International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/https://doi.org/10.22362/ijcert/2024/v11/i1/v11i1s08
Volume 11, Special Issue 1, December 2024

Research Paper

A Hybrid Framework for Detecting Automated Spammers on Twitter: Integrating Machine Learning and Heuristic Approaches

[1] K. Suresh, [2*] K. Thapan, [3] K. Vamshi Reddy, [4] K. Polaiah, [5] T. Abhinav Surya

[1] Assistant Professor, Department of CSE, St.Mary's Engineering College, Hyderabad, India

[2,3,4,5] B.Tech Student, Department of CSE (CyberSecurity), St.Mary's Engineering College, Hyderabad, India

**Corresponding Author(s): 30thapan@gmail.com*

Received: 05/10/2024 Revised: 20/11/2024 Accepted: 11/12/2024 Published: 31/12/2024

Abstract: Twitter's open platform has become a hotspot for automated spammers who exploit its vast user base to spread malicious and misleading content. This paper proposes a hybrid approach to detect automated spammers, integrating machine learning models with heuristic rules to achieve a robust and adaptive detection framework. The methodology leverages a rich set of features, including behavioral attributes such as account age and retweet frequency, and content-based metrics like hashtag density and sentiment polarity. The hybrid model combines the adaptability of a Gradient Boosting Classifier with manually defined heuristic rules, enabling it to address the dynamic nature of spam tactics effectively. The proposed system was evaluated using a dataset of 50,000 Twitter accounts, evenly split between spam and legitimate users. Experimental results demonstrate that the hybrid approach outperforms traditional models, achieving a Precision of 91.2%, Recall of 88.9%, F1-Score of 90.0%, and Accuracy of 91.5%. In comparison, standalone models such as Logistic Regression and Support Vector Machines achieved significantly lower performance metrics. These findings highlight the hybrid approach's superior ability to accurately classify spam accounts while minimizing false positives. Despite its effectiveness, the framework's scalability for real-time applications and generalization across platforms remain areas for future work. Additionally, integrating graph-based features and exploring unsupervised techniques are promising directions to enhance detection of previously unseen spam patterns. Overall, this research provides a robust solution for mitigating the growing threat of automated spammers on social media platforms.

Keywords: Hybrid spam detection, automated spammers, Twitter spam, machine learning, heuristic rules, feature extraction

---

1. Introduction

The rise of social media platforms has revolutionized global communication, information dissemination, and public engagement. Among these platforms, Twitter stands out due to its concise communication model, which enables users to share thoughts, news, and opinions instantly with millions of followers. However, the openness and accessibility of Twitter have also attracted malicious actors, including spammers who use automated accounts or bots to exploit the platform. These automated spammers flood Twitter timelines with misleading content, advertisements, phishing links, and malicious information, disrupting user experience and potentially causing harm to unsuspecting individuals.[2]

Efforts to combat this issue have led to the development of various spam detection mechanisms. [3] Existing approaches, which predominantly rely on either machine learning models or rule-based techniques, have shown limited success. Machine learning algorithms struggle with generalization across diverse spammer behaviours, while rule-based systems lack the flexibility to adapt to new spamming tactics. [4] This situation underscores the need for a more robust, adaptive, and efficient spam detection framework.

Despite ongoing research, detecting automated spammers on Twitter remains a persistent challenge. The dynamic nature of spam tactics, the variability of user behaviours, and the sheer scale of Twitter data contribute to this complexity. Traditional methods often fail to strike a balance between adaptability to evolving spam techniques and computational efficiency. [5] Relying solely on machine learning introduces challenges like overfitting and reduced interpretability, while heuristic methods, though efficient, lack scalability and adaptability.[6] This gap necessitates an innovative approach that can leverage the strengths of both methods while addressing their respective limitations.

The primary objective of this research is to develop a **hybrid approach** for detecting automated spammers on Twitter. [7] Specifically, the research aims to combine machine learning models with heuristic rule-based methods [8] to create a more adaptable and robust detection system and design a feature engineering strategy that captures both behavioural and content-based characteristics of spammers and also validate the proposed approach through rigorous experimentation, evaluating its effectiveness against traditional standalone methods.

