Research Article


Luo GuangJun1, Shah Nazir2, Habib Ullah Khan3, and Amin Ul Haq4

1Education Science Department, Xianyang Normal University, Xianyang, Shaanxi, China
2Department of Computer Science, University of Swabi, Swabi, Pakistan
3Department of Accounting & Information Systems, College of Business & Economics, Qatar University, Doha, Qatar
4School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 61173, China

Correspondence should be addressed to Luo GuangJun; 1653446869@qq.com

Received 14 May 2020; Revised 4 June 2020; Accepted 6 June 2020; Published 9 July 2020

1.Introduction

Mobile message is a way of communication among the people, and billions of mobile device users exchange numerous messages. However, such type of communication is insecure due to lack of proper message filtering mechanisms. One cause of such insecurity is spam, and it makes the mobile message communication insecure. Spam is considered to be one of the serious problems in e-mail and instant message services. Spam is a junk mail or message. Spam e-mails and messages are unwanted for receivers which are sent to the users without their prior permission. It contains different forms such as adult content, selling item or services, and so on [1]. The spam increased in these days due more mobile devices deployed in environment for e-mail and message communication. Currently, 85% of mails and messages received by mobile users are spam [2]. The cost of mails and messages are very low for senders but high for receipts of these messages. The cost paid some time by service providers and the cost of spam can be measured in the loss of human time and loss of important messages or mails [3]. Due to these spam mails and messages, the values able e-mails and messages are affected because each user have limited Internet services, short time, and memory [4].

To handle these problems caused by the spam, researchers proposed different techniques to detect the spam e-mails and messages and secure the communication. Details of some of the techniques are presented in this article. Sharaff [5] proposed a method based on machine learning classifiers to classify ham and spam. In the proposed methods, they used four classifiers including Logistic regression (LR), K-nearest neighbor (K-NN), and decision tree (DT) are used for classification of ham and spam messages in mobile device communication. The SMS spam collection data set is used for testing the method. The dataset is split into two categories for training and testing the research. The results of the experiments demonstrated that the classification performance of LR is high as compared with K-NN and DT, and the LR achieved a high accuracy of 99%. Additionally, the proposed method performance is good as compared with the existing state-of-the-art methods.
designed ensemble methods based on techniques such as bagging, boosting, and stacking for classification of spam and ham. The data set used in the study was collected from Facebook. The experimental results demonstrated that the bagging ensemble learning approach, using J48 (decision tree) base classifier, performs well than its individual model, and the method achieved high performance in terms of detection accuracy. In [8], a method is proposed for ham and spam detection and principle components analysis and support vector machine were used in the designing of the system. Additionally, the performance evaluation and cross validation methods were used in the system. The proposed technique achieved high performance, and the method effectively detected the spam. Kumar [9] used various classifiers for ham and spam detection. They used different feature selection algorithms for selection of suitable features. The experimental results show that the classifier random tree with fisher algorithm achieved high results. The proposed method achieved 99% accuracy. In [10], the spam detection method was proposed using machine learning classifiers and 92% accuracy was achieved. Yang at al. [11] proposed spam detection approach based on multimodal fusion (SDAMF). They used the deep neural networks model for detection of spam and achieved 98.48% accuracy. In [12], a spam detection method was proposed based on the artificial immune system (ISAIS) and 98.05% accuracy was achieved. In [13], the Phishing e-mail detection system framework was proposed based on supervised and unsupervised methods. Ruano-Ordás et al. [14] proposed the spam detection method. They used evolutionary computation for discovering spam patterns from e-mail samples.

In this research study, we proposed a spam detection method using machine learning algorithms such as LR, k-nearest neighbor, and decision tree for classification of ham and spam messages. The SMS spam collection dataset was considered for testing of the current research. The dataset was divided into two categories: 30% for testing and 70% for training purpose for the predictive models. The evaluation metrics for performance such as specificity, accuracy, and sensitivity were considered evaluating the proposed study. The results obtained from experiments confirmed that the proposed research achieved high accuracy.

