SPAM DETECTION & ALGORITHM PREFERENCE MODEL USING MACHINE LEARNING

This project report is submitted to

Rashtrasant Tukadoji Maharaj Nagpur University

In partial fulfillment of the requirement

For the award of the degree

Of

Bachelor Of Engineering in Information Technology

By

AYUSHI UMREDKAR DEWAL ATKARE SHWETA TAYDE

SAKSHI MANDURKAR SHASHANK PAWSEKAR

Under the guidance of

Prof. Alok Chauhan

DEPARTMENT OF INFORMATION TECHNOLOGY

Nagar Yuwak Shikshan Sanstha's

RAJIV GANDHI COLLEGE OF ENGINEERING & RESEARCH, NAGPUR-441110

(An institution affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur)

Session 2021-2022

CERTIFICATE OF APPROVAL

Certified that the project report entitled "**SPAM DETECTION & ALGORITHM PREFERENCE MODEL USING MACHINE LEARNING**" has been successfully completed by **AYUSHI UMREDKAR, DEWAL ATKARE, SAKSHI MANDURKAR, SHASHANK PAWSEKAR, SHWETA TAYDE** under the guidance of **PROF. ALOK CHAUHAN** in recognition to the partial fulfillment for the award of the degree of Bachelor of Engineering in Information Technology, Session 2021-2022, **Rajiv Gandhi College of Engineering & Research, Nagpur** ***(An Institution Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)***

(Signature)
Dr. Sachin Mahakalkar
Dean Academics, RGCER

(Signature)
Prof. Alok Chauhan
Guide, IT Dept.

(Signature)
Dr. Manish Goswami
HOD, IT Dept.

(Signature)
Dr. Manali M. Kshirsagar
Principal, RGCER

DECLARATION

I certify that

a. The work contained in this project has been done by me under the guidance of my supervisor(s).
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in preparing the project report.
d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

1) Ayushi Umredkar

2) Dewal Atkare

3) Sakshi Mandurkar

4) Shashank Pawsekar

5) Shweta Tayde

ACKNOWLEDGEMENT

We are extremely thankful to our guide **Prof. Alok Chauhan**, Prof., Information Technology department, RGCER, under whom our project took the shape of reality from mere idea. we are thankful to our guide for enlightening us with his precious guidance and constant encouragement. we thank our guide for providing us with ample support and valuable time. we are indebted to our guide who constantly provided a stimulus to reach our goals.

We are grateful to **Dr. Manish Goswami**, HOD, Information Technology department, RGCER, for his kind co-operation and timely help.

We express our gratitude towards **Dr. Sachin Mahakalkar**, Dean Academics, RGCER, for his never ending support, planning and motivation.

We express our gratitude towards **Dr. Manali M. Kshirsagar**, Principal RGCER, for her support and motivation.

Lastly, we would like to thank all those who were directly or indirectly related to our project and extended their support to make the project successful.

Name of Projectees

1. Ayushi Umredkar
2. Dewal Atkare
3. Sakshi Mandurkar
4. Shashank Pawsekar
5. Shweta Tayde

CONTENTS

PAGE NO.

LIST OF PUBLICATIONS

Sr. No.	Authors	Title of Paper	Name of International Journals / International Conference	Place and date of Publication
1.	Shashank Pawsekar, Ayushi Umredkar, Dewal Atkare, Sakshi Mandurkar, Shweta Tayde	SPAM DETECTION & ALGORITHM PREFERENCE MODEL USING MACHINE LEARNING	Journal of Machine Learning & Artificial Intelligence Research	UGC CARE, July 2022

List of Figures and Tables

ABSTRACT

SPAM is the Digital Junk Mail – Unsolicited communications sent in bulk over the internet with Rapid Growth of Internet Users. Email spam has steadily grown since the early 1990s and by 2014 was estimated to account for around 90% of total email traffic. Sending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning. Machine learning methods of recent are being used to successfully detect and filter spam emails. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. Our review compares the strengths and drawbacks of existing machine learning approaches and the research problems in spam filtering. We recommended deep leaning and deep adversarial learning as the future techniques that can effectively handle the menace of spam emails. Though there are several email spam filtering methods in existence, we explained below the different categories of spam filtering techniques that have been widely applied to overcome the problem of email spam.

CHAPTER 1

INTRODUCTION:

Overview

SPAM the name comes from a Monty Python Sketch in which the name of the canned pork product Spam is Ubiquitous, Unavoidable and repetitive. In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world. Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam.

Unwanted commercial bulk emails called spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic. Users who receive spam emails that they did not request find it very irritating. It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers.

Problem Statement

Emails have two sub-categories, i.e., Spam and Ham. Unwanted commercial bulk emails called spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic. Users who receive spam emails that they did not request find it very irritating. It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers. Taking control on these types of spam emails by detection and classification with the help of modern techniques should be more focused on accuracy and precision.

