Email Spam Detection: An Empirical Study of Different ML Algorithms

Sowjanya Sunkavalli
Dept. of Computer Science and Cyber Security
University of Central Missouri, MO
Warrensburg,MO, USA
sxs18960@ucmo.edu

Sathya Sai Satish Kothamasu
Dept. of Computer Science and Cyber Security
University of Central Missouri, MO
Warrensburg,MO, USA
sxk14490@ucmo.edu

Samskruthi Velamakanni
Dept. of Computer Science and Cyber Security
University of Central Missouri, MO
Warrensburg, MO, USA
sxv13090@ucmo.edu

Praveen Gyarala
Dept. of Computer Science and Cyber Security
University of Central Missouri,MO
Warrensburg, MO, USA
pxg10030@ucmo.edu

Abstract—The most popular form of official communication for company is email. Despite other communication channels, email usage keeps growing. In today's environment, when email volume is increasing daily, automated email management is crucial. More than 55% of the emails overall are classified as spam. This demonstrates how these spams waste email users' time and resources while producing nothing helpful.

With the rapid development in internet users, email spam has grown to be a significant issue in today's society. For fraud, phishing, and other immoral or criminal activities, people use them. Spam emails that contain harmful links that can damage our system and try to access your system. For spammers, it is very simple to set up a phony profile and email address. They then appear to be a real person in their spam emails and prey on those who are unaware of these scams.

When utilizing spam emails to carry out their illegal actions, spammers use sophisticated and inventive techniques, Therefore, it is necessary to recognize spam emails that contain fraud. This project will recognize spam emails that contain fraud by utilizing machine learning techniques.

Therefore, it's critical to comprehend the various spam email classification methods and how they work. This study primarily focuses on the spam classification method utilizing machine learning techniques. Additionally, this study offers a thorough overview and evaluation of previous research on various machine learning methodologies and email characteristics. Models are constructed for email spam detection and classification in the current research. Different machine learning classifiers, such as Naive Bayes, SVM, Decision Tree, KNN, Bagging and Boosting (Adaboost), and Ensemble Classifiers, have been used in this study. On the email spam dataset from the Kaggle website, classifiers are evaluated and tested. There are several accuracy metrics utilized, including Accuracy Score, F measure, Recall, Precision, Support, and ROC. And with the highest accuracy scores and other important factors, we have found the best machine learning classification approach.

Keywords— Machine learning Classifiers, E-mail, Spam Filter, Supervised Learning, Email Classification, Machine Learning Algorithms, Accuracy Measure

I. INTRODUCTION

For the majority of internet users, email has taken over as the most popular method of formal communication. Spam problems have gotten worse in recent years as a result of more people using email.

Bulk unsolicited message distribution is referred to as spam or junk email . Emails that are meaningful but of the opposite sort are referred to be "Ham." A typical email subscriber receives between 40 and 50 emails daily.

Every year, spammers lose money to both personal and institutional fronts, earning about 3.5 million USD through spam [4]. Users waste a lot of their working time responding to these emails as a result. Spam transmits a significant volume of unwanted and unsolicited bulk emails, accounting for more than 50% of email server traffic, according to.

They impair productivity by consuming user resources with no helpful results. The spam that spammers send out has the intention of spreading malicious criminal actions like identity theft, financial disruptions, stealing sensitive information, and reputational harm for marketing goals.

These factors make email management, including the classification of spam emails, a crucial requirement for businesses looking to boost productivity and cut costs.

Git Hub Link for Project: https://github.com/Satish-10/MLFinalProject

A. Relevant Spam Statistics: We will focus on some global statistics data on spam vs. financial impact in the subsection that follows. In the analysis, some metrics for Australia that are country-specific are also covered. A little over 4 billion email accounts are actively used as of 2020, according to [5], and this figure is expected to increase to 4.48 billion by the year 2024.

This indicates that by the year 2020, email will be actively used by over half of the world's population. Based on this, spam represents 57.26% of all email traffic in 2019 [5]. This demonstrates that more than half of all emails sent globally are unwelcome, uninvited spam emails. According to a recent FBI report, businesses worldwide suffered a financial loss of $12.5 billion in 2019 as a result of compromised business email accounts and spam and phishing attacks [6]. The financial losses suffered by organizations are anticipated to soar in the approaching years as email usage continues to expand at an exponential rate. Following are digital fraud data for the prior years for the Australian context [7] (Australian Competition and Consumer Commission, 2019). This data demonstrates that these online scams are a worldwide concern, and Australia is affected significantly as a result.

