Available online at www.asric.org

ASRIC Journal on Natural Sciences 1 (2021) 63-68

Using Machine Learning Techniques for predicting Email Spam

Hager Mohey Abohalfaya[1], Someya Mohsen[1]

[1]Department of Information System, Faculty of Computer & Information,
Al-minia University - Minia, Egypt

Corresponding Authors Email: Hagerabohalfaya95@gmail.com

Received 10 June 2021; revised 10 July 2021; accepted 22 August 2021

Abstract:
Email has become one of the most efficient and cost-effective methods of communication in recent years. However, as the number of email users grows, so does the number of spam emails. Email management has become a big and rising concern for both people and companies as a consequence of its sensitivity to abuse. Spam, or the unsolicited sending of unwanted email messages, is one example of misuse. Spam is defined as unsolicited bulk email, or email sent to a large number of people without their consent. Half of users receive 10 or more spam emails each day, while some users receive hundreds of unwanted emails per day. Online spiders are used by many spammers to discover email addresses on web pages. Because of spam emails can fill up the storage space of a file server quickly, they could cause a very severe problem for many websites with thousands of users for this in this study, we present a method for spam filtering using some machine learning techniques to predict whether an email is spam or no.

I. INTRODUCTION

Millions of individuals use email on a daily basis. Email is used by them for a number of purposes, including employment, research, and other activities. E-mail is a kind of electronic communication that allows two or more individuals who are connected to the Internet to communicate with each other. Due to the growing use of email and the incursions of online marketers, unwanted commercial email has become a problem on the internet. Unsolicited and undesired junk email delivered in bulk to an indiscriminate recipient list is known as spam email. Spam is typically sent for commercial objectives. Botnets, or networks of infected machines, may send it in large quantities. A spammer sends an email to millions of individuals with the expectation that just a small fraction of them will respond or interact with it. Email spam takes several forms, the most common of which is to advertise blatant frauds or shady business ventures. Emails are being utilized for more than simply communication; they are also used for work management and customer service. Email categorization was inspired by text classification in machine learning, and it is now accepted in a variety of forms, such as classifying emails into a spam folder, blocking spam email, and detecting the user's mood from the email body. Most recent email apps and services, such as Gmail and Hotmail, allow users to easily filter received emails based on the email subject and key tokens in the email body. This technique is suitable for individual work or home operators, as it eliminates the need to create token-based rules to sort emails into different folders. As the problem with which we are working is a classification problem, we not only need to have models that maximize the accuracy results of correct classified samples.

In e-mail filtering, two main techniques are used: knowledge engineering and machine learning. A set of rules must be established in the knowledge engineering method, according to which emails are classified as spam or ham. A collection of such rules should be developed either by the filter's user or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool).Because the rules must be continually updated and maintained, which is a waste of time and inconvenient for most users, this technique yields no promising outcomes. Machine learning is more efficient than knowledge engineering since it does not need the specification of any rules. Instead, a set of training samples is used, which consists of a collection of pre classified e-mail messages. The categorization criteria are subsequently learned from these e-mail messages using a particular algorithm. Machine learning has been extensively researched, and there are several algorithms that may be utilized in e-mail filtering, Support vector machines, Neural Networks, K-nearest neighbour, Rough sets, and the artificial immune system are among them.

The paper is organized as follows: Section 2 provides some related studies. Section 3 gives the dataset that used, followed by the experimental design and Methodology obtained in Section 4. Section 5 depicts results of used algorithms. Finally, conclusions are drawn in Section 6.

II. RELATED WORK

There have been several studies that utilize data mining and machine learning methods and techniques to classify spam e-mails. For example, in one study [1], four classifiers were evaluated to filter spams from a dataset of emails: Neural Network, SVM, Nave Bayesian, and J48. All of the emails were categorized as spam (1) or not spam (1). (0).

The study [2] utilized the Word Stemming or Word Hashing Technique to improve the performance of the content-based spam filter used in the SMTP server, which correctly classified ham and spam emails.

The paper [3] offers a novel spam detection approach based on text clustering and a vector space model. It automatically computes disjoint clusters for all spam/Not-SPAM messages using a spherical k-means algorithm and produces cluster centroid vectors for extracting the cluster description.

