International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Identify the Human or Bots Twitter Data using Machine Learning Algorithms

Nambouri Sravya[1], Chavana Sai praneetha[2], S. Saraswathi[3]

[1,2]Student, Department of Computer Science and Engineering, Panimalar Engineering College, Tamil Nadu, India
[3]Associate Professor, Dept. of CSE, Panimalar Engineering College, Tamil Nadu, India

---

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: **Machine learning, fake accounts, data sets, social platforms**

1. 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 visualised 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 an real time data analytics.

1.2 Existing System

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 malicious 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%. Its 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.

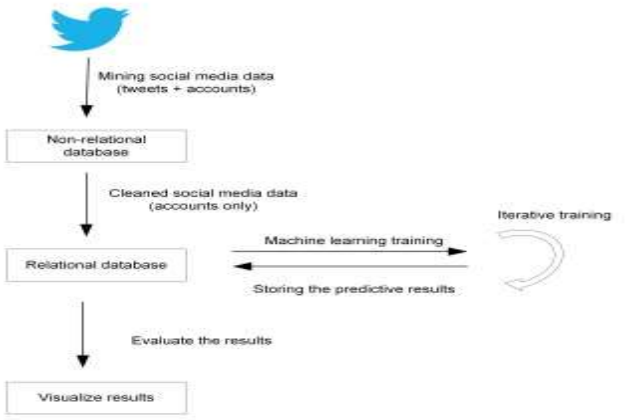


Fig. 1 process architecture

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 visualised and evaluated.

2. 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

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. 2 Data collection (Bots data)

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 pre-processing 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.

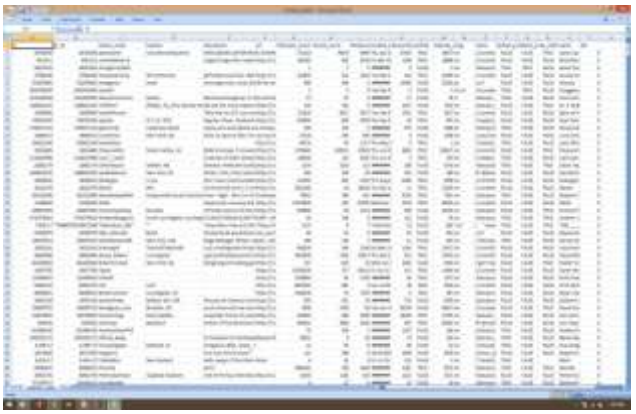


Fig.3 Non-bots data

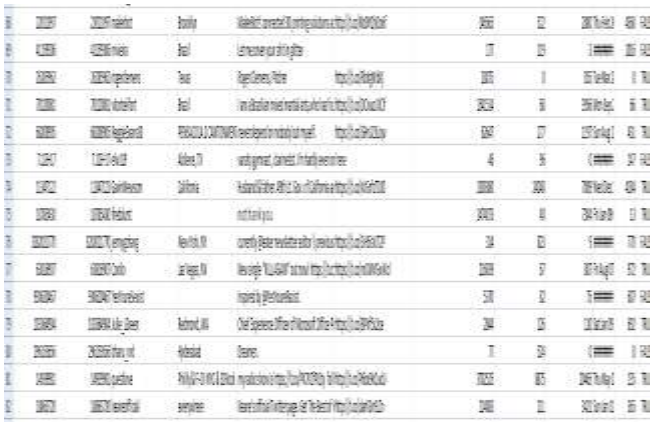


Fig.4 Sample source code

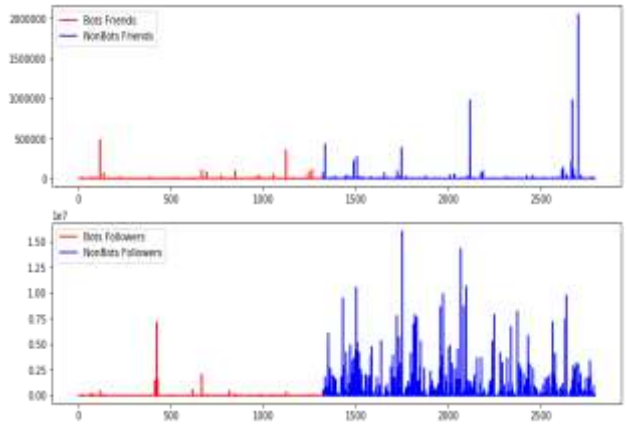
MACHINE LEARNING ALGORITHM TRAINING:

