International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Aug-2022)

https://dx.doi.org/10.12785/ijcds/120135

Spam Message Detection: A Review

S Aditya Chaturvedi[1] and Lalit Purohit[1]

[1]

Information Technology, Shri G S Institute of Technology and Science, Indore (M.P), India

Received 25 Jan. 2022, Revised 16 May 2022, Accepted 16 Jul. 2022, Published 1 Aug. 2022

Abstract: A Message is an information exchanged for personal or business purposes. These messages are generally targeted by spammers, resulting in forfeits in financial or monetary. Spam messages have grown significantly in different fields. Various machine learning based techniques have been used in the past for the detection of spam. A very few review works are available on spam detection techniques in the field of SMS, Email, Twitter and Online reviews. However, these studies have limitations of study of limited techniques from machine learning fields only. Also, an in-depth evaluation of performance for each of the suggested techniques are missing. In this paper, a detailed review of spam message detection techniques in five domains - SMS, Email, Twitter, Instagram and Online Reviews is done. Based on the reviews of state-of-the-art in the five domains, a generalized model for spam message detection is perceived and presented. Additionally, this paper provides a thorough review of the past researching the domain and detailed analysis is presented. The paper concluded with the future trends which can be used for message spam detection in near future.

Keywords: Detection, Email Spam, SMS Spam, Instagram Spam, Online Review Spam, Twitter Spam Detection, Natural Language Processing,Machine Learning

1. Introduction

Online social media users have increased enormously since the past few years. There are currently 4.66 billion active users over the Internet across the globe [1]. With this increasing number of active users, the malicious activities and spam are also increasing. Spam refers to the kind of call/message/information/comment etc. created to result in loss of the recipient [2]. Some losses such as monetary loss, data loss, compromising identity, time and infection of devices can occur due to spam messages. Spammers spam the users in various fields including Health Sector, finance marketing, promotional products, online media, social media, spam calls, education spam etc. [3]. With the passage of time, various platforms emerged which became victims of spam. Many online media platforms were affected by spammers. As they all have one thing in common i.e., "Message". They all contain a message either in the form of Electronic Mail (Email), tweets, posts, reviews, or comments. Therefore, spam detection is an important area of study.

The message exchange is done frequently in the five major domains- Short Service Message (SMS), Email, Twitter, Instagram and Online reviews [4] [5] [6] [7] [8]. SMS is the most fundamental and reliable source of messaging without using the Internet. For the past decade, various users and organizations have used SMS as a medium for providing high priority messages including bank transactions, one time passwords, etc. [9] [10]. Spammers spam the users by various techniques using the SMS resulting in financial, identity and monetary loss [11]. Around 5 billion users use SMS for exchange of information [12]. Major spammers use this platform to perform spam either by sending malicious URL's or by sending false information. E-Mail has become a common and most reliable platform for the exchange of information over the internet [13]. According to the report [14], there are 4147 million E-Mail users around the globe [14]. Currently, there are 1.8 billion users in the most widely used Email platform - Gmail [15] and over 3 million E-Mail exchanges per second [5]. E-mail is the major spamming platform and various ways are used to do so. Total of 122 billion spam Emails exchanged daily [16]. Across the globe around 400 million users are using Twitter platform [17]. There are currently 200 million active users on Twitter. It holds 8% of the Social Media market across the globe [17]. Over 500 million tweets are shared on the platform daily [18]. Twitter is among the most used social media platforms for the purpose of exchanging information by various officials, celebrities, and users. Spammers use phishing links, fake tweets with false information to distract users' attention [6]. India holds the maximum number of Instagram users around the globe. There are around 1074 million active users of Instagram all over the world out of which around 180 million users of Instagram are in India [19]. Instagram is the most widely used platform by users for business and personal means. Users add posts, comments and pictures of the moment of their life. Also, it is used to promote

E-mail address: gsit20r05@sgsitsindore.in, purohit.lalit@gmail.com

the business too. Due to its high popularity, this platform also became the home of various malicious activities [17]. Reviews play an important role if you are either buying something or booking a restaurant or hotel, etc. over the Internet. Online Reviews are very effective for e-commerce and business as the majority of users read reviews. Fake Reviews can result in decreasing sales and values too. Spammers use various techniques and flood the reviews section with fake/bad reviews resulting in a decrease in sales. According to the report, 93% of customers analyze reviews before buying products over e-commerce websites [8]. Around 600 million reviews including hotels reviews, airlines reviews, cabs reviews and many more is generated monthly [20]

Hence, these five major fields carry their own importance and spammers target them frequently. Supervised learning and unsupervised learning are two major models used for spam message detection. Supervised learning-based classification algorithms find their application in SMS [2] [9] [21], Email [3] [10] [13] and Instagram [7] [19] [22]. However, unsupervised learning-based clustering algorithms are extremely used for spam tweets detection. Hybrid approaches are also considered for spam message detection where classification/ clustering-based algorithms are combined with other approaches. Hybrid models are used for E-mail [5] [23] and Online review [7] [24] spam detection. In [25] and [26], for Email spam detection a detailed comparison is made using various word embedding techniques along both machine and deep learning algorithms using Apache Spam Assassin. A detailed analysis is done using various evaluation parameters and it's concluded that Deep learning-based algorithms using appropriate word embedding techniques perform better than traditional machine learning based algorithms for spam Email detection [26].

There are survey works in the domain of SMS spam detection [27], Email spam detection [28], Twitter spam detection [29] and spam detection in online reviews [30] are available separately. In [31], a comprehensive review of spam detection is presented. Variety of domains where spam is found is included. The review work included spam detection work done using hybrid approaches only and analysis is done based upon accuracy. A comprehensive review for SMS spam detection based on deep learning methods is presented in [27]. The review work highlights the application of convolutional neural network and recurrent neural network for SMS spam detection. But, the coverage of review work is very limited. In another review work [28], spam classification in the domain of Email is presented. Review of total 12 papers on Email spam classification is done. The presented review is limited to the domain of Email and performance comparison among various techniques are missing. In [3], a review on spam tweets detection using machine learning techniques is presented. The work includes review on four parameters – method used for spam message detection, feature selection / extraction, data-set used and result evaluation on the accuracy. A common model was presented which displays the steps involved in twitter spam detection. A survey for spam detection in the domain of online review's is presented in [30]. Spam online reviews are categorized in Type-I, II III reviews. However, the presented review is limited to data-sets used, techniques used and evaluation metrics. features and computational complexities are not considered.

