

BOT IDENTIFICATION ON SOCIAL MEDIA NETWORKS USING DEEP LEARNING

Medisetti Siva Sahitya¹, Sirigineedi Jaidheer², Nookala Manoj³, K. Govinda Raju⁴

^{1,2,3}CSE, B. Tech Aditya Engineering College, Surampalem

⁴Guide, M. Tech, (Ph.D., Associate Professor Aditya Engineering College, Surampalem

ABSTRACT

Social media is a web-based technology to facilitate interaction between a large group of people. Twitter is the popular social websites which allow the users to manifest their opinion on different topics like politics, sports, stock market, entertainment and so on. These applications provide the fastest means of conveying information. It highly effect the people viewpoint. So it is very important to know that data should send by authentic users and not by the bot accounts. This project intends to predict the presence of bots in social media like twitter. It introduces a deep learning technique of Count vectorizer for training the NLP model and ANN algorithm to perform identifying the bots exists in social networks. It uses API Based web crawlers to automatically update the datasets and we framed the output in a convenient way that can be understood by anyone. And also this is useful for online polls and review gathering software. Therefore we can protect the public users in identifying the authorization of a particular social media account. Through a front end interface and API we will identify the tweet made by human or bot from twitter, which is used by online surveys and polls.

Keywords: Deep Learning, Bot identification, NLP model, ANN Algorithm

1. INTRODUCTION

In the present world, everyone tends to use social media websites. Twitter became popular social website that allows users to manifest their opinion on different topics like politics, sports, stock market, entertainments and so on. These software applications provide the utmost scalability to the associated users however it can also spread the spam or fake information through bot accounts. Bot accounts are created on social media networks, and they are used to perform the certain functions. They will automatically generate messages, ideas, views over the platform to increase the followers for themselves. This project introduces deep learning approach that can effectively identify bots in any of the social networking websites and eliminate them using various techniques. It uses Count Vectorizer for training the NLP model we perform identifying the bots exists in social networks. It uses API Based web crawlers to automatically update the datasets and we framed the output in a convenient way that can be understood by anyone. And also this is useful for online polls and review gathering software. To identify the bot task we consider parameters such as account activity, date of creation and origin will also be considered in the new dataset. So by removing the fake news spreading accounts like bot will improve the confidentiality and reliability of using the platforms like twitter. It will also help to the organizations which conduct the online surveys and polls.

2. EXISTING SYSTEM

The existing model shows a multilingual method for addressing the bot detection task in Twitter using Deep learning approaches to support public accounts of end users when checking the integrity of a particular twitter account. Many experiments were conducted using state-of-the-art Multilingual Language Models to cause an encoding of the text based features of the user public account. This system also using API for gathering the user information with the help of user id.

Disadvantages:

- Trained on the dataset which is outdated and limited to only twitter.
- The main disadvantage of this paper is it shows the output in Graphical Format which serves little use in real life situations.

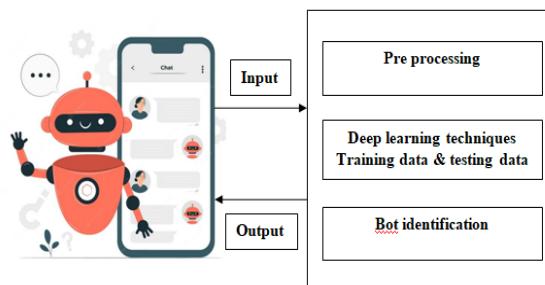
3. LITERATURE REVIEW

Heng Ping, and Sujuan Qin (2018) propose a deep-learning algorithm for the detection of social bots. In this paper, a social bots discovery model discussed on the deep literacy algorithm DeBD is proposed. This model includes three layers, The first sub-caste is the joint content point birth sub-caste, which focuses on the point birth of the tweet's content and the relationship between them will be maintained. The alternate level is the tweet metadata temporal point birth sub-caste, which regards the tweet metadata as temporal data and uses the temporal data as the input of the LSTM to prize the stoner social exertion temporal point. The third level is the point fusing caste, which combine the uprooted joint content features with the temporal features to descry social bots. The main disadvantage of this model

