National College of Ireland

# Ensemble Classification Method for Email Spam Prediction

MSc Research Project
Cyber Security

Chinedu Timothy Udogwu
Student ID: 19222360

School of Computing
National College of Ireland

Supervisor: Imran Khan

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | CHINEDU TIMOTHY UDOGWU |
| **Student ID:** | 19222360 |
| **Programme:** | MSc Cyber Security **Year:** 2021 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Imran Khan |
| **Submission Due Date:** | 16/12/2021 |
| **Project Title:** | Ensemble classification method for email spam Prediction |
| **Word Count:** | **20 Pages** Count 5500 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Chinedu Timothy Udogwu |
| **Date:** | 15/12/2021 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# ENSEMBLE CLASSIFICATION FOR EMAIL SPAM PREDICTION

**CHINEDU TIMOTHY UDOGWU**

**19222360**

**ABSTRACT**

Abstract Every day, email users receive a great deal of spam emails from unknown senders in their inboxes. Spamming has indeed been linked to social engineering, which has resulted in online cyber fraud. This usually starts with an email from an untrustworthy source that contains a URL that, when opened, might compromise one's personal data. The concept of machine learning has been well investigated, and there are numerous algorithms that can effectively do this task. Pre-processing, feature engineering, and the machine learning algorithm are the three steps in the pipeline of an email spam filtration system based on machine learning. Some words, such as combination words, articles, and others, are removed from the email composition in the first step of the training filter, which is the pre-processing of e-mails, because they play no part in categorization. Following the receipt of an email, complete this step. Feature engineering is a decision-making method that employs some previously learnt features from a set of training instances. The values of the features may differ. The authors of the email spam detection built the various machine learning methods. In this study, an ensemble machine learning algorithm for email spam prediction will be established.

*Keywords: Security, Machine Learning, Email Spam, Ensemble Classification*

## 1. INTRODUCTION

The use of E-mail keeps on developing alongside different strategies for interpersonal communication. In 2017, the absolute number of business and client Emails delivered and obtained every day arrived at 269 billion. This number is likely to increase continuously at a normal yearly rate of 4.4% throughout the following four years, arriving at 319.6 billion before the finish of year 2021. Email spam is an expanding issue that influences ordinary web clients as well as turns out to be a significant issue for organizations and businesses.

Email is a powerful, quick, and modest mode of communication. Thusly, spammers want to send spam through such sort of communication. These days, almost all clients have an E-mail, and subsequently they are confronted with spam issue. Spam is a major issue both for clients

and for Internet Service Providers (ISPs). The reasons include are emergence rate of electronic communications from one viewpoint and increment of spam transferring innovation from other point of view. The accessibility of email makes it prone to numerous dangers triggered by hackers. Spam is a huge danger to email; basically, all email clients on the planet endure spam. The word ‘spam’ characterizes the unwanted message, junk mails delivered to the inbox of internet users. Therefore, Email Spam can be defined as non-requested data transferred to the E-letter boxes.

It is thus, highly favorable for email spammers to send heaps of messages to huge number of clients in simple and cost-effective way [1]. It generalizes this issue for all internet clients to get spontaneous email routinely. Spam emails turns out to be the reason of lower efficiency; consume space in letter boxes; expand bugs, Trojans, and materials carrying possibly dangerous data for a specific group of clients; thrash stability of ingoing mails, and consequently clients invest their precious time for arranging approaching mail and erasing in convenient messages. In view of the light of considerable amount of spam mails coming to E-letter boxes, it can be said that spammers don't work alone; it is worldwide, coordinated, making the virtual interpersonal networks. They launch attack on mails of clients, entire organizations, and even states.

Filtering spam using conventional techniques such as dark-white lists (domains, IP addresses, mailing addresses) is practically not possible. Applying text mining strategies to an E-mail may increase productivity of email spam filtration [2]. Additionally, forestalling spam messages will be conceivable to build up topical reliance from topographical (e.g., what subjects are most highlighted in the spam-messages transferred from the specific nations) features. In the past decade, lots of techniques for text clustering and classification have been effectively applied to tackle spam issue.

Many technical and practical methods in the past decade have been applied to filter out spam emails from a bunch of emails. Many machine learning algorithms have been proposed for classifying the emails into spam or ham. This paper is based upon many classification algorithms including an ensemble method to classify the emails into spam or ham.

### 1.1 Research Question

Can the performance of various ML models be analyzed and compared?
Does the ensemble email spam ML technique offer better performance as compared to a single ML model?

### 1.2 Research Objectives

1. To review and examine different types of data mining-based email spam prediction algorithms.
2. To implement existing machine learning algorithms for classifying email messages as spam or ham.

3. To compare performance of existing and proposed methods in terms of certain parameters.

4. To compare the training and testing time of the various ML models.

### 1.3 Contribution

There are several challenges in email spam detection. As shown in literature survey, most of the authors have used the dataset of Twitter for the email spam detection. But the major problem with twitter dataset is that usually tweets are limited to some words that sometimes not express the exact meaning. There may be some challenges that people express their review in different ways. Generally, people have different view regarding a similar text. The majority of remarks are based on other things. For improving this, it is required to implement a technique that can perform classification as well as analysis of the email spam detection. So, in this proposed work, we are considering the concepts of NLP and ML algorithm to the learning of emails that have the capability to precise the comment by extracting the useful information present in the sentence. The performance of the algorithms will be analyzed based on parameters like accuracy, precision, and recall. There will also be a case study to compare the training time of the different ML algorithms.

