

# Spam social media profile detection using hybrid positive unlabelled learning

Nidhi A. Patel<sup>1</sup>, Nirali Nanavati<sup>2</sup>

<sup>1</sup>Gujarat Technological University, Ahmedabad, India

<sup>2</sup>Department of Computer Engineering, Sarvajani College of Engineering and Technology, Surat, India

## Article Info

### Article history:

Received Jan 18, 2025

Revised Sep 30, 2025

Accepted Oct 18, 2025

### Keywords:

Machine learning

PU-learning

Semi-supervised

Social media

Social spam

Spam profile

## ABSTRACT

Online social networks (OSNs) are a communication medium of social interaction for people, where social activities, entertainment, business-oriented activities, and information are exchanged. It creates an environment with worldwide connectivity where groups of individuals may discuss their interests and activities on social media platforms. Billions of people routinely interact with social content, opinion sharing, recommendations, networking, scouting, social campaigns, alerting on OSNs. The increase in popularity of OSNs creates new challenges and perspectives to the researchers of social networks, which is of interest in various fields. One of the most popular networking platforms for microblogging is X (formerly Twitter). Millions of spam accounts have inundated the X network, which could damage normal users' security and privacy. Hence, the research in this filed has become essential for enhancing real users' protection and identifying spam profiles. In this manuscript, we propose hybrid approach based on semi-supervised learning to detect the spam profiles. The proposed work is based on the positive and unlabeled (PU) learning algorithm, which learns from an unlabeled dataset and a small number of positive instances. Simulation results demonstrate that our approach outperformed existing PU learning approach by 17.39% and 17.51% improvement respectively in spam detection rate on X and Instagram datasets.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Nidhi A. Patel

Gujarat Technological University

Ahmedabad, India

Email: nidhi.patel0051@gmail.com

## 1. INTRODUCTION

Online social networking (OSN) services sometimes referred to as social media networking is an online platform used by individuals to establish social networks and interactions with others, to converse their thoughts and feelings about different subjects [1]. The use of OSNs has increased due to development in technology especially, the internet, and hence transformed the way people interact with each other [2]. These OSNs link with other individuals based on similar activities, real-life backgrounds or connections, individual, and career interests [3]. It is simple to access and use OSNs like LinkedIn, Instagram, X (formerly Twitter), and Facebook. The result is, individuals interact with one another on these social media platforms [4].

As per the survey, approximately 5 billion people use different social media platforms to communicate with family, friends, and colleagues [5]. When users post content like text, images, and videos, maintaining privacy and security becomes crucial [6], [7]. Spam profile users are meant to conceal the identity and personal information of individuals for unethical purposes [8]. Our society is at risk from such spam profile accounts. It is important to find such spam profiles on OSNs [9].

A spammer is a user who performs malicious activities on online platforms with the intent to disrupt social media environments and compromise user privacy. According to X, spam accounts typically exhibit behaviors such as posting harmful links, excessive retweeting, creating multiple accounts, suspicious following behavior, overuse of the @mention and #hashtag functions to gain attention, and frequent posting on trending topics to attract casual user interaction [10]. X is one of the well-known social networking sites that allows for interaction and collaboration between millions of users. X was established in 2006. X has around 611 million and Instagram has around 2,000 million active users monthly for a variety of motives [11].

According to a recent analysis, between 8.8% and 14.6% of all X profiles are spam. This corresponds to between 29 and 48 million profiles. For Instagram, the market researchers at ghost data have now also conducted a survey. On average, there are 95 million Instagram spam accounts, which would be around 10% of the platform's active users [12]. In this study, we focus on detecting spam profiles on two major OSN: X and Instagram. X is primarily a text-based platform with features such as tweets, retweets, hashtags, user mentions, follower, and following [13]. This makes X suitable for analyzing content-based and profile-based features in spam profile detection. Instagram, on the other hand, is more focused on visual content and user profile attributes such as profile pictures, usernames, and follower relationships [14]. The differences in data types and feature sets between these two platforms provide a comprehensive environment to evaluate the effectiveness and generalizability of our hybrid positive-unlabeled learning approach for spam profile detection. This diversity in both allows us to evaluate the robustness and generalizability of our hybrid positive-unlabeled learning method across varied social media environments.

It is difficult to identify groups of spammers and hence individual spam account using heuristic and classical methods [15]–[18]. Machine learning helps to achieve good accuracy to identify the same [19]–[21]. Spam profile detection has been performed in different ways [7], [15], [22]. Efforts have been done to detect spam profiles on OSNs using heuristic and various machine learning techniques using classification and clustering techniques [23], [24]. Classification techniques require a labeled dataset, and in clustering, the model is trained with an unlabeled dataset. Hybrid approach resolves constraints of classification and clustering and leads to accurate predictions with improved decision-making.

The paper is organized as follows: the summary of related work on spam profile detection in OSNs is mentioned in section 2. Section 3 describes the proposed hybrid approach. Section 4 highlights the performance of evaluation results, and finally section 5 concludes this work with future research directions.

## 2. SPAM PROFILE DETECTION IN ONLINE SOCIAL NETWORK

The rise in social activities among registered members of the X social network has contributed to its increased popularity. X serves as both a microblogging OSNs and a platform for news updates at the same time. Cybercriminals have recently become more interested in X because of the increase in social contacts. X has been used by spam users to disseminate spam content, publish phishing links, inundate the network with bogus accounts, and carry out other illegal operations. Finding the spam accounts that are part of the network of spammers that carry out these actions is an essential beginning. Numerous methods to identify a group of spammers have been put forth by researchers. However, each of these strategies focused on a particular group of spammers. This study proposes an alternative method for detecting spam accounts on X based on the characteristics of spam accounts. To enhance the performance, we used a hybridization method for spam profile detection.

