

DETECTION OF SPAM MAIL USING MACHINE LEARNING

ABSTRACT

Spam email is unwanted junk email sent out in bulk to a random recipient list. There are different ways that spam can be sent. It could be sent by humans but the most common way is sending them through a network of computers called botnets (spambots). These botnets are often infected with malware and other viruses. The other way in which spam can be shared is by using text messages or social media. Spam Emails tend to choke our inboxes and often cause the mail box to become clumsy. Some Spam Emails might even cause a threat to our systems. Email spammers target victims and trick them into downloading malware, erasing data, or more dangerously getting access to private data. Spammers try and make false claims and trick the victims into believing it. Email spamming continues to grow and as of 2021 it stands at 45% approximately. Through this project we aim to detect or spot spam emails and classify them using different Machine Learning algorithms. We would also compare the algorithms and check the best algorithm which has a better precision and accuracy value.

1. INTRODUCTION

1.1. Introduction:

Spam Email looks like a modern problem to deal with, but it goes way back in history to 1978 where a spam email to promote a new product was sent by Gary Thuerk. To solve this problem of spotting or detecting a Spam Email, there are various Machine Learning techniques in both Supervised and Unsupervised learning. For this particular project we would be focusing on using three Supervised Machine Learning techniques:

- KNN or K-nearest neighbors
- SVM or Support vector machine
- Decision Trees

1.2. Data Preprocessing:

The process of converting data to something a computer can understand is referred to as pre-processing. For the Machine to understand, analyze and operate, the data, the texts (emails/message in the dataset) should be readable. One of the major forms of pre-processing is to filter out useless data.

Extraction of Text Data using Tf-id Vectorizer:

Machines don't understand human language so we need to preprocess the data to make our data understandable by machines. In natural language processing, useless words (data), are referred to as stop words which are commonly used words (such as "the", "a", "an", "in") that are usually ignored since they take up unwanted space. Feature extraction refers to the mapping from string or textual data to real vectors.

The objects we are trying to classify are text messages or strings and strings are unfortunately very hard to handle and preprocess. Many of the ML algorithms can only handle numerical data and therefore there is a need to convert the string data into a set of numerical values or vectors so that we can apply the ML algorithms on them. We are first required to convert all the data into a vector of numbers and then classify them.

For this we use Term Frequency-Inverse Document Frequency. Term frequency (TF) is used to measure the number of times of occurrence or simply the frequency of a word in a file and is found by dividing the frequency of a word by the total number of words in that document.

Inverse Document Frequency (IDF) helps reduce the weightage of terms that are very common in a set of documents and is calculated by taking the log of the total number of documents divided by the number of documents in which that specific term is present.

$IDF = \log(N/n)$

$TF-IDF = \text{Term Frequency (TF)} \times \text{Inverse Document Frequency (IDF)}$.

2. LITERATURE SURVEY

Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges.

Naeem Ahmed, 1 Rashid Amin, 1 Hamza Aldabbas, 2 Deepika Koundal, 3 Bader Alouffi, 4 and Tariq Shah1

Published 3 February 2022

Email spam, called as junk emails or unwanted emails, are a type of email that can harm the user victim by wasting computing resources, and stealing valuable information. Spam filtering is a very important task. The paper provides a broad study of Machine Learning and its types. It gives a review of many of the Machine Learning techniques that can be used for Spam Filtering. A detailed comparison including different parameters like accuracy, precision, recall have been discussed in depth. Both Supervised and Unsupervised Machine Learning techniques are used and they are compared based on their accuracy, precision value and other important parameters. Insights of what spam emails are and the different types of spam emails and spam filtering methods are elucidated. It also provides a comprehension or understanding of future spam detection or filtering methods that are open to research and provide better security email platforms. Finally the paper gives us research gaps, challenges of spam detection and also future space and area of research.

A Comparative Analysis of SMS Spam Detection employing Machine Learning Methods.

