International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-08, Nov 2019

Using Machine Learning to Detect Fake Identities Bots vs Humans

[1]Pushpalata Aher, [2]Reena Sahane

[1,2]Dept. of Computer Engineering, Sandip University, Nashik, Maharashtra, India.

[1]saeebhadane@gmail.com, [2]reenamojad@gmail.com

Abstract - We implement this technique to identify the malicious activities in social contact. The increasing number of accounts in social media platforms is a serious threat to the internet users. To detect and avoid fake identities it is need to understand the dynamic contagion. In exist; there are many models to detect the fake identities by bots or humans. Sybil identities are generally focused on famous social media platforms. The proposed system discussed in this paper is to detect the Sybil and troll identities using machine learning engineered techniques.

Key Words: Big data, bots, data science, fake accounts, fake identities, identity deception, social media, and veracity.

I. INTRODUCTION

The platforms of social media have a great impact on many areas today. In this we are focusing to identify the Sybil and troll identities in the platforms of social networks. There are many identities that are threats and malicious to the people on internet. So to identify the platforms of fake identities we use this supervised machine learning techniques to overcome of these fake identities.

In this the data sets are collected by the large data collection blogs. The data is stored and if any data is found malicious the data is cleaned and stored again. This gets the data more accurate of the user whether the account is a Sybil or troll identities/accounts using advanced techniques. This makes the platforms free of malicious activities to some extent.

Once the data is cleaned the spaces where the data is missing is filled. This shows that the missing spaces are fake identities and filling space are the cleaned fake identities. Before, the data is cleaned it is stored in non-relational database. Therefore, gets the data sets in a collection for future reference and remove the fake profiles.

Then they predict the accounts of social networks that are threats or ward. Using machine learning helps to find the fake identities of many social platforms. This growth in areas of internet makes the accounts more reliable and trustworthy for the users. Then the accounts are iterated in machine learning algorithms to identify the fake profiles over the internet. There is iterative training in machine learning to get the data and store in database. The activities in the accounts are identified as menace or protected in SPM. Finally, the results of identifying bots and troll identities are visualized and resulted by supervised machine learning algorithms.

1.1 Proposed System

Create a social media tweets, hash tags, social media posts, feeds, comments. Create non-relational databases. Using a data set preparation and cleaning. Then create a dataset.

Applying the Ml supervised machine learning algorithms. Finally evaluate and visualize the results. It gives accuracy more than 90%. It is a real time data analytics.

There is a growing number of people who hold accounts on social media platforms (SMPs) but hide their identity for malicious purposes. Unfortunately, very little research has been done to date to detect fake identities created by humans, especially so on SMPs. In contrast, many examples exist of cases where fake accounts created by bots fake or computers have been detected successfully using machine learning models.

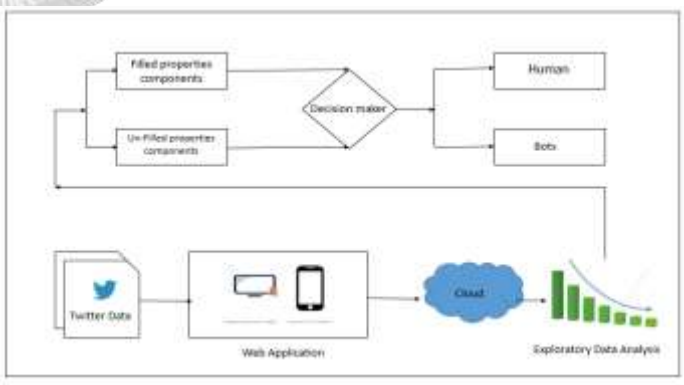


Fig. 1 System Architecture

1.2 Existing System

Earlier, they identified number of characteristics that distinguish fake and genuine followers such as number of tweets and number of followers. Then, we used these characteristics as attributes to machine learning algorithms to classify users as fake or genuine. We achieved high detection accuracy using machine learning algorithms.

