Email Text Analysis for Fraud Detection through Machine Learning Techniques

Rahaf Al-Haddad [1], Fatimah Sahwan [1], Amanh Aboalmakarem [1], Ghazanfar Latif [1, 2*], Yasmeen Mansour Alufaisan [1]

[1] Department of Computer Science, Prince Mohammad bin Fahd University, Al Khobar, Saudi Arabia.
[2] Cybersecurity Center, Prince Mohammad bin Fahd University, Al Khobar, Saudi Arabia.
{201501834, 201402453,201400861}@pmu.edu.sa, glatif@pmu.edu.sa, yalufaisan@pmu.edu.sa

Keywords: Email Text Analysis; Email Fraud Detection; Spam Detection; Machine Learning; Support Vector Machines

Abstract

Technology is improving and developing every day, new techniques and tools are releasing daily. Due to the development of technology, the number of new software and the websites are growing every day. As the numbers of software are increasing the number of users using them is increasing as well. Hackers and malicious people will take this chance to do fraud, hack or trick especially the naïve users. The email has been the best way for communication for several fields such as education, business, and entertainment. Most companies rely on email for communication with customers or with other companies, due to this malicious people focus more on sending fraud emails or phishing emails. Even though people have become more aware of spam and fraud emails, still hackers are improving their email format and content to look like a legitimate email. In this research paper, four machine learning techniques are applied to detect illegitimate fraud email from legitimate email. Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine classifiers are used for the experiments. These classification algorithms are applied on a new Fraud emails dataset consist of 11926 emails, where 6742 are Fraud (spam) emails and the rest are normal (ham) email for their classification. The results show SVM has the best performance where the achieved accuracy is more than 98%.

1 Introduction

Emails are an important method of communication, especially in specific areas such as business and educational institutes. It is used because of how cheap, easy and accessible it is. However, one of the huge disadvantages of emails are frauds that are represented in spam emails, phishing emails, and fake emails. The idea of this research is using data mining techniques because it provides help in finding hidden information within the data that will be able to provide us with useful information and statistics that allow us to develop new knowledge [1]. The objective of this paper is to test several machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest and Support Vector Machine to generate results and then compare them to define and choose the most effective algorithm for fraud detection. We would like to discover the most efficient that can be implemented in the future email filtering applications. So that those applications will have more intelligence while doing fraud detection. Furthermore, by applying machine learning techniques it will give us the advantage of having a machine learn and improve using from the data set experience instead of using old coded techniques [2]. It is the development of programs where they can use the data to improve themselves.

The rest of the paper is as fallows; section 2 discusses the recent literature review, section 3 summarizes the description of the used dataset. Section 4 discusses the used machine learning algorithms. Section 5 represents the experimental results and section 6 concludes the paper.

2 Recent Literature

There are different machine learning-based classification algorithms are available which are used in different domains like image recognition, disease classification, health diagnosis, and automated medication systems, optical character recognition, text analysis, and network security monitoring [3-5]. In this paper, machine learning techniques are chosen based on previous research papers for email analysis, spam detection. In [6], the authors talked proposed model for phishing email as a supervised classification problem to detect spam email from non-spam email. They used Decision Tree, KNN, Naive Bayes, Random Forest, SVM, and logistic regression. The results of Random Forest in Training dataset with TDM and SVD with sub-task1 are Accuracy 95.2%, Precision 0.914, Recall 0.597, F1-Score 0.722. Sub-task 2 results are Accuracy 99.3%, Precision 1.00, Recall 0.948, F1-Score 0.970.

In another article [7], the authors talked about how the increase in the number of email users led to increasing the number of spams in the email in recent years. “Spam mail has become an increasing menace as it increases the chances of virus threats, communication overload, wastage of time, irritation and disturbance, etc., to the users”. They used four classifiers one of them is J48. The results are Accuracy of 85.06%. The other results divided as 0 and 1, which means

spam and non-spam. Spam emails in Precision 0.831, Recall 0.945, F1-Score 0.885. For non-spam email Precision 0.893, Recall 0.705, F1-Score 0.788.

In paper [8] authors' goal was to detect fraud transactions from non-fraud transactions using machine learning techniques. They used binary classification such as Logistic Regression, Linear SVM and SVM Transactions using Scikit-learn library. The results show that SVM with RBF kernel has better performance than the other algorithms in both training set and CV set. However, Logistic Regression detects more fraud transactions than SVM with RBF kernel. In CV set Transfer the results of SVM with RBF kernel are Recall 0.9934, Precision 0.5871, F1-measure 0.7381 and AUPR 0. 0.9855. The SVM with RBF in Train set Transfer have Recall 0.9958, Precision 0.6035, F1-measure 0.7515 and AUPR 0.9895.