Humaira Yasmin Aliza; Kazi Aahala Nagary; Eshtiak Ahmed; Kazi Mumtahina Puspita;

The aim of this study is to detect spam message to prevent different cybercrimes as spam messages have become a security threat nowadays. In this paper, studies on SMS spam problems to perform a better accuracy using several different techniques such as Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Random Forest, Logistic Regression and some more are performed. The result indicated that Support Vector Machine achieved the highest accuracy of 99%, indicating it might be useful as an effective machine learning system for future research.

3. SYSTEM MODEL:

The following are the Machine Learning models that we have used to detect spam mails:

3.1. K-NEAREST NEIGHBORS:

KNN (K-Nearest Neighbors) is one of the simplest supervised learning algorithms. “K” stands for number of data set items that are considered for the classification. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

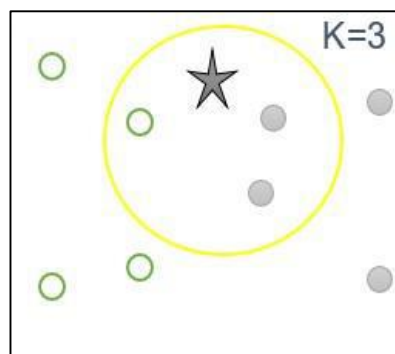


Fig 1. KNN model

3.2. SUPPORT VECTOR MACHINE:

Support Vector Machine Learning Algorithm Support Vector Machine is used for classification and for regression problems. The datasets are used to train the SVM to classify any new data that it receives. It is a supervised machine learning algorithm that works by finding a hyperplane that classifies the dataset into different classes. The SVM maximizes the distance between different classes because of the existence of many linear hyperplanes which is called as margin maximization.

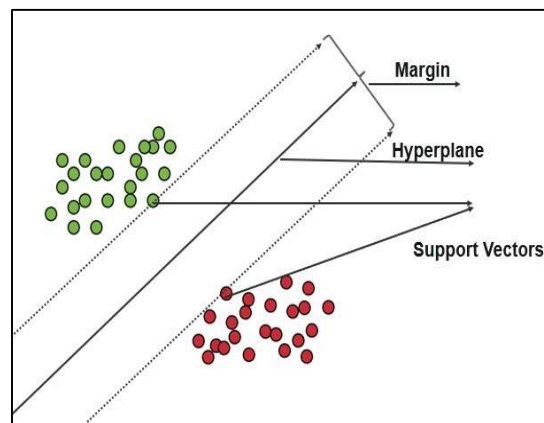


Fig 2. SVM model

SVM algorithm can be used for Face detection, image classification, text categorization, etc. SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier. The SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyper plane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