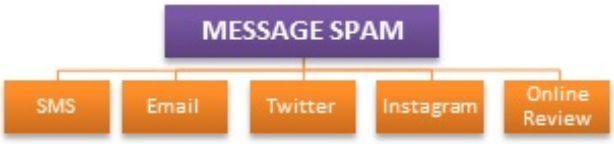


Figure 1. Spam Message Category

The existing review works focus majorly on algorithms belonging to one domain e.g. any one from Email, SMS, Twitter etc. Review work for spam detection techniques in all the domains (SMS, Email, twitter, online reviews and Instagram) is missing. Also, it will be interesting to review various spam detection techniques with reference to feature selection and feature extraction. Also, to the best of our knowledge, review work in the domain of Instagram spam detection is not available. Thus, looking into the uses of messages for information exchange in these five domains and limitations of the present review works on spam message detection, a review work is carried out. Figure 1 shows five domains considered for the review of existing state-of-the-art for spam message detection. In this study, an in-depth analysis on spam message detection is presented. Based on the observation on the process used for spam detection in each of the five domains, a model is presented to understand the generalized steps for spam message detection. Also, the review is carried out on efficiency of the proposed approaches, learning model used, major features used by each approach, etc. by comparing and contrasting their works on various parameters. The further road map of the paper is as follows: In Section 2, an observed spam detection model is presented with the literature review on existing works in the domain of spam message detection. In Section 3 analysis on the existing works and a summary of various subcategories along with the results of study are presented with discussion In Section 4 conclusions drawn from the study are discussed with the future works

2. LITERATURE REVIEW

The past works on spam message detection used a well known data-set for performing experimentation [2] [3] [7] [10]. Each of the work also uses text pre-processing, feature extraction/selection, classification/clustering / other methods, for identification of Spam and Ham data, and Performance evaluation followed by comparison. In this section, similar observations are drawn in all five domains (SMS, Email, Twitter, Instagram and Online Reviews). Based on this observation; a model is presented to elaborate the generalized set of steps followed for spam message

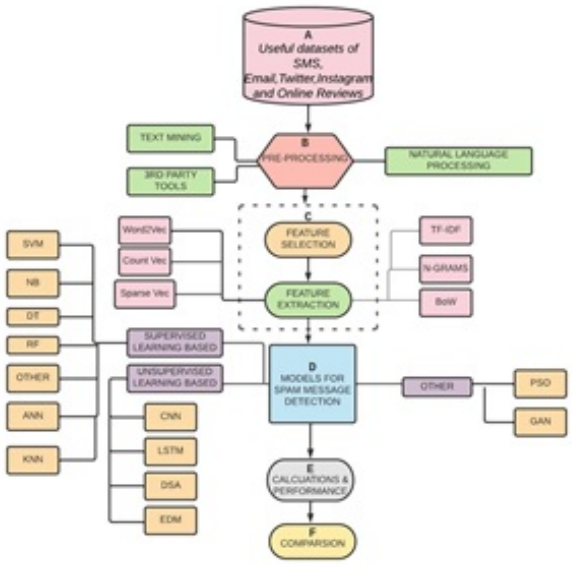


Figure 2. Observed Model

detection as shown in Figure 2. Further, literature review is presented for spam message detection. Table II presents a literature review based on model, methods and domain used for spam message detection.

It can be observed from Figure 2 that the observed model for spam message detection is divided into 6 steps- Step A- Data-sets used for spam message detection in SMS,Email,Twitter,Instagram and Online Reviews. Step B- After the data-set selection, text pre -processing is required to remove the unwanted data from the data-sets. Step C- Feature selection/extraction is the major part after the pre-processing of the data-set. Various techniques are available for feature selection/extraction as it affects the overall performance of the model. Step D- Machine and Deep learning models can be supervised or unsupervised. Different algorithms lie under them are used to detect spam and ham messages.Some other approaches are also used for spam message detection. Step E- After the implementation of spam message detection model, calculation and performance is done involving various evaluation parameters like-accuracy, precision, F-score and recall. Step F- Comparison is made between different research works based on evaluation parameters. Section 2 is now updated to include these details

A. *Useful datasets of SMS, Email,Twitter,Instagram and Online Reviews*

Numerous data-sets are available from various sources to perform experimentation related to spam message detection. There are open-source data sets available from various websites and platforms. Table I represents the review of various datasets used for spam message detection in detail. Few major platforms/websites where datasets for spam message detection can be found are as follows:

1) *Kaggle*

It is a well-known platform where data-sets are easily available. Thousands of data-sets are available for ready-to-use. From Kaggle, various data-sets available such as SMS spam dataset [2], Email spam dataset [13], Twitter Spam dataset [32], Hotel Reviews dataset for online reviews [24], etc.

2) *UCI*

It is a repository for plenty of data-sets related to SMS [9] [10], Email [5], Twitter [33] and Online Reviews [8] Using these datasets, machine learning related experiments for spam message detection are carried out mostly by students and research scholars [9], [10].

3) *Manual*

With the passage of time, data-sets become older, thus, various authors prefer to collect data-set own their own. The data-set available and the manually collected latest data-set have a vast difference [34]. In the study carried out by [6] and [35], data-set for Twitter spam detection and for Instagram [7] [19] [22] [34] spam message detection, data-set collected manually. Similar observations are found for manual data-set collection for spam detection on Online reviews [8] [36].

B. *Text Pre-processing*

Removal of unwanted noise from the data is certainly necessary. After removal of noise, the filtered data improves the efficiency of model [36]. In order to remove unwanted noise from the data-set, pre-processing techniques are used in the past [13] [21] [24] [37]. Mainly three pre-processing techniques were reported (1) Natural Language Processing [9] [35] [38] , (2) Text Mining [6] [21], (3) Third Party Applications [4] [5].

1) *Natural Language Processing*

The machine is not itself capable enough to understand the text like humans do. Here the need for pre-processing of data arose. This results in the removal of unwanted or useless data. In the past, Natural Language Processing (NLP) was used as a pre-processing technique to convert the text into the real vectors [9] [35] [38]. Segmentation, Tokenization and Stops Words Removal steps are generally considered for spam message detection [3] [8] [9] [10] [13] [21] [39]. Also Stemming and Lemmatization generally used in Twitter and Instagram spam detection [19] [32]. While all the major steps of NLP are used in SMS, Email and Twitter spam detection [35] [38] [40].

