University of Maryland, College Park

Twitter Bot Classification

Literature Review

Mohammad Subhan, Sahil Dev, Gabe Margolis, Joshua Goldberg, Jordan Foster

CMSC396H

September 30, 2020

Background

Our group's project for this semester is to research and create a program which will correctly identify bot accounts on one of the most popular social media platforms, Twitter. Social media is a website or application that enables users to create and share content (*Oxford Dictionary*). There are many different social media platforms which exist on the internet today. Among them are Facebook, Youtube, Instagram, and Whatsapp which sit at the top of the most heavily used networking sites. As of July 2020, Twitter has the 15th most monthly active accounts with approximately 326 million users (Clement). Twitter is a well known social media platform which allows users to post and respond to "Tweets". Tweets can range anywhere from thoughts and questions to videos, GIFs, advertising, anecdotes, and news updates. Tweets are initially shared to the account's followers directly. Followers will then like or retweet content they are interested in to share it with their followers. This chain pushes popular tweets to reach millions of people. The largest percentage of twitter users come from the United States and Japan (Clement). However, in recent years there have been reports of fake bot accounts creating and sharing tweets under the guise of being a real human. Our group will find a way to identify which accounts are bots using machine learning techniques with a variety of parameters and training methods.

Social media bots have been around for many years. A bot is an automated software agent which assists other bots or users with some specified pre-programmed task (Edward). One of the most commonly used virtual bots is a chatbot. A chatbot is used to communicate recorded responses to customers. For example, major banks like Capital One and Bank of America use chatbots so users can ask for help with their financial needs. Social media bots also virtually interact with people. These bots may automatically post or respond to people who meet some set criteria. They may also be set to follow certain accounts or share certain posts. Reports have shown that nearly 48 million accounts on Twitter are bots (Efthimion). Unfortunately, social networking sites do not distinguish between bot accounts and real users. Therefore, bots have the ability to deceive real users. A post from a bot could be difficult to distinguish compared to a real user. Bots have been found posting and responding to political tweets. In 2016, about one-fifth of tweets about the U.S election have come from a group of bot accounts (Efthimion). These bots greatly influenced the content that real Twitter users see on their timeline. They falsely depict the

response that the majority of citizens and real users want to show. Many bots can show fake support or anger over the tweet of a politician or content of a news article while other bots spread false information. Twitter Bots which attempt to skew public perception of a person or event cause problems and unrest on social networking sites. Studies have been done to identify accounts as bots and our group's goal for the semester is to improve on those methods and research further into solving the problem.

Datasets

Having a dataset of confirmed origin of tweets and accounts is useful to verify correctness of solution, compare a solution to existing approaches, and to potentially train a machine learning algorithm. The OSoMe project Botometer provides multiple datasets related to Twitter bots. Examples of datasets include manually verified human accounts and tweets, self-identified bots, fake follower accounts that can be purchased, authentic celebrity accounts, and more ("Bot Repository"). Many of the datasets are from 2019 and newer. They are of varying size with thousands of verified human and bot accounts or tweets.

A large source of tweets is FiveThirtyEight's list of Russian troll tweets. These tweets came from a "troll factory" designed to spread false information about politics on social media. The dataset includes almost 3 million tweets from almost 3,000 twitter accounts (Roeder). While the tweets are no longer on Twitter publicly, the dataset includes tweet content, the number of followers and following users for the account, among other information that an algorithm might be able to use to classify a tweet.

A challenge with some datasets is that not all datasets include tweet content, but only the tweet id. By Twitter's Developer Policy, in many scenarios a dataset can only publicly share the ids of tweets. While the content of tweets still on Twitter can be downloaded using the Twitter API, there are rate limitations in the API.

Approaches

Detection of Novel Social Botsby Ensembles of Specialized Classifiers

Twitter bots are extremely prevalent on the platform, as a result the term Twitter bot or social bot cannot fully describe the purpose of these bots. The authors of "Detection of Novel

Social Botsby Ensembles of Specialized Classifiers" seek to differentiate the behaviors of bots and separate them into different categories. Each model is trained to recognize three specific categories of bots, which includes, but is not limited to, traditional spambots, social spambots, and fake followers. Traditional spambots bombard users with content, social spambots support and/or attack users, and fake followers frequently follow users (Sayyadiharikandeh et al., 2020). After being trained on these categories the models are combined to create a general rule. Additionally a model is created to be specifically trained for recognizing the behavior of accounts used by actually humans.

