Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Spam email detection using machine learning algorithm

Rakesh Nayak, Salim Amirali Jiwani *, B. Rajitha

Department of CSE, Vaagdevi Engineering College, Warangal, India

ARTICLE INFO

Article history:
Received 8 February 2021
Accepted 7 March 2021
Available online xxxx

Keywords:
Spam Mail Detection (SPM)
Hybrid bagged approach
Email classification

ABSTRACT

Nowadays the communication between the organizations or any individuals became easier by use of Electronic mail method. The internet users are increasing rapidly day by day and also spams are increasing with the emails. It was an easy task for the spammers to create an email account and making a fake profile. Therefore, detecting of these spam mails that were fraud is of most important. This paper aims to develop a proposed approach of data science for spam email detection (SMD) using machine learning algorithm. A hybrid bagging approach is used in this proposed method for the detection of spam emails which implements the two, Naïve Bayes and J48 (i.e. decision tree) machine learning algorithms. Each of these algorithms is applied as an input with a data set which is partitioned in to different sets using the data science. Emails classification can be accomplished depends on the patterns of repetitive keywords and several additional parameters like Cc (carbon copy) or Bcc (Blind carbon copy), domain, header etc. that are present in their structure. Whenever the machine learning algorithm is applied for the email classification every one of those parameters are considered as a features. This proposed approach of data science for spam mail detection using machine learning algorithm achieved a 88.12% of overall accuracy with the hybrid bagged approach implementation.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

An electronic mail or an Email is referred to the use of emails for sending the advertised emails and unsolicited emails to the group of recipients. The unsolicited email is signified that the permission has not granted by the recipient for receiving those emails. Since from the previous decades, the usage of popularity in spam emails is increasing day by day. In the internet, this spam became a immense adversity. Generally, the waste of message speed, time and storage is known as a Spam. One of the most efficient methods for the detection of spam is Automatic email filtering. But at the present day spammers evade all of those spam filtering applications simply. In the previous years, majority of the spam mails receiving from a finite email addresses can manually be blocked. Then the spam mails can be detected by using the Machine learning approaches. This Machine learning approach is the most effective and efficient one which uses a training dataset. Training datasets are the samples that take a set of pre-classified emails. Email filtering can be accomplished by using the several algorithms with which Machine learning approaches have. This includes the algorithms of "Neural Networks, support vector machines, K-nearest neighbor, Naïve Bayes, Random Forests etc.

The machine learning algorithmic method's affectivity inspires the proposed spam email detection system. Firstly collection of email data is accomplished in the spam mail detection system. The nature of unconstructed and raw is having in those collected email data. It is required to pre-process the email data for the optimal computations in addition to achieve the accurate results. The stop word removal and stemming processes are performed under the pre-processing of data. Then the valuable information can be acquired by performing the tokenization of words on the data while pre-processing. After that selection of the best features is performed from a group or set of features by using a correlation based feature selection (CFS) method. Experiments can be performed by using the email dataset collected from the "kaggle" website library. This can capable of editing the models, conducting the preprocessing and calculating the accurate results from the experiments. The hybrid bagged approach provides combination of two machine learning based classification algorithms such as Naïve bayes and j48 classification algorithms. Each of these classifier algorithms are applied with the email dataset as an input

* Corresponding author.
E-mail address: salimj06@gmail.com (S. Amirali Jiwani).

https://doi.org/10.1016/j.matpr.2021.03.147
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

which is divided into different sets of data randomly. Finally, overall result of classification can be calculated by averaging the both individual results of two machine learning algorithms using a bagging approach.

