Twitter Bot Detection: A Case Study

Aneri Shah
Indiana University
Bloomington
annishah@iu.edu

Dhruva Bhavsar
Indiana University
Bloomington
dbhavsar@iu.edu

Hely Modi
Indiana University
Bloomington
helymodi@iu.edu

ABSTRACT

Bots are machine-generated social entities that are capable of influencing and manipulating information which can be both beneficial as well as malicious. They can be created to target specific individuals for political reasons, gaining personal information, or for increasing the popularity of certain entities.

The aim of this paper is to analyze the Twitter platform for the presence of such bots by looking into the tweets of several random users. For this study, we particularly take the period of U.S. Presidential Elections in 2016. Using the information obtained from numerous users and their tweets, we try to achieve the following objectives:

- Analyze the tweets of the bots
- To classify users as bots or real users and the determine between those accounts, if any

1. INTRODUCTION

Basically, bots are software applications used for running automated tasks over the internet. They are algorithms designed to converse with humans who are capable of replicating content, sentiments, diffusion patterns, and temporal activities. They are used for providing some of the useful services such as weather updates, sports scores, etc.

These bots are used to automatically generate messages, post some contents over the site and advocate ideas. Though they are programmed to carry out a number of tasks efficiently and assist humans, sometimes they also impose a great risk in providing safe and precise data. So, the integrity of the contents present over the internet becomes a concern of matter. Thus, one of the most important questions that we need to ask ourselves is the authenticity and credibility of the data presented on the Internet. Many phenomena occur on social media for the very reason that users have no idea of the identity of such people.

Many companies are prone to using such automated bots primarily for customer care. Although they might seem responsible, they can contribute to spreading misinformation since they do not check source credibility. Some are designed in a way to spread biased views like in the case of U.S. elections. This can lead to thinking that the information presented in popular media is true in spite of it being untrue and therefore, making it harder to detect such bots.

Various reports and studies have suggested the extensive use of social media for social and political conversations. Also, it is proved that quite a few accounts on Twitter are Social Bots that engage in such conversations. Twitter became one such tool to engage people with the topic and to increase voter turnout by spreading popularized views about the party candidates. They are prone to manipulate public opinion and stir up controversies for political gain. Numerous attempts have been made previously to detect the bots on social media with the help of machine learning algorithms. Most of the widely used solutions include analyzing the tweets, the number of followers and following users, mentions, retweets, frequency of the posts, time and date of the tweet and so on.

In this study, we aim to analyze some of the attributes from the data to check whether there are indeed any differences between real and fake accounts and if so, what they are. We will identify the most common hashtags used to follow the people tweeting the most. The following sections will cover the details regarding this study. Section 2 provides some details of the related work done in this area, particularly covering the U.S presidential elections. Section 3 lists out the methods of data collection and analysis from Twitter with the help of Python programming language and gives insights using exploratory data analysis.

2. RELATED WORK

A study by Ferrara et al. focuses on the rise of social bots and their sophisticated and menacing presence. The paper lists out the advantages as well as disadvantages of the prevalence of such bots and methods to prevent them from tampering with the huge network of interconnected users.

A study by Grinberg et al (2019) discusses how the spread of fake news over social media became a concern in the United States immediately after the 2016 Presidential election. They examined how registered users of Twitter were

engaged in fake news sources and showed that the individuals that were most likely engaged with fake news were conservative-leaning, older and highly engaged with political news.

One of the widely known works in the case of presidential elections has been done by Bessi and Ferrara (2016). They explore about 20 million tweets to find the temporal dynamics of the conversation and how it had been affected by the presence of bots. They were also able to distinguish between real and fake users using advanced machine learning techniques.

Research by Messias et al (2013) explores the opposite way around where they created bots to share information and analyze the level of influence created by them to trick influence scoring systems. Another paper studies the similarity between user accounts using Euclidean distance which is then used to cluster similar communities.