This approach to detecting bots on Twitter proved to be fairly accurate when compared to other successful bot detection methods. The main metric used for determining the accuracy of the method of bot detection used in the paper, the Ensemble of Specialized Classifiers (ESC), is the area under the ROC curve. This metric, in this case, is how successful a model can determine a human account from a bot account with 1.0 and 0.0 being the highest and lowest scores respectively. ESC produced an AUC of 0.96 which is comparable to the AUC of Botometer which produced one of 0.97 (Sayyadiharikandeh et al., 2020).

One shortcoming of the paper is that due to the models in this paper's dependency on categorizing accounts some problems can occur during the training phase. For example, say that one Twitter bot being used for training is attacking a real politician's account and is categorized as a social spambot. If somewhere in the futuring during training this bot behaves more along the lines of a traditional spambot by promoting products the data can become invalidated. The data can also be invalidated if the bot account becomes terminated or inactive. Another shortcoming is that if a new category of bots is discovered or created in the future the models used in this paper will fail to detect them. However, because each model is tailored to a specific category of bots similar steps can be taken to add the new category to the general rule.

Given the shortcomings mentioned previously there are improvements that could be made to the work done in this paper. Although the research conducted is fairly recent, new Twitter bots that do not fit the classifications defined in the paper are constantly being created. As a result, one improvement that could be made is to increase the scope of Twitter bots that can be detected by adding more classifications. Another Improvement that could be made is to allow for bot accounts that pivot their behavior into another category to be switched into the training

data for that respective category. This however, would have to be done towards the end of training once the models can successfully categorize accounts.

Language-Agnostic Twitter Bot Detection

One interesting approach addressed in a paper by Jürgen Knauth is "Is it possible to detect bots at account level only, without taking into account the content provided by these accounts" (Knauth). To answer this question they took a robust and detailed approach. The paper focuses on the fact that there are no human limitations by twitter bots, They opted to use account features such as number of friends, geo enabled, and protected as a means to determine if a twitter account is a bot not. They used the MIB dataset which contained 8375 twitter accounts. The methodology was to use a machine learning algorithm and input these features in a way that algorithm can accurately and precisely distinguish between account types. Even more so they were able to test which specific features were the most helpful in making that decision. It concluded the most helpful in order was Levenshtein distance, geo enabled, and statuses count.

The results were good. First they established that a data set of 8385 is a large enough training classification for their machine learning algorithm. Second they were able to predict the account type with an accuracy of 98-99 percent. However, they note that bot developers will pick up on research like this and future researchers must use new data as existing data may not work so well. This is an important thing to note as the paper was written over a year ago. The 2020 election is fast approaching so this problem is at an all time high.

I think the largest benefit of this paper is the fact that the researchers were able to explicitly say which features were the most accurate. Using this we can add only the features we think will have the largest benefit to our algorithm.

Twitter Bot or Not?

In the paper "Twitter Bot or Not?", NYU undergraduate students Mattapalli et. al approach the problem of identifying Twitter bots via classification with several models, comparing the performance between them (2017). The paper begins with an intuitive exploration of the dataset, highlighting the breakdown of bot vs. human occurrences under various data distributions. For example, the paper demonstrates that a bot account is about twice as likely to

have a default profile picture as a human account. This is useful because when evaluating a model, one can compare the expectation of the model's parameters with its actual parameters. It is useful to analyze any disparity here to determine whether the model is performing properly and to gain further insight into the dataset.

The paper compares the performance of six machine learning models, with 28 account features, 12 features for individual tweets, and 19 features for aggregate data over the past 100 tweets for each account, providing for a somewhat language-aware solution that is only partially capable of interpreting the context of tweet. Here are the six models used along with the cross-validation accuracy for each:

Decision tree: a tree structure where at each step a comparison is made to determine which path to take and the leaves represent a final decision; achieved 89% accuracy

Random forest (with adaptive boosting): a set of decision trees where the majority predicted class is selected; achieved 92% accuracy

Logistic regression (L1): uses a linear gradient descent model with a sigmoid component for binary classification with lasso regularization; achieved 89% accuracy

Logistic regression (L2): uses ridge regularization; achieved 88% accuracy

Multilayer Perceptron: a multilayer logistic regression model, where each layer feeds as input into the next; achieved 88% accuracy

Voting classifier: combines several models and chooses the majority predicted class, extending the concept of a random forest; achieved 93% accuracy

In terms of F1 score, the voting classifier performed the best (0.93), followed closely by the random forest with adaptive boosting (0.92) and L1 logistic regression (0.90). Their analysis uses the majority class classifier as a baseline with accuracy of 50.4%.

