Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 04, April 2021: 2276-2285

Research Article

Detecting Malicious Tweet Bots using Machine Learning Algorithms

Sandra Johnson[1], Sai Charan K[2], Sai Chetan K[3], Sri Saideep V[4]

[1,2,3,4]Department of Computer Science and Engineering, R.M.K. Engineering College

Abstract

Social bots are social accounts operated by computer programs in online social networks, which can execute corresponding operations based on a set of procedures. With the large increase in the number, speed and variety of user data in online social networks, attempts, have been made to collect and analyze these big data. Social bots, for example, have been used to conduct automated analytical services and provide improved quality of service to consumer. Nonetheless, malicious social bots were often used to disseminate false information, and this can have real world implications. It is important to identify, delete malicious social bots in online networks. This introduces a novel approach for detecting malicious social bots for features' selection based on the transfer likelihood of click stream sequences and semi-supervised clustering. This approach not only analyses the likelihood of change from user interaction click sources but also considers the interactions time features. Finding from our studies on real online social network platform show that the accuracy if detection of different types of malicious social bots based on transformation likelihood of user activity click streams improves by an average of 12.8 percent compared to the method of detection based on quantitative user behavior analysis.

Introduction

Social bots are social accounts operated by computer programs in online social networks, which can execute corresponding operations based on a set of procedures. The growing usage of mobile devices (e.g. Android and iOS apps) has also helped to increase the frequency and complexity of user engagement across social networks. It is demonstrated by the tremendous quantity, speed and variety of data generated from the user base of the broad online social network. Internet bots have been widely deployed to boost the consistency and reliability of the social network services data collection and analysis. The social bot SF QuakeBot is designed to produce San Francisco Bay earthquake data and to analyze earthquake-related information in social networks in real time. Public opinion on social networks and large user data can however also be used or disseminated for malicious or unethical purposes. Automatic social bots cannot reflect the real needs and interest of actual human beings in

online social networks, so they are usually seen as malicious ones. For instance, some fake social bots accounts created to imitate a normal user's profile, steal user data and violate their privacy, disseminate malicious or false information, comment maliciously, promote or advance some political or ideological agenda and propaganda, and influence stock markets and other societal and economic markets. These practices may have an adverse effect on social networking site security and stability. In this paper, we aim to detect malicious social bots on social network in real time by algorithm. The second segment looks briefly at related research. The third section introduces the method of detecting algorithms for malicious social bots, followed by the fourth section of the experiment and the outcome review.

Diverse approaches have been used in previous research to protect online social network security. User behavior is the most direct representation of user intent, as different users have different routines, interests, and online behaviors (e.g., how one clicks or type, as well as typing speeds). In other words, to profile and classify specific users, we will be able to mine and analyze information concealed in the online actions of the user. We do need to be mindful, however, of situational factors that can play a role in changing the online behavior is complex and their environment continues to evolve. Observable external device context environment and secret user knowledge system. In order to accurately differentiate social bots from real users, identify malicious social bots and minimize the harm of malicious social bots, we need to acquire and evaluate user activity social situation and compare and understand the discrepancies between malicious social bots and real users in dynamic behavior.

Literature Review

Surveying literature is the most important phase of the software development process. Before the tool is developed it is important to measure the time factor, economy and intensity of the market. When these problems are addressed, the next ten steps will be to determine which operating system and language to build the device to be used. Once the programmers begin designing the method, they will need a lot of external help. This support is from senior programmers, from books or from websites. Before constructing the system, the above consideration for the creation of the proposed structure shall be considered.

The authors in [10] used with an attribute-based encryption; the user is marked for data encryption and decryption with the aid of certain attributes and their functions. Current techniques based on attribute-based encryption have shown that if the access structure of the user contains a large amount of attribute information labeled as Don't Care, then the

encryption pairing process has poor calculation efficiency and cipher text information redundancy. In this paper, we suggested a hierarchical multi-authority attribute-based encryption on groups of prime order to solve these problems. The encryption technique is based on a polycentric attribute authorization scheme based on an AND gate access structure, with a single attribute index set in the scheme by each attribute authority throughout to form a binary tree. The parent node's state value can be calculated by the condition of its child node in an access tree attribute. The attribute-based encryption was designed to decrease the effort for decryption and compress the redundant information in the cipher text. Our strategy for encryption has both theoretical and functional importance in the constructions of the “big universe” scheme.

