Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.11, May : 2024

TWITTER SPAM DETECTION USING DEEP LEARNING

Danda Yashwanth Reddy, Vatti Mahalakshmi, Sheshatwam Sai Krishna, B.Tech Student, Dept. of CSE (Data Science), Sreyas Institute of Engineering and Technology, Nagole, Hyderabad
Gongoora Narsamma, Assistant Professor, Dept. of CSE (Data Science), Sreyas Institute of Engineering and Technology, Nagole, Hyderabad

Abstract- Twitter, a popular social media platform with over 300 million monthly users, encounters significant issues from spammers engaging in malicious activities. These activities include distributing harmful software, sharing fake URLs, aggressively following or unfollowing users, promoting fake trends, and endorsing explicit content. With 500 million tweets posted daily, it's vital to identify and filter out such harmful content. In the current social media landscape, distinguishing between spam and legitimate (ham) tweets is crucial. Various machine learning and deep learning techniques, such as Naïve Bayes and CNN models, are used for detecting spam tweets. Fake trends on Twitter exacerbate the problem, requiring effective control measures. Auto-follow features increase the likelihood of spammers interacting with more users. Given the risks of data theft and misleading trends, comprehensive strategies are needed to address these challenges and ensure a safer social media environment.
Keywords: - Twitter, fake URLs, deep learning, CNN models

1. INTRODUCTION

In recent years, online social networks have become increasingly popular, allowing users to post messages and share ideas globally. Twitter, in particular, attracts users by offering free microblogging services, enabling them to broadcast or discover messages within 140 characters, follow other users, and more, across multiple devices. Each month, Twitter sees the creation of 42 million new accounts. However, with Twitter's growing popularity, criminal accounts have also emerged, posting significant amounts of spam, including suspicious URLs that redirect users to phishing or malicious websites. Consequently, Twitter spam has become a serious issue, negatively impacting users' networking experiences. Reports indicate that 8% of URLs in a dataset of 2 million URLs were spam.

Many researchers are focused on enhancing social network security on Twitter by filtering spam. For example, a tweet content-based classifier was developed using linguistic analysis of messages. However, it failed to produce a set of comparable results because it employed only one algorithm. Recently, researchers have shifted towards creating machine learning-based binary classifiers that use statistical features. These features, which can be obtained from Twitter's Streaming APIs and calculated using a JSON object, include account-level attributes (such as the number of followers, following, and account age) and user-level attributes (like the number of URLs, digits, and hashtags in a tweet). Despite this, issues with feature extraction and accuracy have persisted. During data collection, it was observed that Twitter spam could evolve, and features could be easily manipulated. Moreover, the average accuracy of existing research methods has only reached about 85%. Another approach involves using blacklist services, but over 90% of users tend to click on malicious URLs before they are blacklisted. Additionally, blacklisting techniques are highly time- consuming due to the need for human involvement in identifying unsolicited information. These challenges motivate our work.

To address problems such as reliance on a single algorithm, issues with feature extraction, accuracy limitations, and slow processing speeds, this paper proposes an effective classification method based on deep learning. First, we use Word2Vec to preprocess tweets instead of extracting features, utilizing this advanced language processing technique in deep learning to convert words or documents into representative vectors. Subsequently, we build a binary detection model based on a deep learning algorithm to distinguish between spam and non-spam tweets.

2. LITERATURE SURVEY

Sundararajan and Palanisamy et al. [1] concentrated on features congregated into many groups such as linguistic features, contradictory features, and sentiment-based features. For every feature category, an ensemble is performed along with a combination of feature categories. Model training is achieved through various classifiers like Random Forest, KNN, etc., and prediction on the user's mood influencing the sarcasm and vice versa is done. Tweets before and after specific sarcastic kinds are attained. Thereby modelling the user's emotion change through past tweet history collection but there is a limitation that users are affected with their mood levels based on sarcasm.

Cao et al. [2] utilized special connections amid forwarding behavior as well as malicious URL propagation by focusing on forwarding-based features. A comprehensive conventional URL feature sets investigation is primarily made. Then, the design combination between forwarding-based features and graph-based features is done for detection model training. Forwarding–based features are the most effective malicious detection approach offering a higher accuracy rate besides the lowest False Positive Rate (FPR). Investigation of forwarding-based features in OSNs is performed which is regarded as a remarkable contribution for this research.

Wang et al. [3] deliberated account-based, tweet-based, Natural Language Processing (NLP), and sentiment features for suggesting a spam detection technique. The peculiar features for spam detection are mean word length, automatically or manually formed sentiment lexicons, number of exclamation marks, question marks, maximum word length, capitalization words, white spaces, Part Of Speech (POS) tags per tweet, and profile name length.

