ISSN 2277-2685
IJESR/Sep. 2023/ Vol-13/Issue-3/1-09
Mr.K.Vijaya Babu *et. al.*, / International Journal of Engineering & Science Research

Tweet Based Bot Detection Using Big Data

Mr.K.Vijaya Babu [1], Mantha Hari Dinesha Bharadwaj[2], Pamu Sushma[2], Dhawrali Rishad[2], Pechetty Hemanth[2]*[1Asst. Professor,Computer Science and Engineering, CMR Engineering College, medchal, T.S, India[2]B.Tech, Computer Science and Engineering, CMR Engineering College, medchal, T.S, India*

Abstract: With millions of users, Twitter is one of the most well-liked microblogging social networking networks. Twitter has been the victim of several assaults due to its popularity, including rumours being circulated, phishing links, and malware. Users are seriously threatened by tweet-based botnets since they may carry out massive assaults and deceitful operations. Big data analytics methods, notably shallow and deep learning methods, have been used to counteract these threats by properly differentiating between human accounts and tweet-based bot accounts. In this article, we explore current methods and provide a taxonomy that categorises the most advanced methods for tweet-based bot identification. Along with their performance results, we also discuss shallow and deep learning methods for tweet-based bot detection. Finally, we outline and explore the difficulties and unresolved problems related to tweet-based bot identification.

1. INTRODUCTION

Social media is one of the most widely utilised instruments in today's world for communication. Additionally, organisations use it extensively to communicate with their clientele. There are 3.5 billion active social media users worldwide, according to [1]. Organisations utilise social media sites like Facebook, Twitter, LinkedIn, and others to raise brand awareness and increase sales. One of the most widely used social networking sites is Twitter. It has 340 million active users who may converse widely and express their thoughts on a variety of subjects. Various attack types may be aimed at Twitter. For instance, prominent Twitter accounts were taken over in July 2020 as a result of a spear phishing assault [2]. Additionally, fake accounts could be made to pose as reputable users and organisations.

Twitter may also be abused by a botnet, which is a collection of fraudulent accounts run by a botmaster and managed by computer programmes rather than real people.

Social media bots that utilise tweets to communicate seriously endanger Twitter users' security. These bots are used to disseminate spam, phishing links, and bogus information. They may be used as Command and Control (C&C) infrastructure to plan DDoS assaults, even when they are not employed as bots to perform DDoS attacks [3,] [4]. They have the ability to communicate with human accounts in order to trick people and take over their accounts. These bots are also used as instruments to start extensive campaigns of opinion-shaping manipulation. A research [5] found that botnets provide 52% of web traffic, with the remaining 48% coming from legitimate users. It is also important to know that some bots have over 350,000 fictitious followers. It is necessary to create detection algorithms that can precisely differentiate between Twitter bot accounts and real accounts in order to address the

aforementioned problems. One example of huge data is Twitter data, which generates 6,000 tweets per second, or around 500 million tweets each day [6].

A Description Of The Project: Twitter has made considerable use of artificial intelligence to provide user-specific tweet suggestions. Deep neural networks are really used on Twitter data to identify the relevant information for users and so enhance their platform experience. Fighting offensive content has benefited greatly from artificial intelligence. In 2017, rather of using people, artificial intelligence algorithms were used to identify and suspend around 300,000 accounts.

This paper seeks to provide a general overview of several tweet-based bot detection approaches that employ shallow and deep learning methodologies to discriminate between real accounts and automated accounts. The following are the paper's significant contributions in particular:

1) A taxonomy is offered that categorises the most advanced machine learning methods for tweet-based bot detection.

2) A thorough analysis of shallow and deep learning methods for tweet-based bot identification is offered, including the solutions as of the year 2020.

3) The difficulties and unresolved problems with tweet-based bot identification are emphasised and examined.

2. LITERATURE SURVEY

2.1 Existing System

Author [7] provided a short comparative survey of the research work in the field of Twitter spam detection within the year range of 2009-2015. They provided descriptions of several detection techniques under the headings of account-based, tweet-based, graph-based, and hybrid-based approaches. The account-based approaches were shown to take use of user profile information, including followers and following count, as well as other derived attributes, including account age. While it has been demonstrated that features like the distance and degree of connectivity between users can be used for spam detection in graph-based methods. The study, however, largely focused on identifying spam utilising URL and its derived properties, such as length and domain name, in tweet-based approaches. Posting URLs[8] were examined and categorised as dangerous or benign in order to find spam users. In addition, the authors emphasised underutilised characteristics that they said would enhance spam identification.

