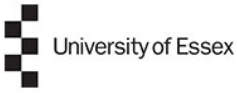
Spam Detection Using Machine Learning



Sruthi Belaganti

Department of Mathematical Sciences

University of Essex

MA902-7-FY: Research Methods

Registration number: 2010557

Supervisor: Dr Alexander Partner

Date of submission: 6 August 2021

Word count: 3844

Contents

Abstract

Spam is identified as inappropriate or unnecessary communications sent over the Internet. These are often sent to a wider audience for reasons like promotions, identity theft, spreading malware etc. Links produced with the intent deceive users through their computers are known as phishing attacks. When the links are clicked unintentionally, a variety of harmful activities can occur, ranging from the download of malware to the theft of personal information.A spam-detector algorithm must detect out how to filter out spam while avoiding marking legitimate items as spam that people wish to receive in their mailbox.Machine learning is utilised in a variety of applications, including online recommendation systems, Facebook friend suggestions,email spam filters among others.This study summarizes different machine learning classifiers like Support Vector Machine(SVM) , Random Forest Classifier, Naive Bayes Classifier to detect spam and filter the spam emails. A classification algorithm is a function that weights the input features and divides the output into different classes.The functions that offer the most accurate and best separation of the groups of data are identified during classifier training. In this research, a data set is identified and three different classifiers, Naive Bayes, Support Vector Machine, Random Forest classifiers are applied on the data and accuracy of the algorithms is checked. Based on the accuracy score, a new email is passed to the algorithm to check if the algorithm classifies it as spam or ham.

Keywords: Spam · Random Forest · Naive bayes · Classifier

1 Introduction

The internet and creation of digital data have transformed information transmission. Newspapers are online, majority of the people receive news from social media, and informal blogs which have turned into primary sources of information.The amount of data has expanded considerably as a result of these changes. Those who couldn't read or write faced significant social and economic barriers in the past. Those that are digitally illiterate will face similar challenges in the future. Millions of news articles, billions of Tweets, and trillions of web pages have to be understood and processed to make sense of the enormous amount of data included in written language. Machine Learning(ML) and Artificial Intelligence(AI) can be used to automate the process of searching and categorising data so that the things that really important can be processed.Every day, a shocking amount of spam is delivered to clients' inboxes. According to the anti-spam organisation, spam was responsible for 62 percent of all email in 2004(1).It is assumed that more than 70 percent of business emails are spam(2) in the modern times. Spam emails not only waste users' time and effort in identifying and deleting unwanted communications, but they also cause a slew of other issues, such as clogging mailboxes, wasting network capacity, and consuming vital personal information.Some spammers use email addresses found in publicly accessible newsgroups. Others employ webbots, often known as spambots, which are programmes that automatically search the internet for email addresses. In most cases, Spambots retrieve email addresses by using keyword matching techniques(3).Spam filtering based on the textual content of e-mail can be considered as a specific example of text categorization from the machine learning perspective, with the categories being spam or non-spam(4).

User feedback and content-based methods are the two basic ways for distinguishing or classifying spam emails. While content-based approaches are more prevalent, feedback mechanisms aid in the prediction of undesired emails(5).Users can flag or report mails with specific actions if they are fraudulent or actionable. Content based spam filtering algorithm works in two stages: training and classification. Individual users' emails are taken from training datasets

throughout the training phase. The email data is passed to a classification algorithm and predictions are made on the test data.An alternate method of content-based filtering is Filtering at the network level, which usually consists of a blacklist. A blacklist comprises of known bad IP e-mail addresses that have previously been used to distribute spam a message.When a connection is received from a server, it is checked against a blacklist.The receiving server compares the IP address of the sending server to the blacklist to see if the sending server's IP address is listed. If it's on the list, the mail will be denied(6). The important spam filtering algorithms are as follows:

1.1 Spam filtering using Naive Bayes

Naive Bayes classification is one of the most successful spam filtering techniques. The initial concept of determining if an email is spam or not is by examining which terms appear in the message and which do not.This method starts by analysing the content of a large number of emails that have already been identified as spam or legitimate. The information gained from the "training set" is then used to compute the probability that a new email is spam or not based on the words in the email when it arrives in a user's mailbox(2).

Sahami(7)may have been the first to use machine learning algorithms for the creation of spam filters in 1998(8). He used a naive Bayes classifier to train and got satisfactory results.End-user applications such as the Mozilla e-mail client and the free software project Spam Assassin have used the naive Bayes technique in order to filter spam emails.

The Naive Bayes classification technique is a conditional probability based algorithm which assumes that all model features are independent. It is assumed that every word in the message is independent of all other words.This assumption is known as conditional Independence. Based on a set of words, the classification system generates probability for the message to be spam or not spam. Bayes formula is used to determine the likelihood, and the components of the formula are derived using the frequencies of the words in the entire set of messages.

The formula for conditional independence is given by

$$P(M|N) = \frac{P(N|M)P(M)}{P(N)} \tag{1.1}$$

where

$P(M \mid N)$ is posterior probability of the target variable.

