Measuring Social Bot Usage in the 2020 Presidential Election

Hunter Dyer
hadyer2@illinois.edu

Eric Cao
ejcao2@illinois.edu

***Abstract*—In this paper, we analyze our novel data set that was collected over a time window that covered the weeks before and after the United States presidential election, and at the height of the COVID-19 pandemic. We utilize Botometer [3] in order to gauge how bot accounts were used, and try to estimate the extent at which they were used in this time window. We also explore additional heuristics in order to see how generalizations fare in discussing bot behavior, and to see if these generalizations actually correlate with bot behavior. The paper is concluded with our anecdotal observations of accounts that were maximally rated by Botometer, and how these accounts differ and what similarities between them we noticed.**

I. Background + Related Work

In recent years, social media has become more impactful in terms of discourse for world events, as well as impacting offline events and views. Subsequently, this opens the door for malicious actors to manipulate this virtual discourse for perceived real-world benefit. A well-known method for performing this manipulation are automated accounts on social media platforms, which are called social bots.

Social bots simply refers to any automated accounts on a social media platform, and the term is behaviour-independent [2]. An account that automatically tweets out new posts from a blog or a news outlet can be considered a social bot, as well as an automated account that helps boost the reputation of a particular account or post. Networks of these bots can be used for the latter purpose in order to have a very noticeable impact.

Networks of bots allow for a 'fake popularity' to be created on social media platforms, which constantly utilize trends and popularity of posts and topics to recommend new content for users. This has various use cases, including reputation boosting for the operator, misinformation spreading, financial influence, etc. In recent estimates (2018 + 2019 respectively), it was thought that 15% of all Twitter accounts and 11% of all Facebook accounts were bot accounts. While these accounts are not all malicious, 10% of the userbase accounts for a large portion of both social media platforms.

Social bots of the malicious variety, as of late, tend to act in coordinated behaviors [2]. The description coined within community discourse and mentioned in Cresci's survey that succinctly describes the behavior of recent malicious bots is "coordinated inauthentic behavior." While this behavior can vary, the behavior usually results in some benefit for the operator of these 'networks'. The University of Indiana's Observatory of Social Media (OSoMe) has done a lot of work in regards to bot detection, and even has a general purpose bot classifier for Twitter, which they call Botometer [3]. Botometer evaluates a Twitter account on numerous features and then gives the account a score in six different areas, which they consider the main behaviour categories for social bots. The categories are astroturf, fake follower, financial, self-declared, spammer, and other.

Astroturfing accounts relate to accounts that are political in nature and follow politicians and boost posts. While fake followers, spammers, and self-declared bot accounts are rather self explanatory, the financial bots are the bots that post within the financial sections of Twitter. They are likely used for attempted market manipulation, or for other purposes where the benefit for the operator is some form of financial gains. Botometer classifies any other sort of behaviors that have occurred in known bot data sets within the 'other' category [3].

These categories cover most uses for social bots, however behavior isn't always neatly classified within these labels. With the rise of COVID-19 and with the impending 2020 election in the United States, it is very likely that bots have been used in some capacity to promote misinformation or otherwise sow general discord.

Early work in bot detection focused primarily on individual accounts, and the appearances of these accounts [2]. Bot detection as a whole is an arms race between operators and the detectors, as early detectors were focused on the appearances and behaviors of individual accounts, which made detectors easier to circumnavigate as they were overtuned for individual actions. From this, Cresci notes that focusing on the behavior of groups stands a better chance of not being made obsolete as fast, as it is assumed that the group behavior will still remain relatively unchanged, and the changes made to the individual accounts don't matter.

From this information, we utilize Botometer in order to provide preliminary analysis of bot usage as it pertains to the 2020 United States presidential election. We also work to gather a data set that can be released in the future for continuing bot detection work, as well as for other similar applications.

II. Methodology + Results

In this section, we present our methodology and the results of the different aspects of our data set that we looked at.

Notably, we reserve analysis and discussion of data for the next section.

A. Data Collection and Labeling

The data set that we utilized in later sections for analysis was collected as part of another project by both of us, in addition to one other collaborator (Mark Cockburn) for CS598 GW (Machine Learning for Systems, Networks, and Security). We collected two sets of tweets from October 12th to November 22nd. The first, which we refer to as our Control set, was collected from a Twitter API endpoint that provides a random 1% selection of all tweets made in a day. This was collected in order to give us a baseline of normal Twitter activity. The other set that we collected, which we refer to as our Filtered set, collected tweets that were made from a list of 83 politicians who were referenced multiple times in PolitFact articles, which is a independent fact-checking organization that focuses on political statements.

The rules we utilized for the Filtered set would capture any tweets in which a user was tagged (i.e. @realDonaldTrump) or mentioned by name in (tweets that included the string "Donald Trump"). It also captured any tweets that the specified user made as well. The list of names includes politicians on both sides of the political spectrum, and we feel that the authors, and the discussion surrounding them, captures a well-focused snapshot of political discourse during the time in which we captured the tweets. Our Control set had 346,570 unique authors, while our Filtered set had 359,362 unique authors.

As we discuss in the next sections, we first utilize these sets to collect baseline measurements before we perform further exploration and analysis. After exploring some baseline measurements and metrics, we explore heuristics on our Filtered set in order to see how well the heuristics generalize to capturing bot behavior. It is performed on our Filtered set, as we are specifically interested in exploring the use of social bots within the political discourse surrounding the 2020 United States presidential election, and not Twitter at large.

In order to label our data, we utilize Botometer [3], as it has an open API to query names with and it was one of the more recent classifiers produced. Botometer is an ensemble classifier that is trained on different data sets which each serve different roles. This variety can be seen in the results that are returned from Botometer when a name is evaluated. These fields can be seen in figure 1.

The Astroturf score is a measure of the political involvement of the user. The Fake Follower score indicates the general composition of how many fake followers the account has. Financial scores correlate with the use of cashtags ($AAPL), which is a construct on Twitter that can be used to discuss stocks and other financial entities. Self Declared measures the similarity of the queried account to other accounts that have self-registered as a bot. Spammer scores correlate to how many times the account reposts the same message. Any other behavior is measured and encoded in the Other score.