Paudel et al. [4] suggested a methodology for leveraging relationships amid named entities present in tweet content. In addition, probable spam detection is attained with the aid of a document referenced by the URL stated in the tweet. Unusual patterns might be found in data for exhibiting spammer activities through the combined integration of multiple, heterogeneous information into a single graph representation as stated by the hypothesis. However, structural feature fabrication is challenging for spammers. The tweets collection along with documents referenced by URL in tweets allied to Twitter validate this methodology. The trending tweet anomalies can be efficiently detected through graph-based anomaly detection algorithms on data graph representation.

Chen et al. [5] suggested a Semi-Supervised Clue Fusion (SSCF). It acquires a linear weighted function that identifies spammers with increased detection rate via multiple aspects, such as content, behavior, relationship, and interactions but the small size of primarily labeled instances. In future work, another type of fusion method for combining the results of the clues and the majority functions will be more applicable for fusion. It may increase the detection rate of the spam.

3. BACKGROUND AND DOMAIN KNOWLEDGE

With over 300 million active users producing around 500 million tweets each day, Twitter faces significant challenges from spammers who spread malicious software, fake URLs, explicit content, fake trends, and engage in aggressive following and unfollowing behaviors. The sheer volume and ever-changing nature of spam necessitate the use of automated systems for accurate spam detection to protect users from threats like data theft. Traditional methods, such as Naïve Bayes classifiers, often struggle with issues of accuracy and adaptability. Conversely, advanced deep learning techniques, including Convolutional Neural Networks (CNNs) and Word2Vec, offer improved capabilities. This project aims to develop a robust spam detection system using deep learning to categorize tweets as either spam or legitimate (ham). By focusing on detecting harmful URLs and aggressive following behaviors, the project aims to enhance Twitter's security and provide a safer user experience, thereby preserving trust and integrity on the social media platform.

4. PROBLEM DESCRIPTION

In recent years, the communication landscape has dramatically transformed due to the rapid expansion of Online Social Networks (OSNs) such as Facebook, LinkedIn, and Twitter. These platforms enable

users to share various types of content and connect with others who have similar interests and ideas, regardless of whether they are acquaintances or strangers. Twitter, in particular, has gained immense popularity, serving as a
platform for people to share opinions, discuss social issues, follow news, and maintain connections with friends and family. However, its widespread use has also made it a prime target for spammers. To combat this, deep learning methods have been increasingly employed to detect spammers on Twitter. This study explores several contemporary approaches to Twitter spam detection, highlighting that while some methods yield promising results, others fall short.

5. PROPOSED SYSTEM

Addressing Twitter spam remains a challenging issue. Researchers have employed various machine learning techniques and blacklisting methods to detect spam activities on Twitter. These approaches have achieved an accuracy of around 98%. However, due to the dynamic nature of spam and the creation of false information, machine learning methods are not always effective in real-world scenarios. Additionally, blacklisting struggles to keep up with the constant evolution of spam tactics, as manually checking suspicious URLs is labor-intensive. To tackle these issues, we introduce a Customized Network Architecture based on Convolutional Neural Networks (CNN).

