[[1]](#footnote-1)

RANDOM FOREST BASED SPAM MAIL CLASSIFICATION

Muhammed Swalihu, *P2703090*, *P2703090@my365.dmu.ac.uk*

***Abstract*—** **Spam emails are a regular incident in everyone’s life who uses emails. In everyday, there are more than hundreds of spam emails having a different content from unidentified addresses and which are produced automatically by some software agents. Black lists and white lists are some traditional email spam classifiers but which are not work perfectly. On recent, the machine learning techniques works perfectly applied to finding the spam emails. The objective with spam email classification is to filter the Spam and not spam emails and taking out the spam mails.**

**The classification algorithms such as Support Vector Machines, Naïve Bayesian and many more are used in different data sets and they are displaying good classification results. This report refers the classification of emails using Random Forest (RF). RF consist of consist of individual trees. Each individual tree votes the classification of the data that given and the RF algorithm chooses the classification depends on the most vote. It’s a data mining technique known as ensemble learning. From the dataset spam email classification from the Kaggle (named as sed.csv) consisting of 5172 rows and each row containing emails. Coming to the columns, there are 3002 columns. In the first column indicates the email name and last column indicates the predictions that is: 1 for spam, and 0 for the not spam. The rest of 3000 columns which is for the 3000 common words in each email. As a result, the data regarding of all 5172 are stored in a solid data frame rather than a text file.**

***Index Terms:* Machine Algorithm, Algorithm classification, Random Forest Classifier, Accuracy, Data mining, Ensemble learning, spam.**

I. INTRODUCTION

T

he email is become one of the powerful tools in the world for exchanging the messages and conversations. In addition to the growth of email, there has been a powerful growth in the spam on these recent years. On the recent years there is a hike in the anti-spam agencies and technologies. which makes us surprise, at same the growth of spam emails also increased rapidly. Which results many users must now spend an amount of time to go through the unwanted junk emails. In addition, spam emails not only waste your time but also finish the server space. Looking to the large websites with a large number of users who may get the same copies of the spam emails. This makes us to make automated technique for filtering such spam emails. There are lot of challenges to create an automated spam email classification because of unstructured information and the large number of documents.

The mystery behind the spam emails is hard to define.

Spammers collect the email addresses from websites, chat rooms and many other sources which collects the users addresses and then sold to the other similar spammers. [1] In these years, every user on the cyberspace get 20-50 spam messages in every day and about 22.43 billion unsought commercial messages are sent every day. According to the latest spam traffic statics the average volume of email globally is about 122.32 billion. From that there are nearly 85% of all emails are spam. The most common spam emails are from the advertisement messages and the second one is for the adult related content nearly 31.7% [2].

It was reported that, United states receives 7.71 billion spam messages in 2020 and then gradually increased to 8.61. China is the second country who receives more than 8.53 billion spam messages in 2021 and then Russia and Brazil in the 3rd and 4th positions [3].

The algorithm classification for spam emails have lot of challenges because of large number of documents in the data set and the different features. The classification algorithm like Support Vector machine (SVM), Naïve Bayesian (NB) are currently used in these datasets and having a good accuracy result.

This report discussing the classification of junk emails by the method of Random forest (RF). Random forest is a data mining tool which is also known as ensemble learning. Ensemble learning, which creates several classifiers like decision trees, and choose the result by using weighted vote on the predictions. At first, the body of emails assessed and then pre-processing and extracting the token. Secondly, by using selection technique the more refined terms will hold there and others were rejected. Then the iterative patterns were returned and then the feature vector of every sample were formed. At last, Random Forest (RF) is used to filter the data.

The Random Forest is good example for regression technique and ensemble learning and this is well fit for data filtering problems. It was founded by the Breimam and Cutler. The Random Forest (RF) has good impact over the years and now it is used for filtering issues. Here RF method uses decision trees to distinguishing the data to make it in to definite categories. On this technique, which considered only based on the elected class of individual trees, where the selection will be considered only basis of which class receives more votes than other. When Random Forest compares with the other machine learning algorithms, RF offers several advantages, which includes lower classification errors. Moreover, RF performance is better than that of SVM. Random Forest is an efficient algorithm for measuring the values of the missing data and securing accuracy in every situation when a large amount of the data is missing. It can balance the data sets with missing values. Compared to Support Vector Machine the training time was exceptionally low. Concerning the accuracy, Random Forest exceeds the more existing machine learning techniques. It can process a thousand variables effectively and it can perform in big data sets. Random Forest has the capacity to handle unlabeled data and makes them a good method for grouping the unlabeled data. In addition, Random Forest requires low parameters when comparing with the whole amount of data, as a result Random Forest is straightforward. The Random Forest allows the user to grow many trees. Here the each trees perform its own grouping function. Then which is marked as the trees votes for the class. Then the forest (each trees are in the forest) chooses the class with the highest votes.

The sections of this report consist of eight parts. The first part was the outline research. The second part was the discussion of the previous research and the study objectives. The third part refers working procedure of Spam. The fourth part discuss the Proposed Algorithm which includes the preparation of datasets, pre-processing of the data, Feature extraction and many more. The fifth part is going with implementation and the sixth part refers the results and discussion. The seventh part assign with conclusion and future scope. The eight part is going with acknowledgment. And last part refers with the bibliography

# II. A LITERATURE SURVEY ON SPAM FILTERING

There are lot of studies on spam filtering using machine learning algorithms. This section covers the number of experiments. According to the [4] G. Mujtaba (2017), he proposed the basic three steps for every classification process. The first one was preprocessing, where the text which is given in the email is converted into the tokens and the avoiding of stop words. The second step was the learning process, and this step is particularly important for the optimization of emails. The third was classification of the messages as spam or ham by using capable algorithms. He considered support vector machine, Random forest and logistic regression as classifications. On his project he used Corpus dataset.

In accord with Divesh palival (2018), [5] the best communication method on cyberspace is messages and emails. Where most of the messages were marked as spam. On these spam emails appear to be real links, while pointing those links moves to website and then those website redirect into undesirable websites. As a result, Divesh approach which ID3 based on the decision tree to reduce these problems. He constructed the decision tree by using Enron data set.

Dhanendra Kumar Dewangan (2018) makes a unique method by using the Support Vector Machine which identify the spam emails. Emails and messages are a major tool of online communication. Spam is a major issue and spreads the virus through the medium. Moreover, which wastes the users time to find out real messages from those junk messages. As a result, spam detection is necessary to minimize the shock. Looking at the multiple series of data sets, the Support vector Machine performs well, and the accuracy was great. He proposes that most of the datasets are too old. As a result, the spammers are changing the content material to avoid the filters.

Rathii. (2013), [6] She proposed a technique using the data mining for analyzing the accurate classifier for the email spam classification. She goes through the many data mining techniques for finding the the performance of different classifiers with the feature selection and without the feature selection algorithm. After finding the best feature selection algorithm they studied selected algorithm for the feature selection purposes. She done her experiment by using several machine algorithms such as J4, Naïve Bayes, Random Forest and Random Tree, bayest Net and Support Vector Machines. Her datasets consist of 58 attributes and 4601 instances. Random Tree performs well among these classifiers with highest score of 99.7% and the lowest score was for the Naïve Bayes and the score was 78.94% accuracy.

Abdulhamid Muhammad shafi & M.S Osho (2014) [7] were done an analysis on different machine learning classification tools such as Random Tree, Bayesian Logistic Regression, Lazy Bayesian rule and J48. Moreover, they compared all these algorithms based on precision, recall, accuracy, and F-score. Here they used the dataset from UCI machine learning repository. For analyzing the recall and precision here they used F-measure as a method. The rotation forest algorithm has the highest F-measure and the Naïve bayes algorithm has the lowest F-measure. For the Rotation Forest algorithm, they got 87.9% and they used Kappa statics for statistical results. The accuracy score for the Rotation Forest algorithm is 94% and the lowest accuracy score is about 89% obtained by REP Tree algorithm. Accuracy for Naïve Bayes is about 88.5% and accuracy for the J48 is around 92.3%. So here the Rotation Forest algorithm works well in all terms of accuracy, precision, recall and F-measures.

Linda Huang & Julia Jia (2018), [8] they produce a solution to increase the efficiency of Naïve Bayes and they shorten the false detection rate. Already discussed above the Naïve Bayes is based on Bayes theorem and it can be used as probabilistic model for the optimization of emails. The Naïve Bayes have good accuracy and at same time spammers may be able to ignore the filter by using diacritics and Leetspeak. Diacritic is an accent, where this written on above or below of a letter shows different pronunciation from the same letter when it is differently marked. Coming with Leetspeak, it’s a coded system where language used is here is very intimate communication. Featuring the alphabet with special characters or numbers may feature and then include miss spelling and slang. Linda & Julia have made some changes in the Naïve Bayes to change the symbols in the text into new possible letters and then checked the spell to make sure that is word from the symbol. After that it passed through an algorithm for the optimization. During this experiment they got 62% accuracy.

III. HOW ANTI-SPAM WORKS

This works in two phases. One is Classification and the second one is Action.

**Classification**

This is phase one, where the emails are subjected to analysis to determine whether it's junk or not. During this analysis it's not only look at the characteristics that makes the email spam, but also look at the characteristics that makes it not a spam.

**Action**

This is the second phase of working, on this phase the classified emails can be avoided, altered in somewhat way or emails can be delivered if the message was not a spam one.

Here are some basic steps in phase 2, which are mentioned below

Step 1

Collecting spam and non-spam data.

Step 2

The second step is all about the preprocessing of the data. Which includes the reduction of noise and cleaning of the data.

Step 3

Identifying the specific and repeated spam sender, by reading the header body.

Step 4

Now it's time to Applying the classification tools. Here we have to apply specific classification tools which depend on the requirement.

Step 5

Store the result in the form of text, content or image based.

TEST DATA

CLASSIFIER MODEL

TRAINING DATA

CLASSIFIERS

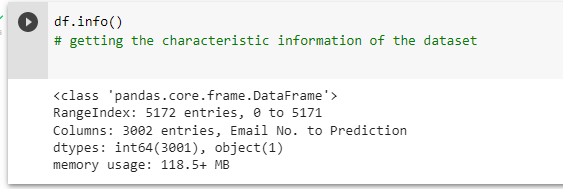
# IV PROPOSED ALGORITHM

METHODS AND DATA SET PREPARATION

Google colab is the platform using here. Uploading the dataset named as “sed.csv” in google colab. The “sed.csv” contains 5172 rows and 3002 columns. Each raw indicates each email. The below table is showing the sample of email imported to colab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Email No. | the | to | ect | and |
| Email 1 | 0 | 0 | 1 | 0 |
| Email 2 | 8 | 13 | 24 | 6 |
| Email 3 | 0 | 0 | 1 | 0 |
| Email 4 | 0 | 5 | 22 | 0 |

This is the information of the datasets



PRE-PROCESSING DATA

Pre-processing is one of the important stages in email filtering. In pre-processing there are three phases Tokenization, stop word removal and stemming. During the Tokenization phase all the symbols like (& %, #) and the numbers will be removed, and the remaining words will be tokenized. Stop words are the common words which are used in messages. These kinds of words are meaningless and provide information such as “we,” “are,”. These kinds of words are removed during phase 2. Eliminating these words frees up the storage and then reduces the search process. The third phase, “Stemming” is the process for finding the origin of the words and then removing the prefixes and post fixes. These phases convert words such as verbs and adjectives to similar words.

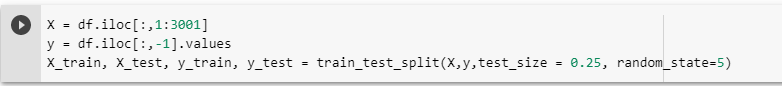
FEATURE EXTRACTION

This comes after the pre-processing phase. Feature selection is the procedure of selecting the subset of the words occurring in the training set. Feature extraction built the classifier more efficiently by reducing the range of vocabulary and boosting the classification accuracy by deducting the noise features.

TEXT VECTORIZATION

Here we will use the Term frequency & Inverse Document frequency (TF & IDF). These are similar embedding tools which are considered in each term of the file. There are a lot of advantages to the vectorizers. In the case of term Frequency & Inverse document Frequency, which are simpler than the other vectorizers and have good efficiency for vectorizing the emails. The tokenizer splits the email based on the special characters and the white space. When comparing it to word analyzer, this pulls more data. The tokenizer fails with informal English and fail to separate URLs. So here we use word-level analyzer to categorize the words. The above mentioned vectorizer, here we divide the data into two sets. The two sets are testing and training.





EVALUATION MEASURES

The evaluation measure which are used in our testing process are below

**True Positive (TP):** True positive refers to the number of junk emails classified as junk.

**True Negative (TN):** This refers to the number of non-spam emails classified as non-spam.

**False Positive (FP):** This refers to the no spam classified as non-spam emails.

**False Negative (FN):** This refers to the number of non-spams classified as non-spam.

**Precision:** Precision is the ratio between true positives to true and the false positive. This identifies the objects in the class.

*Precision = True positive / (True positive + False positive)*

**Recall:** Recall is the ratio of true positives to the true positive and the false negative. Which shows how many objects in the class disarranged.

**Accuracy**: Accuracy is the total sum of the positive and true negative to the that of total number.

*Accuracy of classifier: (True positive + True negative)/ (True positive+ True negative + False positive + false negative)*

**Confusion Matrix:** This is the one of the methods to check out the performance of classifiers using the confusion matrix. The number of accurately classified instances is the sum of diagonals in the confusion matrix and all others are classified incorrectly.

RANDOM FOREST ALGORITHM

We already discussed the Random Forest; it is a metal learner having lot of individual trees. Each tree in the forest votes for overall classification on the given data set and then the random forest chooses the classification with the most votes. The each tree assembled from a subset of the training dataset. Some entities appeared more than once or two in the sample, and the other doesn’t appear. For creating each decision tree, a model is based on individual random subsets of the training dataset and the random subset of applicable variable to regulate the division of data sets. Here each decision tree is built to the complete potential size without any clipping.

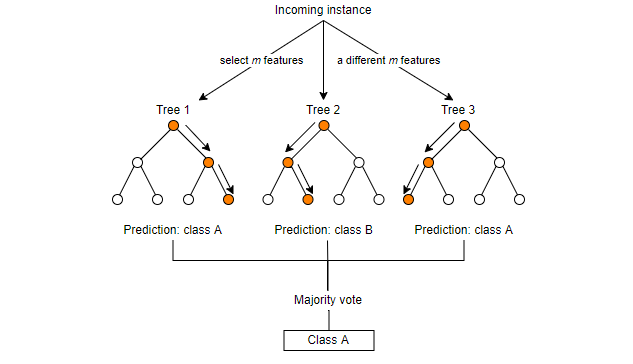


Diagram for the Random Forest classifier

V IMPLEMENTATION

We already discussed the ensemble classification earlier, it’s a learning algorithm that builds a group of classifiers instead of a single classifier and then classifies the data points by the vote of the predictions. It’s a reliable mapping that can find output of the multiple classifiers. [9]. The Random Forest was developed by Leo Breimans and Adele Culter. Figure 2 indicates ensemble learning

Classifier 3

Classifier 2

Classifier 1

Combiner

Final classifier

Ensemble learning (2)

The data categorized in the form of the vectors and the action of categorization is the movement of each vector down along with the binary tree. Here a special condition is given at every node to choose right or the left direction. Here is the pictorial representation:

The estimated category is placed at base parts of the tree, and it is found when the branch is reach nearer. The above already discussed about the forest, that is the group of trees is called the forest. If the tree provides multiple categories, then the right one will select according to the higher votes. Here the word random is mentioned due to the nodes and switching conditions assembled by using the method random sampling from training data sets.

Each tree is built using the following algorithm:

**Start RF Algorithm**

**Input:**

X: it is the no of nodes

N: no of features

Y: no of trees to grown

**Output:**

G: it is the class with leading votes

**While stopping criteria is not true do**

Now selecting a safe starting sample S randomly from training data and create tree Ri from the above selected starting S. By using the steps below.

1, Now selects the n features from N(randomly), Where the n is less than that of N (n < N).

2. Calculating the finest dividing value for the node d among the n features

3, Now divides the parent node by two spring nodes by optimal divide,

4, Now it's time to execute the first three steps to get the maximum number of nodes (x). Create the forest by repeating steps 1 to 4 in Y several times.

**End while**

The Produce the output for each individual tree.

Now use an advanced sample for each separate created tree from the root nodes.

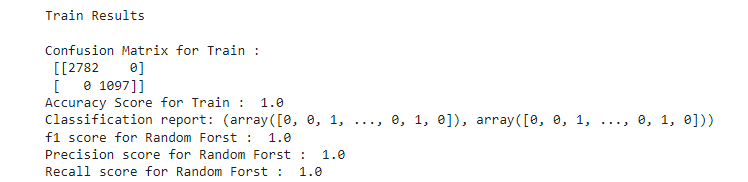
Now merge the votes from each tree

It's now time to find the class with highest number of votes.

VI RESULTS AND DISCUSSION

In this section presents the result of email spam classification by using the random forest. The data set “sed.csv” was evaluated by the random forest algorithm, this was done by an approach assessing predictive models which classify the original set into the training sample set to train the model and test set for calculation. Firstly, the training of data set was achieved with the vectors and then analyzing each email with header, and then checking of the keywords. Classification accuracy is one of the main performance metrics for email spam optimization. It is the ratio of the correctly classified instances in the data set and absolute of test results.

 The train results are given below:

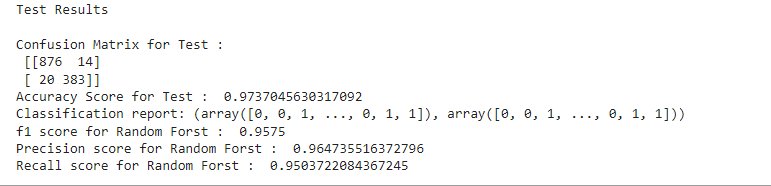


For train results we get the accuracy score of 1.0 and for f1 score, recall and precision score are 1.0.

Confusion matrix for train results:

|  |  |  |
| --- | --- | --- |
| Predicted class  Actual class | Spam | Legitimate |
| Spam | 2782 | 0 |
| Legitimate | 0 | 1097 |

Test results is the main results, which are shown below:



Result for the Test, where accuracy is 0.973 and the f1 score is 0.9573 and for precision is 0.9647 and the recall was 0.9503. As a result, the random Forest performed well in the above result.

Confusion matrix for the test results

|  |  |  |
| --- | --- | --- |
| Predicted class  Actual class | Spam | Legitimate |
| Spam | 876 | 14 |
| Legitimate | 20 | 383 |

VII CONCLUSION AND FUTURE WORK

In this report, email spam classification is done by using random forest algorithm and spam email classification dataset from Kaggle is used here (named as “sed.csv”). Here is the result analysis in terms of accuracy to build the model using the random forest. Random forest contains numerous trees, and it is based on the electing vote that we already discussed above. Spam classification which is used for finding spam messages and legitimates. The advantages of Rand om Forest will handle the extensive number of the input variables. We also captured the Receiver Operating Characteristic (ROC) for the datasets by RF.  The results reported in terms of False positive and True positive. The conclusion from the above results is that the number of junk messages is a very serious problems and which is spreading awfully. Random Forest classifier shows higher accuracy in terms of train data and test data. The train accuracy for the “sed.csv” file is 1.0 and for the train data which is 0.97% accuracy. Random Forest runs well on large data sets. Moreover, it can estimate the missing data and still maintain the accuracy score when the large portion of the data is away.

In future I am planning to do different classifiers to evaluate the performance and accuracy score. The parallel algorithm may be helpful for reducing the time appropriate for the classification. The Feature selection may help to reduce the redundancy of the of the dataset. I am planning to apply this on other tools such as Tangra, Rapid Minor and Matlab.

VIII REFERENCES

[1] (Winder, 2022)

Winder, D., 2022. *https://www.forbes.com/sites/daveywinder/2020/05/03/this-surprisingly-simple-email-trick-will-stop-spam-with-one-click/?sh=12d58037915d*. [online] Forbes. Available at: <https://www.forbes.com/sites/daveywinder/2020/05/03/this-surprisingly-simple-email-trick-will-stop-spam-with-one-click/?sh=12d58037915d> [Accessed 25 May 2022].

[2] (What's on the Other Side of Your Inbox - 20 SPAM Statistics for 2022, 2022)

Ref1 Dataprot. 2022. *What's On the Other Side of Your Inbox - 20 SPAM Statistics for 2022*. [online] Available at: <https://dataprot.net/statistics/spam-statistics/> [Accessed 25 April 2022].

[3] (Email Classification Research Trends: Review and Open Issues, 2022)

Statista. 2022. *Highest number of daily spam emails by country 2021 | Statista*. [online] Available at: <https://www.statista.com/statistics/1270488/spam-emails-sent-daily-by-country/> [Accessed 25 April 2022].

[4] (Email Classification Research Trends: Review and Open Issues, 2022)

Ieeexplore.ieee.org. 2022. *Email Classification Research Trends: Review and Open Issues*. [online] Available at: <https://ieeexplore.ieee.org/document/7921698> [Accessed 25 May 2022].

[5] (EMAIL SPAM FILTERING USING DECISION TREE ... EMAIL SPAM FILTERING USING DECISION TREE ALGORITHM Divesh Palival, Kevin Printer, Ramchandra Devre, Asst.Prof. Nikita Lemos Abstract&mdash;, 2022)

pdfslide.net. 2022. *EMAIL SPAM FILTERING USING DECISION TREE ... EMAIL SPAM FILTERING USING DECISION TREE ALGORITHM Divesh Palival, Kevin Printer, Ramchandra Devre, Asst.Prof. Nikita Lemos Abstract&mdash;*. [online] Available at: <https://pdfslide.net/documents/email-spam-filtering-using-decision-tree-email-spam-filtering-using-decision.html?page=1> [Accessed 25 May 2022].

[6] (Spam Mail Detection through Data Mining – A Comparative Performance Analysis, 2022)

Mecs-press.org. 2022. *Spam Mail Detection through Data Mining – A Comparative Performance Analysis*. [online] Available at: <https://www.mecs-press.org/ijmecs/ijmecs-v5-n12/IJMECS-V5-N12-5.pdf> [Accessed 25 May 2022].

[7] (Comparative Analysis of Classification Algorithms for Email Spam Detection, 2022)

[8] (Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection, 2022)

Ieeexplore.ieee.org. 2022. *Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection*. [online] Available at: <https://ieeexplore.ieee.org/document/8455990> [Accessed 25 May 2022].

[9] (ML Classifier Performance Comparison for Spam Emails Detection- Part 1, 2022)

Medium. 2022. *ML Classifier Performance Comparison for Spam Emails Detection- Part 1*. [online] Available at: <https://towardsdatascience.com/ml-classifier-performance-comparison-for-spam-emails-detection-77749926d508> [Accessed 25 May 2022].

1. [↑](#footnote-ref-1)