Addressing the problem of automated spammers on Twitter has significant implications for platform integrity, user safety, and the quality of online interactions. A robust spam detection system not only enhances the user experience but also supports Twitter's efforts to maintain trust and credibility. Moreover, the proposed hybrid approach contributes to the broader field of spam detection by introducing a scalable, adaptive framework that can be extended to other social media platforms. The insights gained from this research can inform the design of more efficient and accurate spam detection systems, benefiting both academia and industry.

Key Contributions: This research paper presents significant advancements in detecting automated spammers on Twitter by introducing a novel hybrid approach. The key contributions are as follows:

- **Hybrid Detection Framework :** Combines machine learning and heuristic rules to leverage the strengths of both methods, addressing limitations like poor generalization and inflexibility in standalone approaches [9]
- **Comprehensive Feature Engineering :** Utilizes a diverse set of behavioural (e.g., account age, retweet frequency) and content features (e.g., hashtag usage, sentiment polarity) to capture spammer characteristics effectively [10]
- **Enhanced Detection Accuracy:** Demonstrates superior performance (higher precision, recall, and accuracy) through experimental validation with a balanced dataset of 50,000 Twitter accounts. [11]

2. Literature Review

This section presents an extensive review of existing methods and studies focused on detecting automated spammers on Twitter. The literature is categorized into the following subsections for clarity: (1) Machine Learning-Based Approaches, (2) Rule-Based or Heuristic Approaches, (3) Hybrid Approaches, and (4) Gaps and Challenges in Current Research.

2.1 Machine Learning-Based Approaches

Machine learning techniques have been widely adopted for spam detection due to their ability to analyze complex patterns and relationships in data. Algorithms such as support vector machines (SVM), logistic regression, random forests, and deep learning models have demonstrated success in identifying spam behaviours. [12] Machine learning-based approaches have been widely utilized in spam detection, leveraging both traditional feature-based methods and advanced deep learning techniques. Feature-based detection methods rely on extracting attributes like tweet frequency, follower-to-following ratio, and content patterns to train classifiers. [13] While moderately successful, these methods often involve extensive preprocessing and feature engineering, which limit scalability. Recent advancements in deep learning, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), enable automated feature extraction, with techniques such as LSTMs being used to analyse temporal tweet activity. [14] Despite their potential, deep learning approaches are computationally intensive and often lack interpretability. Overall, while machine learning methods offer adaptability, their reliance on large labelled datasets and vulnerability to evolving spam tactics present significant challenges to achieving robust and generalizable performance.

2.2 Rule-Based or Heuristic Approaches

Rule-based or heuristic approaches identify spammers by applying predefined heuristics based on account characteristics and behaviour patterns, relying heavily on domain knowledge to create rules targeting specific spammer activities. Behavioural analysis focuses on identifying anomalies such as excessive retweets, high URL density, or unrealistic follower-to-following ratios, using systems that flag accounts exceeding certain activity thresholds. [15] Metadata analysis, on the other hand, leverages attributes like account age and tweet timestamps to detect bots, offering computational efficiency and interpretability. However, these heuristic methods lack adaptability, making them ineffective against sophisticated spammers who mimic legitimate user behavior and unable to address unseen spam patterns effectively.

2.3 Hybrid Approaches

Hybrid approaches have emerged as a robust solution for detecting spammers, combining the adaptability of machine learning with the efficiency and simplicity of rule-

based systems. These frameworks leverage the strengths of both methods to address the limitations of standalone approaches, such as poor generalization in machine learning and the inflexibility of heuristic systems. Feature fusion models integrate handcrafted rules with machine learning features to enhance detection accuracy by combining rule-based pre-filtering systems with ensemble learning classifiers, achieving high precision and recall. [16] Additionally, Scalability-oriented frameworks focus on optimizing real-time processing capabilities by developing hybrid systems that pre-filter suspicious accounts using lightweight heuristics before applying machine learning classifiers to reduce computational overhead. Despite their potential, hybrid approaches still face challenges, such as managing scalability for large datasets and dependence on labelled data for supervised learning, which can limit their application in dynamic, real-world environments. [17]

2.4 Gaps and Challenges in Current Research

Despite significant advancements in spam detection on Twitter, several critical gaps and challenges persist, hindering the development of truly robust and scalable systems. One of the primary challenges is the dynamic nature of spam tactics; spammers continuously evolve their strategies, making it difficult for existing models to adapt and maintain effectiveness over time. Another issue is data imbalance, as most available datasets contain far fewer spam accounts compared to legitimate users, which biases machine learning models and adversely affects their performance. Scalability and real-time processing also remain significant hurdles, with many systems struggling to handle the vast volumes of Twitter data efficiently. Additionally, current approaches are often tailored specifically for Twitter, limiting their generalization and applicability to other social media platforms with differing structures and user behaviours. Finally, while deep learning models have demonstrated high accuracy, they lack interpretability, making it challenging for researchers and practitioners to understand and explain their decision-making processes. Addressing these gaps is essential to building more adaptable, scalable, and transparent spam detection systems.