Aside from that, statistics shows that both the losses and the number of cases for each year are trending rising. Investment fraud solicits money from investors in exchange for enticing but false business prospects. Scams involving dating target those who use the internet to hunt for romantic relationships. Scammers' preferred method for disseminating malware and other fraudulent frauds is via email. Due of the spammers' creativity and problem-solving abilities, the remedies currently being deployed are frequently inadequate. These factors highlight the need of comprehending and creating systems that can distinguish between spam and legitimate commercial communications.

TABLE I
TOP THREE DIGITAL SCAM LOSSES(AU$) AND FREQUENCY (NUMBER OF CASES) IN AUSTRALIA FROM 2017 – 2020. [1]

Year	Total Loss	Digital Scam Amount	Digital Scam Frequency
2017	$90801407	Investment Scams $31,326,476 Dating and Romance $20,530,578 Other business & employment $ 5,270,948	Phishing 26,386 Identity Theft 15,703 False billing 13,455
2018	$107001471	Investment Scams $38 846 635 Dating and Romance $24 648 024 False Billing $ 5 512 502	Phishing 24,291 Threats to Life 19,455 Identity Theft 12,800
2019	$142898217	Investment Scams $61,813,801 Dating and Romance $28,606,215 False Billing $ 10,110,753	Phishing 25,170 Threats to Life 13,375 Identity Theft 11,373
2020	$52971358	Investment Scams $20,650,486 Dating and Romance $14,708,686 False Billing $ 4,378,559	Phishing 10,689 Threats to Life 4 255 Identity Theft 4,237

Fig 1. Table Data on Email Spam Statistics

B. Spam and Ham

Email spam, often known as electronic mail spam, is the practice of sending unwanted emails or commercial emails to a list of subscribers. Unsolicited emails signify that the recipient has not given consent to receive them. Since last decade, using spam emails has grown in popularity. Spam has grown to be a significant online problem. Spam wastes space, time, and message delivery. Although automatic email filtering may be the best way to stop spam, modern spammers may quickly get around all of these apps. Prior to a few years ago, the majority of spam that came from particular email addresses could be manually stopped.

Spam detection will use a machine learning approach. The major techniques used to filter junk mail include "text analysis, white and blacklists of domain names, and community-based strategies." A widely used technique to combat spam is text assessment of email content. There are numerous solutions that can be implemented in terms of the server and buyer. One of the most widely used algorithms in these techniques is naive Bayes. In the case of false positives, it may be challenging to reject sends that are largely based on content analysis. Regular customers and businesses wouldn't want any important communications to get lost.

The boycott strategy was most likely the first one used for the separation of spam. The strategy is to acknowledge every send, excluding those from local or electronic mail ids. declared a boycott. This method no longer functions as well as it formerly did since more modern regions entered the category of spamming domain names.
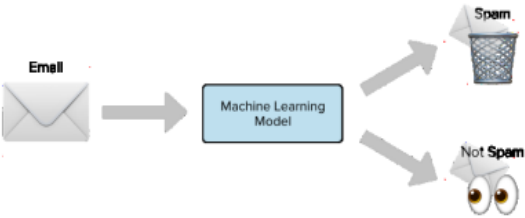

Fig 2. Spam and Not Spam Classification

The whitelist approach is the practice of accepting emails from domain names and addresses that have been publicly whitelisted while placing all other emails in a much lower priority queue. Emails are delivered most successfully after the sender acknowledges a confirmation request sent by the "junk mail filtering system."

The use of electronic messaging systems to transmit unsolicited bulk messages, including mass advertisements, harmful links, etc., is referred to as spam, according to Wikipedia. Unsolicited refers to communications from sources that you did not request. Therefore, if you don't recognize the sender, the email may be spam.

Usually, when downloading any free services, software, or while updating the program, people are unaware that they have just signed up for those mailers. Around 2001, Spam Bayes coined the phrase "Ham," which is defined as "Emails that are not commonly sought and are not labeled spam."