 DOI : 10.35291/2454-9150.2019.0498

Machine learning algorithms are essential to the detection of fake accounts on Twitter and other similar social media. Knowing the key features and behavioral differences between humans with real accounts as opposed to bots operating via fake accounts is key to the detection and elimination of fake followers.

During the process of detecting the fake identities humans and bots have same behavior. These are applied to many supervised machine learning models. Many engineered features are existing but are not much successful in implementing to detect the malicmalicious accounts. Existing system use only two parameters.

‘‘Friend-to-followers ratio.’

Friend count

Less prediction accuracy

Not an real time analysis

Existing system not used for an long dataset.

Accuracy in supervised algorithm is 68 %

The existing system is not much featured to detect troll accounts then the bots accounts. The prediction of identity is not much accurate. The existing system focused on twitter to identify the fake identities. Create a social media tweets, hashtags, social media posts, feeds, comments. Create non-relational databases. Using data set preparation, cleaning .Then create a dataset. Applying the Ml supervised machine learning algorithms. Finally evaluate and visualize the results. Its gives accuracy more than 90%. It’s an real time data analytics. The existing system detect fake identities to 50% of accuracy. Three types of machine learning algorithm are used to detect the fake identities. The model is dependent on features (name, location, profile image). Cross validation and resampling methods are used in machine learning to detect fake identities.

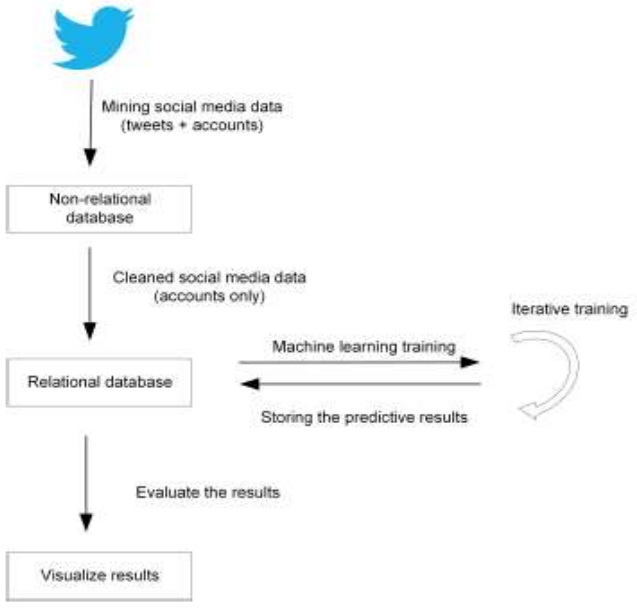


Fig. 2 Process Architecture

Data collection is the first activity. It is collected from various social media networks (twitter, kaggle, data.gov) etc. Then create non-relational databases. Then cleaning process is started after that the data is stored in relational databases. Then train the dataset using supervised machine learning algorithms (Linear regression, Navies Bayes). Finally the results are visualized and evaluated.

II. MODULES

1) **Data collection:** Real time data collected from Twitter, kaggle, UCI , Data.gov

2) **Data Cleaning:** fill the missing data and cleaning the noise data.

3) **Machine learning algorithm:** In this module we use linear regression and Naive bayes supervised algorithms

4) **Compare the machine learning model***:* Finally we create a compare model for other algorithms and also visualize the results.

2.1 DATA COLLECTION:

Real time data collected from Twitter, kaggle, UCI,

Data.gov. Collection of data is one of the major and most important tasks of any machine learning projects. Because the input we feed to the algorithms is data. So, the algorithms efficiency and accuracy depends upon the correctness and quality of data collected. So as the data same will be the output

Fig. 3 Data collection (Bots data)

2.2 DATA CLEANING:

Collecting the data from one task and making it useful to another data is an-other vital task. Data collected will be in an unorganized format and there may be lot of null values, in-valid data values and unwanted data from various means. Cleaning all the data and replacing them with the approximate data and filling the null and missing data with some fixed alternate values are the basic steps in preprocessing of data. Even data collected may contain completely garbage values. It is not necessary to be in exact format what it want to be can be in any format. This process is made to keep the data meaningful and for further processing. Data must be kept in an organized format.

