Int. j. inf. tecnol. (August 2023) 15(6):3103–3111
https://doi.org/10.1007/s41870-023-01352-1

ORIGINAL RESEARCH

TwitterGAN: robust spam detection in twitter using novel generative adversarial networks

Mohammad Diqi[1]

Received: 5 November 2022 / Accepted: 13 June 2023 / Published online: 23 June 2023

Abstract As social media platforms like Twitter continue to evolve, the proliferation of spam content has become a pressing issue, undermining the credibility of shared messages. Traditional spam detection methods, such as black-and-white listing and rule-based learning techniques, struggle to efficiently handle large datasets and adapt to dynamic environments. To address these challenges, we propose a novel spam detection model that leverages generative learning techniques, offering improved performance on vast datasets and changing circumstances. Using a substantial Twitter dataset with an 80% training and 20% testing split, our innovative model demonstrates remarkable effectiveness. Experimental results show a G-Loss score of 8.1207, significantly outperforming the D-Loss score of 0.0081, indicating the model's exceptional accuracy and low error rate. Consequently, our groundbreaking approach emerges as a highly promising solution for real-world spam identification, raising the bar for spam detection research.

Keywords Twitter · Spam detection · Generative learning · RunGAN · Evaluation metrics

1 Introduction

Over the last few years, online social networks (OSNs) have assumed a central role in global connectivity and the sharing of ideas, with industry behemoths like Facebook and Twitter exerting significant influence in worldwide networking [1]. These platforms' staggering popularity has connected an astonishing 2.46 billion individuals, with projections indicating that one-third of the world's population will be online by 2020 [2]. Twitter, in particular, has experienced an influx of over 42 million new accounts created each month, rendering it a primary target for spammers [3]. It has been estimated that one in every 200 social media messages contains spam, with automated bots or programs executing specific tasks accounting for 15% of active Twitter users [4].

The explosive growth of Twitter has led to a concomitant surge in spam content, posing considerable challenges to both individual and corporate users [5]. As the platform's popularity skyrockets, spammers have become increasingly adept at creating nefarious messages and penetrating legitimate conversations. Studies indicate that spam tweets are twice as prevalent as spam emails, heightening their menace to users [6]. To tackle this problem, Twitter has instituted several mechanisms, including allowing users to flag spam accounts and banning them once confirmed [7].

However, the struggle against spam on Twitter and other social media platforms remains a relentless fight, as spammers continuously alter their tactics and generate counterfeit accounts to evade detection [8, 9]. Multiple studies have endeavored to devise approaches for detecting and categorizing spam messages, employing diverse machine learning algorithms and classification methods [10]. Notwithstanding these endeavors, the obstacles presented by spam on social media platforms endure, underscoring the necessity for more sophisticated spam detection techniques.

In this investigation, we present an innovative approach to address spam identification obstacles by leveraging a novel generative learning model architecture, dubbed TwitterGAN, which utilizes deep learning techniques to identify spam on Twitter [4, 11]. Our research makes several crucial contributions to the field of spam detection:

✉ Mohammad Diqi
diqi@respati.ac.id

[1] Universitas Respati Yogyakarta, Yogyakarta, Indonesia

1. We devise an effective detection model using a cutting-edge Generative Adversarial Network (GAN) architecture and a new activation function, demonstrating superior accuracy and reduced loss in comparison to conventional machine learning techniques.
2. Through comprehensive analysis, we achieve state-of-the-art outcomes in identifying fake profiles and introduce an evaluation metric to verify the model's caliber.
3. We propose a new function for selecting the optimal activated value in the discriminator block of GAN, augmenting the GAN's efficacy in spam detection assignments.

The study is structured as follows: In Part I, we provide the context for the research. In Part II, we review the relevant literature. Part III defines the research problem, and Part IV outlines the experimental design, including feature learning, data collection, and data processing. Part V presents the results and conducts a comprehensive analysis. Finally, Section VI concludes by summarizing the findings and discussing remaining issues related to Twitter spam classification.

Through our groundbreaking approach, we aim to make a significant contribution to the field of spam detection in social media and address the growing challenges posed by spam on platforms such as Twitter. As social media platforms continue to shape our world, it is crucial to develop robust and adaptable models capable of effectively combating the dynamic and evolving nature of spam.

2 Related work

Although traditional spam detection methods, like blocklists and permitted listings, have been somewhat effective in reducing social media spam, their efficacy is becoming increasingly limited as spamming techniques evolve and spam content grows [12]. Blocklists evaluate whether a Twitter account has sent spammy tweets and are useful in filtering known spam accounts, but they struggle to adapt to new threats or identify more sophisticated spammers [8]. Permitted listings or apps can help prevent unauthorized software execution, but they may not sufficiently address spam that exploits existing platform features or imitates legitimate user behavior [13]. Consequently, researchers have started to explore machine learning and deep learning approaches to supplement and enhance traditional spam detection techniques, addressing the ever-changing landscape of social media spamming [14].

