User Behavior Based Bot Detection System Using Machine Learning Classifier

Shweta Jain[1], Prof. Sapna Jain Choudhary[2]
[1]Dept of CSE
[2]Professor, Dept of CSE
[1, 2] Shri Ram Group of Institutions Jabalpur, Madhya Pradesh, India.

Abstract- *Fake accounts and automated bot activities on social media have also had significant political impact in the recent years. The interference of Russian government in the 2016 United States Elections has been a huge topic of discussion and debate in the last two years. Various instances of Russian bots spreading fake news and propaganda on Facebook and Twitter have come to light since the elections. Both Facebook and Twitter have come under heavy scrutiny since then, for not taking appropriate actions against these malicious actors. These online social media networks have a very large sphere of influence with their presence in the life of the general public globally and the presence of automated bots on these platforms and their actions have had a negative impact which calls for actions to solve the problems mentioned earlier. Typical features used in some of the methods need a long duration of activities (e.g. weeks) which makes the detection process useless, as the bots can initiate a fair amount of harm before being detected. Moreover, bots are becoming smarter. They mimic humans to avoid being detected and suspended and increase throughput by creating many accounts. The different sets of features focusing on the various aspects of user behavior, content and basic user profile information were considered for training and testing on two labeled datasets of Twitter accounts. We use string/ tweets compression technique for detecting the behavior of Bot accounts on network for classifying as Bots or Nonbots activities.*

Keywords- Detection of Bots; User Behaviors, Machine Learning; Twitter Accounts, String Compression, Logistic Regression.

I. INTRODUCTION

In recent years, social media platforms such as Twitter or Facebook have gained a large level of both popularity and influence among millions of users due to the benefits of publishing, propagating and exchanging large volumes of multimedia content along the network. Therefore, these platforms allow users to establish a digital community as remarked in [1], which has made possible not only to discover and embrace new relationships but to maintain and boost existing ones. On the other hand, due to both the great influence these platforms have on the lifestyle of people and its evolving as a potential communication tool, they have exponentially promoted its attraction for marketing and commercial purposes by analyzing the behavior and opinion of users in different topics or events such as political elections. Consequently, numerous research studies have been fostered in the social media field with different purposes including sentiment analysis [2], traffic control [3], or consumer behavior mining [4]. However, the considerable growth of social media platforms has also provoked the desire of altering people's opinion in certain topics by spreading propaganda or bias information. Many of these controlling procedures are carried out by Bots which are widely described in numerous investigations [5], such as automatic systems which are capable of generating and spreading multimedia content throughout the network without the supervision of a human being. Furthermore, with the disruptive growth of Artificial Intelligence (AI) algorithms, the identification of bots or non-reliable sources has become a crucial challenge to be investigated. It raised many studies and publications with the goal of building robust automatic systems to improve the quality of experience of consumers in such platforms by reducing their privacy risks as well as increasing the trustworthiness on the platform itself at the same time.

The social media platform is a base to exchange information; it is confined to differentiate between human user posts and tweets generated by bots. Bots gradually spread information on the social platforms to create a trend that could change public opinion [6]. Social media platforms can be a big source of user messages, and user's private opinions can be disclosed in social platforms [7] and may misused by social bots to create serious threats to financial systems [8]. Bots can distribute falsify in social media and create rumors in a community of users [9]. Social media platforms are the main source of news and narratives about some events in the world [10] and bots can significantly affect views of these events. Bots unfurl low-quality information and equivocal news, which can be complex to detect based on fulfilled. Social media bots can be created to target various audiences [11]. One study identified multiple types of spambots; including

promoter bots, URL spam bots, and fake followers [12]. Promoter bots spend several months promoting specific hashtags to create fake trends or promote specific products. For example, they can promote an item for sale on Amazon or help a political candidate win an election. URL spam Bots spread scam URL links by embedding them in retweets they create from legitimate user posts.

Other types of bots include fake followers on social media and fake reviewers of specific products. Many studies have attempted bot detection in recent years. For example, [13] used an unsupervised learning approach to detect bots that distribute malicious URL links using URL shortening services. Based on this study, URL sharing bots use constant tweet duplication of legitimate users at a specific time to spread malicious URLs. Their results suggest that about 23 percent of accounts that use URL shortening services are bots. Another popular bot detection service is "Botmeter" 1, which uses a supervised learning approach to detect social bots. Botmeter uses metadata related to each twitter account, such as network, user, and temporal features, to feed a Random Forest classifier algorithm. Network features show how information diffusion happens among multiple groups of users. User features are user name, screen name, the creation time of account, and geographic location. Temporal features show patterns in a tweet's time generation.