P(M) is the prior probability of target.

$P(N \mid M)$ is the probability of predictor variable.

P(N) is the prior probability of predictor variable.

1.2 Spam Filtering using Support Vector Machines(SVM)

SVM is a supervised machine learning technique that can be used to solve classification and regression problems.SVM has been extended to regression problems after it was originally developed for classification problems.Each data item is plotted as a point in n-dimensional space where n is the number of features and the value of each feature is the value of the coordinate. SVMs are part of the generalised linear classifier family. They have the unique virtue of minimising the empirical classification error while also maximising the geometric margin(9).Thus, they are also known as maximum margin classifiers.The concept of hyperplanes supports classification in SVM-based techniques. In conventional binary categorization, such as spam or ham in spam filtering, the hyperplanes operate as class segregators. SVM training is a time-consuming and computationally intensive procedure(10).Classification is performed by locating the hyper-plane that clearly distinguishes the two classes. SVM creates a classifier by looking for an optimal separating hyperplane (optimal hyperplane) that maximises the margin between the categories (in this case spam and ham). SVM has consistency and is efficient when the number of dimensions is greater than the number of samples.The choice of kernels is a critical issue when using SVM because it directly affects the separation of emails in the feature space.

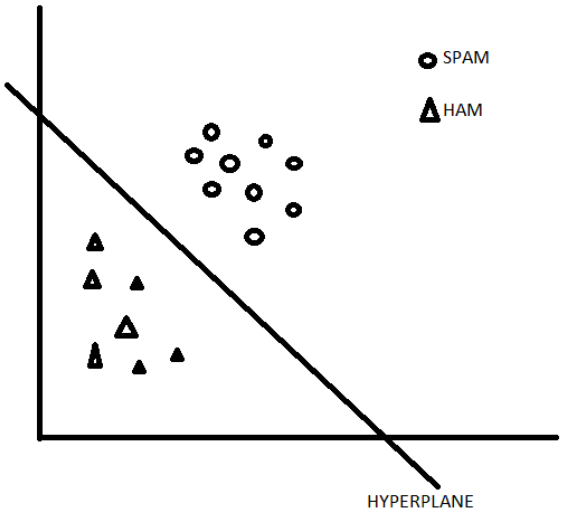


Figure 1: Classification using SVM

1.3 Spam filtering using Neural Networks

Neural networks are a type of machine learning technique that can be effectively utilised to solve classification issues. They can be used to translate features into nonlinear decision boundaries that are fairly complex.Neural networks attempt to replicate how the human brain forms classification rules. A neural net is made up of multiple layers of neurons, each of which receives input from previous layers and passes output to subsequent layers.The neural network iterates for a set number of iterations, referred to as epochs. The cost function is examined after each epoch to see where the model might be improved. Based on the information provided by the cost function, the optimising function then changes the internal mechanics of the network, such as the weights and biases, until the cost function is minimised.

The network uses the weights and functions in the hidden layers of neurons to process the records in the training Set one at a time, then compares the generated outputs to the desired outputs. The system then propagates the errors back through the system, causing the weights to be adjusted for the following entry.As the weights are adjusted, this process repeats itself. The same set of data is processed several times throughout the training of a network as the

connection weights are continually improved.

1.4 Spam filtering using random Forest Classifier

The Random Forest classifier is a tree-based classification algorithm. It uses a bootstrapping mechanism to create random numbers of samples with replacement(11).On each sample, an unpruned decision tree is modelled. A tiny fraction of input data is randomly selected during the modelling of a decision tree to determine the split at each node of the tree. The ensemble of decision trees is used to classify a test sample by a majority vote.When compared to a single decision tree classifier modelled on the training data, Random Forest can usually enhance classification accuracy significantly.When using Random Forests on classification data, Gini index oe Entropy are used to determine how nodes on a decision tree branch are connected. the formula for Gini index and entropy are given by equations 1.2,1.3.Entropy determines how the node should branch based on the probability of a specific outcome. Because of the logarithmic function used to calculate it, it is more mathematically complex than the Gini index.

$$Gini = 1 - \sum_{n=1}^{C} (p_i)^2 \tag{1.2}$$

where

p_i is the relative frequency of observed class

C is the number of classes.

$$Entropy = \sum_{n=1}^{C} -p_i \times \log_2(p_i) \tag{1.3}$$

2 Background Study

Traditional spam filtration systems have used supervised machine learning since the beginning of Internet spam. In supervised learning, a set of labelled spam and ham training data is fed into a classifier in the training phase, which uses the training corpus to develop a model.The model is then utilised to identify unfamiliar emails as spam or ham.The training corpus is crucial to supervised learning because it provides useful information, but creating the training corpus is time-consuming and expensive.

In a paper by M Sahami on Bayesian approach to filter junk emails, Vector Space model is employed, in which each dimension of the space corresponds to a single word in the corpus of emails examined(7).Each email is represented as a binary vector that indicates which words are present and which are not.With this structure, a probabilistic classifier is implemented to detect junk mail from a pre-classified set of training emails.

