COMPARISON OF NAIVE BAYES AND SVM CLASSIFIERS FOR DETECTION OF SPAM SMS USING NATURAL LANGUAGE PROCESSING

N. Krishnaveni[1] and V. Radha[2]

[1]Department of Information Technology, Avinashilingam Institute for Home Science and Higher Education for Women, India
[2]Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, India

Abstract

Day today's innovative world observers an extraordinary possibility in the communication sector. Individuals will in general utilize various approaches to speak with individuals around the world. The regular methods for sharing short data in an exceptionally simple manner and is cases recorded now a days. This desires a need to recognize Spam SMS to stay away from digital wrongdoing robbery and extortion exercises. A labeled dataset is utilized for recognition reason and two classifiers to be specific Support Vector Machine and Naïve Bayes are utilized to make a correlative examination for the location of spam accomplished by utilizing of Short Message Service. SMS doesn't require any web charges yet, it is unsurpassed utilized methods for remote correspondence. Each versatile client has this office of course. It has an incredible monetary effect on the clients just as the specialist co-ops. Then again SMS spam is one of the major digital wrong doing SMS and the exhibition of classifiers are thought about.

Keywords:

Spam SMS, Support Vector Machine, Naïve Bayes, Classification, Natural Language Processing

1. INTRODUCTION

The trend of day's world focuses on advertising almost everything to the common man in different ways. So the economical sectors require a common means to communicate with their target costumer [1]. The easiest way to achieve this is to send short messages related to their intention. Individual's phone number is not only known to the people whom they wish to give their number. In terms of customer feedback, registrations, shopping, etc., a common man's mobile number is given to various sorts of organizations. Some companies try to promote their brand by sending SMS (Small Messaging Services) to the mobile numbers which they have obtained from the above listed ways. On the other hand, there are some unethical companies who buy or sell these numbers for illegal offences. These companies try to manipulate the common man's interest by sending spam messages like lot selection, cash award, lottery, fake bank messages, etc., criminal cases are recorded as cyber theft because of loss of money. A common man cheated because of the greed, lack of awareness of these kinds of theft and interest towards their luck. Hence, detection of spam SMS is very important to safe to avoid these kinds of cyber thefts [2].

Human nature is always interested in sophisticated life. The main source for such a lifestyle is money. This greed leads one group of people to make money in unethical and illegal ways like sending spam messages and make innocent people fall for their threat. On the other hand, the unaware community is affected and loses their money or security details to the spammers. But this research work focuses to detect spam SMS with the contents using machine learning and implemented in python [23] [24].

The proposed algorithm uses natural language processing to process and analyses the insights of the SMS text. Two classifiers namely Support Vector Machine and Naïve Bayes is selected to analyses the processed text and to classify it into spam SMS (fake) or ham SMS (true). The mission of this exploration is effectuated to pursue by the following course of actions. (i) To develop a system that accurately classifies the spam SMS and ham SMS. (ii) To make a comparative study between two important classifiers namely Naïve Bayes and Support Vector Machine and check for its accuracy. (iii) To figure out an approach that can be subsumed into the mobile phone as an option to detect spam automatically.

The section of this paper is made therefore like section 2 gives out the best in class on survey recognition on different domains and examines about the destinations of this work. Section 3 talks about methodology and implementation of the technique. Subsequently, section 4 features and talks about the essentially acquired outcomes and future bearings.

2. LITERATURE SURVEY

Krishnaveni et al. [1] proposed a methodology for identifying the Spam Reviews using the Natural Language Processing for Preprocessing techniques and Neural Networks Classifier. The Features of Dataset is considered for classification with Multiple Features based on NLP and the Reviewers characteristics. The Polarity of the Text is also considered as a Feature [1].

Dipak et al. [2] have used the dataset of spam SMS to predict whether the messages is spam or ham. Natural language processing (NLP) steps like are done in the case of content based text message to detect whether the messages send is spam or ham. Error messages are predicted using different Statistical Techniques. The algorithms like Support Vector Machine, Neural Network and Relevance vector Machine are used and analyzed the best accuracy rate.

Ahmed et al. [3] used N-gram analysis to predict online fake news. The analysis was based on the Term Frequency-Inverted Document Frequency and Linear Support Vector Machine are used for the Machine Learning techniques and NLP process are included for the evaluation of the text. Term Frequency-Inverted Document Frequency and Term Frequency are calculated for Uni-gram, Bi-gram, Tri-gram and Four-gram.