2) *Text Mining*

To identify the hidden patterns and information, the transformation of the unstructured data into the structured data is needed. Text Mining aims to fetch the high quality of data from the text provided and it separates the unwanted data and identifies the useful words from the data-set which are used later for classification of spam or ham [41]. It is used for SMS [21] and Email spam detection [41].

TABLE I. REVIEW OF DATASETS USED FOR SPAM MESSAGE DETECTION

REF. NO	DOMAIN	DATA SET			
		Name	*Quantity*	*Total Ham*	*Total Spam*
[2]	Da	Kaggle	5574	4827	747
[9]	Da	UCI	5572	-	-
[10]	Da	UCI	5574	4827	747
[11]	Da	Kaggle	5574	4827	747
[20]	Da	Kaggle, SMS Spam V.1	5574 11968	4827	747
[50]	Da	UCI	5574	4827	747
[3]	Db	Spam Assassin	4950	2551	2399
[5]	Db	Google/ Yahoo	2200	-	-
[13]	Db	Kaggle Enron	5574 30207	4827 16545	747 13662
[22]	Db	Ling Spam	1000	500	500
[45]	Db	Enron	5500	1500	4000
[49]	Db	Manual	800	400	400
[6]	Dc	Manual	2483	2334	149
[28]	Dc	Kaggle	11968	-	-
[31]	Dc	Manual	70000	62000	8000
[34]	Dc	Manual	467480	-	-
[39]	Dc	Manual	10000	-	-
[7]	Dd	Manual	24602	22743	1859
[19]	Dd	Manual	1400	700	700
[21]	Dd	Manual	2600	1875	625
[30]	Dd	Manual	21099	10609	10490
[8]	De	Manual	45531	1650	350
[23]	De	Kaggle	1600	800	800
[32]	De	OTT YELP	1600 2000	800	800
[33]	De	Manual	1600	800	800
[52]	De	Cornell University	800	400	400

3) Tools and Applications

There are various tools and applications used for pre-processing of data. Rapid Miner [6] and Minhash [34] are among them. They found their uses in Twitter and Instagram spam detection respectively.

C. Feature Selection/Extraction

After pre-processing of data is completed, sometimes many features are associated with the data to describe them like length, count, hyperlinks, special characters, common-uncommon words etc. From among these, only useful features are to be used for caring out further processing of the data. Therefore, the useful features are required to be extracted from the data set. Feature selection and extraction is performed on the pre-processed data-set. From the available set of features, few important features are selected. For SMS spam detection, unique/ common words and length/frequency feature is generally considered [9] [10]. In case of Email spam detection, length/frequency, hyperlinks, sender's address, unique words are the commonly used features [3] [13] [23] [42]. For spam tweet

detection, features mentioned for SMS and Email category along with number of tags/likes, retweets, number of followers/following observed are used [6] [35] [37] [43]. Hyperlinks, no of tags, no of followers/following, unique words, no of likes/comments and length/frequency are features taken into consideration for spam messages detection over Instagram [7] [34]. Features for detection of spam Online Reviews includes length/frequency and unique-common words [8] [24] [39]. Feature Extraction is done for converting the data-set in to real time matrix which is thus the input for the algorithm for classification of spam and ham values [11] [24] [43]. Bag of words [3] [7], N-grams [39] [36], Word2Vec [7] [39], TFIDF [2] [9] [10] along with Hashing [3] [24], GloVE [21] [32] [43], Word cloud [11], Genism Package [13], Confusion matrix [8] [19] are used in the past as feature extraction techniques used for spam message detection.

D. Model used for Spam Detection

The models used for spam detection lie under three categories (1) Supervised Learning, (2) Unsupervised Learning Based and (3) Other Approaches.

1) Supervised Learning

Supervised learning can be defined as the algorithm which learns from a pre-labelled data set and predicts the classification of unlabeled data [44]. This type of learning uses the training data set to predict the output of new input. [9] [10]. For message spam detection, a classification model is prepared with an algorithm to classify spam or ham. Performance evaluation is done to analyze the accuracy, F-Score and other values and then compared with the other models [45]. Few important approaches based upon supervised learning used for spam message detection discussed below:

a) Naïve Bayes(NB): This method is based on the "Bayes Rule" which determines the occurrence probability of an event based upon the previous knowledge. In a feature vector space, the NB method can easily classify high dimensional data points [2]. The probability distribution is calculated for tokens with selected features using the NB method to classify messages (SMS/Email/Tweet/Comments/Review) as ham or spam [13] [28]. The classifier chooses the class having the highest value of posterior probability. If Spam Messages (SM) and Ham Messages (HM) and Message (M) are three parameters then following NB method [13] [23]. Equation (1),(2),(3),(4),(5) and (6) demonstrate its concept for spam message detection:

1) Prior probability of Ham message Its is defined as the ratio of total ham messages to the to number of messages and can be formulated as show in (1):

$$= "Total\ no.\ of\ HM"/"Total\ M" \tag{1}$$

2) Prior probability of Spam message Its is defined as the ratio of total spam messages to the to number of messages and can be defined as show in (2):

$$= "Total\ no.\ of\ spams"/"Total\ M" \tag{2}$$

3) Likelihood of N- Message given is Ham Its is defined as total number of ham messages in the vicinity of n-message to the total number of ham messages and can be formulated as show in (3):

$$= \frac{'No.\ of\ HM\ in\ the\ vicinity\ of\ n-Message'}{Total\ no.\ of\ HM} \tag{3}$$

4) Likelihood of N-Message given is Spam Its is defined as total number of spam messages in the vicinity of n-message to the total number of spam messages and can be defined as show in (4):

$$= \frac{No.\ of\ SM\ in\ the\ vicinity\ of\ n-Message}{Total\ no.\ of\ SM} \tag{4}$$

5) Posterior probability of n-Message being Ham It can be defined as the multiplication of prior probability of ham messages to the likelihood of n-message is ham and can be formulated as displayed in (5):

$$= Equation(1) * Equation(3) \tag{5}$$

6) Posterior probability of n-Message being Spam It can be defined as the multiplication of prior probability of spam messages to the likelihood of n-message is spam and can be formulated as displayed in (6):

$$Recall = (TP)/(TP+FN) = Equation(2)*Equation(4) \tag{6}$$

Finally, we classify n-Message as ham or spam based on whose class membership has a largest posterior probability.