In light of the evolving nature of spam on social media platforms, traditional spam detection methods face numerous significant challenges and limitations. The rapid growth of spam content poses a problem for manual detection techniques and static blocklists, making it difficult for them to keep up with the volume of new threats [15]. Furthermore, spammers continuously change their tactics and employ more sophisticated techniques to evade detection by using stealthy, low-volume spamming sites [16]. Traditional methods may also suffer from false positives, blocking legitimate accounts accidentally, and false negatives, allowing spam accounts to go undetected [17]. Additionally, as social media platforms become more interconnected, spammers can exploit cross-platform vulnerabilities, making platform-specific detection methods less effective [7]. Consequently, advanced machine learning and deep learning techniques are increasingly being utilized by researchers to overcome these challenges and enhance spam detection on social media platforms [18].

Advanced machine learning-based techniques offer promising avenues for more effective and efficient spam detection on social media platforms like Twitter. Researchers can further optimize these techniques by exploring novel feature engineering methods, such as leveraging user profile information, message content features, or graph-based features to enhance detection accuracy [19]. Combining multiple classifiers into an ensemble model can also improve overall performance [10]. Deep learning algorithms have demonstrated promising results in spam detection [20], and adopting a continuous learning framework that adapts to the evolving spam landscape will ensure that spam detection models remain effective in detecting new spam tactics and patterns [21]. By advancing machine learning-based approaches, researchers can make significant strides in combating spam on social media platforms.

To develop machine learning-based spam detection models for social media platforms, it is essential to consider various critical features. User profile features, such as account creation date, number of followers, and profile description, can help identify suspicious accounts or profiles that are more likely to engage in spamming activities [22]. Message content features, including the use of specific keywords, URLs, or hashtags, can offer valuable insights into the nature of spam messages and help classify them more accurately [16]. Graph-based features, which examine relationships between users on the platform, can provide a broader context for detecting spam activity by analyzing the connections and interactions within social networks [22]. Additionally, features related to embedded URLs, such as domain reputation and redirection patterns, can help pinpoint spam messages that aim to lure users to malicious websites or promote scams [23]. By carefully selecting and incorporating these features into machine learning-based spam detection models, researchers can significantly improve the accuracy and effectiveness of spam detection on social media platforms.

To enhance the accuracy and adaptability of spam detection models, deep learning techniques such as Convolutional Neural Networks (CNN) [24] and Long Short-Term

Memory (LSTM) networks can be effectively integrated [14]. A method involves using CNNs to extract relevant features automatically from raw data, such as user profile information or message content, thereby reducing the need for manual feature engineering and enhancing the model's spam detection ability [11, 25]. LSTMs can be employed to analyze temporal patterns and dependencies in user-generated content, enabling the model to better understand the context and detect spam in situations where traditional methods may struggle [26]. Combining these deep learning techniques, such as CNN-LSTM, can further improve spam detection by capturing both spatial and temporal features, providing a more comprehensive understanding of the data [27]. Integrating these deep learning techniques into spam detection models can help researchers develop more accurate and adaptable solutions that effectively address the evolving challenges of spam detection on social media platforms.

In the realm of social media, spam detection can benefit from the unique capabilities of Generative Adversarial Networks (GANs) [28, 29]. With their capacity to generate synthetic data that mimics real-world samples, GANs offer a promising approach for scenarios where labeled data is scarce or costly to obtain [30]. Moreover, GANs can enhance the resilience of spam detection models by leveraging adversarial training, whereby the generator and discriminator components compete to generate and detect spam, respectively [28]. Through this adversarial process, the model is prompted to learn more distinctive features, thereby increasing its ability to generalize to previously unseen spam instances [31]. By utilizing GANs for spam detection in social media, researchers could potentially address certain limitations of conventional machine learning and deep learning methods, such as the need for extensive labeled datasets and susceptibility to adversarial attacks, resulting in more potent and adaptive spam detection models.

