Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 124–132

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

A hybrid Data-Driven framework for Spam detection in Online Social Network

Chanchal Kumar[a],Taran Singh Bharti[a], Shiv Prakash[b]

[a]Department of Computer Science, Jamia Millia Islamia, New Delhi, India
[b]Department of Electronics and Communication, University of Allahabad, Prayagraj, India

Abstract

Twitter is one of the most prominent online social networks(OSN) used by celebrities, politicians, ordinary people, and organizations to enhance their popularity and brand value. The popularity of Twitter makes it a prime target for spammers to attack. Earlier, spammers used to focus on email and web-search engines; therefore, extensive research in spam detection for email and web search engines has been done. However, the humongous traction of OSN has taken away the focus of spammers from email and web search engines towards OSN and micro-blogging sites. The study is inspired by the need to verify the legitimacy of a profile on OSN to avoid any fake information or rumors on OSN. This paper modulates a hybrid framework for spam detection on Twitter. The sampling algorithm SMOTE-ENN combines SMOTE and Edited Nearest Neighbors (ENN) to generate balanced data that is further fed to various deep learning classification techniques to identify whether the tweet is spam or ham. The efficacy and performance of various state-of-the-art algorithms are evaluated through simulation and are compared through various performance metrics to determine the best spam detection framework for OSN.

Keywords: Online Social Networks; Twitter; Sampling; Security; Spam detection; Random Forest, KNN; Decision Tree; SMOTEENN

1. Introduction

This 21st century is the world of online social networking. In the age of globalization, people living in every nook and corner of the world would like to share every moment of their lives with their friends and family, making online social networks a popular and influential communication platform. Online Social Networks (OSN) such as LinkedIn, Twitter, Instagram, Flickr, Facebook, etc., have emerged as popular platforms that are used for personal and commercial reasons. As per the statistical report, there are 330 million monthly active users of Twitter worldwide[1]. Among various popular social media platforms, Twitter is a popular microblogging platforms. It has around 200 million users [2]. Due to the extensive use of online social networks to share information, a tremendous amount of

10.1016/j.procs.2022.12.408

data is readily available to other users on social networks. The freely available high volume of data is being analyzed by many organizations to enhance their acceptance and popularity to gain an edge over their competitors. Though social network sites have evident usefulness, spamming diminishes the credibility of information shared on social networks. A massive number of accounts are not genuine at all and exist only to exploit this readily available data and sometimes may harm society severely by spreading false or negative information or news.

Spam can be detected at two-levels, i.e., at the spammer level or tweet level. In this paper, we addressed the problem of Twitter spam detection. The simulation is carried out on several classification techniques such as Random Forest, decision tree, KNN, etc. have been implemented. Researchers have employed a variety of methodologies in developing a spam detection system that incorporates the machine learning concept. Spam accounts are being created on Twitter at an increasing rate. While researchers attempt to detect spam, spammers attempt to prevent detection[3]. Researchers are working on machine learning-based algorithms that detect Twitter spam accounts. As a result, researchers are still working on developing a 100 percent accurate detection system that would ensure that Twitter is entirely spam-free. An efficient spam detection framework only classifies the account as spam or ham but also improves privacy and guarantees protection for non-spam users.

The rest of this paper is structured as follows. The relevant literature and models are explored in Section II. The proposed framework is thoroughly explained in section III. In section IV, the outcomes of the adopted model are presented, and its performance is contrasted with the outcomes of its adopted counterparts. Section V finally ends the paper, and some outlines are offered for subsequent work.

2. Related Work

Spam detection is not an unexplored or new field of research. Researchers have worked at tweet-level spam, spam account, email spam, web spam, etc. However, over the past few years, the research interest of researchers has shifted toward spam detection in social networking, and various models to filter spam have been developed. Some of the studies referenced here are specific to spam detection in Twitter.