The different strategies have been proposed depending on various features. Some methods for identifying spam based on user profile and message content features [25]–[27]; some work relied on graph-based features, specifically the connection of a social graph and distance [28]–[30]; and some research based on embedded URLs as a method of characteristics of spam detection. According to the survey, spam profiles can be identified by various methods, namely compromised profile/account, content/tweet, graph/network/friendship, URLs, blacklist, and hybrid based.

Profile/account-based features are developed based on the properties and relationship of user accounts. As these features are related to user profile, it recognized all the attributes that were connected to accounts of users. Research in [31]–[38] has identified spam profiles based on profile/account based features using different social media datasets using classification. The research in [39], [40] has used classification and cluster separately for identifying spam profiles. Spam profile detection using content/tweet based investigates content features which are connected to the tweets users published. Gupta *et al.* [41] have identified spam profiles based on content/tweet based features using different social media datasets using classification. Chu *et al.* [42] have used classification and clustering separately for identifying spam profiles.

Spam profile detection based on graph/network/friendship analyzes the connectivity and distance of the graph of social relation. Research in [43]–[47] have identified spam profiles based on graph/network/friendship features using different social media datasets using classification. The research in [48], [49] have used clustering for identifying spam profiles. Ahmed and Abulaish [50] have used

classification and cluster separately for identifying spam profiles. The length restriction on tweet descriptions makes it more lucrative for spammers to submit URLs for sharing harmful content than plain text. Research in [51]–[54] has identified spam profiles based on URL based features using different social media datasets using classification of URLs. The blacklist method is to check a tweet with harmful links. Any tweet with at least one harmful link is marked as spam. The topic modeling approach and the keyword extraction approach are used to generate the blacklist [55]. Swe and Myo [55] have identified spam profiles based on blacklist based features using different social media datasets using classification of blacklist methods.

Spam account deception is not only identified by one feature, but combination of any two or more methods of profile based, content based, network based, URL based or blacklist based. The research in [13], [15], [21], [26], [27], [30], [56]–[66] have identified spam profiles based on different features using different social media datasets using classification. Research in [14], [16], [67] has used clustering for identifying spam profiles. Gupta and Kaushal [68] have used classification and cluster separately for identifying spam profiles. Few of the research that used a hybrid of classification and clustering [17], [69] for spam profile detection.

### 2.1. Research gap

As per the literature survey and analysis, most of the work has been done either only using profile/user based or content/tweet based or graph/network/friendship based or URL or blacklist based. Using only single feature is not capable to detect spam profiles due to challenges with new accounts of profile [31], evolving spammer tactics [32], [33], limited context [39], dependency on user cooperation [40] in profile based features, limited historical behavior [41], ambiguous content, sensitivity to content changes [36], limited user interaction analysis [36] in content/tweet based features, limited profile connectivity [43], [48], incomplete graph data [44], [50], dependence on graph connectivity [45], [46], sensitivity to friend-finding [56], [67] strategies in graph/network/friendship based features, URL obfuscation [50], dependence on URL databases [52], privacy concerns [53], limited URL access [54], resource intensive [54] in URL based features. In blacklist based features, incomplete and outdated blacklists [55], dependency on historical data, limited contextual understanding, privacy concerns in blacklist based so there is a need to investigate more on hybridized features.

The hybridized features that can lead to better decision-making [69], deeper understanding of complex systems, identify the patterns effectively, comprehensive view of a user's behavior [15], multiple data points, increase the redundancy of spam detection [16], and more accurate predictions [17]. The profile/account comprises key elements that can be applied on any social media platform. For use of any social media without any additional information, a profile is required. Using the profile, we can also determine a detail that has been posted on social media. The hybridization of “profile and content” [59] is important because it's combining information from both the user's profile (such as their account details, history, and activity) [32], [34], [36] and the content they generate (such as posts, comments, or messages) [16], [17] to better identify and flag potentially spam or malicious accounts on social media. This approach is used by online communities, social media platforms, and websites to enhance their spam detection systems. Authors in [13], [15], [21], [26], [27], [30], [56]–[66] have either used classification or clustering techniques separately while few of the researchers [17], [69] have combined both approaches to get the advantages of both.

Further, research in [17], [69], the model is not trained using multiple iterations. To overcome the above gaps, we hereby proposed a hybrid approach using semi-supervised based on the positive and unlabeled (PU) learning mentioned in [70], [71]. Moreover, the existing PU-learning approach [70] is classification-based. This approach considers unlabeled instances as negative instances because at the initial level labeling, unlabeled classes are done only with positive instances. To address the above issue, there is a need for an accurate detection model with a good learning paradigm. In our proposed approach, we used hybrid technique with classification and clustering instead of the above assumption. We applied clustering on negative instances and got actual results, and so at the initial level, we got the labeled instances of spam and non-spam profiles.

## 3. RESULTS AND DISCUSSION

The proposed PU-learning approach is a hybridization-based classification and clustering approach for spam profile detection in OSN, and the same has been mentioned in detail in this section. Figure 1 represents the flow of our proposed approach. The process starts with the data selection stage. Next, in the feature selection stage, we used profile and content-based features of social profiles. After that, the featured data will be passed in the machine learning model. This process will be stopped with genuine and deceptive profiles with accuracy, precision, recall, and F1-score.