Despite the promise of GANs for spam detection, there are several challenges and limitations to their application. Firstly, GANs necessitate a careful balance between the generator and discriminator networks during training, which can lead to instability and convergence issues, consequently compromising the accuracy of the spam detection model [32]. To circumvent this concern, researchers can investigate alternative training strategies, like curriculum learning, which gradually increases the complexity of the training process [21]. Secondly, there is a risk of GAN-generated synthetic data unintentionally containing sensitive information during spam detection [33]. To minimize this risk, researchers can integrate differential privacy methods into the GAN training process [34]. Thirdly, evaluating GAN-generated synthetic data for spam detection is challenging, as the lack of ground truth poses an obstacle in comparing the generated samples [35]. One possible solution for this is for researchers to utilize different evaluation metrics, such as the Inception Score, to evaluate the generated data's diversity and quality [36]. Finally, the computational expense of GANs is a limitation, as significant computational resources are required for training [37]. To overcome this, researchers can explore more efficient GAN architectures or leverage transfer learning techniques to reduce the training time required [32]. Addressing these challenges would enable GANs to become a more feasible and effective solution for spam detection on social media platforms.

To enhance spam detection performance on social media platforms, anomaly detection and clustering techniques can be used in conjunction with other machine learning and deep learning approaches. Clustering techniques like K-means or DBSCAN can serve as a preprocessing step to group similar users or messages together, simplifying the problem and providing a structured input for machine learning classifiers such as SVM or Naive Bayes [38]. By focusing on distinct patterns within each cluster, classifiers' performance can be improved. Anomaly detection methods, such as Isolation Forest, can be combined with deep learning techniques like CNN to identify unusual patterns in feature spaces that may indicate spam activity [5]. This hybrid approach can yield more accurate and robust spam detection models. Furthermore, clustering and anomaly detection techniques can be incorporated into ensemble learning methods, which amalgamate multiple classifiers' predictions to achieve better overall performance [39]. Leveraging the strengths of various methods, ensemble learning can enhance spam detection accuracy and adaptability. Lastly, incorporating clustering and anomaly detection techniques in unsupervised or semi-supervised learning settings can help leverage unlabeled data to improve spam detection models [40], which can be especially beneficial in scenarios where labeled data is scarce or expensive to obtain. By integrating these techniques, researchers can develop more advanced and effective spam detection models for social media platforms.

Time series analysis and natural language processing (NLP) techniques can significantly enhance the development of advanced spam detection models for social media platforms such as Twitter. Firstly, time series analysis enables researchers to identify patterns in user activity such as the frequency and timing of posts that can serve as important features for distinguishing between genuine users and spammers [31]. By analyzing tweet production time series, it is possible to more effectively detect spammer bots and reduce their prevalence on social media platforms. Secondly, NLP techniques, including word embeddings like Word2Vec and GloVe, and sentiment analysis can extract valuable information from the textual content of tweets and messages [41]. These features can then be fed into machine learning and deep learning models to improve their accuracy in detecting spam. Thirdly, NLP-based approaches can be combined with other methods, such as graph-based features or URL analysis, to create more comprehensive and robust spam detection models [22]. By leveraging the strengths of various techniques, researchers can better address the evolving nature of spam on social media platforms. Finally, time series analysis and NLP can track and analyze spam campaigns, identify commonalities in content or temporal patterns, and develop more targeted and effective countermeasures [42]. By incorporating time series analysis and NLP techniques, researchers can build more advanced spam detection models that can better adapt to the ever-changing landscape of social media spamming.

With the continuous evolution of spamming techniques and the dynamic nature of social media platforms, it is imperative to identify promising future research directions for spam detection in social media. First, exploring novel deep learning architectures and their combination with traditional machine learning methods can lead to more accurate and adaptable spam detection models [43]. By leveraging the power of deep learning, researchers can potentially uncover complex patterns and relationships in social media data that are difficult to detect using traditional techniques. Second, investigating the use of GANs for spam detection presents a promising avenue, as GANs have shown remarkable success in various domains and can potentially outperform traditional machine learning and deep learning methods in spam detection tasks [5]. This research direction could lead to the development of more effective and robust spam detection models that can better adapt to the ever-changing landscape of social media spamming.

To enhance the performance of spam detection models, it is crucial to focus on the integration of various data sources and feature types, such as graph-based features, content analysis, and user behavior patterns [22]. By incorporating a wider range of features and data sources, researchers can develop more comprehensive models that can better adapt to the ever-changing landscape of spamming techniques. Moreover, investigating the role of time series analysis and NLP in tracking and analyzing spam campaigns can facilitate the development of more targeted and effective countermeasures against spam [42]. By comprehending the temporal patterns and linguistic characteristics of spam, researchers can create more advanced spam detection models that can aptly respond to the dynamic nature of social media platforms.

Table 1 presents a meta-analysis of the available work on which the proposed work is based.

3 Background

In this section, a formal statement of the topic as well as some of the concepts discussed in this article will be presented.