Based on the body of the emails, two popular machine learning approaches, Naive Bayes Classifier and Support Vector Machine, were used to classify emails as spam or ham in "A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification" by Thae Ma(12).In this approach, different sizes of training data was passed to both the classifiers and the model was evaluated based on the parameters Precision, Recall, F-statistic. It is observed that as the training data size increases, the accuracy of the model improved and SVM demonstrated better results when training data was large.

In a study of spam filtering using SVM, Amayri investigated a variety of distance-based kernels and used SVM to determine spam filtering behaviours(13). Majority of the kernels employed in recent studies address continuous data and ignore the text's structure. Rather than using traditional kernels, the author proposed using a variety of string kernels for spam filtering and demonstrated how well string kernels work for spam filtering. A comparison of seven distinct versions of Naive Bayes classifiers like Boolean Naive bayes, Multinomial Naive Bayes, Basic naive bayes etc and the Support Vector Machine has been done by Almeida(9) to automatically filter e-mail spams. For all investigated data sets, SVM had

the best average performance, with an accuracy rate of more than 90% for all datasets tested indicating that SVM is a better classifier for the data sets considered for the analysis. Similar results are obtained in comparison of Naive bayes and SVM in the work done by Thae Ma(12). Chandra(5) offered a new approach for detecting Spam and Ham from the Spam SMS Collection using Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) using Keras models with Tensorflow in the backend and compared the results to Naive Bayes and SVM classifiers.The overall accuracy is 98 percent, which is better than Naive Bayes and SVM whose accuracy is 80 and 97 respectively. Tang(11) performed a study on comparison of SVM and Random Forest Classifier to analyze spam email.He concluded that SVM and Random Forest(RF) are both useful tools for creating accurate classifiers. Between the two, RF is marginally precise but costs more in terms of time and space. If the modelling parameters are fixed, SVM provides similar accuracy in a much more efficient manner.Hence SVM can be used for classification. Enrico Blanzieri proposed a classifier combining SVM and K-Nearest Neighbours(KNN) algorithms and compared the results with KNN and SVM individually.Given a sample to classify, the KNN algorithm finds the k closest samples in the training set using Euclidean metric for determining nearest neighbors , which are then used to train an SVM classifier. The unknown sample is then classified using the learned SVM classifier.On the small dimensions of the feature space, the combined classifier outperformed SVM considerably. On larger dimensions, the advantage is less evident; one probable reason for this is because, in comparison to SVM, it is more sensitive to irrelevant features(14).

3 Method

3.1 Dataset

The dataset consists of 5171 observations with the columns label,text and label number. The label indicates if the email is spam or ham, text represents the content of the email and label number represents the label. The label number is 0 if the email is not spam and 1 if the email is spam.

3.2 Preprocessing

The modifications done to the data before fitting an algorithm to it is referred to as preprocessing. Data preprocessing is a method for converting unclean data into a clean dataset.Some Machine Learning models require data in a specific format.For instance, the Random Forest technique does not take null values, therefore null values must be handled from the original raw data set in order to run the algorithm.The data set should be organised in such a way that any Machine Learning algorithm can be run in parallel and best one can be chosen.
The dataset is read into the python environment and all the necessary libraries required for the analysis are imported. The first 5 observations are examined to check the structure of the data in figure2.

	label	text	label_num
0	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	spam	Subject: photoshop , windows , office . cheap ...	1
4	ham	Subject: re : indian springs\r\nthis deal is t...	0

Figure 2: Spam and Ham emails

The dataset is checked for missing values and it is observed that there are no missing values in the dataset. The number of spam and ham emails in the dataset are visualised in figure3.

