ICIC Express Letters
Part B: Applications
Volume **10**, Number **9**, September **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **805–811**

A ROBUST SYSTEM FOR MESSAGE FILTERING USING AN ENSEMBLE MACHINE LEARNING SUPERVISED APPROACH

Atik Mahabub, Mohammed Innat Mahmud and Md Faruque Hossain

Department of Electronics and Communication Engineering
Khulna University of Engineering and Technology
Khulna 9303, Bangladesh
{ atikmahabub1209042; innat1994 }@gmail.com; fhossain97@yahoo.com

Received February 2019; accepted May 2019

Abstract. *With the increasing use of messages as a basic and mainstream communication implies over the Internet, there comes a risk of spam that affects the Internet and the public. By getting spam messages, Internet clients are presented to security issues and in some cases are presented to unseemly substance. In addition, spam messages squander assets as far as storage, transfer speed and profitability. What exacerbates the issue is that spammers continue designing new methods to evade spam filters. On the opposite side, the huge information streams from a huge number of people and the huge number of traits make the issue increasingly lumbering and complex. In this way, proposing transformative and versatile spam recognition models is a need. In this paper, based on Ensemble Voting Classifier, an intelligent detection system based on EVC is proposed to deal with Email detection on both ham and spam cases. Here, eleven mostly well-known machine-learning algorithms like Naïve Bayes, K-NN, SVC, Random Forest, Artificial Neural Network, Logistic Regression, Gradient Boosting, and Ada Boosting are used for detection. After cross-validation, the suited best three machine-learning algorithms are selected and used in Ensemble Voting Classifier. The experimental outcomes affirm that the proposed framework can accomplish to wonderful outcomes as far as accuracy, precision, and recall. Besides, the proposed recognition framework can effectively find the most important highlights of the messages.*

Keywords: Spam message, NLP, Voting classifier, Machine learning, Data mining

1. **Introduction.** Mails are very popular for faster and inexpensive communication. With time, its utilization for wrong purposes additionally expanded and one of the real issues is sending mass messages for promoting or other such purposes. These mass messages, ordinarily known as spam mail, have turned into a colossal issue. The most compelling motivation for the ascent of spam mail is that it is quicker and less demanding to publicize worldwide as opposed to utilizing publications or TV ads for same. As per a study, spammers can send thousands of messages without any cost everyday [1]. Such large number of spams cause the using of vast amount of storages, contributing network trafficking, requiring a lot of inbox spaces, losing of service provider, wastage of time, etc. Therefore, it is very essential to manage the spam messages effectively.

On the off chance that a spam mail is controlled, it can spare a ton of assets and consumption of the organizations. Beforehand, a technique called "Knowledge Engineering" was utilized to isolate spam mail from vital messages; however, its prosperity rate was not as much as that of spammers as they could discover a route around it by changing a letter in the catchphrases which is utilized to separate spam mail from real mail [2]. To defeat this, different strategies for classifying spam mail utilizing order were presented. These strategies utilized machine learning, man-made reasoning, and different databases to build up a framework to counter spam mail. These techniques were more effective than

DOI: 10.24507/icicelb.10.09.805

knowledge bank as the information was refreshed all the more much of the time and were equipped for learning individually [3]. With the expanding system transmission capacity and enhancing innovation spam messages have turned out to be increasingly refined and it is important to utilize propelled calculations to make effective spam filters. In spite of the gigantic measure of research works that have occurred in this circle, there is no spam filter which is 100% productive. Consequently, there is a need to grow increasingly complex and exact classifier model to take out the issue of spam messages.

Comprehensive research has been done in the field of spam filtering and numerous calculations have been utilized for the equivalent. Support Vector Machine, Bayesian, Random Forest, Decision Tree classifications for extraction of highlights are the regular methodologies utilized by researchers. The consistently changing conduct and properties of spam messages have been a subject of interest. A number of analysts have proposed different strides to upgrade the execution of spam filters.

Harris et al. connected Support Vector Machines (SVM) to the spam discovery issue [4]. The exploratory outcomes demonstrated that their strategy can altogether beat different classifiers as far as identification rate and preparing time. Another work set up by Amayri and Bouguila additionally connected SVM classifier to this issue [5]. They contemplated the effect of SVM kernels on partition of spam messages from ham messages. They demonstrated that string kernels can accomplish to better outcomes than distance-based portions.