The authors in [14] have applied Social experiences occur in situations that influence the attitudes and expectations of the people. Online Social Network (OSN) users now create a vast amount of content that is focused on social interactions. In order to detect automated information transmitted in OSN, we built a wavelet-based model that classifies users as human, legitimate robot, or malicious robot due to spectral patterns obtained from the textual content of users. We construct a Lexicon-based Coefficient Attenuation. Using Random Forest Algorithm, we can define two set of real Twitter datasets. The result using this model will achieve an average of 94.47 percent, taking this into account two scenarios: single theme and miscellaneous.

The authors in [18] has described following. With the tremendous growth and volume of online social networks and their apps, along with the large number of socially linked users, the true semantic importance of published content for user activity detection has become difficult to understand. By knowing the contextual history, differentiating between different groups is unworkable in terms of their importance and mutual relations, or to recognize the most important group members at large. In this paper, we suggest an integrated framework for the analysis of social media content that leverages three layers of functionality, i.e. user-generated content, social graphic interactions and user profile activities, to evaluate and identify anomalous behaviors that vary significantly from the rules in large-scale social networks. For a deeper understanding of the various user activities in the identification of highly adaptive malicious users, many forms of analyses were carried out. We tried a new approach to the data extraction and classification method to contextualize large-scale networks properly. We have also compiled a large number if twitter and YouTube user profiles, along with about 13 million channel events. Extensive analyses were performed for

both social networks on the real-world app behavior datasets. The findings of the assessment demonstrate the efficacy and usefulness of the solution suggested.

The authors in [21] described, there are few numbers of individuals keeping accounts on social media platforms (SMPs), using for malicious purpose and illegal uses. Sadly, to date, truly little work has been done to detect human-created fake identities, especially on SMPs. Conversely, there are several examples of detecting malicious social bots in twitter platform are identified using machine learning algorithm. To detect fake human accounts using machine learning models by the help of programmed features such as the 'friend-to-followers ratio' in the hope of advancing the effective detection of man-made false identities on SMPs.

The authors in [9] have mentioned Online Social Networks (OSN) slowly incorporate financial capabilities by allowing real and virtual currency use possible. These function as new channels for hosting a range of business activities such as online marketing events, where participants can potentially receivevirtual currency as incentives for participating in these events. All OSNs and business partners are worried when attackers use a series of accounts to obtain virtual currency from these events, rendering these events useless and leading to substantial financial losses. Proactively identifying these fake accounts before the online marketing activities is of considerable importance and consequently reduces their priority to behonoured.

The authors in [11] considered Social bots are the most popular type of social network malware. They can create false messages, spread misinformation and even exploit public views. Social bots are developed to provide false information around internet. Bot detection aimed at separating bots from humans and in recent years it is gaining more and more attention. In this paper we propose an improved deep model of behavior (BeDM) for bot detection. The proposed model considers user content as transient text data to derive latent temporal patterns instead of plain text. In addition, BeDM fuses knowledge about content and actions using deep learning methods. To the best of our knowledge, this is the first trail that applies bot detection to deep neural networks. Twitter-collected studies on real world dataset often show the feasibility of our proposed model.

The authors in [22] have described Random Forest and Decision Tree as two models are used to compare the classification results of two models. Random Forest created by Leo Breiman is a collection of unpruned grading or regression trees made from the random selection of training data samples. In the induction process, random traits are picked. Prediction is made by aggregating the ensemble’s predictions (majority voting for classification or averaging for regression). Random Forest typically demonstrates a substantial increase in efficiency as

opposed to a single tree classifier like c4.5. The rate of generalization error it produces contracts favorably with adaboost, though it more resilient to noise.

Methodology Used

This paper explains detecting malicious social bots using 3 different algorithms (Decision tree, Random forest, Pattern). Using this algorithm, we detect 5 values for each. They are Recall, F1, FPR, Precision, and Accuracy. This defines an individual graph, at final it compares the graph and represents accuracy among three.

3.1 Algorithm Used

We are proposing a novel approach for detecting malicious social bots, Using Decision tree, Random forest, Pattern Algorithm to find Accuracy to detect bot.

3.2 Algorithm Description

Random Forest algorithm:

Random forest is best learning method for classification, regression. Random Forest will construct a decision tree at training time and prediction of mean for individual trees. Random decision forest will train data set with decision trees. This was first invented by Tin Kam Ho, he used random subspace method. “Stochastic Discrimination” approach was proposed by Eugene Kleinberg.