The remaining paper is organized as follows: Section 2 is about the related work to the methodology. In Section 3, experimental work is analyzed and presented in detail. The paper concludes in Section 4.

## 2. Methods and Materials

This section shows the research methods and materials of the paper.

### 2.1. Dataset

The dataset considered in the current research is available on kaggle, a machine learning repository [15]. The dataset “SMS spam collection dataset” contains 5572 instances and two attributes v1 and v2. The v2 is the input messages which are either spam or nonspam. The predicted label v1 has two classes: 0 = nonspam and 1 spam. In the data, 4900 are nonspam samples and 672 are spam samples. The dataset is given in Table 1.

### 2.2. Classification Algorithms

The following machine learning algorithms were considered for classifications of ham and spam.

#### 2.2.1. Logistic Regression

LR is a classifier [16, 17]. The problem in binary classification is computing the value of predictive y while \( y \in [0, 1] \); 0 and 1 are for class negative and positive. The LR predicts the variable value of multi-classification such as \( y \in [0, 1, 2, 3] \).

#### 2.2.2. Decision Tree

A DT is a supervised machine learning algorithm [18, 19]. Its shape is like a tree in which each node is a decision node or leaf. This technique of DT is easily understandable and simple for making the decisions. A DT contains external and internal nodes interlinked with each other. Decision can be made based on the internal nodes and the child node to access the preceding node. There is no child of the leaf node and is linked with a label.

#### 2.2.3. K-Nearest Neighbor

K-NN is a classification supervised learning algorithm [18]. It predicts the label of class as a fresh input and utilizes the same to its inputs in the training set. The performance of K-NN is not enough good. Let \( (x, y) \) be the the training observation and the learning function \( h: X \rightarrow Y \), so that an observation \( x, h (x) \) can establish \( y \) value.

#### 2.2.4. Division of Dataset

The set data were split into 30% and 70% for validation and training of the predictive model.

#### 2.2.5. Measure for Evaluation of Performance

To validate the classifier performance, we used metrics such as specificity, accuracy, sensitivity, and execution time which are expressed in equations (1), (2), and (3) which are computed from confusion matrix as given in Table 2.

The formulation of measures is as follows:

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100%, 
\]

\[
\text{sensitivity (Sn)} = \frac{TP}{TP + FN} \times 100%, 
\]

\[
\text{specificity (Sp)} = \frac{TN}{TN + FP} \times 100%. 
\]

### 3. Experiments and Result Analysis

Diverse approaches are used for spam detection. Abayomi-Alli et al. [22] presented a comprehensive review of the soft techniques in spam classifications. Acceptability users of

We performed experiments to classify the ham and spam using the SMS spam collection dataset. Classifiers LR, decision tree, and k-nearest neighbor were used for the classification in this study. The dataset is divided as follows: 30% for validation and 70% for training. The results obtained from experiments are shown in tables and presented in figures graphically. The python on an Intel (R) Core™ i5 -2400CPU and Windows 10 were used for the experiments and setup to obtain the computation results of the experimental work.

3.1. Visualization of SVM Spam Collection Dataset. In the data, 4900 are ham samples and 672 are spam samples which are shown in Figure 1.

Figure 2 shows the ratio of spam and ham messages.

Table 1: Description of SMS Spam collection dataset.

<table>
<thead>
<tr>
<th>Sample no number</th>
<th>V1</th>
<th>V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ham</td>
<td>Go until jurong point, crazy.. Available only in bugis n great world la e buffet...</td>
</tr>
<tr>
<td>2</td>
<td>Ham</td>
<td>Cine there got amore wat...</td>
</tr>
<tr>
<td>3</td>
<td>Spam</td>
<td>Free entry in 2 a wkly comp to win FA cup final tkt 21st May 2005. Text FA to 87121 to receive entry question (std txt rate) T&amp;C’s apply 08452810075over18’s.</td>
</tr>
<tr>
<td>5572</td>
<td>Ham</td>
<td>Rofl. Its true to its name</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix [20, 21].

