Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

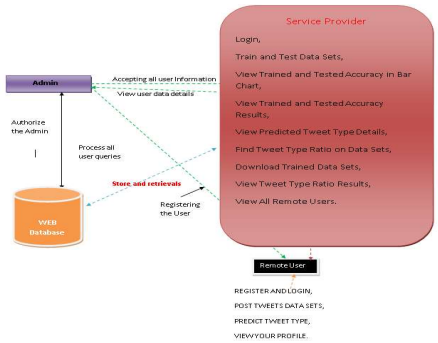
BOTWISE : HARNESSING BIGDATA TO DETECT TWITTER BOTS

[1] MOHAMMAD KOWSAR, [2] MRS. SYED ZAHADA
[1] PG Scholar in the department of MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.
[2]Professor in the department of MCA at QIS Collegeof Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

ABSTRACT:
In nowadays world lots of people like a businessman's, Media, politicians, etc., uses Twitter daily & have become an important part of life. Twitter is one of the favourite social networking sites which let the individuals to express their sentiment on various topics like politics, sports, stock market, entertainment etc. It is one of the quickest means of transmission information. It extremely affects people's viewpoint. An increasing number of people on twitter but hide their identity for malignant purpose. It is dangerous for other users hence the necessity for identifying the twitter bots. So it is essential that tweets are sent by authentic users and not by twitter bots. A twitter bot transmits spam subject matters. Therefore detecting of bots helps to determine spam messages. The characteristics of twitter accounts are utilized as Features in machine learning algorithms to label users as genuine or fake. In this paper, we used three machine learning algorithms to detect the account is fake or real, which are Decision Tree, Random Forest, and Multinomial Naive Bayes The classification performance of the algorithms is compared with their accuracy. The accuracy given by the Decision tree algorithm is 93%, the Random Forest algorithm is 90% and the Multinomial Naive Bayes is 89%. Hence it is seen that the Decision tree gives more accuracy as compared to Random Forest and Multinomial Naive Bayes

INDEX: Social media, Twitter, big data analytics, shallow learning, deep learning, tweet-based bot detection.

INTRODUCTION:
Twitter is one of the most popular micro-blogging social media platforms that has millions of users. Due to its popularity, Twitter has been targeted by different attacks such as spreading rumors, phishing links, and malware. Tweet-based botnets represent a serious threat to users as they can launch large-scale attacks and manipulation campaigns. To deal with these threats, big data analytics techniques, particularly shallow and deep learning techniques have been leveraged in order to accurately distinguish between human accounts and tweet-based bot accounts. In this paper, we discuss existing techniques, and provide a taxonomy that classifies the state-of-the-art of tweet-based bot detection techniques. We also describe the shallow and deep learning techniques for tweet-based bot detection, along with their performance results. Finally, we present and discuss the challenges and open issues in the area of tweet-based bot detection.

SYSTEM ARCHITECTURE:


METHODOLOGY:
Data Preprocessing:
Once collected, the Twitter data undergoes preprocessing to ensure its quality and consistency. This involves tasks such as

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

removing duplicates, handling missing values, and cleaning noisy or irrelevant data. Additionally, text data may be processed using techniquesliketokenization,temming, and stop-word removal to prepare it for analysis.

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Architecture/ Approach	Accuracy	FP	TP	Precision
[24]	Cresci-2017	11.4 m	8,386	N/A	GloVE SMOTE	Tweet & account Metadata	LSTM	Discriminative /Supervised	96%	N/A	N/A	96%
[25]	Cresci-2017	11.4 m	4,929	2,910	GloVe	Tweet Text&Tweet & Account Meta-data	BiLSTM	Discriminative/ Unsupervised	95%	6%	94%	93%
[26]	Own	446,334	62,762	1,000	Word em-bedding	Tweet Metadata	Autoencoder LSTM	Generative/ unsupervised	87%	N/A	N/A	93%
[27]	Moestatter	5,658	5,092	566	DeepTalk	Tweet Text & metadata	CNN + LSTM	Discriminative/ Supervised	N/A	N/A	N/A	88.4%
[28]	CLEF-2019	N/A	2,873	1,240	Word em-bedding	Tweet Text & metadata	CNN	Discriminative/ Unsupervised	85%	N/A	N/A	97%
[29]	Twitter	20	500	25,817	Spam de-tection	IDs, screen name, location	Bayesian classifica-tion	Supervised	N/A	N/A	N/A	89%
[30]	Twitter(Fake Project, Social Honeypot, User Popularity Band)	9,987,698, 5,613,166, 150,336	N/A	N/A	Social bot detection	Tweet-based, user profile-based and social graph based attributes	Deep Q-Learning (DQL)	Unsupervised	93%	N/A	N/A	80%
[31]	ASW EC2	6M	N/A	N/A	N/A	Twitter statistics + Category vector + Sentiment + LDA	Graph neu-ral network	Unsupervised	89%	N/A	N/A	N/A
[32]	Bots and Gender Profiling 2019	412,000	288,000 144,000	Bot	human tweet differenti-ation	Bi-LSTM	Deepbot	Unsupervised	79.64%	N/A	N/A	N/A
[28]	CLEF 2019	3110	2873	1240	Twitter bot	N/A	Convolutional neural network	Unsupervised	N/A	N/A	N/A	97.02%
[33]	ISIS dataset	9M	N/A	N/A	N/A	N/A	Deep neural network	Discriminative/ Unsupervised and Semi-supervised	82%	N/A	N/A	90%