This paper makes a few decisions that appear inconsistent or baseless. There is no clear reason for using two logistic regression models where the only hyperparameter changed is the regularization method, when only one multilayer perceptron architecture and one random forest is evaluated. Evaluating a deeper or shallower multilayer model could provide insight as to how architecture affects the performance of the model. While it is noted that this particular

architecture performed best out of the multilayer perceptron models, without extra evaluation data, it is impossible to determine which architectures perform better than others. Additionally, it is unclear as to why the tanh activation function is used. The paper states that this choice is made due to stronger gradients, but this information is not given in comparison to any other activation function. A ReLU activation is equally capable of strong gradients but takes far fewer resources to calculate. The training method for these models is also unclear, including the cross validation method and number of epochs, making it difficult to build off of the work done here. Since some models take longer to converge than others, it is possible that more complex models were underfit and could have performed better under increased training. Ideally, a graph with the number of epochs on the x-axis and the accuracy of each model on the y-axis, color coded to differentiate between models, would provide incredibly valuable insight into the performance of these models. The majority class classifier used in this paper does not provide a useful baseline. While it can help indicate whether a model performs better than just guessing, a slightly more complex statistical model, such as Naive Bayes would provide a much more tangible baseline.

Despite the best performing model only accurately classifying 94% of accounts, which falls short of other models discussed earlier, this paper provides a useful general approach to this problem, which we will adopt in our method. In particular, we will perform exploratory data analysis to develop an idea of what features should be prioritized in any model. We will then develop and evaluate several models, and compare the results with the expected results. We can then fine tune these models based on the results to achieve greater accuracy.

Timeline

Below is a relative timeline for our research.

1. Compile, clean, and prepare relevant datasets.
2. Analyze data to develop a concept of what our models will look like.
3. Create a simple baseline model.
4. Develop, evaluate and fine-tune multiple classification models.
5. Create a final much more complex, context sensitive model refined from the previous results, if time permits.

References

"Bot Repository." *Botometer.Osome.Iu.Edu*, 2020, botometer.osome.iu.edu/bot-repository/datasets.html. Accessed 29 Sept. 2020.

Clement, J. "Most Used Social Media Platform." Statista, 21 Aug. 2020, www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Clement, J. "Twitter: Most Users by Country." Statista, 24 July 2020, www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

Edwards, Chad, et al. "Is That a Bot Running the Social Media Feed? Testing the Differences in Perceptions of Communication Quality for a Human Agent and a Bot Agent on Twitter." Computers in Human Behavior, vol. 33, Apr. 2014, pp. 372–76. DOI.org (Crossref), doi:10.1016/j.chb.2013.08.013.

Efthimion, Phillip George; Payne, Scott; and Proferes, Nicholas (2018) "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots," SMU Data Science Review: Vol. 1 : No. 2 , Article 5. https://scholar.smu.edu/datasciencereview/vol1/iss2/5

Knauth, Jürgen. "Language-Agnostic Twitter Bot Detection." *Proceedings - Natural Language Processing in a Deep Learning World*, 2019, doi:10.26615/978-954-452-056-4_065.

Liberman, Neil. "Decision Trees and Random Forests." Medium, Towards Data Science, 21 May 2020, towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991.

Mattapalli, Revanth, et al. "Twitter Bot or Not?" 2017, github.com/Vignesh6v/Twitter-BotorNot/blob/master/twitter-bot_or_not.pdf.

Nagpal, Anuja. "L1 And L2 Regularization Methods." *Medium*, Towards Data Science, 14 Oct. 2017, towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c.

Roeder, Oliver. “Why We’re Sharing 3 Million Russian Troll Tweets.” *FiveThirtyEight*, 31 July 2018, fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/. Accessed 29 Sept. 2020.

Sayyadiharikandeh, Mohsen, et al. "Detection of Novel Social Bots by Ensembles of Specialized Classifiers." *arXiv preprint arXiv:2006.06867* (2020).

“Social Media: Definition of Social Media by Oxford Dictionary on Lexico.com Also Meaning of Social Media.” Lexico Dictionaries | English, Lexico Dictionaries, www.lexico.com/definition/social_media.