A framework for spam detection for Twitter named TwitterSpamDetector is developed in [4]. A total of 77033 real-time tweets by 50490 users were collected. TwitterJ4, an open-source java library, is used to provide Twitter Streaming API for collecting real-time tweets to create their own database. A Naïve Bayes classifier is used for training TwitterSpamDetector. The sensitivity and accuracy of TwitterSpamDetector were calculated as 0.913 and 0.943, respectively. In order to achieve accuracy, the author in [2] suggested modification in-stream clustering method by substituting Euclidean distance with a set of classifiers. A group of incremental Naive Bayes classifiers is trained to attain the boundary and mean of the micro-cluster. Here, only the mean of clusters is considered to calculate Euclidean distance. INB-DenStream framework based on DenStream is proposed in this paper. The proposed model was evaluated and compared in terms of recall, sensitivity, F1 measure, precision, purity, and computational with CluStream, DenStream, and StreamKM++.

A deep learning-based approach that uses both user's metadata and text in the tweet for spam detection is proposed in [5]. The text of the user's tweet and the metadata of the user account were considered in developing a combined classifier, while the text of the tweet is considered to develop a text-based classifier. To prove the efficiency of the proposed model, the author compared the results with five machine learning base approaches and two deep learning-based approaches and obtained the maximum accuracy of 93.12% and 99.68% for two different datasets. In [6], the author proposed a Trust-Distrust-Rank algorithm by integrating the trust and distrust of a webpage. Two matrices, T-Rank and D-Rank, are used to indicate each webpage's trustworthiness and untrustworthiness, respectively. Spam is demoted using a T-Rank score, whereas spam detection is indicated using a D-Rank score. A web page circulates more T-Rank/D-Rank ratio to other neighbors with higher T-Rank/D-Rank as compared to other neighboring with a lower value of this ratio. To identify the dubious community, a new differential trust propagation scheme that uses a random-walk-based community discovery method is proposed in [7]. These identified that dubious communities consist of almost all spam pages and are further used to reduce cross-community trust flow.

A supervised machine learning-based algorithm was proposed and tested on data set obtained from Sina Weibo [8]. A semi-supervised model has been proposed by applying a ranking scheme with supervised classification on social graphs. A rank scheme model is combined with supervised classification to develop a semi-supervised framework to detect social spammers [9]. Initially, a small number of labeled data is fed to the classifier, and then the ranking model is applied to spread trust and distrust in any social graph. In [10], the author considered spam from multi-angles. Much work to detect spam has been done through text analysis. However, the author considered multi-views of spammer's behavior to isolate URLs, Hashtags, and tagging features are also kept in consideration. A model known as (MVSD) Multiview Learning for Social-Spammer-Detection is prepared by integrating the classification model with new socialization terms and the information about networks and users gathered from multiple views. A three-stage framework is proposed to detect promotional campaigns and spam in [11]. The similarity between users' accounts that tweets URL for similar or identical purposes is computed. Shannon Information theory-based two methods are used to compute the similarity index. One method uses only the URL posted by the user, while the second one uses the URL as well as the time stamp of the post. A dense subgraph is extracted from account graphs as candidate campaigns. Several different features are also being introduced to classify spam campaigns from promotions.

Spam is not only a Twitter account but also can be a tweet by any user. To identify tweet spam, the author in [12] presented classical content-based techniques to filter spam tweets from Twitter. The well knows tool WEKA is used as a filtering library. Here, a combination of compression and machine learning algorithm filters undesired tweets [13]. In addition, some lightweight features have been identified for tweet representation. In [14], the author proposed S3D semi-supervised spam detection algorithm for tweet level detection. Spam detection and model update are the two modules that construct the algorithm. Spam detection is employed in real-time mode, whereas the module update module is run in batch mode. Dependable ham detector, Near-duplicate detector, Blacklisted domain detector, and multi-classifier-based detector are four lightweight detectors. The information in the detection module is updated in a batch mode based on previously recorded label tweets. To detect Twitter spam, the author in [15] devised a technique where the collection of tweets is pre-processed with the help of the word2Vec technique, and then a binary detection model is applied to categories into spam, and non-spam.