In this research paper [9], the authors used four different classes of spam filtering methods (Content-Based Filtering method, Case Base Filtering Method, Previous Likeness Based Filtering method and Adaptive Filtering method) and applied different machine learning classifiers for spam filtering. They recommend using a content-based filtering technique for detecting spam email. The authors in [10] proposed the idea of using surveys to gather data and perform data mining techniques on them to detect frauds. The paper summarizes the past 10 years technique of automated fraud detection with data mining to achieve high-cost saving. For the methods, they have used a supervised approach method for the labeled data such as neural networks, decision trees and support vector machine (SVM). However, another approach is to use a hybrid approach with labeled data where you combine two methods in one to get the best results such as combining decision tree with Bayesian networks.

In research [11], the authors perform an analysis of spam email detection performance assessment using machine learning. They aim to get the best method for spam email detection. They are comparing the efficiency of three methods and they are Logistic Regression, Decision Tree and Random Forest. They stated that Logistic Regression is a method that creates a prediction module like linear regression. In article [12]. A method called rotating forest was applied to the data to classify if the emails are spam or non-spam in addition to using the whale optimization algorithm-based email spam feature selection (WOA). They introduced the idea of the whale optimization algorithm (WOA) as a metaheuristic optimization algorithm that cut down computation time by avoiding certain criteria.

In [13], the authors presented an effective model for spam detection. They are using the random forest for classification and active learning. Before applying the random forest, they had gone through a process where they first applied term frequency and inverse document frequency (TF, IDF), they calculated find the differences between the measure of their performance which is the area under the curve (AUC) where Random Forest gave 95.2%, Naive Bayes 66.7%, SVM 66.7% and kNN with 66.7%. In research [14], authors used Bayesian text classification that has been used from the day it was created and even with the new methods that occurred recently it is still used because of its simplicity. The article also explains the use of Naive Bayes usage and advantages. Authors in this research paper [15] used four different classification algorithms to classify emails as fraud (Spam) or Ham (Normal) email. They used Decision Tree, Lazy algorithm, Probabilistic and Vector Machines. They used WEKA to analyze the performance of these algorithms. The results show J48 has the best performance among the other algorithms. Accuracy 93.3%, Precision 0.93, Recall 0.933, FP Rate 0.284 and F-measure 0.93.

3. Data Description

In this research, two different datasets are used for the experimental results. The first dataset contains 11926 labeled emails in total from which 5183 fraud (spam) emails and the remaining 6742 are normal (ham) emails [16]. Another dataset was used for fast testing since it contained less values than the main one. Therefore, we were able to run the algorithms fast in the computers that do not have the huge processing power and to try and compare how will the algorithm performance change depending on the size of the dataset. The second dataset's main topic is spam and ham messaged but instead of emails it was composed of SMS messages. The dataset was obtained also Kaggle website that uploaded it from the Machine Learning Repository in 2012 [17], and it contained 5573 labeled SMS messages. Where 747 (13%) messages are spam and the remaining 4825 (87%) are ham (not spam).

4. Algorithms for Email Fraud Detection

There are different machine learning-based classification algorithms are available which are used in different domains. Based on the recent literature on the similar topic, four different machine learning techniques are studied and applied to both datasets. Decision Tree is a supervised machine learning algorithm, it utilized for classification and regression. Decision tree is one of the easiest classification algorithms to interpret and understand. However, Decision Tree structure is similar to a flowchart like trees where it consists of three main parameters are Nodes, Edges (Branch) and Leaf nodes. Selecting attributes will be in the top node (root), decision rule in the Edge, and the outcome in leaf node [18]. To select attributes, it needs to split the dataset into smaller one then pruning it to decrease the size of the tree. Selecting attribute done by attribute selection measures such as Information Gain used in the ID3 algorithm, Gini Index is used in the CART algorithm, and Gain Ratio is used in the C4.5 algorithm.

The Naïve Bayes Classifier is a simple classifier of the machine learning classifiers. It is one of the simplest Bayesian network models which work based on the assumptions among the features [19]. It works on the principle of conditional probability, as given with the assistance of the Bayes theorem [20]. Bayes' theorem mathematically represented as the following Equation 1.

$$P(c \mid x) = (P(x \mid c))/P(x) \qquad (1)$$

Random Forest (RF) algorithm is an ensemble learning technique that consists of several classifiers that are uncorrelated and work individually to make a prediction [21]. Those classifiers are a group of random decision trees and each is built on a bootstrap sample of the training data by the use of randomly selected subset of a variable. The high number of trees in a forest will lead to having high accuracy.