Jabbar et al. [4] has predicted the spam e-mails which can contain the phishing or malware that can harm the system or it can steal confidential information, so it is important to analysis the fake emails. Negative Selection Algorithm (NSA) is used for the anomaly detection for spam filtering techniques. E-mails are scanned to analyses the text content and the process of tokenization and stop word removal process are implemented for the analysis of the e-mail. Based on the content and based on the true positive and true negative values spam emails are detected.

Alkahtani et al. [5] used the spam SMS dataset to analyses the messages based on the text content and by using filtering techniques. Based on techniques like backlist, white list, challenge response system and origin diversity analysis the detection of spam SMS are found. Filters like Heuristic filter, Rule based filter and Genetic algorithm, Artificial Neural Networks, Decision tree techniques and Clustering Techniques are used for the prediction of the spam SMS messages.

Sarit Chakraborty et al. [6] has analyzed the e-mails and predicted out of which it is spam or ham e-mails. The detection is based on the machine learning techniques and used Cumulative Weighted Sum (CWS) for the higher level of accuracy rate. E-mails are classified into content based and image based mails. Basis of weight fixation techniques like Frequency based weight fixation, Matrix based weight fixation, Tree based weight fixation are used for Cumulative Weighted Sum techniques for analysis of spam e-mails.

Shafigh et al. [7] has used the email spam to predict whether the content is spam or ham. Based on the blacklist and white list filtering techniques and the process of Multilayer Perceptron (MLP) and the algorithms like Naïve Bayes and C4.5 decision tree classifier are used for the prediction. In Multilayer Perceptron, the neural networks and the activation of the neurons are calculated for the output. Email header information analysis and keyword matching and messages are implemented to analysis and to predict the output.

Torabi et al. [8] used Support Vector Machine (SVM) for classification and filtering then other machine learning processes are used to detect the spam SMS. End-User techniques and Server side techniques and content base learning spam filtering architecture and spam detection are analyzed for the prediction spam.

Meli et al. [9] has used the Reverse Polish Notation (RPN), Naïve Bayes, Linear Genetic Programming and Genetic programming are used for the spam detection. Blacklist and heuristics used for the machine learning and text classification methods are used. Feature extraction, fitness evaluation, feature results are used for the multi-threading.

Khan et al. [10] used text mining techniques to detect spam SMS and the preprocessing techniques and RapidMiner are used for the prediction of the spam SMS. Preprocessing techniques and classifiers and Performance Evaluation are used. Performance Evaluation are calculated to obtain some terms like Error rate, Accuracy, Recall, Precision, Execution Time and F Measure. The algorithms like Naïve Bayes, Decision tree and Support Vector Machine are used for the detection of spam messages.

Atanasova et al. [11] has processed the email messages based on the preliminary processing. The spam filtering process and neural networks and Multilayer Perceptron are used for the detection process. Based on the trained set the remaining are also classified and analyzed to predict the spam email.

Sajedi et al. [12] used machine learning techniques for the detection of the spam SMS. Based on the Performance Measurement Criterion like Recall, Precision are calculated for evaluation.

Dada et al. [13] predicted by using content based filtering techniques, Case base spam filtering method, Heuristic or Rule based spam filtering techniques are used for the prediction. NLP process are performed for the text based content to analyses the spam SMS. The performances is carried out based on the Classification Accuracy and Classification Error. Classification is done to classify the message into spam or non-spam.

Daisy et al. [14] have predicted the hybrid spam filtration by using machine learning techniques by implementing the Naïve Bayes and Markov Random Fields algorithm to provide accurate rate. The probability rate is used in the Naïve Bayes algorithm. Markov Random Fields uses the property of the classifiers to classify the messages.

Hijawi et al. [15] used spam features techniques and other feature techniques for the detection of the spam SMS. Feature Extraction tool is used for the analyses of the text which is based on the occurrence of the words and the frequency of the words. The rate of accuracy is determined by using confusion matrix, Precision and Recall.

Senthil Murugan et al. [16] used the machine learning techniques for detecting the spam messages through social networks. Based on the algorithms like Naïve Bayes, Rule induction, Decision tree and SVM the prediction is done by using machine learning techniques in it to provide more accuracy.

Ibrahim et al. [17] used E-mail dataset for predicting the spam mails by using Bayesian Classifier and curing techniques for analyses purpose. Based on the probability rate the Bayesian classifiers workout for the prediction process and establish the correct accuracy rate.