Table 1 Meta-analysis of available work

Technique	Advantages	Limitations and challenges
Blocklists and permitted listings	Useful in filtering known spam accounts and preventing unauthorized software execution [8, 13]	Struggle to adapt to new threats or sophisticated spammers; may not sufficiently address spam exploiting platform features or imitating legitimate user behavior [8, 13]
Machine learning	Can optimize feature engineering and improve overall performance with ensemble models [10, 19]	Faces challenges with the rapid growth of spam content, changing tactics, false positives/negatives, and cross-platform vulnerabilities [7, 15–17]
Deep learning (CNN, LSTM)	Can extract relevant features automatically, reduce manual feature engineering, and analyze temporal patterns [11, 26]	Requires large labeled datasets and is computationally expensive [14]
Generative adversarial networks (gans)	Can generate synthetic data for training and enhance model resilience through adversarial training [28–31]	Instability during training, risk of containing sensitive information, difficulty in evaluating generated data, and high computational cost [32–34]
Clustering & anomaly detection	Can improve classifier performance, identify unusual patterns, and leverage unlabeled data [38–40]	May require careful selection and integration of techniques
Time series analysis & NLP	Can identify patterns in user activity and extract valuable information from textual content [31, 41]	Requires careful integration with other methods and tracking of spam campaigns [22, 42]

A. Problem definitions

GANs employ a competitive strategy to create a generative model, which comprises a generator (G) and a discriminator (D). The distribution p in the real data space x is computed using the generator model G. The generator G masterfully crafts a new adversarial sample $G(z)$ from the same $X(z)$ distribution, utilizing the input interference variable p. Meanwhile, the discriminator model D confidently determines the probability $D(x)$ that a specific sample x originates from the authentic G dataset [28]. Table 2 presents the mathematical notation used in this paper.

The primary objective of the GAN model is to discern between real and fake samples using the discriminator feature D. This classification challenge necessitates an objective function with a value of $V(D, G)$. The generator G skillfully transforms the latent noise space z into input data for D. The term $[log(1\text{-}D(G(z)))]$ suggests that the sample originates from genuine data, and D aims to maximize this result. However, when D detects the sample generated by G, its output diminishes. G strives to maximize D's output while supplying D with counterfeit samples, ultimately achieving D's prowess in deceptive discrimination.

B. Proposed method: RunGAN.

GAN represents a state-of-the-art framework that harnesses the principles of zero-sum games to train two models simultaneously [6, 8]. The application of GANs has surged in popularity across numerous fields due to their versatility and efficacy [11, 12]. In our research, we utilize the GAN approach to detect Twitter spam by introducing the innovative architecture RunGAN, which encompasses both the generator G and the discriminator D. Figure 1 illustrates the sophisticated structure of RunGAN, comprising a cutting-edge generator G and a novel discriminator D.

Table 2 Mathematical notation

Notation	Description
$V(D, G)$	Value function
D	Discriminator
G	Generator
E	Denotes the expectation
P	Represent the number of samples x
$P_{data}(x)$	Input data
$P_z(z)$	Noise variables
x_i	Input vector
z_i	Noise vector
$D(x_i)$	discriminator of x_i (real) sample
$G(z_i)$	generator of z_i (fake) sample
$D(G(z_i))$	discriminator of generator output
F	Accuracy function
$F(D(x))$	Discriminator model output

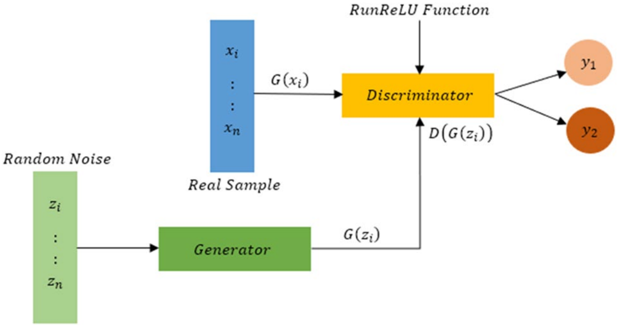

Fig. 1 RunGAN architecture with generator G and novel discriminator D

Within the RunGAN framework, the discriminator D identifies the presence of x in the input $P_{data}(x)$ and proceeds to the subsequent stage.

$$E_{x \in P_{data}(x)} log(F(D(x))) \tag{1}$$

The discriminator model's outcome, denoted as $F(D(x))$, is a real number ranging between 0 and 1 that assesses the likelihood of data correctness. By maximizing Eq. 1, the discriminator can precisely predict the normal value, whereby $F(D(x))$ equals 1 when x belongs to the set of real data $P_{data}(x)$. Thereafter, the generator's data accuracy is rigorously verified.