In [16], the author prepared a massive collection of 600 million live tweets from Twitter, further labeled as spam and non-spam. The prepared dataset has 12 lightweight extracted features over which the performance of six spam detection techniques was evaluated and proved that Random forest performed better than other algorithms with the highest F1 score. In order to understand the characteristics of spam profiles, authors in [17] developed a spam profile detection model by exhibiting publicly available features. To distinguish spam profiles, ten simple and binary features were identified with the help of 82 Twitter profiles. These binary features are selected by using the information gain and reliefF model. Once the feature selection procedure got completed, four classification algorithms were applied to the dataset. The results are compared, and it is observed that selecting the feature for the analysis of spam accounts is better than analyzing the simple language of the tweets. Similarly, in [18], the authors utilized the various models on real-time Twitter data and demonstrated that the Random forest gives better results in spam detection on Twitter.

In [19], 14 million real-time tweets were collected from some popular hashtags for analysis and named HSpam14 dataset. Their annotation process comprises four stages: (i) Heuristic base selection of spam tweets. (ii) clustering of near-duplicate tweets and labeling them. (iii) label the ham tweet after detection, (iv) apply Expectation-Maximization (EM) labeling on remaining tweets. A detailed review of various new spam detection models and methods is presented in [20]. The literature has been reviewed categories-wise, i.e., syntax analysis-based detection, feature analysis-based detection, and blacklisting of spam accounts on Twitter. The author also presented a comparative study of the performance of various models and methods.

In [21] author, a detection approach for fake and clone Twitter profiles has been proposed. A collection of criteria that distinguish between bogus and genuine profiles are used. Their approach uses two methods to identify profile cloning is proposed. One utilizes the C4.5 decision tree technique, and the other utilizes similarity measures, i.e., Similarity of Network relationships and Attributes.

In [22] The Improved Incremental Fuzzy-kernel-regularized Extreme Learning Machine (I2FELM), a technique built on regularised extreme-learning-machine, is utilized to identify Twitter spam reliably. The Author showed validation results that the suggested I2FELM could effectively distinguish between balanced and unbalanced datasets. It is also shown that by reducing the variable taken into account, the I2FELM can more accurately detect spam. In [23] author proposed encoder-decoder technique combined with the vectorizer-converter on the given tweets and their connected URL, resulting in the estimation of a similarity score between them, making it possible to determine whether the tweet indicated by the user is ham or spam

3. Proposed Model

3.1. Dataset

For any classification or detection model, it is essential to obtain a suitably labeled dataset. The Twitter spam dataset used here to evaluate the various state of art models is downloaded from NSCLab [24]. The original dataset has 12 features, and some features were derived from existing features which results in 15 features in the dataset. The dataset used here has 5k spam and 95k ham tweets. Table 1 records the features of the dataset that were utilized.

Table 1 Feature List

S. No	Features	Description	Feature Category
F1	Account_Age	How much old (in no of days) the account is till its most recent tweet	Profile-based
F2	Followers_Count	The followers of Twitter account	Profile based
F3	Following_Count	Total number of accounts he/she follows	Profile based
F4	Userfavourites_count	Total count of favorites	Content based
F5	Lists_count	No of the Lists created	Profile based
F6	Tweets_count	Total count of tweets of the user	Content based
F7	Retweets_count	Total count of retweets	Content based
F8	Hashtag_count	Total count of hashtags incorporated	Content based
F9	Usermention_count	Total count_of_user references	Content based
F10	Urls_count	Total count of URLs	Content based
F11	Char_count	Total count of characters in this tweet	Content based
F12	digits_count	Total count of digits in this tweet	Content based
F13	Is_Retweet	Tweet has re-tweet	Content based
F14	No_HasTags	No of HashTags	Content based
F15	F_ratio	Followers to Followings ratios	Profile based

3.2. The problem of Imbalance Data in the Twitter dataset

With regard to the problem of spam, no real social network dataset is uniformly disturbed. Any social network's live and authentic dataset will not be evenly distributed. For example, on Twitter [19], around 5% of all tweets are spam in nature. Data imbalance might make it challenging to train a data science model. When there are imbalance class problems, the model is trained predominantly on the majority class, which biases the model's forecast toward the majority class. Therefore, handling the problem of imbalance class is essential before proceeding on to the modeling phase and to obtain a suitable model for spam detection.