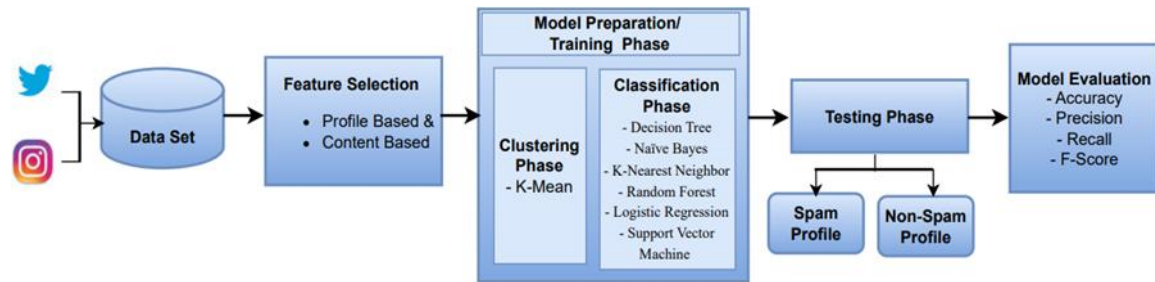


Figure 1. The flow of methodological approach to detect identify deception

### 3.1. Dataset

In most of the existing work, authors have created their own dataset. Some authors have used dataset, which is publicly available. We used the X dataset that has been built and utilized by the researchers in [27], [59]. For Instagram, we considered the dataset used in [72].

### 3.2. Feature selection

Feature selection method functions as a preprocessing phase for classification and prediction algorithms [73]. In order to differentiate between spam and non-spam, it is important to analyze the tweet content [42] and profile details [39], [40] together. The proposed work considers user profile/account and content-based features. There are six user profile related features in X dataset which represent the user account. The seven additional features which are related to the tweet's content are also considered in our work. In this way, for the development of accurate detection models, a total of 13 features which are mentioned in Table 1 have been considered. Table 1 gives a summary of each feature. Table 2 represents the details of 11 features of the Instagram dataset.

Table 1. List of features name and description of X dataset [59]

Feature name	Feature description
Profile based features	
account_age	Account's age
no_follower	The follower counts of the X users
no_following	The followings count of the X users
no_userfavorites	The count of favorites X user received
no_lists	The count of lists X user added
no_tweets	The count of tweets X user sent
Content based feature	
no_retweets	The count of retweets the tweet
no_tweetfavorites	The count of favorites this tweet received
no_hashtag	The count of hashtags included in the tweet
no_usermention	The count of users mentions included in the tweet
no_urls	The count of URLs included in the tweet
no_chars	The count of characters in tweet
no_digits	The count of digits in the tweet

Table 2. List of features name and description of Instagram dataset [72]

Feature name	Feature description
profile pic	Presence or absence of a profile picture for the user
nums/length username	The ratio of numeric characters to the total length of the username
fullname words	Full name represented as a sequence of word tokens
nums/length full name	The ratio of numeric characters to the total length of the full name
name==username	Check for exact string match between username and full name
description length	The number of characters in the user bio
external URL	Has external URL or not
Private	Private or not
Posts	Number of posts
Followers	Number of followers
Follows	Number of follows

### 3.3. Clustering phase

Clustering is useful to find the effective instances from the large data available in the dataset. Those instances which play a great role in determining classification results are sorted and extracted from training samples. This reduces a large data which has no greater impact in producing results, hence there is an

increase in system efficiency. In our work, we used the K-Means clustering algorithm. The labeled instances are extracted, and from them a labeled training sample is generated. After analysis of the different values of  $K$ ,  $K$  equals 2 gives the best results. The result of the clustering algorithm is passed to the next phase to perform the classification of spam and non-spam profiles.

### 3.4. Classification phase

Classification is an algorithm or model that takes a set of labeled samples with features as input. The algorithm or model is trained depending on classifying the set of instances with different classes, and testing gets performed on unlabeled dataset. We applied different six classification algorithms namely decision tree classifier, naïve Bayes classifier, k-nearest neighbor classifier, random forest classifier, logistic regression, and support vector machine for detecting spam profiles.

### 3.5. Proposed algorithm

The proposed spam profile detection algorithm is explained in this section. The outcome of the clustering algorithm will pass to a classification model that classifies spam and non-spam classes. Here, we propose a semi-supervised hybrid approach based on the state-of-the-art PU-learning approach [70] for spam profile detection in OSNs. Initially, we considered PU instances as an input to the algorithm. For labeling the unlabeled instances, we applied clustering on them which results in two clusters. One cluster represents the positive review as spam profile and the second cluster represents non-spam profile as genuine reviews. The classifier is then generated based on the initial set of positive instances and the output of the clustering phase. This generated classifier is applied on the cluster of negative reviews (non-spam profile) which further generates two classes viz. spam profile and non-spam profile. Here, some of the non-spam profiles may not be classified correctly and may belong to spam profile class. To identify those spam profiles, again the classifier is generated based on the initial set of positive instances and output of newly generated negative instances and applied on a new set of negative profile classes. Until the stop requirement is satisfied, this process is repeated. In the end, generated two sets of classify spam and non-spam profiles. Algorithm 1 show the pseudo code of the proposed hybrid approach which is a combination of classification and clustering for identifying spam profile detection. Algorithm 1 represents the proposed ideology.

