A Comparative Study on Twitter Spam Detection Using Deep Learning Techniques

Pinnapureddy Manasa [a], Arun Malik [b], Isha Batra [c] and Ashish Kr. Luhach [d]

[a] *Computer Science and Engineering, Lovely Professional University, Jalandhar, India,*
[b] *Computer Science and Engineering, Lovely Professional University, Jalandhar, India,*
[c] *Computer Science and Engineering, Lovely Professional University, Jalandhar, India,*
[d] *The PNG University of Technology, Lae, Papua New Guinea.*

Abstract

In recent years, the communication field has redefined by the immense evolvement of Online Social Networks, particularly Twitter, Facebook, and LinkedIn. They facilitate the users to make friends of known persons or unknown people having a similar ideology, sharing of contents, etc. Among them, Twitter turns out to be a widely-regarded social networking platform that enables the users to share the ideas, discuss the social issues, read the news, and stay in touch with families and friends. Whereas, it has also targeted by the spammers, because of its immense recognition. Recent times, for the detection of spammers on Twitter, it has presented with the approach on the basis of Deep Learning . This paper focused on the comparisons of various prevailing techniques in twitter spam detection in which some methods provided good results and some degrades the functionality of the corresponding method.

Keywords

Spam detection, social networks, deep learning, Twitter Spam Detection

1. Introduction

A social network is a framework that allows the user through specific visual computer applications to share personal activities, history, interests, and real-life associations. A social network consists of primary user information, social interactions, additional information, such as academic and career information, etc. The following are considered to be the key aspects for the spam detection.

1.1. Online Social Networks (OSNs)

Nowadays, OSNs are the most significant across society, combining with people's everyday life and being unavoidable. Being a social structure, it comprises individual and organizations, which has elements known as ‘nodes’, whereas various kinds of interdependency have considered as the links that exist within nodes. In general, the social network includes two major parameters, namely nodes and links, where nodes relate to the content of relationships in connection with their presence, interest, background, such as trade, relatives, economic trade, shared heritage, transmission of disease, sexual relationships. Social networks have focused on web-based systems in recent days, so there is the possibility of improvements in the behaviour of nodes and the interaction behaviours of OSN users that arise from the mode of communication.

The information shared on OSNs has played a crucial role in the declarations issued by international organisations and institutions. In addition, the rapid development of OSNs helps a large number of

Algorithms, Computing and Mathematics Conference, August 19 – 20, 2021, Chennai, India.
EMAIL: reddymanasa24@gmail.com (Pinnapureddy Manasa)
ORCID: 0000-0002-3074-732X (Pinnapureddy Manasa)

individuals to exchange activities in the form of related posts across various OSNs due to different scopes of followers and there are several navigation keys on several websites, from which users can share the page with several OSNs. All of these make one entity's unique OSN accounts reveal extraordinary links. The user will remind the team of the relevant OSN through the feature, namely, report spam on Twitter and Facebook, to delete the posts suspected of someone being spam. The spam filter system, however, was developed independently by certain social networks with regard to the recognition of spam types. As these filters have predominantly increased spam recognition, there is still an enormous amount of spam in various OSNs at distinct intervals. For this reason, scientists and researchers have developed many common schemes with respect to spam detection and spam classification approaches, for that, it is important to confer the details of Twitter spam detection.

1.2. Twitter Spam Detection

OSNs, such as Twitter, Facebook, and LinkedIn, have been the most popular and highly recognized networking platforms in recent years. They spend much time on OSNs because it makes it easy for users to keep in contact with their family/friends, to post hobbies, events, current affairs. Twitter occupies a significant role among widespread micro blogging social network sites, as it enables the user to post a message in which the number of characters can be up to 280, often referred to as 'tweet.' Spammers take advantage of the implicit trust relationships among users to achieve malicious purposes. Consequently, numerous academics, as well as Twitter, have recommended different methods of spam detection to make twitter a spam-free platform. To be specific, as per the message linguistic analysis, the classifier has interpreted in terms of tweet text. However, this method is inadequate to conduct a set of comparable results due to the limited use of a single algorithm. Features can be identified and evaluated by a JSON object from Twitter's Streaming APIs where tweets with URL are considered. In addition, the feature attributes like time behavior, user profile, followers and following accounts, average time between tweets, social activities etc., are also considered which are listed in the backlist look up table [1]. From there the spam tweet and the non spam tweets are identified. Nevertheless, in terms of feature extraction and compromised accuracy, a few flaws still exist.