Objective

To effectively handle the threat posed by email spams, leading email providers such as Gmail, Yahoo mail and Outlook have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters. These ML techniques have the capacity to learn and identify spam mails and phishing messages by analysing loads of such messages throughout a vast collection of computers. Since machine learning have the capacity to adapt to varying conditions, Gmail and Yahoo mail spam filters do more than just checking junk emails using pre-existing rules. They generate new rules themselves based on what they have learnt as they continue in their spam filtering operation. The machine learning model used by Google have now advanced to the point that it can detect and filter out spam and phishing emails with about 99.9 percent accuracy. The implication of this is that one out of a thousand messages succeed in evading their email spam filter. Statistics from Google revealed that between 50-70 percent of emails that Gmail receives are unsolicited mail. Google's detection models have also incorporated tools called Google Safe Browsing for identifying websites that have malicious URLs. The phishing detection performance of Google have been enhanced by introduction of a system that delay the delivery of some Gmail messages for a while to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analysed collectively. The purpose of delaying the delivery of some of these suspicious emails is to conduct a deeper examination while more messages arrive in due course of time and the algorithms are updated in real time. Only about 0.05 percent of emails are affected by this deliberate delay.

Though there are several email spam filtering methods in existence, the state-of the-art approaches are discussed in this. We explained below the different categories of spam filtering techniques that have been widely applied to overcome the problem of email spam.

Content Based Filtering Technique: Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naive Bayesian classification, this method normally analyses words, the occurrence,

and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spams.

Case Base Spam Filtering Method: Case base or sample base filtering is one of the popular spam filtering methods. Firstly, all emails both non-spam and spam emails are extracted from each user's email using collection model. Subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process. The data is then classified into two vector sets. Lastly, the machine learning algorithm is used to train datasets and test them to decide whether the incoming mails are spam or non-spam.

Heuristic or Rule Based Spam Filtering Technique: This approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message. Several similar patterns increase the score of a message. In contrast, it deducts from the score if any of the patterns did not correspond. Any message's score that surpasses a specific threshold is filtered as spam; else it is counted as valid. While some ranking rules do not change over time, other rules require constant updating to be able to cope effectively with the menace of spammers who continuously introduce new spam messages that can easily escape without been noticed from email filters. A good example of a rule-based spam filter is Spam Assassin.

Previous Likeness Based Spam Filtering Technique: This approach uses memory-based, or instance-based, machine learning methods to classify incoming emails based to their resemblance to stored examples (e.g., training emails).

Adaptive Spam Filtering Technique: The method detects and filters spam by grouping them into different classes. It divides an email corpus into various groups, each group has an emblematic text. A comparison is made between each incoming email and each group, and a percentage of similarity is produced to decide the probable group the email belongs to.

A particular machine learning algorithm is then used to learn the classification rules from these email messages. Examples of such algorithms include Deep Learning, Naïve Bayes, Support Vector Machines, Neural Networks, K-Nearest Neighbour, Rough sets, and Random Forests. The contributions of this work are given as follows:

a. We did a comprehensive evolutionary survey of the most important features of email spam, the evolution and developments. Through this, we highlighted some interesting research gaps and research directions.

b. We discussed the architectures of spam filters and the application of ML techniques to spam filtering process of Gmail, Yahoo mail and Outlook mail. The different components of the email spam filter were vividly discussed.

c. We presented an elaborate study of several techniques applied to email spam filtering and presented a phenomenal review of literatures on spam email filtering over the period (2004–2018).

d. We exposed researchers to some powerful machine learning algorithms that are not yet explored in spam filtering.

e. We stated in clear terms our findings on some open research problems in relation to spam filtering and recommended proactive steps for the development of machine learning techniques to curb future evolving of new variants of spam that might find it easy to evade filters.

CHAPTER 2

Review Of Literature:

In the following subsections, we will highlight some current worldwide statistical observations. Besides, some country- specific metrics will also be discussed. The statistics relating to the adoption of email as a means for communication is quite staggering. As of 2017, there were nearly 5.5 billion email accounts which are actively in use, this number is projected to grow over 5.5 billion in 2019; nearly one third of the population are estimated to use email by the dawn of 2019. As of 2018 approximately 236 billion emails are exchanged daily, of which around 53.5% are just spams. In fact, 2018 saw an average of 14.5 billion spam emails daily. FBI recently reported a loss of USD 12.5 Billion to business email consumers in 2018 incurred by spam emails. The financial loss incurred by the businesses due to this spamming attack may just skyrocket in few years' times, hitting an accumulated figure of around USD 257 Billion from 2012 by the mid of 2020. The estimated yearly damage will be around USD 20.5 Billion. United States has traditionally been the largest source of spam, however, in recent times it is not the case anymore. Though there were legislations such as CAN-SPAM (Controlling the Assault of Non-Solicited Pornography and Marketing Act) to protect the users, it did not have the expected deterrent effect on the spammers. USA houses world's top 70% spam gangs, responsible for coordinated worldwide spamming. Scam watch reports portrays a grim figure in financial losses for Australian consumers due to verities of scam types, primarily carried out through phones and emails in the last three years as portrayed in following table.