Another mainstream machine learning strategy utilized in spam detection in the writing is the Bayesian classifier. Metsis et al. assessed four unique types of Naive Bayes (NB) classifier, which are Multinomial NB, Multi-variate Bernoulli NB, Multi-Variate Gauss NB and Flexible Bayes [6]. Their test assessment dependent on ROC curves uncovered the palatable execution of Flexible Bayes and Multinomial NB with Boolean qualities. Sahami et al. [7] connected standard Bayesian Network (BN) classifier to garbage messages discovery undertakings. The test results on genuine datasets indicated BN can discover promising outcomes as far as location rate. Other standard machine learning classification techniques, for example, K-Nearest Neighbor [8], Artificial Neural Networks [9,10] and Decision Trees [11] have been likewise used to manage spam identification issues. Aski and Sourati [12] represented a spam filter utilizing three surely understood machine learning procedures, to be specific, Multi-Layer Perceptron (MLP), Decision Tree classifier (C4.5) and NB classifier. Consequences of this exploration demonstrated that utilizing MLP with spam filters can prompt progressively exact outcomes contrasted with different systems. Another examination [13] displayed in focusing on gathering the most utilized highlights in spam detection frameworks and researched the significance of each component concerning data gain method.

By and large, it tends to be comprehended from articles that there is a solid inclination towards the utilization of machine learning strategies to spam detection. This is because of their potential learning abilities. From the other side, it is seen that the majority of the proposed methodologies concentrated on enhancing the identification exactness of spam detection models, be that as it may, just few of them have given careful consideration to distinguishing the impact of the highlights and their sorts on the precision of spam detection.

This work considers data mining techniques to detect spam. Depend on the idea of data mining which is gathering important information from a pool of data which is constructed utilizing consistent preparing sets and perception in related fields [14]. This aids in building a superior framework as we have all the data of past disappointments, achievements and current issues [15]. Data mining includes different techniques for research and learning like man-made reasoning or machine learning. Ensemble Voting Classifier (EVC) is one of the Machine Learning (ML) algorithms, which is the mixture of different ML

algorithms. Here, Ensemble Voting Classifier has been used to get maximum output from the top performed ML classifiers.

The whole work is presented on four sections as follows. Segment 1 depicts the introductory speech and the related researches in the field of spam filtering. A review about likelihood workflow or methodology and the ensemble voting algorithm are talked about in Segment 2. Segment 3 presents the experimental results. In this part, further discussions and analyses are also presented. Segment 4 is the brief summary of this work and the blueprint of the future works.

2. **Methodology.** The proposed ensemble architecture can be divided into several subsections, which is illustrated by a flowchart as shown in Figure 1.

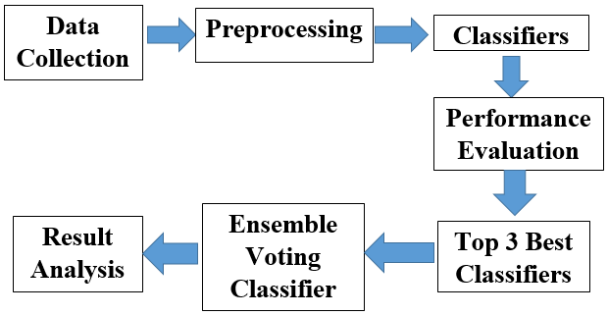

FIGURE 1. Flowchart of the ensemble architecture

Data Collection: At first, a dataset is needed with spam and ham. The proposed system is tested on the dataset of 6000 data in which about 1000 data are spam. It is a dataset of SMS spam.

Preprocessing: In reality, data index which comprises numerous missteps, needs to be refreshed and expelled so as to have exact results of the data index. In this progression data collection, it is changed and coordinated into a proper arrangement before the classifiers are connected in the data index. The data index has appropriately handled before classifiers are connected on it.

Classifier: Subsequent to having the preprocessed document, all the known classifiers, in particular, K-Nearest Neighbors, Support Vector Classifier (SVC), Ada Boost, Multi-Layer Perception (MLP), Decision Tree, Multinomial Naïve Bayesian, Random Forest, Extra Trees, Gradient Boosting, Extreme Gradient Boosting and Logistic Regression have been applied to discovery based on which spam being detected.

Performance Evaluation: Subsequent to applying all classifiers, each one of them was assessed based on execution measurements likewise test score, ROC score, precision score, and recall value so as to make sense of the best classifier.

Selection of Top 3 Classifiers: After the performance evaluation of the different well-known and commonly used classifiers like: Random Forest, Ada Boosting, Gradient Boosting, Extra Trees, Logistic Regression, K-Neighbors, Decision Tree, Multinomial Naïve Bayes, Multi-Layer Perception (MLP), Support Vector Classifier and Extreme Gradient Boosting, the top three best classifiers have been identified based on the performance. Then these top three classifiers were utilized for the next step to ensemble.