3.3. Proposed Spam Detection Framework

Figure 1 depicts the suggested spam detection framework's flow chart, which is discussed in the following sections.

3.3.1. Data Collection - A plethora of websites on the internet will give you a dataset. NSCLab[24], financed by initiatives from the Australian Research Council (ARC), Deakin University, and specific business partners, provides the dataset of tweets and related information. The dataset consists of ground truth labeled by Trend Micro's Web

Reputation Technology. The files are in ARFF format, which Weka can immediately open. Each line represents a single tweet. The feature values are in the remaining columns, and the tweet class (spammer or non-spammer) is in the last column. The Twitter spam dataset has features mentioned in Table 1. The dataset consists of 100 thousand (Number of Spam tweets: 5000 and Number of Ham tweets: 95000) records of a mix of spam and ham tweets.

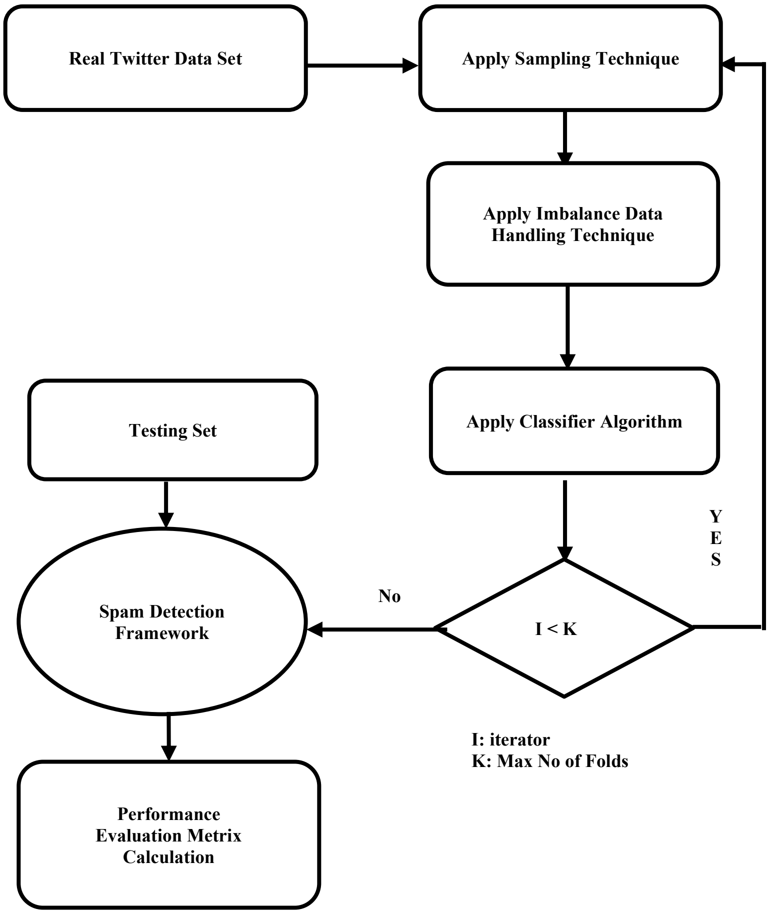

Figure 1 Flow chart of Spam detection Framework

3.3.2. *Dataset Imbalance Handling* - In this paper, to acquire better and concise results on imbalance dataset, a hybrid technique i.e., SMOETEEN is employed. SOMETEEN integrates oversampling and under sampling to achieve more thorough data cleaning.