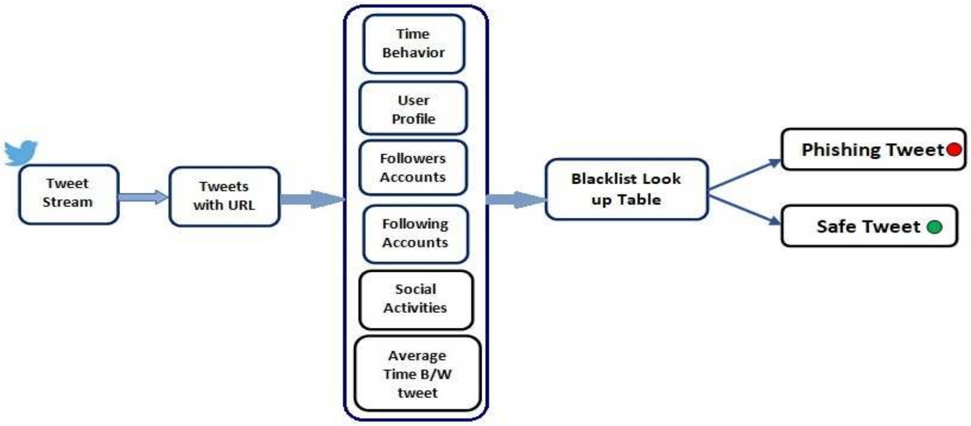


Figure 1: Spam detection framework based on Tweets with URL.

As per observation, Twitter spam would drift [2], and it is easy to fabricate the features, during the data collection process [2] [3]. Whereas, solely the average accuracy can get obtained around 85%, even by organizing the performance of prevailing research works. The further method encompasses blacklist services, in which the possibility of clicking the malicious URLs before they get blacklisted has depicted as more than 90% [4]. Meanwhile, Blacklisting methods have considered being extensively time-consuming, because of the participation of an individual for recognizing unsolicited information. These challenges drove the contributions of this study.

1.3. Twitter Spam Detection via Deep Learning

The framework of Twitter spam detection includes training on vector-based features by Word Vector techniques and the implementation of binary classifiers by multiple machine learning algorithms. Due to inadequacies of Natural Language Processing for traditional machine learning algorithm that uses raw strings, the deep learning has established for obtaining the capability of interpreting and scrutinizing the text through deep neural network accompanied by multiple layers [5]. Throughout the process of network, the output of every preceding layer becomes the input for the next layer in the network process. Specifically, in terms of language processing, the robust functionality was related to deep learning neural language strategies, in addition to distributed vectors trained under the Word Vector process. Unlike the conventional detection strategies that consider feature selection and generation, there is a need to implement the attributes in accordance with the tweet content with the support of Word2Vec.The workflow of an approach to differentiate the Twitter spam has depicted in Figure 4.

