International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)
ISSN: 2305-0543 (Online), ISSN: 2411-6173 (Print)
Vol. 11(40), Apr. 2021, PP. 4931-4936

# Short Message Service (Sms) Spam Detection and Classification Using Naïve Bayes

Christine Bukola Asaju,
*Department of Computer Science,*
Federal Polytechnic Idah,
Idah, Kogi State, Nigeria.
chrisamaju02@gmail.com

Ekuma, James Ekorabon,
*Department of Computer Science,*
Federal Polytechnic Idah,
Idah, Kogi State, Nigeria.
ekuma_ejn@yahoo.com

Richard Ojochegbe Orah
College of Information and
Communication Technology
Salem University Lokoja
Kogi State, Nigeria.
orahseun@gmail.com

***ABSTRACT*— The dynamic nature of technology has caused an unprecedented technological and socio-economical development in everyday life. This development is making everyone to be highly vulnerable to diverse threats. Short Message Service (SMS) spam is one of such threats that affect the security of mobile devices. These spam attempts to deceive users into providing their private information which could later result in a security breach. The major problem of SMS spam is that attackers, hackers, email phishing, ransomware, etc., used it to exploit the victims. The urge to curb this has necessitated this work. Different models have been developed to detect SMS spam, some of these models include Support Vector Machine, Linear Classifier, Decision Trees, Random Forest, Logistic Regression, Naive Bayes, etc. However, most of these techniques have not addressed the point that focuses on SMS spam detection and classifies new SMS spam. The goal of this research is to develop a machine learning model for the detection and classification of new SMS spam. This paper presents a model for SMS spam detection and classification that employs the Naïve Bayes machine learning methodology. String to word vector feature extraction was used to extract the SMS Spam text file from the contents that were collected via UCL repository in its original form. At this point, the proposed system is set to perform data preprocessing, dataset feature extraction, and model training as well as model evolution. The model was learned based on an SMS dataset that consists of 5525 samples collected from an online resource and utilized effectively. The experimental results indicate classification accuracies of 99.42%, for correctly classified and 0.57% for incorrectly classified, respectively in the best cases.**

***Keywords*— *Short Message Service (SMS) spam, Naïve Bayes, String to word, Machine Learning***

## I. INTRODUCTION

With the development of technology, there have been various means used in communication. One of such is a text messaging on mobile devices. Text messaging also known as Short Messaging Service (SMS) is a text communication platform that allows mobile phone users to exchange short text messages. These messages are usually less than 160 seven-bit characters [1].

A major problem that cell phone users encounter is the reception of unsolicited SMS messages. Unsolicited messages are regarded as spam messages. These messages usually come from advertisers and other sources.[2] As the popularity of the platform increases, there is a surge in the number of unsolicited commercial advertisements sent to mobile phones using text messaging [3].

An automated technique for identifying spam to prevent its delivery is regarded as spam filtering [4]. Spam filtering exists both for text messages and emails but has some differences. One of such disparity is that emails, which have a variety of large datasets available, but real databases for SMS spams are very limited.

Another difference is that text messages are short in length of words, therefore, the number of features that can be used for its classification is far less compared to emails. In text messages, there is no header as well. Additionally, text messages are full of abbreviations and have much less formal languages than what is experienced from emails. All of these factors may result in a serious degradation in the performance of major email spam filtering algorithms applied to short text messages.

Consequently, SMS spam detection and classification requires an effective and efficient method. Although models have been developed to detect SMS spam, many of these techniques have not addressed the point that focuses on SMS spam detection and classifies new SMS spam. The is a major gap this research is seeking to solve. The research focus is to develop a machine learning model for the detection and classification of new SMS spam.

The objective of the research is to apply machine-learning algorithm in SMS spam detection and classification problems. The research further explores the design of an the application based on an algorithm that can detect and classify new SMS spam with high-performance accuracy.

## II. DATASET DESCRIPTION

The data set adopted for the work comprised of 5574 text messages from the UCI Machine Learning repository gathered in 2012 [5] (SMS Spam collection Dataset from the UCI Machine Learning Repository.