Table 1: Summary of Deep learning-based detection method

Classification Accuracy:
Classification accuracy is defined the numbers of the correct predictions are divided by the total numberof inputs samples or total number of predictions made. It is mathematically given as:

$$CA = \frac{(NCP)}{(TNI)}$$

Allowing to this experiment the performance of artificial neural network is capable with highest accuracyis 98.9%. Decision tree is much closed result to ANN, having 92.7% accuracy.

Confusion Matrix:
Confusion matrix is doing well performance for the binary classification methods; we are also usingbinary classification in this research. Confusion matrix is giving the result as form of matrix, wheredescribe the full performance of the proposed model. It gives us 4 values and two classes (actual classand predict class) in output.The mathematical representation of the average accuracy of confusion matrix shown below, where 'N' isthe total number of inputs:

$$Accuracy = \frac{TP + FN}{N}$$

F1 Measure:
F1-score is also called F-measures. It is used for the measure of test's accuracy and identifying thenumber of true and positive of the precision and recall. It is the harmonic means value of the precisionand recall. In this experiment the highest if values of AAN are 96%.Mathematically represent as:

$$FM = 2 \times \frac{precssion * recall}{precission - recall}$$

Precision:
Precision define is the fraction of true positive values among number of positive values predicted by theclassifier. It is expressed as:

$$Precision = \frac{(TP)}{(TP) + (FP)}$$

Recall:
Recall, also referred to as sensitivity or true positive rate, and represents the ratio of correctly predictedpositive outcomes to the total number of samples that are actually positive.Mathematically, it can be expressed as:

$$Precision = \frac{(TP)}{(TP) + (FN)}$$

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 7, July : 2024

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Approach	Accuracy	FP	TP	Precision
[65]	Twitter	3.5 million	274,820	31,171	Emotions, sentiment	Cox regression technique (eemotionnegative, positive))	Zero-truncated negative binomial (ZTNB) regression	Supervised	N/A	N/A	N/A	N/A
[66]	Twitter	77,033	35,343	15,147	Spam detection	The collector service, database, and analyzer service	Naive Bayes	Supervised	94.3%	N/A	N/A	N/A
[67]	Twitter	2,031,081	227,196	10,280	Spam detection	Follower ratio, Account reputation, Hashtag position aggregate, URL ratio	Random forest	Unsupervised	96.30%	4.3%	96.2%	96.2%
[68]	Twitter real dataset	600 million	N/A	N/A	Class imbalance of spam detection	account_age, no_follower, no_following, no_userfavourites, no_lists, no_tweets, no_hashtag, no_retweets, no_urls, no_char, no_digits	heterogeneous stacking-based ensemble learning	Supervised	N/A	0.03	0.70	70%
[37]	Kaggle	N/A	70%	30%	N/A	name, status, description, follower count, listed count, screen name, verification status, and average semantic score	Deep Neural Network	Unsupervised	67.7%	N/A	N/A	85%
[69]	Koobface, Twitterbot, TWebot, Yazanbot, FixNazbot, Wbbot, Fbbot	4.2M	N/A	N/A	infection, pre-defined host behaviors, Command and Control	host behavior monitoring and analysis and identification	Behavior detection	Unsupervised	N/A	29.6%	4.5%	N/A

Table 2: Summary of shallow learning-based detection methods

Service Provider:

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Diabetic Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Primary Stage Diabetic Prediction, Find Primary Stage Diabetic Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users:

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorize the users.

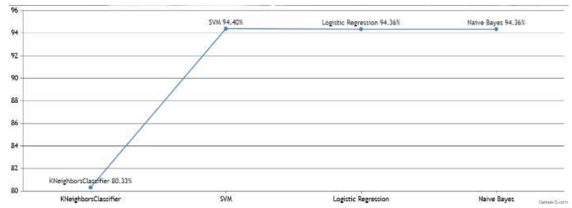
Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT TWEET TYPE, VIEW YOUR PROFILE.