Ensemble Voting Classifier: For the Ensemble Classifier, this article considered the approach of Voting Classifier. Selected top three classifiers were utilized for these Voting Classification to get the best performance and output.

Results: In the last step, the performance of the Voting Classifier will be assessed based on execution measurements likewise test score, ROC score, precision score, recall value and F1. The results will then be compared with other relevant works for evaluating the results.

Voting Classifier Algorithm: The Ensemble Voting Classifier [16] is a meta classifier for consolidating comparative or adroitly extraordinary machine learning classifiers for classification and detection. The Ensemble Voting Classifier executes "hard" and "soft" voting.

Hard Voting: Hard ensemble voting is the easiest case of majority voting. Here, the class label Y is determined through majority voting of each classifier C_j:

$$Y = \text{mode}\{C_1(x), C_2(x), \ldots, C_m(x)\} \qquad [j = 1, 2, 3, \ldots, m]$$

Soft Voting: In soft ensemble voting, the class names are anticipated depending on the anticipated probabilities P_{ij} of each instance 'i' classifier. This methodology is usually prescribed if the classifiers are very much aligned.

$$Y = \arg\max_i \sum_{j=1}^{m} W_j P_{ij} \qquad [j = 1, 2, 3, \ldots, m;\ i = 1, 2, 3, \ldots, n]$$

where W_j is the load that can be distributed with the jth classifier.

3. **Experimental Data.** As previously mentioned, a dataset containing about 6000 messages is used for experiments in which about 1000 messages are spam and about 5000 messages are ham. The representation of the pie chart of this dataset is given in Figure 2. The dataset used in this work is verified and analyzed using previously listed eleven different machine learning base classification techniques. These methods are again used for the cross-validation of the corresponding results.

After preprocessing the dataset like cleaning the missing values, natural language processing and vector transforming the data, the training data is split into 10 folds. Then the cross-validation scores of these eleven classifiers are measured and are given in Table 1. For better comparison, the listed values are shown as bar graph in Figure 3.

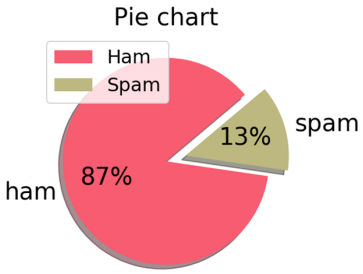

FIGURE 2. The pie chart representation of dataset

TABLE 1. Cross-validation scores of several ML classifiers

Classification type	**Cross-validation score (%)**
K-Neighbors	91.83
Ada Boosting	97.69
Decision Tree	97.51
Random Forest	97.59
Extra Tree	97.77
SVC	86.34
Gradient Boosting	97.11
Logistic Regression	98.24
Multi-Layer Perception (MLP)	98.42
Multinomial Naïve Bayes	98.07
X-Gradient Boosting	97.25

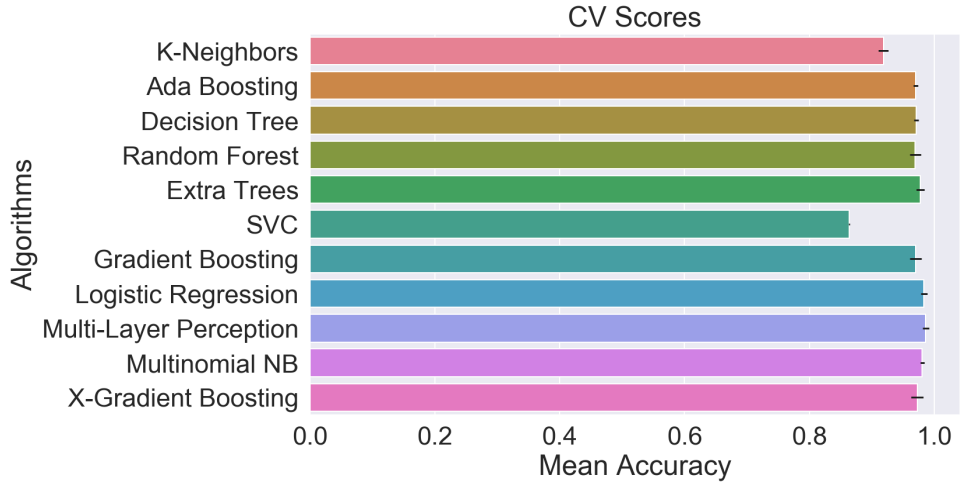

FIGURE 3. Bar chart representation of Cross-Validation (CV) scores of studied eleven ML classifiers as listed in Table 1