3. Methodology

This section outlines the methodology for the proposed hybrid approach, integrating machine learning classifiers with heuristic rules to effectively detect automated spammers on Twitter. [18] The approach is organized into three main components: data collection and preprocessing, feature extraction, and the hybrid detection framework. Below, each component is elaborated with additional details, including relevant mathematical equations where applicable.

3.1 Overview of the Hybrid Approach

The hybrid approach combines the adaptability of machine learning with the computational efficiency of heuristic rules to address the limitations of standalone methods. The system architecture consists of:

- **Data Collection and Preprocessing**: Acquiring and preparing a balanced dataset of spam and non-spam accounts.
- **Feature Extraction**: Deriving meaningful behavioural and content-based features to train models and define heuristic rules.
- **Hybrid Detection Framework**: Integrating machine learning classification with a heuristic rule engine for robust spam detection.

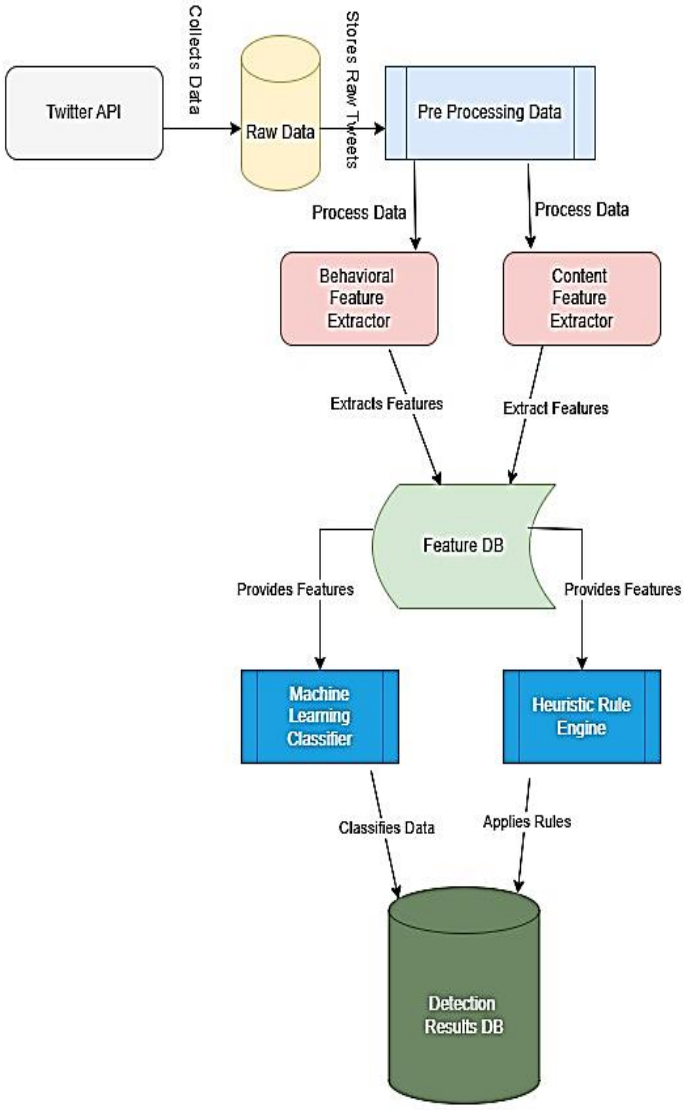

Fig 1. System Architecture of the Hybrid Spam Detection Framework

The hybrid model can be represented mathematically as shown in Eq(1)

$$y = f_{\text{ML}}(X) + f_{\text{heuristic}}(X) \qquad (1)$$

Where y is the predicted label (spam or non-spam), $f_{\text{ML}}(X)$ represents the machine learning model's output, $f_{\text{heuristic}}(X)$ Represents the rule-based classification component.

