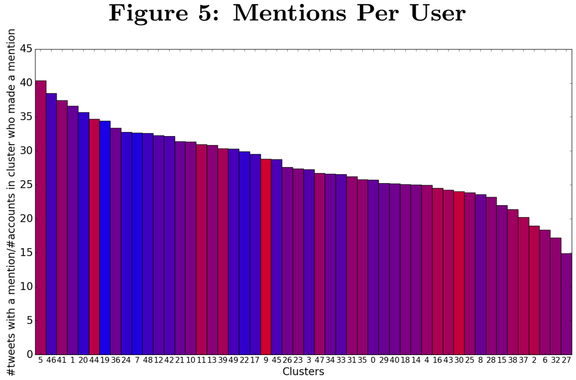


Figure 5: Mentions Per User



Similar to automated mentions, Twitter's guidelines explicitly forbid automated retweeting:

> Automation of Retweets often leads to spam and other negative user experiences; therefore, Retweeting in a bulk or automated manner is prohibited.[9]

Very few users classified as bots mention other users or retweet their tweets, which indicates that Twitter is closely monitoring these methods of user interaction; either very few bots are being created to automate these actions or they are rapidly banned from the service.

4.3 Hashtags

Due to some selection bias for the Twitter accounts we scanned, a disproportionate number of users tweeted using hashtags related to a shared interest, such as independent game development.

However, examining how many times each user in a cluster used a hashtag, as shown in Figure 7, gave a more indicative view of hashtags used for spamming, such as "fifa15coins" and a number of sexually explicit hashtags. Many of these spammed hashtags were only tweeted by a single user, which indicates that the creators of these spam accounts attempt to avoid overlap in which hashtags they are spamming. While Twitter's terms of service forbids automatically posting into the trending topics, it does not forbid automation using hashtags, allowing for these bots to find commonly popular hashtags and spam them without much apparent risk.

A large number of users tweeted with the hashtag of "Finances," but under further inspection a majority of these accounts were verified, indicating that while these accounts may exhibit spamming behavior, they were likely controlled by humans.

4.4 Following

While Twitter forbids automated following, it's one of the most common methods bots use to gain attention from human users. Most bots are following over 10,000 users; they also have similarly exaggerated numbers of followers.

We noticed one cluster of users sharing an extremely large number of both followers and accounts they followed; upon further inspection, the majority of these accounts were promotional accounts. Since they tend to follow back, these accounts become a prime target for sockpuppet accounts that want to appear legitimate; both the following and follower tabs are filled with users who have the default Twitter profile picture and no tweets, with many of their creation dates within the last month.

4.5 Favoriting

Examining how often users favorited tweets gave an interesting trend; bots tended to either favorite very few tweets, or a very large number of them, with human users hovering

Figure 7: Hashtags Per User

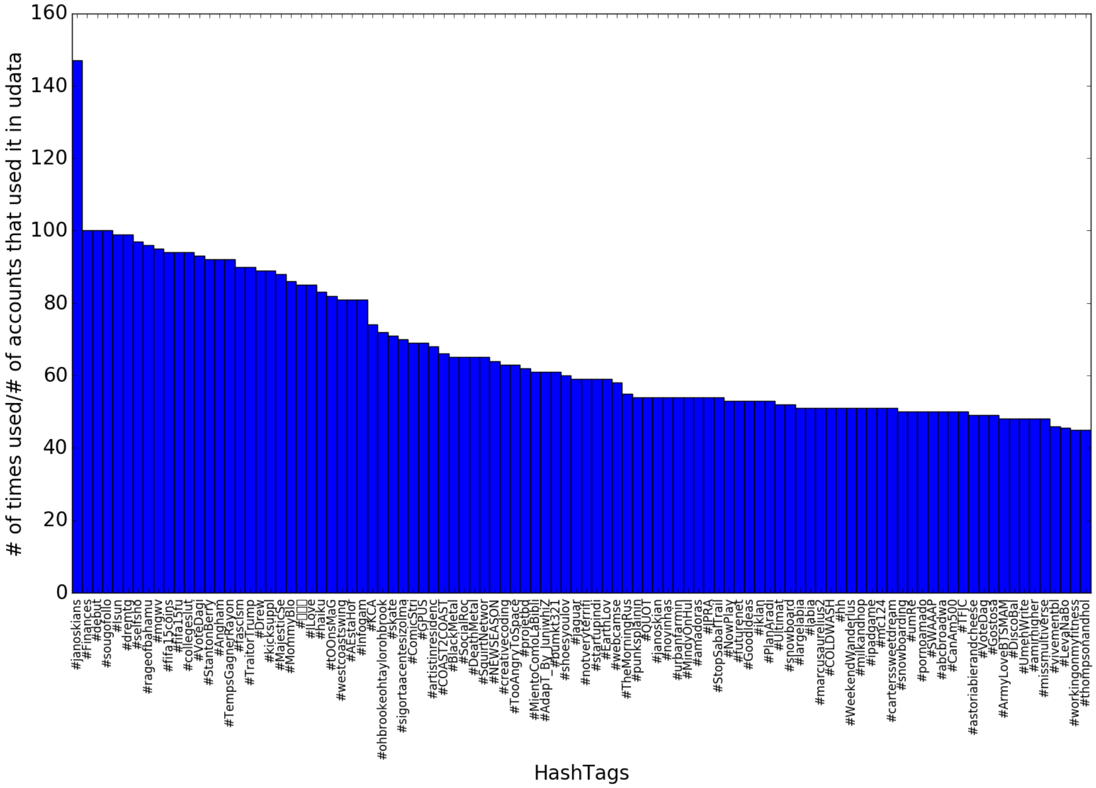
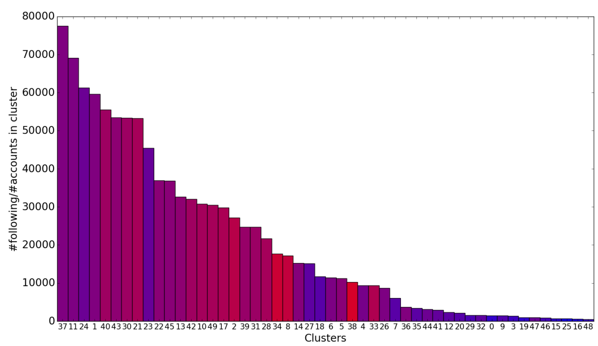