The last two important fields, for our purposes, are the "cap" field and the Overall score, which are utilized together. The


```
{
    "cap": {
        "english": 0.5288664776112156,
        "universal": 0.7089830786080018
    },
    "display_scores": {
        "english": {
            "astroturf": 1.0,
            "fake_follower": 0.6,
            "financial": 0.7,
            "other": 1.4,
            "overall": 0.6,
            "self_declared": 0.1,
            "spammer": 0.0
        },
```


Fig. 1. Return fields for a Botometer query.

"cap" field signifies the conditional probability, as evaluated by Botometer, of other accounts that are given an Overall score greater than that given to the account in question being labeled as a bot [1]. The Overall score is the score given to the account with the highest confidence from all the models used within Botometer. It is important to note that the Overall score may not necessarily appear in the other fields, and this is a consequence of the ensemble nature of Botometer. For measurement in later sections, we utilize the Overall score as our main metric, as it gives a score that is independent of any one behavior pattern and has the highest confidence associated with it. A score closer to 0 is more akin to a human account, while a score closer to 5 is more akin to bot behavior.

During our analysis, we often partition results based on thresholds defined by the Overall score. As previously mentioned, this is used as our primary metric for partitioning our results between bots and humans for further analysis due to its independence of specific types of behaviors, and since it serves well as a general metric for the account. In particular, we will often discuss results after a cutoff of (an overall score of) 4.5, as the cap score corresponds to 90% probability, and we felt this was a good balance of cut offs. Using scores above this would often result in a very small number of accounts to look at, compared to our sample size. A threshold of 4.5 offered, in our opinion, the best balance between the quantity of accounts and the cap score. As can be seen in figure 2, the general CAP score that corresponds with .9 is 4.5. There is a general plateau in probability between 1.5 and 4.5 which corresponds with .8.

B. Control and Filtered Set

The baseline measurements we took were a breakdown of each score for all authors, and then a breakdown of all authors that were awarded an overall score greater than 4.5. As mentioned prior, our whole data set had 346,570 and 359,362 unique authors for the Control and Filtered sets respectively. Due to the time cost associated with querying Botometer

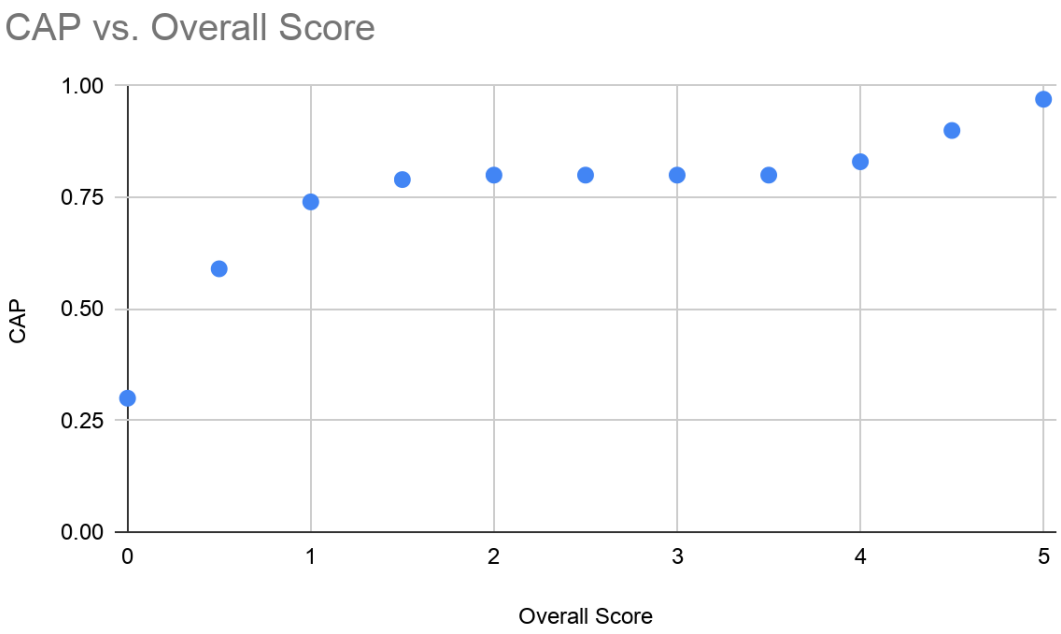

Fig. 2. Mean cap scores of accounts at given thresholds

through their API, we sampled 6,000 names from each set. The baseline measurements can be seen in table VI and VI for the Control and Filtered set respectively (at the end of the paper in the Appendix section).

Within each table, there are three sets of values. The top most values are of those accounts that were given an Overall score greater than 4.5, the middle values correspond to all users in our 6,000 user sample, and the final values are the cap scores that were discussed earlier. It is worth noting that of the 6,000 names that we sampled, generally around 10% of these accounts are unable to be given a score, as they are either turned to private mode, suspended, or deleted since the time that we collected the tweets. Our final total count for Control and Filtered were 5,342 and 5,604. Both had a similar number of accounts that were over our threshold of 4.5, as the Control set had 220 and the Filtered set had 211.

C. Heuristic - Username Structure

In addition to exploring baseline measurements, we also explored three heuristics about bot behavior, and wanted to see how well they generalized to our data set. The heuristics were general ideas about bot behavior and we wanted to see if we could utilize these ideas to effectively generalize bot behavior. In this section, we discuss a heuristic regarding the username structure, and in the following sections we discuss heuristics relating to hashtags and retweet proportions.

This particular heuristic was inspired by a tweet made by Professor Sibin Mohan earlier in the semester. It can be seen in figure 16 within the Appendix section. While Professor Mohan was claiming that users with the name structure of a text based name followed by a large string of numbers were paid trolls, and human, we wanted to see if there might be bot based behavior in this form. We had anecdotally encountered such accounts when manually labelling misinformation on this data set for the project that it was originally collected for. As seen in Professor Mohan's tweet, the number is often thought of as an identification number; however, it may also simply be the default account name. Regardless, we investigate if there is any correlation between these types of usernames and bot behavior as many bots may just use the default handle given by Twitter.