Chakraborty et al. provided a further comparison study in the area of multiplatform spam user detection. The authors understood that in order to accomplish precise detection, various platforms—such as e-mails, blogs, or microblogs—need various methodologies and characteristics. As a result, suggested methods for the years 2011 through 2015 were categorised according to the platform the dataset is on. Each set of approaches was subjected to a qualitative comparison on the same platform[9].

According to Besel et al. , the botnet exploited redirections and URL network shortening providers to mask the real destination sites. They admitted that when users clicked on these URLs, the Bursty botnet's botmaster could be spotted setting up landing pages on phishing websites. They attested to the botmaster's continued success in running businesses that cater to Twitter bots. This research examines and provides insight into Twitter's cyberspace operations, criminality, and dark markets[10].

Recent studies on Twitter social botnet detection were compiled by Alothali et al. in their article published in 2015 . Each approach that was suggested was given an analytical examination along with its merits and disadvantages. The methods were divided into three primary groups, including crowdsourcing, machine learning, and graph-based methods. The crowd sourcing method, which is considered to be the most error-prone of the three, employs human intelligence to find different patterns. Additionally, it was discovered that random forest classifiers, in particular, are the most frequently employed machine learning techniques for identifying social bots among Twitter users[11].

A thorough assessment of harmful social bots' covert behaviour and their detection methods was provided by Latah The author thoroughly examined new, machine learning-based, and graph-based detection methods. The report also examined these tactics' advantages and disadvantages as well as the strategies the bots used to evade detection. As a result, the report offered strategies that may improve the defence mechanisms against rogue bots.

2.2Proposed System

In order to improve comprehension of the total text context, the suggested model utilised the bidirectional technique in which twitter sentences are processed both forward and backward for each layer. The public dataset Cresci-2017, which comprises of tweets from 3,474 human accounts and 1,455 bots for a total of 11.4 million tweets, is used to train the model. Each tweet was preprocessed and tokenized to suit the word embedding model before training. Text was transformed into network-acceptable numerical vectors using a pre-trained GloVe model. The vectors were input into a three-layer model with an initial setting of 0.5 for the decreasing dropout layer. Their model was evaluated using two subsets of testing datasets, each made up of 1,982 and 928 accounts, and it had an accuracy and precision of 93% and 95%, respectively[12].

A unique deep learning model called RTbust was developed by Mazza et al. to discriminate between bots and people based on their retweet patterns. The team first examined the behavioural patterns of both people and bots before developing the model. The investigation revealed a particular time pattern for retweeting, which was divided into four types. The droplet pattern, which represents typical users and has a reasonable lag between the time a tweet is written and when it is retweeted, comes in first. The three remaining patterns were potentially bot-related because of their erratic and strange retweeting behaviour.

2.2 Preliminary cryptographic steps

BOT DETECTION BASED ON TWITTER

Although detecting social bots is a difficult task, some studies have examined the traits and behaviour of bots , , and [and provided various features that are frequently seen in the majority of studies. Verified accounts, for instance, may be trusted to belong to real people. Additionally, since bots typically mass-follow and have a short lifespan , the ratio of followers to following and the age of the account are considered discriminative characteristics in detecting bots. To differentiate between tweet-based bots and human accounts, tweet-based bot detection systems primarily leverage the following characteristics :

• ID: This stands for the tweet's distinctive identity.

• User: It is a representation of the tweet's author.

• Created_at: This field displays the tweet's creation time in UTC.

• Text Tweet: This alludes to the tweet's main content.

• Tweet Length: It provides the tweet's character count.

• #Hashtags: This identifies how many hashtags were used in the tweet.

• #URLs: This identifies how many URLs are included in the tweet.

• in_reply_to_status_id: This feature shows the ID of the original tweet when the tweet is a reply.

• in_reply_to_user_id: This feature identifies the author of the original tweet when a tweet is a reply.

• Coordinates: These show where the tweet was geographically located.

• Favorite_Count: This number shows how many times Twitter users have liked the tweet.

The number of times a tweet has been retweeted is shown by the field "Retweet_Count."

• Reply Count: This indicates how many times the tweet has received replies.