It is comprised of a collection of 425 SMS spam messages that were manually extracted from the Grumble text Web site, a subset of 3,375 SMS that was chosen randomly, non-spam (ham) messages of the NUS SMS Corpus (NSC), a list of 450 SMS non-spam messages collected, and the SMS Spam Corpus v.0.1 Big (1,002 SMS non-spam and 322 spam messages publicly available)[5].



**Article History:**
Received Date: Dec. 08, 2020, Accepted Date: Jan. 20, 2021, Available Online: Apr. 01, 2021

## III. SMS SPAM DETECTION AND CLASSIFICATION WORKFLOW

Below is the working flow of SMS Spam detection and classification.

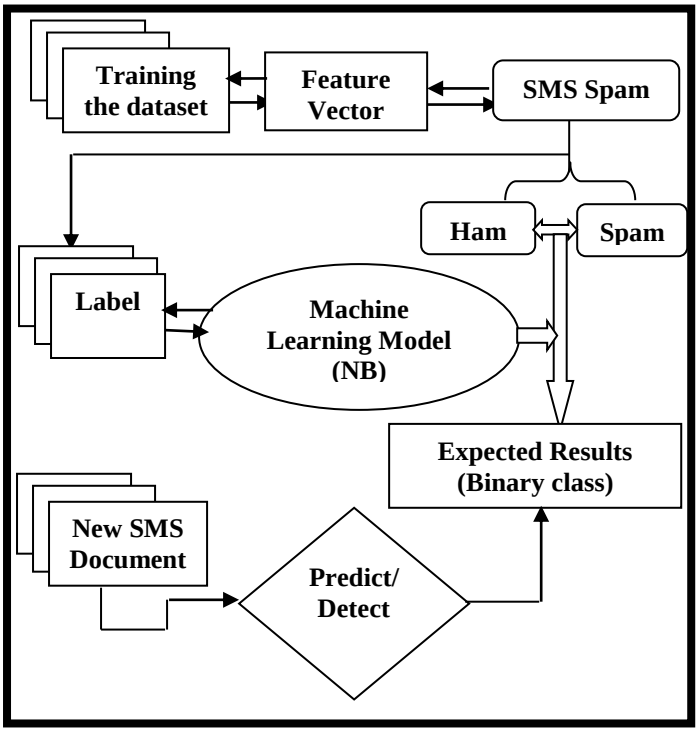


**Figure 1: The working flow of SMS Spam Processes**

The performance of the classifier was summarized and evaluated. Feature extraction and initial analysis of data were done with library Weka. The application of machine learning algorithms (weka plugin) was done in the NetBeans IDE for the implementation of the model. The paper is organized as follows: Section IV discusses the related works on SMS data mining, using Naïve Bayes algorithm. Section V describes the methods of approach, and VI discuses the implementation , while VII discusses the result. The conclusion and future work is found in VIII.

## IV. RELATED WORKS

Short messaging Services (SMS) has become an important means of communication today between millions of people around the world. SMS services, which are a must-have service nowadays for telecom operators, transmit their messages using standardized communications protocols. At the same time, problems are being faced which is caused by SMS spamming. Previously, it has been explained that SMS spam and email spams have similar features.

Therefore, spam filtering methods used in emails can be adopted for SMS spam filtering to deal with the spread of mobile phone spams [6, 7]. Nonetheless, the properties exhibited by SMS spams is different from the email spams. SMS spam in email has few characters and is limited to 140bytes by standard text messages [8]. Due to restrictions, many times, mobile devices spams are written in less formal languages such as abbreviations or idioms. Unfortunately, SMS spams irritate users the same way as e-mails [ 9, 10, and 12].