Support Vector Machine (SVM) is one of the most well-known models in machine learning, and it is a very potent model. SVM is a type of supervised learning models used for classification and regression difficulty. SVM is a very good model to classify complex datasets, especially small and medium datasets [22]. Kernel Trick is a method in machine learning to keep away from some intense computation in some algorithms, which makes a few computations go from infeasible to feasible. SVM is the best-known member of kernel methods in machine learning [23].

5 Results and Discussion

In this study, four machine learning classifiers are applied to the spam email datasets to detect whether an email or a message is a Fraud (spam) or normal (ham). Table 1 presents the result of four classifiers applying on Dataset 1, while Table 2 shows the result of four classifiers applying on Dataset 2. Both Table 1 and Table 2 represent in terms of accuracy, precision, recall, and F1-score. Our data was split into 20% for testing and the rest of the 80% was used for the training.

Table 1: Comparison the results between four classifiers applying on Dataset 1

Classifier	Accuracy %	Precision	Recall	F1-score	Class
Naive Bayes	97.32%	0.98 0.96	0.97 0.97	0.97 0.97	Ham Spam
Decision Tree	96.19%	0.95 0.98	0.98 0.94	0.97 0.96	Ham Spam
Random Forest	97.64%	0.96 1.00	1.00 0.94	0.98 0.97	Ham Spam
Support Vector Machine	98.22%	0.97 1.00	1.00 0.96	0.98 0.98	Ham Spam

As the baseline for comparison, we indicate that the Support Vector Machine (SVM) classifier has the highest accuracy as shown in Table 1 and Table 2, while Decision Tree classifier has the lowest accuracy, notice that when applied to our datasets it took less time to process than Support Vector Machine (SVM). The Random Forest (RF) and Naive Bayes (NB) come in the middle which, amazed us knowing that people usually get low accuracies from the Naive Bayes classifier. The accuracy of SVM is 98.22%, the accuracy of NB and RF is 97.64%, and the accuracy of Decision Tree is 96.19% based on the Database 1 as shown in Table 1. For the Dataset 2, the accuracy of SVM is 98.47%, the accuracy of RF is 97.13%, the accuracy of NB is 96.63%, and the accuracy of Decision Tree is 95.72% as shown in Table 2. According to the accuracies given, the accuracies do not have that huge differences between them. However, SVM is considered the most effective classifier with our data compared with the other classifiers where the SVM works well with unstructured and semi-structured data like text, Images, and trees, and is the best-known member of Kernel Trick in machine learning.

Table 2. Comparison the results between four classifiers applying on Dataset 2

Classifier	Accuracy %	Precision	Recall	F1-score	Class
Naive Bayes	96.63%	0.96 1.00	1.00 0.71	0.98 0.83	Ham Spam
Decision Tree	95.72%	0.96 0.87	0.98 0.75	0.97 0.81	Ham Spam
Random Forest	97.13%	0.97 0.98	1.00 0.79	0.98 0.87	Ham Spam
Support Vector Machine	98.47%	0.98 0.98	1.00 0.89	0.99 0.93	Ham Spam

The experimental results show that our proposed method Support vector machine-based method gives better results as compared to the recent studies as shown in Table 3. The results show that the SVM based method got 98.22% and 98.47% accuracies respectively for Database 1 and Database 2 as compared to the other studies where got the accuracies in between 85.06% to 96.2%.

Table 3: Comparison of proposed method results with other recent studies

Classifier	Accuracy %	Precision	Recall	F1-score
Proposed Dataset 1	98.22%	0.985	0.97	0.97
Proposed Dataset 2	98.47%	0.98	0.945	0.97
Decision Trees [15]	93.3%	0.93	0.933	0.930
J48 Trees [7]	85.06%	0.831	0.945	0.885
RF + SVD [6]	95.2%	0.914,	0.597	0.722

6 Conclusions

In this research paper, four machine learning algorithms are applied to detect fraud emails from normal emails. We used SVM, Decision Tree, Random Forest, and Naïve Bayes classifiers. The first dataset used in this article contains a total number of 11907 emails where 5138 are spam email and the reset are legitimate emails. The results show SVM has the best results and performance, the accuracy was 98.22%, precision 0.985, recall 0.97 and F1-score 0.97. Meanwhile, Decision Tree got the lowest results among the other classifier with an accuracy of 96.19%. For future research, we will try another machine learning algorithm to get better results, and we will use a better analysis tool that will evaluate their performance.

References

[1] Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. "A comprehensive survey of data mining-based fraud detection research." *arXiv preprint arXiv:1009.6119* (2010).

[2] Fogel, Alexander L., and Joseph C. Kvedar. "Benefits and risks of machine learning decision support systems." *Jama* 318, no. 23 (2017): 2356-2356.