requires large datasets which makes training harder and more expensive which is not suitable. [1] This paper aims to introduce an automatic system for the discovery and removing of bots that live on social media platforms. S. Gannarapu, A. Dawoud, R. S. Ali and A. Alwan proposed this model in the year 2020. The exploration has the purpose of removing the non genuine accounts like bot accounts, their affiliated information, and the data which are posted by these accounts and to make these platforms free of deceiving information. Bots discovery and discarding will increase the authenticity of the contents presented on different social media applications. Also, It'll improve the position of privacy and authenticity of these platforms and related druggies. The exploration uses the bot discovery fashion grounded on machine literacy algorithms. The factors used to study this model are data, point selection, and bot discovery. The survey performs web development and hosting on the collected data with a machine- learning algorithm to perform bot discovery in social media networks. The proposed system provides a more accurate and effective system for bot discovery it helps to terminate the false news spreading bot accounts. [2] Pradeep Kumar Tiwari, T Velayutham were proposed this methodology in the year 2017. For better and fair prophetic approach, the quality of data is important. Low quality content may affect into predicting of indecorous cause of an event, deceiving trending issues. They are targeting twitter for the identification of similar bots, which are created to perform the certain functions in social media. As it largely used by data scientists for operations related to scientific prediction and sentiment analysis. In this paper, they overcome the earlier approaches and used a machine literacy grounded approach for the bracket between a bot profile and public profile. And also they have linked 10 attributes of stoner profile and tweet pattern for an account and calculated a score called bot Score for each profile to model as bot or as human. They have expanded the list of features in determining between bot and mortal to further fine- granulated marker. The proposed system was set up to be more accurate than traditional modals. The main disadvantage of this model it achieved a relatively low accuracy of 75%, so it is not affordable.[3]

4. PROPOSED METHODOLOGY

The proposed methodology mainly deals with social media networks to provide the integrity to the associated users while using any software platforms. While using these apps individuals can share destructive content in a network there are millions of automated accounts exists, commonly known as bots, they prepare fraud news or information and impact public opinion without human involvement. It works on the basis of deep learning to detect the bot actions and eliminate them using various techniques. So by removing the fake news spreading accounts like bot will improve the confidentiality and reliability of using the platforms like twitter. It Use updated and latest dataset for better accuracy of deep learning models. It uses API Based web crawlers to automatically update the dataset. This keeps the model always updated to new kind of bots which is not possible in the old system.



More parameters such as account activity, date of creation and origin will also be considered in the new dataset. Previous papers show the output as Graph which has little significance in the real life usage. Our proposed system will show the output which can be understood to laymen. The new system will also be trained on dataset from different social media network like twitter. This model used for more robust to various kinds of automated accounts. This new system can be used in tandem with an API so it can be used with different secondary web crawlers. This is useful for online polls and review gathering software

Hardware requirements:

- Processor: i5 / Ryzen 5 Processor
- RAM : 4GB (minimum)
- Hard Disk : 160 GB (Min)
- Input Devices - Keyboard, Mouse

Software requirements:

- Operating System: Windows 10, Linux (Ubuntu/ Debian)

- Backend Scripting :Python, Flask, Fast API, Ajax, Socket IO, JQuery
- Frontend Scripting: HTML, CSS, JavaScript
- IDE : Google colab or Jupyter Notebook
- Testing Tool : Postman API

5. ALGORITHMS USED

Logistic Regression: Logistic regression is one of the supervised learning algorithms that is used for binary classification problems. It is a statistical method used to produce and model the relationship between a dependent variable and one or more independent variables. The aim of logistic regression algorithm is to discover the best fit between the input data and a sigmoid function that mapping the input values to a probability score between 0 and 1. That probability score can be used to make binary predictions.

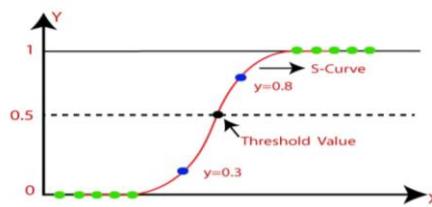


Fig 4.1 Logistic Regression

Decision Tree: It is also one of the deep learning techniques called decision tree used to solve the issues that related to regression and classification purpose, and also it is regularly preferable to achieve the relation between nodes. This algorithm is in the form of tree structure classification, in which nodes will be present in different levels. It is used for taking the effective decisions using these nodes. The network nodes will be used for dataset features, each leaf node making the trained model. The decision tree consists of decision node and child nodes.

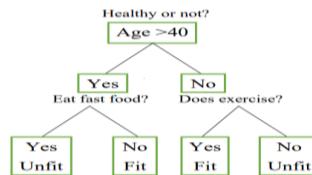


Fig 4.2 Decision Tree

Random Forest: Random Forest is also one of the frequently used algorithms in deep learning. It consists of group of decision trees of various datasets and it brings the average to increase the prediction accuracy of datasets.