Susila Devi et al. [18] have used email spam filtering techniques like Naïve Bayesian Classifier, K-Nearest Neighbor, Boosting, Neural Network and Support Vector Machine to analyses spam detection. Machine based learning are implemented for the analysis of spam SMS.

Revar et al. [19] have predicted the spam E-mails through different types of spam filtering techniques by using SVM. The parameters like e-mail address, content, URL are used for the analysis. The information is extracted and normalization process done and the statistical analysis are performed for the accurate result.

Sharama et al. [20] have analyzed based on the origin based technique which is related to the blacklist, whitelists and Real-time Blackhole List (RBL) to predict the E-mail spam. Content based spam detection techniques are also implemented to detect the content based messages].

Bhowmick et al. [21] used machine learning techniques and other spam filtering and image spam for the detection of the spam emails and to analysis the content and images in emails. Word obfuscation, Bayesian Poisoning attack, Backscatter spam and Image spam are analyzed for the detection of the email spam.

By and large investigation of the papers inspected above, served to be a useful factor to clarify choice for the proposed system. The papers [1]-[3] [5] [8] [10] [12] [13] [16]-[21] which utilized the significant characterization calculations for the location of spam surveys demonstrated that Support Vector Machine calculation functioned admirably with higher precision rate than other regular algorithms. On the other hand, papers [14] Support Random Forest algorithms and papers [2] [6] [10] [13] [14] [16] [21] support Naïve Bayes to the best classifiers. The paper [8] [12] [20] are survey papers, which gave extraordinary comprehension towards the different calculations utilized for recognition of spam SMS and messages. Subsequently this dataset has been utilized for the proposed procedure too.

3. METHODOLOGY AND IMPLEMENTATION

The following processes are carried to classify the SMS data into Spam or Ham.

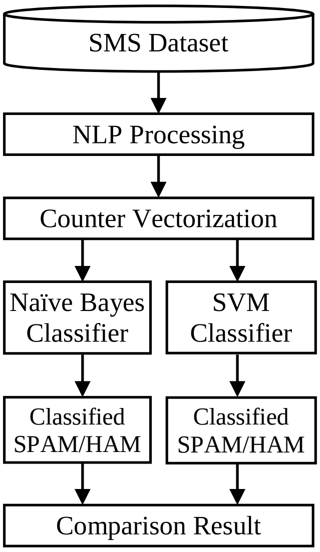

Fig.1. Overall Methodology

This includes i) Data collection ii) NLP Preprocessing iii) Counter vectorization iv) Classification using Naïve Bayes and SVM Classifiers v) Evaluation. The Fig.1 shows the overall methodology for attaining result.

3.1 DATA COLLECTION

This procedure includes the assortment of information from primary source. The dataset can be disconnected by different techniques with the end goal of examination. A few techniques are i) Download from online sources, ii) Use crawlers to flock together data from web pages. iii) Manually gather information from every client as far as survey, archives, and interviews [1].

3.1.1 Dataset Description:

The SMS spam collection dataset is taken from Kaggle repository. It contains 2 attributes with 5573 instances namely ham/spam and text message.

3.2 NLP PREPROCESSING

The natural language processing have the capacity for putting down multifarious pragmatic functions that can be engaged in various activities serving as preprocessing in terms of stemming, stop word removal, lemmatization, POS tagging, bag-of-words, n-gram analysis, etc [1].

3.2.1 Tokenization:

It alludes to separating a bigger assortment of text into littler lines, words for a non-English language. Each element that is a piece of words break off hinge on rules is called as a token. Kinds of tokenization of text are Sentence Tokenization and Word Tokenization [1] [3]. Sentence tokenization parts each sentence from a section and spares as independent sentence, when it experiences a full stop (.) or a semi-colon (;). Word tokenization parts each word in a sentence and stores it, when it experiences a whitespace or an accentuation with the exception of punctuation (‘) [24].

3.2.2 Stop Word Removal:

A stop word is an ordinarily utilized word, (for example, “the”, “an”, “an”, “in”) that a web crawler has been modified to disregard. At the point when information is oppressed for preparing, these stop words can be evacuated effectively [24].