The remaining sections of the research article are as follows: Section 2 discusses similar work, whereas Section 3 describes the research technique. The design specification is discussed in Section 4, and execution is discussed in Section 5. Section 6 deals with the results evaluation, whereas Section 7 deals with the conclusion and future work.

## 2. RELATED WORK

Machine learning techniques have been utilized in various research of the prediction as well as prevention of Spam emails. These studies accomplished great outcomes as well. This section presents in detail, previous investigations and studies carried out to predict spam emails.
The review is divided into the following subsections. 2.1 Spam filtering and 2.2 ensemble classification and model running time

### 2.1 Spam filtering

Asif Karim, et.al (2019) presented a comprehensive analysis in which several observations were studied under ML based proposition [3]. The study suggested that the supervised approaches were proved more suitable and provided superior consistency in the performance of the model. It was analyzed that some classification techniques namely Support Vector machine (SVM) were more utilized. The single algorithm anti-spam systems were extensively utilized. Therefore, the strength of research was proved efficient into hybrid and multi-algorithm systems. Moreover, there was a necessity of more inventive technique which would consider diverse directions of the problem.

Akash Iyengar, et.al (2017) discussed that an integrated approach was implemented to detect the spam emails [4]. Every email provider had its own filtering methodology however, many methods from these were unable to work with their full potential. In some instances, an email

available in native language or any other language than English was considered authentic. Thus, the major intent was that these issues were tackled with the deployment of presented approach for enhancing the efficacy to delete and filter the emails. The accuracy obtained from the integrated approach was maximized up to 98.1% on real-time dataset as compared to the classic technique. This approach assisted the interns in avoiding the spam mails. The future work would aim at enlarging the precision while detecting the spam for the users in dynamic manner for which the presented technique would be combined with URL detection model.

Semih Ergin, et.al (2014) designed a solution to tackle the issue of Turkish spam e-mail for which the special attributes of Turkish e-mails were considered [39]. There were 4 phases included in the designed model in which morphological decomposition was performed; features were selected, training and testing had also carried out. The attributes of an email were extracted using fixed prefix stemming scheme. Afterward, the features were selected with the employment of MI methodology. The Decision Tree (DT) and ANN classification algorithms were implemented, and their accuracies were found adequate. The Artificial Neural Network (ANN) provided the precision of 91.08% and Decision Tree (DT) offered the accuracy of 87.67% in case of selection of dimensions of feature vectors.

Bilge Kagan Dedeturk, et.al (2020) recommended a new spam detection technique in which the ABC algorithm was deployed along with a LR classifier [6]. Three public datasets were employed to carry the evaluation. The outcomes exhibited that the recommended technique could handle the high-dimensional data using its highly effectual local and global search potential. The performance of recommended technique for detecting the spam was compared with SVM, LR, NB and existing techniques. The recommended approach was performed better than other methods based on accuracy obtained in classification.

### 2.2 Ensemble classification and model running time

ShubhangiSuryawanshi, et.al (2019) emphasized on implementing various ML classification algorithms such as NB, SVM, Adaboost and Ensemble classification models with a voting method [7]. The dataset taken from UCI and Kaggle was employed to test and compute these models regarding several parameters such as accuracy, recall and ROC. The outcomes indicated that all the suggested classification models were performed according to the number of attributes and size of the dataset. The initial outcomes validated that the Ensemble classification algorithm with voting method was proved the most appropriate as least FPR and greater precision was obtained from it. In the future, ensemble classifier would be put together with Drift Detection method for dealing with the issue of drift issue while filtering the email spam.

Shrawan Kumar Trivedi, et.al (2016) aimed at developing a spam classification framework with or without implementing ensemble of classifiers [8]. An effective and sensitive classifier was generated for differentiating the spam emails from the ham emails and a fine precision was acquired with least PR. The informative attributes of the Enron dataset were searched to integrate the Greedy Stepwise feature search technique. Several ML classification algorithms

namely Bayesian, NB, SVM, J48, Bayesian with Adaboost were compared. The parameters such as FPR etc. were considered for testing and computing these classification algorithms. All these aspects were analyzed in their entirety. This indicated that the Support Vector machine (SVM) was the most suitable classification algorithm as a superior accuracy and least FPR was obtained from it.

Simranjit Kaur Tuteja, et.al (2016) aimed at implementing the Back Propagation Neural Network (BPNN) filtering algorithm that was an ANN-FF with BP planned based on text classification for classifying the imperative emails from unsolicited emails [9]. There was necessity of all the background for executing this algorithm. The training time was found superior in comparison with other classification algorithms. The future work would focus on utilizing the presented algorithm along with K-Means in the pre-processing phase to maximize the efficacy while detecting and filtering the email spam and phishing.

S. Nandhini, et.al (2020) proposed an effective approach in which some algorithms were deployed to construct a ML model which had potential to classify the email as spam or ham [10]. The experiments were conducted on the UCI data set. The constructed ML model was trained and set up by quantifying the performance of 5 ML classifiers such as LR, DT, NB, KNN and SVM. The dataset was trained and tested applying Weka tool. This evaluation exhibited that the Random Tree (RT) performed better as compared to other classifiers concerning various parameters. The K-Nearest Neighbor provided the similar outcome, but it consumed much time for constructing the ML model in contrast to RT.