On the other hand, research [4] proposes and implements a new better model that combines the negative selection algorithm (NSA) with particle swarm optimization (PSO).

The usage of several learning algorithms for identifying spam messages from e-mail is investigated and identified in study [5]. In addition, a comparison of the algorithms has been given. Rapid Miner was developed on a common dataset after a thorough examination of several classifiers using various software packages such as WEKA.

The researchers in [17] compare the performance of Non Linear SVM based Classifiers with various kernel functions on the Enron Dataset in order to evaluate as many attributes as possible. Because of its sparse data format and acceptable Recall and Precision Values, SVM has proven to be a good classifier. SVM is also recognized as a fundamental example of "kernel techniques," one of the most significant topics in machine learning.

The Random Forests (RF) Algorithm is used to classify emails, and the authors claim that ensemble learning gives a more reliable mapping that may be produced by merging the output of several classifiers [18].

III.DATASET DESCRIPTION

The dataset for this study was compiled over the course of two months from multiple e-mail ids. Around 57 spam email attributes were detected and used in the sample. The attributes used ranged from address to address, type of spam received, and the organization from whom the spam was received. The Kaggle Machine Learning Repository contains datasets for machine learning techniques. Kaggle's spam dataset is made up of data gathered from 4601 email messages. In the Spam dataset, each instance has 58 properties. The frequency of a certain word or character in the email that corresponds to the instance is represented by the majority of the characteristics.

- Word freq w: 48 attributes describing the frequency of word w, the percentage of words in the email.
- Char freq c: 6 attributes describing the frequency of a character c, defined in the same way as word frequency.
- Char freq cap: 3 attributes describing the longest length, total numbers of capital letters and average length.
- Spam class: the target attribute denoting whether the email was considered spam or no spam. [21].

IV. METHODOLOGY

This section includes the following approaches:-

4.1- preprocessing step:-

This is the first stage that is executed whenever an incoming mail is received. This step contains Min-Max scaler. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

4.2 Feature selection:-

The feature selection stage comes after the pre-processing stage. Feature selection is a type of spatial coverage reduction that effectively exemplifies intriguing email message fragments as a compressed feature vector. When the message size is enormous and a condensed feature representation is required to make text or image matching quick, this technique comes in handy. The recognition of spam e-mails with minimum number of features is important in view of computational complexity and time. Feature selection involves processes like stemming, noise removal such as (check skewness whether high or moderate or low) and remove features with high correlation and stop word removal steps [22].

This step can be implemented using SelectKBest Method that is a technique where we choose those features in our data that contribute most to the target variable. SelectKBest then simply retains the first k features of X with the highest scores.

This method evaluates each of the classification algorithms on the training set and selects the best one for application on the test set.

V.RESULTS

To produce an predictive model based on previous approaches , we used the following classification algorithms representing in :- SVM (Support Vector Machine) , NB (Gaussian Naïve- base) , DT (Decision Tree classifier), KNN (K-nearest neighbors) with total accuracy of each one of them (91.2 ,91% ,90.7%, 88%).

Before applying these algorithms we need to do two steps:-

5.1 Result Evaluation

The data set was separated into two parts; one part is used as training data set to produce the prediction model, and the other part is used as test data set to test the accuracy of our model. The training data set contains feature values as well as classification of each record. Testing is done by 10-fold cross validation method [21].

5.2 Cross Validation

There are several means of estimating how well the classifier works after training. The easiest and most straightforward means is by splitting the dataset into two parts and using one part for training and the other for testing. This is called the holdout method. The disadvantage is that the evaluation depends heavily on which samples end up in which set. Another method that reduces the variance of the holdout method is k -fold cross-validation.

Table 1:
Accuracy of algorithms

. Algorithm	Training	Testing
SVM	91.4%	91.2%
NB	89%	91%
KNN	90.6%	88%
DT	91%	90.7%