 DOI : 10.35291/2454-9150.2019.0498

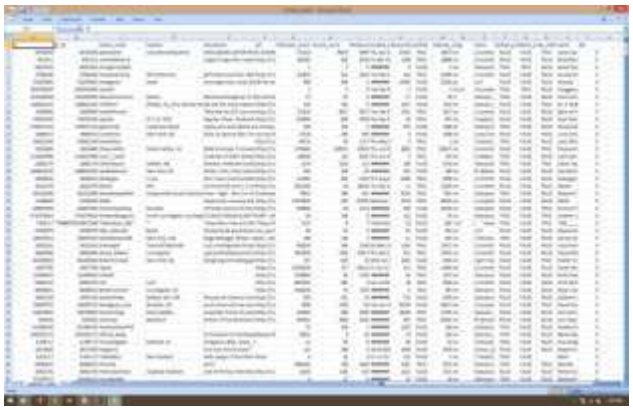


Fig.4 Non-bots data

Fig.5 Sample source code

2.3 MACHINE LEARNING ALGORITHM TRAINING:

The next step is algorithms are applied to data and results are noted and observed. The algorithms are applied in the fashion Mention in the diagram so as to improve accuracy at each stage.

2.4 TRAINING AND TESTING:

Finally after processing of data the next task is obviously testing. In this process where performance of the algorithm, quality of data, and required output all appears out. 80 percent of the data is utilized for training from the huge data set collected and 20 percent of the data is reserved for testing. Training is the process of making the machine to learn and giving it the capability to make further predictions based on the training process. Whereas testing means already having a predefined data set with output also previously labeled and the model is tested whether it is working properly or not and is giving the right prediction or not. If maximum number of predictions is right then model will have a good accuracy percentage and is reliable to continue with otherwise better to change the model.

2.5 Experimental Results

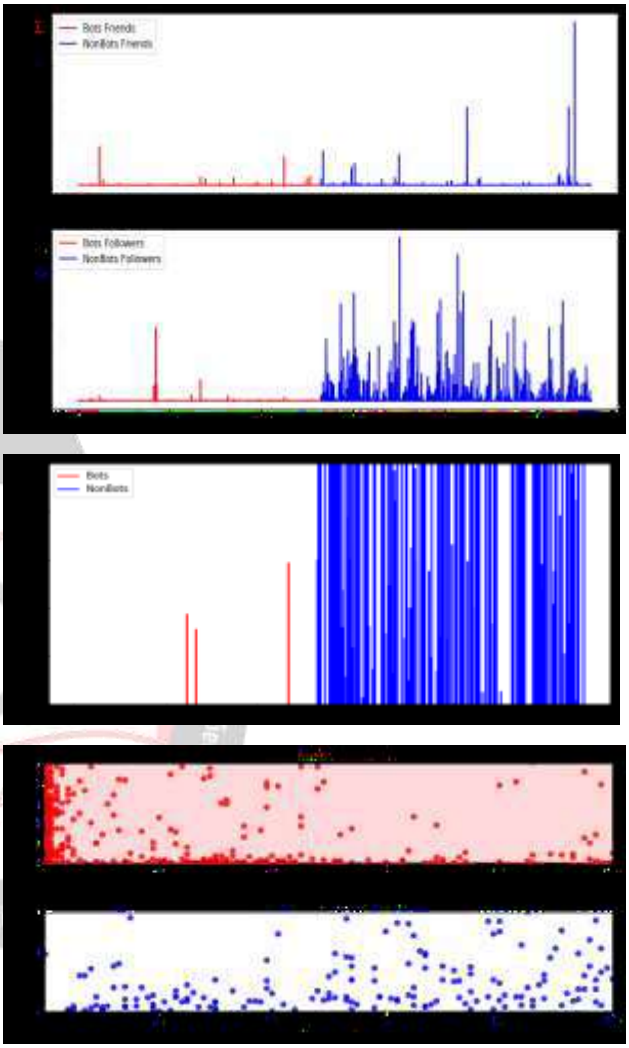
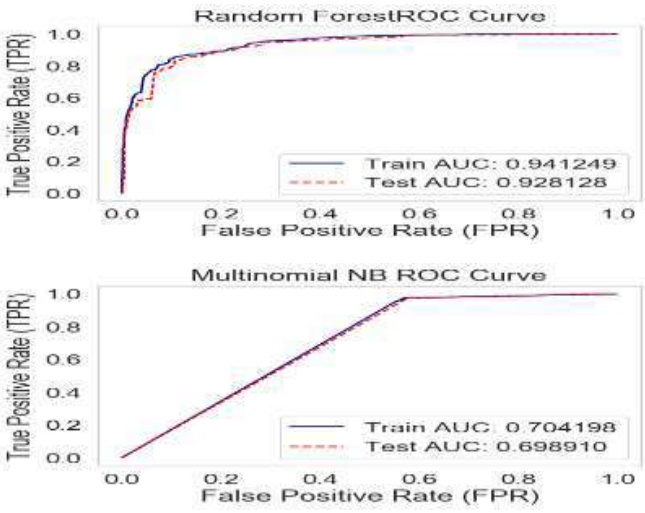