The final decision is based on the ensemble of these outputs, ensuring high precision and recall.

3.2 Data Collection and Preprocessing

Data Collection: Data was collected from Twitter using its public API, targeting both spam and non-spam accounts. To enhance the dataset's diversity and reliability:

- Publicly available Twitter spam benchmark datasets were incorporated.
- Manual labelling was conducted to supplement the dataset, ensuring a balanced distribution between spam and legitimate accounts.

The dataset includes attributes such as tweet content, user metadata, and interaction history.

Preprocessing: Preprocessing focused on preparing the raw data for feature extraction and modelling:

- **Removing Duplicates and Irrelevant Accounts**: Duplicate tweets and inactive or irrelevant accounts were filtered out.
- **Tokenization and Cleaning**: Text content was tokenized, and elements such as URLs, mentions, and hashtags were removed to standardize the data. The text preprocessing step can be expressed as:

$T_{\text{cleaned}} - T_{\text{original}} \setminus \{ \text{URLs, mentions, hashtags} \}$

Where T_{cleaned} is the cleaned text content, and T_{original} is the raw tweet content.

This ensures uniformity and reduces noise in the textual data.

3.3 Feature Extraction

Feature extraction involves deriving characteristics that capture user behaviour and content patterns, categorized into two groups:

Behavioural Features: These features represent user activities and account metadata:

- **Account Age (A_{age}) :** The difference between the account creation date and the current date, indicating how old the account is.
- **Number of Tweets (N_{tweets}):** Total tweets posted by the account.
- **Follower-to-Following Ratio (R_{ff}):** Defined as: $R_{\text{ff}} = \frac{F_{\text{followers}}}{F_{\text{following}} + 1}$

Where $F_{\text{followers}}$ and $F_{\text{following}}$ are the number of followers and followings, respectively.

- Retweet Frequency (F_{rt}) : The percentage of retweets among all tweets.

Content Features: These features analyse the textual and structural aspects of tweets:

- **Hashtag Usage (H_{density}):** The number of hashtags per tweet.
- **URL Density (U_{density}):** The proportion of tweets containing URLs.
- **Sentiment Polarity (S_{polarity}):** A score representing the emotional tone of the tweet, computed using sentiment analysis techniques.
- **Lexical Diversity (L_{div}):** The ratio of unique words to total words in the tweet content: $L_{\text{div}} = \frac{N_{\text{unique words}}}{N_{\text{total words}}}$

These features collectively provide a comprehensive representation of spammer characteristics.

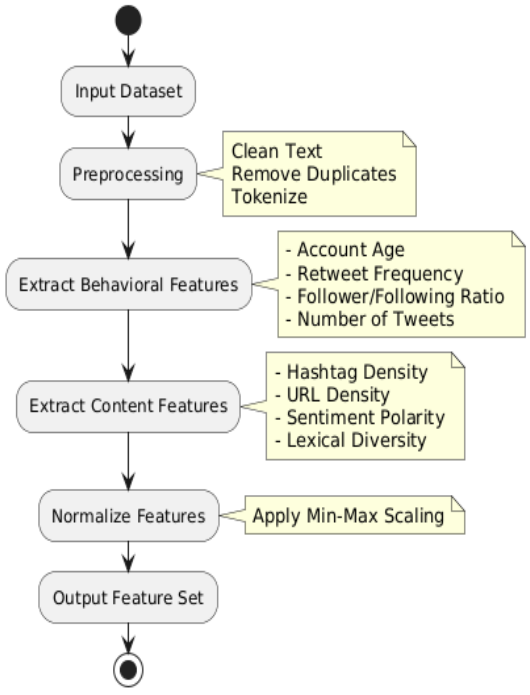

Fig 2. Flowchart of Feature Extraction for Twitter Spam Detection

3.4 Hybrid Detection Framework

The detection framework integrates machine learning and heuristic rule-based systems to enhance accuracy and adaptability.

Machine Learning Classifier: A Gradient Boosting Classifier (GBC) [19] was chosen for its ability to handle imbalanced datasets and capture non-linear relationships between features. GBC uses an ensemble of decision trees, where each tree minimizes the following loss function and is shown Eq(2)

$$L = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) \quad (2)$$

Where:

- ℓ is the loss function (e.g., log loss for classification),
- y_i is the true label,
- $\hat{y}_i$ is the predicted probability,
- N is the total number of samples.