b) Support Vector Machine (SVM): It is another learning model widely used for spam message classification. Firstly, SVM is trained with 'n' data points which are already labeled as spam or ham. Then visualization of the data points in n-dimensional space is done for spam and ham classes. After that SVM breaks the space into regions and locates the new data-point using regional classification. A hyperplane separating different classes (ham or spam) is obtained [28] [35].

c) Random Forest (RF): RF consists of many individual trees. Each tree here votes for classification result of the data-set given and the result with most votes is considered. For spam message detection first of all, categories and samples are taken from the training data-set matrix. For building forest, node values are calculated using classifier and threshold function. After getting the node value, the left and the right node is determined using the split function. This process of calculating node value and splitting go on until a single sample remains and the category of the remaining sample is determined from known categories

(such as spam or ham). One tree is completed once all nodes are gone through splitting process. All trees are generated similarly. When all the trees are generated, the forest is built successfully. F-value function is used for classification of spam or ham messages (SMS/ Email /Tweet/ Comments / Review) [9] [10].

d) Recurrent Neural Networks(RNN): RNN finds several features/characteristics from the given spam data-set. Patterns are extracted from the features/characteristics and prediction is made for either spam or ham. Basically, it has 2 main layers- input layer and output layer. Several hidden layers exists between the two to carrying information from the previous layer. Biases and weights are assigned to each layer and output is generated.

2) Unsupervised Learning

Unlike the supervised learning approach this model does not require a training data-set. Different patterns, features are identified from the data-set itself [10]. Message spam detection using unsupervised learning involves clustering for detecting spam messages. Some approaches based upon Unsupervised Learning are discussed below:

a) Encoder-Decoder Model (EDM): The EDM method is based on Recurrent Neural Networks (RNN) to predict the result of sequence-to-sequence problems. EDM is mainly used to summarize the large vector data spam message detection into smaller vectorized data. Here, the encoder is fed with the large vectorized data as input and summarized vectorized data is obtained as output from the decoder. After that similarity score is calculated on vectorized result and a message is categorized as spam if similarity score is less than 75 % and ham if similarity score is greater than 75 % [43]

b) K-Means Clustering: This is the most common unsupervised learning approach used for spam message detection. Here, each cluster has the centroid and repeated calculations are done to optimize positions. In spam message detection it is used to group similar/duplicate messages (SMS/Emails/Tweets /Posts/Reviews) into the same cluster and marked as ham or spam. Then classification is done using a supervised learning model to finally categorize the data to spam or ham [34].

3) Other Approaches

There are various other approaches used for spam detection. These approaches are not commonly or widely used but effective. These are used to classify the ham and spam messages but also to optimize the existing models. Some approaches are:

a) Particle Swarm Optimization (PSO): PSO is the optimization technique based on swarm intelligence. Just Like the bird searching for food randomly can optimize her search if she works with the flock, PSO consists of particles searching for the best optimal solution in space [46]. PSO is used in hybrid approaches for spam message detection as a technique for feature selection [47] while as a classifier in [23] and [48].

b) Generative Adversarial Network (GAN): GAN consists of 2 main parts: Generator and Discriminator. The Generator create fake messages(SMS/Email /Tweets/ Comments/ Review) using the noise from the dataset and those fake messages are provided to the discriminator with the real data-set [49]. Task of the discriminator is to predict real and fake values correctly. Discriminator keep improving its ability to separate values until it predicts maximum fake messages from generator correctly [45].

E. Calculation and performance evaluation

There can be various parameters for calculation and performance evaluation for the method used for spam message detection. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are basic parameters used for calculation [3] [21] [22] [50]. TP [13] is defined as number of spam messages (SMS/Emails /Tweets /Comments /Reviews) identified correctly as spam while TN refers to the number of ham messages correctly detected as ham. FP can be defined as the number of ham messages classified as spam while FN refers to the number of spam messages classified as ham [2] [11].

Four major parameters used for calculation are as follows:

1) Recall

It is defined as the actual spam messages detection probability [2]. Recall can be formulated as shown in (7):

$$Recall = (TP)/(TP + FN) \tag{7}$$

2) Precision

It is defined as the probability for true value of spam messages detection [10]. Precision can be formulated as defined in (8):

$$Precision = (TP)/(TP + FP) \tag{8}$$

3) Accuracy

The ratio of true values of spam messages detected with all four values is known as accuracy [11].It can be formulated using (9):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{9}$$

4) F Measure

It defines the overall performance of the method [13]. It can be defined using (10)

$$FMeasure = 2 * (Precision)/(Precision + Recall) \tag{10}$$

The performance evaluation is done based upon the obtained result of calculation from (7) [2], (8) [10], (9) [11] and (10) [13].

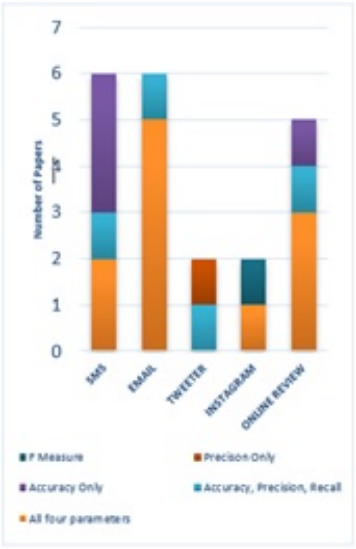


Figure 3. Various parameters used for of evaluating performance of spam message detection techniques

F. Comparison

Lastly, among various parameters involved in calculation, few or all parameters are considered and comparison is done based on models and methods used for spam message detection [9] [13] [22]. Figure 3 represents the use of various parameters including accuracy, precision, recall, F score used together or used few in five domains for evaluating performance of existing spam message detection techniques. Total of 21 papers are considered for performing this study. On X-axis of Figure 3, five domains are shown and on Y-axis number of papers which have used various parameters for evaluating performance are shown.

It can be observed from Figure 3 that for SMS domain [2] [10] [11], accuracy parameter is widely used for spam email detection. Whereas accuracy, precision, recall and F-measure parameters are majorly used in comparison for Email spam detection model [3] [13] [23] [50]. Comparison using precision parameter alone is done for spam tweet detection [43] and F-measure parameter solely used for Instagram spam detection [7]. Accuracy, precision and recall parameters used in comparison for all domains [5] [9] [32] [36] except Instagram.