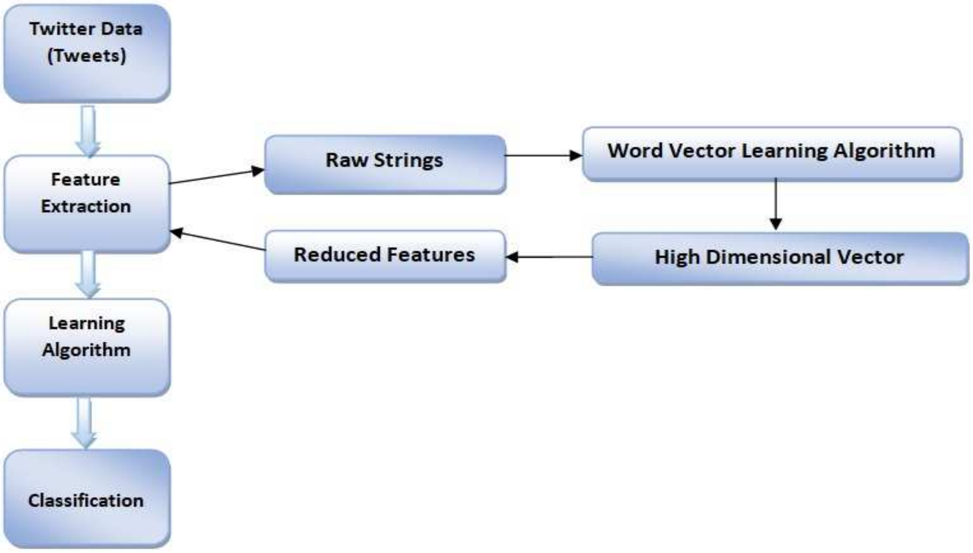


Figure 2: New twitter classification workflow based on Deep Learning.

1.4. Feature Selection Methods for Twitter Spam Detection

In many proposed methods, ML algorithms prove their importance. Once they are qualified, they get the opportunity to identify a user by considering a collection of features associated with the account and tweets of the corresponding user as a spammer or non-spammer [1]. There are five forms of this method as shown below table 1.

Table 1: Different forms of methods in spam detection using features selection

Methods	Feature Selection
Profile-Based	It reflects the user's account-based features
Content-Based	It quantifies the user's tweet-based features
Time-Based	It takes into account the speed of a Twitter user following others
Graph-Based	It takes into account features based on the Twitter user community
Automation-Based	takes into consideration the level of usage associated with the Twitter API

2. Related Work

This section consists of the reviews of the various articles on Twitter spam detection:

Sundararaja and Palanisamy [6] concentrated on features congregated into many groups such as linguistic features, contradictory features, and sentiment-based features. For every features category, ensemble is performed along with combination of features categories. Model training is achieved through various classifiers like Random Forest, KNN etc., and prediction on user's mood influencing the sarcasm and vice versa is done. Tweets before and after specific sarcastic kinds are attained. Thereby modelling the user emotion change through past tweet histories collection but there is a limitation that users are affected with their mood levels on the basis of sarcasm.

Cao et al [7] utilized special connections amid forwarding behavior as well as malicious URLs propagation by focussing on forwarding-based features. A comprehensive conventional URL feature sets investigation is primarily made. Then, the design combination between forwarding-based features and graph-based features is done for detection model training. Forwarding – based features are the most effective malicious detection approach offering higher accuracy rate besides lowest False Positive Rate (FPR). Investigation of forwarding-based features in OSNs is performed which is regarded as remarkable contribution for this research.

Wang et al [8] deliberated account-based, tweet-based, Natural Language Processing (NLP), and sentiment features for suggesting a spam detection technique. The peculiar features for spam detection are mean word length, automatically or manually formed sentiment lexicons, number of exclamation marks, question marks, maximum word length, capitalization words, and white spaces, Part Of Speech (POS) tags per tweet and profile name length.

Paudel et al [9] suggested a methodology for leveraging relationship amid named entities present in tweet content. In addition, probable spam detection is attained with the aid of document referenced by URL stated in the tweet. Unusual patterns might be found in data for exhibiting spammer activities through the combined integration of multiple, heterogeneous information into a single graph representation as stated by hypothesis. However, structural features fabrication is a challenging for spammers. The tweets collection along with documents referenced by URL in tweet allied to Twitter validate this methodology. The trending tweets anomalies can be efficiently detected through graph-based anomaly detection algorithms on data graph representation.