As discussed in Table I, the trend is rising each year with digital theft and email spam will only increase due to the increasing acceptance of these media as mentioned in the figures mentioned above. Investment scams basically provide fraudulent but promising business opportunities for exchanging large sums of money, while dating scams harass potential dating partners in digital space. When it comes to delivering malicious software to spread such scams, emails are still the go-to choose for fraudsters. Recent reports indicate that Australian businesses and consumers have already lost about AUD 56,000 due to email fraud during the first two and a half months of 2019.

Since April 2019, Brazil and Russia have easily reached the USA and China (another major country from spam), to produce about 16% and 14% of the total volume of global spam.

Table I

FINANCIAL LOSS INCURRED IN AUSTRALIAN MARKETS DUE TO DIGITAL SCAMS		
Year	Total Loss (AUD)	Losses in Top Three Categories (AUD)
2018	107,025,301	Investment Scams – 38,846,635 Dating & Romance – 24,648,024 False Billing (Fake Invoices) – 5,512,502
2017	90,928,622	Investment Scams – 31,327,476 Dating & Romance – 20,530,578 Other Business & Employment – 5,270,948
2016	83,561,599	Dating & Romance – 25,480,351 Investment Scams – 23,631,338 Advanced Pay & Fee Fraud – 6,499,604

(Source: "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms", 2021 International Conference on Information Networking (ICOIN), 2021)

As we are reviewing spam detection systems which uses Machine Learning (ML) algorithms, it is important to review on the history of ML in the field and the different algorithms that are used in the current context to classify spam. Researchers have pointed out that, content, and the operational mechanisms of the spam emails changes over the time. Therefore, the techniques that are working now may not be useful soon. This phenomenon is identified as the conceptual drift. Machine Learning is the engineering approach formulated to enable computational instruments to act without being programmed explicitly.

In the following section, we will discuss on number of ML techniques, approaches and algorithms and their associated benefits with Supervised, Unsupervised and Semi Supervised Machine Learning Algorithm Approaches.

Supervised Machine Learning Algorithms:

Supervised machine learning algorithms learns from a set of pre-labelled data, with the possible outputs for the corresponding spends have already been given. This

algorithm learns gradually using the labelled data provided and eventually builds up its own probabilistic mapping system to use for new inputs. This technique has two different subtypes called, Regression and classification. This technique is mostly used to generate the outputs which are in categorical nature.

Unsupervised Machine Learning Algorithms:

The name describes, in this technique there are no labelled data or explicit instructions to pre trained the designed model. Therefore, these systems are not provided with a training. In this the analysis is carried out based on the dataset and feature out the common characteristics, structures and features in a group. Then rearrange the output data in different based structure or the pattern. The output data can be organized in different types such as clustering, anomaly detection, association and autoencoders.

Semi-Supervised Machine Learning Algorithms:

Another one approach the system is trained with both labelled and unlabelled data in the testing phase and the system analysis are carried out using both techniques. The main objective of this approach is to achieve better accuracy and precision than the traditional supervised and unsupervised approaches. In the following section we will be focusing on the different machine learning algorithms that have been used in the reviewed studies. These have been analysed after categorizing them under the above discussed machine learning algorithm approach. In a system one or several ML algorithms have been used to achieve the expected performance measurements.

SUPERVISED LEARNING BASED MACHINE LEARNING ALGORITHMS.

Artificial Neutral Network (ANN):

It Has used a system using ANN approach to classify spam and ham emails. This developed model is based on thirteen pre-labelled-fixed email features which are associated with spam emails. ANN is built using artificial neutrons, Hence the name come from. The number of artificial neutrons that are been used in the system can be varied and depend on the requirements of the system. These neurons are connected to different layers such as Input layer, Hidden layers, and output layers. ANN systems ‘learns’ through a process named, ‘BackPropagation’. The produced new output of the

network is compared and matched with the ideal match that should have been produced. The variation is considered and adjust the weights between the neutron connections with many iterations.

Naive Based Machine Learning Algorithm:

It has been developed using the Bayes' rule which tries to derive the probability of an event occurrence based on even related prior knowledge and conditions. This approach is highly scalable, fast and easy to implement into a system. This has been used in the system developed by to provide the solution to the problem independence of random variables with 23 different classification rules. This system uses Decision tree along with Naïve Based to generate the expected outcome. The main drawback of this algorithm is this can be only used if the input features are 'completely independent on each other'. In the practical scenario, this is not always possible.

Support Vector Machine (SVM):

It is another well established and most frequent used Machine learning classification algorithm. Some of the systems have used only SVM as their system classification algorithm while some researchers have used combination of algorithms including SVM. The weights are reflecting the importance of different analysis categories; 'classes. As per the researchers, the advanced weighted SVM algorithm has higher performance metrices. In the SVM algorithm a hyperplane is created generating different classes to analyse various features derived from the dataset. SVM can be adopted into any number of vector dimensions. In the 2D dimension the approach would be a line. In the 3D dimension it would be a hyperplane.