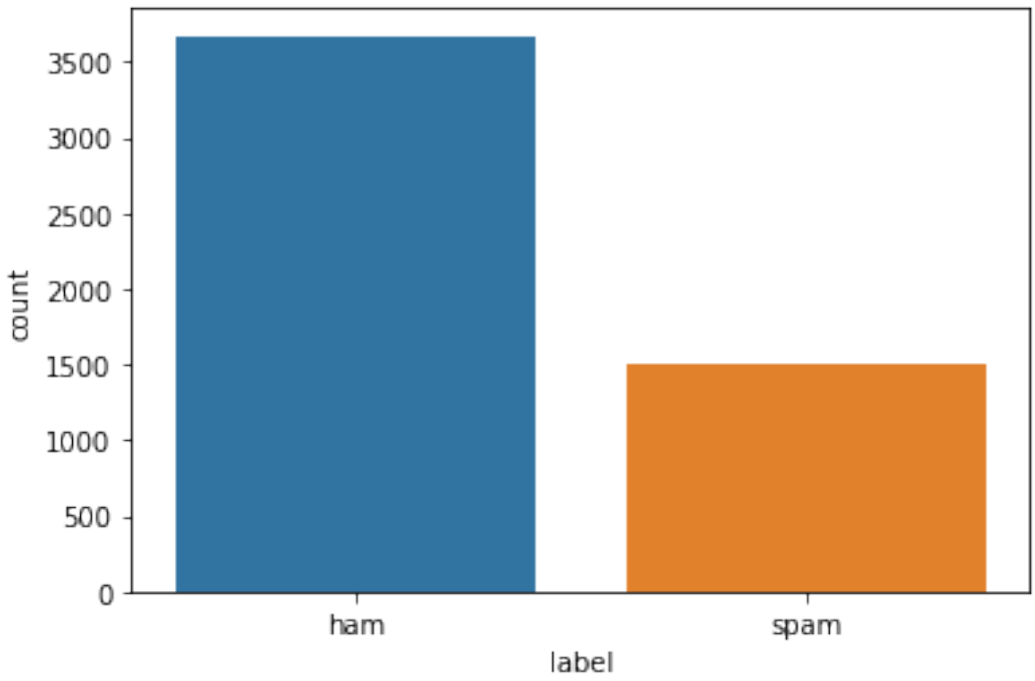


Figure 3: Spam and Ham emails

The number of spam emails is 1499 which comprises of 29% of the total emails and the rest 3672 , 71% of the emails are observed as legitimate ones.To categorise emails as ham or spam, a natural language processing (NLP) algorithm which is also called as bag-of-words is used. Features from textual data are extracted using a bag-of-words model. As the algorithm does not grasp language, the words in the corpus must be represented numerically. For analysis, this numeric form can be fed into any algorithm.For converting text data into features, there are a variety of feature engineering methodologies. Some of them include assigning a feature to each distinct word and calculating the number of occurrences per training sample. In the process of cleaning the data, Natural Language Toolkit(NLTK) library is used. Each word of the text is split into tokens using the tokenization and all the words are converted into lower case. The punctuation marks between the sentences and stop words are deleted from the text

data. Stop words in English is a list of words in nltk library which consists of the frequent words which do not add much meaning to the sentence. Any stopwords can be added to the predefined list depending on the content of the text. Eliminating the stop words saves computational space and time.

The most important words which are used to differentiate spam and ham emails are visualized using word clouds. A word cloud is an effective visual representation for text that displays the most significant used words in varied font sizes and colors. If size of the words is smaller, it indicates that the words are less influential. Figure4 demonstrates the word cloud of spam emails.

Figure 4: Prominent words of spam emails

Figure5 demonstrates the word cloud of legitimate emails.

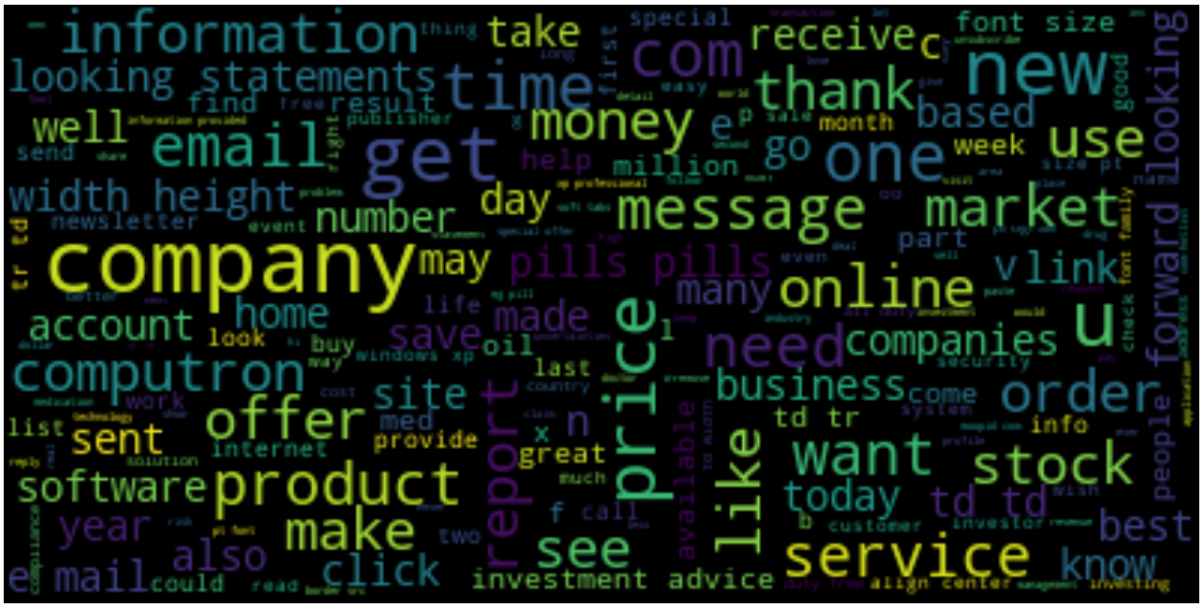


Figure 5: Prominent words of ham emails

The input and output variables are separated into two different data frames X and y respectively. The input variable is the text variable and the output variable is the label_num which is 1 when the email is spam and 0 if it is not.