$$E_{Z \in P_z(z)} log(1 - F(D(G(z)))) \tag{2}$$

The optimization process adjusts three factors to ensure $F(D(G(z))) \approx 0$, making it challenging for G to produce high-quality deceptive data, as shown in Eq. 2. The data generated by the generator aims to deceive the discriminators, leading them to believe they have discovered something novel. During network training, the discriminator's objective function can be expressed as in Eq. 3.

$$\max_D E_{Z \in P_z(z)} \left[log(1 - D(G(z))) \right] + E_{x \in P_{data}(x)} \left[log(D(x)) \right] \tag{3}$$

The objective function aims to maximize the sum of two expressions below to determine a discriminator function D. Therefore, Eq. 4 illustrates the value function $V(D, G)$:

$$\min_G \max_D V(D, G) = E_{X \sim P_{data}(x)} [logD(X)] + E_{z \sim p_z(z)} \left[log(1 - D(G(z))) \right] \tag{4}$$

RunReLU: In order to develop the RunGAN architecture, we introduce a novel function called RunReLU, designed to enhance the activation function's performance in the discriminator computation, as shown in Eq. 5.

$$\text{Common}\, ReLU = max(0, x) \tag{5a}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{5b}$$

$$RunReLU = max\left[ReLU \times f(x)\right] \tag{5c}$$

$$RunReLU = max\left[ReLU \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right] \tag{5d}$$

The Gaussian Distribution, represented by $f(x)$ in Eq. 5, is based on x. To devise a fresh activation function for the discriminator classifier, we incorporate a Gaussian Distribution that ranges from 0 to 1 with the final ReLU score computation [44]. Algorithm 1 outlines the process for training the model with the innovative RunGAN architecture.

Algorithm 1: RunGAN

```
procedure TRAINING WITH RunGAN
    Input:
    input features $x_1 \ldots x_n$
    input space (vector) $z_1 \ldots z_n$
    Output:
    $RunReLU(x_i) = update\ activation\ function$
    Calculate:
    $Generator: G(z_i) = p(y \mid z_i) = \{0,1\}$
    $Discriminator: D(x_i) = p(y \mid x_i) = \{0,1\}$
    $D(G(z_i)) = D(G(z_i) * RunReLU(x_i))$
    $Loss = \frac{1}{m}\sum_{i=1}^{m} log(D(x_i)) + log(1 - D(G(z_i)))$
end procedure
```

4 Experimental setup

In this section, we outline the main idea behind our experimental setup, discuss the dataset used for our research, and detail the pre-processing steps undertaken to ensure reliable and accurate results.

A. Main idea

The primary objective of this study is to pioneer a state-of-the-art spam detection model by harnessing large datasets and generative learning techniques. A plethora of approaches, including machine learning, have been explored to augment spam detection and protection. However, the dependence on human-driven feature engineering in machine learning poses a challenge when it comes to training extensive datasets within vast, dynamic ecosystems like Twitter. This calls for the adoption of more sophisticated algorithms. Generative Adversarial Networks (GANs) have garnered recognition for their effectiveness in addressing a wide range of issues, including those related to online social network security. Therefore, we employ generative learning in our model development to attain heightened accuracy in Twitter spam detection [35].

B. Dataset

In our research, we harnessed a Twitter spam dataset from NSClab for our analysis. This dataset consists of 10,000 instances and various features such as *account_age*, *no_follower*, *no_following*, *no_userfavourites*, *no_lists*, *no_tweets*, *no_retweets*, *no_hashtag*, *no_usermention*, *no_urls*, *no_char*, and *no_digits*. Additionally, it includes a label that classifies each instance as either *spam* or *not spam*. By scrutinizing these features, we can fashion a robust model for detecting spam on Twitter. Prior to training, we partition the dataset into training and testing subsets to evaluate the performance of our learning model. We present a vast labeled post dataset that encompasses both normal and spam-labeled data, enabling a successful training and testing process. We input vector data of user profiles into the learning model using specific attributes as model inputs. In this study, we allocate 80% (8000) of the dataset for training and 20% (2000) for testing. The distribution of the dataset is elaborated in Table 3 as follows:

C. Pre-processing *data*

Our study on Twitter spam detection involved training a dataset of 10,000 samples using the GAN model to capture the typical data representation. To evaluate each test sample, the GAN model was utilized to compute a detection score. Instances with high recency ratings are considered problematic and flagged. To optimize the detection value and reduce errors, we applied normalization to the dataset by transforming actual values to interval range values [0,1] using the min–max scaler.

5 Result and analysis

In this section, we present the results of our experiments and provide a detailed analysis of the findings. We first discuss the detection test employed to evaluate the performance of our proposed method, followed by a description of the

Table 3 Wall post dataset

Dataset	Sample
Dataset training	8.000
Dataset testing	2.000
Total	10.000

evaluation metric used to measure the effectiveness of the detection algorithm.