Another study which is on spam filtering methods and measures was carried out by Chang et al. [13]. Spam filtering methods and features reweighting methods based on good word attack strategies were developed by them. This approach depends on the limited character and short text messages on top of the weight values which are evaluated against a dataset (i.e. SMS and comment dataset). They opined that a good word attack strategy will mislead the classifier's output with the least number of input characters. The authors in the work also introduce a feature reweighting method in which a novel rescaling function is proposed. This function is aimed at reducing the significance of the features characterizing a short word. This method is performed to rescale the weights and increase the linear classifier's robustness against a good word attack strategy.

Sethi et al. [14], compared different machine learning algorithms that filters and detect SMS spam messages using three basic procedures, namely; the raw text messages, the length of the messages, and the information gain matrix. Naïve Bayes, Random Forest, and Logistic Regression algorithms were adopted in the experiment.

Meanwhile, in [15] different machine learning algorithms were used to train four features derived from SMS text messages. These features include the size of the message, monograms with the highest frequency in the text messages, frequently occurring diagrams, and a class of messages (i.e. ham and spam as 0 and 1 respectively). The authors were able to make a discovery that the Naïve Bayes algorithm outperforms the other classifiers used in the study.

Different features for SMS spam classification were also explored and analyzed by Choudhary and Jain [16]. Numerous features were extracted from SMS text messages including mathematical symbols, special symbols, and emotions, among many others. In their study, characteristics and behaviors of SMS spam messages for classification with a successful result was achieved.

An SMS spam filtering method using non-content features was proposed by [17]. Static features were used instead of the content of an SMS text message as features, (i.e. number of messages and message size), temporal features (i.e. number of messages sent in one day, size of messages in one day and time of day) and network features (i.e. number of recipients and clustering coefficients) in the study. In the detection of spammers, it was discovered that incorporating network and temporal features into conventional static features, a better performance was achieved.

Warade et al. concept of spam detection in [18] was based on the relationship between the sender and receiver and the message contents. For lack of mutual relations between the senders and the receivers, a text message is classified as spam while the SMS displays the contents of the spam. The message is then automatically transferred into the spam box. Whereas, with mutual relation between senders and receivers, an SMS text message is considered legitimate with no visible spamming content. The relationship between senders and receivers will be examined through the inspection of SMS logs and the direct relationship between the two.



IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

Safie et al. [19] applied string to word vector feature extraction to prove that SMS spam detection and classification work better. They achieved this through a vector space and Artificial Neural Network (ANN) algorithm. Accuracy shows a significant improvement.

[20] aims at comparing the performance of different algorithms with feature selection and algorithms without a feature selection. The first approach was that the sampled data was being examined without any filters or features selection, then the classifiers were tested each time beginning with the best-first feature selection to be able to select the most beneficial features and then apply various classifiers for classification.

Using Random Tree classifier, achieved 99.72% accuracy which means it works best to detect spam emails. In conclusion, the accuracy of email filters was improved greatly when the algorithm with feature selection was applied to the entire process and that classifiers of tree shape are more efficient in detecting spam emails [21].

[22] in their work, also detected an unknown zero-day phishing email that relies on an evolving connectionist system. The system was named the phishing dynamic evolving neural fuzzy framework (PDENFF). This framework follows a hybrid learning approach (supervised/ unsupervised) and is supported by an offline learning feature to achieve its purpose. Adopting this system helped in enhancing the detection of zero-day phishing e-mails was improved between 3% and 13%. Moreover, it used rules, classes, or features to enhance the learning process using ECOS which provided the system with the advantage of distinguishing phishing emails from a legitimate one [22].

[23] developed a model for classifying phishing emails. They adapted the forest machine learning mechanism. The dataset used comprises of 2000 phishing emails with advanced features. This model was able to achieve an accuracy of 99.7% classification with low false negative(FN) and false positive(FP). It was further reiterated that this model is more efficient because it requires fewer features to detect phishing.

A fraudulent detection model was proposed by [24] using an advanced selection of features where the different categories were compared in terms of the fraudulent email detection rate. The study was conducted applying several classification approaches and algorithms, such as SVM, NB, J48, and CCM, in addition to different features sets. An accuracy percentage of 96% was achieved and the results indicated that the level of accuracy was affected by the type of selected features rather than the classifiers' type, [24].