3.3.3. Spam detection framework

Input: Initialization of parameters Twitter dataset TDB from NSCLab with T_L attributes and R_m Records
Output: Analysis of algorithm through m metrics on Twitter spam dataset
Assumption: Spam and Non-Spam ratio is 1:19 as per [19]

```
1. Initialization of the dataset
2. Apply cross-validation strategy Stratified10Fold
3. For each fold do
     i. Split data in training and testing
```

```
        b. For each fold do
            • Fed the data obtained in step ii to the classifier.
            • Compute metrics to analysis the performance of the classifier algorithm
            • Apply SMOTEENN (random_state=2) and obtain the balanced dataset
            • Compute metrics to analysis the performance of the classifier algorithm
            • Compare results computed in step iii and iv.
        End
    End
4. Compute means value of all the metrics values obtained.
5. Compare each classifier's findings to determine the best spam detection model.
```

SMOTEENN **(Wi, Tn, TDB, Hk)**

```
1. Randomly select data from the minority class.
2. Compute the distance between the data generated randomly and the k closest neighbors.
3. Add the result acquired by multiplying the difference by a random value (0-1) to the minority.
4. Repeat steps 2–3 until the required minority class proportion.
5. Calculate K to be the number of closest neighbors. If K cannot be determined, it will be three.
6. Find the K-nearest neighbor observations, then return the majority class.
7. If the observation's class and its K-nearest neighbor’s majority class are not the same, the observation and its
   K-nearest neighbors are removed from the dataset.
8. Repeat steps 2 and 3 until the expected results are received.
```

4. PERFORMANCE EVALUATION AND RESULT DISCUSSION

This section includes a comparative analysis of the proposed model with some classical spam detection models. The proposed algorithm is implemented in python. The following parameters and metrics are used for comparative analysis.

4.1. Simulation Environment

System with configuration mentioned below running Windows 10 python using google-colab.

Table 2 System Configuration

Processor	:	Intel(R)-Core (TM)-i7 8665U@1.90GHz ,2.11GHz
CPU Speed	:	
Logical Processor	:	8
Core	:	4
Ram	:	16

4.2. Criteria for evaluating performance

One of the most common snags for researchers when dealing with unbalanced datasets is the metrics they utilize to evaluate their model. In actuality, spam tweets account for about 5% of all tweets on Twitter [19]. Most data sets in prior studies were relatively uniformly distributed, i.e., the spam to non-spam ratio was nearly 1:1. However, the dataset used here has a 1:19 spam to the non-spam ratio, which makes the dataset imbalanced. It is essential to find the optimum value of precision and recall in such a classification problem. These metrics can be conjoined using F1 score.

Table 3 Performance Matrices used

Evaluation metrics		Formula
Accuracy (A)	:	$A = \frac{(TP + TN)}{(TP + TN + FP + FN)}$
Precision (P)/Positive Predictive Value	:	$P = \frac{TP}{TP + FP}$
Recall(R):	:	$R = \left(\frac{TP}{TP + FN}\right)$
F Score(F):	:	$F = 2 * P * \frac{R}{(P + R)}$
Area Under the Curve (AUC)	:	$ROC = \oint_{Roc} f(x)$

In a balanced classification problem, accuracy is one of the parameters to evaluate the efficiency of any model. It is the ratio of correctly predicted outcomes to all the samples taken into the consideration. In the case of the Imbalance classification problem, accuracy will not be the best metric to be used. This value is especially important for information retrieval when positive predictions are more important than negative predictions. It represents what percentage of all the positive predictions is genuinely positive. Alike precision, recall is another metric that is preferred in information retrieval when positive predictions are more important than negative predictions. It represents the ratio of predicted positives to total positives. In reference of Twitter spam detection, it's fine if some spam tweets go undetected (false negative), but it is not desirable to classify non-spam tweets as spam tweets (false positive). Therefore, False Positives should be kept to a minimum as much as possible. Precision takes precedence over recall in this situation.

4.3. Spam detection result of various models

The previous research reveals that spam identification in unbalanced datasets is the primary issue in spam identification which is more representative of real-world datasets. As a result, a hybrid strategy was presented in this paper to boost the spam detection rate. Table 3 summarises the findings of the random unbalanced dataset experiment. As shown in table 3, random forest precision is the highest, which is followed by the Decision Tree algorithm and KNN. When compared to other cutting-edge algorithms, Gaussian Nave Bayes performed the poorest in terms of accuracy and precision. .

