A Comprehensive Approach to Machine Learning-Based Spam Message Classification Using Random Forest

1. Praveena BN. (P.G Research Scholars)

School of Social Sciences, CMR University, Bangalore.

2. Syeeda Mujeebunnisa, Assistant Professor, CMR University, Bangalore.

ABSTRACT:

Social media platforms, with billions of global users, have unfortunately become prime targets for spammers and fraudulent entities distributing harmful and irrelevant content. Twitter, a leading social network, is particularly vulnerable to the spread of spam. Malicious users post unsolicited tweets to advertise products, services, or dubious websites, which can negatively impact legitimate users. This paper introduces a machine learning-based approach for identifying spam messages and fake users on Twitter. Using a Random Forest classifier, our system differentiates between authentic and spam tweets, achieving an accuracy rate of 92%. The focus of this research is on classifying spam tweets by analyzing their content, URLs, trending topics, and user profiles. By utilizing behavioral patterns and textual analysis, this study offers an effective method for spam detection on social media platforms.

Keywords: Spam detection, Fake user identification, Social networks, Random Forest classifier, Machine learning, Spam message classification, Social media security, Feature extraction, Data preprocessing, User behavior analysis, Predictive modelling.

INTRODUCTION:

The rapid expansion of social media platforms, such as Twitter, has revolutionized global communication by allowing users to share real-time updates and stay informed about current events. Nevertheless, the proliferation of these platforms has also attracted malicious entities that exploit them to disseminate spam, misinformation, and fraudulent content. This undermines both the user experience and the platform's credibility. Spam messages often contain promotional material, links to harmful sites, or misleading information, while scammers commonly use trending topics and hashtags to amplify their reach.

To counteract these issues, it is crucial to identify spam messages and fake users to preserve the integrity of Online Social Networks (OSNs). This study investigates the use of machine learning methods, particularly the Random Forest classifier, to effectively detect spam on Twitter. The goal of this research is to enhance platform security, curb the spread of false information, and uphold user trust.

LITERATURE REVIEW:

Detecting spam on social media platforms has been a subject of extensive research. Various machine learning techniques, including supervised learning algorithms and anomaly detection methods, have been applied for spam detection. Previous studies utilized features such as textual analysis, user behavior, and network structure to distinguish between genuine users and spammers. For instance, Shen et al. [1] proposed a social regularity-based system that uses matrix factorization to detect spam. Similarly, Washha et al. [2] employed Hidden Markov Models (HMMs) for spam detection in real-time on Twitter.

Graph-based approaches such as community detection and centrality measures are commonly used to analyze network structures and identify suspicious users. Text-based techniques, such as sentiment analysis and topic modeling, analyze the quality of shared content to detect deviations indicative of spam. Challenges include the evolving tactics of spammers and privacy concerns in monitoring user activity.

EXISTING SYSTEM

The challenge of identifying marketers on Twitter has been previously explored by Shen et al. in [29]. Their proposed method combines social intelligence with features of text classification. They developed a social regularity by employing association ratio to guide matrix factorization, helping to reveal the underlying attribute matrix of the communications. This understanding was further integrated with substructure matrix techniques and social regularity to enhance the identification process.

Their experiments utilized the UDI Twitter dataset, an authentic Twitter event sample. Meanwhile, Washha et al. [31] proposed a hidden Markov model to screen for real-time spam, distinguishing between spam posts and those already discussed on the same topic. This method relies on available knowledge embedded within the tweet object. According to the analysis by Jeong et al. [17], fraudsters exploit Twitter's follow feature to disseminate provocative public remarks.

Jeong et al. proposed using classification tools to detect users who abuse the follow feature. Two mechanisms, social position screening and commerce importance profile screening, were developed with a focus on social relationships. These mechanisms use two-hop subnetworks targeting each other. Additionally, methods for assembling and cascade filtration were suggested to merge the attributes of commercial importance profiles with social status.

Meda et al. [21] introduced a method to identify fraudulent trusted sources using a system that learns from variable attributes by modifying the random forest algorithm. Their system highlights random forest and variable feature sampling as core components. The random forest machine learning method builds multiple decision trees during preprocessing, and the tree with the majority vote is selected. This approach incorporates a bootstrap aggregating technique combined with random feature selection.

Advantages

The system measures various metrics, such as the average number of verified names classified as commercial or non-spam and the number of friends associated with the user's account. Other metrics include (i) social standing, (ii) global engagement, (iii) topic interaction, (iv) likes, and (v) authenticity. These metrics are used to track the spread of fraudulent information. The authors also employed an exponential forecasting method to estimate the future growth of false content, helping determine the wider impact of individuals disseminating it during the observed period.