A community detection approach [14] is used to detect online activities of a group of online users who share similar ideas. For example, DeBot [15] is a bot detection service that uses the correlation of activities between different accounts. Application of Benford's law is used for bot detection by analyzing online behaviors of bots in [16]. One of the drawbacks of previous models is that we need different information about each user's account, such as user and network features, to differentiate a human account from a bot account. However, in real-world scenarios, we need to detect bot accounts in the early stage of posting comments on social media to prevent the spread of misinformation in online communities. In this work, we improve previous models for social media bot detection by using minimum information about each online user's posts to detect Social Media bots.

II. RELATED WORK

Yang et al. (2013) [16] proposed ten novel features three graph-based, three neighbor-based, three automation-based, and one timing-based to infer whether a Twitter account is genuine or a spambots. Graph-based and neighbor-based features were useful for finding malicious bots that attempt to evade profile-based features by adjusting their own social behaviors, whereas automation-based and timing-based features were used to detect social bots that attempt to evade content-based detection approaches by increasing the number of their human-like tweets. These novel features were evaluated using four different classifiers, namely RF, decision trees (DTs), Bayes networks (BNs), and Decorate (DE).

Dickerson et al. (2014) [17] employed an ensemble of classifiers including SVMs, Gaussian NB, AdaBoost, gradient boosting, RFs, and extremely randomized trees based on tweet syntax, tweet semantics (at the individual user and neighbourhood levels), user behavior, and network-centric user features. They also applied sentiment analysis on a per-user basis over a variety of topics. Moreover, they identified topics discussed by employing latent Dirichlet allocation (LDA). They employed kernel principal component analysis (PCA) for de-noising and dimensionality reduction. Their sentiment features improved the accuracy of the classifier. Interestingly, they found that when a user's proportion of tweets with sentiment is between 0.5 and 0.9, they are much more likely to be a human than a bot.

Oentaryo et al. (2016) [18] utilized four classifiers NB, RF, SVM, and logistic regression (LR) to distinguish between human accounts and three types of bots; namely, broadcast, consumption, and spambots. They considered profile and follow features, and both static (i.e., time-independent) and dynamic (i.e., time-dependent) tweet-based features.

Fazil & Abulaish (2018) [19] identified six new features and redefined two features. The newly identified features included one content-based, three interaction-based, and two community-based features, while the redefined features were content-based. These features were fed to RF, DT and Bayesian network (BN) classifiers to distinguish between automated spammers and legitimate users. They found that interaction- and community-based features were the most effective for spam detection, whereas metadata-based features were the least effective ones. The interaction based features focused on the followers of a user rather than on the other users they follow, since these features cannot be determined by the user. This approach can be considered to represent a hybrid approach as it depends on graph-based features as well as content-based features.

Begenilmi¸s & Uskudarli (2018) employed supervised ML algorithms to detect organized behaviors based on RF, SVM, and LR. They employed user- and temporal-based features to distinguish between three different categories. Their method utilized features of collective

behavior in hashtag-based tweet sets, which were collected by querying hashtags of interest.

Al-Qurishi et al. (2018) used three levels of features namely, content, graph, and profile activities in order to detect anomalous behaviors in OSNs. The key concept in this study was leveraging contextual activity information among OSN users. They considered iterative regression, RF, J48, regression, and SVM classifiers. They also employed PCA, along with a ranking methodology to weight these features according to their relative importance in the examined dataset. Moreover, to detect a topic-based behavior, they used LDA. Accordingly, they found that all OSN users appear to be remarkably similar until their corresponding activity traits are considered, at which point significant contradictions arise.

Although the detection of social bots is a challenging task, there are some works that analyzed the characteristics and behavior of bots and offered various features that are recurrent in the majority of works. For example, verified accounts are guaranteed to be human users. Moreover, the ratio of followers to following and the age of the account are considered discriminative characteristics in detecting bots since bots generally mass-follow and have short life span. The following features are mainly used by tweet-based bot detection techniques to distinguish between tweet-based bots and humans accounts:

- ID: It represents the unique identifier of the tweet.
- User: It represents the user who posted the tweet.
- Created_at: It indicates the UTC time when the tweet is created.
- Text Tweet: It refers to the body of the tweet.
- Length of Tweet: It gives the number of characters in the tweet.
- #Hashtags: It indicates the number of hashtags in the tweet.
- #URLs: It indicates the number of URLs in the tweet
- in_reply_to_status_id: If the tweet is a reply, this feature represents the original tweet's ID.
- in_reply_to_user_id: If the tweet is a reply, this feature represents the author of the original tweet.
- Coordinates: It represents the geographic location of the tweet.
- Favorite_Count: It indicates how many times the tweet has been liked by Twitter users.
- Retweet_Count: It is the number of times the tweet has been retweeted
- Reply Count: It is the number of times the tweet has been replied to.
- Favorited: Boolean feature, which holds true when the tweet is liked by the authenticating user.
- Retweeted: Boolean feature, which holds true when the tweet is retweeted by the authenticating user.
- Possibly_sensitive: Boolean feature, which holds true when the tweet contains a link.