In [25], Kathirvalavakumar et al proposed a multilayer neural network to detect phishing emails. The proposed network depends on a feedforward pruning algorithm that extracts distinguished data and features from the email and applies a weight trimming strategy. This pruning strategy helps in reducing the number of features that go through the algorithm resulting in minimum computation required for classification of emails into phishing or not. The network has provided fair results in terms of false positives and false negatives. This network was tested on data from 2007, thus, using this network for current data requires identifying the new features to the algorithm incorporating them into input domain for training to be useful, [25].

Consequently, effective and efficient methods are required for SMS spam detection and classification. Although different models have been developed to detect SMS spam, many of these techniques have not addressed the point that focuses on SMS spam detection and classifies new SMS spam, and that is a major gap this work seeks to address. The researchers' effort is to develop a machine learning model for the detection and classification of new SMS spam.

## V. MATERIALS AND METHOD

This section focuses on the concept of spam detection using machine learning tasks. The approach was based on the samples of SMS Spam concerning their classes. Based on this fact, the system will be built with the available data set collected with other related literature reviews such as journals or articles.

### *A. Machine learning Approach*

1) Collect the sample data (SMS Spam/historical data)

2) Pre-processing (that is the data were provided with two labels, spam, and ham), since it is a supervised learning approach, then it is a binary classification.

3) Apply feature extraction with Weka library (to convert the SMS Spam into binary classification analysis)

4) Resample the dataset by applies training set and testing set during system development analysis using Weka tools.

Develop the model with Weka plugin and used java Netbeans IDE environment to implement the system with all the requirements stated above and used the proposed algorithm to perform the classification model and structured data analytics.

### *B. Naïve Bayes Classifier*

The Naïve Bayes classifier provides a new way of analyzing data based on Bayes theorem. It is based on evidence by maintaining a relation between the target and the problem space. It is a probabilistic classifier that uses Bayes theorem with some solid assumptions. It is used for text classification. The real-world applications use Naïve Bayes classifier for email sorting, spam detection, document categorization, etc. It is a very efficient method because it is less computationally intensive in CPU and memory use as it uses a small amount of training data. In general, a Naïve Bayes classifier assumes the presence and absence of a particular feature required to classify a data set. The probability model for a classifier is denoted by p(C|F1 ,.....,Fn). Where c denotes the class variable which is used to classify the sample dataset, and $F_1$ to $F_n$ is the number of features. If the number of features n is large or a single feature is containing a large number of values, the probability table becomes infeasible. So, the Bayes theorem is rewritten as equation (1) below:

$$P^{(C|F_1;\ldots;F_n)} = \frac{p(C)\,p(F_1;\ldots;F_n\,F|C}{p(F_1;\ldots;F_n)} \quad (1)$$



IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

*C.* System Design

The method adopted to achieve this work is as follows:

SMS Spam data collection

SMS Spam data pre-processing

SMS Spam Feature extraction

Training set and Test set

Build the model

Based on this concept above, supervised learning will be used for training of the algorithm with a label of the class it belongs., the algorithm learns the relationship between the feature sets and the output by using the labeled data and hence it is then able to classify the unlabeled data from the learned relationship. Figure 3 shows the conceptual framework of the model.