Algorithm 1: Proposed hybridized PU-learning for spam profile detection

```

1. Input: collection of PU instances.
2.  $i \leftarrow 1$ ;
3.  $|W_0| \leftarrow |U_1|$ ;
4.  $|W_1| \leftarrow |U_1|$ ;
5.  $U_L \leftarrow \text{Cluster}(U)$ ;
6.  $N \leftarrow \text{Extract\_Negative}(U_L)$ ;
7.  $G \leftarrow \text{Generate\_Classifier}(P, U_L)$ ;
8.  $N_L \leftarrow G(U)$ ;
9.  $L_i \leftarrow \text{Extract\_Negative}(N_L)$ ;
10. while  $|W_i| \leq |W_{i-1}|$  do
11.    $C_i \leftarrow \text{Generate\_Classifier}(P, L_iL)$ ;           // Generate classifier
12.    $L_iL \leftarrow C_i(L_i)$ ;                             // Train classifier
   applied in the unlabeled dataset
13.    $W_i \leftarrow \text{Extract\_Positives}(L_iL)$ ;           // Extract positive (spam) instances
   from labeled dataset
14.    $L_{i+1} \leftarrow L_i - W_i$ ;                         // Remove positive (spam)
   profiles instances
15.    $i \leftarrow i + 1$ ;
16. End while
17. Return  $C_i$ 
18. Output: testing instances classified as positive and negative opinion reviews.
```

Here,  $P$  is number of positive instances,  $U_1$  is original unlabeled data set,  $C_i$  is classifier at iteration  $i$ ,  $N$  is negative instances,  $G$  is generated classifier,  $L_i$  is unlabeled set at iteration  $i$ , and  $W_i$  is unlabeled instance classified as positive by classifier  $C_i$ .

## 4. RESULTS AND DISCUSSION

For the performance evaluation of our approach with existing work on the dataset mentioned in sub-section 3.1, we used different Python libraries. For training and testing data, we considered the ratio of spam and non-spam profiles to be the same viz. 1:1. For the simulation and comparison, different models were trained based on randomly selected 1.5 million tweets out of 2 million tweets. For testing, we randomly selected 25,000 tweets for spam and remaining 25,000 tweets for non-spam profiles. For the Instagram

dataset, we considered 70% data for training and the remaining 30% for testing after analysis with different training and testing ratios.

We utilized the following six classifiers in the existing and proposed approach and trained the model. We used: decision tree classifier, naïve Bayes classifier, k-nearest neighbor classifier, random forest classifier, logistic regression, and support vector machine. We trained the model and then tested it using a test dataset. For evaluation, we compared the results by accuracy, precision, recall, and F-score parameters. We have implemented the existing approach [70] and collected results using various datasets. The graphs in Figures 2 and 3 depict the accuracy of the X and Instagram datasets respectively. The x axis denotes the different types of algorithms, and the y axis represents the corresponding accuracy scores.

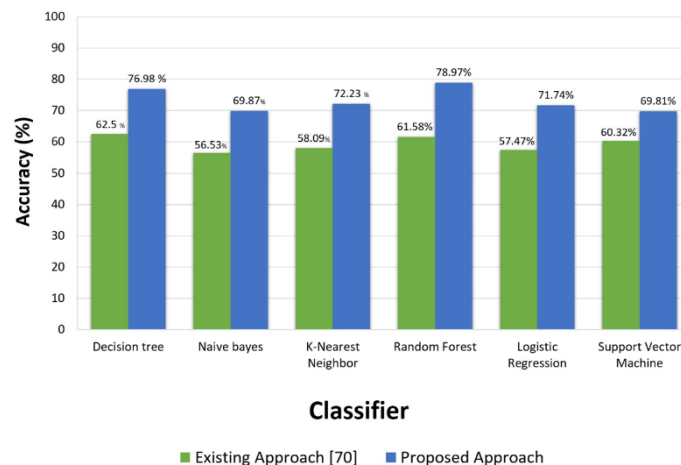


Figure 2. Accuracy comparison of existing and proposed approach for X dataset

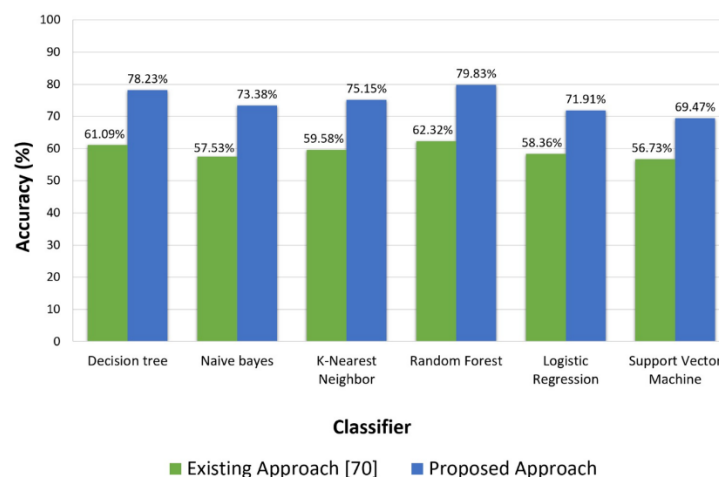


Figure 3. Accuracy comparison of existing and proposed approach for Instagram dataset

Our approach has been compared with existing PU learning algorithm [70] on X profile and content-based dataset with identified features [59]. Our approach which is a hybrid clustering and classification technique outperforms the existing approach in terms of accuracy, precision, recall, and F-score. In Table 3, the highest level of accuracy we have is 62.50% using a decision tree in an existing approach, and 78.97% accuracy using random forest in the proposed approach. In Table 4, the highest level of accuracy we have is 62.32% in an existing approach and 79.83% accuracy in the proposed approach using the random forest. In both datasets, we got higher accuracy in the proposed approach using random forest. The random forest is an ensemble learning algorithm by combining multiple trees to improve performance and reduce the risk of overfitting.