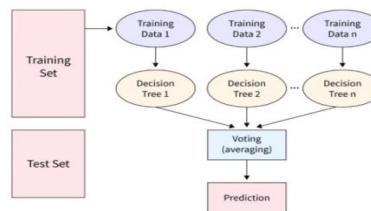


Fig 4.3 Random Forest

The previous algorithms can have the disadvantage like prone to over fitting that is more complex and high variance. Those models are not suitable to get generalizing of other datasets. To overcome the above mentioned problem we can optimized it with using random forest algorithm.

MLP Classifier: To solve the classification problems in learning algorithms, an Artificial Neural Network algorithm was implemented in deep learning with Multiple Layer Perceptron (MLP) classifier. It is also one of the supervised learning algorithms. The MLP classifier uses neural network as it having many layers of nodes, in that each node was processed by weight of sum of inputs and outputs along with an activation function. It contains three layers of data called input layer, hidden layer and output layer. The output layer will produce the final classification result after being processed by the input layers data. The data representations can be done with learning in hidden layers, which are present in between input layer and output layer. And also it uses the back propagation that is used for change the weights during training of both predicted and actual data. It can handle the high dimensional input data and used with various data source.

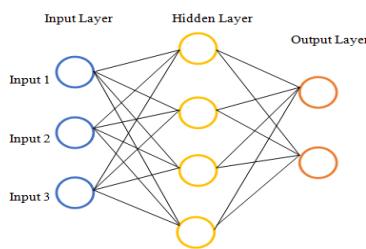
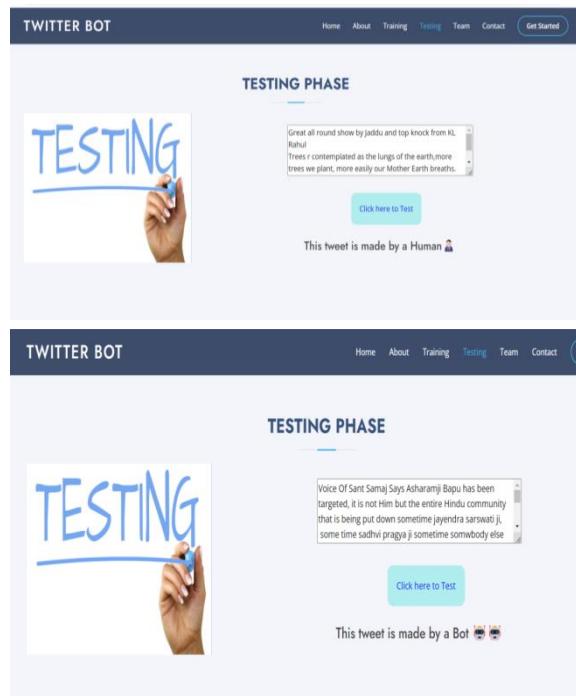
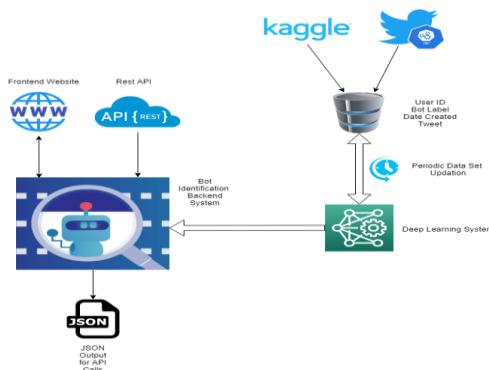


Fig 4.4 MLP Classifier

6. EXPERIMENTAL RESULTS



Architecture diagram



7. ADVANTAGES

- Using dataset which can update itself using API based crawlers. This helps in identification of different new kinds of bots.
- This helps to train the model to different kinds of bots. This makes the system more robust across different platforms
- Our system can be integrated to different other subsystems like web crawlers.
- By using the ANN algorithm we can assess Parallel processing capability, Storing data on the entire network.
- It gives the best results through front end interface and API created, which is mainly used by organizations when surveys and online polls were conducted.