<table>
<thead>
<tr>
<th>Actual spam (1)</th>
<th>Predicted spam (1)</th>
<th>Predicted ham (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Classification of spam and ham messages.

Figure 2: Ratio of ham and spam messages.
3.2. Classification Results of Classifiers. To perform the classification of the ham and spam messages, in this paper, we used the classification algorithms such as LR, decision tree, and K-nearest neighbor with essential basic hyperparameters. The dataset was divided into two parts for training and testing. The classifiers were trained with 70% of the samples and validated with 30% samples of the data set. All the experiential results are reported in Table 3. According to Table 3, the LR at hyperparameter $C = 1$ achieved 99% accuracy, 100% specificity, sensitivity 86%, and MCC, 93% and the processing time is 0.494 seconds. The classifier decision tree obtained 98% accuracy, 94% specificity, sensitivity 86%, and MCC 95%, and the processing time is 46.032 seconds. Similarly, the k-nearest neighbor classifier achieved 95% accuracy, 100% specificity, sensitivity 60%, and MCC 80%, and the processing time is 0.630 seconds. The experimental results (according to Table 3), the classification performance of LR is high as compared with the decision and k-nearest neighbor in terms of accuracy. The classification accuracy of classifiers is shown in Figure 3. Similarly, the computation time of LR is low as compared with k-NN and DT. Figure 4 shows the processing time graphically for better understanding. From these experiential results analysis, we concluded that the LR effectively classifies the ham and spam because the achieved accuracy is high. The 100% specificity of the LR model correctly detected the ham messages. Similarly, 86% sensitivity shows that LR spam message capability is good. Thus, the experimental results suggest that LR is a the best classifier for the classifications of ham and spam successfully.

Figure 3 shows the performance of classifications for classifiers including LR, K-NN, and DT.

The classifier processing time for K-NN, LR, and DT is shown in Figure 4.

### Table 3: Classification performance of classifiers.

<table>
<thead>
<tr>
<th>Predictive model</th>
<th>Evaluation performance measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>Logistic regression ($C = 1$)</td>
<td>99</td>
</tr>
<tr>
<td>K-nearest neighbor ($K-NN, K = 1$)</td>
<td>95</td>
</tr>
<tr>
<td>DT</td>
<td>98</td>
</tr>
</tbody>
</table>

![Figure 3: Performance of classifications for classifiers.](image)

![Figure 4: Processing time of classifiers.](image)

### Table 4: Comparison of the current study with existing approaches in terms of accuracy.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>ISAIS</td>
<td>98.48</td>
</tr>
<tr>
<td>Our study 2019</td>
<td>LR</td>
<td>99</td>
</tr>
</tbody>
</table>

3.3. Comparison of Performance with Existing Methods. The comparison performance of classifications of the current approach is done with the existing approaches in term of accuracy. The current approach achieved an accuracy of 99% which is high as compared with the available approaches. Table 4 shows the accuracy obtained from the current approach along with other approaches available.

The performance comparison is graphically shown in Figure 5.
4. Conclusion

Detection of spam is important for securing message and e-mail communication. The accurate detection of spam is a big issue, and many detection methods have been proposed by various researchers. However, these methods have a lack of capability to detect the spam accurately and efficiently. To solve this issue, we have proposed a method for spam detection using machine learning predictive models. The method is applied for the purpose of detection of spam. The experimental results obtained show that the proposed method has a high capability to detect spam. The proposed method achieved 99% accuracy which is high as compared with the other existing methods. Thus, the results suggest that the proposed method is more reliable for accurate and on-time detection of spam, and it will secure the communication systems of messages and e-mails.

Data Availability

No data were used to support the study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the Key scientific research of the Ministry of Education, China (Grant no. DCA190332).

References