DewiWardani, et.al (2018) described that the spam mail had disturbed the email service [11]. These spam mails contained the contents which lead to create trouble before the users regarding electronic news, advertisements, and other things. Thus, a major purpose was to design a general model to detect the spam. The URLs that were listed within the email were traced to extract the standard information. Afterward, the metadata keywords were gathered from the page that had been later pointed by URLs. A notification was sent for dumping the email into a spam folder in automatic manner at the time of detection of spam after calculating the similarity. The average time was counted 5.15 seconds and the running time was 0.15 seconds for the URLs existed in the database.

Aisha Zaid, et.al (2016) investigated that the privacy and security of great amount of sensitive data about staff and students was threatened due to the malicious spam in case of educational institutes [12]. This study focused on enhancing the detection of malicious spam by selecting the attributes in the educational sector. A framework was established in which a new dataset was applied to carry out feature selection that was a step to enhance the classification in later phase. Afterward, 3 classification algorithms namely NB, SVM and MLP were employed for authenticating the chosen attributes while detecting the email spam. This data set was proved unique because there was any research suggested in the literature for detecting the malicious spam in a particular domain such as the educational field. The integrated technique was utilized against the existing approach and the accuracy was enhanced up to 97.3% on the real time

dataset. The feature selection process assisted in enhancing the training time and accuracy while detecting the malicious spam.

Despite the observations from the state of the art, past solutions are still insufficient, which is why I am conducting this research. More research is needed to assess the accuracy, precision, recall, f1score, as well as the training and testing time of the specified classifiers- decision tree, naive bayes, SVM (support vector classifier), and their ensemble (voting classifier) using my proposed unique dataset that has not been used in any of the previous studies.

## 3. METHODOLOGY

To collect meaningful information from data in this study, we used the knowledge discovery database methodology. Data selection, data pre-processing, data transformation, data mining, and data evaluation are some of the procedures involved.[13] Every phase has its own set of operations. The KDD method is illustrated in the diagram below. In architecture, the remaining steps are considered.

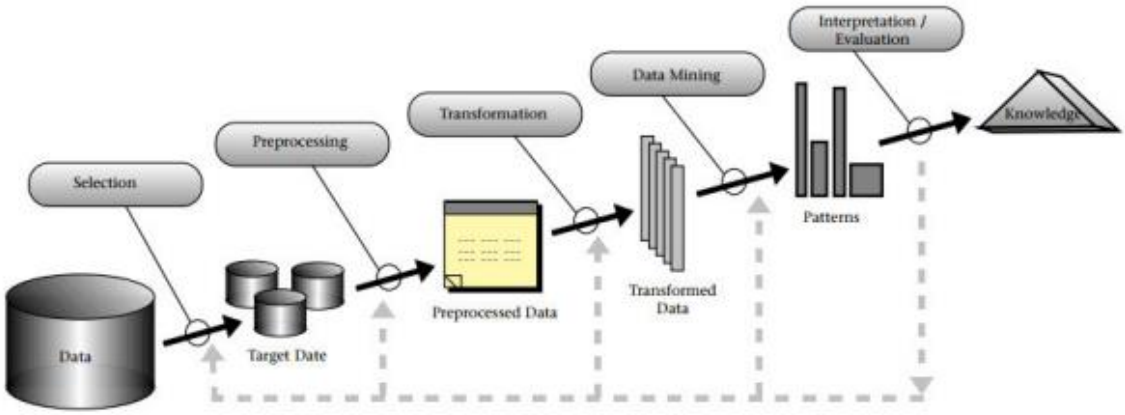



Fig 1. Knowledge Discovery in Databases (KDD) methodology

**Data Collection:**

The collection of data is done from microblogging sites by using google API and store data as a.csv file. The dataset contains 5171 samples, 3672 non-spam messages called "ham", and 1499 messages labeled as "spam". The data is restricted to a single column, categorizing the email between "spam" or "ham"

**Data Pre-processing**:

In Preprocessing, the following tasks are performed.

Removing Punctuation: The 14 punctuation marks such as question marks, exclamation mark, comma etc. are removed.

Stop word Removal: The stopwords such as the, have, etc. are removed since they don't have any meaning.

Tokenization: This divides the incoming mail into a series of demonstrative symbols termed as tokens. These demonstrative symbols are taken out from the structure of email, the header, and subject.
Stemming: The words are reduced to the root words.

**Data Transformation**:
The algorithms used in machine learning purposes are generally based upon statistics and computations. The raw email messages are now converted to the vectors for analysis and processing. For this research work, NLP toolkit is used for Text-based Feature vector extraction from the collected dataset.

**Data Mining:**
Here, the matrix tokens are fed to the classification models for implementation. The dataset will be classified into three branches- positive, negative, and neutral. The different existing types of machine learning models I considered for this research experiment includes support vector machines, decision tree classifier, naïve bayes classifier and the ensemble classification (voting classifier). I have used 80% of the dataset for training, whereas 20% of the data was used for testing. The following data mining techniques have been employed for this research work.

**Support Vector Machine (SVM) classifier:** Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a certain coordinate in the SVM algorithm. [14] Then we accomplish classification by locating the hyper-plane that clearly distinguishes the two classes. The primary goal is to improve generalization ability.