Fig. 6 Experiment Result



 DOI : 10.35291/2454-9150.2019.0498

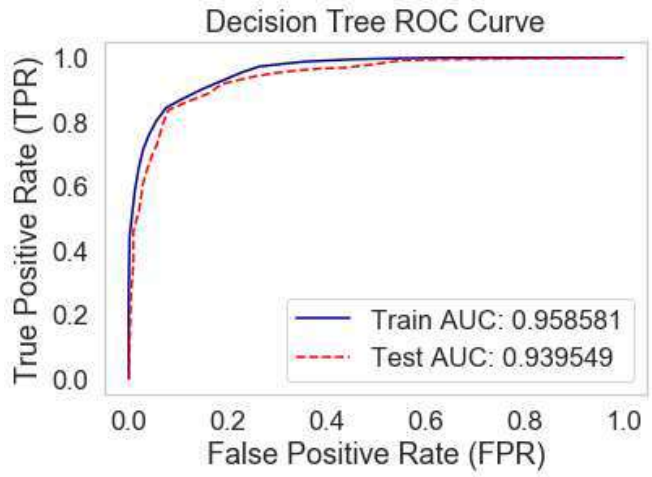


Fig. 7 Multinomial NB ROC Curve

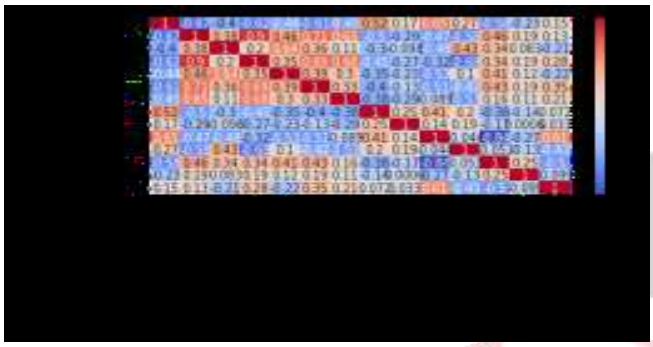


Fig.8 Decision Tree

III. CONCLUSIONS

A dataset of all media stages are gathered and kept up. In this paper the general idea of the procedure has been clarified as a synopsis. The exactness of the procedure is guaranteed. For the future use they may expand their exactness much more with another calculation.