One of the most developed programs is the Botometer which is based on the random forest algorithm. It uses more than 1000 features for classification. These features can be classified into 6 classes: Network, User, Friends, Temporal, Content and Sentiment. It makes use of all the attributes beginning from the text of the tweet to the count of followers and following, likes and retweets, etc. It had been employed to detect the influence of bots around the discussion of vaccination. Nevertheless, the research continues around the world to develop a model to detect bots.

Clark et al. improvise new methods for detecting these bots using three factors to measure the similarity between user's tweets and the URL count. The designed algorithm results in an accuracy of over 90% which also leads them to observe the behavior differences in these various types of bots.

Cresci et al went deeper into their analysis of data crawled from Twitter to find a new generation of such bots that were much more sophisticated compared to the previous ones. Their classifier along with their clustering algorithms was able to achieve an accuracy of 97.6% in detecting the bots.

3. DATA AND METHODOLOGY

For conducting our study, we have obtained the dataset from Kaggle which consists of two data files. One contains the information about the tweets and the other file stores information about all the users. The tweets are contained between a specific time period of 2014 to 2017. The dataset consists of tweets from the 3000 Twitter accounts that were believed to be connected to Russia's Internet Research Agency, a company known for operating social media troll accounts and which was believed to have influenced the 2016 US Election. The accounts were suspended by Twitter but this dataset was reconstructed by NBC and consists of a subset of the deleted data which NBC used for their investigation. This dataset was specifically used for exploratory data analysis but we fetched our own dataset which was used for classification purpose.

3.1 Data Collection

Since we already had a public dataset of all the bot tweets and their users from Kaggle, we began with analyzing the attributes of both the files. For the tweets data, we had over 200k rows after the removal of duplicates and NAN values. For the users data, the count stood at around 600 unique users. The users dataset consists of the following columns: created_at, description, favorites_count, friends_count, id, language, listed_counts, location, name, screen_name, statuses_count, time_zone, verified, known_bot. The tweets dataset consists of the following columns: created_at, favorite_count, id, id_str, is_tweet, lang, retweet_count, source, text, truncated, user_screen_name.

Along with this dataset, we also fetched a new list of verified and unverified users, picked randomly, and extracted a hundred tweets for each of them using Tweepy and Twitter API. The attributes: lang and time_zone of users are now deprecated so we were not able to fetch this for our new dataset.

We then merged the bot, verified and unverified datasets to get our final dataset we used for classification. Thus, the final dataset consists of around 40k bots tweets, 16k verified account tweets and 5k unverified accounts tweets. In this dataset, we maintain a column of 'known_bot' to distinguish between the real and fake accounts. The value of known_bot is true for the bots fetched from the public dataset and false for the other accounts. Below we describe the important parameters used further in the code analysis:

- retweet_count: the number of retweets for the tweet
- followers_count: the number of followers of the user
- favourites_count: the number of favourites on the tweets
- statuses_count: the number of retweets and tweets issued by the user
- friends_count: the number of people following by the user
- listed_count: the number of public lists that the user is a member of
- tweet_length: the length of the tweet
- text: the content of the tweet
- created_at: the date when the tweet/user was created
- user_screen_name: the twitter handle of the user
- verified: Boolean value for accounts verified on Twitter
- known_bot: created boolean column for identifying known bots

3.2 Data Preprocessing and EDA

After the preliminary steps, we began visualizing our tweets' data. Plotting the number of tweets against the dates gave us an idea about the peaks in the activity of users. These peaks were noticed around the important dates for the whole presidential campaign as shown in Figure 1. Using predefined important dates in the election, we have marked red dots which coincide with the spikes in the tweet activity.