Table 3. Comparative results for X dataset

Approach	Accuracy (%)	Precision (P)	Recall (R)	F-score (F)	Classifier
Existing [70]	62.50	0.6316	0.6308	0.6312	Decision tree
Proposed	76.98	0.7829	0.7521	0.7672	
Existing [70]	56.53	0.5654	0.5578	0.5616	Naïve Bayes
Proposed	69.87	0.7067	0.6928	0.6997	
Existing [70]	58.09	0.5981	0.5873	0.5927	K-nearest neighbor
Proposed	72.23	0.7110	0.7146	0.7128	
Existing [70]	61.58	0.6250	0.6012	0.6129	Random forest
Proposed	78.97	0.7812	0.7729	0.7770	
Existing [70]	57.47	0.5935	0.5681	0.5805	Logistic regression
Proposed	71.74	0.7389	0.7105	0.7244	
Existing [70]	60.32	0.6101	0.5931	0.6015	Support vector machine
Proposed	69.81	0.7242	0.7058	0.7149	

Table 4. Comparative results for Instagram dataset

Approach	Accuracy (%)	Precision (P)	Recall (R)	F-score (F)	Classifier
Existing [70]	61.09	0.6047	0.5835	0.5939	Decision tree
Proposed	78.23	0.7636	0.7813	0.7723	
Existing [70]	57.53	0.5923	0.5512	0.5710	Naïve Bayes
Proposed	73.38	0.7531	0.7458	0.7494	
Existing [70]	59.58	0.5808	0.6042	0.5923	K-nearest neighbor
Proposed	75.15	0.7437	0.7713	0.7572	
Existing [70]	62.32	0.6317	0.6346	0.6331	Random forest
Proposed	79.83	0.8198	0.8246	0.8222	
Existing [70]	58.36	0.6193	0.5719	0.5947	Logistic regression
Proposed	71.91	0.7391	0.7251	0.7320	
Existing [70]	56.73	0.5841	0.5691	0.5765	Support vector machine
Proposed	69.47	0.7153	0.7015	0.7083	

## 5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The rapid growth of spam accounts has been allowed by the ongoing expansion in the variety of social network content. X is gaining popularity on social platforms for marketers and fraudsters who frequently use spam accounts to further achieve their goals because of its openness, the capacity to influence featured taglines, and the capacity to inflate profiles. Spammer techniques for evasion must be developed in order to minimize the impact of spam accounts on genuine users. This research conducted a thorough analysis of machine learning classifiers for identification of spam account in X and Instagram. We have suggested a semi-supervised hybrid method that includes clustering and classification. Our work would help to create a useful system that can identify spam X and Instagram accounts and secure genuine users from spammers. As a part of future work, this model can be extended for additional sites, which are afflicted by massively produced illegal accounts, such as other web forums, email services with more features, and OSN services. In order to find related information at a given moment. Tweet timestamp can also be integrated in the investigative approaches for spam profile detection.

## ACKNOWLEDGMENTS

Authors would like to express our sincere gratitude to all those who have supported and contributed to this research. Primarily, first author extends our heartfelt thanks to our Ph.D. supervisor (Dr. Nirali Nanavati) for her unwavering guidance, invaluable insights, and encouragement throughout the research process.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nidhi A. Patel	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Nirali Nanavati	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest related to this work.

## DATA AVAILABILITY

Datasets utilized in this research are cited in reference [27], [59], [72].