Table 4 Performance matrix of Spam detection framework

S.No	Accuracy (A)	Precision (P)	Recall (R)	F Score(F)	ROC_AUC
RF without SMOTEENN	99.04	97.73	82.68	89.46	97.89
RF with SMOTEENN	99.26	99.49	99.07	99.28	99.97
DT without SMOTEENN	98.34	83.39	83.6	83.16	91.23
DT with SMOTEENN	98.29	97.92	98.79	98.36	98.27
KNN without SMOTEENN	97.22	75.9	65	69.95	86.5
KNN with SMOTEENN	98.76	97.91	99.76	98.83	99.81
GNB without SMOTEENN	19.78	5.75	97.82	10.87	82.33
GNB with SMOTEENN	59.92	56.47	98.13	71.68	85.46

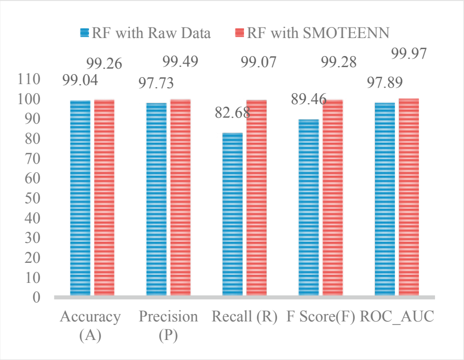

Figure 2

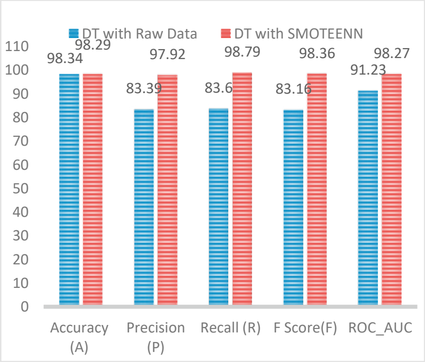

Figure 3

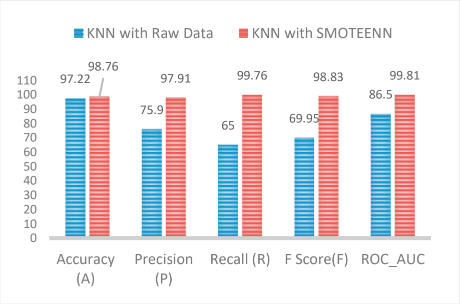

Figure 4

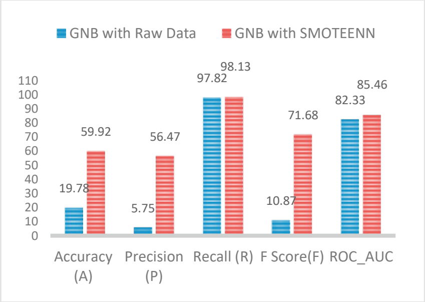

Figure 5

5. Conclusion and Future Scope

A hybrid Twitter spam detection framework was proposed. The proposed framework considered the problem of unbalanced data in real-time Twitter dataset and resolve it for better results. In the literature, many classification and regression techniques such as decision tree, SVM, LR, etc. exist however, most of the literature covered in related work was carried out without accounting for class imbalance, therefore, their findings are based on balanced datasets. In this paper, an empirical assessment of proposed detection framework is evaluated on a real organizations dataset with four deep learning algorithms, and it is concluded that the random forest algorithm with accuracy (99.26), recall (99.07) and precision (99.49) performed better as compared to other state-of-art algorithms while the performance of Naïve Bayes(gaussian) with accuracy (59.92), recall (98.13) and precision (56.47) performed the least.

In future work, the impact on the performance of various classifiers by increasing the no of spam tweets and tuning of hyperparameters of various algorithms can be observed. In the future, high-performance computing (HPC), deep learning (DL), machine learning (ML), approximation algorithms, and statistical testing approaches will be used to improve the efficiency of the solution to this problem.

References

[1] Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[2] Tajalizadeh H, Boostani R. A Novel Stream Clustering Framework for Spam Detection in Twitter. *IEEE Transactions on Computational Social Systems* 2019; 6: 525–534.