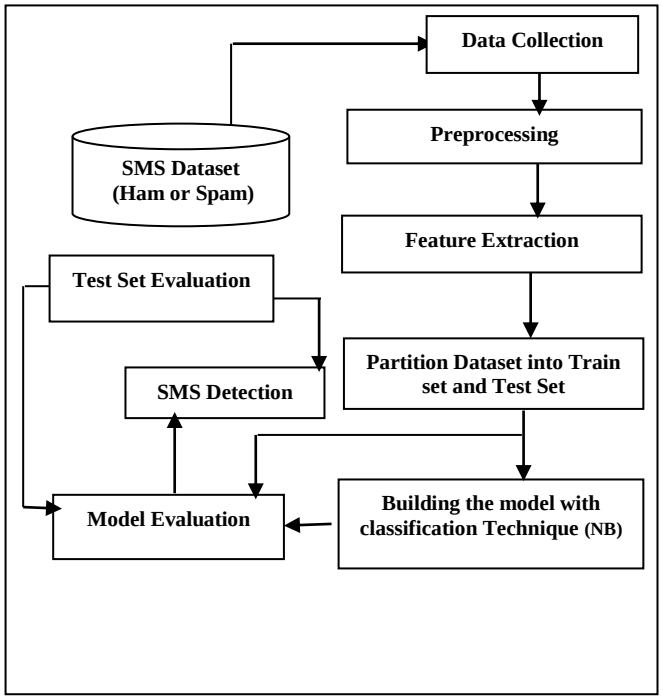


Figure 3. Steps for SMS Spam classification

*Pre-Processing*

In this step, complete geometric correction and filtering are done. The preprocessing uses the output of the classifier to take the required action to improve the performance.

*Dataset Description*

The dataset used in this paper is freely available on the internet. This was collected via the UCI repository was used for this analysis. This dataset consists of 2 attributes and 5525 instances. The first attribute, class_att has two possible values spam and ham which are nothing but class labels. Spam has 747 instances whereas ham has 4778 instances. Attribute one represents the name of the label. The second attribute is text, whose values are nothing but a text message, [26].

*Experimental Set-Up*

All the experiments that were carried out in this section are computed using open source tool Weka 3.8.0[27] and java programming language with Netbeans IDE under the OS Windows for implementation of the machine learning model and processor with 4GB of main memory. Weka is a collection of a machine learning algorithm for data mining tasks. These algorithms may be applied directly using the default algorithm in the tool itself or we can call the algorithm using java code [28]. The following subsection discusses more content of dataset, pre-processing of the dataset, and performed classification and detection using naïve Bayes.

*Selection of Training Data*

In this step, the particular attributes are selected which best describes the pattern for detecting either the messages is spam or ham.

*Classification of Outputs*

The output of the expected result is classified as to different categories accordingly namely ham or spam.

## VI. IMPLEMENTATION

This section focuses on the general implementation and results of the research. The system was implemented with a sample of SMS datasets, for a binary classification technique namely ham and spam. Based on these concepts the data was collected and utilized for the detection and classification of the machine learning model. This was achieved with the proposed model namely Naïve Bayes. This algorithm was used to train the data collected and the model was built by calling the java code without using the default algorithm. These SMS datasets were pre-processed and features were extracted before applying a classification algorithm on it. The classification method used for this work was based on the Naïve Bayes algorithm which was able to capture all the required training sets and used the test set to make prediction/detection and classification.

The system was implemented with the set of two (2) features to distinguish their performance. When those factors were structured into the Weka model and java NetBeans for parameter turning, the proposed model was achieved successfully with five thousand, five hundred and twenty-five (5525) instances, which was used to perform this analysis. This result was used to build the model for predicting a promising result. Here the binary value from a given sample was transformed into an excel format with an extension of csv and arff for the machine-readable task.

*Model Evaluation*

The experiment of SMS detection or classification was done on two folds, which are the sample of the dataset collected that was used to perform SMS classification. And the training set was used to build the model and then used the test set for predicting the result with an unknown class label to predict a new class label with their respective classes as shown below



IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

| | | |
|---|---|---|
| *Correctly Classified Instances* | *522* | *99.4286 %* |
| *Incorrectly Classified Instances* | *3* | *0.5714 %* |
| *Kappa statistic* | | *0.9763* |
| *Mean absolute error* | | *0.011* |
| *Root mean squared error* | | *0.0742* |
| *Relative absolute error* | | *4.5086 %* |
| *Root relative squared error* | | *21.3327 %* |
| *Total Number of Instances* | | *525* |

Table 1. Detail Performance Evaluation by class

| class | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|
| spam | 0.986 | 0.973 | 0.980 | 0.997 |
| ham | 0.996 | 0.998 | 0.997 | 0.997 |

Notably, Naïve Bayes achieved a good performance in this experiment in terms of accuracy by class with 0.986-0.996 in precision, 0.973-0.998 in recall, 0.980-0.997 in F-measure while ROC is 0.997-0.997 in both cases.