To investigate the heuristic, we collected all names within our data set that matched the Python based regular expression $[a-zA-Z]^{+}[0-9]\{5,20\}$. From our initial 346,570 unique, this captures 40,589 authors, which is close to 11.7% of the data set that we collected. Due to time restrictions, we were able to run only 10,000 names of these 40,589 names. Of the 10,000 we sampled, 8,853 were valid due to the attrition factors we discussed earlier (deleted account, suspended account, or account moved to private). Notably, we set the minimum number of digits beyond four as we wanted to omit names that included someone's birth year.

We present our results in a fashion similar to our baseline measurements we conducted for our Control and Filtered data sets in Table I. In addition to looking at the distribution of scores for this heuristic, we also thought it might be enlightening to explore the correlations between the number of digits in the username number and their score. These can be seen in figure 4 and figure 5 in which the mean and median Overall scores are computed for each possible value of numbers in the username. Figure 3 shows a distribution of

Numbers in Username	Mean	StD	Median	Max	Min	Count
5	2	1.54	1.4	5	0	690
6	2.07	1.49	1.6	5	0	343
7	2.41	1.54	1.8	4.9	0.1	72
8	2.37	1.4	1.8	5	0	7635
9	2.28	1.22	1.95	5	0.6	62
10	2.37	1.21	1.9	4.6	0.2	41
11	1.78	1.2	1.6	3.9	0.2	5
12	2.47	0.99	2.2	3.8	1.4	3
13	0.8	0.2	0.8	1	0.6	2

TABLE I
METRICS FOR NUMBER COUNT DISTRIBUTION OVERALL SCORES IN THE USERNAME HEURISTIC

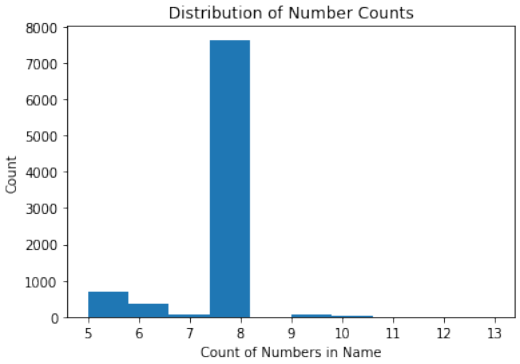

Fig. 3. Distribution of Numbers in Username.

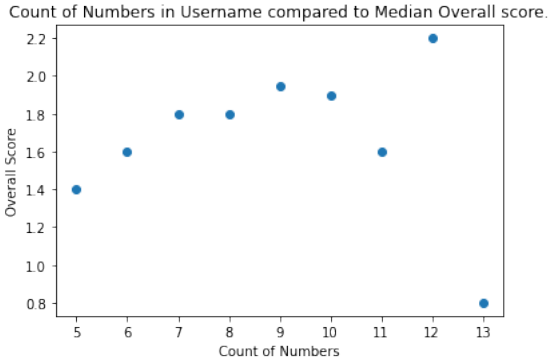

Fig. 5. Median Overall Score of Numbers in Username

	Mean	StD	Median	Count
#MAGA	2.63	1.57	3	2650
#Trump2020	2.08	1.57	1.6	1178
#Trump	2.13	1.54	1.6	890
#Election2020	1.98	1.59	1.4	842
#Georgia	2.38	1.58	1.95	668
#BidenHarris2020	1.25	1.35	0.8	662
#COVID19	1.7	1.56	1.05	524
#Biden	2.25	1.57	1.7	391
#MAGA2020	1.91	1.53	1.4	349
#StopTheSteal	2.33	1.56	1.8	346

TABLE II
MEAN, STANDARD DEVIATION, MEDIAN, AND COUNT OF TOP 10 MOST POPULAR HASHTAGS USING OVERALL SCORE.

the count of numbers occurring in the usernames.

D. Heuristic - Hashtag Involvement

We also explore the usage of hashtags on Twitter, and how they correlate with bot usage and activity. For this heuristic, we indexed all hashtags that were used within our Filtered data set and looked at the top 10 most common hashtags. From all users that had utilized one of the top 10 hashtags, there were a total of 9,335 authors. Some authors had used multiple occurrences of the most popular hashtags so we counted them as a duplicate within these 9,335. After attrition factors and only counting unique authors, there were 8,500 unique occurrences. In table II we present a breakdown of general score metrics broken down on a hashtag basis.

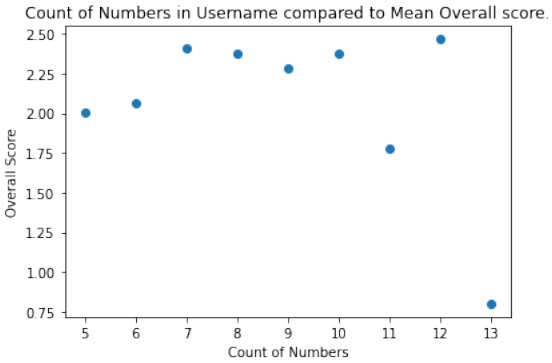

Fig. 4. Mean Overall Score of Numbers in Username

This heuristic was mainly explored as we wanted to see if there was any correlation between bot activity and the nature of hashtags used. We felt that this might be a particularly effective heuristic within our Filtered data set as anecdotally, it seems as if accounts whose sole purpose to is to sow discord, typically associate with political extremities, and utilize hashtags that correspond to their political extremity.

E. Heuristic - Retweet Proportions

The final heuristic that we investigated was the retweet proportion of an account. This was chosen as a heuristic since a common behavior for bots is to retweet a target tweet in order to boost the reach that the tweet has. We felt this simplification might be effective for detecting bots. For this heuristic, we only considered users that we had observed making 4 or more tweets. We did this to increase some of the significance of the results, as we didn't want ratios of 1.0 and 0.0 to be included by users who we observed only making 1 or 2 tweets. As with the other heuristics, we collected 10,000 names randomly from those eligible. After attrition, we were left with 9,450 users. We did not force any particular proportions when sampling our data.