3.3. DECISION TREES:

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

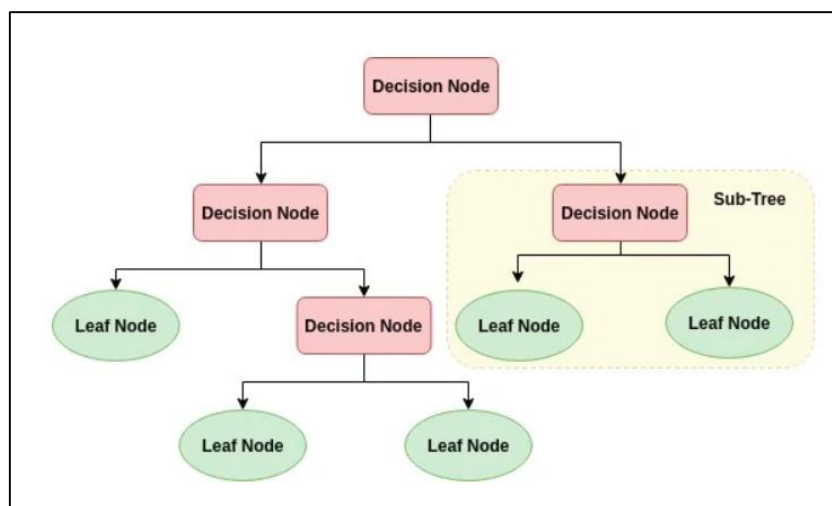
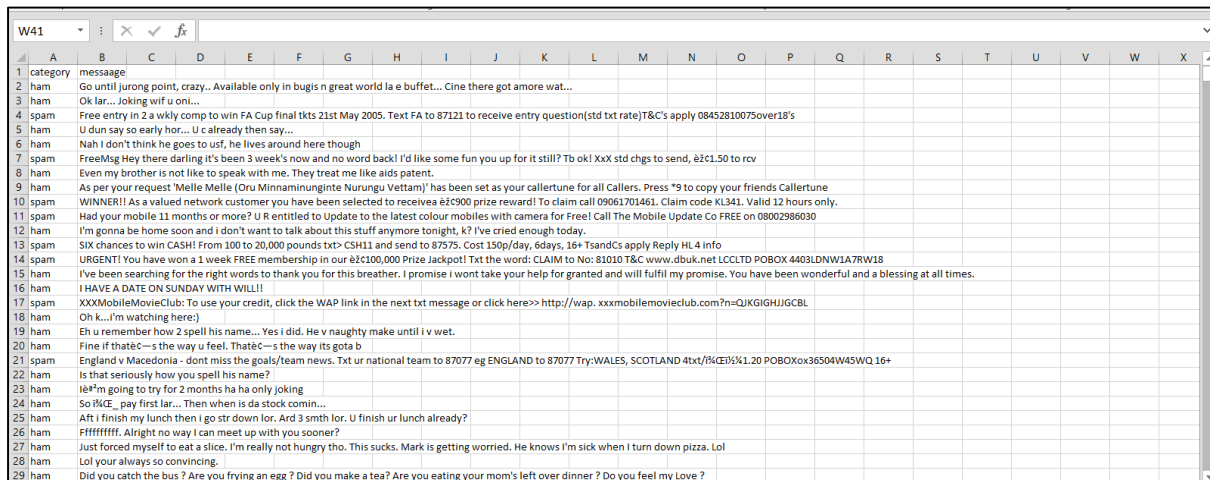


Fig 3. Decision tree model

4. IMPLEMENTATION

4.1. DATASET AND SOFTWARE:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	category	message																						
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...																						
3	ham	Ok lar... Joking wif u oni...																						
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's																						
5	ham	U dun say so early hor... U c already then say...																						
6	ham	Nah I don't think he goes to usf, he lives around here though																						
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, €£1.50 to rcv																						
8	ham	Even my brother is not like to speak with me. They treat me like aids patient.																						
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nuringu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune																						
10	spam	WINNER!! As a valued network customer you have been selected to receive a €1000 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.																						
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030																						
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.																						
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info																						
14	spam	URGENT! You have won a 1 week FREE membership in our €100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7R1W18																						
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.																						
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!																						
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJGIGHUJGCB																						
18	ham	Oh k..i'm watching here)																						
19	ham	Oh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.																						
20	ham	Fine if that's the way u feel. That's the way its gota b																						
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/160p 1.20 POBOXox36504W45WQ 16+																						
22	ham	Is that seriously how you spell his name?																						
23	ham	Ie*I'm going to try for 2 months ha ha only joking																						
24	ham	So i'll pay first lar... Then when is da stock comin...																						
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?																						
26	ham	FFFFFFF. Alright no way i can meet up with you sooner?																						
27	ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows i'm sick when i turn down pizza. Lol																						
28	ham	Lol your always so convincing.																						
29	ham	Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?																						

Fig 4. Dataset

- We have used ANACONDA and Jupyter Notebook to run and simulate the code and produce the output and graphs.

4.2. TRAINING AND TESTING

The 'train_test_split' function is used and the train:test ratio is taken to be **8:2**.