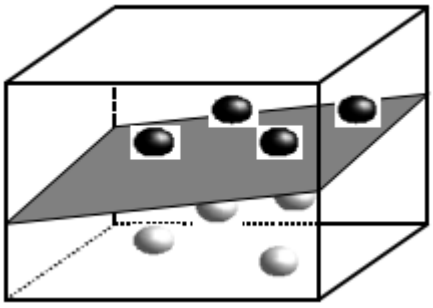


Fig. 1.1 An SVM dividing black and white points in 3 dimensions.

**Decision Tree Classifier:** In both regression and classification applications, this is often preferred for overcoming classification issues. It's a tree classification in which the core nodes indicate the properties of a data collection, the branches represent the rules of selection, and each leaf node represents the outcome. [15] [18] The Decision node and the Leaf node

appear to be the only two nodes in a decision tree. Decision nodes are used to make decisions and have many branches, whereas leaf nodes are the outcome and have no branches.

**Naïve Bayes:** The supervised learning method of nave Bayes is based on the Bayes theorem and is used to solve classification problems. For mail categorization, it's usually combined with a large training data set. The Bayesian Classifier is a basic classification technique that aids in the creation of machine prediction models. [16] It's a probabilistic classifier, which means it predicts an item's chances of happening.

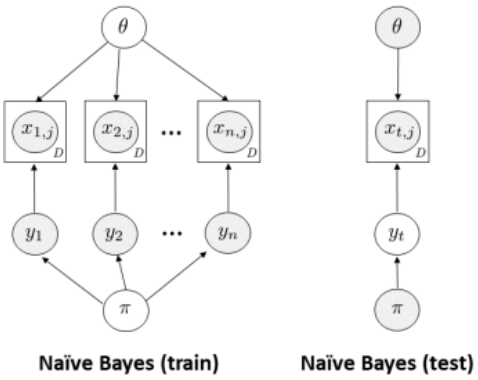



Fig. 1.2 Naïve Bayes

The final mining method with regards to this research is the **Voting classifier** which I will discuss in detail in the next chapter.[17]

**Interpretation and Evaluation**

To find the best classifier for the research purpose, I performed empirical analysis and calculated various performance parameters like those in the related works [12] which includes accuracy, training and testing time, precision, recall, and f1 score. The higher the score, better the performance of the model.

## 4. DESIGN SPECIFICATION

In this section, the architectural design along with the overall process of the progression for this research has been presented. In the first stage data gathered by the sources. In the next phase, data pre-processing is carried out, including various tasks such as removing stop word, punctuation marks, stemming and then tokenization. Later, the pre-processed data is used to extract the features. Next, we split the data into two categories training and testing with the ratio 80:20 respectively. The next step is about training the machine learning model using existing ML algorithms and the proposed ML approach in this work.

Moreover, the voting classifier is my proposed algorithm. To establish an ensemble, I simply trained the other three models (Decision tree classifier, Nave Bayes, and SVM), then integrated them to create an ensemble of models with a simple per-class voting system that is

then used to classify the email sample into ham or spam classes. Finally, the performance of the model has been tested and analyzed using the test data.