Machine learning techniques are more effective; these samples are a set of pre-classified emails that are used as training data. There are numerous methods available for email filtering using machine learning techniques. The following algorithms are among them: "Naive Bayes, support vector machines, neural networks, K-nearest neighbor, random forests, etc."

C. Different ML Techniques

Due to the recent technology's quick development, more people are using the internet and email to communicate, which brings up the problem of unwanted spam emails. Spam email causes several problems, including privacy concerns [1], clogs email systems [1], spreads viruses, and is linked to numerous harmful and commercial websites.

Spam email traffic that is expanding quickly uses up a lot of bandwidth and memory, takes up a lot of users' time, and costs money. Users must spend a significant amount of time screening and deleting junk email and getting rid of spam can be expensive [3]. Reading the text or subject line of a mail to determine if it is spam or not takes a lot of time and effort.

Therefore, a spam classifier is required to identify and appropriately classify spam email [1][2].

There is a need for a method that can identify spam emails by considering a set of features [2][7]. A significant number of both necessary and optional features are present in the spam sample. The best and most crucial features are chosen using the Feature Selection approach. The features are chosen using Bag of Words, Word Count, and Term Frequencies [3].

Support Vector Machine [14], Nave Bayes [10, 16], Boosting Classifier (Adaboost)[12], and Ensemble Classifier[2] are only a few of the classification algorithms used to categorize email spam. By creating hyperplanes using support vectors, several classification techniques, such as Support Vector Machine, are used in this study to differentiate different class situations [14] [15].

In Naive Bayes, classification is done by considering the probabilities of various classes [10]. In the Ensemble Classifier, the prediction is carried out by considering the voting process, which employs the two variables hard and soft. When voting is "hard," the class label is determined by the majority of classifier predictions; when voting is "soft," the class label is determined by averaging the classifier predictions [2].

An email spam dataset from the Kaggle website is used for the experiment. AUC curve, Accuracy Score, Precision and Recall, F-Measure, Support, and Confusion matrix are some of the different comparative measures that are used to compare the accuracy of all classifiers

II. MOTIVATION

With the rapid development of internet users, email spam has grown to be a significant issue in today's society. They are being used for fraud, phishing, and unlawful and immoral behavior. Infecting your system and ours both occurs when spam emails are used to distribute harmful links. For the spammers, it is very simple to set up a phony profile and email address. They target those people who are unaware of these scams by making themselves appear to be real persons in their spam emails. Spam emails must be found and categorized, which is a need. As a medium of communication, email has grown in importance. Around 196 billion emails are sent daily around the world, and there are projected to be 4.1 billion email accounts there in 2014. The main danger facing email users is spam.

In 2013, spam accounted for 69.6% of all email flows. Users may follow links in spam emails to phishing or malware-infected websites, which can access and impair the recipient's computer system. Additionally, some websites have the ability to collect private data. As a result of lost productivity, spam also costs companies $2000 per employee annually. Consequently, an efficient spam filtering technology makes a substantial contribution to both our society and the long-term viability of the internet. There are now several methods for detecting spam. These techniques include mass email detection, message heading scanning, blacklisting, greylisting, and content-based filtering.

Blacklisting is a method for identifying IP addresses that frequently send spam. Future email sent from the IP addresses on the list is denied, and these IP addresses are added to a domain name system-based blackhole list. Spammers are getting around these lists, though, by employing more IP addresses.

Another method of detecting bulk emails is to filter spam. To assess if an email is spam or not, this method looks at the total number of recipients. However, a lot of trustworthy emails can receive a lot of traffic. The ability to detect spam by scanning message headings is reliable. Spammers' programs create email subject lines. These headings occasionally have errors that prevent them from adhering to the rules for standard headings. If there are mistakes in these headings, the email is probably spam. However, spammers are becoming more adept at learning from their mistakes and do so less frequently.
Greylisting is a technique where the email is rejected, and the sender is informed of the problem. Spam software will disregard this and not resend the email, whereas people are more inclined to do so. This method is not optimal because it is irritating to people.

The efficiency of spam filtering techniques could be increased by combining them with current spam tactics. To identify whether an email is spam, content-based methods examine the email's content. Our project's objective was to examine machine learning techniques and assess how effective they would be as content-based spam filters.