8. CONCLUSION AND FUTURE SCOPE

Present project summarize the different applications of deep-learning of ANN algorithm in social media sector. The main motive of this study was to brief the applications and available techniques of deep-learning is a type of machine-

learning and artificial intelligence (AI) that solves the problem caused by the bot accounts. The paper also highlights the different literatures, which reflects different technologies to identify the bots. It gives the benefits identify the different tweets made by human and bot accounts. And it supports two main contributions for the scientific groups which includes a Deep-Learning algorithm to automatically identifying bots as well as APIs. Finally, this project will gives the outstanding performance of transformers in downstream NLP model system as the one discussed, by producing a more robust input vector which leads to the final classifier system to be more capable of getting relevant low-level data features from this system. This paper highlights limitations of existing systems. The future scope of this project is to help the society that one can expect massive demand for security over the social media. Because as we know social websites can turned into our daily routine to use the APIs of the particular social media network account in view to identify the bot actions with high accuracy. Deep Learning should be the major tools for the researchers to address the above mentioned issues. These methods can be major implementations to solve the future crisis. As we have limited our scope to bot identification through interface, but in future it can be extended further up to APIs connection that can be useful when online polls and surveys conducted. Hence future researchers should organize a proper dataset covering all area fields of social media websites in order to protect the users from fraud spreading bot automated accounts. These methods can be major implementations to solve the future crisis.

9. REFERENCES

- [1] [Https://Ieeexplore.Ieee.Org/Document/8344722](https://ieeexplore.ieee.org/document/8344722)
- [2] [Https://Ieeexplore.Ieee.Org/Document/10020919](https://ieeexplore.ieee.org/document/10020919)
- [3] [Https://Ieeexplore.Ieee.Org/Document/6280553](https://ieeexplore.ieee.org/document/6280553)
- [4] [Https://Ieeexplore.Ieee.Org/Document/9470605](https://ieeexplore.ieee.org/document/9470605)
- [5] [Https://Ieeexplore.Ieee.Org/Document/9579883](https://ieeexplore.ieee.org/document/9579883)
- [6] H. Ping And S. Qin, "A Social Bots Detection Model Based On Deep Learning Algorithm" 2018 Ieee 18th International Conference On Communication Technology (Icct), 2018, Pp. 1435-1439, Doi: 10.1109/Icct.2018.8600029.
- [7] S. Gannarapu, A. Dawoud, R. S. Ali And A. Alwan, "Bot Detection Using Machine Learning Algorithms On Social Media Platforms" 2020 5th International Conference On Innovative Technologies In Intelligent Systems And Industrial Applications (Citisia), 2020, Pp. 1-8, Doi: 10.1109/Citisia50690.2020.9371778.
- [8] T. Velayutham And P. K. Tiwari, "Bot Identification: Helping Analysts For Right Data In Twitter" 2017 3rd International Conference On Advances In Computing, Communication & Automation (Icacca) (Fall), 2017, Pp. 1-5, Doi: 10.1109/Icaccaf.2017.8344722.
- [9] 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" 2016 Ieee/Acm International Conference On Advances In Social Networks Analysis And Mining (Asonam), 2016, Pp. 533-540, Doi: 10.1109/Asonam.2016.7752287.
- [10] J. Oh, Z. H. Borbora, D. Sharma And J. Srivastava, "Bot Detection Based On Social Interactions In Mmorgs" 2013 International Conference On Social Computing, 2013, Pp. 536-543, Doi: 10.1109/Socialcom.2013.81.
- [11] H. Shukla, N. Jagtap And B. Patil, "Enhanced Twitter Bot Detection Using Ensemble Machine Learning" 2021 6th International Conference On Inventive Computation Technologies (Icict), 2021, Pp. 930-936, Doi: 10.1109/Icict50816.2021.9358734.
- [12] S. M. Attia, A. M. Mattar And K. M. Badran, "Bot Detection Using Multi-Input Deep Neural Network Model In Social Media," 2022 13th International Conference On Electrical Engineering (Iceeng), 2022, Pp. 71-75, Doi: 10.1109/Iceeng49683.2022.9781863.
- [13] M. Sevi And İ. Aydin, "Detection Of Fake Twitter Accounts With Multiple Classifier And Data Augmentation Techniqu," 2019 International Artificial Intelligence And Data Processing Symposium (Idap), 2019, Pp. 1-6, Doi: 10.1109/Idap.2019.8875944.
- [14] A. Bhattacharya, R. Bathla, A. Rana And G. Arora, "Application Of Machine Learning Techniques In Detecting Fake Profiles On Social Media," 2021 9th International Conference On Reliability, Infocom Technologies And Optimization (Trends And Future Directions) (Icrito), 2021, Pp. 1-8, Doi: 10.1109/Icrito51393.2021.9596373.
- [15] R. Bailurkar And N. Raul, "Detecting Bots To Distinguish Hate Speech On Social Media," 2021 12th International Conference On Computing Communication And Networking Technologies (Iccnt), 2021, Pp. 1-5, Doi: 10.1109/Iccnt51525.2021.9579883.