INTERNATIONAL JOURNAL OF
TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

A HYBRID APPROACH FOR DETECTING AUTOMATED SPAMMERS IN TWITTER

GUBBALA TIRUMALA
PG Scholar, Department of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India
E-Mail: tirumalagubbala@gmail.com

Mr. NAGA. SRINIVASA RAO*
Asst. Professor, Dept of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India
E-Mail:naagaasrinu@gmail.com

Abstract—
Twitter is one among the most popular micro blogging services, which is usually used to share news and updates via short messages limited to 280 characters. However, its open nature and enormous user base are often exploited by automated spammers, content polluters and other malicious users to commit various cyber crimes like cyber bullying, trolling, rumor spreading and stalking. Accordingly, variety of approaches has been proposed by researchers to address these issues. However, most of those approaches are based on user characterization and completely ignore mutual interactions. During this whitepaper, we present a hybrid approach to detecting automated spammers by merging community-based features with other feature categories, namely metadata-based, content-based, and interaction-based features. The novelty of the proposed approach lies within the characterization of users based on their interactions with their followers, since a user can evade features associated with his/her own activities, but evading those supported the followers is difficult. Nineteen different features, including six redefined features and two redefined features, are identified for learning three classifiers, namely random forest, decision tree, and Bayesian network, on a true dataset that includes benign users and spammers. The discriminatory power of various categories of features is also analyzed and it is found that interaction- and community-based features are the most effective for spam detection, while metadata-based features are shown to be the least effective.

Index Terms— **Social network analysis, spammer detection, spambot detection, social network security.**

I. Introduction

Online social media is one of the defining phenomena in this technology-driven era. Platforms like Facebook and Twitter play a key role in enabling global connectivity. An estimated 2.46 billion users are now connected, and by 2020, a third of the world's population will be connected. Users of these platforms freely generate and consume information, resulting in unprecedented amounts of data. Several sectors have already recognized the critical role of social media analytics in improving productivity and gaining competitive advantage. Information gleaned from social media has been used in healthcare to support effective service delivery, in sports to connect with fans, in the entertainment industry to complement intuition and experience in business decisions, and in politics to track electoral processes and encourage broader engagement with supporters and predict poll results. Despite the benefits, however, the rapid increase in social media spam content calls into question the credibility of the research based on the analysis of this data. A Nexgate report estimates that there is an average of one spam post for every 200 social media posts, and a recent study reports that around 15% of active Twitter users are automated bots. The growing volume of spam posts and the use of autonomous accounts (social bots) to generate posts raise many concerns about the credibility and representativeness of the data for research. Focusing on Twitter, this report proposes a novel, effective approach to detecting and filtering unwanted tweets that complements previous approaches in this direction. Previous studies are based on historical characteristics of tweets, which are often no longer available on Twitter after a short time and are therefore not suitable for real-time use. Our approach leverages an optimized set of readily available features, independent of historical text features on Twitter. The functions used are categorized in relation to the Twitter account, the user or Recently, Vosoughi et al. discover that both genuine and false news spread at equal rate. False news on Twitter spread rapidly. Social bots are deployed to accelerate the process and human users further amplify the content. To detect spam tweets, numerous detection systems have been proposed, using various techniques that are reviewed in this section. in relation to the pairwise interaction between users. A number of machine learning models were trained. A recursive feature elimination was used to determine the robustness and distinctiveness of each feature. Compared to a previous study, the proposed traits demonstrate greater power of discrimination with more consistent performance across the different learning models. Users who post spam employ some evasive tactics such as B. posting an average of 4 tweets per day, and tricks to balance the follower-followee relationship. Our analysis shows that an average automated spam posting account posts at least 12 tweets per day within well-defined periods of activity. The pattern of activity is similar to the staircase function, showing waves of intermittent activity. Our study contributes (a) a new set of lightweight features suitable for real-time detection of spammers on Twitter, and (b) an additional dataset1 that provides insight into the behavior of spam users on Twitter.

II. RELATED WORKS

Spam entails any form of activity that causes harm or disrupts other online users. The increasing amount of spam tweets can be attributed to humans' inclination to spread misleading information, even if such information originated originated from unreliable sources, such as a social bot account.

Thomas et al. and Lee and Kim analysed streams of URLs used by spam users and studied how spammers exploit URLs obfuscation to redirect users to malicious sites.

Grier et al. analysed a large number of distinct URLs pointing to blacklisted sites due to their involvement in scam, phishing and malware activities. Although the approach is effective, it is often slow and fails to detect URLs that point to malicious sites but have not been blacklisted previously.

Gao et al. also studied URL usage on Facebook to detect spamming activity and observed that this form of spamming is mostly associated with compromised accounts rather than accounts created solely for spam activity.