III. MAIN CONTRIBUTIONS AND OBJECTIVES

- This paper offers a thorough analysis and overview of the research on various machine learning algorithms and email properties that have been employed in those approaches.

- Additionally, it outlines potential paths for future study as well as difficulties encountered in the field of spam classification.

- This study primarily focuses on the machine learning-based spam classification approach.

- The best algorithm for email spam detection is chosen in this

study based on its precision and accuracy after being applied to our data sets and discussed in detail.

- On the email spam dataset from the Kaggle website, classifiers are evaluated and tested. Numerous accuracy metrics are employed, including Accuracy Score, F measure, Recall, Precision, Support, and ROC.

IV. RELATED WORK

Since the usage of email has increased, there are severe problems that need to be addressed. Therefore, comparing various classification techniques is necessary because numerous studies have been done to discover the best and most appropriate classifier for spam detection. This section discusses various classification techniques.

"Ensemble learning for data stream analysis: A survey," by Jerzy Stefanowski and Joao Gama [11] They have employed a variety of classifiers in this paper, including Decision Trees, Naive Bayes, Ripper, SVM, and KNN. SVM outperformed other classifiers in terms of accuracy, scoring 98.3 out of 100[6][9]. The problems with notion drift are addressed. Target class is concept, while drift is the changing patterns in a data stream. This method enhances performance and even performs better with tiny datasets that include the spam email dataset.

A Bayesian technique to screening spam email by Sahami, Dumais, Heckerman, and Horvitz [10]. The bag of words is employed as a representation method in this study, which employs naive Bayesian analysis. The probability of classes is how the Naive Bayes classifier functions. In this study, a few numerical and a few non-numeric features from the dataset were used, and encouraging results were obtained after applying the Na ve Bayes method.

S. K. Trivedi and S. Dey published a paper titled "Interplay between Probabilistic Classifiers and Boosting Algorithms for Detecting Complex Unsolicited Emails" [12]. They have employed the Nave Bayes and Boosting algorithm in this study. The boosting technique increases the classifiers' prediction accuracy. Decision tree estimators and AdaBoost have been employed. When using Naive Bayes, a class chooses the features.

A Survey on Supervised Classification on Data Streams, by Vincent Lemaire, Christophe Salperwyck, and Alexis Bondu, Springer International Publishing Switzerland 2015 [13]. They employed a method that is memory-efficient in this paper. Local Data Analysis is carried by using this algorithm. It is carried out on devices with limited resources. The set of attribute values a1, a2, a3,... and a with class labels are contained in each record.

The Euclidean Formula is used to calculate the distance between recently saved entries. If the distance is below the cutoff, it will look up the class labels of the closest neighbor and assign the same class label to the new entries.

Future Generation Computer Systems Elsevier B.V., "Scalable real-time classification of data Streams with idea drift"[8]. This study use the Naive Bayes method.
Simple to use and incredibly powerful is Naive Bayes. Using Naive Bayes, you can build models very quickly and generate extremely accurate predictions.
Support vector machines for spam categorization by Vapnik and Drucker was published in IEEE Transactions on Neural Networks [14]. SVM is frequently used for text classification and produces encouraging outcomes. It uses a hyperplane to split the feature set. Compared to other classifiers, boosted decision trees and SVM provide improved accuracy. SVM training is completed more quickly. SVM testing takes time and is unreliable for small datasets.

Nien-Feng Li, Cheng-Chi Lee, Jyh-Jian, Sheu, Ko-Tsung Chu, For detecting idea drift in spam filtering, "an effective incremental learning mechanism"[4] has been developed. The emails in this study are categorized using a Decision Tree classifier. All qualities supported by decision trees are categorical. Using the maximal gain ratio, the Critical attribute is chosen. Accuracy Matrix is composed of precision, recall, and F measure.

We compared various classifiers to improve classification accuracy of spam emails as a result of the classifiers outlined in the background research.