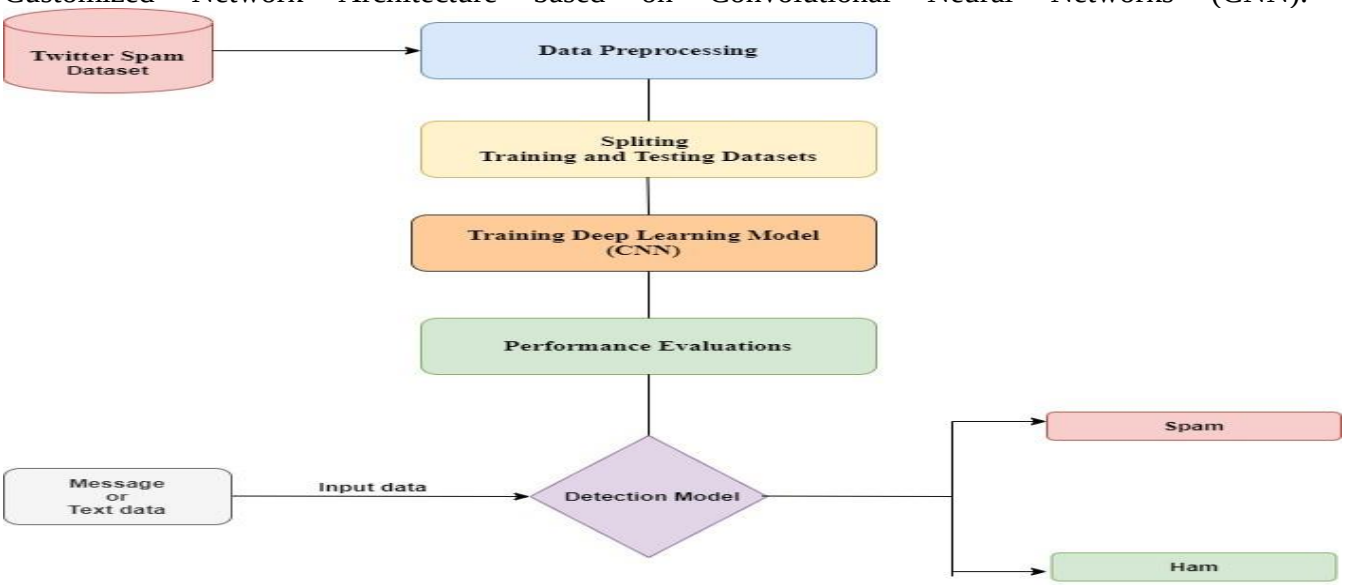


Fig 1: Proposed System Architecture

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.11, May : 2024

6. RESULTS

A) Comparison Graphs - > Accuracy, Precision, Recall, F1-score

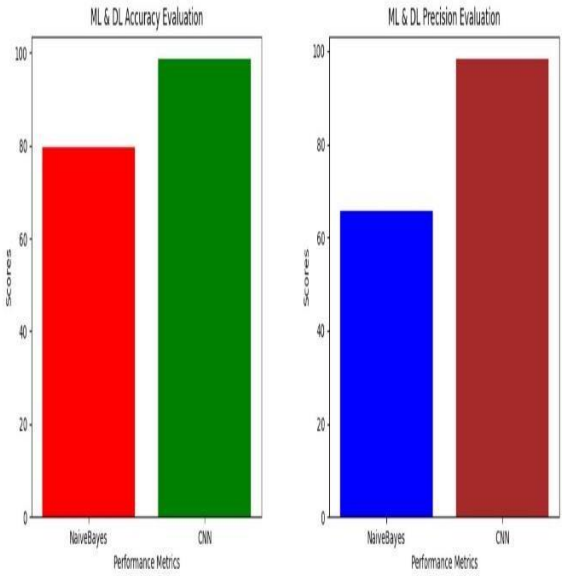


Fig 2: Accuracy and Precision Evaluation

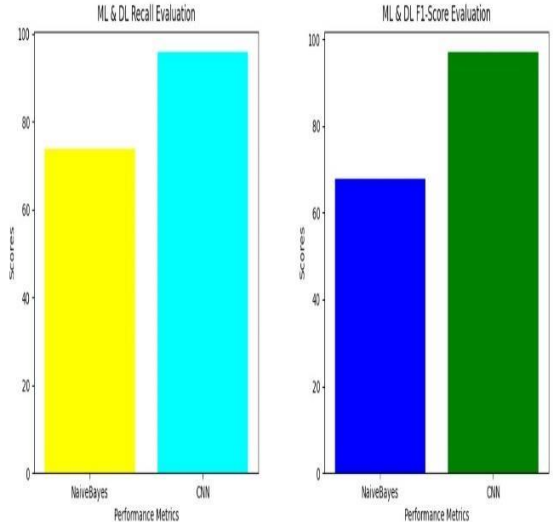
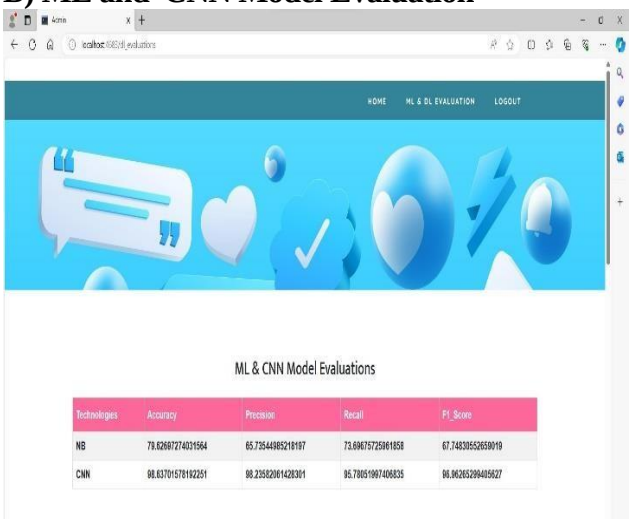