Chen et al [10] suggested a Semi-Supervised Clue Fusion (SSCF). It acquires a linear weighted function which identifies spammers with increased detection rate via multiple aspects, such as content, behavior, relationship, and interactions but the small size of primarily labeled instances. In future work, other type of fusion methods for combining the results of the clues and the majority functions will be more applicable for fusion. It may increases the detection rate of the spam.

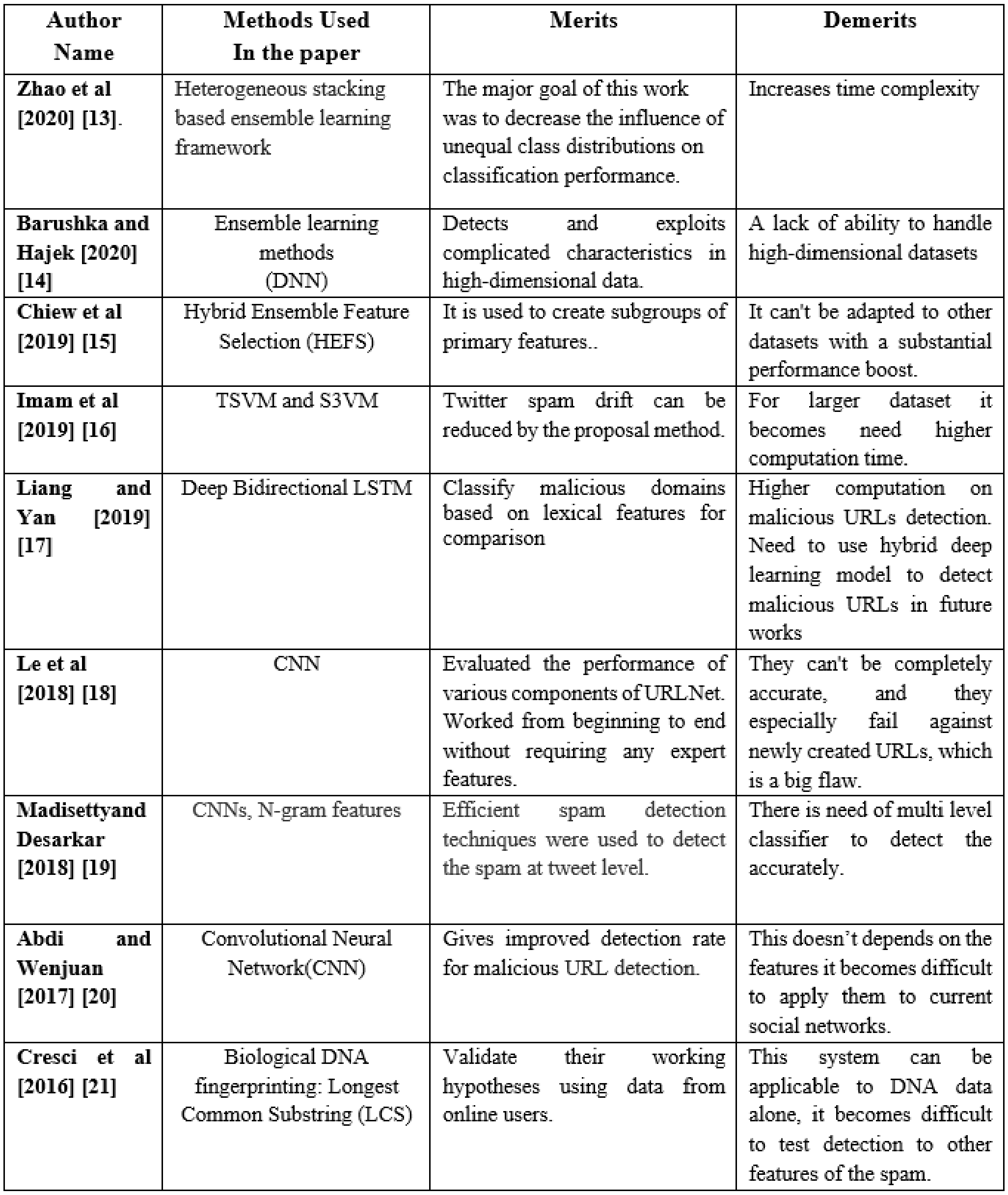
Le et al [11] proposed Convolutional Neural Networks to both characters along with URL String words for URL embedding learning in a joint optimized framework. Numerous rare words exist here various URLNet. Components need to be examined no expert features is necessitated but it is not capable for semantic or sequential patterns capturing. They might not be entirely exhaustive, and they fail against newly generated URLs.