Decision tree (DT):

Decision tree machine learning algorithm is another algorithm that have been used more commonly in the reviewed supervised learning approach studies. The reasons to use this more often are this is an algorithm that can be used easily, easier explanations and visualizations. This can be used with both large and small data sets. Has the ability

to handle both numerical data and the categorical data in the system. In the developed system, they have used DT along with other algorithms in their system. DT has been used in the tier three stage with binomial categorization of spam and ham emails. The model could classify the spam in real time, for this feature DT has provides significant insights as it has simple computational mechanism which is required for efficient real time computational requirements.

UNSUPERVISED LEARNING BASED MACHINE LEARNING ALGORITHMS:

In this section we are focusing on the unsupervised machine learning algorithms that have been used in the reviewed systems. The adoption of unsupervised machine learning algorithms is low compared to supervised machine learning algorithms. There are two algorithms used by the researchers.

K-nearest Neighbour machine learning algorithm (KNN):

This algorithm is effective to use when there is noise in the input dataset. This can be used to generate both classifications and regression outputs for the developed system. The main drawback of this algorithm is it is highly sensitive for the outliers in the data set. Apart from that, computational cost for this algorithm is comparatively higher with regard to other machine learning algorithms. This may be the main reason that this has not been adopted more commonly in the reviewed studies. K- means Clustering machine learning algorithm This algorithm has straightforward implementation mechanism, and the computational cost is comparatively lower than KNN ML algorithm. These are the reasons for this algorithm to be one of the commonly used unsupervised machine learning algorithm in spam classification field. In the K means clustering the data mining process initiates with the first group which is selected randomly. There is a randomly selected centroid for each cluster to begin the process. Repetitive calculations are carried out starting from that centroid to generate the optimized position.

How Gmail, Yahoo and Outlook emails spam filters work

Different spam filtering formulas have been employed by Gmail, Outlook.com and Yahoo Mail to deliver only the valid emails to their users and filter out the illegitimate messages. Conversely, these filters also sometimes erroneously block authentic messages. It has been reported that about 20 percent of authorization-based emails usually fail to get to the inbox of the expected recipient. The email providers have designed various mechanisms for use in email anti-spam filter to curtail the dangers posed by phishing, email-borne malware and ransomware to email users. The mechanisms are used to decide the risk level of each incoming email. Examples of such mechanisms include satisfactory spam limits, sender policy frameworks, whitelists and blacklists, and recipient verification tools. These mechanisms can be used by single or multiple users. When the satisfactory spam thresholds are too low it can lead to more spam evading the spam filter and entering the users' inboxes. Meanwhile having a very high threshold can lead to some important emails being isolated unless the administrator redirects them. This section discusses the operations of Gmail, Yahoo and Outlook emails anti-spam filters

Gmail filter spam

Google's data centers make use of hundreds of rules to determine whether an email is valid or spam. Every one of these rules depicts specific features of a spam and certain statistical value is connected with it, depending on the likelihood that the feature is a spam. The weighted importance of each feature is then used to construct an equation. A test is conducted using the score against a sensitivity threshold decided by each user's spam filter. And consequently, it is classified as a lawful or spam email. Google is said to be using state of the art spam detection machine learning algorithms such as logistic regression and neural networks in its classification of emails. Gmail also use optical character recognition (OCR) to shield Gmail users from image spam. Also, machine-learning algorithms developed to combine and rank large sets of Google search results allow Gmail to link hundreds of factors to improve their spam classification. The evolving nature of spam over time revolves around factors such as domain reputation, links in message headers and others. These can make messages to unexpectedly end up in the spam folder. Spam filtering principally works on the foundation of “filters”

settings that are continuously updated with the emergence of state-of-the-art tools, algorithms, discovery of new spam and the feedback from Gmail users about likely spammers. Many spam filters employ text filters to eradicate hazards posed by spammers depending on the senders and their history.

Yahoo mail filter spam

Yahoo mail is the first free webmail providers in the world with over 320 million users. The email provider has its own spam algorithms that it uses to detect spam messages. The basic methods used by Yahoo to detect spam messages include URL filtering, email content and spam complaints from users. Unlike Gmail, Yahoo filter emails messages by domains and not IP address. Yahoo mail uses combination of techniques to filter out spam messages. It also provides mechanisms that prevent a valid user from being mistaken for a spammer. Examples are ability of the users to troubleshoot SMTP Errors by referring to their SMTP logs. Another one is the complaint feedback loop service that helps a user to maintain a positive reputation with Yahoo. Yahoo whitelisting (internal whitelisting and Return Path Certification) is also provided. Unlike blacklisting, a whitelist blocks by letting the user specify the list of senders to receive mail from. The addresses of such senders are placed on a trusted-users list. Yahoo mail spam filters allows the user to use a combination of whitelist and other spam-fighting feature to reduce the number of valid messages that are erroneously classified as spam. On the other hand, using whitelist alone will make the filter to be very strict and the implication is that any unapproved user would be blocked automatically. Many anti-spam systems use automatic whitelist. In this case, an anonymous sender's email address is checked against a database; if there is no history of spamming, their message is sent to the recipient's inbox, and they are added to the whitelist.