There is some related research by A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab that uses machine learning techniques for email spam identification. [ii] They discuss a concentrated literature review of machine learning and Artificial Intelligence (AI) algorithms for email spam detection. The "picture and textual dataset for the e-mail spam identification with the use of multiple approaches" has been employed by K. Agarwal [3] and T. Kumar. Harisinghaney et al. (2014) [4] and Mohamad & Selamat (2015) [v] are two other studies that have done this. With experiments on a dataset, Harisinghaney et al. (2014) [iv] applied the KNN algorithm, Naive Bayes, and Reverse DBSCAN algorithm. OCR library" [iii] is used for text recognition, although it doesn't work well. The feature selection hybrid strategy of TF-IDF (Term Frequency Inverse Document Frequency) and Rough pure mathematics is used by Mohamad & Selamat (2015) [v].

Machine learning models are now being used by researchers to identify scam emails. Using six alternative machine learning algorithms—Naive Bayes (NB) classification, K-Nearest Neighbor (K-NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Artificial Immune System, and Rough Sets—the authors of the research [3] have conducted tests. The experiment's objective was to mimic human perception and recognition.

Training and filtering were the two processes offered by the concept of tokenization. They divided their method into four sections: performance evaluation, description of the feature,

classification of spam, and email pre-processing. It was determined that the maximum accuracy, precision, and recall was obtained by the Nave Bayes method.

A hybrid system using SVM-NB and another machine learning technique is described by Feng et al. [1]. They suggest using the SVM technique, creating a hyperplane between the specified dimensions, and reducing the training set by removing datapoints. The NB algorithm will then be applied to this set in order to estimate the likelihood of the result. Chinese text corpus was used in this experiment. When compared to NB and SVM alone, their new technique was applied successfully, and accuracy increased.

By testing with several classifiers including NB, SVM, KNN, Tree and Rule based algorithms, Mohammed et al[4] .'s goal was to identify the unwanted emails. They created a lexicon of spam and ham emails, which is then used to the training and testing data to remove them. They used the Email-1431 dataset and the Python computer language to carry out their investigation. They came to the conclusion that Support Vector Machine and NB were the best operational classifiers.

A hybrid-based approach is proposed by Wijaya and Bisri [5] that combines Decision Tree with Logistic Regression as well as a False Negative threshold. They were successful in boosting DT's efficiency. The outcomes were compared to earlier studies. The Spam Base dataset was used for the experiment. The accuracy of the suggested method was 91.67%.

V. PROPOSED FRAMEWORK

In the Proposed framework we will train Email Spam data set with different supervised and unsupervised machine learning models and identify the model which gives us the more accurate values in terms of classification of Spam Email by considering different model output parameters like Accuracy, Precision, Recall etc.,

CLASSIC CLASSIFIERS

By extracting the models representing significant data classes, classification is a type of data analysis. For example: "A loan application as dangerous or safe," a classifier or model is built to predict class labels. Data classification is a two-step process that involves learning (building a classification model) and categorization.

1.NAÏVE BAYES Theorem

Spam detection was accomplished in 1998 using a naive Bayes classifier.

For supervised learning, there is an algorithm called the Naive Bayes classifier. The Bayesian classifier uses dependent

events and calculates the likelihood that an event that already happened can predict an event that will happen in the future.

On the basis of the Bayes theorem, which presumes that features are independent of one another, naive Bayes was developed. The Naive Bayes classifier method can be used to categorize spam emails because word probability is the key factor at play. Any word that frequently appears in spam but not in ham indicates that an email is spam.

The Naive Bayes classifier algorithm is now considered to be the best method for email filtering. In order for the model to function well, it is very well trained using the Naive Bayes filter. The Naive Bayes algorithm always calculates the likelihood of each class, and the class with the highest likelihood is then selected as an output. A correct result is always produced by naive Bayes. It is applied in numerous areas, such as spam screening.

Bayes Theorem

E2 = E1 + P(E1/E2) P(E1)/P(E2) ---- (1)

P(E1) = prior probability

[Probability that event E1 occurs before event E2]

Posterior Probability = P(E1/E2) --- (2)

When event E2 occurs, [E1 event Probability] occurs.

The probability computation for a feature f1, f2, ... fn is for any class. A distribution is recommended because it has a probability of p(fi/c) that a feature belongs to a specific class.