3.2.3 Stemming:

It is a Text Normalization method in the field of Natural Language Processing [24]. Stemming is the way toward attaching the inferred words to their promise stem, base or root structure. Stemmers expel morphological fastens for the most part additions from words, leaving just the root word called as stem. The stemmer, which is utilized in this strategy is Snowball Stemmer. Snowball is a bit of string getting ready language proposed for making stemming estimations for utilizing Information Retrieval [1], [24]. A portion of the principles are:

- **3+ies→Y**: this rule removes the ‘ies’ and replaces by ‘y’ (applies → apply)
- **4+ing→**: this rule removes the ‘ing’ and does not replace anything (fishing→fish)
- **3+s→:** this rules removes ‘s’ from the end of the word (cats →cat)

This is the means by which stemming works. An intriguing reality about English language is in spite of its negligible use of i-postfixes, it has such an unpredictable stemmer.

3.3 COUNTER VECTORIZATION

The number of each word occurrence in a text or article, book or in a document is called as counter vectorization. The number of occurrences of each word is counted. Count Vectorizer, Tf-idf Vectorizer, Hashing Vectorizer are some of the variants of vectorization.

- **Count Vectorizer:** It converts a collection of text documents to a matrix of token of unique words counts. It discovers every single one of a kind word in text-set and makes as one vector. It changes over every content to a variety of exceptional words in an array considers and an outcome, we have one vector of special words and numerous exhibits as an array with many tally of zero.

The way toward changing over content into vector is called vectorization. By utilizing Count Vectorizer work we can change over content report to framework of word tally. Network which is created here is Sparse Matrix. In the wake of applying the Count Vectorizer we can plan each word to feature. This can be changed into Sparse Matrix.

- **TF-IDF:** TF-IDF speaks to Term Frequency-Inverse Document Frequency which in a general sense tells hugeness of the word in the corpus or dataset.
- **Term Frequency (TF):** Term Frequency is described as how customarily the word appears in the report. As each sentence is definitely not a comparable length so it may be possible a word appears in long sentence happen extra time when stood out from word appear in sorter sentence.

TF = (No of time word appear in the document)/(Total no. of word in the document)

- **Inverse Document Frequency (IDF):** It is an idea which is for discovering noteworthiness of the word. It relies upon how less successive words are more helpful.

IDF = $\log_{10}$(Number of Document/Number of document in which word appear)

3.4 CLASSIFICATION PROCESS

Classification is the Supervised Learning process in Machine Learning Technology to characterize the perceptions through measurable or relapse examination. There are numerous Classification Algorithms are there in the Literature. In this technique the generally utilized and acknowledged Algorithms for Text Classification procedure, for example, Support Vector Machine (SVM) and Naïve Bayes are utilized.

3.4.1 Support Vector Machine Classifier:

Support Vector Machine (SVM) is a supervised machine learning algorithm accomplished of performing classification, regression and to detection outlier. The linear SVM classifier works by splitting into two classes. Based on the features selected the data points will be grouped into one class and the other features will be labeled into another group of class. It can deal with multiple continuous and categorical variables. SVM creates a hyperplane in multidimensional space to spilt different classes. SVM provides optimal hyperplane to reduce the errors in problem. The fundamental idea of SVM is to discover a maximum marginal hyperplane (MMH) that most rightly dissociate the dataset into classes. In the SVM calculation, it is anything but difficult to group utilizing linear hyperplane between two classes. Yet, the inquiry emerges here is this aspect can be reckon up of SVM to distinguish hyper-plane. So the appropriate response is no, to take care of this issue SVM has a method that is normally known as a Kernel trick. Kernel is the capacity that changes information into a reasonable structure. There are different kinds of Kernel Functions utilized in the SVM calculation for example Polynomial, Linear, non-Linear, Radial Basis Function, and so on. Here utilizing portion stunt low dimensional information space is changed over into a higher-dimensional space.

3.4.2 Naïve Bayes Classifier:

Naive Bayes relies upon Bayes' Theorem with a doubt of independence among pointers. In fundamental terms, a Naive Bayes classifier acknowledges that the proximity of a particular segment in a class is irregular to the closeness of some other component. Whether or not these features depend upon each other or upon the nearness of various features, these properties openly add to the probability that this natural item is an apple and that is the explanation it is known as 'Guileless or Naive'. Naive Bayes model is definitely not hard to gather and particularly accommodating for astoundingly enormous instructive assortments. Close by ease, Naive Bayes is known to beat even especially complex portrayal procedures. Bayes speculation gives a technique for finding out back probability *P*(*c*|*x*) from *P*(*c*), *P*(*x*) and *P*(*x*|*c*) [22].

$$P\left(c|X\right)=\frac{P\left(x|c\right)P\left(c\right)}{P\left(x\right)} \tag{1}$$

$$P\left(c|X\right)=P\left(x_1|c\right)\times P\left(x_2|c\right)\times\ldots\times P\left(x_n|c\right)\times P\left(c\right) \tag{2}$$

Here, (*c*|*x*) is the posterior probability of class (*c*, target) given predictor (*x*, attributes). *P*(*c*) is the prior probability of class. *P*(*x*|*c*) is the likelihood which is the probability of predictor given class [22]. *P*(*x*) is the prior probability of predictor [22].