Outlook email spam filter

After Gmail and Yahoo mail, we discussed Outlook from Microsoft in this section and how it handles spam filtering. In 2013, Microsoft changed the name of Hotmail and Windows Live Mail to Outlook.com. Outlook.com was patterned after Microsoft's Metro design language and directly imitates the interface of Microsoft Outlook. Outlook.com is a collection of applications from Microsoft, one of which is Outlook webmail service. Outlook webmail service allows the users to send and receive

emails in their web browser. It allows the users to connect cloud storage services to their account so that when they want to send an email with file attachments, they can select files from not only their computer and OneDrive account but also from Google Drive, Box, and Dropbox account. Moreover, Outlook webmail service also allows users to encrypt their email messages and disallow the recipient from forwarding the email. Whenever a message is encrypted in Outlook.com, it is only the person with the password that will be able to decrypt the message and read it. This is a security measure that guarantees that only the intended recipient is permitted to read the message. The main difference between Outlook.com webmail service and the MS Outlook desktop application is that Outlook desktop application allows you to send and receive emails, via an email server, while Outlook.com is an email server. Outlook.com webmail service on-the-other-hand is for business and professionals who rely on email. Moreover, MS Outlook desktop application is a commercial software that comes along with the Microsoft Office package. It is a computer software program that provides services like email management, address book, notebook, a web browser and a calendar which allows users to plan their programs and arrange upcoming meetings. About 400 million users are using Outlook.com. Statistics shows that their site receives about eight billion emails a day and out of which 30%–35% of those emails are delivered to the users' inboxes. Outlook.com have its own distinctive methods of filtering email spams.

CHAPTER 3

WORK DONE:

3.1 Architecture:

Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails. Mail filters are generally used to manage incoming mails, filter spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware. The workings of email are influence by some basic protocols which include the SMTP (Simple Mail Transfer Protocol).

An email message is made up of two major components which are the header and the body. The header is the area that have broad information about the content of the email. It includes the subject, sender, and receiver. The body is the heart of the email. It can include information that does not have a pre-defined data. Examples include web page, audio, video, analogy data, images, files, and HTML markup. The email header is comprised of fields such as sender's address, the recipient's address, or timestamp which indicate when the message was sent by intermediary servers to the Message Transport Agents (MTAs) that function as an office for organizing mails. The header line usually starts with a “From” and it goes through some modification whenever it moves from one server to another through an in-between server. Headers allow the user to view the route the email passes through, and the time taken by each server to treat the mail. The available information has to pass through some processing before the classifier can make use of it for filtering. Fig. below depicts a mail server architecture and how spam filtering is done.

3.2 Process:

The necessary stages that must be observed in the mining of data from an email message can be categorized into the following:

Dataset:

Since the classification algorithm needs a dataset upon which to perform its functionalities, it is of utmost importance that emails are retrieved with a hundred percent accuracy from their respective servers, irrespective of their domain. This retrieval is done using protocols such as IMAP (Internet Message Access Protocol),

SMTP (Simple Mail Transfer Protocol), POP3 (Post Office Protocol 3) and so on. For this model we have the database contains around 5572 messages.