The detection of bots, as well as their interactions with their communities and the rest of the world, is essential. The paper focuses on plotting a network of Twitter users, based on a particular hashtag, and detecting the communities in it, followed by detecting and locating the bots in these communities. In addition to this, sentiment analysis is conducted on normal users' tweets as well as the bots in these communities. This paper also aims to identify the overall sentiment of the communities, and thus provide promising conclusions relating to bot behavior in the overall network.

III. PROPOSED WORK

We initially plan for building a model for classification of Bots and Nonbots using a supervised learning model. The different subsets of features which can be used for training a model based on following three aspects a user account profile:

1. User Profile Data
2. Content
3. User Behaviour

Classifying text via other classifiers may take more processing and time. This is convenient since we can just pass feature values as inputs into the network and get a classification.

1. The user profile data display certain basic information about the user account such as name, location, profile picture and a short biography along with number of tweets, retweet and replies posted. These details provide a rough idea about the account which might be indicative of whether a Twitter account is being operated by a bot or a genuine human user.

2. Contents are the statistics and patterns about some of the Twitter-specific attributes such as #hashtags, @mentions and URLs as well as the actual content of the tweets. The advantage of using these features is that they are independent of the language in which tweets are posted and provide some idea about the nature of content posted by the account under consideration

3. To gain more insights about a Twitter user, some subtle features need to be defined that capture the behaviour of the

account over a longer period of time. It basically describes the posting behavior of the Bots.

3.2 Proposed Algorithm

Step 1: Import important libraries.
Step 2: Read the dataset of user's profiles. Four types of datasets are used.

a) genuine_accounts-- Stores the information of genuine user profiles.
b) fake_followers-- Stores the information of fake follower's profiles.
c) Social_spambots—Information about social spam profiles.
d) Traditional_spambots—Information about traditional spambots profiles.

Step 3: Read the dataset of tweets. Four types of datasets are used.

a) genuine_accounts_tweets-- Stores the information of genuine tweets.
b) fake_followers_tweets-- Stores the information of fake tweets.
c) Social_spambots_tweets—Information about social spam profiles.
d) Traditional_spambots_tweets—Information about traditional spambots tweets.

Step 4: Display the shape of all the datasets.
Step 5: Create a Sequence string for tweet (T), retweet (R) and reply/answer to tweets. (A).
Step 6: Crete the string for each account based on behavior activity for each account.
Step 7: Return a compressed string "R" for the sequence string.
Step 8: Convert string object to byte object.
Step 9: Create separate strings of T, R and A as "OS". Also create a separate compressed string "CS".
Step 10: find the string Ratio.
String ratio = (Original string size of T, R and A)/ Compressed String.
= OS/CS.
Step 11: Create a confusion matrix for OS and String ratio.
Step 12: Split the dataset for training and testing.
Step 13: Train the classifier using logistic regression model.
Step 13: evaluate the matrix using result.

3.3 Flowchart of Proposed Method

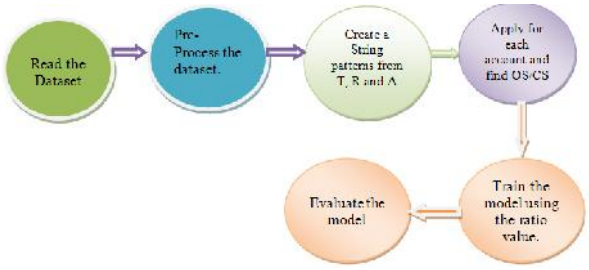

Figure 3.1: Proposed Architecture.

IV. RESULTS WORK

We began with over 40 possible features; we found that the most informative features are lexical diversity, friend to follower ratio, replies count, quote count, statuses count, and tweet frequency. The fact that these attributes are the most informative was expected. The OS to CS ration is the most valuable facts to evaluate the models based on accuracy.

[1] The system has been realized by implementing the pre-processing stage of feature extraction. The machine learning models and the data visualization are implemented in python using Jupyter. For the machine learning library, we used scikit and tensor flow.
[2] A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 4.1: Confusion matrix.