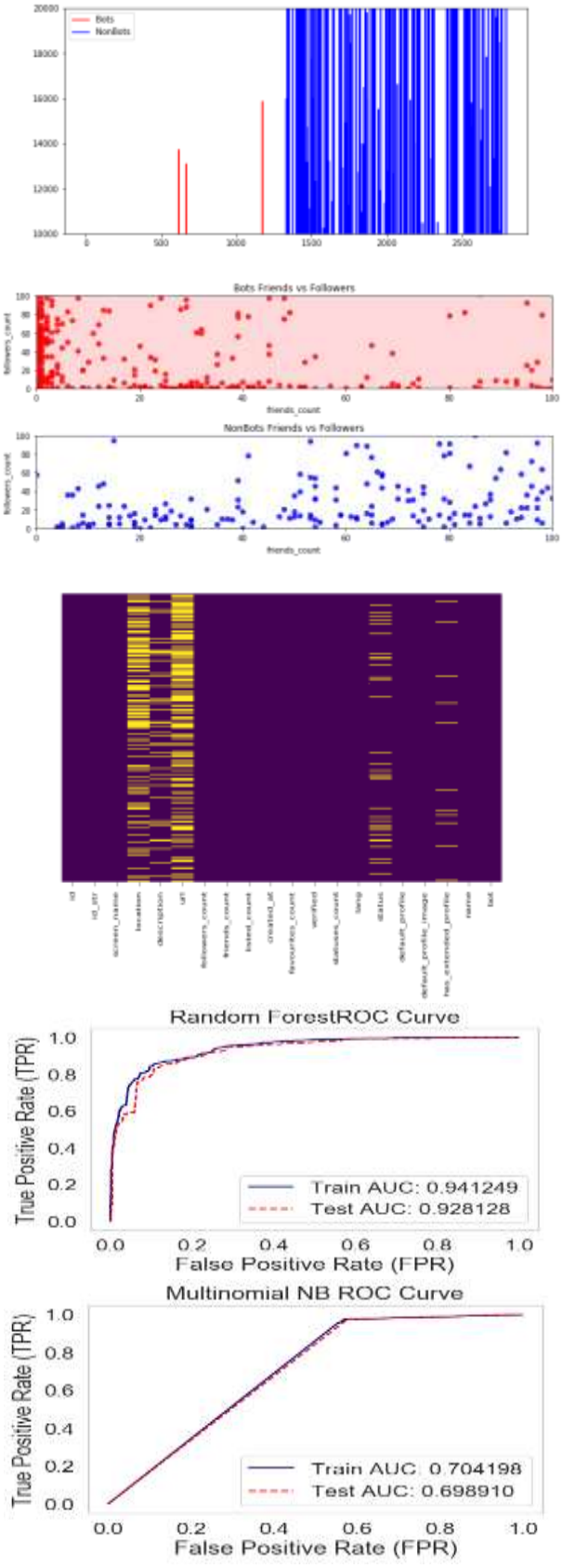
The next step is algorithms are applied to data and results are noted and observed. The algorithms are applied in the fashion Mention in the diagram so as to improve accuracy at each stage.