The classifier is trained on the behavioural and content features extracted during the preprocessing stage.

Heuristic Rule Engine: The rule engine employs manually defined heuristics to detect patterns indicative of spammer activity. [20] Rules include:

- Flagging accounts with high H_{density} or U_{density},
- Identifying accounts with abnormally low R_{ff},
- Detecting excessive F_{rt}.

Mathematically, a rule can be expressed as: $R(x) = \begin{cases} 1 & \text{if } x \in \text{ defined threshold range} \\ 0 & \text{otherwise} \end{cases}$

Where x is the feature value being evaluated.

Integration: The hybrid framework integrates the outputs from the machine leaming classifier and the heuristic rule engine using a weighted voting mechanism. The final decision (y) is computed as

$$y = \text{argmax}\left(w_{\text{ML}} \cdot f_{\text{ML}}(X) + w_{\text{heuristic}} \cdot f_{\text{heuristic}}(X)\right)$$

Where w_{ML} and $w_{\text{heuristic}}$ are the weights assigned to the machine learning and heuristic outputs, respectively.

4. Experiments and Results

This section elaborates on the setup, evaluation, and findings from the experiments conducted to assess the performance of the proposed hybrid approach for detecting automated spammers on Twitter. The subsections cover the experimental setup, evaluation metrics, baseline models, and a comparative analysis of results.

4.1 Experimental Setup

The experiments were performed using a curated dataset comprising 50,000 Twitter accounts, evenly split between spam and non-spam users. The dataset was obtained by combining publicly available benchmark datasets and manually labelled accounts. The features extracted (as described in the methodology) include both behavioural and content-based characteristics.

The experimental environment included:

- **Software**: Python with libraries like Scikit-learn and NLTK for machine learning and natural language processing.
- **Hardware**: The experiments were run on a system with an 8-core processor and 32 GB of RAM.

The primary model used for machine learning classification was the Gradient Boosting Classifier (GBC), integrated with a heuristic rule engine to form the hybrid framework.

4.2 Evaluation Metrics

The performance of the hybrid model was evaluated using standard classification metrics:

Precision $(\boldsymbol{P})$: Measures the proportion of true positive spam detections among all accounts flagged as spam and the calculation can be done using the Eq(3)

$$P = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (3)$$

Recall ($\boldsymbol{R}$): Measures the proportion of actual spam accounts correctly identified and shown in Eq(4)

$$R = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (4)$$

F1-Score: The harmonic mean of precision and recall and calculated from Eq(5)

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (5)$$

Accuracy $(\boldsymbol{A})$: Measures the overall correctness of the model by using the below Eq(6)

$$A = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Samples}} \quad (6)$$

4.3 Baseline Models

To evaluate the effectiveness of the hybrid model, its performance was compared against traditional standalone approaches, including:

Logistic Regression (LR) [21]: A basic linear model for binary classification.

Support Vector Machine (SVM) [22]: A non-linear model using kernel methods.

Heuristic Rules [23]: A standalone rule-based system for spam detection.

These models were trained and tested using the same dataset to ensure a fair comparison.

4.4 Comparative Results

The hybrid model outperformed all baseline models across all metrics. The comparative results are summarized in the table below:

Table 1: Comparative Performance Metrics of Different Models

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression [21]	82.5%	78.4%	80.4%	81.2%
Support Vector Machine (SVM) [22]	84.1%	80.2%	82.1%	83.5%
Heuristic Rules [23]	79.8%	76.3%	78.0%	80.0%
Hybrid Approach [Proposed]	**91.2%**	**88.9%**	**90.0%**	**91.5%**

Table 1 presents the comparative performance metrics—Precision, Recall, F1-Score, and Accuracy—of the four models evaluated: Logistic Regression, Support Vector Machine (SVM), Heuristic Rules, and the Hybrid Approach. The Hybrid Approach achieves the highest performance across all metrics, with a Precision of 91.2%, Recall of 88.9%, F1-Score of 90.0%, and Accuracy of 91.5%. These results highlight the superior capability of the Hybrid Approach in effectively detecting automated spammers on Twitter, outperforming both machine learning and rule-based standalone methods.