Moving on, the text had to be processed so that the data could be better used for further analysis. This included:

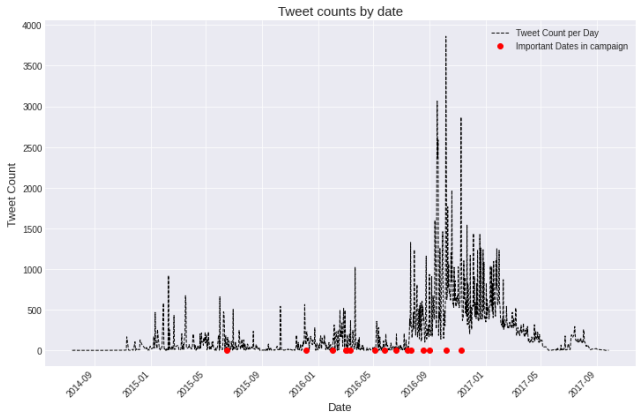

Figure 1: Tweet Count by Date

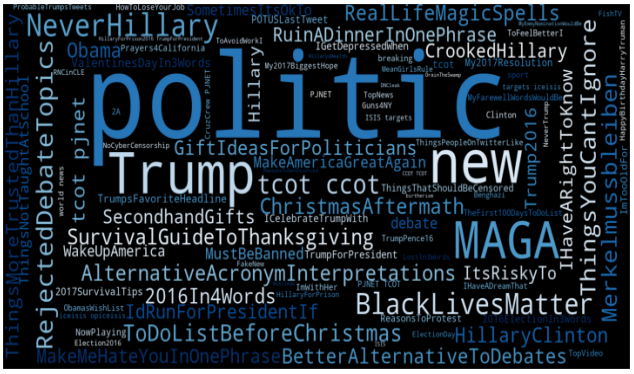

Figure 2: Most Common Hashtags

- Removing punctuation marks and special characters using the 're' library
- Using the techniques of stemming and lemmatization
- Removal of stop words
- Removing special characters, mentions, RT etc

Moving on, we looked at the most popular hashtags used in all the tweets (Figure 2). Through this, we found that most of those were in the support of Trump and his presidency. Next, we wanted to look at the time period when most of these fake accounts were created given in the data. The figure showed that the maximum count was obtained in the years 2013 and 2014. A closer look at the total users based on their type, the percentage of bots, real users and unknown accounts in the dataset were 66%, 25% and 9% respectively.

As mentioned previously, we merged the tweets with their respective users based on the type and created a separate word dataset for the known bots and the remaining ones for text analysis. After applying the text processing methods, we obtained a list of the most commonly used words by these two groups of users as shown in the Table 1. A deeper analysis using these words against the number of users showed that a large number of non-bot accounts did not use the most common words used by bot accounts (Figure 3).

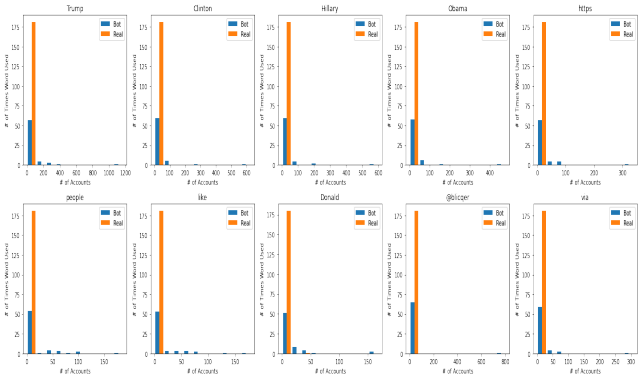

Figure 3: Number of Accounts vs Word used

Words By Real Users	Words By Bots
like	Trump
people	Clinton
one	Hillary
new	Obama
time	https
get	people
Trump	like
know	Donald
thank	@blicqer
@realDonaldTrump	via
Biden	@realDonaldTrump
today	I'm
see	get
love	@midnight
day	one