True Positives (TP) - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Once you understand these four parameters then we can calculate Accuracy, Precision, and Recall.

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = TP+TN/ (TP+FP+FN+TN)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = TP/ (TP+FP).$$

In the table below, the results are analyzed for the existing method and proposed method. The results are analyzed by calculating accuracy of the model.

Table 4.1: Performance Evaluation.

Method	Testing Accuracy (%)
Random Forest	**91.77**
CNN with hybrid Model	**97.04**
Proposed Method	**97.89**

It is observed that proposed classifier achieved best accuracy due to the past tweet behavior features for which string patterns are created.

REFERENCES

[1] J. Knauth, ‘‘Language-agnostic Twitter-bot detection,’’ in Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP), 2019, pp. 550–558.

[2] K. Shuang, H. Guo, Z. Zhang, J. Loo, and S. Su, ‘‘A word-building method based on neural network for text classification,’’ J. Exp. Theor. Artif. Intell., vol. 31, no. 3, pp. 455–474, May 2019.

[3] J. Zhu, C. Huang, M. Yang, and G. P. Cheong Fung, ‘‘Context-based prediction for road traffic state using trajectory pattern mining and recurrent convolutional neural networks,’’ Inf. Sci., vol. 473, pp. 190–201, Jan. 2019.

[4] A. S. M. Alharbi and E. de Doncker, ‘‘Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information,’’ Cogn. Syst. Res., vol. 54, pp. 50–61, May 2019.

[5] J. Pizarro, ‘‘Using N-grams to detect bots on Twitter,’’ in Proc. CLEF, Working Notes, 2019, pp. 1–10.

[6] M. Forelle, P. N. Howard, A. Monroy-Hernandez, and S. Savage, ´ “Political bots and the manipulation of public opinion in venezuela,” CoRR, vol. abs/1507.07109, 2015.

[7] A. Mosallanezhad, G. Beigi, and H. Liu, “Deep reinforcement learning based text anonymization against private-attribute inference,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (Hong Kong, China), pp. 2360–2369, Association for Computational Linguistics, Nov. 2019.

[8] N. Yousefi, M. Alaghband, and I. Garibay, “A comprehensive survey on machine learning techniques and user authentication approaches for credit card fraud detection,” arXiv preprint arXiv: 1912.02629, 2019.

[9] Z. Rajabi, A. Shehu, and H. Purohit, “User behavior modelling for fake information mitigation on social web,” in Social, Cultural, and Behavioral Modeling (R. Thomson, H. Bisgin, C. Dancy, and A. Hyder, eds.), (Cham), pp. 234–244, Springer International Publishing, 2019.

[10] T. A. Oghaz, E. c. Mutlu, J. Jasser, N. Yousefi, and I. Garibay, “Probabilistic model of narratives over topical trends in social media: A discrete time model,” in Proceedings of the 31st ACM Conference on Hypertext and Social Media, HT ’20, (New York, NY, USA), p. 281–290, Association for Computing Machinery, 2020.

[11] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” CoRR, vol. abs/1701.03017, 2017.

[12] Z. Chen, R. S. Tanash, R. Stoll, and D. Subramanian, “Hunting malicious bots on twitter: An unsupervised approach,” in Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II (G. L. Ciampaglia, A. J. Mashhadi, and T. Yasseri, eds.), vol. 10540 of Lecture Notes in Computer Science, pp. 501–510, Springer, 2017.

[13] R. Abdolazimi, S. Jin, and R. Zafarani, “Noise-enhanced community detection,” in Proceedings of the 31st ACM Conference on Hypertext and Social Media, HT ’20, (New York, NY, USA), p. 271–280, Association for Computing Machinery, 2020.

[14] N. Chavoshi, H. Hamooni, and A. Mueen, “DeBot: Twitter bot detection via warped correlation,” in IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, pp. 817–822, 2016.

[15] L. Madahali and M. Hall, “Application of the benford’s law to social bots and information operations activities,” in 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp. 1–8, 2020.

[16] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., & Dai, Y. (2014). Uncovering social network Sybils in the wild. ACM Transactions on Knowledge Discovery from Data (TKDD), 8, 1–29.

[17] Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. (2014), "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?" In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM '14 (pp. 620–627). IEEE.

[18] Oentaryo, R. J., Murdopo, A., Prasetyo, P. K., & Lim, E.-P. (2016),"On profiling bots in social media", In International Conference on Social Informatics (pp. 92–109). Springer.

[19] Fazil, M., & Abulaish, M. (2018), "A hybrid approach for detecting automated spammers in twitter", IEEE Transactions on Information Forensics and Security, 13, 2707–2719.