3.3 Modelling the data

In Supervised learning, the data is usually split into subsets called training and test sets.Machine learning algorithms' performance is estimated using the train-test split technique. The training dataset is the first subset, which is used to fit the model.The second subset is the test set which is not used to train the model; instead, the model makes predictions on the test set and compares them to the actual values. Depending on the performance of the model on the test dataset,predictions can be made for new data where the output or target values are not known.

To use textual data for predictive modelling, words in the text must first be converted to integers or floating-point numbers in order to be used as inputs for machine learning models.The scikit-learn's CountVectorizer is used to convert text into a vector based on the frequency of each word that appears throughout the text.

In this dataset, 70% of the data is considered as training data and the rest is considered as test data. Naive Bayes classification algorithm is applied on the training data. When the algorithm encounters a word, it calculates the probability of a word being spam or ham. When the probability of the spam words in a mail is greater than ham words, the email is identified as spam. Equations 3.1 and 3.2 give the formulae to calculate the probabilities of spam and not spam.

$$P(x_i \mid Spam) = \frac{Nx_{i|Spam} + \alpha}{N_{Spam} + \alpha.N_{total}} \tag{3.1}$$

$$P(x_i \mid Ham) = \frac{Nx_{i|Ham} + \alpha}{N_{Ham} + \alpha.N_{total}} \tag{3.2}$$

where

x_i=x_1,x_2,x_3...x_n, x_1 is the first word.

$Nx_{i|Spam}$ is frequency of word x_i in spam emails

$Nx_{i|Ham}$ is frequency of word x_i in Ham emails

N_{Spam} is the number of words in spam email

N_{Ham} is the number of words in Ham email

N_{total} is the total number of words in the text

α is Laplace smoothing parameter which prevents the problem of zero probability. The value is usually 1.

The Naive Bayes model which is fit on the training data is evaluated using test data set.

Similarly SVM classifier object is fit on the training data using linear kernel and sigmoid Kernel.SVM classifies the data using hyperplane and the support vectors which define the

hyperplane.The is used to make predictions based on hypothesis function h defined as follows:

$$h(x_i) = \begin{matrix} +1 & if w.x + b \geq 0 \\ -1 & if w.x + b \leq 0 \end{matrix} \tag{3.3}$$

where

w is width of the margin

b is the intercept

x is the vector

4 Results

Evaluating a model is a crucial step in creating a successful machine learning model. Accuracy is the most common classification evaluation metric.Other metrics like recall and precision can be included as well to get a complete view of the model evaluation.The count of correctly and incorrectly predicted test records by a classification model are used to evaluate the model's performance. The confusion matrix provides a detailed picture of not only a predictive model's performance, but also which classes are being predicted correctly and wrongly, as well as the kind of errors committed.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Table 1: Structure of Confusion Matrix

TP represents true positive

FP represents false positive

FN represents false negative

TN represents true negative

A true positive is when the model predicts the positive class correctly. A true negative, on the other hand, is a result in which the model predicts the negative class accurately.A false positive occurs when the model predicts the positive class inaccurately. A false negative is an outcome in which the model predicts the negative class inaccurately.

Figure 6 represents the confusion matrix of the data using Naive Bayes classifier.