This distribution performs well for text data since it is simple to count the number of times a word appears in a specific email.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

---(3)

$$P(B) = \sum_y P(B|A)P(A)$$

---(4)

2. SUPPORT VECTOR MACHINE

It's a method that is overseen. To classify data, SVM uses labeled information. The text categorization task lends itself best to SVM (Support Vector Machine). With the right optimization, SVM may be utilized for large datasets effectively. In comparison to other classification techniques, SVM is one of the discriminating methods and is more accurate.

Both positive and negative training sets are required for SVM. A negative training set is not used by any other classification technique. With large dimensional datasets, SVM performs effectively. To distinguish between positive and negative data values in multi-dimensional datasets, SVM uses the hyperplane[5].

It is determined which support vectors are nearest to the decision surface. If an email is not included in the support vectors used in spam detection, it is eliminated from the data and has no impact on the performance of the SVM classifier.
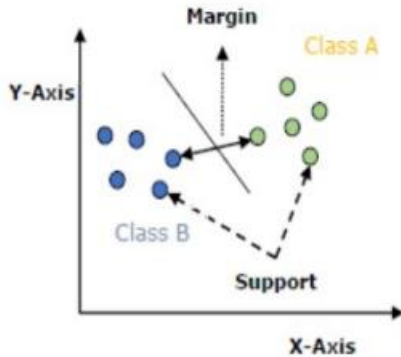
Fig. 3 SVM

3. DECISION TREE

Another algorithm that has been utilized more frequently in the research using the supervised learning technique that have been evaluated is the decision tree machine learning algorithm. The simplicity of this approach and its simpler explanations and visuals are the justifications for using it more frequently. Both large and small data sets can be used with this. Possesses the capacity to manage both the system's categories and numerical data [4]. They combined DT with other algorithms in the system they built in [16], along with other algorithms. In the tier three stage, DT has been applied to the binomial classification of spam and ham emails. As DT contains a feature that allows the model to categorize spam in real time, this feature offers important insights because DT has a simple computational mechanism, which is necessary for effective real-time computational requirements.

The process of learning a decision tree from class-labeled training tuples is known as "decision tree induction." A decision tree is built similarly to a flowchart.

Test on attribute for internal nodes or non-leaf nodes.

Branch = displays the test's results

A class label is held by a leaf node.

Root node is the top node.
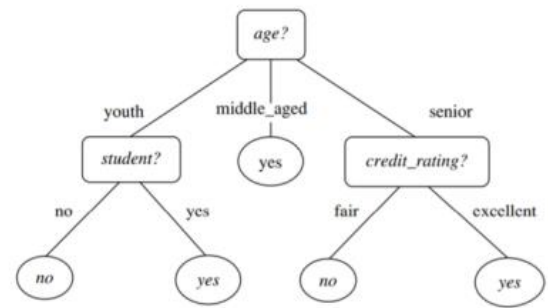
Fig. 4 Decision Tree

Decision tree Induction:

No "domain knowledge or parameter setting that is adequate for examining knowledge" is required for the construction of "decision tree classifiers." It manages information that is multifaceted. The decision tree induction process's learning and categorization phases are easy and quick. To select the feature that will best divide the tuple into different classes, characteristic choice events are used.

When a decision tree is created, a sizable portion of the branches may show disturbance and irregularities in the creation of the data. In order to increase the classifier's accuracy on obscure information, tree pruning tries to identify and remove such branches.

Entropy utilizing a single attribute's frequency table:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

----(5)

Entropy utilizing the two properties' frequency tables:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

----(6)

4. K- NEIGHBOR TO THE EAST

"The supervised classification method K-nearest neighbors. In order to forecast how a new sample point will be classified, this algorithm uses certain data points and a data vector that have been divided into a number of classes.

An inefficient algorithm is K- Nearest Neighbor. LAZY algorithms only attempt to memorize the steps that they cannot learn on their own. It doesn't make decisions on its own. A new point is classified using the K-Nearest Neighbor algorithm using a similarity metric, which can be Euclidian distance.

The neighbors of an object are determined by the Euclidean distance measurement.

$(x - a)2 + (y - b)2 = dist.((x, y), (a, b))$ ----(7)

5. ENSEMBLE LEARNING METHODS

In order to reduce variability by using bagging bias by using boosting predictions by stacking, ensemble approaches in machine learning use numerous base models to generate a predictive model. Two Sorts Base classifiers are constructed sequentially in this case. Base classifiers are running in parallel here.