Table 2. The New SMS Test set for model evaluation

| | **New SMS Detection and Classification** | **Binary Class** |
|---|---|---|
| 1 | Hello, brother, you have won a brand new car for yourself, call this landline for claiming your price | spam |
| 2 | Congratulation!!! You have won yourself a free brand new laptop, contact this 0807890786323 for collection | spam |
| 3 | Sorry, I was unable to make it yesterday | ham |
| 4 | Wow!!! special offer for you get 4GB for 1000 to enjoy this amazing offer, just recharged above 200 | spam |
| 5 | Please if you reach inform me about your school fees to tell your dad | ham |
| 6 | It's ok, I will keep you abreast | ham |
| 7 | ATM BLOCK: Dear customer your ATM card has been blocked due to BVN upgrade of the year quickly call 09052207076 to reactivate within 24h | spam |

## VII) Discussion of Results

The results of this work were achieved with the SMS dataset and the model evaluations were done with training data as showing above in under section VI. The result obtained is based on the proposed model that was used to classify the total number of 5525 instances and the accuracy of the model is 99%. This could be inferred that the system was able to learn well and captured all the required sample data for effective utilization.

## VIII) CONCLUSION AND FUTURE WORK

The SMS spam messages problem is of the increase in almost every country today. The increase is without a sign of slowing down, as the number of mobile users increases. The cheap rates of SMS services are also a contributing factor to this increase. Therefore, this paper presents the spam detection and classification using Naïve Bayes algorithms.

Before the analysis was done the proposed system used a weka tool to perform the resampling technique, where the dataset where partition into the training set and testing set with the following file format (arrf and txt). Our partition was done with 90% of the training set and 10% of the testing set. After this, the researchers implemented the model with java code written in weka machine learning that was able to handle feature extraction with embedded n-grams feature technique as well as string to word vector function. The data set was trained in both arrf and txt file format. Based on this concepts the model was built with the training set and evaluation was done using the test set and this was able to classify the message into binary classification with our proposed model (Naïve Bayes) and detect the SMS spam being sent to the receiver and the expected results is either spam or ham.

However, it is not enough to evaluate the performance of the model based on the accuracy alone, since the dataset is experiencing imbalanced; therefore, the precision, recall, f-measure and ROC Area of the algorithms must also be observed. After some examinations, Naïve Bayes provides good accuracy by class with 0.986-0.996 in precision, 0.973-0.998, in recall, 0.980-0.997 in F-measure while ROC is 0.997-0.997 in both cases and the results based on the features used. For future works, adding more discretization feature set as a necessary step in the model estimation for better performance.

From the analysis of the results obtained, Naïve Bayes copes well with the SMS dataset. It was well concluded that these results are of sufficient accuracy to be of much practical use. Hence, the effort of future research is to improve classification performance with the discretization feature set as a necessary step in the model estimation for better performance.

It is therefore recommended that SMS Spam message detection and classification using the Naïve Bayes will be of help to the society from such messages that can deceive them to supply all their necessary personal details for easy tracking.

## References


[1] Shirani-Mehr, Houshmand. "SMS spam detection using machine learning approach." (2013): 1-4.

[2] Alzahrani A, Rawat DB. Comparative Study of Machine Learning Algorithms for SMS Spam Detection. SoutheastCon (2019) Apr 11 (pp. 1-6). IEEE.

[3] Qian, Wang, Han Xue, and Wang Xiaoyu. "Studying of classifying junk messages based on data mining." Management and Service Science, 2009. MASS'09. International Conference on. IEEE, 2009

[4] Cormack GV. Email spam filtering: A systematic review. Now Publishers Inc; 2008.