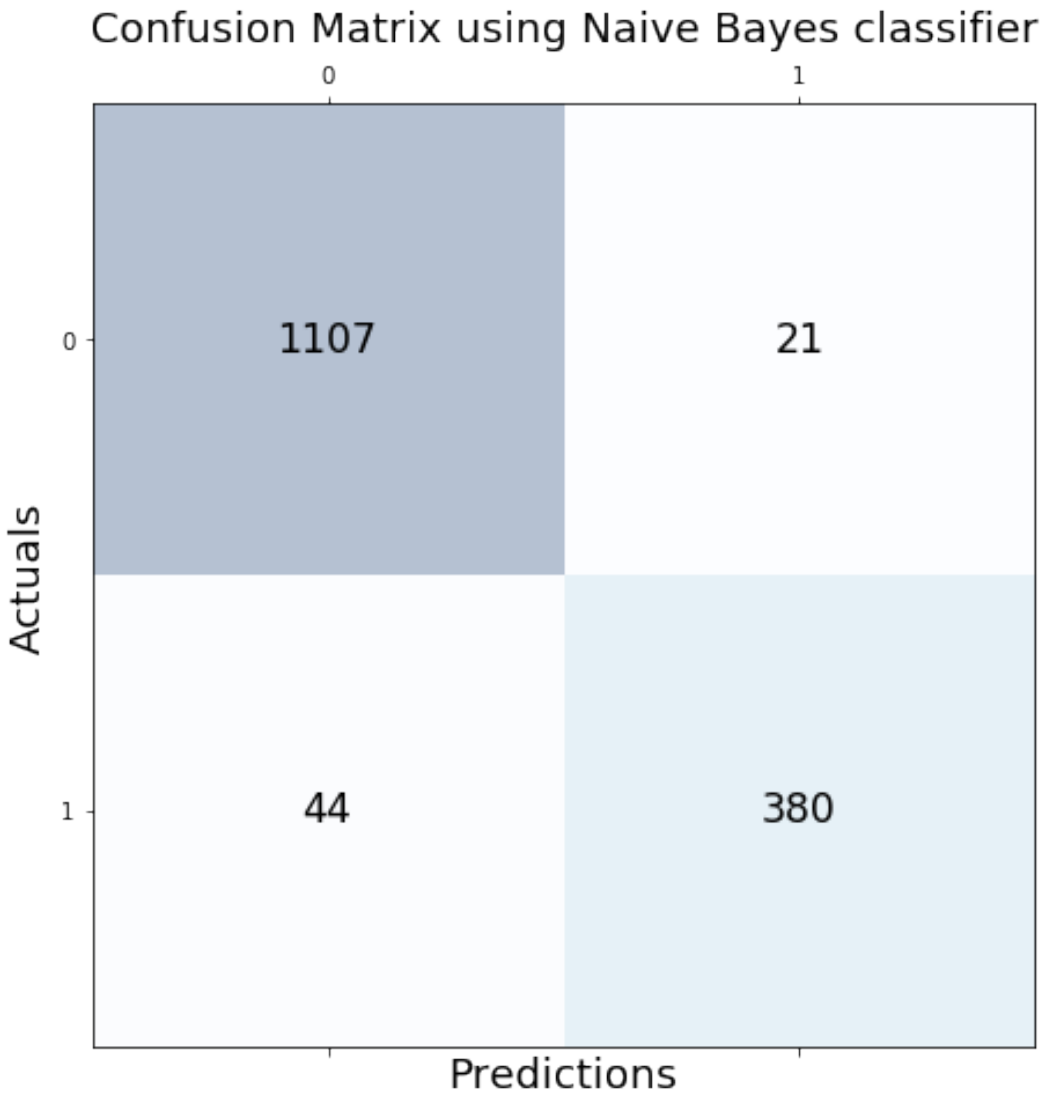


Figure 6: confusion matrix using naive bayes classifier

1107 ham emails were correctly classified by the model.380 spam emails were correctly classified by the model.21 ham emails were incorrectly classified as belonging to spam by the model.44 spam emails were incorrectly classified as belonging to ham by the model.

Figure 7 represents the confusion matrix using SVM linear kernel. It is observed that 1091 ham emails were correctly classified by the model.402 spam emails were correctly classified by the model.37 ham emails were incorrectly classified as belonging to spam by the model.22 spam emails were incorrectly classified as belonging to ham by the model.