MODEL, METHOD AND DOMAIN USED FOR SPAM MESSAGE DETECTION

Most of the available works on spam message detection are domain specific. For spam SMS detection, a machine learning based model using text mining is proposed [41]. In this work, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Decision Tree and Naïve Bayes methods were used for message classification. The proposed model is compared based on five performance parameters - accuracy, time to train the model, time of prediction, accuracy-time ratio (train) and accuracy-time ratio (test). In another similar work [4], a spam detection on case study of messages in the Indian region is done by using both classification and clustering based methods. Separately collected Indian spam and ham messages were mixed with the regular SMS spam data-set and performance evaluation is done between classification and clustering methods. An Email spam detection model using supervised learning is proposed by [13]. The model follows a two step methodology to correctly class Emails as spam or ham. Firstly, classification is done followed by URL analysis and filtering to correctly classify an Email. A PSO based Hybrid model presented by [42] where PSO combined with PEGSOS (primal estimated sub-gradient solver for SVM) for email spam detection. Comparison was done later on results in terms of accuracy by using the hybrid approach against PEGSOS alone. An integrated approach using Naïve Bayes and K-Means clustering presented by [33] for spam tweet detection. Using the integrated approach results in better detection rate and F-measure. Hybrid feature selection model using machine learning for spam tweet detection is presented in [37]. In this work, a combination of user based, graph based and content-based features are used for classification. This resulted in the increase in the accuracy as compared to user, graph, content-based features alone.
An implementation of a model for Instagram spam detection using classification is done by [19]. Naïve Bayes classifier is used over the manually collected Indonesian language data-set using scrapping technique. An efficient model for detecting spam reviews using deep learning and machine learning is proposed in [39]. Whole model is divided into four phases. In phase-I, data-set acquisition is done, text preprocessing of data is completed in phase-II. In phase-III, feature selection is done followed by implementation of various machine learning and deep learning algorithms in phase-IV. Also two different data-sets namely – YELP and Kaggle are used to test the efficiency of the proposed approach. After reviewing state-of-the-art works, it is observed that for comparing proposed models for spam detection in five domains, supervised learning, unsupervised learning and other models are used. Supervised learning models are used by [2] [9] [10] [35] [22] for classification.

Unsupervised learning-based model are used by [6] [10] [34] [41] for clustering. KNN [8] [39], DSA (Den Stream Algorithm) [6], KMS [34] are mainly used methods for spam message detection using unsupervised learning model. Deep learning-based algorithms including ANN [3], RNN (Recurrent Neural Network) [39], MLP (Multi-Layer Perceptron) [39], LSTM (Long Short Term Memory) [9] [32], EDM (Encoder-Decoder Model) [43], and CNN (Convolutional Neural Network) [39] are used for spam message detection. However, hybrid models including – NB-PSO [23], NB-GL (Grey List Filtration) [5], NB-C&H (Cuckoo and Harmony Search) [24] are preferred for spam message detection. Further, EM (Ensemble Model) [8], TM (Transformer Model) [21] and GAN (Generative Adversarial Network) [49] are combined with SVM for spam

TABLE II. REVIEW OF SPAM DETECTION METHODS AND MODEL USED

REF NO	DOMAIN	MODEL			FEATURE
		Supervised Learning	Unsupervised Learning	Other	
[2]	Da	✓			NB, RF, LR
[9]	Da	✓			SVM, LSTM, DT, NB, KNN, RF, LR
[10]	Da	✓			DT, RF, KNN, SVM, LR
[11]	Da	✓			NB, SVM
[20]	Da	✓			TM
[50]	Da	✓			CNN
[3]	Db	✓			NB, LR, SVM, ANN
[5]	Db	✓		✓	NB, GL
[13]	Db	✓			NB, SVM, DT, RF, KNN
[22]	Db	✓		✓	NB, PSO
[45]	Db	✓		✓	NB,HMM
[49]	Db		✓		LSTM
[6]	Dc		✓		DSA
[28]	Dc		✓		LSTM
[31]	Dc	✓			SVM
[34]	Dc	✓			NB, J48Classifier, DecorateAlgorithm
[39]	Dc		✓		EDM
[7]	Dd	✓		✓	NB, SVM, XGBoost
[19]	Dd	✓			K-fold Cross-Validation
[21]	Dd	✓			CNB, SVM
[30]	Dd	✓	✓		K-fold CrossValidation, Clustering, RF
[8]	De	✓		✓	NB, SVM, DT, KNN, LR, EM
[23]	De	✓		✓	NB, C&H
[32]	De	✓	✓	✓	SVM, KNN, NB, RNN,MLP,CNN
[33]	De	✓			RF,SVM,NB
[52]	De	✓			DNN

messages detection.

NB [8] [37], SVM [11] [22] [36], DT [8] [9], RF (Random Forest) [2] [36], LR (Logistic Regression) [2] [3], J48 Classifier [37] and CNB (Complementary Naïve Bayes) [22], DNN (Deep Neural Networks) [31] are majorly used methods for spam message detection using supervised learning model.

3. ANALYSIS AND DISCUSSION

Figure 4 represents different models - supervised learning, unsupervised learning and other models used for spam message detection. From Figure 4, it is clearly evident that the supervised learning-based technique for message spam detection is used mostly. Out of 27, 23 work's preferred supervised learning model based methods for spam detection. However, 6 work's used unsupervised based models. 7 work's used other model based methods for spam detection over messages. Figure 5 summarizes the comparison of state-of-the art on accuracy achieved for spam message detection in each of the five domains. For Figure 5 detailed analysis of message spam detection techniques used in each domain is as follows:

A. *SMS*

For SMS spam detection, In [51], CNN is used and highest accuracy of 99.44% is achieved. Hyper parameters are tuned to make the model most accurate. [9] used several word embedding techniques with different algorithms out of which LSTM results in the accuracy of 98.5%. In [2] different features sets are used and out of NB, RF & LR, NB method has optimal performance with accuracy 98.4%. In [21], a modified TM so used achieved the accuracy of 98.9% by tuning the hyper-parameters. In [11] & [10], SVM method results in the highest accuracy with 98% & 97.8 % respectively. Different features extraction techniques are used in both [10] [11].