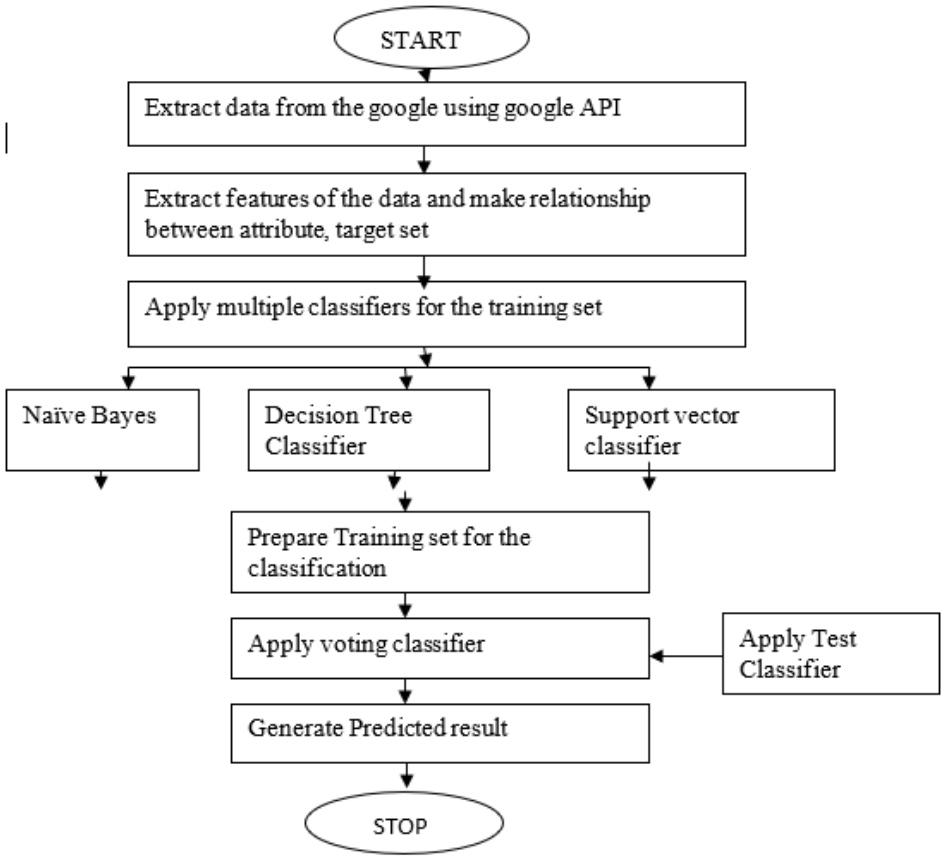



**Fig. 2. Block diagram of proposed research.**

## 5.0 Implementation

This section delves into the specifics of how the study was carried out. The Python 3.8.5 version was used in the implementation. Because of its machine learning capabilities and ease of use, Python is the most popular programming language. Python also comes with a variety of packages for machine learning and natural language processing. For machine learning development, we largely used Nltk, pandas, and other necessary libraries.

The data for this study came from www.kaggle.com. It's basically a .csv file that collects data from microblogging sites using the Google API. There are a total of 5171 email messages, with 3672 "ham" emails and 1499 "spam" messages. The data has been separated into two sets: 80-20 training and testing. An imbalanced data set was used in this experiment. It means that there aren't equal numbers of spam and ham messages in the training data set.

To begin, we uploaded the dataset file using Python's Pandas package, which allows us to read data from.csv files. The NLTK library was used because it provides categorization,

tokenization, stemming, named entity extraction, and parsing features. During data processing, messages are divided into discrete words and tokenized.
The tokenized data is turned into vectors because we need to translate text files into numerical feature vectors to run machine learning algorithms. During the vectorization process, BoW transforms vectors, counts the number of times a word appears in each message, weights vectors, and normalizes them.
In this experiment, the train_test_split and KFold functions from the Python Sci-kit learn package were used to split the data into training and test datasets. We chose an 80:20 ratio for database selection. The training data set makes up 80% of the total, whereas the testing data set makes up 20%.
Machine learning models are fed the pre-processed data, which forecast the outcomes. The outputs of the model are expressed as evaluation metrics, such as accuracy, F1 score, recall, precision, running time, and so on, with accuracy being the most significant parameter in this study.
The table below shows the accuracy of various machine learning classifiers.

| **Model** | **Accuracy** |
|---|---|
| Decision tree | 94.97% |
| Naïve bayes | 95.45% |
| SVM | 96.52% |
| Ensemble classifier | 97.97% |

Table 1. Accuracy table

In addition, a comparison table was produced and used to evaluate the proposed methodology' outcomes. Finally, using the matplotlib library in conjunction with the seaborn library, correlation tests were performed. The seaborn library is used to make a graph that compares the results of several models including SVM, Decision tree, naive bayes, and voting classifier.

## 6. EVALUATION

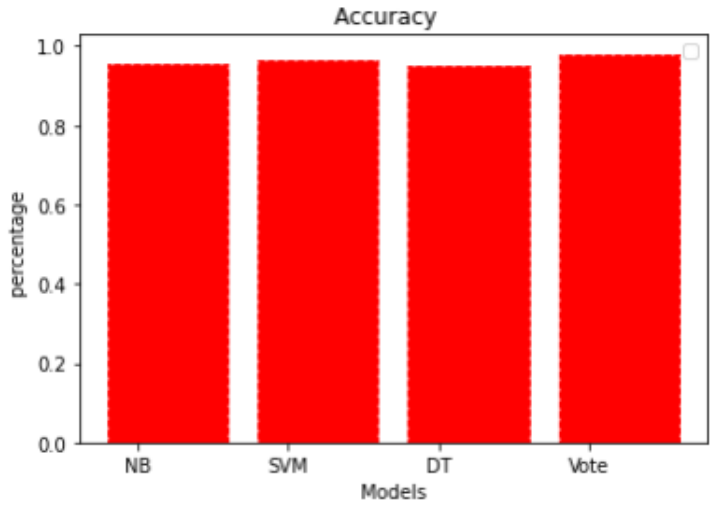
In this section, there will be 2 case studies. In the first case, the results of the different algorithms will be analyzed based on the performance metrics which are shown in table 1.1, whereas in the second case, the training and testing time of different algorithms will be compared.

Table 1.1 **Performance Parameter**

| | |
|---|---|
| False Positive Rate (FPR) | FPR = FP/(FP+TN) |
| False Negative Rate (FNR) | FNR = FN/(TP+FN) |
| True Positive Rate (TPR) | TPR = TP/(TP+TN) |
| True Negative Rate (TNR) | TNR = TN/(TN+FP) |
| Accuracy | (TP+TN)/(TP+TN+FN+FP) |
| Precision | TP/TP+FP |
| Recall | TP/TP+FN |
| F-Measure | 2*Recall*Precision/Recall + Precision |