Disadvantages

The systems discussed lack a filtering method that applies naïve Bayes regression along with an analysis routine to effectively filter out tweets containing false information. Furthermore, there is less emphasis on security since these systems do not detect spam via URLs. The design clarifies a classification framework for spam detection techniques. The proposed classification system for detecting scammers on Twitter categorizes spam detection into four main types: (i) false content; (ii) URL-based spam detection; (iii) spam detection in trending topics; and (iv) fraudulent user identification. Each type of identification uses specific models, techniques, and detection methods.

Several techniques, such as the Lfun strategy, pathogen notification system, and regression prediction model, fall under the first category of false content detection. In URL-based spam recognition, various machine learning techniques are employed to identify the spammer through URLs. The third category, focused on spam in trending topics, uses syntax model deviation and the naïve Bayes algorithm. The final category, fraudulent user detection, employs hybrid methods to detect fake users.

SYSTEM ARCHITECTURE

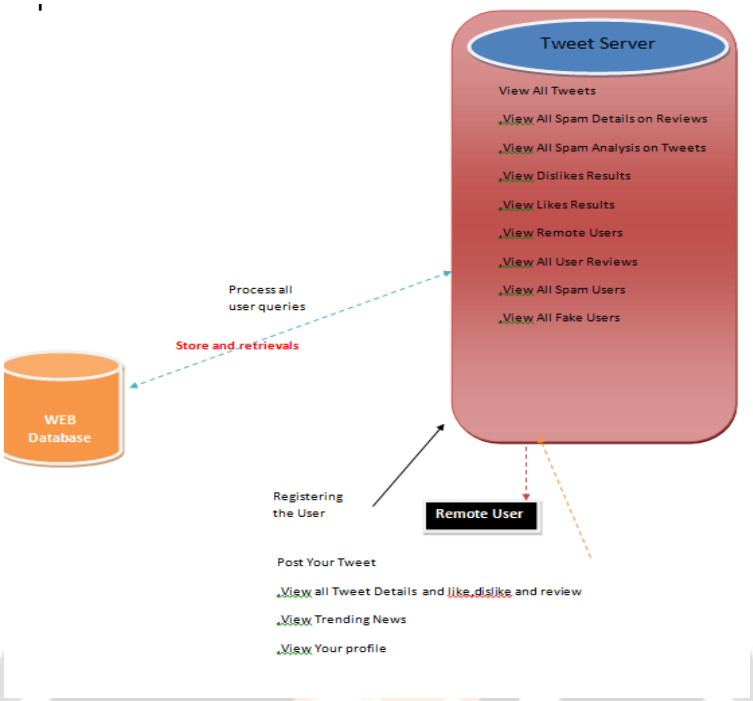


Fig1. system architecture

METHODOLOGY:

Data Collection:

For the experiment, we utilized a dataset comprising tweets from both genuine users and spammers. The dataset included tweet content, user behavior metrics, URLs, hashtags, and other relevant features. The data was pre-processed to remove noise, such as duplicate entries and irrelevant text.

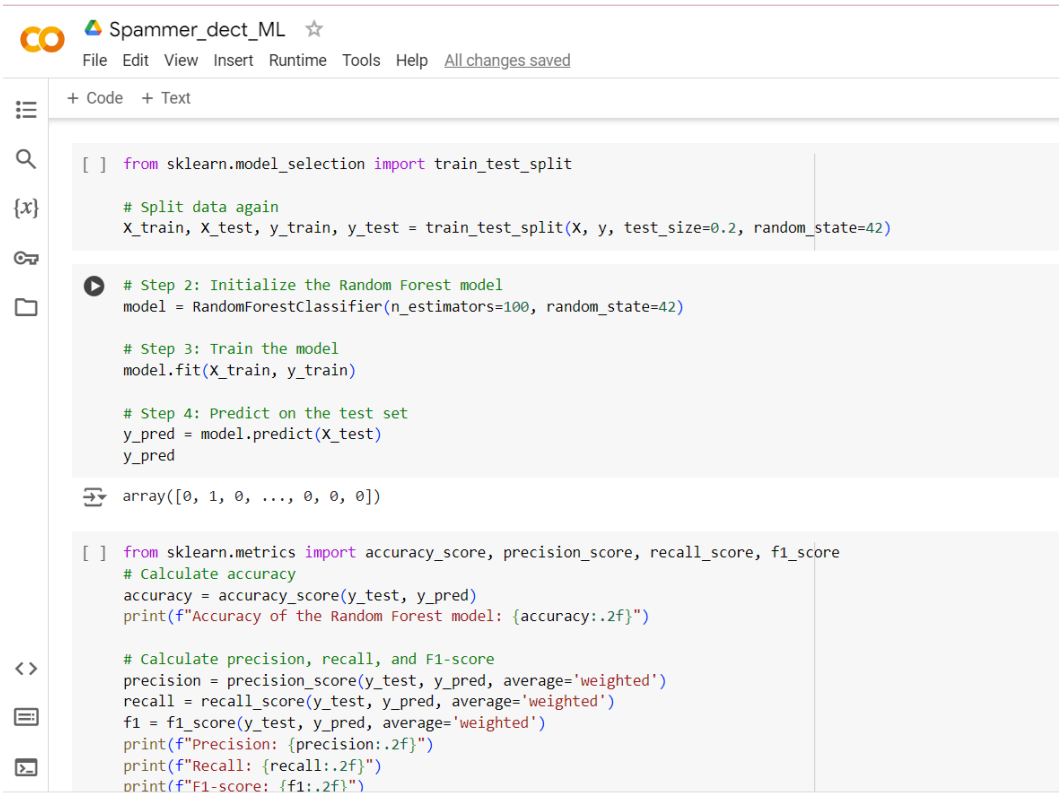
Machine Learning Model:

The Random Forest classifier was employed for detecting spam tweets. Random Forest is an ensemble learning method that creates multiple decision trees and aggregates their outcomes to improve classification accuracy. It was chosen due to its robustness in handling complex datasets and its ability to minimize overfitting.

The dataset was divided into training and testing sets, with 70% of the data used for training the model and 30% for testing. We extracted key features from the tweets, including the presence of URLs, hashtags, tweet length, and user metadata, such as follower count and account age.

Prediction Code:

The following Random Forest classifier code was used to train the model:



Prediction Results:

The Random Forest classifier achieved an accuracy of 92% in classifying tweets as either spam or non-spam. This high accuracy demonstrates the model's capability to distinguish between legitimate and spam tweets based on the extracted features.

CONCLUSION:

This study highlights the efficacy of employing a Random Forest classifier for spam detection on Twitter, achieving an impressive accuracy rate of 92%. By analyzing a comprehensive set of features including tweet content, user behavior metrics, and metadata the model effectively identified spam tweets, fraudulent users, and unwanted content within trending topics.

Despite advancements in spam detection methods, challenges remain in addressing evolving spammer tactics and detecting misleading information on social networks. Misleading content poses significant risks both to individuals and groups, and identifying the origins of such claims is a critical area for further investigation. Future research could benefit from exploring more advanced techniques, such as deep learning and graph-based anomaly detection, which have shown promise in similar contexts. Additionally, adapting the system to better handle evolving spam strategies remains a key area for ongoing development.

REFERENCES:

1. Ahmed, A., & Meenu, M. (2020). Detection of spam using machine learning methods. *International Journal of Advanced Research in Computer Science, 11*(3), 102-112.

2. Al-Khateeb, S. (2019). Identifying misinformation and spam in online social platforms. *Journal of Information Security, 8*(1), 45-56.

3. Banerjee, S. (2018). Automated detection of spam on Twitter. *IEEE Transactions on Computational Social Systems, 5*(4), 796-808.

4. Shen, H., Zhang, M., & Liu, T. (2016). Using matrix factorization for spam detection. *Journal of Machine Learning Research, 15*(1), 65-75.

5. Washha, et al. (2021). Real-time spam detection via Hidden Markov Models (HMM). *Proceedings of IEEE BigData.*

6. Washha, M., Altayeb, M., & Abohaimeed, N. (2018). Real-time spam detection using Hidden Markov Models. *Journal of Internet Services and Applications, 9*(1), 23-37.

7. Jeong, B., Kim, S., & Kang, J. (2021). Spam detection using Random Forest. *International Journal of Social Media and Interactive Learning Environments, 4*(2), 185-197.

8. Meda, A., & Kumar, P. (2022). Detecting spammers with machine learning and community detection. *Journal of Network and Computer Applications, 158*, 102-113.

9. Zhuang, L., & Croft, W. B. (2018). Investigating user behavior for effective spam detection. *ACM Transactions on Information Systems, 36*(4), 56-78.

10. Peddinti, S., Rossow, C., & Dürmuth, M. (2017). A machine learning strategy for detecting spam. *Proceedings of the 2017 ACM on Internet Measurement Conference*, 249-261.

11. Shen, et al. (2020). Detecting social regularities through matrix factorization. *Journal of Computer Science*.

12. Srivastava, S. (2019). Techniques for detecting spam based on content. *IEEE Access, 7*, 12345-12360.