B. *Email*

For the Email category, [52] achieved the highest accuracy of 99.4% with LSTM model. [13] used NB, SVM, DT, RF methods are Email spam detection along with URL filtering techniques, out of which SVM method outperformed with the accuracy of 97.8%. [50] used several

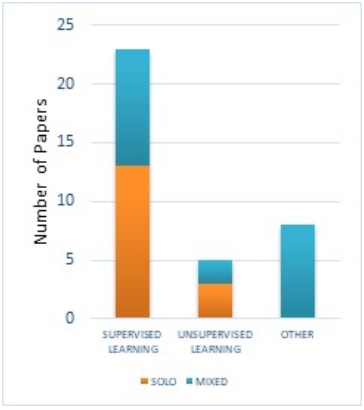


Figure 4. Analysis of various models used for spam message detection

feature selection techniques and a different model i.e., Hidden Markov Model and achieved the accuracy of 91.2%. A NB-PSO based hybrid model is proposed and achieved an accuracy of 95.5% as compared to traditional methods [23]. In [5], an integrated approach of NB with Grey list filter resulted in an accuracy of 97.3%. Out of NB, LR, SVM and Neural Network, the highest accuracy of 98.6% achieved with the Neural Network method [3].

C. Twitter

For twitter spam detection, [35] achieved an accuracy of 93% with SVM method by using different feature extraction methods. DSA methods proposed for spam tweets detection [6]. Rapid miner tool is used and an accuracy of 99% with this method is achieved. [32] used compared machine learning and deep learning based algorithms and LSTM method with integrated inception layer achieved 95.7% accuracy. [43] used EDM along with different word embedding techniques and attained the accuracy of 73%. [37] used different feature selection techniques and J48 classifier outperformed with an accuracy of 97.6% as compared to other approaches.

D. Instagram

For Instagram Spam Detection, [34] used a combination of clustering and K-fold Cross Validation to achieve an accuracy of 96.2% with RF method by using various feature selection techniques. [19] used the NB method with K-fold Cross Validation to achieve an accuracy of 80%. CNB and SVM methods together used by [22] and SVM method results are more accurate with 96% accuracy. [7] results in the accuracy of 96.01% with SVM.

E. Online Review

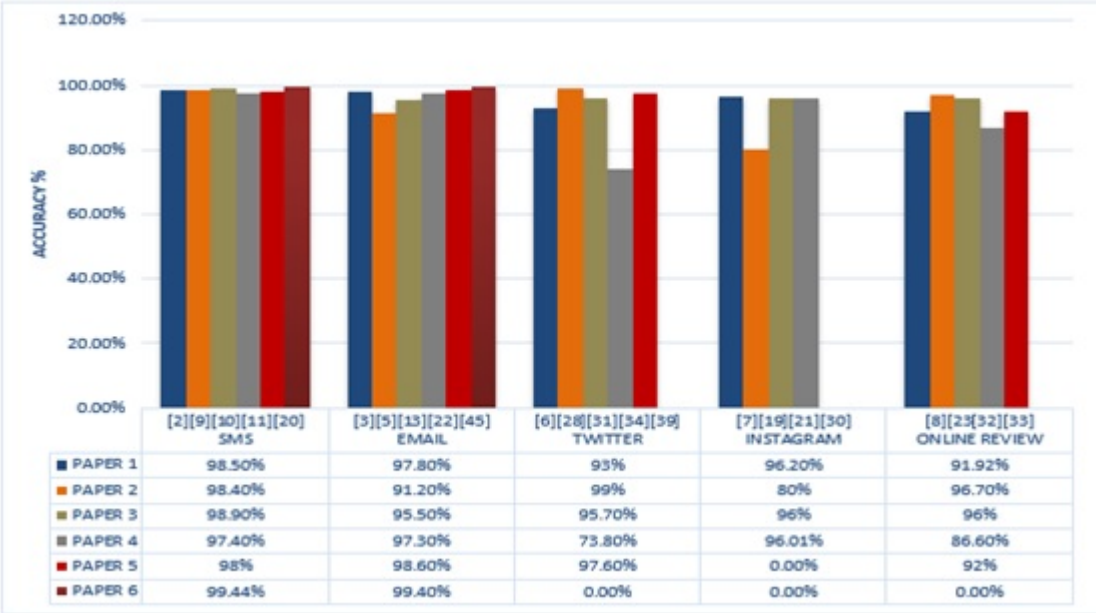
For Online Review Spam Detection [24] proposed integrated approach of NB-CH and 91.92% accuracy was achieved. [39] achieved 96.7% accuracy with LSTM method, [8] achieved accuracy of 96% with EM. [36] achieved accuracy of 86.6% using SVM method. In [20],machine learning and deep learning based algorithms are compared out of which DNN achieved the accuracy of 92.05%.

The comparison of existing-state-of-the-art on accuracy of spam message detection in SMS/Email/Twitter/Online Reviews and Instagram domains is shown in Figure 5. In Figure 5 , X axis represents the domains of spam message detection while Y axis represents the accuracy percentage. It can be observed from Figure 5 that the supervised learning-based models result in the highest accuracy in SMS [51], Email [52], Twitter [3], Instagram [34]. Accuracy of 98.4% using NB for SMS spam detection [2] and accuracy of 80% using NB for Instagram spam detection [19] represents that same method can result differently on different domains due to the change in the data-set. Also, Hybrid methods like - NB-PSO used for email spam detection [23] with 95.5% accuracy and NB-C&H used Online review detection [24] represents that hybrid approaches perform better in terms of accuracy then the most methods based on supervised and unsupervised learning models.

Table III presents the comparison of spam message detection techniques using various parameters. It can be observed form Table III that NLP based Bag of words and N-grams techniques are used for feature extraction in all domains [3] [7] [39] [36] except SMS. Word embedding based Word2Vec taken into consideration for Instagram [7] and Online review domains [39] but TFIDF/Hashing is mostly used in all domains as feature extraction in spam detection [2] [9] [3] [10] [22] [24]. Various researchers preferred GloVE for feature extraction for SMS [21] and Twitter domains [32] [43]. It can also be observed that for Instagram spam detection [19] and Online review spam detection [8], confusion matrix is used for feature extraction. Word cloud [11] and genism package [13] found their use as feature extraction in SMS and Email spam detection respectively. Further form Table III, it can be observed that SVM is the mostly used method for spam message detection [13] [21] [22] [35].Further, from Figure 4 it is clear that unsupervised learning-based model and other model, are less preferred for spam message detection.