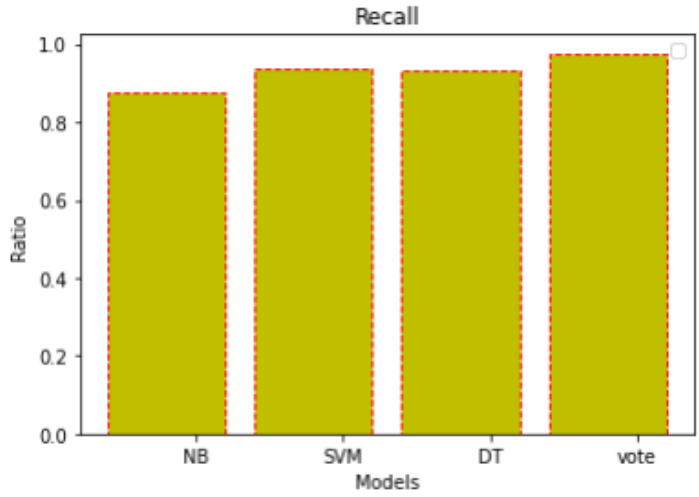
- True Positive (TR) is described as the amount of positive tuples that are accurately labeled by the classifier.
- True Negative (TN) is defined as the amount of negative tuples that are accurately labeled by the classification model.
- False Positive (FP) is defined as the negative tuples that are wrongly labeled as positive.
- False Negative is defined as the positive tuples that are wrongly labeled as negative.

**CASE STUDY 1. PERFORMANCE COMPARISON**



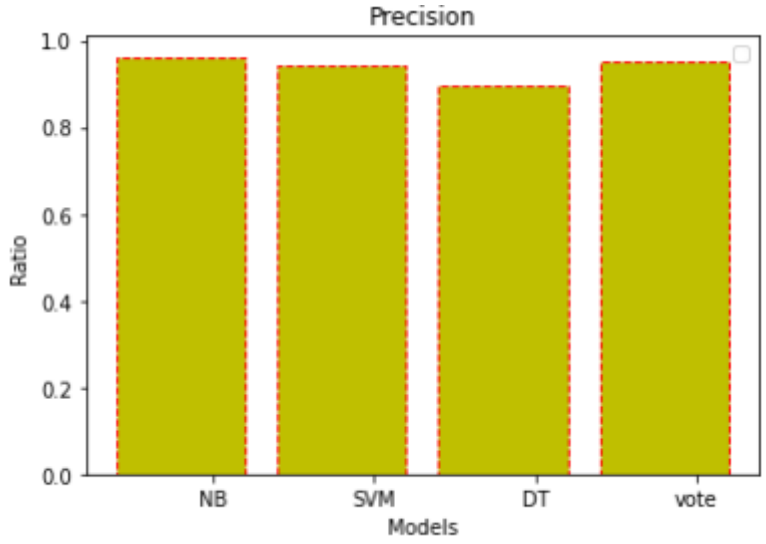

- **Figure 3.1 Accuracy Analysis**

As shown in Figure 3.1, the accuracy of various classifiers like NB, SVM, Decision and voting classifiers are compared for the email spam detection. The voting classifier has maximum accuracy which is 97.97% percent as compared to other classifiers.



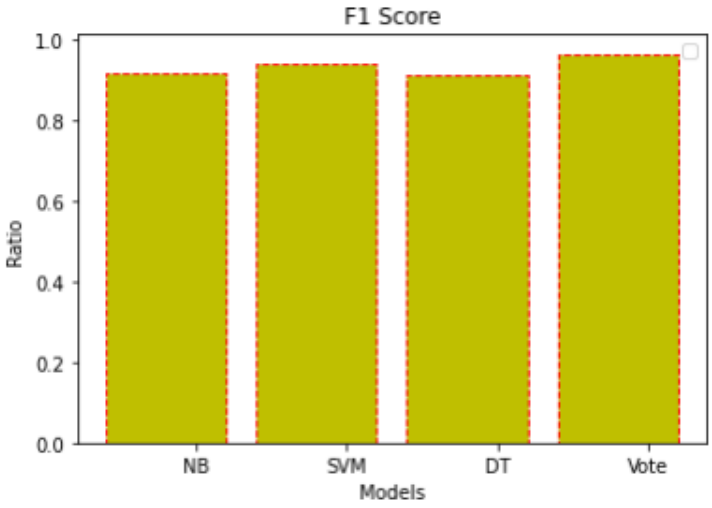

- **Figure 3.2 Recall Analysis**

As shown in figure 3.2, the recall of various classifiers is compared for the email spam detection. The voting classifier has maximum recall value which is 0.97 as compared to other classifiers




- **Figure 3.3 Precision Analysis**

As shown in figure 3.3, the precision of various classifiers is compared for the email spam detection. The Naïve bayes classifier has maximum precision value which is 0.96 as compared to other classifiers. However, the voting classifier comes second best at 0.95.

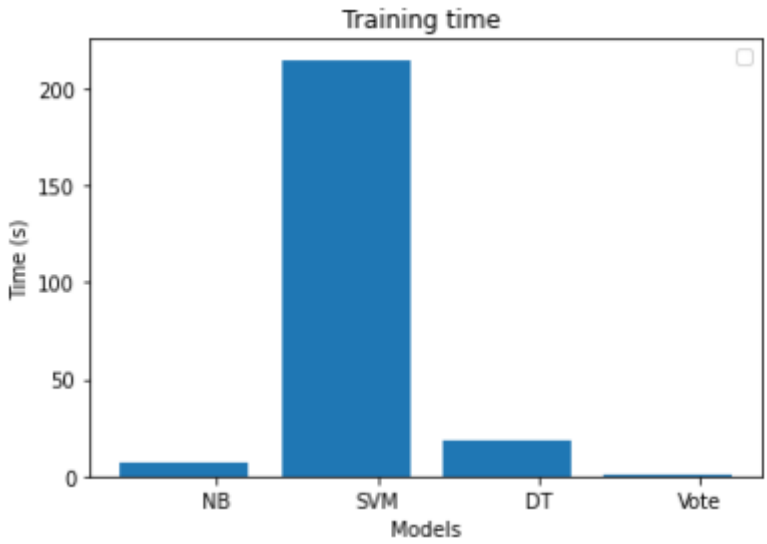



- **Figure 3.4. F1-Score Analysis**

As shown in figure 3.4, the F1-Score of various classifiers is compared for the email spam detection. The voting classifier has maximum f1 score which is 0.96 as compared to other classifiers.