Selvaganapathy et al [12] suggested stacked restricted Boltzmann machine, instance-bases learning with parameter k-nearest neighbor, Binary Relevance, Label Power set with SVM but the network training is rapid and detection with minimal false positives. In future large scale real time dataset can be deliberated in model developing fine tuning of parameters might be done as an optimization task.

3. Comparison Between Different Methods

This comparative study was carried out from the existing literature survey. This comparison is based on the methods, merits and demerits proposed.

Table 2: Demonstrates the Comparison Between the Different Methods of Twitter Spam Detection.



Author Name	Methods Used In the paper	Merits	Demerits
Zhao et al [2020] [13].	Heterogeneous stacking based ensemble learning framework	The major goal of this work was to decrease the influence of unequal class distributions on classification performance.	Increases time complexity
Barushka and Hajek [2020] [14]	Ensemble learning methods (DNN)	Detects and exploits complicated characteristics in high-dimensional data.	A lack of ability to handle high-dimensional datasets
Chiew et al [2019] [15]	Hybrid Ensemble Feature Selection (HEFS)	It is used to create subgroups of primary features..	It can't be adapted to other datasets with a substantial performance boost.
Imam et al [2019] [16]	TSVM and S3VM	Twitter spam drift can be reduced by the proposal method.	For larger dataset it becomes need higher computation time.
Liang and Yan [2019] [17]	Deep Bidirectional LSTM	Classify malicious domains based on lexical features for comparison	Higher computation on malicious URLs detection. Need to use hybrid deep learning model to detect malicious URLs in future works
Le et al [2018] [18]	CNN	Evaluated the performance of various components of URLNet. Worked from beginning to end without requiring any expert features.	They can't be completely accurate, and they especially fail against newly created URLs, which is a big flaw.
Madisettyand Desarkar [2018] [19]	CNNs, N-gram features	Efficient spam detection techniques were used to detect the spam at tweet level.	There is need of multi level classifier to detect the accurately.
Abdi and Wenjuan [2017] [20]	Convolutional Neural Network(CNN)	Gives improved detection rate for malicious URL detection.	This doesn't depends on the features it becomes difficult to apply them to current social networks.
Cresci et al [2016] [21]	Biological DNA fingerprinting: Longest Common Substring (LCS)	Validate their working hypotheses using data from online users.	This system can be applicable to DNA data alone, it becomes difficult to test detection to other features of the spam.

The following Figure 3 shows overall work flow of spam detection models, which accepts raw tweets data from users where the feature selection, feature extraction and data filttering is done based on the user based features,URL based features, Content based features and Graph based features. From this data spam detection models classifies whether the users tweet content is spammers tweet or non spammers tweet.