3.5 DATA EVALUATION

Data evaluation is done to check the performances of the algorithms, comparing analysis on different algorithms can be done, statistical evaluation, calculation of risk factors can be analyzed, development of data visualization can be performed, and grouping of data can be established. Here the performance of the classifier is done by using the Accuracy, Precision, Recall and F-Measure by introducing the confusion matrix.

4. SOFTWARE USED

Python in Spyder Integrated Development Environment (IDE) in Scientific Python Development IDE

5. RESULTS AND DISCUSSIONS

The Dataset is stacked and NLP Preprocessing is finished. The count vectorization is performed and it is done to recognize the quantity of unique words in the dataset. The absolute number of exceptional words in the dataset which is 13504 is found. The all out occurrences 4457 from the Dataset are ordered into Spam or Ham. The accompanying Table.1 gives the quantity of Spam SMS and Ham SMS distinguished by utilizing both the classifiers.

Table.1. Classification of Spam and Ham SMS

Classifier	Ham SMS	Spam SMS
SVM	4788	785
Naïve Bayes	4659	914

The following Fig.2 provides the average percentage of Spam SMS (15%) and Ham SMS (85%) of the overall dataset by using both the classifiers. It shows that the Spam SMS is lesser than the Ham SMS.

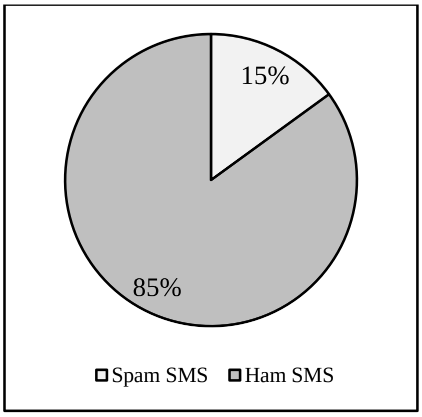

Fig.2. Percentage of Spam and Ham SMS Classification

The Table.2 and Fig.3 the Comparison chart explains the performance of Naïve Bayes and SVM classifiers for the classification process of the given Dataset. The Accuracy, Precision, Recall and F-Measure of both the classifiers are compared and found SVM is higher than the Naïve Bayes in all criteria.

Table.2. Performance Comparison of Classifiers

Classifier	Accuracy	Precision	Recall	F-Measure
SVM	93.02	90.88	91.45	93
Naïve Bayes	94.32	92.84	93.07	94

6. CONCLUSION AND FUTURE SCOPE

The technological world appearances changed sorts of spams each day. However some methods of conveying these spams focus on the people straightforwardly and the effect of such spam violations are tremendous much of the time. One such spams are SMS spam. The untrustworthy gathering of individual or association focuses on the portable starting at any sort and send spam SMS and control their enthusiasm to reaction emphatically to their spam trap. Consequently, the proposed strategy demonstrated to distinguish such spam utilizing current classifiers, for example, Naïve Bayes and Support Vector Machine. Alongside the identification of spam SMS a similar report between both the classifiers utilized were done, which came about that SVM works better than Naïve Bayes with an exactness of 94.32%. The other performance measures are additionally demonstrated that the SVM classifier works better than the Naïve Bayes classifier for recognizing Spam SMS.

This proposed research extensions to utilize diverse datasets which contains spam of various thought process. This model can be joined in cell phones as an alternative to distinguish the got SMS whether it is spam or ham. Further the dataset can be gathered through crawlers to get genuine information from the clients. The scalability of the dataset can be checked in this model. The performance of the classifier can be improved by introducing ensemble classifiers.

REFERENCES

[1] N. Krishnaveni and V. Radha, "Spam Review Predictions through Multiple Features using Sentiment Analysis and Neural Networks", *Journal of Critical Reviews*, Vol. 7, No. 12, pp. 2810-2817, 2020.

[2] K.O. Kawade and K.S. Oza, "Content-based SMS Spam Filtering using Machine Learning Technique", *International Journal of Computer Engineering and Applications*, Vol. 7, No. 4, pp. 1-12, 2018.