[5] SMS Spam Collection Data Set from UCI Machine Learning Repository,"http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection"

[6] Q. Xu, E., W. Xiang, Q. Yang, J. Du, and J. Zhong. (2012) "SMS Spam Detection Using Noncontent Features." IEEE Intell. Syst. 27(6): 44–51.

[7] Sethi, G., and V. Bhootna. (2014) SMS Spam Filtering Application Using Android.

[8] Nagwani, N. K. (2017) "A Bi-Level Text Classification Approach for SMS Spam Filtering and Identifying Priority Messages." 14 (4): 8




IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

[9] Almeida, T. A., J. M. Gómez, and A. Yamakami. (2011) "Contributions to the Study of SMS Spam Filtering: New Collection and Results." p. 4.

[10] Mujtaba, D. G., and M. Yasin. (2014) "SMS Spam Detection Using Simple Message Content Features." J. Basic Appl. Sci. Res. 4 (4): 5.

[11] Delany, S. J., M. Buckley, and D. Greene. (2012) "SMS Spam Filtering: Methods and Data," Expert Syst. Appl. 39(10): 9899–9908

[12] Shirani-Mehr, H. (2013) "SMS Spam Detection using Machine Learning Approach." p. 4.

[13] Chang, P. P. K., C. Yang, D. S. Yeung, and W. W. Y. Ng. (2015) "Spam Filtering for Short Messages in Adversarial Environment." Neurocomputing 155: 167–176.

[14] Sethi, P., V. Bhandari, and B. Kohli. (2017) "SMS Spam Detection and Comparison of Various Machine Learning Algorithms", in 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN). pp. 28–31

[15] Mujtaba, D. G., and M. Yasin. (2014) "SMS Spam Detection Using Simple Message Content Features." J. Basic Appl. Sci. Res. 4 (4): 5.

[16] Choudhary, N., and A. K. Jain. (2017) "Towards Filtering of SMS Spam Messages Using Machine Learning-Based Technique", in Advanced Informatics for Computing Research 712: 18-30.

[17] Q. Xu, E., W. Xiang, Q. Yang, J. Du, and J. Zhong. (2012) "SMS Spam Detection Using Noncontent Features." IEEE Intell. Syst. 27(6): 44–51.

[18] Warade, S. J., P. A. Tijare, and S. N. Sawalkar. (2014) "An Approach for SMS Spam Detection." Int. J. Res. Advent Technol. 2 (2): 4.

[19] Safie, W., N.N.A. Sjarif, N.F.M. Azmi, S.S. Yuhaniz, R.C. Mohd, and S.Y. Yusof. (2018) "SMS Spam Classification using Vector Space Model and Artificial Neural Network." International Journal of Advances in Soft Computing & Its Applications 10 (3): 129-141.

[20] Rathi, M., & Pareek, V. (2013). Spam Mail Detection through Data Mining-A Comparative Performance Analysis. International Journal of Modern Education and Computer Science, (12), 31

[21] Al-Momani, A., Gupta, B. B., Wan, T. C., Altaher, A., & Manickam, S. (2013). Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing email

[22] Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics

[23] Nizamani, S., Memon, N., Glasdam, M., & Nguyen, D. D. (2014). Detection of fraudulent emails by employing advanced feature abundance. Egyptian Informatics Journal, 15(3), 169-174.

[24] Sa'id Abdullah Al-Saaidah (2017), Detecting Phishing Emails Using Machine Learning Techniques, https://www.meu.edu.jo/libraryTheses/590422b4d5dd8_1.pdf, Retrieved 23/06/2020

[25] Kathirvalavakumar, T., Kavitha, K., & Palaniappan, R. (2015). Efficient Harmful Email Identification Using Neural Network, British Journal of Mathematics & Computer Science, (1), 58

[26] Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

[27] Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. The Journal of Machine Learning Research. 2017 Jan 1;18(1):826-30.

[28] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An update; SIGKDD Explorations, Volume 11, Issue 1. [Available Online: http :// www.cs. waikato.ac.nz/ ml/weka/index.html]