BOOSTING AND ADABOOST CLASSIFIER

"Boosting is an ensemble method for creating a strong classifier from a collection of weak classifiers. Boosting is finished by creating a model from training data sets, then creating another model that corrects the original model's flaws." [8] Boosting Models are added until the training set is correctly predicted.

AdaBoost stands for Adaptive Boosting.

AdaBoost is the first successful boosting algorithm designed for binary classification. AdaBoost is used to understand boosting

VI. DATA DESCRIPTION

The Kaggle website is where the dataset for spam emails was obtained. The email files in this csv file were chosen at random, along with their labels for spam or not-spam classification. In the csv file, each email is represented by a single row. This table has 5172 rows and 3002 columns. The name of the email is shown in the first column. In order to preserve privacy, the name was supplied with numbers rather than the receivers' names. The labels for prediction are 1 for spam and 0 for not spam in the last column. After removing the non-alphabetical characters and words, the remaining 3000 columns show the 3000 most used words across all emails. The matching cells for each row contain the number of words in each column of the relevant email for that row. As a result, rather than having their contents saved separately in each one, all 5172 emails are combined into a single compact data frame.

According to Wikipedia, the use of electronic communications networks to disseminate large amounts of unsolicited messages, such as spammy links and advertisements, is referred to as "spam." Unsolicited communications are those that you receive from sources without asking for them. So, if you don't recognize the sender, the email can be spam. Most of the time, when users install any free services, software, or when they update a program, they are unaware that they have merely agreed to receive those mailers. In the early 2000s, Spam Bayes coined the word "Ham," which is described as "Emails that are not typically solicited and are not marked spam".

The dependent variable in email spam classification is a binary variable. For spam emails, the dependant variable has a value of 0, whereas for ham emails, it has a value of 1. In this experiment, "Prediction" is the dependent variable, and the keywords in the emails' messages are the independent variable.

VII. RESULTS/EXPERIMENTATION & COMPARISON/ANALYSIS

Algorithm

1.Insert the Dataset or file for training or testing

2.Check the dataset for supported encoding

3.Select any of the Scaler from the given list of

    1:StandardScaler(),2:Normalizer(),3:MinMaxScaler(),4:MaxAbsScaler(),5:RobustScaler()

4.After Selecting Scaler train the model with one of the given lists of classifiers.

    1:naive_bayes,2:svm,3:decision_tree,4:knn,5:boosting,6:adaboost,7:bagging,8:randomforest

5.After training the model with the required classifiers, check the comparison values for accuracies among them.
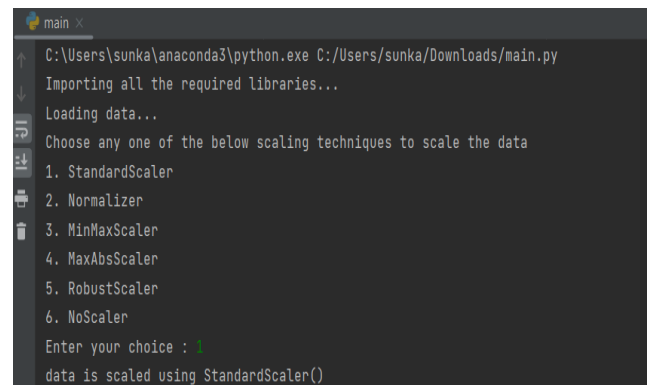
Implementation

    Initially the Email dataset is loaded and encoded to feed to different classifier models.

    Later the data is divided into train and test data in terms of 80:20 ratio. Once the data is divided for training and testing partitions, we need to scale the data using different scaler techniques and after the we feed the scaled data to different Classifier models and we find the Accuracy Values according to the logic provided in the above algorithm.

At the end we compare all the accuracy values and we found that Gradient Boosting Classifier is the optimal classifier among all and has the highest accuracy value.

Results/ Comparison of Classifiers

Fig 5. Selecting Scaler

Fig 6. Selecting Classifier (Naïve Bayes)

Fig 7. Selecting another classifier

Fig 8. Selecting SVM Classifier

Fig 9. Comparison among Classifiers

VIII. CONCLUSION

This study compares various classification algorithms. The studies used email datasets from Kaggle repositories. Following comparison, it is shown that Ensemble Classifier outperforms the other classifiers in terms of outcomes and testing speed. Our findings demonstrate that the Gradient Boosting is a more effective classifier with preset norms.