## REFERENCES

- [1] P. A. Bhat, M. Chaitra, S. R. Thyli, N. Anitha, and Rajeshwari, "Fake Instagram profile detection," *8th IEEE International Conference on Computational System and Information Technology for Sustainable Solutions*, 2024, pp. 1-6, doi: 10.1109/CSITSS64042.2024.10816892.
- [2] V. Pandi, P. Nithiyanandam, S. Manickavasagam, I. M. Meerasha, R. Jaganathan, and M. K. Balasubramanian, "A comprehensive analysis of consumer decisions on Twitter dataset using machine learning algorithms," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1085–1093, 2022, doi: 10.11591/ijai.v11.i3.pp1085-1093.
- [3] S. Chelas, G. Routis, and I. Roussaki, "Detection of fake Instagram accounts via machine learning techniques," *Computers*, vol. 13, no. 11, 2024, doi: 10.3390/computers13110296.
- [4] G. Rajesh, K. P. Kalaivani, D. Hemalatha, V. Prabhu, and N. S. R. Lingham, "Identification of fake and spam users on social networking platforms," *2023 Intelligent Computing and Control for Engineering and Business Systems*, 2023, pp. 1-5, doi: 10.1109/ICCEBS58601.2023.10449109.
- [5] Datareportal, "Global social media statistics," *datareportal.com*. [Online]. Available: <https://datareportal.com/social-media-users>
- [6] T. Qiu, X. Liu, X. Zhou, W. Qu, Z. Ning, and C. L. P. Chen, "An adaptive social spammer detection model with semi-supervised broad learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4622–4635, 2022, doi: 10.1109/TKDE.2020.3047857.
- [7] A. M. Al-Zoubi, J. Alqatawna, and H. Faris, "Spam profile detection in social networks based on public features," *8th International Conference on Information and Communication Systems*, pp. 130–135, 2017, doi: 10.1109/IACS.2017.7921959.
- [8] A. Almaatouq et al., "If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts," *International Journal of Information Security*, vol. 15, no. 5, pp. 475–491, 2016, doi: 10.1007/s10207-016-0321-5.
- [9] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, no. 1, pp. 27–34, 2015, doi: 10.1016/j.neucom.2015.02.047.
- [10] M. Chakraborty, S. Das, and R. Mamidi, "Detection of fake users in SMPs using NLP and graph embeddings," *arXiv:2104.13094v1*, 2021.
- [11] M. Woodward, "Twitter user statistics 2025: what happened after 'x' rebranding?," *searchlogistics.com*. [Online]. Available: <https://www.searchlogistics.com/learn/statistics/twitter-user-statistics/>
- [12] C. Erxleben, "95 Million bots: one in ten instagram accounts is fake," *basichthinking.com*. [Online]. Available: <https://www.basichthinking.com/bots-instagram-accounts-fake/>
- [13] P. Sowmya and M. Chatterjee, "Detection of fake and clone accounts in Twitter using classification and distance measure algorithms," *International Conference on Communication and Signal Processing*, vol. 265, pp. 67–70, 2020, doi: 10.1109/ICCSP48568.2020.9182353.
- [14] K. Zarei, R. Farahbakhsh, and N. Crespi, "How impersonators exploit Instagram to generate fake engagement?," *IEEE International Conference on Communications*, 2020, pp. 1-6, doi: 10.1109/ICC40277.2020.9149431.
- [15] C. Kumar, T. S. Bharti, and S. Prakash, "A hybrid data-driven framework for SPAM detection in online social network," *Procedia Computer Science*, vol. 218, pp. 124–132, 2022, doi: 10.1016/j.procs.2022.12.408.
- [16] T. Jose and S. S. Babu, "Detecting spammers on social network through clustering technique," *Journal of Ambient Intelligence and Humanized Computing*, 2019, doi: 10.1007/s12652-019-01541-6.
- [17] G. Xu, J. Qi, D. Huang, and M. Daneshmand, "Detecting spammers on social networks based on a hybrid model," *IEEE International Conference on Big Data, Big Data 2016*, pp. 3062–3068, 2016, doi: 10.1109/BigData.2016.7840960.
- [18] A. M. Priyatno, "Spammer detection based on account, tweet, and community activity on Twitter," *Jurnal Ilmu Komputer dan Informasi*, vol. 13, no. 2, pp. 97–107, 2020, doi: 10.21609/jiki.v13i2.871.
- [19] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Annals of Mathematics and Artificial Intelligence*, vol. 85, no. 1, pp. 21–44, 2019, doi: 10.1007/s10472-018-9612-z.
- [20] B. Kardaş, I. E. Bayar, T. Özyer, and R. Alhajj, "Detecting spam tweets using machine learning and effective preprocessing," *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 393–398, 2021, doi: 10.1145/3487351.3490968.
- [21] S. B. Abkenar, E. Mahdipour, S. M. Jameii, and M. H. Kashani, "A hybrid classification method for Twitter spam detection based on differential evolution and random forest," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 21, 2021, doi: 10.1002/cpe.6381.
- [22] R. Sharma and A. Sharma, "Fake account detection using the machine learning technique," *Smart Computing*, pp. 197–203, 2021, doi: 10.1201/9781003167488-25.
- [23] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on Twitter," *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1191–1198, 2018, doi: 10.1109/ASONAM.2018.8508495.
- [24] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," *2019 International Conference on Electrical, Computer and Communication Engineering*, vol. 10, no. 7, pp. 1–5, 2019, doi: 10.1109/ECACE.2019.8679186.