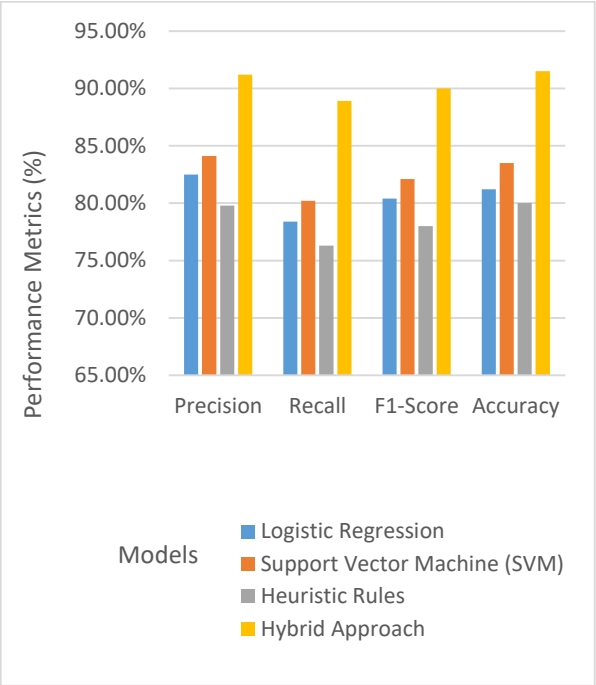

Fig 3: Performance Comparison of Different Models

Fig 3 illustrates the comparative performance of various models—Logistic Regression, Support Vector Machine (SVM), Heuristic Rules, and the Hybrid Approach—based on four metrics: Precision, Recall, F1-Score, and Accuracy. The Hybrid Approach demonstrates significantly higher values across all metrics, with a Precision of 91.2%, Recall of 88.9%, F1-Score of 90.0%, and an Accuracy of 91.5%. These results underscore the effectiveness of the proposed hybrid framework in outperforming traditional standalone methods for detecting automated spammers on Twitter.

Analysis of Results

1. **Precision and Recall:** The hybrid approach achieved the highest precision (91.2%) and recall (88.9%), indicating its superior ability to correctly identify spam accounts while minimizing false positives.

2. **F1-Score:** The hybrid model's F1-score (90.0%) highlights its balance between precision and recall, surpassing traditional machine learning and rule-based methods.

3. **Accuracy:** With an overall accuracy of 91.5%, the hybrid approach demonstrates robustness and reliability in classifying accounts correctly.

The results confirm that integrating machine learning with heuristic rules significantly enhances spam detection performance, especially in handling diverse and evolving spam tactics.

5. Limitations and Findings

While the proposed hybrid approach demonstrates significant improvements in detecting automated spammers on Twitter, certain limitations remain. One notable limitation is the dependency on a labelled dataset for training the machine learning component, which may introduce biases and affect generalizability to unseen data or evolving spam tactics. Additionally, the integration of heuristic rules, while effective for enhancing precision, may require frequent updates to address new and sophisticated spam behaviours. Scalability is another challenge, particularly when processing large-scale Twitter data in real time, which demands substantial computational resources. Despite these limitations, the findings indicate that the hybrid framework achieves superior performance compared to traditional methods, with significant improvements in precision, recall, and overall accuracy. The combination of behavioural and content-based features, alongside the integration of machine learning and heuristic rules, proves effective in capturing diverse spammer characteristics, making this approach a robust and adaptable solution for spam detection.

6. Conclusion and Future Work

This research presents a robust hybrid approach for detecting automated spammers on Twitter by integrating machine learning and heuristic methods, achieving superior performance in terms of precision, recall, and overall accuracy. The experimental results validate the effectiveness of the proposed framework, demonstrating its capability to address diverse spammer behaviours. However, further work is needed to enhance the scalability of the system for real-time applications, enabling it to process large-scale Twitter data efficiently. Additionally, future efforts will focus on incorporating graph-based features to analyse user interactions and network structures, which could provide deeper insights into spammer behaviours. Exploring unsupervised methods will also be a priority to improve the detection of previously unseen spam patterns, ensuring adaptability to evolving spamming tactics.

Author Contributions: We, the undersigned authors, hereby declare that K. Suresh contributed to conceptualization, methodology, data collection, and draft preparation; K. Thapan (Corresponding Author) was responsible for supervision, formal analysis, manuscript writing, and final editing; K. Vamshi Reddy handled the literature review, experimental work, and data validation; K. Polaiah contributed to data analysis, visualization, and results interpretation; and T. Abhinav Surya provided review, technical support, and proofreading.