[3] Chen C, Zhang J, Xiang Y, et al. Asymmetric self-learning for tackling Twitter Spam Drift. In: *Proceedings - IEEE INFOCOM*. 2015. Epub ahead of print 2015. DOI: 10.1109/INFCOMW.2015.7179386.

[4] Kabakus AT, Kara R. "Twitterspamdetector" a spam detection framework for twitter. *International Journal of Knowledge and Systems Science* 2019; 10: 1–14.

[5] Alom Z, Carminati B, Ferrari E. A deep learning model for Twitter spam detection. *Online Social Networks and Media* 2020; 18: 100079.

[6] Zhang X, Wang Y, Mou N, et al. Propagating both trust and distrust with target differentiation for combating link-based Web spam.

ACM Transactions on the Web; 8. Epub ahead of print 2014. DOI: 10.1145/2628440.

[7] Li J, Sun Y. Web-age information management: 16th international conference, WAIM 2015 Qingdao, China, june 8–10, 2015 proceedings. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2015; 9098: 452–453.

[8] Madisetty S, Desarkar MS. A Neural Network-Based Ensemble Approach for Spam Detection in Twitter. *IEEE Transactions on Computational Social Systems* 2018; 5: 973–984.

[9] Li Z, Zhang X, Shen H, et al. A semi-supervised framework for social spammer detection. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2015. Epub ahead of print 2015. DOI: 10.1007/978-3-319-18032-8_14.

[10] Shen H, Ma F, Zhang X, et al. Discovering social spammers from multiple views. *Neurocomputing* 2017; 225: 49–57.

[11] Zhang X, Li Z, Zhu S, et al. Detecting spam and promoting campaigns in twitter. *ACM Transactions on the Web*; 10. Epub ahead of print 2016. DOI: 10.1145/2846102.

[12] Igor Santos, Igor Mi˜nambres-Marcos, Carlos Laorden PG-G, Aitor Santamar´ıa-Ibirika and PGB. Twitter Content-Based Spam Filtering. *Proc Int Joint Conf SOCO'13-CISIS'13-ICEUTE'13 Cham, Switzerland: Springer* 2014; 449–458.

[13] Chen C, Zhang J, Xie Y, et al. A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection. *IEEE Transactions on Computational Social Systems*. Epub ahead of print 2015. DOI: 10.1109/TCSS.2016.2516039.

[14] Sedhai S, Sun A. Semi-Supervised Spam Detection in Twitter Stream. *IEEE Transactions on Computational Social Systems* 2018; 5: 169–175.

[15] Wu T, Liu S, Zhang J, et al. Twitter spam detection based on deep learning. *ACM International Conference Proceeding Series*. Epub ahead of print 2017. DOI: 10.1145/3014812.3014815.

[16] Chen C, Zhang J, Xiang Y, et al. 6 Million Spam Tweets : A Large Ground Truth for Timely Twitter Spam Detection. 2015; 7065–7070.

[17] Al-Zoubi AM, Alqatawna J, Faris H. Spam profile detection in social networks based on public features. *2017 8th International Conference on Information and Communication Systems, ICICS 2017* 2017; 130–135.

[18] Ameen AK, Kaya B. Applied Mathematics , Electronics and Computers Detecting Spammers in Twitter Network. Epub ahead of print 2017. DOI: 10.18100/ijamec.2017436078.

[19] Sedhai S, Sun A. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2015; 223–232.

[20] Wu T, Wen S, Xiang Y, et al. Twitter spam detection: Survey of new approaches and comparative study. *Computers and Security* 2018; 76: 265–284.

[21] Sowmya P, Chatterjee M. Detection of Fake and Clone accounts in Twitter using Classification and Distance Measure Algorithms. *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020* 2020; 67–70.

[22] Zhang Z, Hou R, Yang J. Detection of Social Network Spam Based on Improved Extreme Learning Machine. *IEEE Access* 2020; 8: 112003–112014.

[23] Badola K, Gupta M. Twitter Spam Detection Using Natural Language Processing by Encoder Decoder Model. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021* 2021; 402–405.

[24] DataSet, http://nsclab.org/nsclab/resources/index.html.