TABLE 2. Different performance parameters for selected three classifiers: MLP, Logistic Regression and Multinomial Naïve Bayes after hyper-tuning

Classification type	Best score	Accuracy	Precision	Recall	ROC score
MLP	98.68	98.75	100	61.43	94.36
Logistic Regression	98.52	98.56	97.14	67.62	93.54
Multinomial NB	98.38	97.84	80.04	91.90	95.97

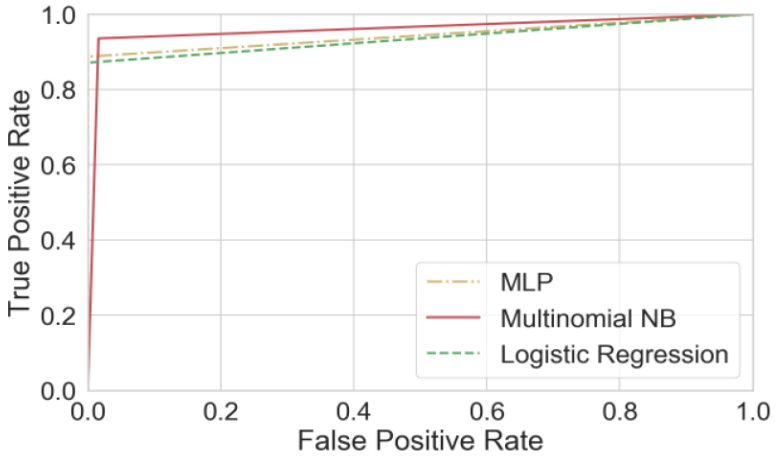

FIGURE 4. ROC curve of top three classifiers

Based on the cross-validation scores, the best three ML classification algorithms are found as: i) MLP (98.42%), ii) Logistic Regression (98.24%) and iii) Multinomial Naïve Bayes (98.07%). These three classifiers are hyper-tuned to get best results from them and then used in the next step of ensemble technique. Table 2 lists the optimum performance of selected top three classifiers and Figure 4 shows the ROC curves of them.

For MLP classification, the parameters were tuned on alpha, hidden layer size and maximum iterations. Thus, the best results were achieved for alpha = 0.01, hidden layer size = 14, maximum iteration = 1000, random state = 0 and solver = 'limited-memory Broyden Fletcher Goldfarb Shanno (lbfgs)'. The corresponding performance parameters are: best score = 98.68, accuracy = 98.75, precision = 100, recall = 61.43 and ROC score = 94.36.

For Logistic Regression classification, the parameters were tuned on tolerance, maximum iteration, concordance statistic (C), intercept scaling and solver. The optimum values of these parameters are found as: C = 100, intercept scaling = 4, maximum iteration = 100, solver = 'liblinear' and tolerance = 0.0002. The corresponding performance parameters are: best score = 98.52, accuracy = 98.56, precision = 97.14, recall = 67.62 and ROC score = 93.54.

For Multinomial Naïve Bayes classification, the parameter was tuned on alpha and the best result was obtained for alpha = 0.01. The corresponding performance parameters are (Table 2): best score = 98.38, accuracy = 97.84, precision = 80.04, recall = 91.90 and ROC score = 95.97.

The obtained best three ML classification algorithms are then used together in voting classifier to get maximum test score. Here, top three classifiers have been chosen because more than three classifiers will increase the complexity without significant improvement of results and less than three would compromise with the performance. Thus, the ultimate test score of Ensemble Voting Classifier is achieved as 98.93. As given in Table 3, the other parameters of this classifier are: Precision = 99, Recall = 100, F1 = 99 for Ham mails and Precision = 100, Recall = 90, F1 = 95 for Spam mails. The ROC score is 95.16. Figure 5 presents the ROC curve of Ensemble Voting Classifier.

TABLE 3. Report for Ensemble Voting Classifier

Type	Test score	Precision	Recall	F1	ROC score
Ham	98.93	99	100	99	95.16
Spam		100	90	95	
Average		99	99	99	