[3] Latif, Ghazanfar, Jaafar Alghazo, Loay Alzubaidi, M. Muzzamal Naseer, and Yazan Alghazo. "Deep convolutional neural network for recognition of unified multi-language handwritten numerals." In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp. 90-95. IEEE, 2018.

[4] Alghazo, Jaafar M., Ghazanfar Latif, Loay Alzubaidi, and Ammar Elhassan. "Multi-Language Handwritten Digits Recognition based on Novel Structural Features." *Journal of Imaging Science and Technology* 63, no. 2 (2019): 20502-1.

[5] Latif, Ghazanfar, Achyut Shankar, Jaafar M. Alghazo, V. Kalyanasundaram, C. S. Boopathi, and M. Arfan Jaffar. "I-CARES: advancing health diagnosis and medication through IoT." *Wireless Networks* (2019): 1-15.

[6] Vazhayil, Anu, N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and A. D. R. Verma. "PED-ML: Phishing email detection using classical machine learning techniques." In *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA)*, pp. 1-8. Tempe, AZ, USA, 2018.

[7] AL-Rawashdeh, Ghada Hammad, and Rabiei Bin Mamat. "Comparison of four email classification algorithms using WEKA." *International Journal of Computer Science and Information Security (IJCSIS)* 17, no. 2 (2019).

[8] Oza, Aditya. "Fraud Detection using Machine Learning." *TRANSFER* 528812, no. 4097: 532909.

[9] Dada, Emmanuel Gbenga, Joseph Stephen Bassi, Haruna Chiroma, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. "Machine learning for email spam filtering: review, approaches and open research problems." *Heliyon* 5, no. 6 (2019): e01802.

[10] Santoso, Budi. "An Analysis of Spam Email Detection Performance Assessment Using Machine Learning." *Jurnal Online Informatika* 4, no. 1 (2019): 53-56.

[11] Shuaib, Maryam, Olawale Surajudeen Adebayo, Oluwafemi Osho, Ismaila Idris, John K. Alhassan, and Nadim Rana. "Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification." *SN Applied Sciences* 1, no. 5 (2019): 390.

[12] DeBarr, Dave, and Harry Wechsler. "Spam detection using clustering, random forests, and active learning." In *Sixth Conference on Email and Anti-Spam. Mountain View, California*, pp. 1-6. 2009.

[13] Olatunji, Sunday Olusanya. "Improved email spam detection model based on support vector machines." *Neural Computing and Applications* 31, no. 3 (2019): 691-699.

[14] Dada, Emmanuel Gbenga, Joseph Stephen Bassi, Haruna Chiroma, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. "Machine learning for email spam filtering: review, approaches and open research problems." *Heliyon* 5, no. 6 (2019): e01802.

[15] Sharaff, Aakanksha, Naresh Kumar Nagwani, and Abhishek Dhadse. "Comparative study of classification algorithms for spam email detection." In *Emerging research in computing, information, communication and applications*, pp. 237-244. Springer, New Delhi, 2016.

[16] Radev, D. "CLAIR collection of fraud email, ACL data and code repository." *ADCR2008T001* (2008).

[17] Almeida, Tiago A., José María G. Hidalgo, and Akebo Yamakami. "Contributions to the study of SMS spam filtering: new collection and results." In *Proceedings of the 11th ACM symposium on Document engineering*, pp. 259-262. 2011.

[18] Navlani, Avinash. "Decision Tree Classification in Python." *Data Camp* (2018).

[19] Xu, Shuo. "Bayesian Naïve Bayes classifiers to text classification." *Journal of Information Science* 44, no. 1 (2018): 48-59.

[20] Latif, Ghazanfar, M. Mohsin Butt, Adil H. Khan, Omair Butt, and DNF Awang Iskandar. "Multiclass brain Glioma tumor classification using block-based 3D Wavelet features of MR images." In *2017 4th International Conference on Electrical and Electronic Engineering (ICEEE)*, pp. 333-337. IEEE, 2017.

[21] Pranckevičius, Tomas, and Virginijus Marcinkevičius. "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification." *Baltic Journal of Modern Computing* 5, no. 2 (2017): 221.

[22] Haque, Md Sarwar M., Ghazanfar Latif, Md Rafiul Hasan, Md Arifuzzaman, Shakib S. Shafin, and Quazi A. Rahman. "Scalable Parallel SVM on Cloud Clusters for Large Datasets Classification." (2019): 13-5.

[23] Alghazo, Jaafar M., Ghazanfar Latif, Ammar Elhassan, Loay Alzubaidi, Ahmad Al-Hmouz, and Rami Al-Hmouz. "An Online Numeral Recognition System Using Improved Structural Features–A Unified Method for Handwritten Arabic and Persian Numerals." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 9, no. 2-10 (2017): 33-40.