- [25] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," *Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 1–9, 2010, doi: 10.1145/1920261.1920263.
- [26] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," *7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2010, pp. 1–10.
- [27] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 million spam tweets: a large ground truth for timely Twitter spam detection," *IEEE International Conference on Communications*, pp. 7065–7070, 2015, doi: 10.1109/ICC.2015.7249453.
- [28] J. Song, S. Lee, and J. Kim, "Spam filtering in Twitter using sender-receiver relationship," *Recent Advances in Intrusion Detection - 14th International Symposium, Berlin, USA*, pp. 301–317, 2011, doi: 10.1007/978-3-642-23644-0\_16.
- [29] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving Twitter spammers," *RAID'11: Proceedings of the 14th international conference on Recent Advances in Intrusion Detection*, pp. 318–337, 2011, doi: 10.1007/978-3-642-23644-0\_17.
- [30] A. H. Wang, "Don't follow me - spam detection in Twitter," *SECURITY 2010 - Proceedings of the International Conference on Security and Cryptography*, pp. 142–151, 2010, doi: 10.5220/0002996201420151.
- [31] E. V. D. Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018, doi: 10.1109/ACCESS.2018.2796018.
- [32] A. K. Ali, F. S. Hanoon, and S. F. Raheem, "The effect of feature selection methods in detecting malicious accounts in social media," *Journal of Advanced Sciences and Nanotechnology*, vol. 2, no. 1, pp. 215–224, 2023, doi: 10.55945/joasnt.2023.2.1.215-224.
- [33] S. D. Munoz and E. P. G. Pinto, "A dataset for the detection of fake profiles on social networking services," *2020 International Conference on Computational Science and Computational Intelligence*, pp. 230–237, 2020, doi: 10.1109/CSCI51800.2020.00046.
- [34] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in Facebook," *ISEA Asia Security and Privacy Conference 2017*, 2017, doi: 10.1109/ISEASP.2017.7976996.
- [35] C. Chen, J. Zhang, Y. Xiang, and W. Zhou, "Asymmetric self-learning for tackling Twitter spam drift," *2015 IEEE Conference on Computer Communications Workshops*, Hong Kong, China, pp. 208–213, 2015, doi: 10.1109/INFCOMW.2015.7179386.
- [36] A. Khalil, H. Hajjdiab, and N. Al-Qirim, "Detecting fake followers in Twitter: a machine learning approach," *International Journal of Machine Learning and Computing*, vol. 7, no. 6, pp. 198–202, 2017, doi: 10.18178/ijmlc.2017.7.6.646.
- [37] A. K. Ali and A. M. Abdullah, "Fake accounts detection on social media using stack ensemble system," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 3013–3022, 2022, doi: 10.11591/ijece.v12i3.pp3013-3022.
- [38] M. Alsaleh, A. Alarifi, A. M. Al-Salman, M. Alfayez, and A. Almuhaayin, "TSD: Detecting sybil accounts in Twitter," *2014 13th International Conference on Machine Learning and Applications*, pp. 463–469, 2014, doi: 10.1109/ICMLA.2014.81.
- [39] A. N. Hakimi et al., "Identifying fake account in Facebook using machine learning," in *Advances in Visual Informatics*, Cham, Switzerland: Springer, 2019, pp. 441–450. doi: 10.1007/978-3-030-34032-2\_39.
- [40] S. Lee and J. Kim, "Early filtering of ephemeral malicious accounts on Twitter," *Computer Communications*, vol. 54, pp. 48–57, 2014, doi: 10.1016/j.comcom.2014.08.006.
- [41] A. Gupta, H. Lamba, and P. Kumaraguru, "\$1.00 per RT #BostonMarathon #PrayForBoston: analyzing fake content on Twitter," *2013 APWG eCrime Researchers Summit*, pp. 1–12, 2013, doi: 10.1109/eCRS.2013.6805772.
- [42] Z. Chu, I. Widjaja, and H. Wang, "Detecting social spam campaigns on Twitter," *Proceedings of the 10th international conference on Applied Cryptography and Network Security*, pp. 455–472, 2012, doi: 10.1007/978-3-642-31284-7\_27.
- [43] V. M. Priyadarshini and A. Valarmathi, "A novel spam detection technique for detecting and classifying malicious profiles in online social networks," *Journal of Intelligent and Fuzzy Systems*, vol. 41, no. 1, pp. 993–1007, 2021, doi: 10.3233/JIFS-202937.
- [44] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "Friend or foe? fake profile identification in online social networks," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–23, 2014, doi: 10.1007/s13278-014-0194-4.
- [45] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying fake accounts on social networks based on graph analysis and classification algorithms," *Security and Communication Networks*, 2018, doi: 10.1155/2018/5923156.
- [46] S. Y. Bhat and M. Abulaish, "Community-based features for identifying spammers in online social networks," *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 100–107, 2013, doi: 10.1145/2492517.2492567.
- [47] S. Y. Bhat, M. Abulaish, and A. A. Mirza, "Spammer classification using ensemble methods over structural social network features," *2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops*, vol. 2, pp. 454–458, 2014, doi: 10.1109/WI-IAT.2014.133.
- [48] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Detection of fake accounts in social networks based on one class classification," *ISecure*, vol. 11, no. 2, pp. 173–183, 2019, doi: 10.22042/iseure.2019.165312.450.
- [49] F. Ahmed and M. Abulaish, "An MCL-based approach for spam profile detection in online social networks," *11th IEEE Int. Conference on Trust, Security and Privacy in Computing and Communications, TrustCom-2012 - 11th IEEE Int. Conference on Ubiquitous Computing and Communications*, pp. 602–608, 2012, doi: 10.1109/TrustCom.2012.83.
- [50] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Computer Communications*, vol. 36, no. 10–11, pp. 1120–1129, 2013, doi: 10.1016/j.comcom.2013.04.004.
- [51] R. R. Rout, G. Lingam, and D. V. L. N. Somayajulu, "Detection of malicious social Bots using learning automata with URL features in Twitter network," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 1004–1018, Aug. 2020, doi: 10.1109/TCSS.2020.2992223.
- [52] P. C. Lin and P. M. Huang, "A study of effective features for detecting long-surviving Twitter spam accounts," *International Conference on Advanced Communication Technology*, pp. 841–846, 2013.
- [53] Y. H. F. Jbara and H. A. S. Mohamed, "Twitter spammer identification using URL-based detection," *IOP Conference Series: Materials Science and Engineering*, vol. 925, no. 1, Sep. 2020, doi: 10.1088/1757-899X/925/1/012014.
- [54] X. Zhang, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in the Twitter social network," *Proceedings - IEEE International Conference on Data Mining*, pp. 1194–1199, 2012, doi: 10.1109/ICDM.2012.28.
- [55] M. M. Swe and N. N. Myo, "Fake accounts detection on Twitter using blacklist," *17th IEEE/ACIS International Conference on Computer and Information Science*, pp. 562–566, 2018, doi: 10.1109/ICIS.2018.8466499.
- [56] I. I. -Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter," *Neurocomputing*, vol. 315, pp. 496–511, 2018, doi: 10.1016/j.neucom.2018.07.044.
- [57] F. S. Alsubaiei, "Detection of inappropriate tweets linked to fake accounts on Twitter," *Applied Sciences*, vol. 13, no. 5, 2023, doi: 10.3390/app13053013.
- [58] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for Twitter," *2017 14th International Bhurban Conference on Applied Sciences and Technology*, pp. 466–471, 2017, doi: 10.1109/IBCAST.2017.7868095.