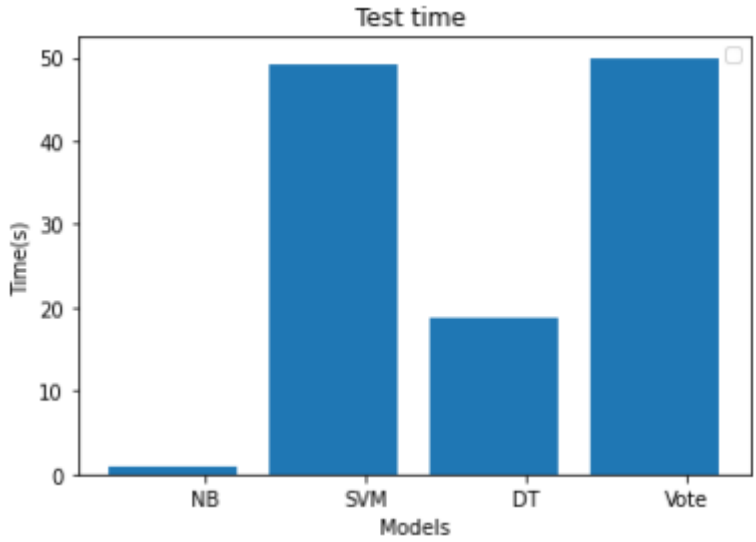
**CASE STUDY 2. RUNNING TIME COMPARISON**

The entire time it takes to develop the training model and the total time it takes to get the results based on the test data are the two components of time performance. Figure 4.10 and Fig 4.11 are graphical representations for the training and testing time of the evaluated methods.




- **Figure 3.5. Training Time comparison**

As shown in figure 3.5, the training time of all the models is compared and it is analyzed that proposed model has least training time




- **Figure 3.6. Testing Time comparison**

As shown in figure 3.6, the testing time of all the models is compared and it is analyzed that the proposed model has greatest testing time at 49.9seconds, with SVC being a near second among the others.

- **Table 2. Comparison Analysis**

| **Parameter** | **Naïve Bayes** | **SVM** | **Decision Tree** | **Voting** |
|---|---|---|---|---|
| Accuracy | 95.45 Percent | 96.52 Percent | 94.97 percent | 97.97 percent |
| Recall | 0.87 | 0.93 | 0.93 | 0.97 |
| Precision | 0.96 | 0.94 | 0.89 | 0.95 |
| F1 Score | 0.91 | 0.93 | 0.91 | 0.96 |
| Training Time | 6.7 seconds | 214.9 seconds | 18.7 seconds | 0.4 second |
| Testing Time | 0.9 second | 49.0 seconds | 0.3 second | 49.9 seconds |

**Table 2** shows the metrics scores given by the 4 different algorithms that are used during the testing phase. It can be **concluded** that the best scores belong to the voting classifier. It provides the best accuracy of 97.97%. Although, it takes the least time for training, the testing time is not so encouraging at 49.9seconds. In sharp contrast, it is a close call to determine which of the other models has the lowest performance, but it does appear from the experiment that SVM is

the second best in terms of performance scores across the evaluation metrics, however the huge disadvantage of its training and testing time at 214.9 seconds and 49seconds respectively does not seem promising. Further, the ensemble classifier fared better across other performance metrics such as recall and f1 score compared to other models (NB, SVM, DT) [19].

### 6.2 Discussion

After experimental implementation and evaluation between multiple algorithms, we can conclude that the proposed model (Voting classifier) gave a higher accuracy of 97.97% along with achieving higher recall and f1 score among all other classifiers evaluated. Although, there are several studies [4] where researchers claim to achieve more than 98% of accuracy. The same can also be said for the F-1 score metrics, where values closer to 1.00 are better. However, my proposed model (voting classifier) gave a relatively high F1 score of 0.96 among all other evaluated classifiers in this experiment.

Also, time performance is strongly influenced by secondary factors such as hardware platforms, software libraries (whether optimized or not), and the quality of one's profiler, rather than the algorithm itself, implying that time differences between the examined algorithms can often be considered negligible. [20]

To highlight the **novelty aspect** of this study, I used a comprehensive novel dataset to analyze the accuracy, precision, recall, f1score, as well as training and testing time of the specified classifiers—decision tree, naive bayes, SVM (support vector classifier), and the voting classifier. Additionally, the results acquired for this experiment are unique, implying that mixing numerous models at once (ensemble) for email spam prediction, produces better and more accurate results than a single ML model, thus, fulfilling my research questions.