In order to understand the results we first bin the proportions and look at corresponding scores in order to understand what the general Overall score for each range of retweet proportions looks like. We then look at the correlation of all names sampled along with their paired Overall Score to understand if the conclusions from binning the proportions gave an accurate

Retweet Proportion P	Mean	StD	Median	Count
0 <P $\leq$ .25	2.24	1.44	1.8	1895
.25 <P $\leq$ .5	2.61	1.61	2.2	510
.5 <P $\leq$ .75	2.95	1.62	3.6	1128
.75 <P $\leq$ 1.0	3.37	1.35	3.9	6471

TABLE III
METRICS OF OVERALL SCORE IN DIFFERENT QUANTILES.

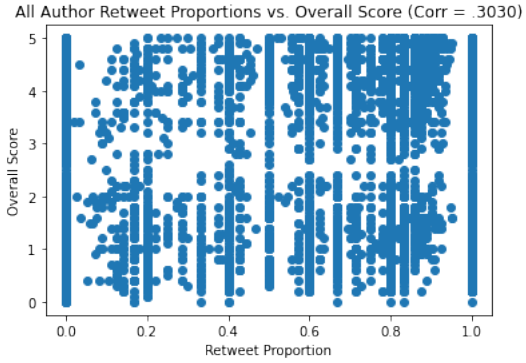

Fig. 6. Retweet proportions plotted against Overall score for all points.

understanding of the data. We present the results in table III, figure 6, figure 7, and figure 8.

F. Manual Validation

Until now, we have primarily focused on the results from Botometer in regards to our election-centric data set that we collected, we also wanted to extend our investigation to evaluate Botometer, as we have placed trust in its scores. To do a soft evaluation, we took 20 accounts from our Filtered set and our Control set (40 total). We specifically select accounts with an Overall score of 5 as these represent the strongest representation of bot behavior, in the metaphorical eyes of Botometer, and we felt this would represent a best-case scenario evaluation of Botometer. We make the distinction between the Filtered set and Control set as we feel they represent two distinct halves of our data.

We labeled 80% of these randomly selected accounts as bots. Our manual labelling of accounts was largely dependent on our understanding of social bot behavior and motives, that we discuss in the background section. While it is difficult to parameterize all factors that went into our decision, we discuss our general findings and observations in the discussion section. Our discussion covers the dominant factors that we used for distinction between bots and humans.

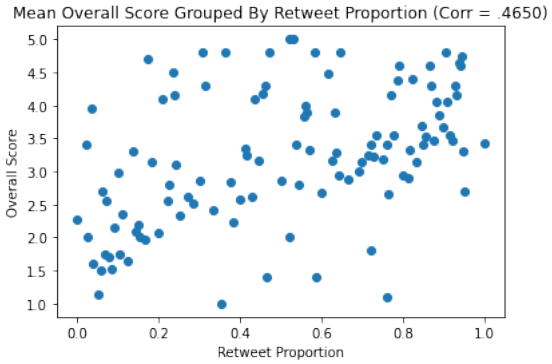

Fig. 7. Retweet proportions plotted against mean Overall score for each unique proportion occurrence.

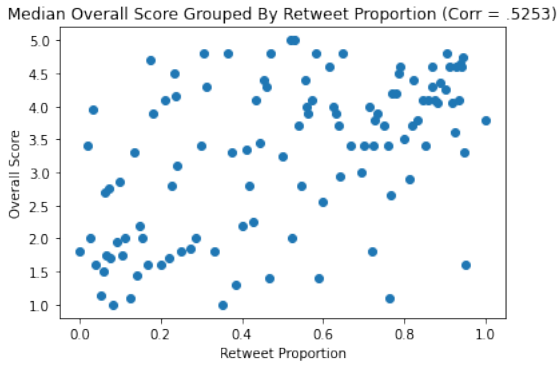

Fig. 8. Retweet proportions plotted against median Overall score for each unique proportion occurrence.

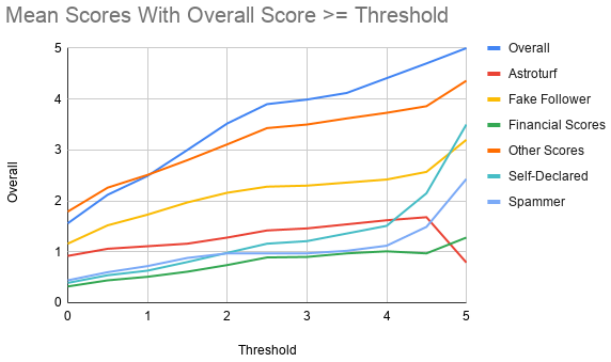

Fig. 9. Average bot score per category over the control data set bot plotted against a threshold indicating a cutoff of what would be considered necessary cap score for being classified as a bot.

G. Election Week

We also wanted to give a glimpse into how general score metrics appear during election week. In this investigation we sampled 500 random names for each day from 10/29 to 11/6. We then looked at a few brief metrics in regards to how the Overall score fluctuated within this time period. We keep our investigation and discussion of this scenario brief as the number of names we were able to run each day was relatively low, and we felt that not as many definitive conclusions could be drawn from such a small sample size. We examine each individual score field within the Control and Filtered set, while also looking at the count within each threshold value. These can be seen in figure 9, figure 10, and figure 11

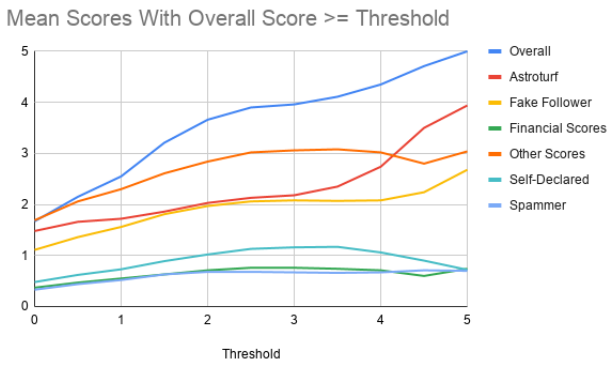

Fig. 10. Average bot score per category over the filtered data set bot plotted against a threshold indicating a cutoff of what would be considered necessary cap score for being classified as a bot.