Table 1: Top 15 words by user

3.3 Classification

Moving forward, we aimed to classify the data into real and fake accounts using a few classification algorithms. The first step involved deciding the key attributes for this task. Our research on the previous work showed that many of the developed models used the number of friends count, followers count, retweets count, favorites count, number of posts, emotion scores, etc. to determine the class of the user. Hence, we agreed on eight features ('retweet_count', 'followers_count', 'statuses_count', 'favourites_count', 'friends_count', 'listed_count', 'tweet_length', 'polarity_score') from our dataset. For calculating the polarity scores of the tweet text, we used TextBlob library. The polarity scores indicate the sentiment of the tweet. Since we had a lot of missing values in the selected columns, we had to perform mean imputation to resolve these values. We used three algorithms to classify our data as described below.

3.3.1 Decision Trees

Decision Trees is a supervised machine learning technique used for classification. Here, the data is split using the most informative parameter which is calculated based on its entropy value. It consists of Nodes, the attribute to split on; the Branches, outcome of the attribute on which the decision is made and the Leaf Nodes, which are the terminal nodes in

the tree predicting the final outcome. In classification, the values of the branches are typically discrete or categorical which makes it easier to divide the attribute in a binary decision. Hence, the key idea is to create subsets of the data, then cluster the data points into dense and sparse regions in the form of a tree. This approach has been derived from the idea of a Divide-and-Conquer algorithm.

To use this approach on our dataset, we had to tune the parameters a bit so that we could achieve maximum accuracy by leveraging our attributes. The parameter n_estimator is used to limit the number of attributes used to construct the tree. Next, we changed the value of the max_depth parameter which controls the depth up to which the tree can be built thus preventing overfitting. We kept the default criterion as 'gini' to calculate the information gain. Applying Decision Trees classifier gave us an accuracy of 90.5% and we found out that the most important feature was listed_count followed by the favorites_count and friends_count sharing almost the same weightage.

3.3.2 *Random Forest*

Random Forest is an ensemble supervised machine learning technique which basically consists of multiple decision trees to form an ensemble. Each decision tree predicts an outcome the class with the most number of votes is selected as the final prediction. The reason why random forest works so well is that it overcomes any bias while grouping independent decisions from multiple decision trees and prevents overfitting.

Same as with decision trees, we had to tune our model for some parameters. We changed the values of default n_estimator and max_depth for training our data and the classifier yielded 93.4% accuracy. To confirm with our predicted results, we tried to search a few random twitter handles predicted as bots and were unable to find those accounts on Twitter. Using the same keyword for fetching the most important features, we obtained listed_count and followers_count with the highest importance followed by statuses_count and the retweet_count.

3.3.3 *Naive Bayes*

Naive Bayes is a supervised probabilistic machine learning model that is used for text classification. This model assumes that any feature in the model is independent of the other features. The model takes as input the counts of different words present in the text field of the Tweets dataset and predicts whether the tweet is made by a bot or not. We first generate a word count vector from the tweet text using CountVectorizer and then transform the train and test data accordingly. Multinomial Naive Bayes is used because the word counts are discrete quantities. The model gives an accuracy of 88.25% using only the text of the tweet. This concluded our code for the study and we present our observations in detail in the next section.

4. DISCUSSION

With the aid of exploratory data analysis, we were able to form our observations. Looking at the data, one of the most prominent observations was the activity of Twitter users and bots around the important dates of the election campaign. Most of the tweets were posted right on and around the