A. Detection Test

In this study, we fine-tune numerous hyperparameters to optimize network training performance. Throughout the training and testing phases, we utilize an epoch value of 50,000 and a batch size of 512. We select a high hyperparameter value to ensure a favorable training model outcome. As the epoch value increases, the training loss diminishes, indicating an improvement in the model's performance.

To demonstrate the effectiveness of RunReLU in spam detection for large samples, we assess the performance of two generative model architectures during training and testing. In comparing GAN with RunGAN, we adjust the same hyperparameters and employ the basic generative architecture with a comparable dataset. The differences between generator and discriminator loss are detailed in Table 4.

According to the findings presented in Table 4, the RunGAN architecture demonstrates superior performance compared to the traditional GAN architecture in terms of generator and discriminator loss. Specifically, the generator loss of RunGAN (8.1207) is higher than that of GAN (6.6529), indicating that the RunGAN generator is better at generating more realistic data, which can be more challenging for the discriminator to differentiate. Additionally, the discriminator-generated loss of RunGAN (0.0081) is lower than that of GAN (0.0186), indicating that the RunGAN discriminator is more efficient in identifying generated data. Furthermore, the RunGAN architecture exhibits a higher discriminator-real loss (0.1001) compared to the GAN architecture (0.0512), indicating that the RunGAN discriminator is better at recognizing real data samples. Lastly, the RunGAN discriminator demonstrates a superior ability to distinguish between real and generated data, with a difference between real and generated losses of 0.0920, compared to 0.0326 for GAN architecture.

Overall, these results demonstrate that the RunGAN architecture shows promising potential for effectively detecting Twitter spam or other similar applications.

B. Evaluation metric

In order to evaluate the performance of the RunGAN model, we utilize the confusion matrix as an evaluation metric. The confusion matrix results for both the GAN and RunGAN models are shown in Table 5, respectively, providing a comparative analysis of their effectiveness in detecting Twitter spam. These figures measure true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP).

In the case of the GAN model, there are 748 TN, signifying that it accurately identified 748 non-spam instances. The model presents 2 FP, indicating it erroneously classified 2 non-spam instances as spam. With only 1 FN, the model misclassified a single spam instance as non-spam. Finally, there are 749 TP, revealing that it correctly identified 749 spam instances.

On the other hand, the RunGAN model demonstrates flawless classification, boasting 750 TN and 750 TP, with no FP or FN. This result suggests that the RunGAN model accurately classified all instances in the test dataset, showcasing its superior performance in Twitter spam detection compared to the GAN model.

In conclusion, the RunGAN model, featuring its innovative architecture, outperforms the standard GAN model in detecting Twitter spam. This heightened performance is evidenced by the impeccable classification results obtained in the confusion matrix, underlining the potential of RunGAN as a valuable asset for tackling spam on Twitter.

6 Conclusion

Twitter serves as a prominent OSN for sharing concise text, images, and videos among individuals and organizations. However, with Twitter's growth comes a surge in irrelevant information. A survey reveals that one in every 200 social media messages contains spam. Moreover, traditional spam detection algorithms struggle to identify spam attacks with unforeseen patterns. Thus, conventional learning architectures face challenges in detecting spam within Twitter posts.

This study introduces RunGAN, a novel generative model designed to establish a learning strategy for detecting spam on Twitter. For our experiment, we separated the training data utilized for model development from the testing examples employed to assess the model's performance. Based on

Table 4 Generator and discriminator loss

Architecture	G Loss	D_Gen Loss	D_Real Loss	D_Real – D_Gen
GAN	6.6529	0.0186	0.0512	0.0326
RunGAN	**8.1207**	**0.0081**	**0.1001**	**0.0920**

Bold values show the Generator and discriminator loss resulted by our proposed architecture

Table 5 Confusion matrix

Model	TN	FP	FN	TP
GAN	748	2	1	749
RunGAN	**750**	**0**	**0**	**750**

Bold values show the Confusion matrix resulted by our proposed architecture the experimental results, RunGAN achieves a higher G-Loss of 8.1207. This increased G-Loss suggests that the discriminator outperforms the generator in identifying deceptive samples. The discriminator score demonstrates the model's ability to effectively detect Twitter spam within extensive datasets. As a result, the proposed method holds potential for addressing spam detection challenges and curbing the spread of spam on Twitter.

To further enhance the model's performance, we suggest implementing innovative algorithms such as Deep Belief Network (DBN) and introducing new hyperparameters and regulators during training. The model's flexibility enables the adjustment of hyperparameters to improve its dynamic learning capabilities and achieve greater accuracy with multiple features. Therefore, we believe that this dynamic model presents a viable solution for enhancing spam detection capabilities.