An incredible web application can be created where information sources are not given legitimately rather understudy parameters are taken by assessing understudies through different assessments and inspecting. Specialized, diagnostic, consistent, memory based, psychometry and general mindfulness, interests and expertise based tests might be planned and parameters are gathered through them with the goal that outcomes will be unquestionably precise and the framework can be utilized realiably.

Additionally choice trees have not many impediments like over fitting, no pruning, absence of capacity to manage invalid and missing qualities and not many calculations have issue with colossal number of qualities. All these can be thought about and significantly progressively solid and increasingly precise calculations can be utilized. At that point the venture will be all the more dominant to rely on and significantly increasingly effective to rely on.

REFERENCES

[1] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews,''Profile characteristics of fake Twitter accounts,'' Big Data Soc., vol. 3,no. 2, p. 2053951716674236, 2016, doi: 10.1177/2053951716674236.

[2] C. Xiao, D. M. Freeman, and T. Hwa, ''Detecting clusters of fake accounts in online social networks,'' in Proc. 8th ACM Workshop Artif. Intell. Secur., 2015, pp. 91–101.

[3] S. Mainwaring, We First: How Brands and Consumers Use Social Media to Build a Better World. New York, NY, USA: Macmillan, 2011.

[4] V. S. Subrahmanian et al. (2016). ''The DARPA Twitter botchallenge.''[Online]Available:https://arxiv.org/abs/1 01. 05140.

[5] A. K. Jain and B. Gupta, "Phishing detection: Analysis of visual similarity based approaches," Security and Communication Networks, vol. 2017, 2017.

[6] S. Venkatesan, M. Albanese, A. Shah, R. Ganesan, and S. Jajodia, "Detecting Stealthy Botnets in a Resource-Constrained Environment using Reinforcement Learning," 2017

[7] T. Tuna et al., ''User characterization for online social networks,'' Social Netw. Anal. Mining, vol. 6, no. 1, p. 104,

2016.

[8] P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, ''Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying,''Logic J. IGPL. vol. 24, no. 1, pp. 42–53, 2015.

[9] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, ''Faking sandy: Characterizing and identifying fake images on Twitter during hurricane sandy,''in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 729–736.

[10] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, ''Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?'' in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM),Aug. 2014, pp. 620– 627.

[11] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews,

''Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach,'' in Proc. Int. Conf. Social Media Soc., 2015,p. 9.

[12] B. Viswanath et al., ''Towards detecting anomalous user behavior in online social networks,'' in Proc. Usenix Secur., vol. 14. 2014, pp. 223–238.

[13] Xiaoyun Wang, Chun-Ming Lai, Yunfeng Hong, ChoJui Hsieh, S. Felix Wu Multiple Accounts Detection on Facebook Using Semi-Supervised Learning on Graphs

[14] Neha M. Yadav1 Prof. Dr. P. N. Chatur2 Compromised Account Detection and Prevention by Profiling Social Behavior and FASS Key Concept 2018 International Conference on Recent Trends in Electrical, Electronics and Computing Technologies.2018

[15] Sneha Rane 2. Asst Prof. Megha Ainapurkar, 3. Asst Prof. Ameya Wadekar DETECTION OF COMPROMISED ACCOUNTS IN ONLINE SOCIAL NETWORK Proceedings of the Second International Conference on Computing Methodologies and Communication (ICCMC 2018) IEEE Conference Record # 42656; IEEE Xplore ISBN:978-1-5386-3452-3

[16] Nitin T Simon Dr. Susan Elias Detection of Fake Followers using Feature Ratio in Self-Organizing Maps 2017 IEEE.

[17]Shubham Patil,Akshay Ingale, Pradeep Ranher, Purushottam Mahakal and Mr. Y. L Hakim

"Detect Fake Identities of Bots vs Human using Machine Learning " IJSRD - International Journal for Scientific Research & Development| Vol. 7, Issue 02, 2019 | ISSN (online): 2321-0613.

 DOI : 10.35291/2454-9150.2019.0498