This algorithm was developed by Leo Breiman and Adele Cutler; they registered “Random Forests” as a trademark. Ho first introduced Breiman's “Bagging” idea and later independently by Geman and Amit, to construct a decision tree with controlled variance.

In training for random forests involves in technique like bootstrap, aggregating, bagging. This involves in the training set X=x1……. xn with responses Y=y1……. yn, bagging will select a random sample with training set and tree fits to these samples:

For b = 1… B:

It will replace and training examples from X, Y; call these Xb, Yb.

After training and predicting samples x’ will be made average of the prediction from all individual regression trees on x’take the majority vote in the case of classification trees.

This bootstrapping process leads to better performance, because it decreases variance of model, it doesn’t increase bias. The prediction of one tree is not possible in its training set, if the trees are notcorrelated. Training many trees based on a single training set it will produce a strongly correlated tree. Bootstrap sampling is way of representing de-correlating trees.

The estimation of uncertainty of a prediction can be made a standard deviation of the predictions from every individual regression tree on x’.

The number of trees or samples is a free parameter. Thousands of trees are used based on the size and nature of training set. Number of optimal trees B can be found using cross-validation. The mean prediction of error on training of data, tree doesn't have xi in their bootstrap sample.

Decision Tree Algorithm:

Decision tree algorithm is part of family of supervised learning algorithm. Unlike other algorithms, decision tree algorithm is also used to solve regression problem and classification problem.Using Decision Tree to create a training model which uses to predict class or value of a target by learning rules from training data.Decision Tree algorithm can be easily compared with other classification algorithms. Decision Tree algorithm tries to solve problems using tree representation. Every internal node of tree corresponds to an attribute, and every leaf node corresponds to a class label.Pseudo code of Decision Tree algorithm. Take best attribute values of dataset at root of the tree.Split training data into subsets. These individual subsets should contain same values for an attribute.Repeat steps 1 and 2 on each subset till you find a leaf node in all of branches of tree.In Decision Tree, we need to start predicting classes from root of the tree until we reach leaf node with predicted class value. We know to use the modeled decision tree too predict target class. Now we know to create decision tree model.

Assumptions while creating decision tree:

There are some assumptions made from decision tree.Firstly, whole training set is considered as root.Feature values are preferred as categorical. If value is continuing, then they are discretized prior to build.Records are distributed based on the recursively on attribute values.Internal node of tree is done by using statistical approach

Pattern Approach Algorithm:

Pattern Approach Algorithm will involve in identification of patterns is the automatic identification of trends and regularities. Recognition of patterns is closely associated with Artificial Intelligence and Machine Learning.

Along with techniques such as data mining and the exploration of information in databases, these concepts are used interchangeably. These are however distinguished: machine learning is one approach to pattern recognition, while other approaches include hand-crafted (unlearned) rules or heuristics; and pattern recognition is one approach to artificial intelligence, while other approaches are described in symbolic artificial intelligence.

The area of pattern recognition involves the automated discovery of data regularities using computer algorithms and the use of these regularities to take steps such as classifying the data into different categories.

Pattern recognition algorithm is trained by sample data, using the trained data it will have set of predictions. By the help of trained data, it will predict spam or non-spam comments and tweets.

The algorithms will involve in the use of trained data to predict testing data. Using this approach, malicious tweets will be detected. It will have high accuracy among other algorithms.

Result

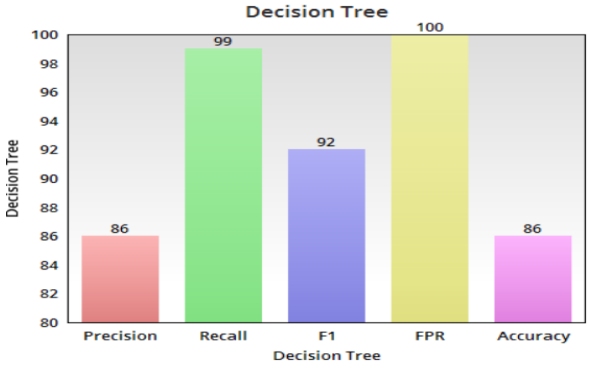

Figure 1 Precision, recall, f1, FPR, accuracy for Decision tree,

In decision tree algorithm, we have very less accuracy and precision. FPR value is 1; Recall value was almost equal to 1. This algorithm was less effective compared to other algorithms.