Data availability The dataset comes from the NSClab/Resources Twitter Spam. (http://nsclab.org/nsclab/resources/).

Declarations

Conflict of interest None declared.

References

1. Elmendili F, Idrissi YEBE (2020) A framework for spam detection in twitter based on recommendation system. Int J Intell Eng Syst. https://doi.org/10.22266/ijies2020.1031.09
2. Inuwa-Dutse I, Liptrott M, Korkontzelos I (2018) Detection of spam-posting accounts on Twitter. Neurocomputing. https://doi.org/10.1016/j.neucom.2018.07.044
3. Wu T et al (2017) Detecting spamming activities in twitter based on deep-learning technique. Concurrency Computat. https://doi.org/10.1002/cpe.4209
4. Fazil M, Abulaish M (2018) A Hybrid Approach for Detecting Automated Spammers in Twitter. IEEE Transact Informat Forensics Sec. https://doi.org/10.1109/TIFS.2018.2825958
5. Karakaşlı MS, Aydin MA, Yarkan S, Boyaci A (2019) Dynamic feature selection for spam detection in twitter. Lecture Notes Elect Eng. https://doi.org/10.1007/978-981-13-0408-8_20
6. Li C, Liu S (2018) A comparative study of the class imbalance problem in Twitter spam detection. Concurr Computat. https://doi.org/10.1002/cpe.4281
7. Çıtlak O, Dörterler M, Doğru İA (2019) A survey on detecting spam accounts on Twitter network. Soc Netw Anal Min 9(1):35. https://doi.org/10.1007/s13278-019-0582-x
8. Zheng X, Zeng Z, Chen Z, Yu Y, Rong C (2015) Detecting spammers on social networks. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.02.047
9. Yurtseven I, Bagriyanik S, Ayvaz S (2021) "A Review of Spam Detection in Social Media," In: Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021. doi: https://doi.org/10.1109/UBMK52708.2021.9558993.
10. Raza M, Jayasinghe ND, Muslam MMA (2021) A comprehensive review on email spam classification using machine learning algorithms. Int Conf Informat Netw. https://doi.org/10.1109/ICOIN50884.2021.9334020
11. Bhuvaneshwari P, Rao AN, Robinson YH (2021) Spam review detection using self attention based CNN and bi-directional LSTM. Multimedia Tools Applicat 80(12):18107–18124. https://doi.org/10.1007/s11042-021-10602-y
12. Kaddoura S, Chandrasekaran G, Popescu DE, Duraisamy JH (2022) A systematic literature review on spam content detection and classification. Peer J Comp Sci. https://doi.org/10.7717/PEERJ-CS.830
13. Wapet L, Tchana A, Tran GS, Hagimont D (2019) Preventing the propagation of a new kind of illegitimate apps. Future Generat Comp Syst. https://doi.org/10.1016/j.future.2018.11.051
14. Jain G, Sharma M, Agarwal B (2019) Spam detection in social media using convolutional and long short term memory neural network. Annals Mathemat Artif Intell. https://doi.org/10.1007/s10472-018-9612-z
15. Faris H et al (2019) An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. Informat Fusion. https://doi.org/10.1016/j.inffus.2018.08.002
16. Fu Q, Feng B, Guo D, Li Q (2018) Combating the evolving spammers in online social networks. Comput Secur 72:60–73. https://doi.org/10.1016/j.cose.2017.08.014
17. Kaur R, Singh S, Kumar H (2018) Rise of spam and compromised accounts in online social networks: a state-of-the-art review of different combating approaches. J Netw Comput Appl 112:53–88. https://doi.org/10.1016/j.jnca.2018.03.015
18. Barushka A, Hajek P (2020) Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. Neural Comput Appl 32(9):4239–4257. https://doi.org/10.1007/s00521-019-04331-5
19. Sharaff A, Jain M, Modugula G (2022) Feature based cluster ranking approach for single document summarization. Int J Informat Technol (Singapore). https://doi.org/10.1007/s41870-021-00853-1
20. Shahariar GM, Biswas S, Omar F, Shah FM, Binte Hassan S (2019) "Spam review detection using deep learning", in IEEE 10th annual information technology. Electron Mobile Commun Conf (IEMCON) 2019:0027–0033. https://doi.org/10.1109/IEMCON.2019.8936148
21. Gopi AP, Jyothi RNS, Narayana VL, Sandeep KS (2020) Classification of tweets data based on polarity using improved RBF kernel of SVM. Int J Informat Technol (Singapore). https://doi.org/10.1007/s41870-019-00409-4
22. Lin G, Sun N, Nepal S, Zhang J, Xiang Y, Hassan H (2017) Statistical twitter spam detection demystified: performance, stability and scalability. IEEE Access. https://doi.org/10.1109/ACCESS.2017.2710540
23. Cao C, Caverlee J (2015) Detecting spam URLs in social media via behavioral analysis. Lecture Notes Comp Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-16354-3_77
24. Diqi M, Mulyani SH, Pradila R (2023) DeepCov: effective prediction model of COVID-19 using CNN algorithm. SN Comp Sci 4(4):396. https://doi.org/10.1007/s42979-023-01834-w
25. Wanda P, Jie HJ (2020) DeepProfile: Finding fake profile in online social network using dynamic CNN. J Informat Sec Appl 52:102465. https://doi.org/10.1016/j.jisa.2020.102465
26. Jain G, Sharma M, Agarwal B (2019) Optimizing semantic LSTM for spam detection. Int J Informat Technol (Singapore). https://doi.org/10.1007/s41870-018-0157-5
27. Ghourabi A, Mahmood MA, Alzubi QM (2020) A hybrid CNN-LSTM model for SMS spam detection in Arabic and English messages. Future Internet. https://doi.org/10.3390/FI12090156
28. Bang D, Kang S, Shim H (2020) Discriminator feature-based inference by recycling the discriminator of GANs. Int J Comp Vision. https://doi.org/10.1007/s11263-020-01311-4