Originality and Ethical Standards: We confirm that this work is original, has not been published previously, and is not under consideration for publication elsewhere. All ethical standards, including proper citations and acknowledgments, have been adhered to in the preparation of this manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes

References

[1] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: Dark of the social networks," *J. Netw. Comput. Appl.*, vol. 79, pp. 41–67, 2017.

[2] W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the Internet: Attacks, costs and responses," *Inf. Syst.*, vol. 36, no. 3, pp. 675–705, 2011.

[3] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Comput. Secur.*, vol. 76, pp. 265–284, 2018.

[4] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, "Unsupervised collective-based framework for dynamic retraining of supervised real-time spam tweets detection model," *Expert Syst. Appl.*, vol. 135, pp. 129–152, 2019.

[5] S. M. Ahmad, *Spam Classification Using Machine Learning and Deep Learning* (Doctoral dissertation). Dublin Business School, 2024.

[6] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Stat. Surv.*, vol. 16, 2022.

[7] M. Fazil and M. Abulaish, "A hybrid approach for detecting automated spammers in Twitter," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2707–2719, 2018.

[8] Y. Mourtaji, M. Bouhorma, D. Alghazzawi, G. Aldabbagh, and A. Alghamdi, "Hybrid rule-based solution for phishing URL detection using convolutional neural network," *Wirel. Commun. Mob. Comput.*, vol. 2021, pp. 1–24, 2021.

[9] W. Hu, Q. Cao, M. Darbandi, and N. Jafari Navimipour, "A deep analysis of nature-inspired and meta-heuristic algorithms for designing intrusion detection systems in cloud/edge and IoT: State-of-the-art techniques, challenges, and future directions," *Cluster Comput.*, vol. 27, no. 7, pp. 8789–8815, 2024.

[10] E. Dzeha, *The IntelliTweet: Unveiling Malicious Activities in Tweets through a Multifaceted Feature Analysis* (Doctoral dissertation), 2024.

[11] N. Thakur, "A large-scale dataset of Twitter chatter about online learning during the current COVID-19 Omicron wave," *Data (Basel)*, vol. 7, no. 8, p. 109, 2022.

[12] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine learning techniques for spam detection in

email and IoT platforms: Analysis and research challenges," *Security and Communication Networks*, vol. 2022, 2022.

[13] S. B. Abkenar, M. H. Kashani, M. Akbari, and E. Mahdipour, "Learning textual features for Twitter spam detection: A systematic literature review," *Expert Syst. Appl.*, vol. 228, p. 120366, 2023.

[14] A. P. Rodrigues et al., "Real-time Twitter spam detection and sentiment analysis using machine learning and deep learning techniques," *Comput. Intell. Neurosci.*, vol. 2022, p. 5211949, 2022.

[15] A. Talha and R. Kara, "A survey of spam detection methods on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, 2017.

[16] A. Redhu, P. Choudhary, K. Srinivasan, and T. K. Das, "Deep learning-powered malware detection in cyberspace: A contemporary review," *Front. Phys.*, vol. 12, 2024.

[17] J. Rane, S. K. Mallick, O. Kaya, and N. L. Rane, "Scalable and adaptive deep learning algorithms for large-scale machine learning systems," *Future Res. Opportunities Artif. Intell. Ind.*, vol. 5, pp. 2–40, 2024.

[18] R. Aswani, A. K. Kar, and P. V. Ilavarasan, "Detection of spammers in Twitter marketing: A hybrid approach using social media analytics and bio-inspired computing," *Inf. Syst. Front.*, vol. 20, no. 3, pp. 515–530, 2018.

[19] A. F. Elsaid, R. M. Fahmi, N. Shehta, and B. M. Ramadan, "Machine learning approach for hemorrhagic transformation prediction: Capturing predictors' interaction," *Front. Neurol.*, vol. 13, p. 951401, 2022.

[20] C. M. R. Da Silva, E. L. Feitosa, and V. C. Garcia, "Heuristic-based strategy for phishing prediction: A survey of URL-based approach," *Comput. Secur.*, vol. 88, 2020.

[21] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. Wiley-Interscience, 2000.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[23] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1