2. Related work

2.1. Mail spam detection

The more popular and widely used communication system is the electronic mail or Email communication system. Many organizations throughout the world have been exerting the efforts in identifying the spam emails. The authors who were discussed about the identification of these spam or ham are described further. The filtering techniques are need for the classification of the emails to be classified as a spam/ham. The spam email filtering system is proposed by the Selamatand Mohamad [3] in which emails are classified by the use of two different features selection methods. This considers the Malay and English and rough set theory method and TF-IDF were used to select the features from the email dataset after completion of pre-processing on the data. By presenting the practically best results, classification process is applied with the machine learning algorithm. Harisinghaneyin [5] proposed another work for the email dataset classification based on the machine learning algorithms. The implementation of algorithms which when they are applied with data for preprocessing presenting the effective results include Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, k-nearest neighbors (KNN) algorithm and Naïve Bayes. The techniques for optimization were also been adapted in this paper. A method of feed forward neural networks is used by the Farisin [4] for the detection of emails as spam and for effective optimization of results. The Krill Herd algorithm was used to train the neural network. Training and testing can be performed by dividing the preprocessed dataset into two equal halves. Then the comparison between the results attained from neural network of optimized classification and other optimization algorithms such as genetic algorithm and back propagation algorithm is done.

2.2. Machine learning

In implementing the machine learning models for the detection spam emails, researchers took their lead. By using the six different machine learning algorithms such as Artificial Neural Network (ANN), rough sets, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), system artificial immune, and Naive Bayes classification (NB), authors conducted experiments in the work [2]. Purpose of their experiment was to mimic the detection and recognition abilities of human. Tokenization was looked into, and the concept went through two stages such are training and filtering. Four steps were included in this algorithm they are: pre-processing of email dataset, feature description, classification of spam and evaluation of system performance. They concluded that the utmost precision, recall and accuracy were offered by using the Naive Bayes algorithm.

A hybrid system is proposed by Feng in [2] which described between two machine learning algorithms of NB and SVM. In this method, the generation of Hyperplane between the specified dimensions is accomplished by applying the SVM algorithm and then data points are removed leads to reduction of training set. Then the result probability was predicted by implementing this training set using the NB algorithm. The corpus text of Chinese was explored to conduct this experiment. Their algorithm proposed was implemented successfully with an increased accuracy in comparison to SVM and NB themselves. A hybrid-based algorithm was proposed by Wijaya and Bisri in [1] that integrates the decision tree together with the false negative threshold in logistic regression. The decision tree performance was increased successfully. On the Spam Base dataset, experiment was performed and results were compared with previous studies.

3. Spam mail detection using machine learning

This proposed approach of data science based email spam detection system should take the input with a email data set and then classified the spam emails among the dataset by using the data science model of text mining along with the machine learning based hybrid bagging approach. Fig. 1 represents the implementation of a proposed approach of data science for email spam detection using ML model.

3.1. Email dataset

The system of spam mail detection (SMD) utilizes the email dataset. The "kaggle" website is considered here for selecting the email dataset in which several emails are screened randomly. This dataset contains the emails of totally 1000 for the classifying of emails as ham or spam. This dataset is divided into two sets of data containing 500 emails in each for each of the classifier algorithm because of the using of hybrid bagging approach method. Each of these data sets are trained with the 300 emails and tested with the remaining 200 emails for both of the classifier algorithms, Naïve Bayes and J48 algorithm.

3.2. Raw email dataset

There is requirement of pre-processing the email dataset before going to the next considerations since it can be considered as a raw in nature. Firstly, the texted data is goes under the tokenization process. Secondly, removal of stop-word is carried out on the words of tokenization output. Lastly, reduction of words into their base words is done by the stemming process. Since the removal of

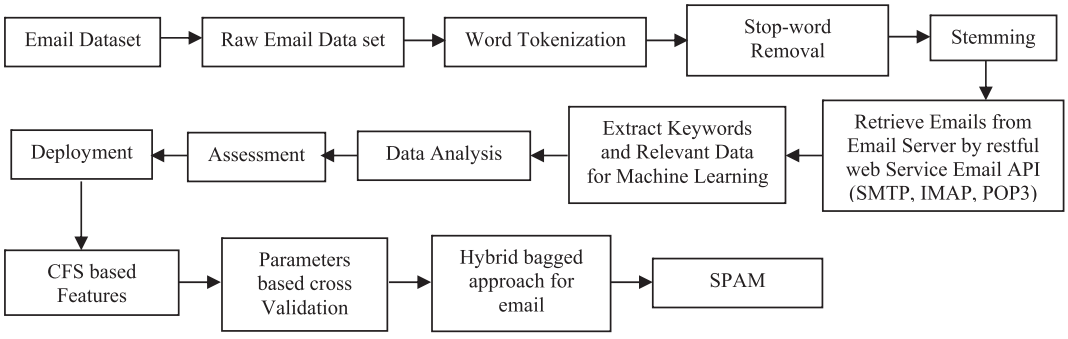

Fig 1. Implementation of a proposed approach of data science for email spam detection using ML model.