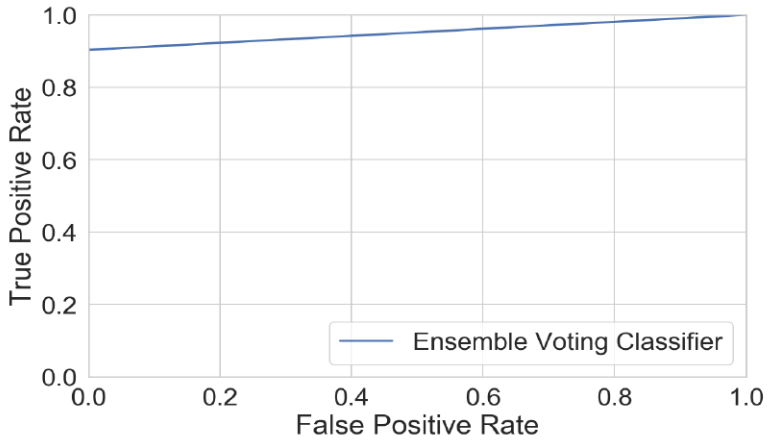

FIGURE 5. ROC curve of Ensemble Voting Classifier

Comparing the results in Tables 2 and 3, it is evident that Ensemble Voting Classifier (Table 3) provides improved test score than the other individual ML classification algorithms (Table 2). This Ensemble Voting Technique would be more helpful for large number of datasets.

4. **Conclusions.** In this article, a novel multi-classifier based Ensemble Voting Classifier technique is proposed for detecting both spam and non-spam messages. Several well-known and mostly used machine-learning classification algorithms have been utilized to an open source given dataset of messages for arranging them into spam and ham. The outcomes demonstrated that the proposed method would be better to use in terms of

accuracy, precision, recall, ROC score, and F1. Besides, the outcomes demonstrated that Ensemble Voting Classifier indicated better test score of about 99% when contrasted with the outcomes acquired by the other classifiers which are below 98%.

In any case, spammers keep thinking of the strategies and methods that ask for an evolvable and versatile spam location framework. For future works, different answers for the imbalanced grouping assignments can be contemplated (for example, cost-delicate learning and additionally preprocessing). Furthermore, the effect of these arrangements on the pertinence of info highlights can be explored.

REFERENCES

[1] A. H. Mohammad and R. A. Zitar, Application of genetic optimized artificial immune system and neural networks in spam detection, *Applied Soft Computing*, vol.11, no.4, pp.3827-3845, 2011.

[2] K. C. Ying, S. W. Lin, Z. J. Lee and Y. T. Lin, An ensemble approach applied to classify spam e-mails, *Expert Systems with Applications*, vol.37, no.3, pp.2197-2201, 2010.

[3] B. Enrico and A. Bryl, A survey of learning-based techniques of email spam filtering, *Artificial Intelligence Review*, vol.29, no.1, pp.63-92, 2008.

[4] D. Harris, D. Wu and V. N. Vapnik, Support vector machines for spam categorization, *IEEE Trans. Neural Networks*, vol.10, no.5, pp.1048-1054, 1999.

[5] O. Amayri and N. Bouguila, A study of spam filtering using support vector machines, *Artificial Intelligence Review*, vol.34, no.1, pp.73-108, 2010.

[6] V. Metsis, I. Androutsopoulos and G. Paliouras, Spam filtering with Naive Bayes – Which Naive Bayes?, *Conference on Email and Anti-Spam (CEAS)*, vol.17, pp.28-69, 2006.

[7] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, A Bayesian approach to filtering junk e-mail, *Learning for Text Categorization: Papers from the 1998 Workshop*, vol.62, pp.98-105, 1998.

[8] D. C. Trudgian, Spam classification using nearest neighbor techniques, *International Conference on Intelligent Data Engineering and Automated Learning*, Heidelberg, Berlin, pp.578-585, 2004.

[9] D. Puniškis, R. Laurutis and R. Dirmeikis, An artificial neural net for spam e-mail recognition, *Elektronika ir Elektrotechnika*, vol.69, no.5, pp.73-76, 2006.

[10] Z. Chuan, X. Lu, M. Hou and X. Zhou, A LVQ-based neural network anti-spam email approach, *ACM SIGOPS Operating Systems Review*, vol.39, no.1, pp.34-39, 2005.

[11] S. Rajput and A. Arora, Designing spam model-classification analysis using decision trees, *International Journal of Computer Applications*, vol.75, no.10, pp.6-12, 2013.

[12] 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.

[13] F. Toolan and J. Carthy, Feature selection for spam and phishing detection, *eCrime Researchers Summit*, pp.1-12, 2010.

[14] V. K. Singh and S. Bhardwaj, Spam mail detection using classification techniques and global training set, *Proc. of the 2nd Int. Conf. Intelligent Computing and Information and Communication*, Singapore, pp.623-632, 2018.

[15] H. Faris, A. M. Al-Zoubi, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah and H. Fujita, An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks, *Information Fusion*, vol.48, pp.67-83, 2019.

[16] S. Raschka, *Python Machine Learning*, Packt Publishing Ltd, UK, 2015.