BOT DETECTION TECHNOLOGY USING BIG DATA

1.VANAM GOPINATH,

ASSISTANT PROFESSOR MATRUSRI ENGINEERING COLLEGE. SAIDABAD.

2.STVSAV RAMYA,

ASSISTANT PROFESSOR MATRUSRI ENGINEERING COLLEGE. SAIDABAD

ABSTRACT- Twitter is a prominent social networking service that allows users to share their thoughts on a variety of subjects such as politics, sports, the stock market, and entertainment. It is one of the most efficient ways of transmitting data. It has a significant impact on people's perspectives. As a result, it is critical that tweets be sent by real people rather than Twitter bots. Spam messages are sent by a Twitter bot. As a result, identifying bots aids in the detection of spam communications. Using machine learning techniques, this article offers a method for detecting Twitter bots. Decision tree, Multinomial Nave Bayes, Random Forest, and Bag of Words are compared.

Key Words: Bot detection, Twitter, Bot, Social bots, Machine Learning.

INTRODUCTION

Twitter is one of the fastest-growing social media platforms. It enables users to exchange news, express themselves, and debate current events. Users may follow individuals who share their interests or have similar viewpoints. Users may send tweets to their followers right away. Re-tweeting allows the content to reach a wider audience. During live events such as sports or award ceremonies, the number of tweets spikes. Smartphones and PCs can both access Twitter. Paid promotions may result in significant income creation as well as an increase in product sales. Students may use Twitter to learn more about the subjects that are covered in class. The message that is shared with followers is referred to as a tweet. The tweet should be short and to the point, with a maximum of 140 characters. The hashtag (#) is used to locate and follow a certain subject. When a hashtag gets popular, it is referred to as a trending topic. Twitter connections are bi-directional, meaning that a person may have both followers and followers. If you follow someone on Twitter, you will be able to view all of their tweets if the account is public; but this does not imply that he or she will be able to see your tweets. If you follow someone back, they will be able to view your tweets. Users get a large number of tweets, some of which are sent by bots. Bot detection is required

to identify fraudulent users and safeguard legitimate users from disinformation and harmful intentions. A Twitter bot is software that automatically sends tweets to people. Bots are created to do tasks such as spamming. Twitter bots' nefarious aim is to: 1) propagate rumours and fake information. 2) To smear someone's reputation. 3) For the purpose of stealing credentials, false conversations are generated. 4) Users are directed to bogus websites. 5) To influence the popularity of a person or a group by changing their views. We're working with a Kaggle dataset. Number of followers, friends, location, screen name (used to interact online), verified (if the user is authorized), favourite (used for liked tweets), URL, id, description, and listed count are among the characteristics. "The spearman correlation coefficient is used to extract features." The data collection has been honed to detect bots. Decision Tree, Multinomial Nave Bayes, Random Forest, and Bag of Words are all being used. Real-time data is tested using the algorithm with the greatest accuracy.

LITERATURE REVIEW

Several efforts in Twitter Bot Detection have been done. The following techniques and work are presented: Machine learning algorithms that are reliant on manufactured characteristics identify false identities generated by people or bots. It was assessed if easily accessible and well-designed characteristics utilized for a successful detection using machine learning models can be employed for the detection of false identities generated by bots or computers. Supervised algorithms need a dataset of features with a label that classifies each row or result. Features are thus utilized to predict an outcome by supervised machine learning algorithms. These characteristics may be the properties obtained via APIs which indicate the number of friends and a particular piece of information on an SMP account. The predicted results from the trained model of the computer only gave 49.75% of the top F1 score. The models have been taught to utilize engineering characteristics without depending on behavioral data [1]. Content polluters or bots who hijack conversations for political or publicity reasons are a well-known issue for the prediction of events, election prediction and differentiating between true and false news in the social media. It is especially difficult to identify this kind of bot. Content polluters are conveyors that aim to undermine a real debate by depriving it for political or publicity reasons. Methods have been developed to detect social problems in real time. Two features of tweets, i.e. temporal information and message variety, were