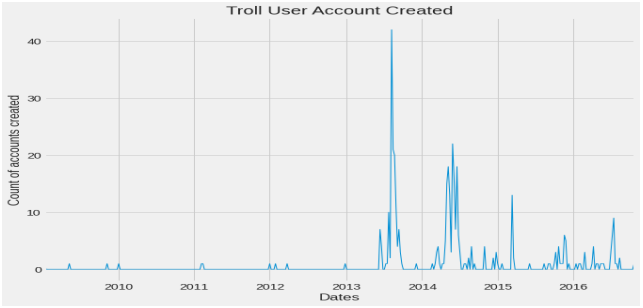

Figure 4: Date of User Accounts Created

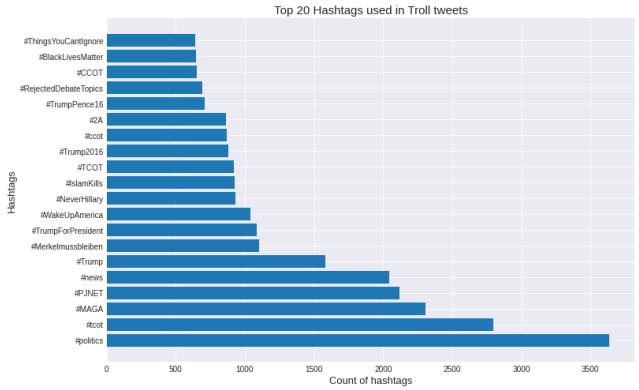

Figure 5: Top 20 Hashtags

important dates of the elections and most of the bots were created around two years before the presidential elections (Figure 4).

The word cloud of the most hashtags used during the time period reveals words such as #Politics, #maga, #tcot, #trump, #pjnet, #NeverHillary, #WakeupAmerica etc. which symbolized the supporters of Trump. These hashtags were used mostly during the month of July in the election year which was around the time when the Russian influence had begun. One of the hashtag which captured our attention was #Merkelmussbeliben which was amongst the top 10 hashtags used and indicated support for a particular German chancellor. These word clouds gave us the basic idea of how the tweets were used during the campaign and where they were revolving around. Then after, in order to quantify the frequency of each hashtag we plotted a bar plot representing the top 20 hashtags that were used in the tweets along with their frequency count which gave us the exact idea about the most frequently used hashtag during the campaign (Figure 5). All these show that the troll tweets during the 2016 US election campaign were more in favor of TRUMP.

Then, we analyzed and plotted the graph for when most of the user profiles were created and the number of such accounts created. This depicted that most of such accounts were created between 2013-2014. This led to determine the number of tweets created by these user profiles which we have mentioned previously. The graph plot depicted the user name and their corresponding tweet counts. Next, we determined the count of each user from each of the time-zone and

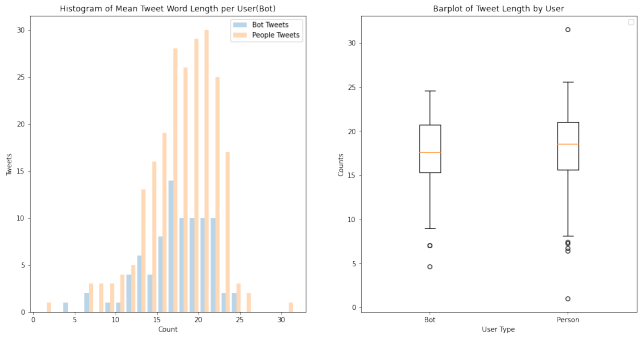

Figure 6: Mean length of the Tweets

Algorithm	Accuracy(%)
Random Forest	93.4
Decision Trees	90.5
Naive Bayes	88.25

Table 2: Accuracy of Models

language combinations. The graph obtained showed that most of the user tweets were in the English language followed by the tweets in the Indonesian language and Deutch language.

These analyses helped us to determine the number of users profile created, the language used for generating tweets and the time period when most of the accounts were created. Thus, the result depicts that most of them were during the election period and analyzing their tweets also shows that most of the tweets were involved with the political discussion and debates.

We also examined the tweet length by the real users as well as the bots. From the figure shown in the mean tweet word length per user, we found that there was more variability in real users' sentences as compared to the bots. Bots were prone to having much more concise statements that have a mean length of 16-20 words whereas the mean length for real users is around 13-23 words (Figure 6). This cannot be considered as a piece of conclusive evidence but the consistency of mean words for the bots can be considered as something to think about.