- [59] N. Sun, G. Lin, J. Qiu, and P. Rimba, "Near real-time Twitter spam detection with machine learning techniques," *International Journal of Computers and Applications*, vol. 44, no. 4, pp. 338–348, 2022, doi: 10.1080/1206212X.2020.1751387.
- [60] I. David, O. S. Siordia, and D. Moctezuma, "Features combination for the detection of malicious Twitter accounts," *2016 IEEE International Autumn Meeting on Power, Electronics and Computing*, 2017, doi: 10.1109/ROPEC.2016.7830626.
- [61] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, "A machine learning approach for Twitter spammers detection," *International Carnahan Conference on Security Technology*, 2014, doi: 10.1109/CCST.2014.6987029.
- [62] H. Shen and X. Liu, "Detecting spammers on Twitter based on content and social interaction," *2015 International Conference on Network and Information Systems for Computers*, pp. 413–417, 2015, doi: 10.1109/ICNISC.2015.82.
- [63] M. A. Albahar, "Detecting fraudulent Twitter profiles: a model for fraud detection in online social networks," *International Journal of Innovative Computing, Information and Control*, vol. 15, no. 5, pp. 1629–1639, 2019, doi: 10.24507/ijicic.15.05.1629.
- [64] J. Kaubiyal and A. K. Jain, "A feature-based approach to detect fake profiles in Twitter," *Proceedings of the 3rd International Conference on Big Data and Internet of Things*, pp. 135–139, 2019, doi: 10.1145/3361758.3361784.
- [65] N. C. Le, M. T. Dao, H. L. Nguyen, T. N. Nguyen, and H. Vu, "An application of random walk on fake account detection problem: a hybrid approach," *2020 RIVF International Conference on Computing and Communication Technologies*, 2020, pp. 1–6, doi: 10.1109/RIVF48685.2020.9140749.
- [66] D. Punkamol and R. Marukatat, "Detection of account cloning in online social networks," *2020 8th International Electrical Engineering Congress*, 2020, pp. 1–4, doi: 10.1109/IEEECON48109.2020.229558.
- [67] N. Eshraqi, M. Jalali, and M. H. Moattar, "Detecting spam tweets in Twitter using a data stream clustering algorithm," *2nd International Congress on Technology, Communication and Knowledge*, pp. 347–351, 2016, doi: 10.1109/ICTCK.2015.7582694.
- [68] A. Gupta and R. Kaushal, "Improving spam detection in online social networks," *2015 International Conference on Cognitive Computing and Information Processing*, 2015, pp. 1–6, doi: 10.1109/CCIP.2015.7100738.
- [69] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *Journal of Supercomputing*, vol. 76, no. 7, pp. 4802–4837, 2020, doi: 10.1007/s11227-018-2641-x.
- [70] R. Narayan, J. K. Rout, and S. K. Jena, "Review spam detection using semi-supervised technique," *Advances in Intelligent Systems and Computing*, vol. 519, pp. 281–286, 2018, doi: 10.1007/978-981-10-3376-6\_31.
- [71] D. H. Fusilier, R. G. Cabrera, M. M.-Y. -Gómez, and P. Rosso, "Using PU-learning to detect deceptive opinion spam," *WASSA 2013 - 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Proceedings*, pp. 38–45, 2013.
- [72] B. Bakhshandeh, "Instagram fake spammer genuine accounts," *kaggle.com*. [Online]. Available: <https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts>
- [73] N. Iqbal and P. Kumar, "Recent developments in soft computing based techniques for feature selection and disease classification," *Suranaree Journal of Science and Technology*, vol. 30, no. 2, 2023, doi: 10.55766/sujst-2023-02-e01872.

## BIOGRAPHIES OF AUTHORS



**Nidhi A. Patel**    is a Ph.D. Research Scholar at Gujarat Technological University in Ahmedabad, Gujarat, India. She is working as an Assistant Professor with the Department of Computer Engineering at Institute of Shree Swami Atmanand Saraswati Institute of Technology, Surat, India. She holds Bachelor of Engineering in Computer Engineering, Master of Engineering in Computer Engineering with a specialized in software engineering. Her research focuses on machine learning, artificial intelligence, deep learning, data mining, and big data. Her academic output includes a strong portfolio of journal articles, review articles, and conference papers. She can be contacted at email: [nidhi.patel0051@gmail.com](mailto:nidhi.patel0051@gmail.com).



**Dr. Nirali Nanavati**    is an accomplished Associate Professor in Computer Engineering at Sarvajani College of Engineering and Technology (SCET), Surat, India. She earned her doctoral degree from SVNIT, Surat, following an M.S. in Computer Science from New Jersey Institute of Technology (NJIT), Newark, USA. Prior to academia, she gained industry experience as a technical consultant with IBM France and Infosys Technologies Ltd. Her research focuses on privacy-preserving data mining, database systems, artificial intelligence, and machine learning. She contributed to patents and has published extensively journal articles, book chapters, and conference papers. Beyond research, she is a celebrated mentor. Under her guidance, student teams won first prizes in international competitions including the CSI-InApp 2021 "Medical image translation" and 2020's "Generative AI based project", along with awards at the Smart India Hackathon and Innovations 2019. She regularly delivers expert workshops on AI, machine learning, privacy, and data analytics. She can be contacted at email: [nirali.nanavati@scet.ac.in](mailto:nirali.nanavati@scet.ac.in).