29. Diqi M, Hiswati ME, Nur AS (2022) StockGAN: robust stock price prediction using GAN algorithm. Int J Informat Technol (Singapore). https://doi.org/10.1007/s41870-022-00929-6
30. Barigye SJ, de la Vega JMG, Perez-Castillo Y (2020) Generative adversarial networks (GANs) based synthetic sampling for predictive modeling. Molecular Inform. https://doi.org/10.1002/minf.202000086
31. Madisetty S, Desarkar MS (2018) A neural network-based ensemble approach for spam detection in twitter. IEEE Transact Computat Soc Syst. https://doi.org/10.1109/TCSS.2018.2878852
32. Lee JY, Choi SI (2020) Improvement of learning stability of generative adversarial network using variational learning. Appl Sci (Switzerland). https://doi.org/10.3390/app10134528
33. Lu PH, Wang PC, Yu CM (2019) Empirical evaluation on synthetic data generation with generative adversarial network. ACM Int Conf Proc Series. https://doi.org/10.1145/3326467.3326474
34. Xu C, Ren J, Zhang D, Zhang Y, Qin Z, Ren K (2019) GANobfuscator: Mitigating information leakage under GAN via differential privacy. IEEE Transact Informat Forens Sec. https://doi.org/10.1109/TIFS.2019.2897874
35. Tang X, Qian T, You Z (2020) Generating behavior features for cold-start spam review detection with adversarial learning. Informat Sci. https://doi.org/10.1016/j.ins.2020.03.063
36. Kumar A, Dabas V, Hooda P (2020) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Informat Technol (Singapore). https://doi.org/10.1007/s41870-017-0072-1
37. Li M, Lin J, Ding Y, Liu Z, Zhu JY, Han S (2022) GAN Compression: Efficient Architectures for Interactive Conditional GANs. IEEE Transact Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3126742
38. Miller Z, Dickinson B, Deitrick W, Hu W, Wang AH (2014) Twitter spammer detection using data stream clustering. Informat Sci. https://doi.org/10.1016/j.ins.2013.11.016
39. Fitni QRS, Ramli K (2020) “Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems,” In: Proceedings - 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2020, doi: https://doi.org/10.1109/IAICT50021.2020.9172014.
40. Soares E, Garcia C, Poucas R, Camargo H, Leite D (2019) Evolving fuzzy set-based and cloud-based unsupervised classifiers for spam detection. IEEE Lat Am Trans 17(09):1449–1457. https://doi.org/10.1109/TLA.2019.8931138
41. Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H (2019) Sentiment analysis based on improved pre-trained word embeddings. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2018.08.044
42. Mohammed MA et al (2019) An anti-spam detection model for emails of multi-natural language. Xinan Jiaotong Daxue Xuebao/J Southwest Jiaotong Univ. https://doi.org/10.35741/issn.0258-2724.54.3.6
43. Singh AB, Singh KM, Chanu YJ, Thongam K, Singh KJ (2022) An improved image spam classification model based on deep learning techniques. Sec Communicat Net. https://doi.org/10.1155/2022/8905424
44. Wanda P (2022) RunMax: fake profile classification using novel nonlinear activation in CNN. Soc Netw Anal Min 12(1):158. https://doi.org/10.1007/s13278-022-00983-9