stop-word and stemming process are facilitate in reducing search space to extract and select the features efficiently, they are important to be pre-processed.

3.3. Word tokenization

The process of dividing a script stream into words, symbols, phrases or any element of expression called a token is known as "Tokenization". This token theory is still used as an input for further processing, e.g. for parsing and content mining. There is significance for the tokenization in both of these semantics (where it is like content mining), in a building and software engineering like a lexical testing. What is meant by the term "word" is sometimes difficult to define. Because tokenization is done at the word level, a token often relies on modest heuristics, such as: For example, tokens are separated by the characters such as "line break" or "space" or by "punctuation marks" or by whitespaces. A part of one token is referred to each one of neighboring stream of alphabetic characters andneighboring stream of numbers. The resultant token lists may or may not contain punctuations and white spaces.

3.4. Stop-word removal

The English words which don't give much meaning to a sentence are called "Stop words". By maintaining the meaning of the sentence, these stop words are ignored safely. Such as, if a query of "how to make a cheese and veggie sandwich" is made for searching then the web pages consists of a "how", "to", "make", "a", "cheese" "and" "veggie" "sandwich" words will be tried to searched by the search engine. However, the words "how", "to", "and" and "a" are the most widely used words of English language, search engine tries to search for the web pages which consists of words as recipes like "cheese", "veggie" and "sandwich". Then from this, by focusing on the result of search interest, those four words can be stopped or removed for retrieving web pages of the actual focus of interest like "veg", "cheese" and "sandwich".

3.5. Retrieve emails from email server by restful web service email API

In regardless of the server domain, retrieving of emails from the respective servers with 100% accuracy is the most important as a dataset is required for the classification algorithm on which its functions can be performed. The protocols like Simple Mail Transfer Protocol (SMTP), Internet Message Access Protocol (IMAP), Post Office Protocol 3 (POP3) etc. can be used for such retrieval operation. The classification problem can remain only once when the emails have been successfully retrieved in a PWA (Progressive Web App) or application.

3.6. Extract keyword and relevant data for machine learning

This process is done by scanning the text that includes the header and body in the every view of the email. The wide range of spam or ham mail classification is performed by scanning the specific keywords with the analyzing of text files in emails. The priority level then determined with reading of these parameters by the algorithm. Therefore, the data preparation contains the retrieved emails and their scanned word.

3.7. Data analysis

The input parameters on which the algorithm runs are taken into account after the completion of data preparation phase. For the recognition of similar patterns in future emails, this "trains" the classification model and save the time which would have been required for further comparisons. The results obtained after training had to be collected for further analysis. This is done primarily to find outliers: the guaranteed results that deviate from the norm in some of the factors are classified differently. The dataset is splitted into two types of results as ambiguous and reasonable by this outlier analysis. The output which is used to perform the further assessment is the correct output. Then debug the result and this result is return to the training phase by the algorithm.

3.8. Assessment

Now, it is required to perform the assessment on the correct result for the one last time before being filtered or classified. The algorithm was trained on the dataset which was already available. This also have a result in real time after retrieval of the users email. If the accuracy is to be optimized then it is required for contrasting and comparing the real time results and results of datasets. By using the same parameters, the algorithm is retrained further. This allows us to analyze the uncertainty of the received output and to further improve the level and accuracy of filtering and prioritization.

3.9. Deployment

A detailed report needs to outline and recapitulate the overall working of the system in a way that experts, as well as laymen, can grasp its basic functionality. The system ubiquity for users is critical for users to the successful deployment of the client online. Lastly, as the number of users increases scalability became necessary with no compromise of the system efficiency. Without compromising of the expected system performance, this algorithm and PWA should be scalable onto multiple devices.

3.10. Feature selection

Various features such as language, alphanumeric words, spellings or grammatical errors, inappropriate words (words related to advertising of products/services, quotes, words for adults, etc.), length of the document, frequency count, etc are contained in the feature set. A method of Correlation Feature Selection (CFS) is used by the SMD (Spam Mail Detection) system. The best features which are supportive for the improvement of system performance are identified by this CFS from a group of features. An assumption can be made to perform the CFS method that is "the features from subsets of good features are strongly correlated with the classification, but are not correlated with each other". Text data with a set of features is initially considered to be a bag of words. For the visualization of number of words per document, the term frequency method is considered. The words are eliminated which were having a frequency below the threshold value after calculating the frequency of total words. The importance of using words is explained and search space also reduced by this method. The method of correlation-based feature selection is used for further reusing of the resulting feature set.

3.11. Correlation based feature selection

A set consisting of features that are mostly relating to a certain class are only selected with this Correlation based feature selection method. The mathematical expression for the correlation based feature selection method is given in the Eq. (1).

$$\boldsymbol{CFS} = \boldsymbol{max}_{\boldsymbol{S}_k}\left[\frac{\boldsymbol{r}_{\boldsymbol{cf}1}, \boldsymbol{r}_{\boldsymbol{cf}2}, \boldsymbol{r}_{\boldsymbol{cf}3}\cdots, \boldsymbol{r}_{\boldsymbol{cfk}}}{\sqrt{\boldsymbol{k}+2(\boldsymbol{r}_{\boldsymbol{f}1\boldsymbol{f}2}, +\cdots \boldsymbol{r}_{\boldsymbol{fifj}}+\boldsymbol{r}_{\boldsymbol{fkf}1},)}}\right] \quad (1)$$

Where, the number of classes is $\boldsymbol{c}$, the number of features with a feature set $\boldsymbol{f}$ is $\boldsymbol{k}$ number of features and the number of classes is $\boldsymbol{c}$. Then in the equation (1), $\boldsymbol{r_{cf}}$ signifies the average correlation between features to class and $\boldsymbol{r_{ff}}$ signifies the average correlation between features to feature.

3.12. Hybrid bagged approach

The classification process considers a Naïve Bayes Multinomial classifier in addition to the J48 (decision tree based algorithm) classifier in a hybrid bagged approach. The combination of multiple repeated sets of the same data set is considered with which the variance is decreased is called as Bagging approach or bootstrap aggregating approach The entire email dataset can be divided randomly into the separate email datasets such as SED1 and SED2 (Spam Email Dataset 1 and 2) thereby generating the multiple models. By taking into consideration of each of this sample email dataset both of the two classifiers are trained individually. The average of two individual classifier algorithm's results will give the result of overall system. The multi class learning followed by classification of spam emails can be carried out by suing the Naïve Bayes and decision tree based J48 classifier algorithms. The values that were predicted in classification are averaged and then regarded as the result of the classification.

3.13. Parameter cross validation

The results are acquired as reliable by applying with the K-Fold cross validation method. However, this K-Fold cross validation method has some limitations of chance of total testing dataset having with the all spam mails or the training dataset having with the majority of the spam mails. Within a distributed set, making a best range of spam/ham while separating the data can resolve the limitations of the K-fold cross validation which was supported by the Stratified K-fold cross validation. Finally, the machine learning algorithm's accuracy can be improved by conducting the tuning of parameters using Scikit-Learn approach.

3.14. Spam

In accordance with the Wikipedia, the spam can be referred as using of electronic message system for sending the huge number of advertisements, malicious links etc. called as unsolicited mails to the group of email addresses. The meaning for unsolicited is that the recipient not granted the permission for sender to send the mails. That means a mail is said to be a spam if a message or mail is coming from the unknown sender email. Generally while downloading of any software or free services and while software update processing, users simply signed in for those mailers without any knowledge about it leads to getting the spam mails.

4. Results

The "kaggle" website is used to acquire the training dataset. The training email dataset consists of totally 1000 numbers of mails and two machine learning algorithm for the each of 500 mails is considered for the classification experiment. The two algorithms that are used are Naïve Bayes and J48 which were brought together for the best of both algorithms using a hybrid bagged approach. The aggregate prediction result of these two algorithms gives the overall result of spam email detection system. Therefore reliable and accurate results are ensured in a system.

The spam mail detection system's performance can be evaluated by using the parameters such as true negative rate, false positive rate, false negative rate, recall, precession, accuracy and F1-measure. Evaluation of these performance parameters can meet the efficiency for the proposed spam mail detection system. In a binary classification point of view a decision theory is used for the detection of mails as spam or ham. For this, a finite probability based mathematical model is used to get an unknown output. According to the amount of data that is referenced, the optimal solution can depends. Accuracy is high for the larger dataset.

4.1. Confusion matrix

Several performance measures can be used to evaluate the spam email detection. The visualization of email detection for those models is being carried out by using the Confusion Matrix. The definition for the confusion matrix can be given as,

$$\boldsymbol{Confusion\ matrix} = \begin{bmatrix} \boldsymbol{TN} & \boldsymbol{FP} \\ \boldsymbol{FN} & \boldsymbol{TP} \end{bmatrix}$$

Where,$\boldsymbol{TP} = \textbf{TruePositive} \rightarrow$ Spam email predicted as spam,$\boldsymbol{FP} = \textbf{FalsePositive} \rightarrow$ Spam email predicted as ham,$\boldsymbol{TN} = \textbf{TrueNegative} \rightarrow$ Ham email predicted as ham,$\boldsymbol{FN} = \textbf{FalseNegative} \rightarrow$ Ham email predicted as spam. Then the experimental results of this Confusion matrix parameters are tabulated in Table 1.

4.2. Precision

The values that are correctly identified are calculated by using the measurement of precision. The spam emails are classified from the known set of positive emails. The precision gives the classified number of spam emails that have been correctly identified.

$$Pecision = \frac{TP}{TP + FP}$$

4.3. Recall

The number of emails that are correctly calculated as a spam among the total number of available spam emails is given by the measurement of recall. It can be defined as,

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

4.4. Accuracy

The main aim of this paper is to get the high accuracy of identify the emails in a correct manner as ham and spam**.**

$$Accuracy = \frac{(TN + TP)}{(TP + FN + FP + TN)}$$

4.5. F1-Score

The recall and precision scores used to calculate the F-measure or value of $\boldsymbol{F_{\beta}}$. Here $\boldsymbol{\beta}$ is recognized as 1 then the value of $\boldsymbol{F_{\beta}}$ or F1

Table 1
Result for confusion matrix parameters.

Evaluation measures	Naïve Bayes	J48	Proposed hybrid bagged approach
True positive	82	86	90
False positive	15	8	7
True negative	87	90	92
False negative	18	15	13

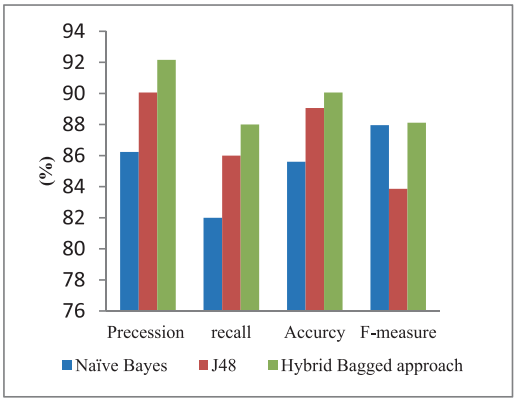

Fig. 2. Comparative performance of naïve Bayes, J48 and proposed hybrid bagged approach SMD.

provides the***F***1 – ***score***. The "Harmonic mean" of recall and the precision values is defined as "***F***1 – ***score***."

The comparative analysis of the results as from Fig. 2 clearly indicates that better results are achieved in terms of precision, recall and accuracy with proposed hybrid bagged approach when compared with the Naïve Bayes and J48 decision tree algorithm.

5. Conclusion

In a security point of view classification of emails as a spam and ham has most important for the users. So, a data science approach of Spam Mail Detection system was proposed in this paper. This uses a hybrid bagged approach for its implementation with the two machine learning algorithms. Naïve Bayes and J48 are of those two machine learning classification algorithms used in this approach. Then by using the dataset that is classified in previous, these two classification algorithms are trained and classified the emails. This function is also extended for classifying and displaying the emails that are received in an organized manner. From the results it can be illustrated that the proposed hybrid bagged approach based SMD system has gained an overall accuracy of 88.12% and achieved a better performance than both the individual classification algorithms.

CRediT authorship contribution statement

Rakesh Nayak: Conceptualization, Methodology, Software, Visualization, Supervision. **Salim Amriali Jiwani:** Writing - original draft. **B. Rajitha:** Data curation, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Bisri, A. Wijaya, "Hybrid decision tree and logistic regression classifier for email spam detection", in Proceedings of 8th Inter., Conf., Info., Tech., Elect., Eng., (ICITEE), pp: 1-4, Oct, 2016.
[2] J. Sun, C. Cao, Q. Yang W. Feng, and L. Zhang, "A support vector machine based Naive Bayes algorithm for spam filtering", in IEEE 35th Proceedings of International Performance Comp. Comm. Conf. (IPCCC), Dec. 2016, pp. 1-8.
[3] A. Selamat, M. Mohamad, "An evaluation on the efficiency of hybrid feature selection in spam email classification", In Proceedings of 2015 Int., Conf., Comp., Comm., & Contr., Tech., (I4CT), Kuching, Sarawak, Malaysia, pp:227-231, 2015.
[4] I. Aljarah, H. Faris, "Optimizing feedforward neural networks using Krill Herd algorithm for e-mail spam detection", In IEEE Proceedings of Jordan Conf., Appl., Elect., Eng., & Comp., Tech., (AEECT), Amman, Jordan, pp.1-5, 2015.
[5] A. Harisinghaney, A. Dixit, S. Gupta, A. Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm", In Proceedings of 2014 Inter., Conf., Opt., Rel., & Inf. Tech. (ICROIT), Faridabad, Haryana, pp.153-155, India, 2014.

Further Reading

[1] Mamoun Alazab, Asif Karim, Bharanidharan Shanmugam, Sami Azam, Krishnan Kannoorpatti, A comprehensive survey for intelligent spamemaildetection, IEEE Access 7 2019.
[2] R. Shams, R.E. Mercer, "Personalized spam filtering with natural language attributes", Proc. 12th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), pp. 127-132, Dec. 2013.
[3] O. Amayri, N. Bouguila, Content-based spam filtering using hybrid generative discriminative learning of both textual and visual features, Proc. IEEE Int. Symp. Circuits Syst. (2012) 862–865.
[4] W. Awad, S. ELseuo, Machine learning methods for spam E-Mail classification, Int. J. Comput. Sci. Inf. Technol. 3 (1) (2011) 173-184.
[5] N. Raad, G. Alam, B. Zaidan, A. Zaidan, Impact of spam advertisement through E-mail: A study to assess the influence of the anti-spam on the E-mail marketing, Afr. J. Bus. Manage. 4 (11) (2010) 2362–2367.
[6] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Trans Pat. Ana. Mach. Intel. 28 (7) (2006) 1088–1099.
[7] G. Wu, E.Y. Chang, KBA: Kernel boundary alignment considering imbalanced data distribution, IEEE Trans. Know. Data Eng., 17 (6) (2005) 786-795.
[8] M. Aery, S. Chakravarthy, eMailSift: email classification based on structure and content, Proc. 5th IEEE Intl. Conf. Data Mining (2005) 1–8.
[9] H. Drucker, D. Wu, V.N. Vapnik, Support vector machines for spam categorization, IEEE Trans. Neural Networks 10 (5) (1999) 1048–1054.
[10] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.