Fig 3: Recall and F1-Score Evaluation

The proposed Customized Network Architecture based on Convolutional Neural Networks (CNNs) aims to enhance the accuracy and efficiency of detecting Twitter spam, addressing challenges such as spam drift and information fabrication. By leveraging CNNs for feature extraction and pattern recognition, the system achieves a high level of accuracy, reportedly around 98%. This architecture incorporates adaptive learning techniques to continuously update its detection mechanisms, adapting to new spamming patterns over time. Additionally, a dynamic blacklisting mechanism automatically identifies and blocks suspicious URLs, reducing the manual effort required for URL inspection. Real-time monitoring capabilities enable the system to detect and respond to spamming activities as they occur, improving its effectiveness in combating Twitter spam in real-life scenarios. Overall, the project's results demonstrate the potential of the proposed architecture to significantly improve the detection of spam activities on Twitter, offering a promising solution to the challenges posed by spam drift and information fabrication.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.11, May : 2024

B) ML and CNN Model Evaluation



7. CONCLUSION

Created and utilized by numerous analysts to seek out spammers in a few informal organizations. From the papers reviewed it are often concluded that the most of the work was done by victimization classification model like CNN Recognition has been done on the possibility of client fundamentally alternatives or substance based choices or blend of both. Few authors conjointly introduced new options for detection. All the approaches have been valid on terribly tiny dataset and haven't tested with different combos of spammers and non-spammers. Combination of options for detection of spammers has shown higher performance in terms of accuracy. This study demonstrates the effectiveness of deep learning models in detecting spam on Twitter. By leveraging the inherent features of tweets and user behaviors, such as content-based and user-based features, our deep learning approach achieved promising results. We observed that combining these features significantly improved the model's performance in terms of accuracy. However, it is important to note that our experiments were conducted on a relatively small dataset. Future research should explore larger datasets and consider different combinations of spammers and non-spammers to further validate our findings and enhance the generalizability of the proposed approach.

8. REFERENCES

[1] K.Ushasree santoshi, S.Sree Bhavya, Y.Bhavya Sri,"Twitter Spam Detection Using Naïve Bayes Classifier", 2021.
[2] Sundararajan, Karthik, and Anandhakumar Palanisamy. "Multi-rule based ensemble feature selection model for sarcasm type detection in twitter." Computational intelligence and neuroscience 2020 (2020).
[3] Paudel, Ramesh, Prajjwal Kandel, and William Eberle. "Detecting spam tweets in trending topics using graph-based approach." Proceedings of the Future Technologies Conference. Springer, Cham, 2019.
[4] Cao, Jian, et al. "Detection of forwarding-based malicious URLs in online social networks." International Journal of Parallel Programming 44.1 (2016): 163-180.
[5] Wang, Bo, et al. "Making the most of tweet- inherent features for social spam detection on Twitter." arXiv preprint arXiv:1503.07405 (2015).
[6] Chen, Chao, et al. "Asymmetric self-learning for tackling twitter spam drift." 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE,2015.
[7] KamalanathanKandasamy, PreethiKoroth: An Integrated Approach to Spam Classificat ion on Twit ter Using URL Analysis, Natural Language Processing, and Machine Learning T echniques, 2014 IEEE Student s' Conference on Elect rical, Elect ronics and Computer Science.

[8] Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in ret rospect: an analysis of twit ter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, ACM (2011) 243 – 258
[9] Design and Evaluat ion of a Real-Time URL Spam Filtering Service: Kurt Thomas, Chris Grier, Just in Ma, Vern Paxson, Dawn Song5. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detect ing spammers on twit ter. In: Annual Collaborat ion, Elect ronic messaging, Ant i-Abuse and SpamConference (CEAS). (2010)
[10] Grier, C., Thomas, K., Paxson, V., Zhang, M.: @spam: The underground on 140 characters or less. : Proceedings of the 17th ACM conference on Computer and communicat ions security, ACM (2010) 27–37
[11] Wang, A.H.: Don't follow me: Spam det ection in twitter. In: Securit y and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, IEEE (2010) 1–10
[12] Bratko, A., Filipiˇc, B., Cormack, G., Lynam, T ., Zupan, B.: Spam filtering using stat istical data compression models. The Journal of Machine Learning Research 7 (2006) 2673–2698
[13] Detecting Spam Bots in Online Social Networking Sites- A Machine Learning Approach: Alex Hai Wang. DBSec'10 Proceedings of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy.