Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 04, April 2021: 2179-2191

Research Article

Detecting Spam Email Using Machine Learning

A Jasmine Gilda, P Ravi Tejaswini, P Sahithi, K Likitha

Department of Computer Science and Engineering, R.M.K. Engineering College,
Thiruvallur, India
Email: ajg.cse@rmkec.ac.in

Abstract

In everyday life, email became the most affordable also simple method of communication for both official works and business applicants because of simple convenience of web access. The spam mails are increasing with the growth in internet users and Gmail end user. People misuse by sending unwanted emails for commercial purposes and fraudulent purposes. Emails are not the only way to send the spam messages, they can also be found in SMS, forums, social media etc. Some spam contains attachments that if it is opened the computer can be infected with viruses or malware. The data classification method is used in the email filtering. When it comes to data classification, choosing the best-performing classifier is a critical step. Many researches provided many techniques to detect these spam emails and improve the accuracy by using machine learning (ML) algorithms. Both naive bayes (NB) classifier and binomial logistic regression had the option to detect the spam mails as naive bayes can be used to classify large data whereas logistic regression is a statistical method respectively. These algorithms are performed on a ling-spam dataset taken from kaggle website. In the proposed system various datasets are performed on the dataset. The result of the proposed model will be compared with the base models to conclude whether the implemented models have improved the performance and evaluation.

Keywords: *Machine learning techniques, naïve bayes, multinomial naïve bayes, particle swarm optimization, bio-inspired algorithm, genetic algorithm, NLP, Logistic Regression*

Introduction

Machine learning algorithms or techniques can be used for multiple purposes in the field of computers. The techniques mainly focus on predicting, analyzing, detecting, recognition. The most common machine learning tasks are clustering, regression, classification. As emails are the most used for the communication purpose, there are many people spamming the contents by sending unwanted junk, spoofing. Email spam detection classifies the mails into spam and not spam, this is identified as supervised learning. Since the early 1990s, email spam has gradually risen, and by 2014, it was projected to account for about 90 percentage of all email network traffic. In 2007, It is assessed that spam cost organizations on the request for $100 billion. We receive approximately 40-50 emails per day with almost 60-70% of spam mails that will be difficult to classify manually. Apart from sending the unwanted mails these can also cause security issues in the computer system. Spammers acquire email addresses from various sources, like chat rooms, blogs, list of customers, newsgroups, and malwares that collects users' contact books. The email addresses are also marketed to other spam business or spammers as well.

[12]Anti-spam methods are classified into four groups:

those that involve individual intervention, automated by email managers, automated by email senders, and those that are used by researchers and law enforcement officials. The **end-user’s methods** are discretion, address munging, avoiding responding to spam, contact forms, disable HTML in emails, disposable email addresses, ham passwords, and reporting spam.

The **techniques automated by email administrators** are authentication, challenge systems, checksum-based and country-based filtering, DNS-based blacklists, URL filtering, Strict enforcement of RFC standards, honeypots, hybrid filtering, outbound spam protection, PTR or reverse DNS checks, rule-based filtering, SMTP callback verification, SMTP proxy, Spam trapping, statistical content filtering, tar pits.

The **techniques automated by email senders** are limit email backscatter, background checks on new users and customers, confirmed opt-in for mailing lists, egress spam filtering, port 25 blocking, port 25 interception, rate limiting, spam report feedback loops, from field control, strong AUP and TOS agreements.

[2]The detection of spam emails is classified into two ways: Knowledge engineering is the traditional approach and machine learning techniques are the advanced models. Knowledge engineering is a network-based strategy in which it consists of some rules along with IP addresses and network addresses. Though we get appropriate results if the approach takes a lot of time and also the maintenance and setting up the rules is not convenient to the users whereas machine learning makes the process easier by recognizing the spam emails accordingly, then applies the trained commands to the incoming emails.

To identify the spam emails in the given network the companies offer many tools and technologies. [1] To handle these spams the email suppliers such as google and yahoo mail worked with different machine learning techniques as it can adapt to any conditions. The filters check the mails by using the existing rules and they even come up with new rules based on what they've learned from spam filtering operations. Although by using these rules the spam mails succeed in evading their spam filtration process. According to google statistics, 50-70% of the emails that users receive are unsolicited mails. Content and case-based spam filtering, heuristic and rule-based spam filtration, previous likeness-based filtration, and adaptive spam filtering are some of the spam filtering techniques used to combat spam emails.

We know that Google has data centers to maintain the customers data. Before the email enters into the mailbox there are hundreds of rules to segregate the mails in the data centers. Gmail/Google uses the following spam filters: blatant blocking, bulk email filter, category filters, null sender disposition, and null sender header tag validation.

In this paper, the multinomial naive bayes algorithm (classifier) is combined with particle swarm optimization and Genetic algorithm as a traditional and existing approach. The Bayes theorem, which has strong independence and probability distribution properties, is the foundation of naive bayes. Particle swarm optimization is a type of intelligence derived from natural species such as birds, fish etc. A genetic algorithm is a search-based natural selection method optimization. The base models such as support vector machine, random forest algorithm, xgboost algorithm and natural language processing (NLP) are performed on the ling spam dataset. These are performed to compare the precision, recall, f-measure, and accuracy of the proposed algorithm i.e, logistic regression and naive bayes algorithms.

Literature Survey

[5]The researchers have conducted studies to identify the unsolicited emails. Most of the review is done by combining the machine learning classifiers as the algorithm and optimizing it to improve the accuracy. Till today finding the most accurate algorithm is the challenging problem.

In [1] study, a research was carried out on some of the most commonly used machine learning (ML) algorithms that were effective in detecting the spam emails. The article explains some of the key concepts, attempts, performance, and review trends in spam classification or filtration. The review is done on the techniques that are applied by the leading service providers like Google, Yahoo Gmail system. The advantages along with disadvantages of the algorithms are discussed.

In [3] Using two machine learning classifiers, SVM-NB, a hybrid framework is proposed. The approach involves using the SVM algorithm to build a hyperplane between the dimensions and removing data points from the training dataset. The dataset is employed to predict the result using the NB algorithm. The accuracy of SVM-NB is higher when compared to the accuracy of SVM and NB.

In [6] an integrated approach of decision tree (DT) and GA is proposed to generate more accurate solutions in detection of spam emails. The problem of high dimensionality curse will arise while trading with any implementation of text classifications. So, the feature extraction step is the key step to reduce and remove the unwanted features. By reducing the feature space, the training of the model’s speedups and improvise the precision and accuracy of the classifier. In this given paper, principal component technique is used to eliminate the unsuitable features.

In [5] The combination of logistic regression (LR) and decision tree is implemented. LR is used to decrease the noise before induction with a decision tree. DT can handle the numerical and nominal attributes, and can handle the training data when the attributes are missing. The drawback of DT is over-fitting or sensitivity to the training dataset, noise data, unwanted data that may reduce the accuracy and performance. As the noise data present in the training dataset the algorithm suffers from over or under fitting and the accuracy might be decreased.

In this existing system there are certain limitations i.e., as the system deals with the text data the time complexity will be high. There are certain proposed systems that face NP-hard problem to select the set of attributes.

Author & Year	**Technique**	**Disadvantages**
[13] Mohamad & selamat [2015]	Machine learning	The shape, texture of the image sent in email is ignored and only text form is considered.
[14] Youn & Mcleod [2007]	Decision tree	The system works on the mails in csv format.

[15] Al-Shboul et al. [2016]	Random forest algorithm	The experiment consists of inconsistent data with 25% of spam emails.
[16] Mujtaba et al. [2017]	Machine learning	The review explains different algorithms but did not explain the tools and feature extraction for classification.
[17] Harisinghaney et al. [2014]	K-nearest Neighbor and naïve bayes	The approach has high time complexity and the text recognition is only for certain fonts,
[18] Faris et al. [2015]	neural network	The algorithm is trained after every repetition.
[19] Tuteja [2016]	Artificial neural network	The system did not specify the pros and cons of the system.
[20] Ajaz et al [2017]	Secure hash algorithm	The proposed system classifies the emails but failed to break the misuse of network band width and the storage.

Basic Concepts

The algorithms we used in detecting the spam is discussed

3.1 Naive bayes approach

Using probability methods, this method is used to solve classification problems. It is capable of handling massive datasets. In naïve bayes, the probability distribution is determined using the dataset's frequency distribution. The MNB classifier, which focuses on term frequency, uses a multinomial distribution for the given function.

3.2 .Bio-inspired algorithms

The metaheuristic optimization that is inspired by biological behaviors of animals or birds and have been used to find ideal solution to the problem statement. We will be using particle swarm optimization and genetic algorithm to get the accurate solution for classifying emails.

3.3 .Particle swarm optimization (PSO)

Swarming-based approaches such as those seen in fish or birds are recognized as the PSO. The particles are judged on their optimal position as well as their overall global position. To find the global best location, the particles in the search space are dispersed. Different calculations and techniques, such as function selection and parameter tuning, are available in library i.e., Pyswarms library. The selection of the feature process can save storage by eliminating features that aren't needed during classification. As a result, the Particle swarm optimization will be used to regulate and find the model's parameters.

3.4 Genetic algorithm

The method is a Darwinian natural selection-based evolutionary algorithm that selects the required individual from a given number of species. It is based on the principles of variation, selection and past generations. The algorithm is fixed with a certain size, and each person is assigned a unique number in binary format. It iterates through a fitness mechanism that selects the best individuals for offspring reproduction. The implementation is done with the TPOT library and cross-validation training.

Methodology

The proposed application should be able to identify the spam and ham emails. The feature representation is done by count vectorizer, to convert the collection of text documentation to vector of terms. To implement the model the steps to be followed are

- Preparing text data
- Feature extraction process
- Creating word dictionary
- Training the classifier
- Running the predictions

At first, to perform the system we need a dataset of emails that contains both spam and ham mails. We need to prepare the text data. We will preprocess the data.

We will extract the features that are needed for training the machine learning algorithm. The punctuations, stop words, numeric data and other symbols will be removed in this preprocessing.

After removing the unwanted features, a word dictionary is created which contains the unique words of spam and ham mails. This data will be used to classify the incoming new emails as spam and ham. In logistic regression, if the probability of the message is less than 0.5 it is considered as ham and if the probability is greater than 0.5 it is considered as ham.

The dataset will be divided into two sets, where we take 80% of dataset to train the machine or classifiers. Later, the remaining 20% of the data is used to test the data and predict the outcome of the algorithm.

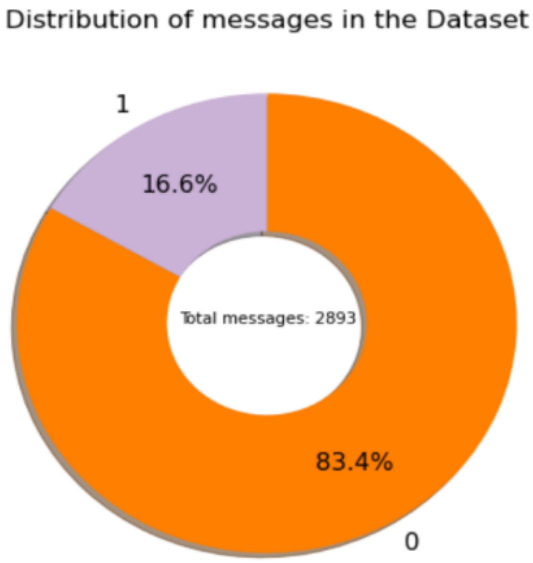
The supervised learning approach will be used since email spam detection is a classification category. We will divide the dataset into two sections i.e., training data and testing data, part of the supervised learning methodology. This takes the training data as input and then evaluates the classifier.

The main goal is to train a classifier with the dataset taken and parameters, then analyses the outcome of a testing dataset that the classifier is unfamiliar with. The advantage of proposed system i.e., The system preprocesses the dataset and extracts the useful feature so the time complexity is low. The selection of the attributes does not cause NP-hard problem. The system provides the accuracy of 99.47% by using logistic regression and naïve bayes classifier.

4.1 Feature Extraction

In this step, the proposed system transforms the dataset into a way that we can train the machine to predict the outcome. The TFIDF vectorizer is used to extract the features of the dataset. The features that are extracted

are stored in dictionary by using bad of words.

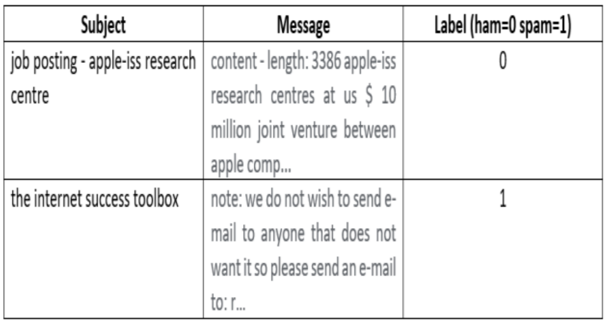


Algorithm

The proposed system will be implemented in python with the following libraries scikit-learn, flask, pandas, keras, tensor flow and some other mandatory libraries. The dataset we are working on is Ling-spam dataset downloaded from Kaggl.com.

It contains 2,893 spam and ham emails taken from the *Linguist List*. These messages focus on the job posting, educational courses, researches and discussions. 2412 ham emails, 481 spam emails.

The dataset contains



Subject	Message	Label (ham=0 spam=1)
job posting - apple-iss research centre	content - length: 3386 apple-iss research centres at us $ 10 million joint venture between apple comp...	0
the internet success toolbox	note: we do not wish to send e-mail to anyone that does not want it so please send an e-mail to: r...	1

Email spam detection is done the taken dataset by applying feature extraction techniques

- Count vectorization
- TF-IDF vectorization

The machine learning algorithms applied are

- Naïve bayes algorithm

- Logistic Regression

The traditional approaches that are combined with multinomial naïve bayes algorithm are

- GA
- PSO

The base model algorithms that are applied to compare with the proposed system are

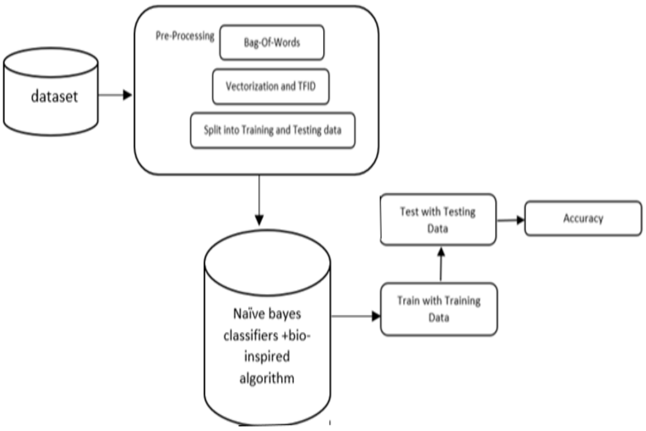
- SVM
- Random forest
- NLP
- Xgboost

The logistic regression is bid to the data that is cleaned before applying the algorithm to predict the appropriate result and avoid unwanted tokens while performing the algorithm. To classify the data into spam and ham comes under binomial logistic regression if it is more than two independent values or categories then logistic function will be used. For now, to divide the emails we use the binomial logistic regression algorithm. We also use naïve bayes classifier as to classify the large dataset.

An integrated definition of Naïve bayes and the traditional approach i.e., particle swarm optimization, with Naïve bayes providing a probability distribution that defines the classification of the email as spam and non-spam emails based on keywords that are in email dataset, and PSO will be used for the future optimization of the NB algorithm to achieve better accuracy. The Count vectorization method is used to get the necessary features from a (BOW) bag of words for text classification.

When an email is extracted from the ling spam dataset, it is assumed to be in the raw format. To complete the feature extraction and classification process. To begin, the dataset's emails should be pre-processed. Tokenization, stemming, and stop words elimination are all steps in the pre-processing process. Initially the tokenization is done by using scikit learn library i.e., count vectorization to tokenize the email into individual words and splitting the words into different tokens. The stop words such as a, an, and etc are to be removed from the tokens. Stemming is the method of reducing a word to a word stem that connects to suffixes, prefixes, and root words (lemma).

BOW (bag of words) is a technique for removing features from text documents. After removing the features, the remaining words or text will be used to train a dataset. It also establishes a vocabulary of all the documents' special words. The TF-IDF is a statistical measure that assesses the relevance of a word in a series of documents. This is achieved by multiplying two metrics: the frequency a word appears in the documents and the word's inverse document frequency over a range of documents. The dataset will be divided in the ratio 80:20 i.e., training and testing data.



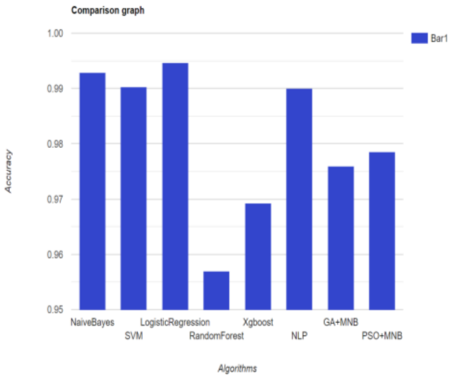
The CFS method is used to choose the required featured words from the cleaned data; it only chooses the feature set that is most related to the class. The probability distribution of the tokens is determined using naive bayes. By using particle swarm optimization method, the parameters of the NB classifier are optimized. The features are taken as particles. These particles will be flying randomly to look for the best match for tokens. The tokens will be matched and find the local solution and global solution. The performance measures of every particle rely on the features related that are to be optimized. Based on the feature evaluation similarity using Particle swarm, the tokens will be classified as spam and ham.

Conclusion

We have successfully implemented the proposed algorithm logistic regression and naive bayes algorithm where naive bayes is used to detect the large set of spam emails which is used in real life and the logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input. The accuracy of 99.29% and 99.47% is obtained. When compared to other base models the proposed algorithm acquired better evaluation measures like accuracy, recall and precision. With the scikit-learn library and its modules, the algorithms were tested and experimented with. The genetic algorithm also worked well with multinomial naïve bayes when the dataset is distributed in the ratio 80:20 for training and testing data.

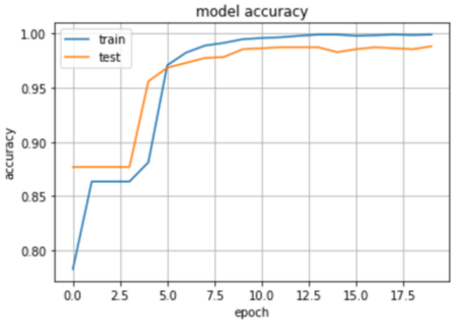
Result

The evaluation measures like precision, accuracy and recall are compared among base model algorithms and proposed algorithm. The base algorithms are SVM (Support vector machine), random forest, xgboost, Natural language processing (NLP). The traditional methods like genetic algorithm and particle swarm algorithm are combined with multinomial naïve bayes. The proposed models are logistic regression and naïve bayes algorithm.



The model accuracy of NLP in terms of epoch and accuracy is represented in the graph. The acc and val_acc is shown in the following graph. The acc(train) is referred to as the training dataset and val_acc(test) is referred to as the work of the model outside the training dataset.

An **epoch** is used to indicate the number of passes of the entire training dataset the machine learning algorithm has completed. When the datasets are large then the data is divided and grouped into batches



References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, ‘‘Machine learning for email spam filtering: Review,approaches and open research Dec. 2016, pp. 1–8, doi: 10.1109/pccc.2016.7820655.

[2] International Journal of Intelligent Engineering and Systems, Vol.11, No.3, 2018 DOI: 10.22266/ ijies2018. 0630.01.

[3] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, ‘‘A support vector machine based Naive Bayes algorithm for spam filtering,’’ in Proc. IEEE 35th Int. Perform. Comput. Commun. Conf. (IPCCC), Dec. 2016, pp. 1–8, doi: 10.1109/pccc.2016.7820655.

[4] W. Awad and S. ELseuofi, ‘‘Machine learning methods for spam E-Mail classification,’’ Int. J. Comput. Sci. Inf.Technol., vol. 3, no. 1, pp. 173–184, Feb. 2011, doi: 10.5121/ijcsit.2011.3112.

[5] A. Wijaya and A. Bisri, ‘‘Hybrid decision tree and logistic regression classifier for email spam detection,’’ in Proc. 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), Oct. 2016, pp. 1–4, doi: 10.1109/ICITEED.2016. 7863267.

[6] A. I. Taloba and S. S. I. Ismail, “An intelligent hybrid technique of decision tree and genetic algorithm for E-Mail spam detection,’’ in Proc. 9th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), Cairo, Egypt, Dec. 2019, pp. 99–104, doi: 10.1109/ICICIS46948.2019.9014756.

[7] R. Karthika and P. Visalakshi, ‘‘A hybrid ACO based feature selection method for email spam classification,’’ WSEAS Trans. Comput, vol.14, pp. 171 – 177 ,2015. [Online]. Available: https:// www. wseas. org/ multimedia /journals/computers/2015/a365705-553.pdf

[8] R. Belkebir and A. Guessoum, ‘‘A hybrid BSO-Chi2-SVM approach to arabic text categorization,’’ in Proc. ACS Int. Conf. Comput. Syst. Appl. (AICCSA), Ifran, Morocco, May 2013, pp. 1–7, doi: 10.1109/ AICCSA.2013.6616437

[9] (2019). 1. Supervised Learning—Scikit-Learn 0.22.2 Documentation. Accessed: Oct. 9, 2019. [Online]. Available: https://scikit-learn.org/stable/ supervised_learning.html

[10] (2020). Google Colaboratory. Accessed: Mar. 18, 2020. [Online]. Available: https://colab.research.google.com

[11] GeeksforGeeks. (2019). Naive Bayes Classifiers-GeeksforGeeks. Accessed: Nov. 10, 2019. [Online]. Available: https://www.geeksforgeeks. org/naive-Bayes-classifiers/problems,’’ Heliyon, vol. 5, no. 6, Jun. 2019, Art. no. e01802, doi: 10.1016/j.heliyon.2019.e01802.

[12] https://en.wikipedia.org/wiki/Anti-spam_techniques. Accessed: March 10, 2021. [Online]

[13] M. Mohamad, and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification", In: Proc. of 2015 International Conference on Computer, Communications, and Control Technology (I4CT), Kuching, Sarawak, Malaysia, pp.227- 231, 2015.

[14] S. Youn, and D. McLeod, "Efficient spam email filtering using adaptive ontology." In: Proc. of Fourth International Conference on Information Technology, Las Vegas, NV, USA, pp.249-254, 2007.

[15] H. Faris, I. Aljarah, and B. Al-Shboul, "A Hybrid Approach Based on Particle Swarm Optimization and Random Forests for E-Mail Spam Filtering", In: Proc. of International Conference on Computational Collective Intelligence, Halkidiki, Greece, pp.498-508, 2016.

[16] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues", IEEE Access, Vol. 5, pp. 9044-9064, 2017.

[17] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm", In: Proc. of 2014 (ICROIT), Faridabad, Haryana, pp.153-155, India, 2014.

[18] H. Faris, and I. Aljarah, "Optimizing feedforward neural networks using Krill Herd algorithm for e-mail spam detection", In: Proc. of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, pp.1- 5, 2015.

[19] S. K. Tuteja, "A Survey on Classification Algorithms for Email Spam Filtering", International Journal of Engineering Science, Vol.6, No.5, pp.5937- 5940, 2016.

[20] S. Ajaz, M. T. Nafis, and V. Sharma, "Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier", International Journal of Advanced Research in Computer Science, Vol.8, No.5, pp.1195-1199, 2017.

[21] C. Yang, R. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers", In: Recent Advances in Intrusion Detection, Springer Berlin/Heidelberg, pp.318-337, 2011.

[22] S. Kumar, and S. Arumugam, "A Probabilistic Neural Network Based Classification of Spam Mails Using Particle Swarm Optimization Feature Selection", Middle-East Journal of Scientific Research, Vol.23, No.5, pp.874-879, 2015.

[23] N. P. DíAz, D. R. OrdáS, F. F. Riverola, and J. R. MéNdez, "SDAI: An integral evaluation methodology for content-based spam filtering models", Expert Systems with Applications, Vol.39, No.16, pp.12487-12500, 2012.

[24] A. K. Sharma, S. K. Prajapat, and M. Aslam, "A Comparative Study Between Naive Bayes and Neural Network (MLP) Classifier for Spam Email Detection", In: IJCA Proceedings on National Seminar on Recent Advances in Wireless Networks and Communications. Foundation of Computer Science (FCS), pp.12- 16, 2014.

[25] W. Ma, D. Tran, and D. Sharma, "A novel spam email detection system based on negative selection", In: Proc. of Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09, Seoul, Korea, pp.987-992, 2009.

[26] T. S. Guzella, and W. M. Caminhas, "A review of machine learning approaches to spam filtering", Expert Systems with Applications, Vol.36, No.7, pp.10206-10222, 2009.

[27] G. Chandrashekar, and F. Sahin, "A survey on feature selection methods", Computers & Electrical Engineering, Vol.40, No.1, pp.16-28, 2014.

[28] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization", IEEE Transactions on Knowledge and Data Engineering, Vol.28, No.9, pp.2508-2521, 2016.

[29] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited", In: Proc. of Australasian Joint Conference on Artificial Intelligence, Vol.3339, Cairns, Australia, pp. 488-499, 2004.

[30] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering", In: Proc. of the workshop on Machine Learning in the New Information Age, Barcelona, Spain, pp.9-17, 2000.