One of the validity risks in our experiment is that testing was done on an email dataset without considering the evolving patterns in the emails, which may alter a classifier's accuracy. In the future, to address the drift issue in email spam filtering, ensemble classifier with Drift Detection approach may be used.

REFERENCES

[1] "Global spam volume as percentage of total e-mail traffic from January 2014 to September 2019, by month." https://www.statista.com/statistics/420391/spam-email-traffic-share/.

[2] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise," Elsevier, vol. 2015, pp. 101–102. [3] O. Saad, A. Darwish, and R. Faraj, "A survey of machine learning techniques for Spam filtering," IJCSNS Int. J. Comput. Sci. Netw. Secur.

[4] K. Asif, A. Sami, S. Bharindhan, and K. Krishan, "A Comprehensive Survey for Intelligent Spam Email Detection," IEEEXplore, 2019.

[5] "Number of e-mail users worldwide from 2017 to 2024." [Online]. Available: https://www.statista.com/statistics/255080/number-of-e-mailusers-worldwide/.

[6] M. Guntrip, "https://www.proofpoint.com/us/corporateblog/post/fbi-reports-125-billion-global-financial-lossesdue-business-email-compromise."

[Online]. Available: https://www.proofpoint.com/us/corporate-blog/post/fbi-reports-125- billion-global-financial-losses-due-business-email-compromise.

[7] "Australian Competition and consumer Commission," Scam Stat., [Online]. Available: https://www.scamwatch.gov.au/scamstatistics?scamid=all & date=2018.

[8] K. Jackowski, B. Krawczyk, and M. Wozniak, "Application of adaptive ´ splitting and selection classifier to the spam filtering problem," Cybern. Syst. An Int. J.

[9] Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification," ResearchGate.

[10] F. Qian, Y. C. H. Abhinav Pathak, Z. M. Mao, and Y. Xie, "A case for unsupervised-learning-based spam filtering," Univ. Minnesota J., 2010.

[11] Y. Alamlahi and A. Muthana, An Email Modelling Approach for Neural Network Spam Filtering to Improve Score-based Anti-spam Systems. Modern Education and Computer Science Press, 2018.

[12] L. Melian and A. Nursikuwagus, "Prediction student eligibility in vocation school with Na¨ıve-Byes decision algorithm," 2018.

[13] A. S. Aski and N. K. Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques," Elsevier, vol. 2016, pp. 145–149.

[14] K. Pawar and M. Patil, "Pattern classification under attack on spam filtering," IEEExplore, 2015. [15] A. K. Rajan, V, and A K. "V, V., & Rajan, "An Improved Spam Detection Method With Weighted Support Vector Machine," IEEE Explor. ." IEEExplore.

[16] H. Kaur and A. Sharma, "Improved Email Spam Classification Method Using Integrated Particle Swarm Optimization and Decision Tree," IEEE Xplore, vol. 2016, pp. 516–521.

[17]Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.

[18] Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. https://doi.org/10.1109/ACCESS.2019.2954791

[19]K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[20]Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014

[21] https://llcapstoneproject.weebly.com/motivation.html

[22] S. Suryawanshi, A. Goswami and P. Patil, "Email Spam Detection : An Empirical Comparative Study of Different ML and Ensemble Classifiers," 2019 IEEE 9th International Conference on Advanced Computing (IACC), 2019, pp. 69-74, doi: 10.1109/IACC48062.2019.8971582.

[23]N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113, doi: 10.1109/ICIRCA48905.2020.9183098.

[24] M. RAZA, N. D. Jayasinghe and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," 2021 International Conference on Information Networking (ICOIN), 2021, pp. 327-332, doi: 10.1109/ICOIN50884.2021.9334020.

[25] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Wozniak, ´ "Ensemble learning for data stream analysis: A survey", Inf. Fusion, vol. 37, pp. 132–156, 2017.

[26]Jyh-Jian Sheu,Ko-Tsung Chu,Nien-Feng Li,Cheng-Chi Lee, " An efficient incremental learning mechanism for tracking concept drift in spam filtering", February 2017.