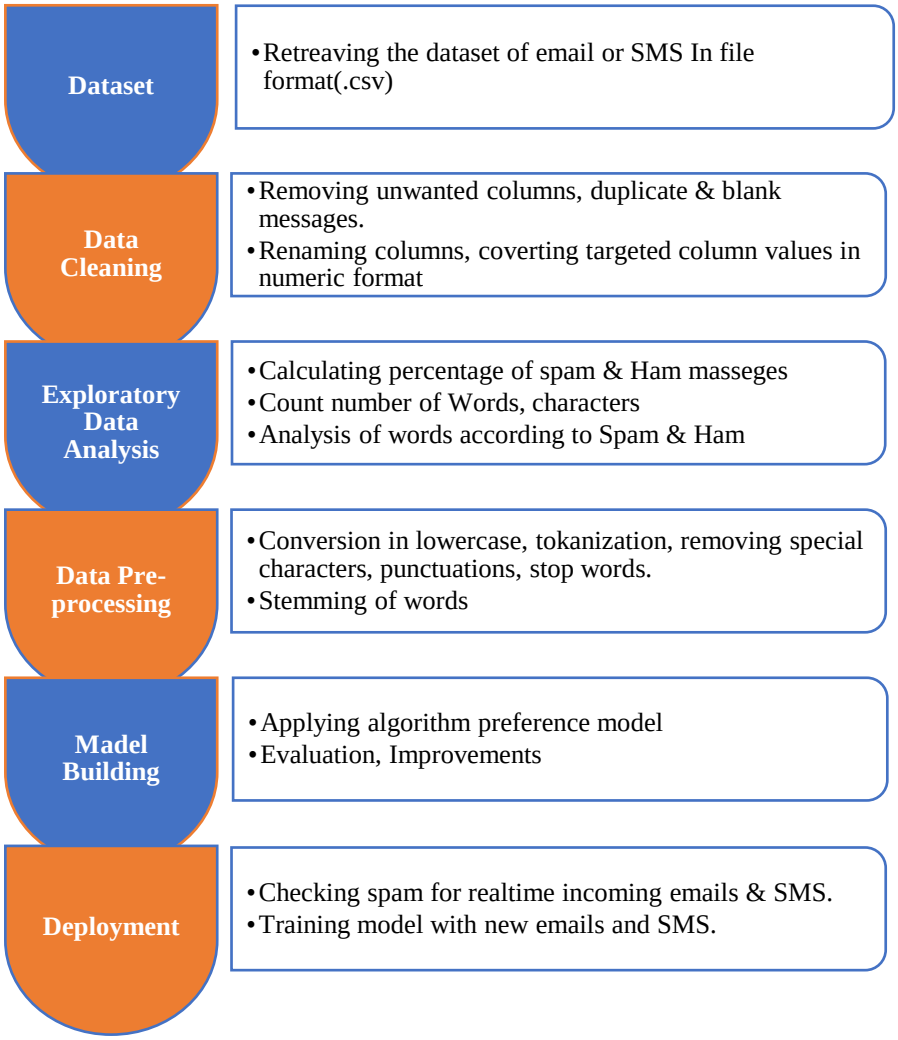


Fig. 1: Flow Diagram

Data Cleaning:

After the data preparation phase, the input parameters are considered upon which the algorithm is to be executed. Data cleansing is also important because it improves your data quality and in doing so, increases overall productivity. When you clean your data, all outdated or incorrect information is gone – leaving you with the highest quality information. For data cleaning performing actions as follows-Remove unwanted

columns, renaming data columns in meaningful way, Convert the data belonging to target in numeric format, Remove missing values and duplicate values.

Exploratory Data Analysis:

Exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling and thereby contrasts traditional hypothesis testing. At first checking the percentage of the ratio of spam and ham emails/SMS, As shown in figure the data is imbalanced, so the further process will vary according that perspective.

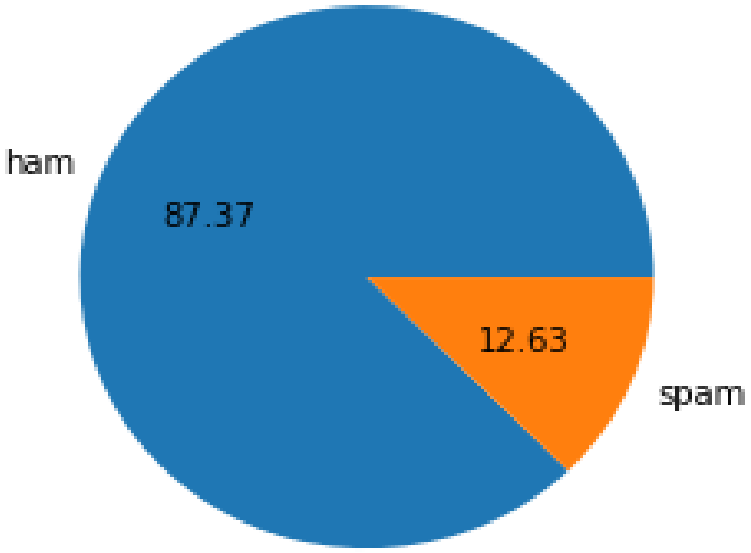


Fig.2: Ratio of Spam & Ham

Now we're counting the number of words, characters and sentences and adding respective columns in our data frame. Describing feature columns Calculating it's mean and overall count.