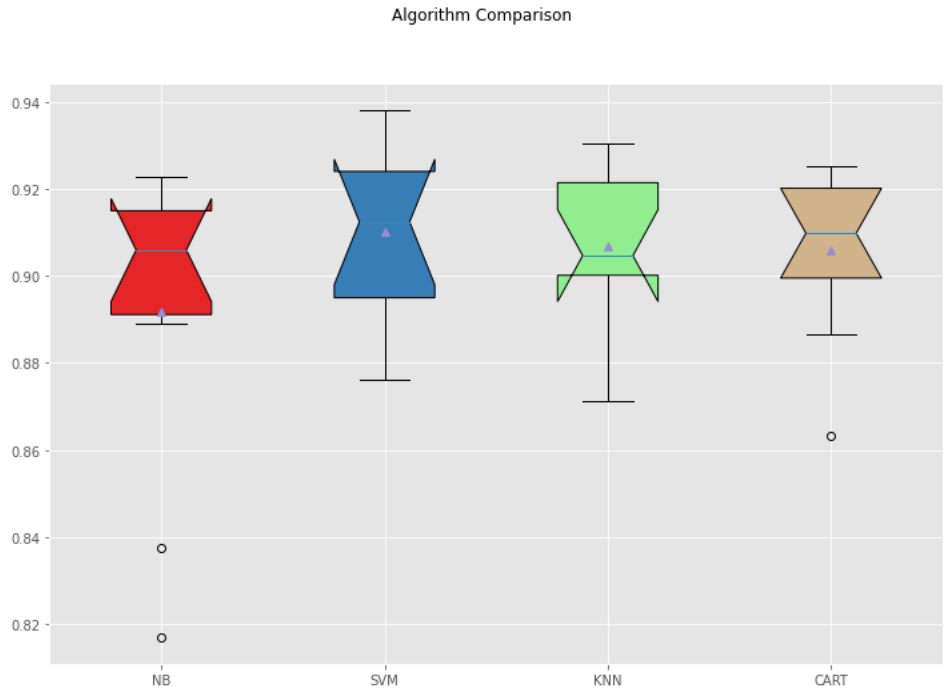
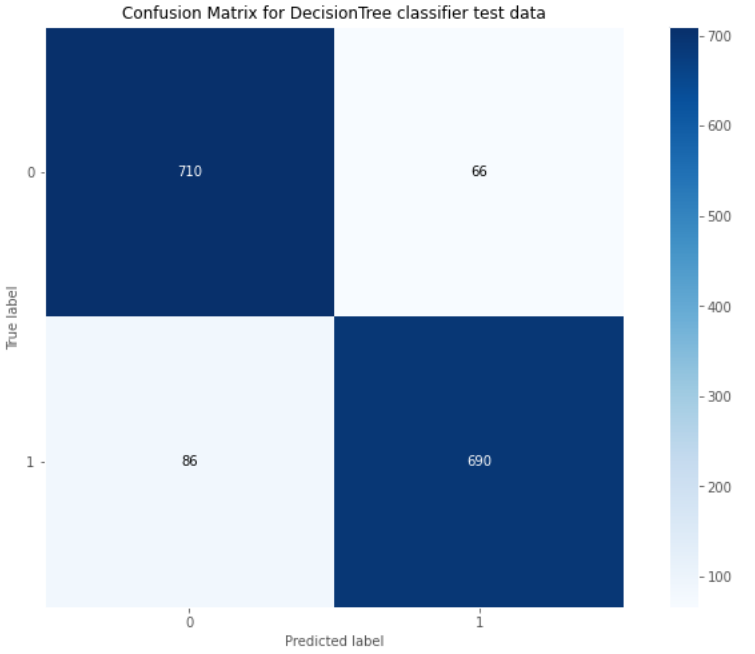
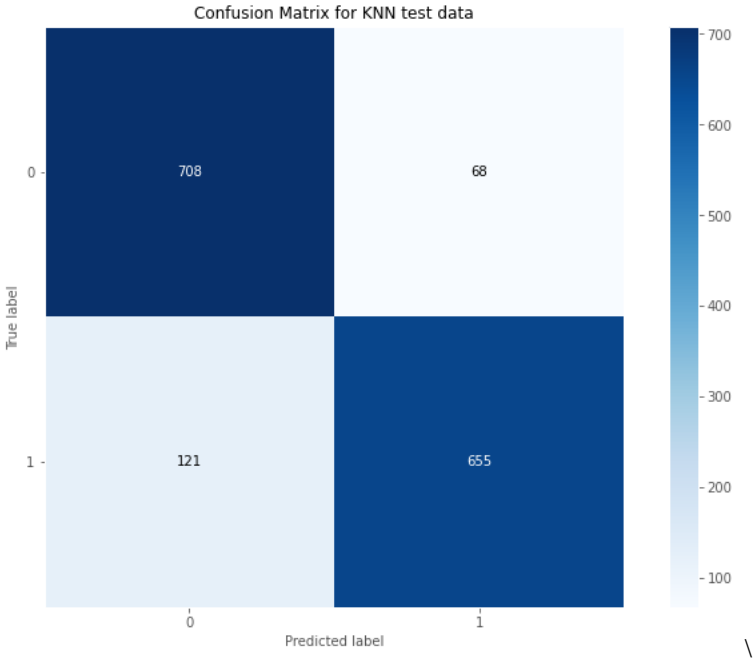