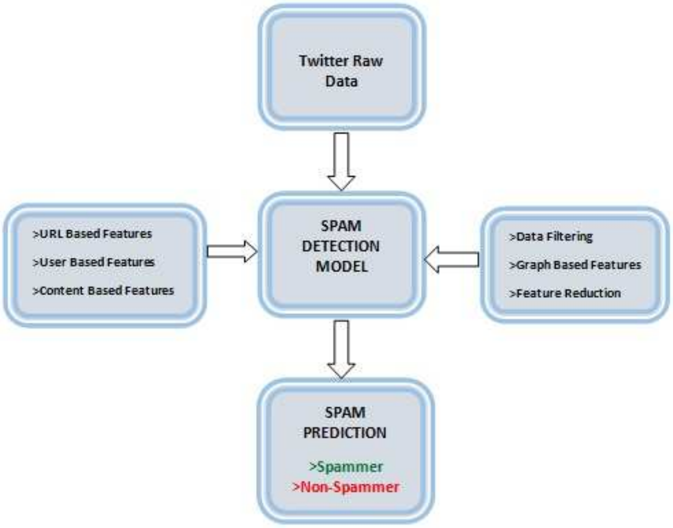


Figure 3: Overall Frame Work of Spam Detection Model

4. Critical Analysis

This section comprehensively scrutinizes the existing methods and strategies performed for detecting spam detection in twitter. Besides, summarizes the associated steps of machine-learning and deep learning based approaches that assist during the process of spam detection. The registered user's social activities have been tremendously increasing because of Twitter social network. The various activities of spammers by utilizing twitter are spreading malicious messages, post phishing links, network flooding with fake accounts, and indulging in further malicious activities. The distinct spam account identification plays a crucial step in the process of spammer's network detection involving in these activities. Several methods have been developed by the researchers for potential spam activities prevention.

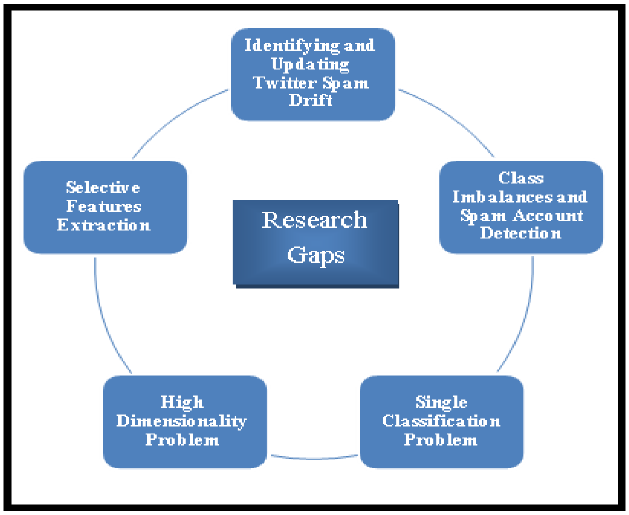

Identifying and Updating Twitter Spam Drift
Class Imbalances and Spam Account Detection
Single Classification Problem
High Dimensionality Problem
Selective Features Extraction
Research Gaps

Figure 4: Research Gaps

Twitter spam detection and social network activities has been accomplished by numerous methodologies. Various methodologies have been suggested by many researchers for recognizing spammers group but these methodologies provide solution for particular spammer category. By means of prevailing techniques, spammers' detection in twitter is again considered to great challenge. There is a need to develop a precise twitter spam detection system for exact spammer and non-spammers detection correctly with improved security of every twitter user. From the analysis of various papers it was observed that, there are still some challenges in spam detection in terms of high dimensionality, class imbalances, twitter spam drift, features selection and single classification problem where the researches need to focus.

5. Conclusion

This paper mainly focuses on the various comparison methods which are used for twitter spam detection. This comparison has been done on the basis of their research objectives, proposed techniques, merits, demerits and features used in spam detection. It is observed that some methods provided good results and some degrades the functionality. While dealing high dimensionality problems, satisfactory classification performances attainment via conventional classification approaches is a challenging task. The comparison reveals that the feature selection and feature decision are necessitated for assuring prediction proficiency for model training and there is lesser accuracy due to high dimensionality problem. It also observed that, Single Classifiers performs in a less superior manner when compared to ensemble learning methods which is revealed through earlier researches. Artificial intelligence technology is deployed by many spammers to falsify spammer's connection relationship and imitating normal users' social relationships. Hence effectual malicious accounts detection becomes a highly challenging task, where the researches need to be focused. Thus, Ensemble Feature Selection methods have to be focused by researchers for increasing the detection rate of the spam detection in the twitter dataset.