Figure 8: Following Per Cluster

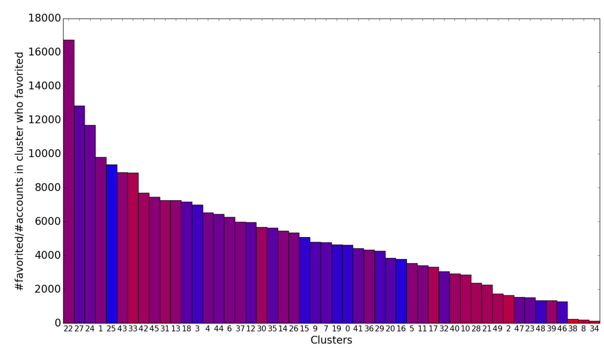


in between the two extremes.

We manually examined the users in clusters 22 and 33. Cluster 22 had the largest number of favorites, while cluster 33 favorited a large number of tweets and scored rather heavily in the *BotOrNot* assessment. Cluster 22 was comprised mainly of businesses and other Twitter "personalities" such as YouTubers. These users likely either search their names and favorite tweets including that text, or favorite the tweets that mention them as a way of quickly responding to fans.

Cluster 33, however, was comprised of accounts that tweeted links to related sites with little commentary besides relevant hashtags. These "aggregator" Twitter accounts search for hashtags and tweets relating to the topic that they post and then favorite those tweets. This behavior may pass under Twitter's radar because the accounts select phrases to search for that are relatively uncommon, simply reflecting the intermittent behavior of other users.

Figure 9: Favorites Per Cluster



5. CONCLUSION

Our study focused on analyzing interaction between human and bot Twitter users. We used *BotOrNot*, combined with unsupervised machine learning, to cluster users and determine how bots gain visibility with their target audience. We determined that bots are generally unlikely to engage with human users beyond simply following them and using hashtags. However, when bots *do* interact with users, they do so without moderation, resulting in bots tending towards behavioral extremes.

5.1 Further Work

Although we were able to manually identify advertising links, when we retrieved data on individual Tweets we did not expand Twitter's shortened URL format. This made media, such as photos, appear identical to other links, since Twitter represents media as URLs. In addition, the sample size for manual classification was small by necessity; setting up a web service such as Mechanical Turk to crowdsource this account classification would improve analysis and clustering.

Furthermore, while several accounts were discovered that exhibited bot-like behavior such as spamming a hashtag, a number of these accounts were verified by Twitter, especially those run by so-called "financial consultants." While a closer examination of these users was outside the scope of this project, they seem closely related to the issue of spambots, and further discussion as to whether these accounts violate Twitter's terms of service seems warranted.

While we sought to provide an analytical overview of our sample users, a deeper statistical analysis of individual clusters and bots is warranted, based on several of the characteristics we identified. In particular, analyzing the distribution of mentions across individual clusters of bots and clusters of humans would lead to greater insight into that aspect of their behavior.

Finally, Twitter allows for users to quote other Tweets, as a commentary layered upon a retweet. This is unlikely to be a vector by which bots interact with humans, but it would provide a further view of Twitter users' behavioral patterns and its absence could possibly indicate automated behaviors.

6. REFERENCES

[1] S. Aral and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011.

[2] A. Bessi and E. Ferrara. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21(11), 2016.

[3] N. Bilton. Social media bots offer phony friends and real profit, nov 2014.

[4] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Dependable Secur. Comput.*, 9(6):811–824, Nov. 2012.

[5] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. *CoRR*, abs/1602.00975, 2016.

[6] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian. 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)*, pages 620–627, Aug 2014.

[7] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media, 2011.

[8] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA, 2010. ACM.

[9] Twitter. Automation rules and best practices, apr 2016.

[10] C. Xiao, D. M. Freeman, and T. Hwa. Detecting clusters of fake accounts in online social networks. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, AISec '15, pages 91–101, New York, NY, USA, 2015. ACM.

APPENDIX

A. CONTRIBUTIONS

Alic Szecsei provided data retrieval methods for Twitter accounts, programmed the unsupervised machine learning, and wrote the data analysis.

Willem DeJong programmed BotOrNot score retrieval, retrieved data for Twitter accounts to store in SQL databases, and created many of the graphs and charts.

B. MISC. DATA

Figure 10: Manual identification of accounts

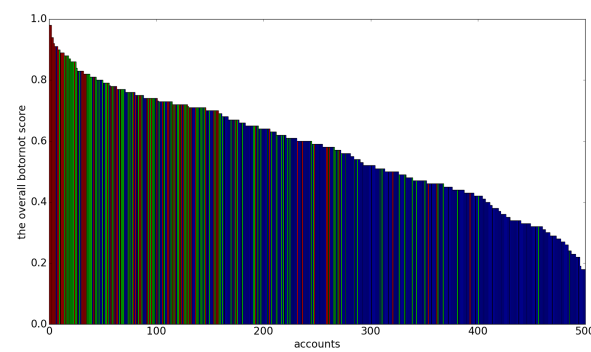


Figure 11: Followers Per Cluster