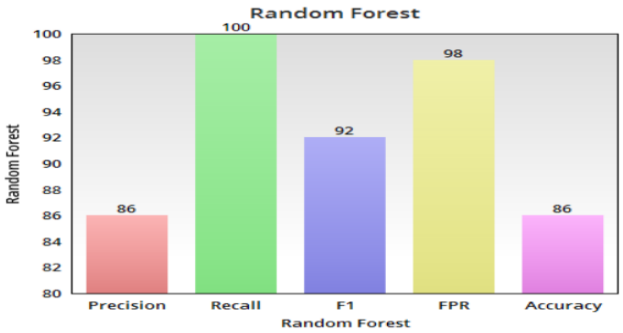

Figure 2 Precision, recall, f1, FPR, accuracy for Random Forest.

In Random Forest algorithm we have very less accuracy compared to Pattern approaches. F1 value is 0.925, Recall value is 1, and FPR value is almost equal to 1. Precision value is 0.86 which is equal to accuracy.

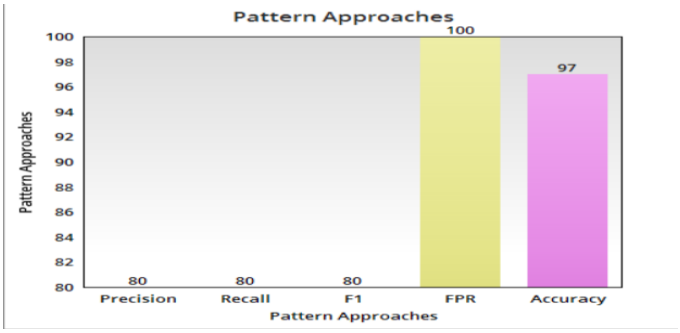

Figure 3 FPR, Accuracy for Pattern Approach.

In Pattern approach algorithm there is no value of Precision, Recall, and F1. The FPR value is equal to 1, Accuracy value is almost 1. This algorithm has the highest accuracy percentage compared to other algorithms.

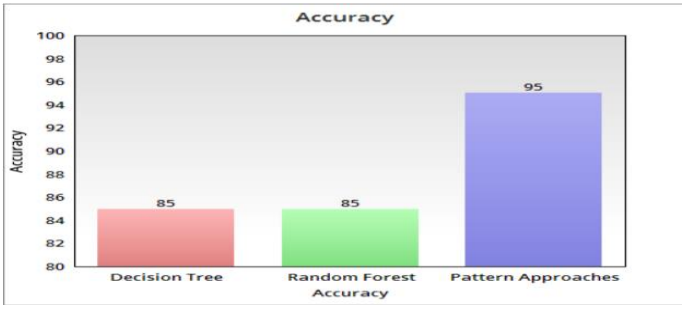

Figure 4. Accuracy for different Machine Learning Algorithms.

Above Fig 4. Represents the Accuracy values of all three Algorithms. They are Decision Tree, Random Forest, and Pattern Approach. Among the three algorithms, Pattern approach has highest Accuracy value, which is used to find the malicious accounts accurately.

Discussion

The existing system involves in detecting malicious social bots-based on click stream sequence. It detects based on user behavior, when it comes to accuracy we can't define. But when it comes to Detecting Malicious Tweet Bots using Machine Learning Algorithms it mainly focuses on accuracy, precision, recall, F1, FPR. Using various algorithms like Decision tree, Random forest, pattern approach it define accuracy for each algorithm and final it will compare with other accuracy values.

Conclusion

Furthermore, we expand this model to consider more fine-grained user activities beyond attacker classification. The new model incorporates the app cluster inherent Hierarchical structure. Service providers should explore manipulating user behaviors and categories as an

outline through a visualization tool, while monitoring fine grained user activity trends throughout each category. Our system does not require prior knowledge or expectations of groups of users (unsupervised), thus capturing unintended or previously unknown behaviors effectively. Using case studies on real-world online social networks we prove its effectiveness.

References

[1] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, ‘‘A new approach to bot detection: Striking the balance between precision and recall,’’ in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, San Francisco, CA, USA, Aug. 2016, pp. 533–540.

[2] C. A. De Lima Salge and N. Berente, ‘‘Is that social bot behaving unethically?’’ Commun. ACM, vol. 60, no. 9, pp. 29–31, Sep. 2017.

[3] M. Sahlabadi, R. C. Muniyandi, and Z. Shukur, ‘‘Detecting abnormal behavior in social network Websites by using a process mining technique,’’ J. Comput. Sci., vol. 10, no. 3, pp. 393–402, 2014.