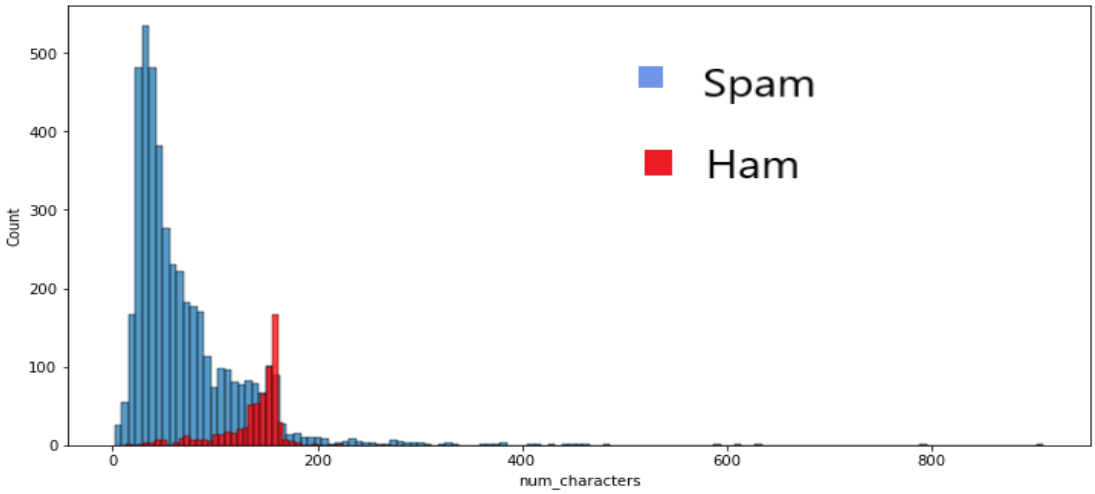


Fig.3: Count of Characters

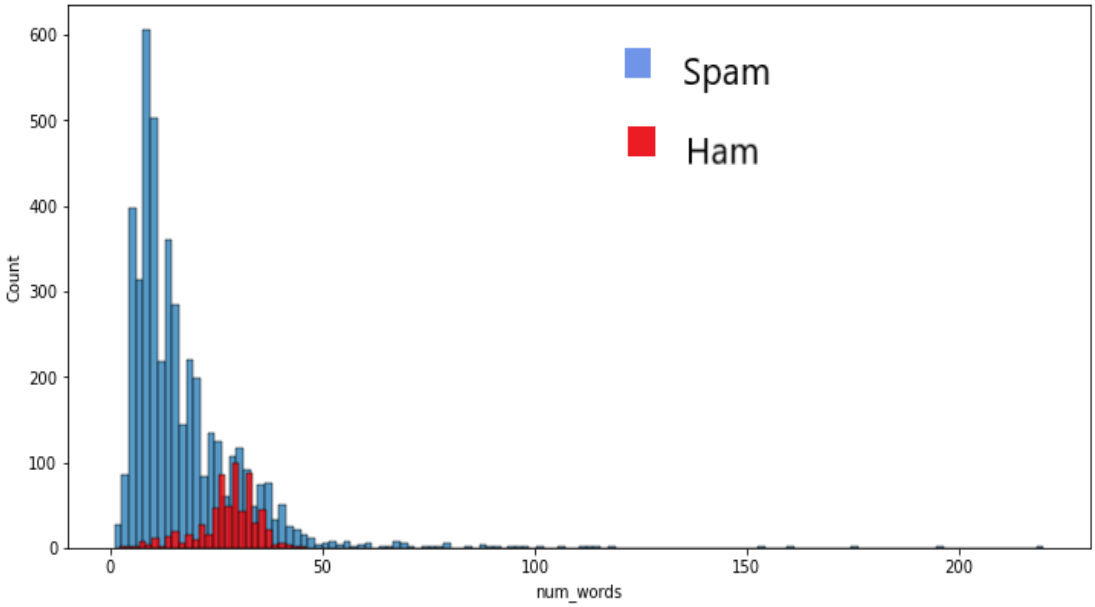


Fig.4: Count of Words

By following graphs for number of words and characters we can consider that spam messages/emails have a greater number of words and characters as compared to Ham messages.

Data Pre-processing:

Converting data that is messages with respect to our need as follows lower case Conversion

• Lower case Conversion

• Tokenization

• Removing special characters

• Removing stop words and punctuation

• Stemming of the words

By scanning the text in each view of the email, Specific keywords are scanned for by parsing the text files to perform broad spam/non-spam classification. The algorithm then reads these parameters to determine the level of priority. The preparation of data thus consists of the retrieval of emails and their scanning. After processing the data, we get the words only format of the messages in the data frame.

For analysis of top used words in spam and not spam messages take a look at word cloud diagrams as follows the bigger size of words shows that these words are majorly used in their respective categories.