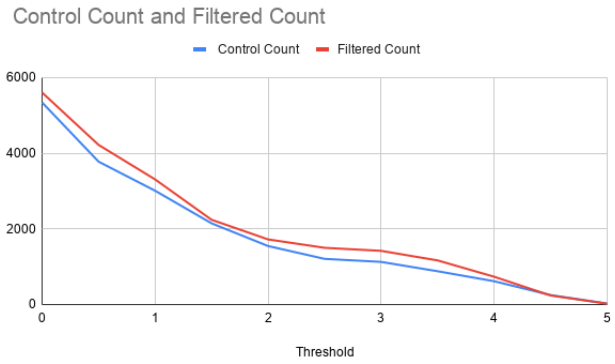

Fig. 11. The total number of accounts with a bot score above a threshold plotted against a threshold indicating a cutoff of what would be considered necessary cap score for being classified as a bot.

III. Discussion

A. Control and Filtered Sets

First we compare some overall aggregate statistics about Botometer's evaluation of the likelihood of accounts being bots. In Figure 11, to avoid making a binary classification about the likelihood of an account being a social bot or not, we use a threshold score and look at the number of accounts that would be considered a bot by the model given that threshold. When comparing the filtered count and control count. we find that their distributions are very similar. When comparing data sets of approximately the same size, there is not a substantial difference in the proportion of bots between the two data sets regardless of threshold. We find that this result is very interesting because our hypothesis going into the project was that we would see much more bot activity in the Filtered set as there was much political discussion about COVID-19 and the election during the time frame that we collected tweets.

We propose that this may be the result of a few phenomena. For one, Botometer may not be adept at classifying certain types of social bots used during the election, which would bias our results. Additionally, bot-like behavior can arise from accounts that are controlled by people, but utilized in a dishonest manner, such as accounts that are paid to retweet certain tweets to increase its exposure and visibility. These accounts may differ in behavior and interactions from normal social bots, but we would ideally also like to track this type of behavior, as they serve a similar purpose. Finally, an important metric not captured is the impact that a given bot has. Even if there is a similar proportion of bots in each data set, they may not have equivalent effectiveness in promoting content and gaining visibility.

We also look at the how the types of bots differ between the control data set and the filtered data set in Figures 9 and 10, respectively. Here, we again use the threshold along the x-axis to represent a cutoff for what we consider a bot and along the y-axis we plot the mean score per Botometer category. We found that while the overall scores across the two data sets were similar, the filtered data set had many more astroturfing bots. While this is the least represented category in the control data set, it is the most represented category in the filtered data set. While this isn't too surprising given that the filtered data set is concerned partly with political topics, it is striking and may also contribute to the public perception around bot usage in regards to the election and COVID-19.

B. Heuristic - Username Structure

As mentioned, there are accounts that may not be social bots, but are illegitimate. These bots pose as users using default usernames, but are instead paid per tweet they make and are given the general purpose of sowing discord. Just like a social bot, these accounts try to gain legitimacy through engagement with the bot coming in the form of retweets, likes, and replies. In order to capture this behavior and observe if their activity is marked as different from bots, we use a simple regular expression in order to match against every account's username. These usernames consist of a string of letters comprising the name and a string of numbers. As many account append a 4 digit string in order to represent their birth year, we match accounts only with a number that is 5 digits or greater. As we can see from I, on average, these account did not exhibit bot-like behaviors and received low scores by Botometer.

Interestingly, there was an overabundance of accounts that had 8 numbers within their name, far more than any other, and we believe this is due to the result of Twitter's default handle generation. However, the general Overall score doesn't seem to indicate a large presence of bot behavior. As Professor Mohan mentioned, it might be paid trolls and other humans responsible for this behavior, or other general paid entities. We think that this deserves a deeper look in the future, but it is outside the scope and time constraints of this paper. The distribution of scores of those with specifically 8 numbers in their name is bimodal, and can be seen in figure 12. The cluster in the lower half of the distribution is larger, which contributes to the lower general scores that we observed, however there were still a notable number of scores that were above 3, but were skewed to the lower end of the range between 3 and 5. Overall this heuristic did not produce the results

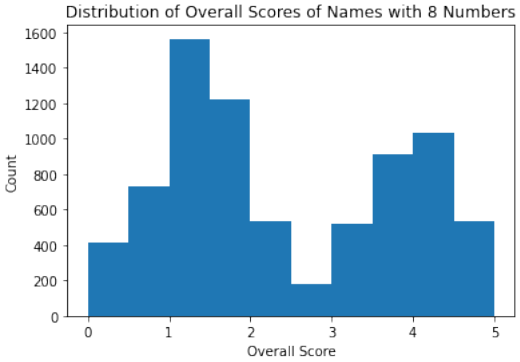

Fig. 12. Distribution of scores for names with 8 numbers.

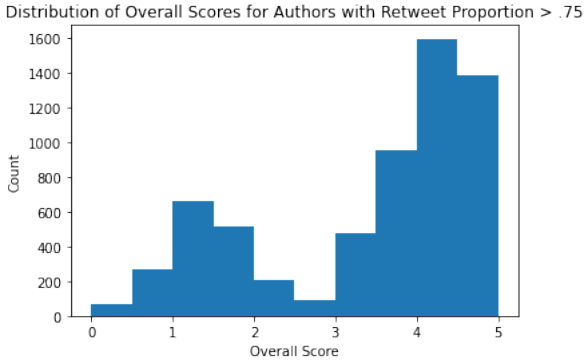

Fig. 13. Histogram of Overall scores within the highest quantile of Retweet proportions.

that we expected, but instead illuminated the possibility of another archetype of accounts on Twitter that we would like to continue investigating.