[3] H. Ahmed, I. Traore and S. Saad, "Detection of Online Fake News using N-Gram Analysis and Machine Learning Techniques", *Proceedings of International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127-138, 2017.

[4] A.J. Saleh, A. Karim, B. Shanmugam and S. Azam, "An Intelligent Spam Detection Model based on Artificial Immune System", *Information*, Vol. 10, No. 6, pp. 209-216, 2019.

[5] Hasan Alkahtani, Paul Gardner Stephen and Robert Goodwin, "A Taxonomy of E-mail Spam Filters", *Proceedings of International Arab Conference on Information Technology*, pp. 351-356, 2014.

[6] D. Sen, C. Das and S. Chakraborty, "A New Machine Learning based Approach for Text Spam Filtering Technique", *Communications on Applied Electronics*, Vol. 6, No. 10, pp. 28-34, 2017.

[7] A.S. Aski and N.K. Sourati, "Proposed Efficient Algorithm to Filter Spam using Machine Learning Techniques", *Pacific Science Review A: Natural Science and Engineering*, Vol. 18, No. 2, pp. 145-149, 2016.

[8] Z.S. Torabi, M.H. Nadimi-Shahraki and A. Nabiollahi, "Efficient Support Vector Machines for Spam Detection: A Survey", *International Journal of Computer Science and Information Security*, Vol. 13, No. 1, pp. 11-19, 2015.

[9] C. Meli and Z.K. Oplatkova, "SPAM Detection: Naive Bayesian Classification and RPN Expression-Based LGP Approaches Compared", *Proceedings of Online Conference on Computer Science*, pp. 399-411, 2016.

[10] Z. Khan and U. Qamar, "Text Mining Approach to Detect Spam in Emails", *Proceedings of International Conference on Innovations in Intelligent Systems and Computing Technologies*, pp. 45-56, 2016.

[11] T. Atanasova, S. Parusheva and E. Kostadinova, "Spam Filtering through Neural Network", *International Multidisciplinary Scientific Geoconference*, Vol. 1, No. 2, pp. 383-388, 2016.

[12] H. Sajedi, G.Z. Parast and F. Akbari, "SMS Spam Filtering using Machine Learning Techniques: A Survey", *Machine Learning Research*, Vol. 1, No. 1, pp. 1-14, 2016.

[13] E.G. Dada, J.S. Bassi and H. Chiroma, "Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems", *Heliyon*, Vol. 5, No. 6, pp. 1-16, 2019.

[14] S. Jancy Sickory and S. Rijuvana, "Hybrid Spam Filtering Method using Machine Learning Techniques", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 9, pp. 1-12, 2019.

[15] H. Faris, J.F. Alqatawna and I. Aljarah, "Improving Email Spam Detection using Content Based Feature Engineering Approach", *Proceedings of Jordan Conference on Applied Electrical Engineering and Computing Technologies*, pp. 1-6, 2017.

[16] N.S. Murugan and G.U. Devi, "Detecting Spams in Social Networks using ML Algorithms-A Review", *International Journal of Environment and Waste Management*, Vol. 21, No. 1, pp. 22-36, 2018.

[17] D.S. Ibrahim, "Hybrid Approach to Detect Spam Emails using Preventive and Curing Techniques", *Journal of Al-Qadisiyah for Computer Science and Mathematics*, Vol. 10, No. 3, pp. 1-16, 2018.

[18] K.S. Devi and N. Supriya, "Overview of Content Based Spam Filters Techniques and Similarity Hashing Algorithms", Master Thesis, Department of Computer Science and Engineering, Raghu Institute of Technology, pp. 1-78, 2017.

[19] Pooja Revar, A. Shah, J. Patel and P. Khanpara, "A Review on Different Types of Spam Filtering Techniques", *International Journal of Advanced Research in Computer Science*, Vol. 8, No. 5, pp. 1-12, 2017.

[20] M. Sharama and S. Sharma, “A Survey of E-mail Spam Filtering Methods”, *Control Theory and Informatics*, Vol. 7, No. 1, pp. 1-8, 2018.

[21] A. Bhowmick and S.M. Hazarika, “Machine Learning for E-Mail Spam Filtering: Review, Techniques and Trends”, *Heliyon*, Vol. 5, No. 6, pp. 1-12, 2016.

[22] Data Science, Available at: www.https://towardsdatascience.com.

[23] S. Bird, E. Klein and E. Loper, “*Natural Language Processing with Python: Toolkit*”, O'Reilly Media, 2009.