Favorited is a boolean feature that is true when the authenticating user likes the tweet.

Retweeted is a boolean characteristic that is true when the authenticating user retweets a particular tweet.

• Possibly_sensitive is a boolean feature that is true if a link is included in the tweet.

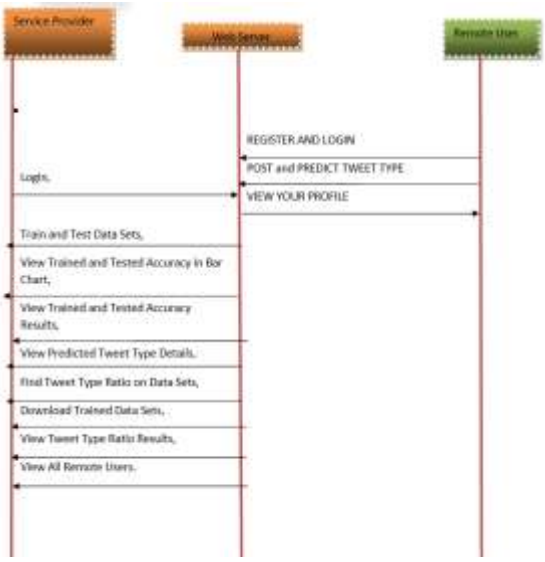

Fig 1: System Flow

3. SYSTEM ANALYSIS&DESIGN

3.1 TAXONOMY

We discuss machine learning methods for tweet-based bot identification in this section. The two main types of state-of-the-art approaches are shallow learning-based detection and deep learning-based detection, as illustrated in Fig. 1. The shallow detection approaches are further divided into three subcategories based on the learning approach: supervised learning, semi-supervised learning, and unsupervised learning. In supervised learning, the learning model is honed using labelled data in order to forecast the results of the incoming data. The model is created using an unsupervised learning strategy using unlabeled data. It looks for patterns and structures within the data itself. Both labelled and unlabeled data are used in semi-supervised learning approaches to train the model. On the other hand, two subcategories of deep learning-based detection methods have been identified: generative architecture-based methods and discriminative architecture-based methods. A generative model learns the joint probability distribution p(x, y) if we have input data x and wish to categorise them into labels y. The conditional probability distribution p(y|x) is learned by the discriminative model, on the other hand. A deep neural network and a generative model are combined to create a deep generative architecture. It often relates to uncontrolled learning. The deep discriminative architecture uses supervised learning and is constructed by fusing a deep neural network with a discriminative model to calculate and optimise p(y|x).

The remainder of the section has a comprehensive examination of each category.

3.2 Detection Methods Based On Deep Learning

Deep neural networks have recently attracted the interest of scientists working in domains ranging from language processing to computer vision. Its potency in terms of textual categorization has been established. It can analyse structured data, such as words, and automatically generate discriminant features, replacing costly and labor-intensive hand-crafted features that need in-depth data expertise. As a result, deep neural networks like

Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) were used as a feature extractor or classifier for many language processing problems, one of which is the tweet-based bot detection, because a classifier's overall performance heavily depends on the quality of its data. The input for neural networks must be a certain format, ideally organised data, but most importantly, it must be a numerical vector that represents the data. For that purpose, a number of pre-trained word embedding models exist, including the well-known Word2Vec model.

As a result, all text tweets are first translated using trained models into a format that the network can understand.

Most people use the Long Short-Term Memory (LSTM) model to process and categorise words.

It is an enhanced version of the RNN vanilla model that is suitable for lengthy text input since it can keep track of previous input for a longer time. As a result, the majority of the research addressed in this section uses an LSTM variant. For instance, Kudugunta and Ferrara acknowledged the drawback of using tweet text or tweet metadata as a single input. As a result, a Contextual LSTM that incorporates both characteristics was suggested for better bot identification. To train the model and lessen the imbalance seen, they utilised the open dataset Cresci-2017. The minority class was filled using the Synthetic Minority Oversampling Technique (SMOTE), which avoided using totally synthetic data that could have an impact on performance.

8,386 users' tweets from training data were tokenized, loaded, and given to the GloVE word embedding model for feature extraction. Additionally, before classifying in the dense layer, tweet metadata like the retweet and reply count were concatenated with the features of the tweet text.