4. CONCLUSION AND FUTURE WORK

After comprehensively analyzing the selected state-of-the-art, several research findings and conclusions are made. It is concluded that the majority of work's preferred manual collection of datasets for spam message detection. For text preprocessing, NLP is a widely used technique. Also, TFIDF is mostly used technique for feature extraction in all five domains. It can be concluded that, for spam SMS detection models, the accuracy ranges between 97 % - 100 %. Also, the preferred dataset in this category is mostly preferred from Kaggle [2] [21] [11]. The accuracy for Email spam detection techniques ranges between 91% - 100%. The accuracy for spam tweet detection techniques ranges between 72% - 99%. However, the accuracy of proposed methods for Instagram spam detection model ranges between 80% -97% and the accuracy for Online



	[2][9][10][11][20] SMS	[3][5][13][22][45] EMAIL	[6][28][31][34][39] TWITTER	[7][19][21][30] INSTAGRAM	[8][23][32][33] ONLINE REVIEW
PAPER 1	98.50%	97.80%	93%	96.20%	91.92%
PAPER 2	98.40%	91.20%	99%	80%	96.70%
PAPER 3	98.90%	95.50%	95.70%	96%	96%
PAPER 4	97.40%	97.30%	73.80%	96.01%	86.60%
PAPER 5	98%	98.60%	97.60%	0.00%	92%
PAPER 6	99.44%	99.40%	0.00%	0.00%	0.00%

Figure 5. Analysis of various models used for spam message detection

Review Spam Detection model ranges between 85% - 97%.Further, it is concluded that, DSA outperform with 99% accuracy for spam tweet detection [6] represents that unsupervised learning models are efficient and can be used for other domains. The supervised learning such as SVM based spam message detection approaches can be followed and results best in terms of accuracy for spam message detection. Some evolutionary computing based methods such as Genetic Algorithm, Particle Swarm Optimization, Non-dominated sorting genetic algorithm and some other methods like Generative Adversarial Networks, Negative Selection Algorithm and many more. and hybrid approaches find their application in spam message detection in future.

TABLE III. COMPARISON OF SPAM MESSAGE DETECTION TECHNIQUE

REFNO	DOMAIN	FEATURE SELECTION												FEATURE EXTRACTION				RESULT	
		Whole Text	Common Words	Unique Words	Hyperlinks	Length/Frequency	Header	Sender's Address	No Of Tags	No Of Likes	No Of Comments	No Of Follower/Following	Retweets	Special Symbols	Tfidf/Hash/Count/Sparse	N-Grams	Bag Of Words	Word2vec	Others	Algorithm Used	Accuracy
[2]	Da	✓				✓									✓					NB	98.4%
[9]	Da		✓	✓											✓					LSTM	98.5%
[10]	Da		✓												✓					SVM	98%
[11]	Da	✓		✓											✓				Word Cloud	SVM	97.4%
[20]	Da	✓													✓				GloVe	Modified TM	98.9%
[50]	Da		✓	✓	✓	✓									✓				Confusion Matrix	CNN	99.44%
[3]	Db		✓		✓	✓	✓	✓							✓	✓			Grey List	NN	98.6%
[5]	Db					✓													Correlation Based	NB-GL	97.3%
[13]	Db	✓			✓										✓				Gensim Package	SVM	97.8%
[22]	Db		✓	✓																NB-PSO	95.5%
[49]	Db		✓	✓		✓									✓					LSTM	99.4%
[45]	Db		✓	✓																HMM	91.2%
[6]	Dc			✓	✓	✓			✓			✓	✓						RapidMiner, MAO	DSA	99%
[28]	Dc	✓												✓		✓			GloVe	IN-LSTM	95.7%
[31]	Dc	✓			✓				✓			✓							Weka	SVM	93%
[34]	Dc	✓	✓	✓						✓		✓								J48	97.6%
[39]	Dc	✓			✓														GloVe	E-DM	73.8%
[7]	Dd			✓		✓								✓	✓		✓	✓		SVM	96.01%
[19]	Dd	✓		✓															Confusion Matrix	NB	80%
[21]	Dd	✓		✓											✓					SVM	96%
[30]	Dd		✓	✓	✓	✓			✓	✓	✓	✓		✓		✓			Minhash	RF	96.2%
[8]	De	✓				✓													Confusion Matrix	EM	96%
[23]	De	✓		✓											✓				C&H	NB-C&H	91.92%
[32]	De	✓				✓									✓	✓		✓		LSTM	96.7%
[33]	De		✓	✓		✓								✓			✓			SVM	86.6%
[52]	De		✓	✓		✓										✓				DNN	92.05%

References

[1] J. Johnson, "Global digital population as of january 2021,statisa,january2021,https://www.statista.com/statistics/617136/digital-population-worldwide/."

[2] P. Sethi, V. Bhandari, and B. Kohli, "Sms spam detection and comparison of various machine learning algorithms," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 28–31.

[3] M. Sethi, "Email spam detection using machine learning and neural networks," ,*International Research Journal of Engineering and Technology*, iRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04 — Apr 2021.

[4] S. Agarwal, S. Kaur, and S. Garhwal, "Sms spam detection for indian messages," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015, pp. 634–638.

[5] A. Iyengar, G. Kalpana, S. Kalyankumar, and S. GunaNandhini, "Integrated spam detection for multilingual emails," in *2017 International Conference on Information Communication and Embedded Systems (ICICES)*, 2017, pp. 1–4.

[6] N. Eshraqi, M. Jalali, and M. H. Moattar, "Detecting spam tweets in twitter using a data stream clustering algorithm," in *2015 International Congress on Technology, Communication and Knowledge (ICTCK)*, 2015, pp. 347–351.

[7] A. Septiandri and O. Wibisono, "Detecting spam comments on indonesia's instagram posts," *Journal of Physics: Conference Series*, vol. 801, p. 012069, 01 2017.

[8] A. K. Suborna, S. Saha, C. Roy, S. Sarkar, and M. T. H. Siddique, "An approach to improve the accuracy of detecting spam in online reviews," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 296–299.

[9] S. Gadde, A. Lakshmanarao, and S. Satyanarayana, "Sms spam detection using machine learning and deep learning techniques," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2021, pp. 358–362.

[10] P. B, "Spam detection using nlp techniques",international journal of recent technology and engineering (ijrte," *ISSN*, vol. 2277-3878, Volume-8, pp. –2 11,.

[11] P. Navaney, G. Dubey, and A. Rana, "Sms spam filtering using supervised machine learning algorithms," in *2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2018, pp. 43–48.

[12] SlickText, "44 mind-blowing sms marketing and texting statistics"." [Online]. Available: https://www.slicktext.com/blog/2018/11/44-mind-blowing-sms-marketing-and-texting-statistics/

[13] A. Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar, and D. Karia, "E-mail spam classification via machine learning and natural language processing," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, pp. 693–699.