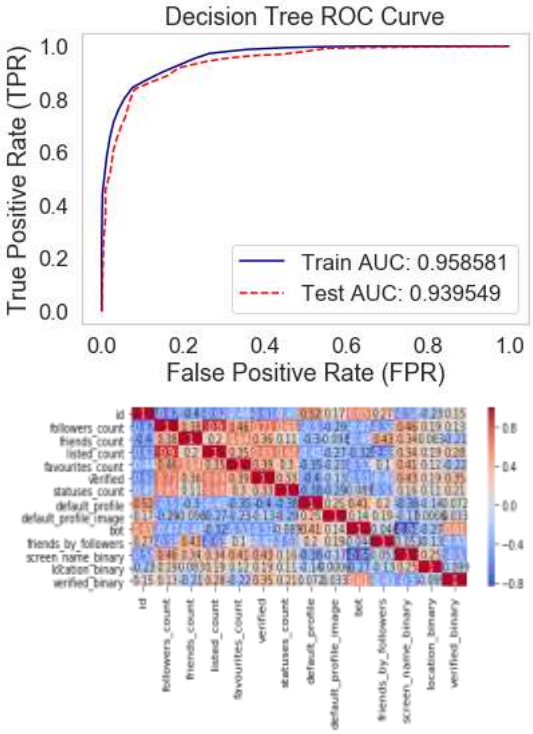
TRAINING AND TESTING:

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.

ExperimentalResults







3. CONCLUSIONS

A dataset of all media platforms are collected and maintained. In this paper the overall concept of the process has been explained as a summary. The accuracy of the process is ensured. For the future use they may extend their accuracy even more with another algorithm.

A powerful web application can be developed where inputs are not given directly instead student parameters are taken by evaluating students through various evaluations and examining. Technical, analytical, logical, memory based, psychometry and general awareness, interests and skill based tests may be designed and parameters are collected through them so that results will be certainly accurate and the system can be used realiably.

Also decision trees have few limitations like overfitting, no pruning, lack of capability to deal with null and missing values and few algorithms have problem with huge number of values. All these can be taken into consideration and even more reliable and more accurate algorithms can be used. Then the project will be more powerful to depend upon and even more efficient to depend upon.

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/1601.05140

[5] Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, "In a world that counts: Clustering and detecting fake social engagement at scale," in Proc.25th Int. Conf. World Wide Web, 2016, pp. 111–120.

[6] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of Twitter spam," in Proc. ACM SIGCOMM Conf.Internet Meas. Conf., 2011, pp. 243–258.

[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] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: Human, bot, or cyborg?" in Proc. 26th Annu. Comput. Secur. Appl.Conf., 2010, pp. 21–30.

[14] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in Proc. NDSS, 2013,pp. 1–17.

[15] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, ‘‘Predicting online extremism, content adopters, and interaction reciprocity,’’in Proc. Int. Conf. Social Inform., 2016, pp. 22–39.VOLUME 6, 2018 6547

E. van der Walt, J. Eloff: Using Machine Learning to Detect Fake Identities: Bots versus Humans [16] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, ‘‘Fame for sale: Efficient detection of fake Twitter followers,’’ Decision Support Syst., vol. 80, pp. 56–71, Dec. 2015.

[17] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, ‘‘Detecting spammers on Twitter,’’ in Proc. Collaboration, Electron. Messaging, AntiAbuse Spam Conf. (CEAS), vol. 6. 2010, p. 12.

[18] H. Kwak, C. Lee, H. Park, and S. Moon, ‘‘What is Twitter, a social network or a news media?’’ in Proc. 19th Int. Conf. World Wide Web, 2010,pp. 591–600.

[19] Z. Zhang and B. B. Gupta, ‘‘Social media security and trustworthiness: Overview and new direction,’’ Future Generat. Comput. Syst., to be published, doi: 10.1016/j.future.2016.10.007.

[20] A. K. Jain and B. B. Gupta, ‘‘Phishing detection: Analysis of visual similarity based approaches,’’ Secur. Commun. Netw., vol. 2017, Jan. 2017, Art. no. 5421046. [Online]. Available: https://doi.org/10.1155/2017/5421046

[21] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, ‘‘On profiling bots in social media,’’ in Proc. Int. Conf. Social Inform., 2016, pp. 92–109.

[22] M. Tsikerdekis and S. Zeadally, ‘‘Multiple account identity deception detection in social media using nonverbal behavior,’’ IEEE Trans. Inf. Forensics Security, vol. 9, no. 8, pp. 1311–1321, Aug. 2014.

[23] J. T. Hancock, ‘‘Digital deception,’’ in Oxford Handbook of Internet Psychology. London, U.K.: Oxford Univ. Press, 2007, pp. 289–301.