Benevenuto et al. studied the statistical properties of user accounts and how URL shortening services affect spam detection mechanisms. However, the universal use of URLs and URL shortening by the vast majority of Twitter users makes it difficult to directly identify potentially nefarious links on a large scale. In general, the use of URLs relies on historical information, limiting the possibilities for real-time detection.

Danezis and Mittal utilised a social network model to infer legitimate user accounts that are being controlled by an adversary.

Lee et al. created social honeypot accounts mimicking naive Twitter users to entice spam posting users. Users who fall prey by engaging with these accounts are assumed to be in violation of usage policy. Users identified using this method were analysed to distinguish different user types focusing on link payloads and features that can capture the dynamics of follower-following networks of users.

Varol et al. employed many features related to users, content and the network to develop a system for social bot account detection.

Chen et al. provides an in depth analysis of deceptive words used by spammers on Twitter. The work of Chen et al. is motivated by Twitter Spam Drift, i.e. the property of statistical features of spam tweets to change over time. Twitter Spam Drift is caused because spammers continuously adopt and abolish various evasive tricks. Features related to this phenomenon were utilised in training machine learning classifiers.

Li and Liu analysed how the effect of unbalance datasets can be mitigated in detection tasks. Standard machine learning methods are sometimes considered as inadequate in capturing the variability of spamming behaviour.

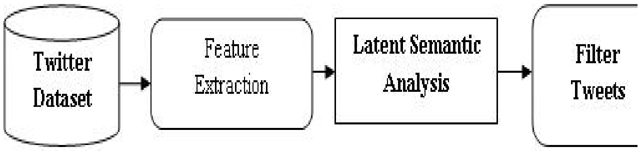
Wu et al. utilised a deep learning technique based on Word2Vec to capture the variation of spam-related challenges. While it is essential to allow detection models to continuously learn features strong enough to distinguish spam from nonspam, methods that solely rely on textual information are be inadequate to draw the distinction between a habitual spam posting account and a non-spam posting account.

III. PROPOSED APPROACH

- To address existing drawbacks, this work proposed a novel, effective approach to detect and filter unwanted tweets. Our approach leverages an optimized set of readily available features, independent of historical text features on Twitter. The functions used are categorized in relation to the Twitter account, the user or in relation to the pairwise interaction between users. A number of machine learning models were trained.

- Recursive feature elimination was used to determine the robustness and distinctiveness of each feature. Users who post spam employ some evasive tactics such as B. posting an average of 4 tweets per day, and tricks to balance the follower-followee relationship.

System Architecture



For the experimental evaluation the proposed of twitter data set is provided.This data set also contains of followers and spammers in twitter along their profiles. Features of extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Latent semantic analysis is a technique in natural language processing.

IV. Modules

1. Load Twitter Dataset

- In this module, a user load twitter dataset for detect automatic spammer posting account.
- This dataset contains user's posting and his account details.
- This post has any contents.

2. Feature Extraction

In the proposed automated spammer detection method, 19 features, including 6 new and 2 redefined features, are identified. The feature set is classified into three broad categories, namely

1. Community-based
2. Metadata-based
3. Content-based
4. Interaction-based.

3. Latent Semantic Analysis:

- This module takes input dataset as Feature extraction module spam detection results.
- This module once again detects spam using latent semantic analysis.
- It provides accurate results.
- Using spam posting, this module identifies spam posting user account.

V. ALGORITHMIC STRATEGY

Algorithm 1: Spam posting account detection

Input: Twitter Dataset (TD)

Output: Detect Spam Account

Step 1: Extract Tweet T From TD

Step 2: Extract Feature from T

Step 3: Apply Latent Semantic Analysis In Each T

Step 4: Classify Tweet into spam or not using LSA Algorithm

Step 5: Find Spam Account.

Step 6: End

VI. EXPERIMENTAL SETUP AND RESULTS.

The performance of the proposed study is analyzed using three machine learning techniques, namely *random forest*, *decisiontree*, and *Bayesian network* on the dataset.

A. Evaluation Metrics

The proposed approach is evaluated using three standard metrics, namely, *detection rate* (*DR*), *false positive rate* (*FPR*), and *F-Score*. *DR* (aka *recall*) represents the fraction of spammers detected from the set of all spammers, and it is defined using Equation , where *TP* stands for true positives and represents the number of actual spammers classified as spammers, and *FN* stands for false negatives and represents the number of actual spammers misclassified as benign users.