[4] F. Brito, I. Petiz, P. Salvador, A. Nogueira, and E. Rocha, ‘‘Detecting social-network bots based on multiscale behavioral analysis,’’ in Proc. 7th Int. Conf. Emerg. Secur. Inf., Syst. Technol. (SECURWARE), Barcelona, Spain, 2013, pp. 81–85.

[5] T.-K. Huang, M. S. Rahman, H. V. Madhyastha, M. Faloutsos, and B. Ribeiro, ‘‘An analysis of socware cascades in online social networks,’’ in Proc. 22nd Int. Conf. World Wide Web, Rio de Janeiro, Brazil, 2013, pp. 619–630.

[6] H. Gao et al., ‘‘Spam ain’t as diverse as it seems: Throttling OSN spam with templates underneath,’’ in Proc. 30th ACSAC, New Orleans, LA, USA, 2014, pp. 76–85.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, ‘‘The rise of social bots,’’ Commun. ACM, vol. 59, no. 7, pp. 96–104, Jul. 2016.

[8] T. Hwang, I. Pearce, and M. Nanis, ‘‘Socialbots: Voices from the fronts,’’ Interactions, vol. 19, no. 2, pp. 38–45, Mar. 2012.

[9] Y. Zhou et al., ‘‘ProGuard: Detecting malicious accounts in social network-based online promotions,’’ IEEE Access, vol. 5, pp. 1990–1999, 2017.

[10] Z. Zhang, C. Li, B. B. Gupta, and D. Niu, ‘‘Efficient compressed ciphertext length scheme using multi-authority CP-ABE for hierarchical attributes,’’ IEEE Access, vol. 6, pp. 38273–38284, 2018. doi: 10.1109/ACCESS.2018.2854600.

[11] C. Cai, L. Li, and D. Zengi, ‘‘Behaviour enhanced deep bot detection in social media,’’ in Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI), Beijing, China, Jul. 2017, pp. 128–130.

[12] C. K. Chang, ‘‘Situation analytics: A foundation for a new software engineering paradigm,’’ Computer, vol. 49, no. 1, pp. 24–33, Jan. 2016.

[13] Z. Zhang, R. Sun, X. Wang, and C. Zhao, ‘‘A situational analytic method for user behaviour pattern in multimedia social networks,’’ IEEE Trans. Big Data, to be published. doi: 10.1109/TBDATA.2017.2657623.

[14] S. Barbon, Jr., G. F. C. Campos, G. M. Tavares, R. A. Igawa, M. L. Proença, Jr., and R. C. Guido, ‘‘Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets,’’ ACM Trans. Multimedia Comput., Commun., Appl., vol. 14, no. 1s, Feb. 2018, Art. no. 26.

[15] J. Y. Park, N. O'Hare, R. Schifanella, A. Jaimes, and C.-W. Chung, ''A large-scale study of user image search behavior on the Web,'' in Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst., Seoul, South Korea, 2015, pp. 985–994.
[16] G. Wang, X. Zhang, S. Tang, C. Wilson, H. Zheng, and B. Y. Zhao, ''Clickstream user behaviour models,'' ACM Trans. Web, vol. 11, no. 4, Jul. 2017, Art. no. 21.
[17] Y. Liu, C. Wang, M. Zhang, and S. Ma, ''User behaviour modelling for better Web search ranking,'' Front. Comput. Sci., vol. 11, no. 6, pp. 923–936, Dec. 2017.
[18] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, ''Leveraging analysis of user behaviour to identify malicious activities in large-scale social networks,'' IEEE Trans. Ind. Informat., vol. 14, no. 2, pp. 799–813, Feb. 2018.
[19] J. Liu, Y. Xiao, K. Ghaboosi, H. Deng, and J. Zhang, ''Botnet: Classification, attacks, detection, tracing, and preventive measures,'' EURASIP J. Wireless Commun. Netw., vol. 2009, Dec. 2009, Art. no. 692654. doi: 10.1155/2009/692654.
[20] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, ''Detecting automation of twitter accounts: Are you a human, bot, or cyborg?'' IEEE Trans. Depend. Sec. Comput., vol. 9, no. 6, pp. 811–824, Nov. 2012.
[21] E. Van Der Walt and J. Eloff, ''Using machine learning to detect fake identities: Bots vs humans,'' IEEE Access, vol. 6, pp. 6540–6549, Jan. 2018.
[22] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Maqsood, "Random Forests and Decision Tree," IJCSI, Vol. 9, Issue 5, No 3, sept 2012.