4.3. LIBRARIES/ MODULES USED:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score
import seaborn as sns
```

- Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
- NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- The `train_test_split()` method is used to split our data into train and test sets.
- `TfidfVectorizer` - Transforms text to feature vectors that can be used as input to estimator.
- `StandardScaler` removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way.

4.4. DATA PREPROCESSING

The original dataset when imported consists of 5 columns. Out of the five columns, three columns are unnamed or not named and they had null values in them. These three columns don't have any use so we have removed them. The other two columns are named. The column that has categorical values- 'spam/ham' are converted to numerical or real vectored values so that we can apply ML models on them.

5. SAMPLE CODE

CODE SNIPPETS:

```
## KNN
knn = KNeighborsClassifier(n_neighbors=9, metric='euclidean')
knn.fit(x_train, Y_train)
pred_knn= knn.predict(x_test)
accuracy_knn= accuracy_score(Y_test, pred_knn)
print('Accuracy of KNN : ', accuracy_knn)
```

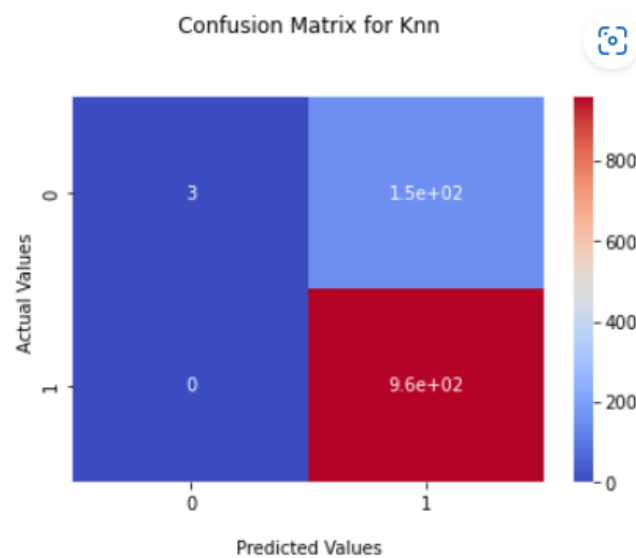


FIG 5: CONFUSION MATRIX FOR KNN

```
## SVM
svm_model=svm.SVC(kernel="linear")
svm_model.fit(X_train_features,Y_train)
pred_svm=svm_model.predict(X_test_features)
accuracy_svm=accuracy_score(Y_test,pred_svm)
print("accuracy of svm is : ",accuracy_svm)
```

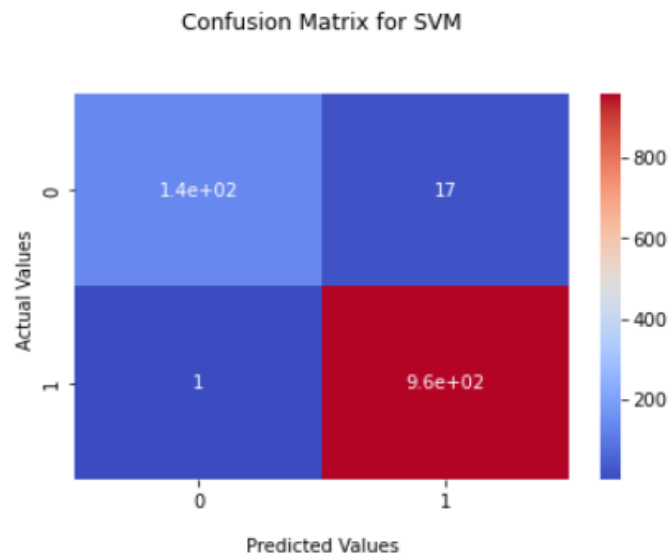



FIG 6: CONFUSION MATRIX FOR SVM