B. *Evaluation Results*

The performance of the proposed approach is evaluated using three classifiers, namely, *random forest*, *decision tree*, and *Bayesian network*, which are implemented in Weka.6 We have used ten-fold cross validation to ensure the participation of each instance in both training as well as testing procedure. The performance of the classifiers is evaluated using standard evaluation metrics, namely, *DR*, *FPR*, and *F-Score.*

VII. CONCLUSION

This study offers an effective method for spam detection and new insights into the sophisticatedly evolving techniques for spamming on Twitter. The proposed spam detection method utilized an optimized set of readily available features. Being independent of historical tweets which are often unavailable on Twitter makes them suitable for real-time spam detection. The efficacy and robustness of the proposed features set is shown by testing a number of machine learning models and on dataset collected orthogonally from the study data. Performance is consistent across the different models and there is significant improvement over the baseline. It was also shown that automated spam accounts follow a well-defined pattern with surges of intermittent activities. The proposed spam tweet detection approach can be applied in any real-time filtering application. For example, it is applicable to data collection pipelines to filter out irrelevant content at an early pre-processing stage to ensure the quality and representativeness of research data. The combination of handcrafted features and features learnt in an unsupervised manner using word embeddings is shown to significantly improve baseline performance and to perform comparably to the best performing feature set using a smaller number of features.

References

1. Improving product marketing by predicting early reviewers on E-Commerce websites
S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, “Improving product marketing by predicting early reviewers on E-Commerce websites,” Deleted Journal, no. 43, pp. 17–25, Apr. 2024, doi: 10.55529/ijrise.43.17.25.

2. Kodati, Dr Sarangam, et al. “Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN.” Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD) 2799-1202, vol. 3, no. 06, 23 Nov. 2023, pp. 32–40, journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798, https://doi.org/10.55529/jpdmhd.36.32.40. Accessed 2 May 2024.

3. V. Srikanth, “CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS,” IJTE, pp. 106–109, Jan. 2023, [Online]. Available: http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf

4. V. SRIKANTH, “DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS,” International Journal of Technology and Engineering, vol. XV, no. I, pp. 201–204, Feb. 2023, [Online]. Available: http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf

5. V. SRIKANTH, “A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES,” IJTE, vol. XV, no. I, pp. 300–302, Mar. 2023, [Online]. Available: http://ijte.uk/archive/2023/A-REVIEW-ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf

6. S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, “Detection of fake currency using machine learning models,” Deleted Journal, no. 41, pp. 31–38, Dec. 2023, doi: 10.55529/ijrise.41.31.38.

7. “Cyberspace and the Law: Cyber Security.” IOK STORE, iokstore.inkofknowledge.com/product-page/cyberspace-and-the-law. Accessed 2 May 2024.

8. “Data Structures Laboratory Manual.” IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-structures-laboratory-manual. Accessed 2 May 2024.

9. Data Analytics Using R Programming Lab.” IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-analytics-using-r-programming-lab. Accessed 2 May 2024.

10. V. Srikanth, Dr. I. Reddy, and Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, 501301, India, “WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3),” journal-article, 2019. [Online]. Available: https://www.jetir.org/papers/JETIRDA06001.pdf

10. V. SRIKANTH, “Secured ranked keyword search over encrypted data on cloud,” IJIEMR Transactions, vol. 07, no. 02, pp. 111–119, Feb. 2018, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/1121_approvedpaper.pdf

11. V. SRIKANTH, “A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION,” IJIEMR Transactions, vol. 06, no. 12, pp. 337–344, Dec. 2017, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf

12 . SRIKANTH MCA, MTECH, MBA, “ANALYZING THE TWEETS AND DETECT TRAFFIC FROM TWITTER ANALYSIS,” Feb. 2017. [Online]. Available: http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170309.pdf

14 Srikanth, V. 2018. “Secret Sharing Algorithm Implementation on Single to Multi Cloud.” International Journal of Research 5 (01): 1036–41. https://journals.pen2print.org/index.php/ijr/article/view/11641/11021.

5. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023.

16. Babu, Dr P. Sankar, et al. “Intelligents Traffic Light Controller for Ambulance.” Journal of Image Processing and Intelligent Remote Sensing(JIPIRS) ISSN 2815-0953, vol. 3, no. 04, 19 July 2023, pp. 19–26, journal.hmjournals.com/index.php/JIPIRS/article/view/2425/2316, https://doi.org/10.55529/jipirs.34.19.26. Accessed 24 Aug. 2023.

17. S. Maddilety, et al. “Grid Synchronization Failure Detection on Sensing the Frequency and Voltage beyond the Ranges.” Journal of Energy Engineering and Thermodynamics, no. 35, 4 Aug. 2023, pp. 1–7, https://doi.org/10.55529/jeet.35.1.7. Accessed 2 May 2024.

18. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023.