6. References

1. Mukunthan, B., and M. Arunkrishna. "Spam Detection and Spammer Behaviour Analysis in Twitter Using Content Based Filtering Approach." Journal of Physics: Conference Series. Vol. 1817. No. 1. IOP Publishing, 2021.
2. Jin, Xin, et al. "Socialspamguard: A data mining-based spam detection system for social media networks." Proceedings of the VLDB Endowment 4.12 (2011): 1458-1461.
3. Yang, Chao, Robert Harkreader, and Guofei Gu. "Empirical evaluation and new design for fighting evolving twitter spammers." IEEE Transactions on Information Forensics and Security 8.8 (2013): 1280-1293.
4. Chen, Chao, et al. "Asymmetric self-learning for tackling twitter spam drift." 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2015.
5. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning. nature 521 (7553), 436-444." Google Scholar Google Scholar Cross Ref Cross Ref (2015).
6. Sundararajan, Karthik, and Anandhakumar Palanisamy. "Multi-rule based ensemble feature selection model for sarcasm type detection in twitter." Computational intelligence and neuroscience 2020 (2020).
7. Cao, Jian, et al. "Detection of forwarding-based malicious URLs in online social networks." International Journal of Parallel Programming 44.1 (2016): 163-180.
8. Wang, Bo, et al. "Making the most of tweet-inherent features for social spam detection on Twitter." arXiv preprint arXiv:1503.07405 (2015).
9. Paudel, Ramesh, Prajjwal Kandel, and William Eberle. "Detecting spam tweets in trending topics using graph-based approach." Proceedings of the Future Technologies Conference. Springer, Cham, 2019.
10. Chen, Hao, et al. "Semi-supervised clue fusion for spammer detection in Sina Weibo." Information Fusion 44 (2018): 22-32.
11. Le, Hung, et al. "URLNet: Learning a URL representation with deep learning for malicious URL detection." arXiv preprint arXiv:1802.03162 (2018).
12. Selvaganapathy, ShymalaGowri, Mathappan Nivaashini, and HemaPriya Natarajan. "Deep belief network based detection and categorization of malicious URLs." Information Security Journal: A Global Perspective 27.3 (2018): 145-161.

13. Zhao, Chensu, et al. "A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data." Applied Sciences 10.3 (2020): 936.
14. Barushka, Aliaksandr, and Petr Hajek. "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks." Neural Computing and Applications 32.9 (2020): 4239-4257.
15. Chiew, Kang Leng, et al. "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system." Information Sciences 484 (2019): 153-166.
16. Imam, Niddal, Biju Issac, and Seibu Mary Jacob. "A semi-supervised learning approach for tackling Twitter spam drift." International Journal of Computational Intelligence and Applications 18.02 (2019): 1950010.
17. Liang, Yuchen, and Xiaodan Yan. "Using deep learning to detect malicious urls." 2019 IEEE International Conference on Energy Internet (ICEI). IEEE, 2019.
18. Li, Wentao, et al. "Lssl-ssd: Social spammer detection with laplacian score and semi-supervised learning." International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2016.
19. Madisetty, Sreekanth, and Maunendra Sankar Desarkar. "A neural network-based ensemble approach for spam detection in Twitter." IEEE Transactions on Computational Social Systems 5.4 (2018): 973-984.
20. Al-Milli, Nabeel, and Bassam H. Hammo. "A convolutional neural network model to detect illegitimate URLs." 2020 11th International Conference on Information and Communication Systems (ICICS). IEEE, 2020.
21. Cresci, Stefano, et al. "DNA-inspired online behavioral modeling and its application to spambot detection." IEEE Intelligent Systems 31.5 (2016): 58-64.