studied. It was observed that content polluters frequently timed their tweets together in this data set. By analyzing the time trends, bot accounts may be deduced. Bots have also utilized a limited number of URLs in their tweets [2]. Twitter users have begun to purchase false accounts followers. This may lead to spam on Twitter. Based on an account, 13,000 fraudulent followers and 5,386 real followers were carefully confirmed. Then many features that differentiate fake and true followers were discovered. These were used to categorize people as false or true as characteristics of machine learning algorithms. The Cumulative Distribution Function (CDF) for the six characteristics [3] has been provided to demonstrate that these features are really helpful to differentiate between fraudulent account followers and genuine user accounts. Bots must be detected to identify false users or bad users and safeguard real users from disinformation or harmful attempts. Twelve characteristics, such as followers count, buddy count, etc. in the bot repository data set are created using the statistical derivation. Other characteristics such as number of hash tags per tweet, preferred count per tweet, and the number of URLs per tweet are determined by adding them to users. Logical regression, neural and gradient-intensive network. By comparing the performance of these three increased gradients, the issue of identifying users as bot or human in a Twitter has been detected [4]. There are three kinds of trustworthy and honest users. Sybil accounts are numerous adversary-controlled accounts. Here the honest and Sybil areas are diminished and Sybils have little link to honest users. The Sybil communities generate a false, credible impression on honest users of the social network. By extensive connections. By doing a profile research of people and bot, the difference was found in tweet content, tweeting behavior, and accounting characteristics such as external URLs [5]. Cross linked actions and no etiquette data were discovered for the connected Twitter accounts unlike current bot identification methods. This technique is 94 percent accurate and effectively detects bots [6]. Studies have revealed that most spam communications are generated automatically by bots. Bot spammer detection thus lowers spam communications. Time entropy and tweet similarity were employed as criteria for the identification of spammers. Precision, reminder and f-measurement of this technique resulted, respectively, in 85%, 94% and 90% [7]. Twitter platforms or theme feed constitute 9% of the tweets. Number of tweets, followers, followees

and date of the first and final tweets have been identified for each account. Average tweets per day to compare the average tweeting activity were computed. Bot or not a scale of 0-100 percent shows that twitter accounts are likely to be human or social bot. Tweets, re-tweets and mentions, tweets and emotions appear or not. Of the 51 accounts, 84 percent had platform feeds, theme feeds and selected accounts four times. Platform and theme feeds generated 4.6 and 7.1 tweets per day per account. Selective accounts posted much fewer than 2.2 tweets each day in automated accounts [8].

PROPOSED SYSTEM AND METHODOLOGY

The block diagram of our system is shown in figure 1 and 2.

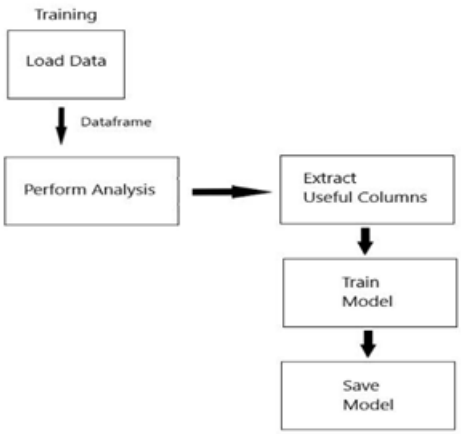

Figure 1: Training of the dataset

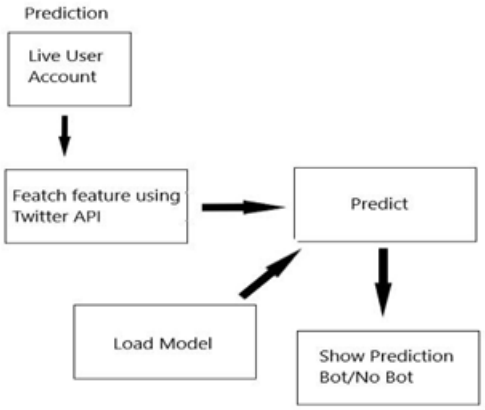

Figure 2: Prediction of Real-time data

There are numerous characteristics to the train data. The necessary functionality is retrieved using the Spear-man correlation technique. Naive Bayes, Decision Tree, Random Forest and Bag of Words are three learning models. Real-time data as displayed in Figures 1 and 2 are used as the optimum learning model. Data are pre-processed and zero values are eliminated using pandas (tool for pre-processing). The dataset is trained and the test data set is the actual Twitter data. The output is in shape 0 or 1 (1 indicating that it is a bot and 0 indicating that it is not a bot).