C. Heuristic - Hashtag Involvement

Another metric we looked at was the impact of hashtags and whether or not bots were more likely to use certain hashtags. We primarily looked at hashtags since they are a staple of Twitter and are often attributed to particular political campaigns and designate the nature and topic of a post made. This heuristic had less interesting results compared to the other two, but results were still somewhat enlightening. The results from this investigation can be seen in table II, in which we present metrics of the Overall score as they relate to the top 10 hashtags by occurrences in our Filtered set.

The notable result from this investigation was that the hashtags with the higher scores (mean or median), generally correlate with topics that are divisive and often have polarized views. We particularly see with #Georgia and #StopTheSeal that there was a a bit more bot activity within these hashtags compared to the rest of the list, which may indicate that there were bot campaigns participating in discourse about the aftermath of the United States presidential election. However, these Overall scores are not much larger than the baseline measurements taken for our Filtered set. So, while there may be a presence, it was not overwhelmingly apparent.

There is also a notable outlier with #MAGA, as the median score was 3, while others were much lower. This, along with some of the other bigger hashtags, all seem to indicate that there may be campaigns that are being performed with bot networks in order to control the discourse. Controlling the discourse gives an actor the power to control perception as well as dictate how views on this topic can influence future events or discourse. Given more time we would have liked to further investigate exactly how the high scoring authors were behaving to possibly understand where the bots originated from, or how expansive their network may be.

D. Heuristic - Retweet Proportions

We present our results for retweet proportions in table III, figure 6, figure 7, and figure 8. Table III shows a breakdown of the Overall score when we bucket the different proportions that occurred within quantiles. From this table, we see that as the proportion of retweets an account makes increases, so does the mean and median score. Interestingly, the concentration of the accounts that we sampled with retweet proportions above .75 was much higher than the other quantiles. We further breakdown the upper quantile in figure 13, which includes the distribution of counts for each Overall score, only considering users that had a retweet proportion above .75. From this we can see that the distribution is close to bimodal. There is a small cluster that has an Overall score between .5 and 2.5, while the other cluster has scores between 3 and 5. The scores between 3 and 5 is noticeably larger in terms of quantity of users, as well as being skewed closer to 5.

We show in figure 7 and figure 8 that there is a noticeable correlation between Overall score and the retweet proportion, and is particularly stronger among higher retweet proportions, which is inline with observations made in figure 13. We plotted all points in our sampled set in figure 6, but the results from this graph are largely inconclusive. It does show that scores generally cover the entirety of the possible range at most proportions, and that there is some slight separation in score distributions

Overall, we feel these results show promise in being able to designate bot behavior. This is mainly justified with decent correlation we see, along with the generally high score in the upper quantile. It could be the case though these results are only due to the politcal nature of the Filtered set, but as we have seen in the general baseline measurements from our Control and FIltered sets, the Filtered set has a slightly higher bot score overall on different metrics, but we don't feel that the magnitude of this difference would have correlated with the distinct results we see when filtering our results with this retweet proportion, and that our results are more of a result of the effectiveness of this heuristic rather than bias in the Filtered data. This heuristic showed the most promise out of the three that we investigated. While the heuristic itself can't be used in totality to detect bots, it aids in further lowering the potential subset.

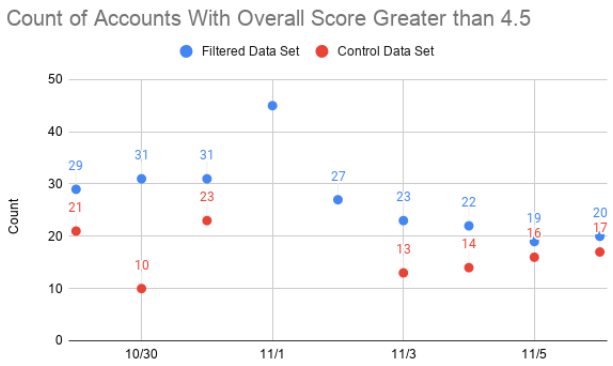

Fig. 14. Over the election week period, we plotted the number of bots with a overall bot score of 4.5 for both the control data set and the filtered data set. This corresponds to around 90% confidence that the account is a bot.

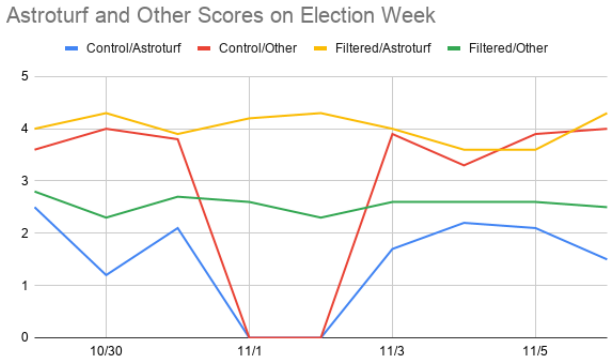

Fig. 15. Here we focus on the most popular classification of bots in both the filtered data set and the control data set. We plot the average bot score in these categories across all accounts identified as a bot in Figure 14.

E. Election Week

As one of the motivations behind collecting and analyzing this data was the 2020 US Presidential Election, we also look at bot activity during the election. Unfortunately due to a technical issue, we were unable to gather data for our control data set on 11/1 and 11/2. For each day, we randomly sampled 500 authors in order to get a representative sample and plotted the total number of accounts that were classified as a bot from Botometer (in which we considered an account a bot if their Overall score was greater than 4.5). In Figure 14 we can see that in general, the number of bots in the filtered data set is slightly higher than the number of bots in the control data set, which mirrors our findings from the entire data collection period, however we are unsure that this slight increase is statistically significant enough to make claims regarding bot use in each set.

Additionally, we looked at how this affected the most common score types for each data set, Astroturf for the filtered data set, and Other for the control data set. What we would expect from increased bot activity is that during Election week, the number of bots in the filtered data set would increase as bot activity increased. From what we observed, the activity is relatively stable. Notably, the large dips in the control data set's bot usage are a result of the aforementioned technical issue.