Nevertheless, the research is still limited, as evidenced by the following points:

- Our model correctly classified the email spam, however it is still susceptible to error, as its accuracy is only 97.97 percent. However, if I wanted to achieve better results, the grid search tool could have been utilized more to achieve optimal performance. On a different set of data, results would have varied, therefore I would recommend using more training data.
- This model's intelligence would be limited, and it would need to be paired with another system to be genuinely useful.

## 7. CONCLUSION AND FUTURE WORK

### 7.1. Conclusion

Online users now have a platform thanks to the introduction of social media. This platform allows users to express and share their thoughts and ideas on a variety of topics and situations. There are about 300 million active users on this platform. In a single day, these users send almost 500 million messages. This platform has become extremely popular. The voting classifier is developed in this study for email spam detection. For assessing email spam detection, the voting classifier combines the three classifiers. The nave bayes, decision tree, and SVM are the three methods. The data is trained using a combination of these classifiers. The voting classifier takes an input test set and predicts a result in one of three categories: positive, negative, or neutral. Python is used to implement the recommended and available techniques. The voting classifier has a maximum accuracy of 97.97 percent, which is higher than the other SA classification models. It also has a better training time as opposed to the other discussed models.

### 7.2. Future Work

Considering the vast areas in cyber security in need of more research study, this study is only restricted for the email spam detection. The proposed work can be further developed with a hybrid or other more complex data mining algorithms. [21] Machine learning filtering for non-English spam words presents another opportunity for further research.in the future.

## References

[1] A. Lakshmanarao, K. Chandra Sekhar, Y. Swath, “An Efficient Spam Classification System Using Ensemble Machine Learning Algorithm”, 2018, Journal of Applied Science and Computations, Volume 5, Issue 9

[2] Apurva Taunk, Srishty Bharti, Sipra Sahoo, “An Ensemble Method for Spam Classification”, 2020, International Journalof Scientific& Technology Research Volume 9, Issue 02

[3] Asif Karim, Sami Azam, Bharanidharan Shanmugam, Krishnan Kannoorpatti, MamounAlazab, “A Comprehensive Survey for Intelligent Spam Email Detection”, 2019, IEEE Access

[4] Akash Iyengar, G. Kalpana, S. Kalyankumar, S. GunaNandhini, “Integrated SPAM detection for multilingual emails”, 2017, International Conference on Information Communication and Embedded Systems (ICICES)

[5] Semih Ergin, SahinIsik, “The investigation on the effect of feature vector dimension for spam email detection with a new framework”, 2014, 9th Iberian Conference on Information Systems and Technologies (CISTI)

[6] Bilge Kagan Dedeturk, BahriyeAkay, “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm”, 2020, Applied Soft Computing

[7] ShubhangiSuryawanshi, Anurag Goswami, Pramod Patil, “Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers”, 2019, IEEE 9th International Conference on Advanced Computing (IACC)

[8] Shrawan Kumar Trivedi, “A study of machine learning classifiers for spam detection”, 2016, 4th International Symposium on Computational and Business Intelligence (ISCBI)

[9] Simranjit Kaur Tuteja, NagarajuBogiri, “Email Spam filtering using BPNN classification algorithm”, 2016, International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)

[10] S. Nandhini, JeenMarseline K.S., “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection”, 2020, International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)

[11] DewiWardani, RetisaSiwi, Bambang Harjito, MaysaMarshallia, “Using Metadata in Detection Spam Email with Pornography Content”, 2018, International Conference on Electrical Engineering and Computer Science (ICECOS)

[12] Aisha Zaid, Ja'farAlqatawna, Ammar Huneiti, “A Proposed Model for Malicious Spam Detection in Email Systems of Educational Institutes”, 2016, Cybersecurity and Cyberforensics Conference (CCC)

[13] M. Chehad, “Knowledge Discovery Data (KDD)” [Online]. Available: https://medium.com/analytics-vidhya/knowledge-discovery-data-kdd-a8b41509bff9

[14] V Vishagini, Archana K Rajan, “An Improved Spam Detection Method with Weighted Support Vector Machine”, 2018, International Conference on Data Science and Engineering (ICDSE)

[15] Adi Wijaya, AchmadBisri, “Hybrid decision tree and logistic regression classifier for email spam detection”, 2016, 8th International Conference on Information Technology and Electrical Engineering (ICITEE)

[16] G.Vijayasekaran, S.Ros, “Spam and Email Detection in Big data Platform using Naives Bayesian classifier”, 2018, International Journal of Computer Science and Mobile Computing, Vol.7 Issue.4, pg. 53-58

[17] Megha Rathi, Vikas Pareek, “Spam Mail Detection through Data Mining – A Comparative Performance Analysis”, 2013, International Journal of Modern Education and Computer Science, Volume 12, PP. 31-39

[18] Ahmed I. Taloba, Safaa S. I. Ismail, “An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection”, 2019, Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)

[19] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Opeyemi Emmanuel Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems”, 2019, Heliyon

[20] YuktiKesharwani, Shrikant Lade, “Spam Mail Filtering Through Data Mining Approach –A Comparative Performance Analysis”, 2013, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 9

[21] Samira. Douzi, Feda A. AlShahwan, Mouad. Lemoudden, and Bouabid. El Ouahid, “Hybrid Email Spam Detection Model Using Artificial Intelligence”, 2020, International Journal of Machine Learning and Computing, Vol. 10, No. 2