Since our dataset for the classification algorithms consisted only of numerical columns, we could not take into account the tweets generated by the bots. Hence, we created a separate data which only contained the tweets by all the users and applied Multinomial Naive Bayes algorithm which yielded an accuracy of 88.25%. The results from all these techniques are summarized in the table below.

Our EDA also involved exploring the most used words in the tweets by these bots before and after the day of the election. So, we filtered the tweets before and after November 8, 2016. We observed that after the elections, the buzz around Trump decreased and we found more tweets including the words 'Christmas' and 'Thanksgiving' (Figure 7).

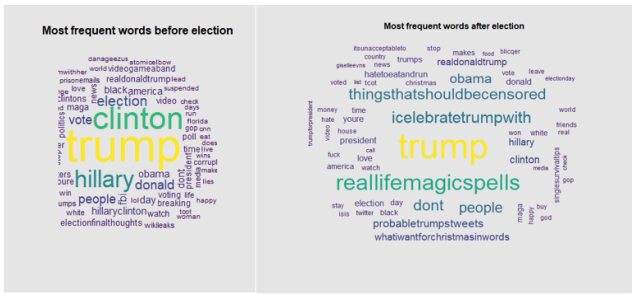

Figure 7: WordCloud for Before and After Elections

5. CONCLUSION

The objective of this study was to analyze whether there were any indicators which stood out while comparing the real users vs the bots. It seems that using the attributes we mentioned above, there does appear a striking difference when it comes to the statistical measures like listed_count and the favourites_count. With that being said, we also achieved a classification accuracy over 80% while only distinguishing the text from the tweets of these two groups of users.Hence, we can say that the performances of all the models were satisfactory but it remains to see how they would translate to a good bot-detection index.

The analysis of bot tweets from the public data showed that most of the bots were created during the US Presidential elections and the tweets spiked up during the major events of the election. But immediately after the elections, the number of tweets decreased drastically, through which it can be stated that many bot accounts come up during such events. When analyzing the bot account with respect to such events, we do get some knowledge about their activities and characteristics. But other than that, in general, it is quite difficult to classify what they do differently than a normal account.

One of the limitations of our model is that it depends on Twitter data for the statistical measures. Approaches based on only the text generated by the bots can be more useful and this can be considered as a future model for improvement.

6. REFERENCES

[1] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104. doi: 10.1145/2818717

[2] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. First Monday. doi: 10.5210/fm.v21i11.7090

[3] Messias, J., Schmidt, L., Oliveira, R., & Benevenuto, F. (2013). You followed my bot! Transforming robots into influential users in Twitter. First Monday, 18(7). doi: 10.5210/fm.v18i7.4217

[4] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot. Proceedings of the 25th International Conference Companion on World Wide Web - WWW 16 Companion. doi: 10.1145/2872518.2889302

[5] Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots. Proceedings of the 26th International Conference on World Wide Web Companion - WWW 17 Companion. doi: 10.1145/3041021.3055135

[6] Vikas. (2018, February 15). Russian Troll Tweets. Retrieved from https://www.kaggle.com/vikasg/russian-troll-tweets

[7] Popken, B. (2018, August 23). Twitter deleted 200,000 Russian troll tweets. Read them here. Retrieved from https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

[8] Housemusic. (2018, September 10). Twitter Bot Detection [Data Collection]. Retrieved from https://www.kaggle.com/housemusic/twitter-bot-detection-data-collection/notebook

[9] Jonathanbouchet. (2018, February 22). Beware of trolls. Retrieved from https://www.kaggle.com/jonathanbouchet/beware-of-trolls#grams-sentiment-analysis

[10] Timeline: Pivotal moments in Trump's presidential campaign. (2016, November 9). Retrieved from https://www.reuters.com/article/us-usa-election-timeline-factbox/timeline-pivotal-moments-in-trumps-presidential-campaign-idUSKBN1341FJ