Performance was improved compared to utilising only tweet metadata. The model was evaluated using single and combined characteristics to demonstrate the superiority of the approach, and the results showed that 96% for both precision and accuracy favoured the suggested technique.

3.3 System Architecture:Supported Learning Methods

For more precise bot identification, Knauth suggested combining account-based and tweet-based characteristics. The assumption is that the bots behave differently from actual users. As a result, characteristics based on tweets were used to infer behavioural and emotional aspects. Among them is the statistical calculation of user data attributes like the minimum, maximum, and mean of tweeting. The consistency of the chosen features was assessed using a number of classifiers, including support vector machines, random forests, logistic regression, and multi-layer perceptrons. On the public dataset Cresci-2017, with 6708 users for training and 1677 for testing, the Adaboost classifier, with 99% for both precision and accuracy, performed best.

An approach to botnet identification proposed by Wang and Paschalidis involves analysing social node interactions. The method was divided into two phases: (1) detecting an anomaly in a social "interaction" graph where different nodes are connected by edges that have strong correlated communication using large results of deviation on the degree distribution, and (2)

detecting community in such a graph. The performance of the suggested method is evaluated against other community detection techniques using actual botnet traffic

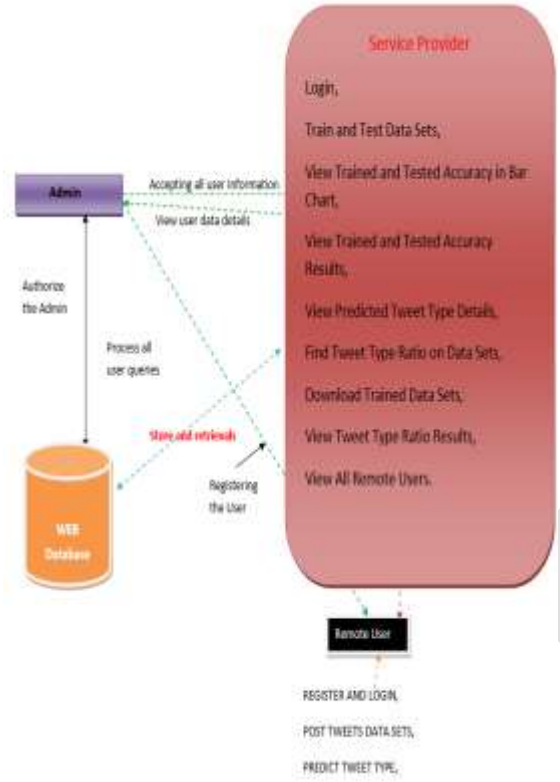

Fig 2 : System Architecture

3.4 Data Flow Diagram : Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed. For this purpose the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