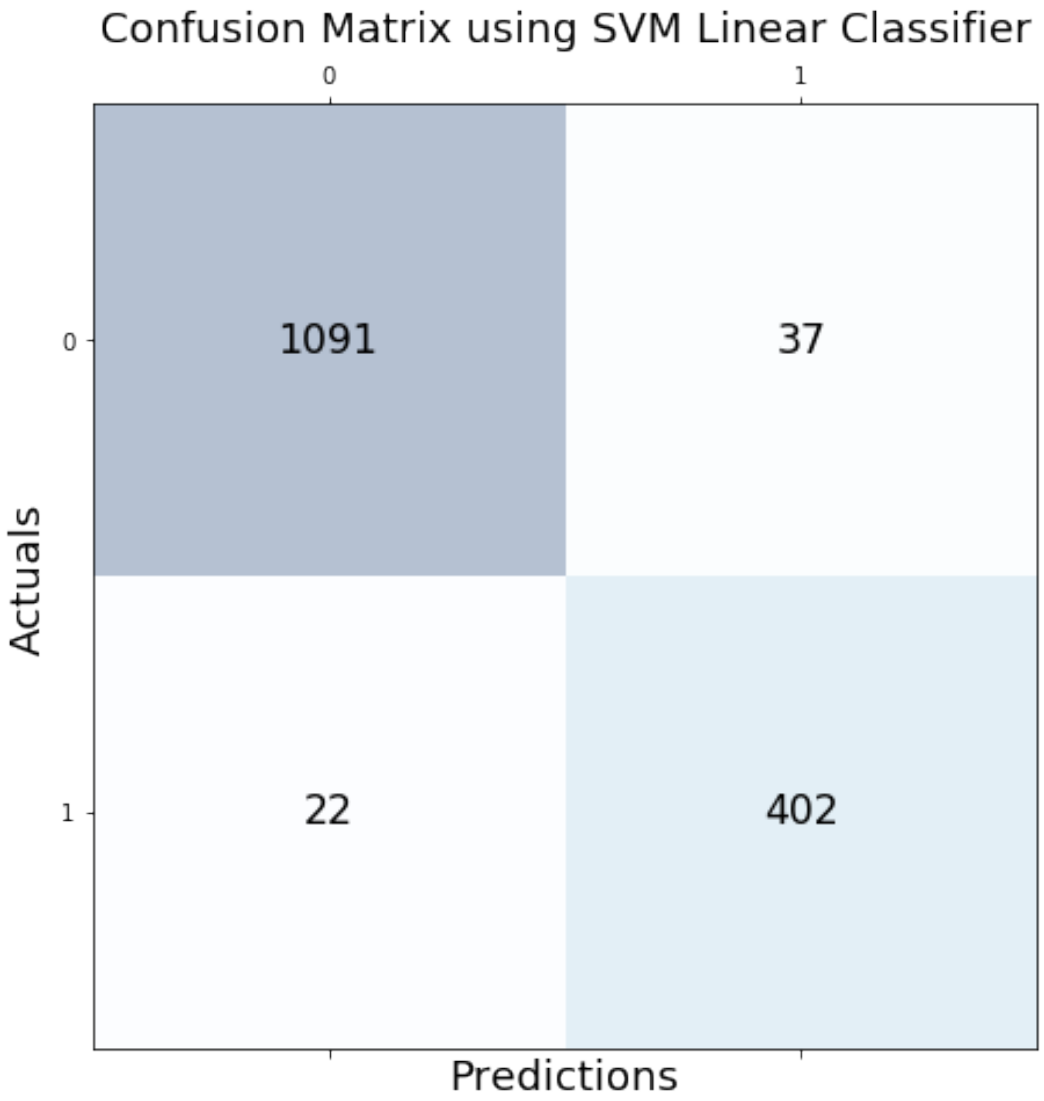


Figure 7: confusion matrix using SVM linear classifier

Figure 8 represents the confusion matrix using SVM sigmoid kernel. It is observed that 1070 ham emails were correctly classified by the model.421 spam emails were correctly classified by the model.58 ham emails were incorrectly classified as belonging to spam by the model.3 spam emails were incorrectly classified as belonging to ham by the model.