```
## DECISION TREE
classifier = DecisionTreeClassifier(criterion='entropy',random_state = 0)
classifier.fit(x_train, Y_train)
pred_dt = classifier.predict(X_test_features)
accuracy_dt=accuracy_score(Y_test,pred_dt)
print("accuracy of Decision tree is :",accuracy_dt)
```

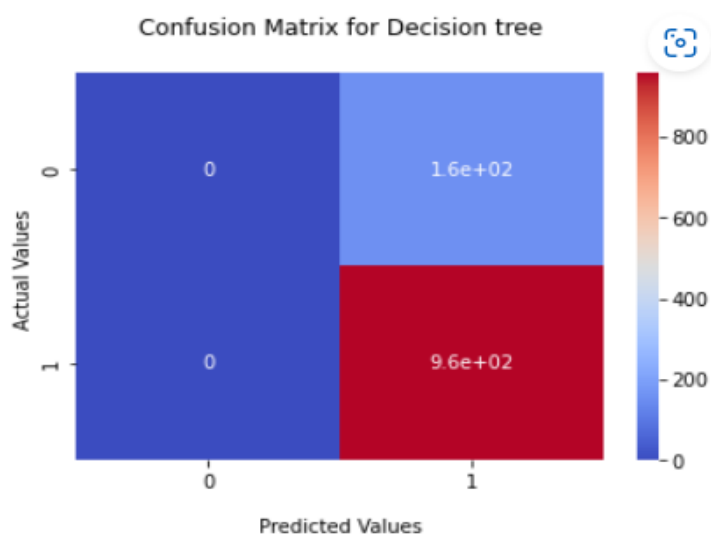


FIG 7: CONFUSION MATRIX FOR DECISION TREE

6. SAMPLE OUTPUT

ALGORITHM	ACCURACY	PRECISION	RECALL
KNN	86.36%	86.33%	100%
SVM	98.38%	98.25%	99.89%
DECISION TREE	86.09%	86.09%	100%

FIG 8: COMPARING THE ML MODELS BASED ON DIFFERENT PARAMETERS

- From the above table it can be analyzed that the accuracy is best for SVM. Even the precision score is the best for SVM whereas recall score is better for both KNN and Decision Trees.

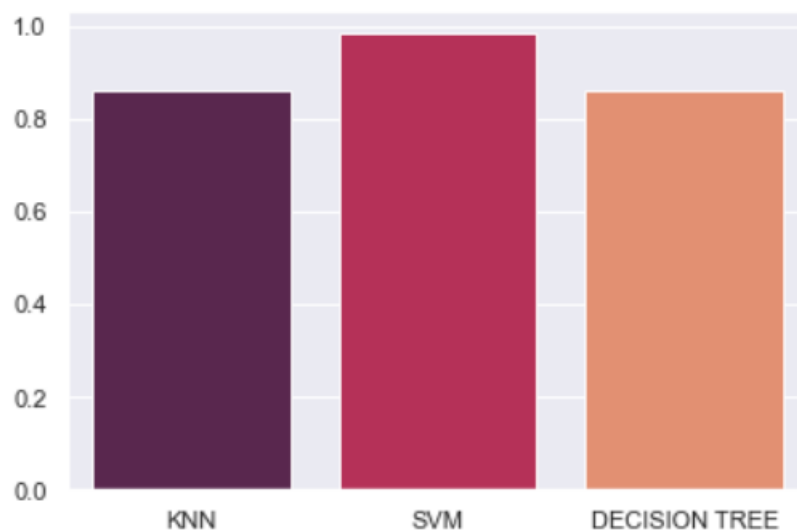


FIG 9: GRAPH COMPARING THE ML MODELS BASED ON ACCURACY

- Even from the bar graph given above it can be analyzed that the accuracy is best for SVM.

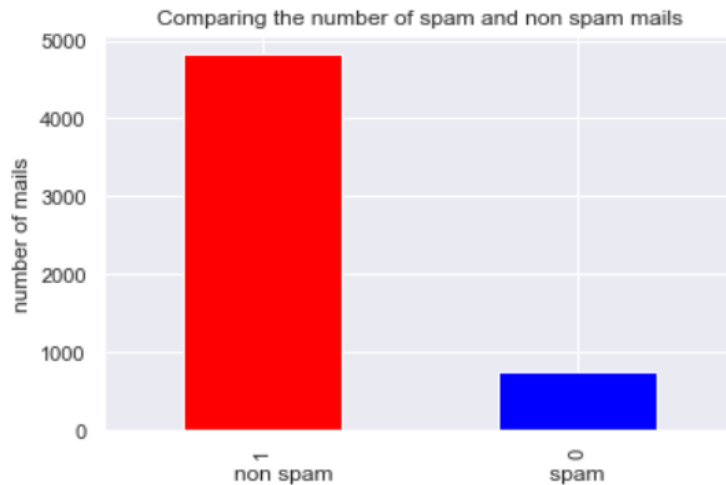


FIG 10: GRAPH OF SPAM AND NON SPAM MAILS

7. CONCLUSION

Thus, through this project a comprehensive analysis of various classifiers was implemented on a common dataset. The results were compared based accuracy, precision, and recall score. Models like SVM are a good example with high accuracy.

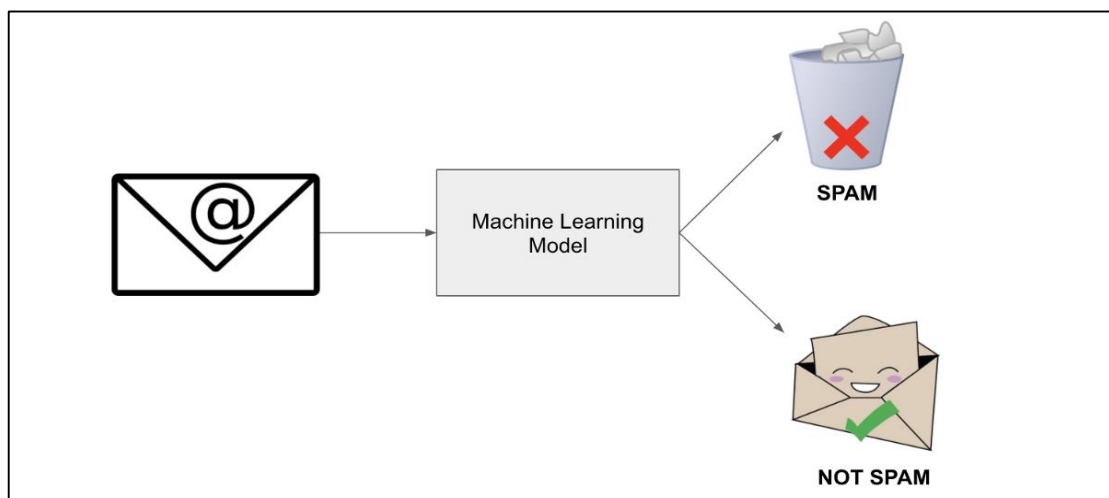


FIG 11: HOW SPAM MAILS ARE DETECTED?

Through various Machine Learning models we can classify mail into spam and non-spam. These algorithms may not be 100% accurate but they still help us a lot in elimination of unwanted or junk mail. There are several other supervised and non-supervised models that we havent covered in this project.

8. REFERENCES

- [1]. Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B. and Shah, T., 2022. Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges. *Security and Communication Networks*, 2022.
- [2]. Aliza, H.Y., Nagary, K.A., Ahmed, E., Puspita, K.M., Rimi, K.A., Khater, A. and Faisal, F., 2022, March. A Comparative Analysis of SMS Spam Detection employing Machine Learning Methods. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 916-922). IEEE.
- [3]. Nandhini, S. and KS, J.M., 2020, February. Performance evaluation of machine learning algorithms for email spam detection. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-4). IEEE.
- [4]. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N. and Al Najada, H., 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), pp.1-24.
- [5]. Trivedi, S.K., 2016, September. A study of machine learning classifiers for spam detection. In *2016 4th international symposium on computational and business intelligence (ISCBI)* (pp. 176-180). IEEE.