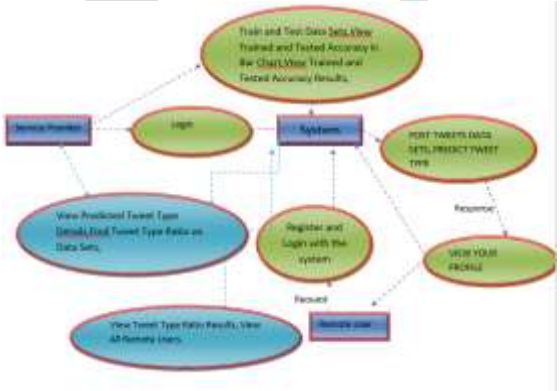

Fig 3: Data Flow Diagram

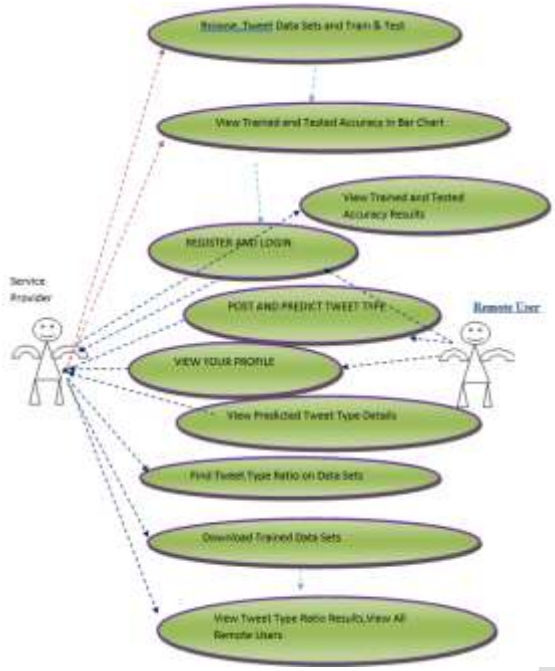

Fig 4 Use Case UML Diagrams

5. CONCLUSION

One of the most widely used social media channels for connecting individuals and assisting businesses in consumer outreach is Twitter. A tweet-based botnet may infiltrate Twitter and establish malicious accounts to carry out widespread assaults and influence operations. To combat tweet-based botnets and properly distinguish between real accounts and tweet-based bot accounts, we have concentrated on large data analytics, particularly shallow and deep learning.

In addition to discussing relevant polls, we have also supplied a taxonomy that categorises the most recent methods for detecting tweet-based bots until the year 2020. The shallow and deep learning approaches for tweet-based bot identification are also discussed, along with performance outcomes. Finally, we outlined and spoke about the unresolved problems and upcoming research difficulties.

6. FUTURE SCOPE

Deep learning algorithms especially generative adversarial networks or semi-supervised techniques may play an important role to leverage anomaly detection approach to address the major challenge of continuous change in the overall social media environment and rapid evolution of social media bots. Retrainable models through real-time processing would be another solution to this issue. Finally, most of the models are confined on twitter now. Therefore, leveraging the DL solutions to overcome similar issues in other platforms may potentially increase the usability and impact of this research to a great extent.

7. REFERENCES

[1] Rajesh Tiwari, Manisha Sharma and Kamal K. Mehta, "Dynamic Load Balancing in Parallel Processing using MPI Environment to Improve System Performance" , International Journal of Advance Research in Computer Science and Software Engineering, Vol. 5, Issue 6, June 2015, pp 730 – 734 , ISSN: 2277 – 128X.
[2] I. Arghire. (2020). Twitter Hack: 24 Hours From Phishing Employees to Hijacking Accounts. hijacking-accounts
[3] The Rise of Social Media Botnets. Accessed: Feb. 21, 2021.
[4] M. Imran, M. H. Durad, F. A. Khan, and A. Derhab, ''Toward an optimal solution against denial of service attacks in software defined networks,'' Future Gener. Comput. Syst., vol. 92, pp. 444–453, Mar. 2019.
[5] Rajesh Tiwari, Manisha Sharma and Kamal K. Mehta, "Improve the Execution Time by using GPU for Complex Application with SIMD " , International Journal of Scientific Research(IJSR), special issue March 2018, pp 227 – 232 , ISSN: 0976 – 2876. List Sr. No. 1240, Journal No. 20876.
[6] S. Aslam. (2021). Twitter by the Numbers: Stats, Demographics & Fun Facts.
[7] A. Aldweesh, A. Derhab, and A. Z. Emam, ''Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues,'' Knowl.-Based Syst., vol. 189, Feb. 2020, Art. no. 105124.
[8] S. Mahdavifar and A. A. Ghorbani, ''Application of deep learning to cybersecurity: A survey,'' Neurocomputing, vol. 347, pp. 149–176, Jun. 2019.
[9] E. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb, ''MalDozer: Automatic framework for Android malware detection using deep learning,'' Digit. Invest., vol. 24, pp. S48–S59, Mar. 2018.
[10] Rajesh Tiwari, Manisha Sharma and Kamal K. Mehta, "Performance Analysis of various High-Performance Computing Models" , Journal of Advanced Research in Dynamic and Control System, Vol. 10, special issue 2, July 2018, pp 2192 – 2200 , ISSN: 1943 – 023X.
[11]Dr. Md. Rafeeq, N. Navneetha, Dr. N. Subhash Chandra, M. Bhargavi, Dr. K Rajeshwar Rao "VM Allocation Technique and Optimized Performance Improvement for the Cloud Architecture " Volume -12 , Special Issue-4.ISSN 2063-5346,doi: 10.31838/ecb/2023.12.si4.169 2023.
[12]D. Palanivel Rajan, C. N. Ravi, Desa Uma Vishweshwar & Edem Sureshbabu ,A Review on Various Cloud-Based Electronic Health Record Maintenance System for COVID-19 Patients, ICCCE April 2023 https://doi.org/10.1007/978-981-19-8086-2_15, April 2023