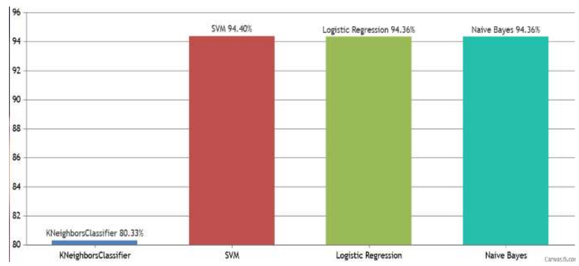
RESULTS ANALYSIS :

Tweet Type	Ratio
Bot	48.0
Quality	52.0

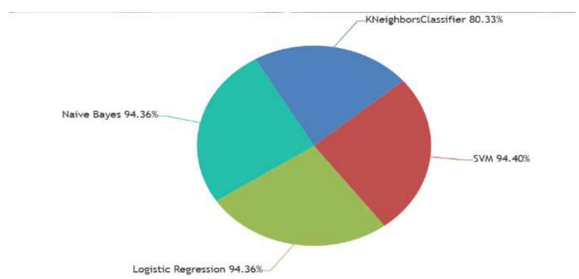
Prediction Ratio Details



Line Chart Prediction Results



Bar Chart Prediction Results



Pie Chart Prediction Results

CONCLUSION :

Twitter is one of the most popular social media platforms that allows connecting people and helps organizations reaching out to customers. Tweet-based botnet can compromise Twitter and create malicious accounts to launch large-scale attacks and manipulation campaigns. In this review, we have focused on big data analytics, especially shallow and deep learning to fight against tweet-based botnets, and to accurately distinguish between human accounts and tweet-based bot accounts. We have discussed related surveys, and have also provided a

taxonomy that classifies the state-of-the-art tweet-based bot detection techniques up to 2020. In addition, the shallow and deep learning techniques are described for tweet-based bot detection, along with their performance results. Finally, we presented and discussed the open issues and future research challenges.

FUTURE ENHANCEMENT :

Feature Extraction: Various features can be extracted from Twitter data, such as tweet frequency, account age, posting patterns, follower count, friends count, profile information, language used, sentiment analysis of tweets, and more. Python libraries like NLTK (Natural Language Toolkit) and TextBlob can be used for text analysis and sentiment analysis.

Bot Detection Rules: Developers can also create rules-based approaches to detect bots. This involves defining specific criteria or thresholds based on patterns observed in bot behavior. Python provides tools for implementing these rules efficiently.

Visualization: Python libraries like Matplotlib and Seaborn can be used to visualize the data and analysis results, making it easier to interpret and communicate findings.

REFERENCES :

1. Srikanth veldandi, et al. “Design and Implementation of Robotic Arm for Pick and Place by using Bluetooth Technology.” Journal of Energy Engineering and Thermodynamics, no. 34, June 2023, pp. 16–21. https://doi.org/10.55529/jeet.34.16.21.
2. Srikanth veldandi., et al. “Grid Synchronization Failure Detection on Sensing the Frequency and Voltage beyond the Ranges.” Journal of Energy Engineering and Thermodynamics, no. 35, Aug. 2023, pp. 1–7. https://doi.org/10.55529/jeet.35.1.7.
3. Srikanth veldandi, et al. “Intelligents Traffic Light Controller for Ambulance.” Journal of Image Processing and Intelligent Remote Sensing, no. 34, July 2023, pp. 19–26. https://doi.org/10.55529/jipirs.34.19.26.
4. Srikanth veldandi, et al. “Smart Helmet with Alcohol Sensing and Bike Authentication for Riders.” Journal of Energy Engineering and Thermodynamics, no. 23, Apr. 2022, pp. 1–7. https://doi.org/10.55529/jeet.23.1.7.
5. Srikanth veldandi, et al. “An Implementation of Iot Based Electrical Device Surveillance and Control using Sensor System.” Journal of Energy Engineering and Thermodynamics, no. 25, Sept. 2022, pp. 33–41. https://doi.org/10.55529/jeet.25.33.41.
6. Srikanth veldandi, et al “Design and Implementation of Robotic Arm for Pick and Place by using Bluetooth Technology.” Journal of Energy Engineering and Thermodynamics, no. 34, June 2023, pp. 16–21. https://doi.org/10.55529/jeet.34.16.21.
7. Srikanth, V. “Secret Sharing Algorithm Implementation on Single to Multi Cloud.” Srikanth | International Journal of Research, 23 Feb. 2018, journals.pen2print.org/index.php/ijr/article/view/11641/11021.
8. V. Srikanth. “Managing Mass-Mailing System in Distributed Environment” v srikanth | International Journal & Magazine of Engineering, Technology, Management and Research, 23 August. 2015. http://www.ijmetmr.com/olaugust2015/VSrikanth-119.pdf
9. V. Srikanth. “SECURITY, CONTROL AND ACCESS ON IOT AND ITS THINGS” v srikanth | INTERNATIONAL JOURNAL OF MERGING TECHNOLOGY AND ADVANCED RESEARCH IN COMPUTING, 15 JUNE. 2017. http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170605.pdf
10. V. Srikanth. “ANALYZING THE TWEETS AND DETECT TRAFFIC FROM TWITTER ANALYSIS” v srikanth | INTERNATIONAL JOURNAL OF MERGING TECHNOLOGY AND ADVANCED RESEARCH IN COMPUTING, 20 MARCH. 2017. http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170309.pdf