[14] S. R. Department, "Number of e-mail users worldwide from 2017 to 2025"." [Online]. Available: https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[15] "Christo petrov "52 gmail statistics to show how big it is in 2022"." [Online]. Available: https://techjury.net/blog/gmail-statistics/#gref

[16] Dataprot, "How much spam is sent worlds's daily email traffic"." [Online]. Available: https://dataprot.net/statistics/spamstatistics/#:~:text=How%20much%20spam%20is%20sent,the%20world's%20daily%20email%20traffic.

[17] [U+FFFD]Dean B, "Many people use twitter in 2022?"."

[18] "Statista research department, "instagram - statistics facts"." [Online]. Available: https://www.statista.com/topics/1882/instagram/#dossierKeyFig.s

[19] B. Priyoko and A. Yaqin, "Implementation of naive bayes algorithm for spam comments classification on instagram," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, 2019, pp. 508–513.

[20] N. Krishnaveni and V. Radha, *A Hybrid Classifier for Detection of Online Spam Reviews*, 01 2022, pp. 329–339.

[21] X. Liu, H. Lu, and A. Nayak, "A spam transformer model for sms spam detection," *IEEE Access*, vol. 9, pp. 80 253–80 263, 2021.

[22] N. Haqimi, N. Rokhman, and S. Priyanta, "Detection of spam comments on instagram using complementary naïve bayes," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, p. 263, 07 2019.

[23] K. Agarwal and T. Kumar, "Email spam detection using integrated approach of naïve bayes and particle swarm optimization," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 685–690.

[24] S. Rajamohana, K. Umamaheswari, and S. Keerthana, "An effective hybrid cuckoo search with harmony search for review spam detection," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017, pp. 524–527.

[25] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," *Procedia Computer Science*, vol. 184, pp. 853–858, 01 2021.

[26] F. Hossain, M. N. Uddin, and R. K. Halder, "Analysis of optimized machine learning and deep learning techniques for spam detection," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–7.

[27] S. Annareddy and S. Tammina, "A comparative study of deep learning methods for spam detection," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp. 66–72.

[28] M. RAZA, N. D. Jayasinghe, and M. M. A. Muslam, "A comprehensive review on email spam classification using machine learning algorithms," in *2021 International Conference on Information Networking (ICOIN)*, 2021, pp. 327–332.

[29] S. Gheewala and R. Patel, "Machine learning based twitter spam account detection: A review," in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, 2018, pp. 79–84.

[30] S. P. Rajamohana, K. Umamaheswari, M. Dharani, and R. Vedackshya, "A survey on online review spam detection techniques," in

2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017, pp. 1–5.

[31] A. Barushka and P. Hájek, *Review Spam Detection Using Word Embeddings and Deep Neural Networks*, 05 2019, pp. 340–350.

[32] M. V. Neha and M. S. Nair, "A novel twitter spam detection technique by integrating inception network with attention based lstm," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1009–1014.

[33] R. Katpatal and A. Junnarkar, "An efficient approach of spam detection in twitter," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018, pp. 1240–1243.

[34] W. Zhang and H.-M. Sun, "Instagram spam detection," in *2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2017, pp. 227–228.

[35] S. Gharge and M. Chavan, "An integrated approach for malicious tweets detection using nlp," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017, pp. 435–438.

[36] W. Etaiwi and A. Awajan, "The effects of features selection methods on spam review detection performance," in *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, 2017, pp. 116–120.

[37] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for twitter," in *2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2017, pp. 466–471.

[38] C. Rădulescu, M. Dinsoreanu, and R. Potolea, "Identification of spam comments using natural language processing techniques," in *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2014, pp. 29–35.

[39] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. Binte Hassan, "Spam review detection using deep learning," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 0027–0033.

[40] D. Pal, P. Verma, D. Gautam, and P. Indait, "Improved optimization technique using hybrid aco-pso," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 277–282.

[41] M. Rubin, "Spam detection insms using machine learning through text mining"," *International Journal Of Scientific Technology Research*, vol. 9, no. ue 02, february 2020 Issn 2277-8616.

[42] L. M. El Bakrawy, *Hybrid Particle Swarm Optimization and Pegasos Algorithm for Spam Email Detection*, 01 2019.

[43] K. Badola and M. Gupta, "Twitter spam detection using natural language processing by encoder decoder model," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 402–405.

[44] H. Gheeseewan and S. Pudaruth, "Categorisation of computer science research papers using supervised machine learning techniques," *International Journal of Computing and Digital Systems*, vol. 9, no. 6, p. 1165.

[45] S. A. Israel, J. Goldstein, J. S. Klein, J. Talamonti, F. Tanner, S. Zabel, P. A. Sallee, and L. McCoy, "Generative adversarial networks for classification," in *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2017, pp. 1–4.

[46] H. Wu, H.-z. Li, G. Wang, H.-l. Chen, and X.-k. Li, "A novel spam filtering framework based on fuzzy adaptive particle swarm optimization," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, vol. 1, 2011, pp. 38–41.

[47] I. Idris and A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization," *Applied Soft Computing*, vol. 22, p. 11–27, 09 2014.

[48] G. Lingam, R. R. Rout, D. V. L. N. Somayajulu, and S. K. Ghosh, "Particle swarm optimization on deep reinforcement learning for detecting social spam bots and spam-influential users in twitter network," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2281–2292, 2021.

[49] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 89–95.

[50] S. R. Gomes, S. G. Saroar, M. Mosfaiul, A. Telot, B. N. Khan, A. Chakrabarty, and M. Mostakim, "A comparative approach to email classification using naive bayes classifier and hidden markov model," in *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)*, 2017, pp. 482–487.

[51] P. Roy, J. Singh, and S. Banerjee, "Deep learning to filter sms spam," *Future Generation Computer Systems*, vol. 102, 09 2019.

[52] E. E. Eryılmaz, D. [U+FFFD] Şahin, and E. Kılıç, "Filtering turkish spam using lstm from deep learning techniques," in *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, 2020, pp. 1–6.

S Aditya Chaturvedi S Aditya Chaturvedi is post-graduation scholar at SGSITS Indore with Information Technology as major. He received his BTech degree from Jaipur National University, Jaipur.

Lalit Purohit Lalit Purohit is working as Associate Professor at SGSITS Indore. He has earned his PhD from IIT Roorkee. His research interest includes, Application of Security in the area of Web Services, IoT and Cloud Computing.