F. Manual Validation

The results of our manual validation weren't exactly what we expected, but were interesting. We ended up labeling 80% of the 40 accounts we looked at as bots (which were all given a 5.0 in Overall score by Botometer). We were limited by the number of names we could manually label by time, but we hope to do more in the future. 80% was a bit lower than we had expected, but the ones that we marked as humans generally weren't obviously human accounts. There was also a stark difference between the accounts in the Filtered and Control sets.

Within our Control set we labeled 3 accounts of 20 as humans. These 3 accounts generally had between 0 and 5 tweets, which is likely the reason they were misclassified, as the general behavior of the account wasn't averaged over multiple interactions, but was rather gauged on a small number. This doesn't give an accurate view of the accounts behavior as they make more interactions in the future. The general trends that we noted with the bots in the Control set were that the accounts usually self-identified as a bot a majority of the time, either in their username, display name, or in the biography section of their account. The rest of the accounts that we labeled as bots were more dubious as to their true nature.

These other bot accounts were often fixated on one topic or trend, and generally only retweeted tweets about this topic or tweets that included hashtags relating to the topic of their account. Sometimes, there would be replies to other tweets mixed in, but they would often be one word exclamations (i.e. 'RIGGED!'). There weren't many measures taken to hide the fact that the account was only fixated on one subject. While there may be a chance that the account was actually human, these account behaviors did not fall into generally conceivable human use cases, and just seemed to amplify a signal on a particular topic (usually whether they were for or against some topic in particular).

In our Filtered data set, general bot behavior was much more difficult to differentiate. In our Filtered set, we labeled 5 accounts of 20 as humans. Of these 5, 3 were marked so due to the lack of activity. In these cases, the interactions that the account had generated were replies to political figures (usually commenting the accounts' thoughts on the particular political figure). The other 2 fell more closely to the previously described behavior, in that they often only retweeted tweets or tweets with hashtags relating to a particular topic. These accounts had the distinction of adding commentary to other tweets very rarely through the quote tweet function (which still counts as a retweet). From what we observed (of our small sample size of 2), the commentary seemed very methodical in its generation, and had weird quirks to it. For example, one account used characters (the dot commonly used for dot

products in mathematical notation) that were not accessible through general keyboards, and used them at a frequency that made it seem intentional. There were also cases where the commentary made would be composed of a sentence, followed by a space, and then followed by three exclamation points (i.e. 'I had a sandwich for lunch !!! It was really good !!! I hope I have a sandwich tomorrow !!!') and the pattern was repeated for an unfixed amount of times.

The remaining 15 accounts that we observed, which we marked as bots, all had very similar behavior. We believe the behavior we observed was common enough between all that we would classify it as an overall 'archetype' of bot accounts. As previously discussed, these accounts all had the commonality that all their interactions consisted of signal boosting different information or misinformation of one particular topic. The most common topic that we observed was election fraud and different claims surrounding it. The accounts we labeled as bots didn't generally have any tweets of their own outside 1 pinned tweet that usually was a joke or some snippy remark about a politician that had a notable amount of interactions to it (replies, likes, retweets). It seems that this may have been included, if they were actually bot accounts, in order to provide some effort in deploying camouflage.

The general biography section of these accounts were interesting. They make it very obvious which political allegiance they operate for, and usually include hashtags (many of which are included in the top 10 list of our hashtag heuristic) in order to give themselves more visibility. There is generally then some statement of their age and family to give them legitimacy. The age generally tends to be greater than 50. We often noticed that there would be a novel location listed on their profile that was intended as humor (i.e. "Don't worry, I'll find you"). There were also many accounts within our small sample that gave explicit instructions to not DM the account, which we found strange. Profile pictures also ranged from just pictures of Flags, to actual photos of humans. As a point in future work, it would be interesting to see if the photos could be traced back to some other source, in the event that the account was actually a bot.

The display names are also changed in order to mock the general discourse on Twitter. Among the accounts we labeled, there were many that included 'President Elect' and 'Dr.'. The 'Dr.' is in reference to a recent opinion piece in the Wall Street Journal by Joseph Epstein about the usage of Dr. within Dr. Jill Biden's title. Aside from the display names, the usernames (the @ handle) were of no common form that we observed in our small sample. The last strange occurrence we noted with these accounts is that they will often have a lot of followers, and follow an equal amount of people. For instance, many of the bot accounts we observed had in the range of 10,000 to 40,000 followers, and generally followed a similar number of people. While the number of users an account can follow is up to the discretion and behavior of the user, the number of followers a user has relies on the behavior of others. There was nothing in the profiles that would give a hint of explanation towards their perceived popularity.

For this evaluation, we only considered accounts with an Overall score of 5.0. While this obviously reduces and ignores the intricacies within the scoring of Botometer, it was necessary due to our limited time (especially with query limits on the Botometer API). We wanted to look at the most extreme case to see what they generally looked like. In future work, we would also like to find better ways to qualify the behavior of these new accounts that we observed, along with the accounts we observed with 8 numbers in their name, as we feel that there might be promising work in further developing a taxonomy of malicious accounts. We also want to acknowledge that in our observations we present in this section, we observed the accounts knowing what their Overall score was. In the future we hope to manually label more accounts without priming ourselves with the knowledge of the scores.

IV. Limitations and Future Work

Overall, we were satisfied with the analysis we were able to do, however we faced a few limitations, primarily related to time that inhibited from completing everything we wanted to look at for this project. The primary limitation was faced was the data collection time, along with the query speed for the Botometer API. Getting our data set fully rated by Botometer was not feasible within our time frame, but we were able to generate scores for 45,500 different authors in our data set, which comprises around 6% of all names that we collected. With more time and effort we hope to generate scores for all names in our data set so that we might be able to fully understand the happenings of our data. We did our best to look at various aspects of our data sets with relatively small samples of names, but more names would give a more solid understanding of the behavior of bots within our data set and would have allowed for more thorough analysis. We also feel that in future work we would like to explore heuristics in a more formalized manner so that we can qualify the behavior in more formal definitions.