IMPLEMENTATION

This provides a short explanation of the system's algorithm deployment specifics. There are four algorithms:

Decision Tree, Multinomial Naive Bayes, Random Forest and Word Bag.

Decision Tree

The technical details of the Decision Tree method are presented in Algorithm 1. Require: identifies bots by number of followers, friends, name of the screen, description, location and verified #tag

Ensure that the above function is converted into binary values.

- The whole training set is regarded as the root.
- Information gathered is utilized to determine which characteristic each node should be labeled with.
- On the training instance, recursively build each sub-tree that would be categorized along the tree's path.
- Label that node yes or no if all positive or negative occurrences remain.
- If no attributes are left, the label with the most votes is retained at that node.
- If no instances remain, label the parent's training instance with a majority vote.

Multinomial Naive Bayes

Method 2 provides Multinomial Naive Bayes algorithm implementation details. Require: System identifies bots based on number of people, friends, names of the screen, descriptions, locations and checks. Ensure that the above function is converted into binary values. Hypothesis is that the particular account is bot.

- P(h|d) is the probability of hypothesis h given the data
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true.
- P(d) is the probability of the data (regardless of the hypothesis).
- P(h|d) =(P(d|h) *P(h)/P(d) return (P(hjd)).

Random Forest

Method 3 provides specifics on the development of bots using the Random Forest algorithm. Require: System identifies bots based on number of people, friends, names of the screen, descriptions, locations and checks.Ensure: Conversion of the aforementioned function to binary values.

- Choose randomly k characteristics for certain m features
- Calculate the node d with the best spilt point among k
- Disperse the node into daughter nodes using the optimal dividing point.
- Repeat 1 to 3 until nodes l have achieved their number

- Forest is constructed by repeating steps 1 to 4 n to generate the n tress number to provide random forest
- Use each tree on the test function and save the results
- Calculate votes for every forecast result
- Consider the top result voted as the final forecast

Bag of Words

Algorithm 4 provides instructions for implementing bots using the algorithm Bag of Words. Require: Detects bots by the number of followers, friends, name of the screen, description, place and verified

Ensure: Converts the above to binary values

- Data about known bot accounts are gathered
- Designed vocabulary. It comprises of one word, two words or more. The representation of hash is utilized.
- Comparing test data to vocabulary and saving it as a binary vector
- Scoring techniques include the number of times that word occurs, and the frequency with which each word appears in a document from all of the words in the document.

CONCLUSION

We developed an algorithm in our research that identifies Twitter bots. Bag of words method for train data was the best model with an accuracy of 96.7 percent compared to Decision Tree, Naive Bayes and Random Forest 96.65 percent for test data. Thus, word algorithms were used to real-time data and the Twitter bots have been detected effectively.

REFERENCES

[1] Van Der Walt, Estée, and Jan Eloff. Using machine learning to detect fake identities: bots vs humans. IEEE Access 6 (2018): 6540-6549.

[2] Sever Nasim, Mehwish, Andrew Nguyen, Nick Lothian, Robert Cope, and Lewis Mitchell. Real-time detection of content polluters in partially observable Twitter networks. arXiv preprint arXiv:1804.01235 (2018).

[3] Khalil, Ashraf, Hassan Hajjdiab, and Nabeel Al-Qirim. Detecting Fake Followers in Twitter: A Machine Learning Approach. International Journal of Machine Learning and Computing 7, no.6(2017).

[4] Wetstone, Jessica and Sahil R. Nayyar. I Spot a Bot: Building a binary classifier to detect bots on Twitter. (2017).

[5] Karataş, Arzum, and Serap Şahin. A Review on Social Bot Detection

Techniques and Research Directions. In Proc. Int. Security and Cryptology Conference Turkey, pp. 156-161. 2017.

[6] Chavoshi, Nikan, Hossein Hamooni, and Abdullah Mueen. Identifying correlated bots in twitter. In International Conference on Social Informatics, pp. 14-21. Springer, Cham, 2016.

[7] Perdana, Rizal Setya, Tri Hadiah Muliawati, and Reddy Alexandro. Bot spammer detection in Twitter using tweet similarity and time interval entropy. Jurnal Ilmu Komputer dan Informasi 8, no. 1 (2015): 19-25.

[8] Haustein, Stefanie, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated bot accounts on T witter. Journal of the Association for Information Science and Technology 67, no. 1 (2016): 232-238.