Figure 1: The classification algorithms comparison.

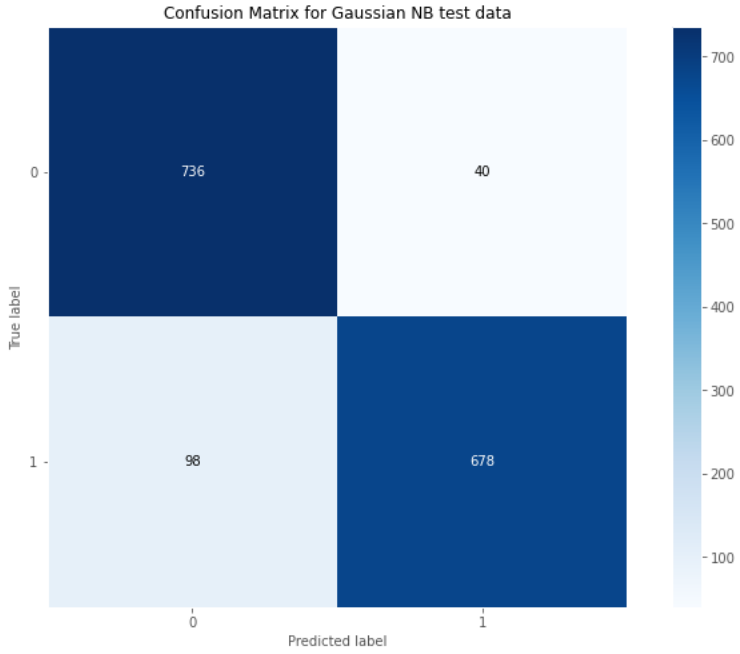
Table [1] depicts Accuracy of training and testing features for machine learning algorithms such that SVM is the best accuracy that achieved 91.4% for training, 91.2% for testing. NB achieved 89% for training and 91% for

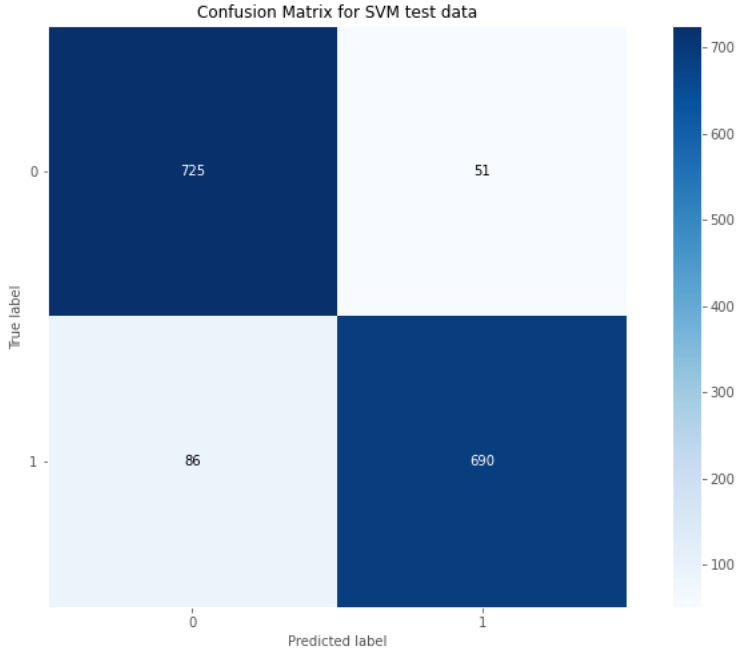
testing. KNN achieved 90.6% for training and 88% for testing. DT achieved 91% for training and 90.7% for testing. Refer to table 1 and figure 1, SVM is the best accuracy but KNN is the worst accuracy.
The blow figures show confusion matrices for four classification algorithms used





\





VI.CONCLUSION

In this paper, we looked at machine learning techniques and how they may be used to spam filtering. A look at how state-of-the-art algorithms have been used to classify communications as spam or not is presented. The attempts of many researchers to utilize machine learning classifiers to solve the problem of spam were highlighted [22]. Data mining tools. Organization, 2(08), pp.2760-2766.