Additionally, we found a substantial increase in the number of bot accounts in a random sample of the control data set as compared to a random sample in the filtered data set. This does not imply that the impact of these bots were the same. In order to determine this, we would need access to several statistics that we did not have access to like views, retweets, favorites, and internal statistics for how likely a tweet is to be recommended. As our data is collected in real time,and Twitter can suspend accounts and modify their recommendation algorithm over the time we curated our data set, any metrics we would collect on this would be inherently biased and observations we made about them may be the effect of Twitter's recommendation algorithm, rather than bot behavior. This is why we instead opted to collect data in real-time, so as to get an accurate reflection of user activity with respect to our previous defined rules.

In regards to future work, we have a few points that we hope to continue working on. The first is that we wish we were able obtain results for all names in order to get a more

thorough understanding of our data. With time we hope to finish generating scores for each user, and then with some manually labeling we hope we are able to produce a data set that is useful for future work on social bot detection that we can publish. While we work to finish labelling our data set, we would have also liked to evaluate our data set with other previously successful bot detection methodologies. It would have been interesting to see if our data set introduces any previous blind spots, as well as giving some rough measure about how similar our data set was to previous data sets. Most of the progress on points of future work were a consequence of time limitations in the semester.

V. Conclusion

In this paper we provided analysis of a data set collected surrounding the United States 2020 presidential election, and provided preliminary measurements of analysis of potential bot usage on Twitter. This was done so with measurements provided by Botometer, a recent ensemble bot classifier. We discuss our findings, along with analysis of the findings, in order to assess the general usage. We also provide a discussion and analysis of Botometer's ratings, and how useful they are in an applied setting through manual validation of Botometer's scores.This was performed on our Control and Filtered set. While bot detection is an admittedly difficult task, we attempted to provide some useful insight into how Botometer fared with manual verification, as well as describing the observations we made within our data set.

VI. Appendix

References

[1] *Botometer by OSoMe*. URL: https://botometer.osome.iu.edu/faq.

[2] Stefano Cresci. *A Decade of Social Bot Detection*. Oct. 2020. URL: https://cacm.acm.org/magazines/2020/10/247598-a-decade-of-social-bot-detection/fulltext.

[3] Mohsen Sayyadiharikandeh et al. *Detection of Novel Social Bots by Ensembles of Specialized Classifiers*. 2020. arXiv: 2006.06867 [cs.SI].

Overall Score >4.5	Mean	Median	Std	Max	Min
Astroturf	1.649545	0.5	1.9544	5	0
Fake Follower	2.593636	2.4	1.1012	5	0.5
Financial Scores	0.97409	0.6	0.998725	4.4	0
Other Scores	3.88681818	4.6	1.10951	5	0
Self-Declared	2.24954	2.5	2.078977	5	0
Spammer	1.558636	1.15	1.43958	4.9	0
Overall Score	4.737727	4.7	0.1423364019	5	4.6

All	Mean	Median	Std	Max	Min
Astroturf	0.9219955	0.6	1.012576	5	0
Fake Follower	1.158742	1	0.94978	5	0
Financial Scores	0.321321	0	0.5754318274	4.4	0
Other Scores	1.78809434	1.6	1.2363467	5	0
Self-Declared	0.392006739	0	0.917408319	5	0
Spammer	0.4390864845	0.2	0.657954189	4.9	0
Overall	1.56025833	1.2	1.4286482	5	0

CAP	Mean	Median	Std	Max	Min
Threshold	0.9077986	0.8995513244	0.031231111	1	0
Population	0.611482665	0.7287106186	0.232922	1	0

TABLE IV
MEASUREMENTS OF CONTROL SET.

Overall > 4.5	Mean	Median	Std	Max	Min
Astroturf	3.5109	4.6	2.00186	5	0
Fake Follower	2.272985	2	1.32	5	0.4
Financial Scores	0.61943127	0.2	0.84298317	4.6	0
Other Scores	2.8255924	2.7	1.0861066	5	1
Self-Declared	0.9142180094	0.1	1.4794744	4.8	0
Spammer	0.7421800948	0.2	1.186	4.8	0
Overall	4.738662559	4.7	0.1383805	5	4.6

All	Mean	Median	Std	Max	Min
Astroturf	1.4815132	1	1.269192	5	0
Fake Follower	1.10603	0.8	0.92582	5	0
Financial Scores	0.36846895	0.2	0.5506374324	4.6	0
Other Scores	1.68909	1.4	1.170459	5	0
Self-Declared	0.47896145	0	0.991355	4.8	0
Spammer	0.3379728765	0.2	0.556575954	4.8	0
Overall	1.673893	1.2	1.4679143	5	0

CAP	Mean	Median	Std	Max	Min
Threshold	0.9062138	0.8995513244	0.027258929	1	0.8748575
Population	0.631729	0.7287106	0.216096	1	0

TABLE V
MEASUREMENTS OF FILTERED SET

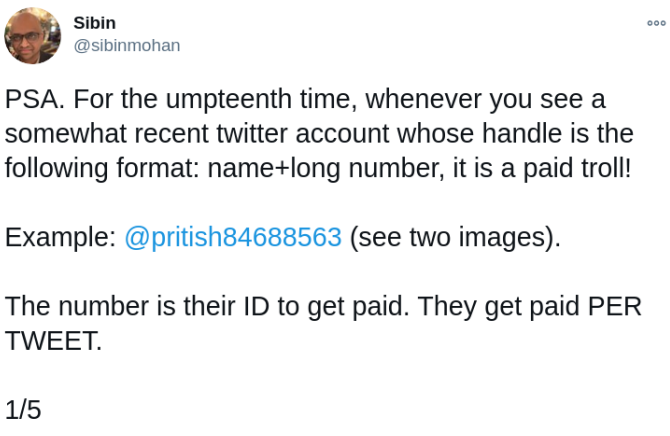

Fig. 16. Tweet made by Professor Mohan