11. V. Srikanth. "A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION" v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 DECEMBER. 2017. https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf
12. V. Srikanth. "SECURED RANKED KEYWORD SEARCH OVER ENCRYPTED DATA ON CLOUD" v srikanth | INTERNATIONAL JOURNAL OF INNOVATIVE ENGINEERING AND MANAGEMENT RESEARCH, 08 Febraury. 2018. http://www.ijiemr.org/downloads.php?vol=Volume-7&issue=ISSUE-02
13. V. Srikanth. "WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3)" v srikanth | Journal of Emerging Technologies and Innovative Research (JETIR), 08 mAY. 2019. https://www.jetir.org/papers/JETIRDA06001.pdf
14. V. Srikanth, et al. "Detection of Fake Currency Using Machine Learning Models." Deleted Journal, no. 41, Dec. 2023, pp. 31–38. https://doi.org/10.55529/ijrise.41.31.38.
15. V. Srikanth, et al. "A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES." 25 Mar. 2023, pp. 300–305. http://ijte.uk/archive/2023/A-REVIEW-ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf.
16. V. Srikanth, "DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS." 25 Mar. 2023, pp. 201–209. http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf.
17. V. Srikanth, "CHRONIC KIDNEY DISEASE PREDICTION USING MACHINELEARNINGALGORITHMS." 25 January. 2023, pp. 106–122. http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf.
18. Srikanth veldandi, et al. "View of Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN". journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798.
19. Srikanth veldandi, et al. "Improving Product Marketing by Predicting Early Reviewers on E-Commerce Websites." Deleted Journal, no. 43, Apr. 2024, pp. 17–25. https://doi.org/10.55529/ijrise.43.17.25.
20. Srikanth veldandi, et al."Intelligents Traffic Light Controller for Ambulance." Journal of Image Processing and Intelligent Remote Sensing, no. 34, July 2023, pp. 19–26. https://doi.org/10.55529/jipirs.34.19.26.
21. Veldandi Srikanth, et al. "Identification of Plant Leaf Disease Using CNN and Image Processing." Journal of Image Processing and Intelligent Remote Sensing, June 2024, https://doi.org/10.55529/jipirs.44.1.10.
22. Veldandi Srikanth, et al. "Human-AI Interaction Using 3D AI Assistant" International Conference on Emerging Advances and Applications in Green Energy (ICEAAGE-2024), 15, Feb 2024, https://www.researchgate.net/publication/380971799_ICEAAGE-2024_Conference_Proceedings_Final
23. Veldandi Srikanth, et al. "Voice Based Assistance for Traffic Sign Recognition System Using Convolutional Neural Network" International journal of advance and applied research (IJAAR) ISSN – 2347-7075, 4th, April 2024, https://ijaar.co.in/wp-content/uploads/2021/02/Volume-5-Issue-4.pdf
24. Veldandi Srikanth, et al. "Convolutional Neural Network Based Heart Stroke Detection" International journal of advance and applied research (IJAAR) ISSN – 2347-7075, 4th, April 2024, https://ijaar.co.in/wp-content/uploads/2021/02/Volume-5-Issue-4.pdf

25. Srikanth veldandi, et al. “Data Analytics Using R Programming Lab | IOK STORE.” IOK STORE, ww.iokstore.inkofknowledge.com/product-page/data-analytics-using-r-programming-lab.
26. Srikanth veldandi, et al. “Data Structures Laboratory Manual | IOK STORE.” IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-structures-laboratory-manual.
27. Srikanth veldandi, et al. “Cyberspace and The Law: Cyber Security | IOK STORE.” IOK STORE, iokstore.inkofknowledge.com/product-page/cyberspace-and-the-law

AUTHOR PROFILE:

Mrs. Syed Zahada, currently working as an Assistant Professor in the Department of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She did her MCA from Azad college of computers, Hyderabad, Affiliated to Osmania University. Her area of interest is Machine Learning,Artificialintelligence, CloudComputing and Programming language.

Ms. Mohammad Kowsar, currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. She Completed B.Sc. in Statistics from Sri Harshini Degree College, Ongole, Andhra Pradesh. Her areas of interest are Machine learning& Cloud computing.