REFERENCES

[1]. Youn, S. and McLeod, D., 2007. A comparative study for email classification. In Advances and innovations in systems, computing sciences and software engineering (pp. 387-391). Springer, Dordrecht.

[2]. Hamsapriya, T. and Renuka, M.D.K., 2010. Email classification for spam detection using word stemming. International journal of computer applications, 1(5), pp.45-47.

[3]. Sasaki, M. and Shinnou, H., 2005, November. Spam detection using text clustering. In null (pp. 316-319). IEEE. https://doi.org/10.1109/CW.2005.83.

[4]. Idris, I. and Selamat, A., 2014. Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22, pp.11-27.
[5]. Scholar, M., 2010. Supervised learning approach for spam classification analysis using
[6]. Abu Naser, S., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2), 221-228.
[7]. Elzamly, A., Abu Naser, S. S., Hussin, B., & Doheir, M. (2015). Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods. Int. J. Adv. Inf. Sci. Technol, 38(38), 108-115.
[8]. Abu Naser, S. S., & Alhabbash, M. I. (2016). Male Infertility Expert system Diagnoses and Treatment. American Journal of Innovative Research and Applied Sciences, 2(4).
[9]. Qwaider, S. R., & Abu Naser, S. S. (2017). Expert System for Diagnosing Ankle Diseases. International Journal of Engineering and Information Systems (IJEAIS), 1(4), 89-101.
[10]. Abu Naser, S. S., & Al-Hanjori, M. M. (2016). An expert system for men genital problems diagnosis and treatment. International Journal of Medicine Research, 1(2), 83-86.
[11]. Kumaresan, T. and Palanisamy, C., 2017. E-mail spam classification using S-cuckoo search and support vector machine. International Journal of Bio-Inspired Computation, 9(3), pp.142-156.
[12]. Sharma, A. and Suryawanshi, A., 2016. A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. International Journal of Computer Applications, 136(6), pp.28-35. https://doi.org/10.5120/ijca2016908471
[13]. Shah, N.F. and Kumar, P., 2018. A Comparative Analysis of Various Spam Classifications. In Progress in Intelligent Computing Techniques: Theory,Practice, and Applications(pp. 265-271). Springer, Singapore.
[14]. Bassiouni, M., Ali, M. and El-Dahshan, E.A., 2018. Ham and Spam E-Mails Classification Using Machine Learning Techniques. Journal of Applied Security Research, 13(3), pp.315-331.
[15]. Sah, U.K. and Parmar, N., 2017. An approach for Malicious Spam Detection In Email with comparison of different classifiers.
[16]. Spambase.documentation at the UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLReposito ry.html, May 01, 2018, 06:54:55 pm.
[17]. Mohamad, M. and Selamat, A., 2015, April. An evaluation on the efficiency of hybrid feature selection in spam email classification. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on (pp. 227-231). IEEE.
[18]. Hall, M.A., Practical Machine Learning Tools and Techniques. United State: Morgan Kauffman, 2011.
[19]. DeBarr, D. and Wechsler, H., 2012. Spam detection using random boost. Pattern Recognition Letters, 33(10), pp.1237-1244.
[20]. Zhang, Y., Wang, S., Phillips, P. and Ji, G., 2014. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based systems, 64, pp.22-31. https://doi.org/10.1016/j.knosys.2014.03.015.
[21]. Hamsapriya, T., D. Karthika Renuka, and M.Raja Chakkaravarthi. "Spam Classification based on Supervised Learning using Machine Learning Techniques." DIGITAL WORLD 2.04 (2011).
[22]. Dada, Emmanuel Gbenga, et al. "Machine learning for email spam filtering: review, approaches and open research problems." *Heliyon* 5.6 (2019): e01802.