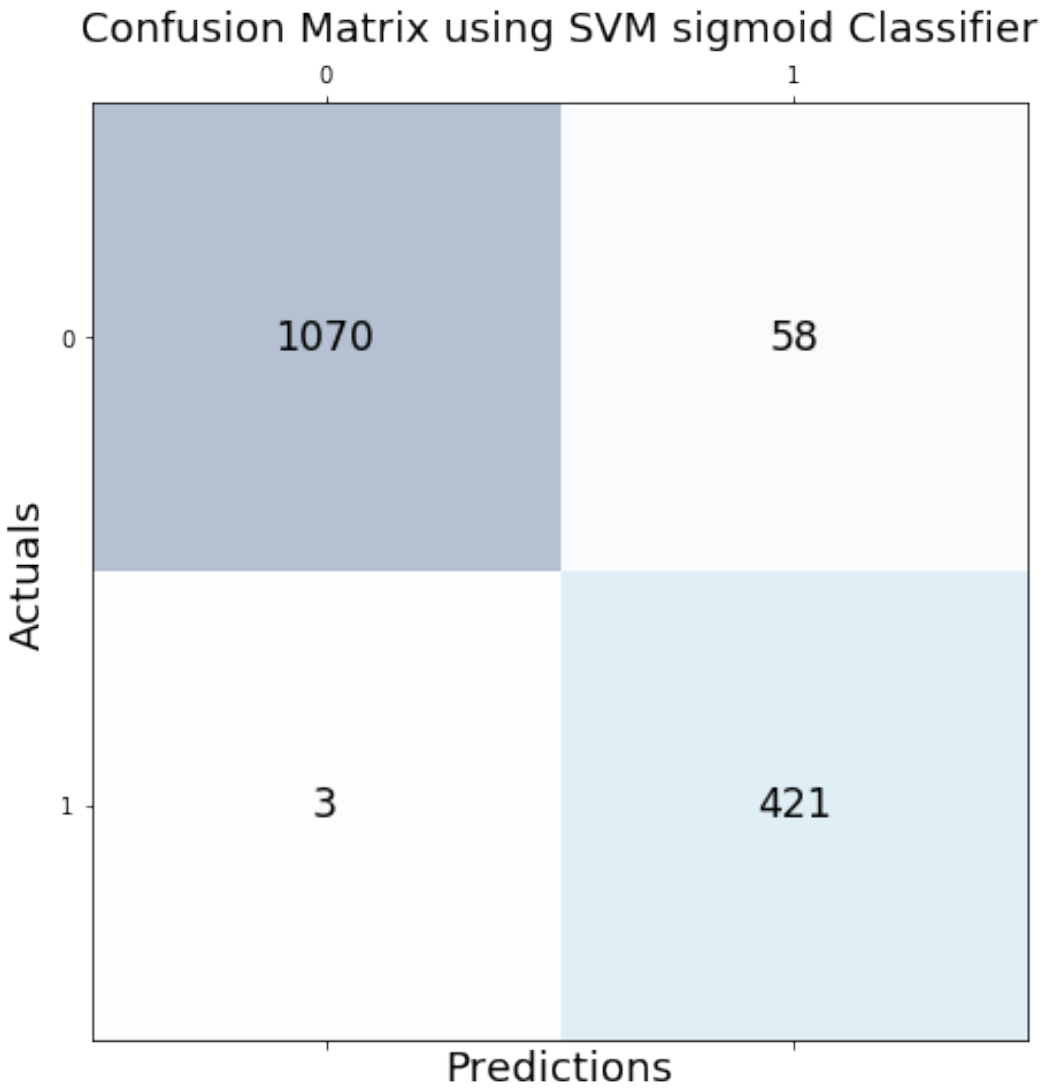


Figure 8: confusion matrix using SVM Sigmoid classifier

The ratio of number of correct predictions divided by the total number of input samples is the accuracy of the model.The accuracy of the three models are calculated and compared in table 2.The formula for accuracy is given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Model	Accuracy
Niave Bayes	95.81
SVM using linear kernel	96.19
SVM using sigmoid kernel	96.06

Based on the experiment, it is found that SVM classifier on a whole has better accuracy than Naive bayes classifier. SVM Linear kernel has marginally high accuracy when compared to SVM sigmoid kernel. Hence SVM linear kernel classifier can be used to classify new emails as spam or not spam.In a study by Agarwal on comparision of different SVM kernels for spam classification,linear kernels with SVC outperformed other kernels (RBF, polynomial, sigmoid)(1). Based on the observations of related works in the background study and experiments on the dataset considered, altogether SVM is regarded as a better classifier than Naive Bayes for spam classification.

4.1 Limitations & Future research

In case of Naive Bayes classifiers, With tiny data sets, the precision will be affected.Because the assumption that all features are independent is rarely true in practise, the naive bayes algorithm is less accurate than more complex algorithms because it oversimplifies the problem. A range of machine-learning based algorithms, including as neural networks, maximal entropy model, could be presented for spam filtering in the approaches considered in this research.It is well established that integrating numerous categorization models yields better results than employing a single, fine-tuned model. However, the methods for merging several classifiers remain a critical aspect in determining aggregated prediction performance(15).In the future combining multiple models to check if it yields better results could be a topic of research.

4.2 Conclusion

In this study, four different classification techniques for email spam filtering , Naive Bayes classifier, Support Vector machines ,Random Forest classifier , Neural Networks have been reviewed. A data set is taken for applying SVM and Naive Bayes classifier and it is observed

from the results that SVM is a better classifier when compared to naive bayes classifier. Overall, Machine Learning has a lot of potential for defending against cyber-threats like spamming emails. Some governments and businesses are employing or considering employing machine learning techniques to combat cyber criminals. While there are valid privacy and ethical concerns about spam filters, governments must ensure that restrictions do not impede enterprises from adopting machine learning for security purposes.

References

[1] D. K. Agarwal and R. Kumar, "Spam filtering using svm with different kernel functions," *International Journal of Computer Applications*, vol. 136, no. 5, pp. 16–23, 2016.

[2] B. Yu and Z.-b. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms," *Knowledge-Based Systems*, vol. 21, no. 4, pp. 355–362, 2008.

[3] A. Khorsi, "An overview of content-based spam filtering techniques," *Informatica*, vol. 31, no. 3, 2007.

[4] C.-C. Lai, "An empirical study of three machine learning methods for spam filtering," *Knowledge-Based Systems*, vol. 20, no. 3, pp. 249–254, 2007.

[5] A. Chandra and S. K. Khatri, "Spam sms filtering using recurrent neural network and long short term memory," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, 2019, pp. 118–122.

[6] O. Kufandirimbwa and R. Gotora, "Spam detection using artificial neural networks (perceptron learning rule)," *Online Journal of Physical and Environmental Science Research*, vol. 1, no. 2, pp. 22–29, 2012.

[7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62. Citeseer, 1998, pp. 98–105.

[8] J. Hovold, "Naive bayes spam filtering using word-position-based attributes." in *CEAS*, 2005, p. 41.

[9] T. A. Almeida and A. Yamakami, "Content-based spam filtering," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–7.

[10] G. Caruana, M. Li, and M. Qi, "A mapreduce based parallel svm for large scale spam filtering," in *2011 eighth international conference on fuzzy systems and knowledge discovery (fskd)*, vol. 4. IEEE, 2011, pp. 2659–2662.

[11] Y. Tang, S. Krasser, Y. He, W. Yang, and D. Alperovitch, "Support vector machines and random forests modeling for spam senders behavior analysis," in *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*. IEEE, 2008, pp. 1–5.

[12] T. M. Ma, K. Yamamori, and A. Thida, "A comparative approach to naïve bayes classifier and support vector machine for email spam classification," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2020, pp. 324–326.

[13] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artificial Intelligence Review*, vol. 34, no. 1, pp. 73–108, 2010.

[14] E. Blanzieri and A. Bryl, "Instance-based spam filtering using svm nearest neighbor classifier." in *FLAIRS Conference*, 2007, pp. 441–442.

[15] Z. Yang, X. Nie, W. Xu, and J. Guo, "An approach to spam detection by naive bayes ensemble based on decision induction," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 2. IEEE, 2006, pp. 861–866.