Fig.5: Word cloud for Not Spam Words

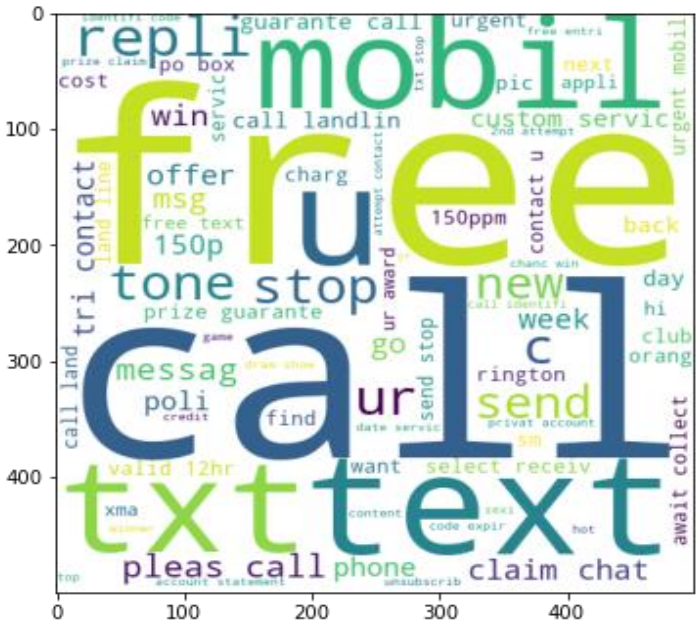


Fig.6: Word cloud Spam Words

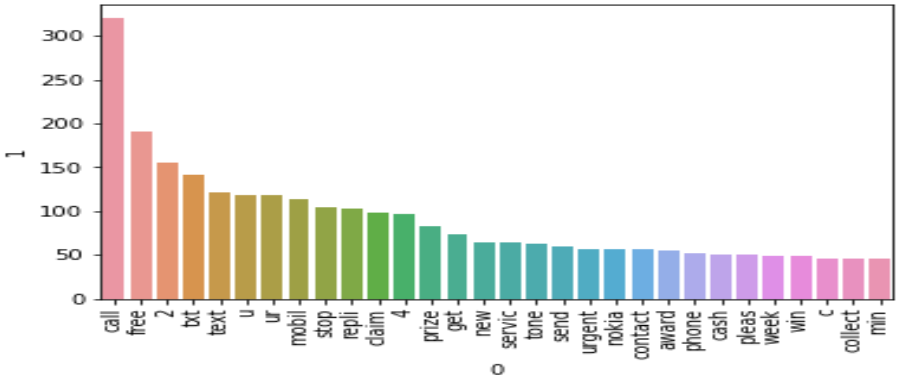


Fig 7: No. of frequent words used in Spam messages

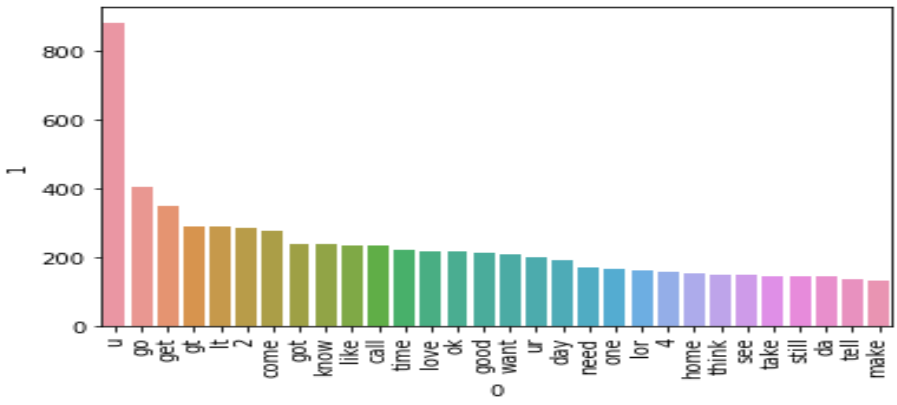


Fig 8: No. of frequent words used in not Spam messages

The above bar graphs show the word counts for each category.

Model Building:

The classification model is 'trained' to recognize similar patterns in future emails and conserve the time that would've been required for further comparisons. The accuracy of the final algorithm is directly proportional to the amount of training it receives. The results obtained after training need to be surveyed for further analysis. This is done primarily to find 'outliers': results that deviate from the norm in enough factors that they warrant being classified differently. This outlier analysis bifurcates the dataset into two types of outputs: ambiguous and proper. The proper output is the one on which further assessment is performed. If the output obtained in a certain case is ambiguous, it is not ideal. The result is debugged, and the algorithm reiterates back to the training phase. Performing algorithms requires numeric input, So we are using TFIDF to convert or vectorize the text. Tfidf Vectorizer uses an in-memory vocabulary to map the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix.

By our previous research conclusion, the naïve bayes algorithm is most Frequent used Classification algorithm, so we are Performing the most used naïve bayes algorithms and some other latest algorithms as follows-

• GaussianNB

As the name suggest, Gaussian Naïve Bayes classifier assumes that the data from each label is drawn from a simple Gaussian distribution. The Scikit-learn provides sklearn.naive_bayes.GaussianNB to implement the Gaussian Naïve Bayes algorithm for classification.

• MultinomialNB

It is another useful Naïve Bayes classifier. It assumes that the features are drawn from a simple Multinomial distribution. The Scikit-learn provides sklearn.naive_bayes.MultinomialNB to implement the Multinomial Naïve Bayes algorithm for classification.

• BernoulliNB

Bernoulli Naïve Bayes is another useful naïve Bayes model. The assumption in this model is that the features binary (0s and 1s) in nature. An application of Bernoulli Naïve Bayes classification is Text classification with 'bag of words' model. The Scikit-learn provides sklearn.naive_bayes.BernoulliNB to implement the Gaussian Naïve Bayes algorithm for classification.

• LogisticRegression

Logistic regression, despite its name, is a classification algorithm rather than regression algorithm. Based on a given set of independent variables, it is used to estimate discrete value (0 or 1, yes/no, true/false). It is also called logit or MaxEnt Classifier.

Basically, it measures the relationship between the categorical dependent variable and one or more independent variables by estimating the probability of occurrence of an event using its logistics function.

sklearn.linear_model.LogisticRegression is the module used to implement logistic regression.

• SVC

Support vector machines (SVMs) are powerful, yet flexible supervised machine learning methods used for classification, regression, and, outliers' detection. SVMs are very efficient in high dimensional spaces and generally are used in classification problems. SVMs are popular and memory efficient because they use a subset of training points in the decision function.

• DecisionTreeClassifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

• KNeighborsClassifier

The K in the name of this classifier represents the k nearest neighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors. The choice of the value of k is dependent on data.

• RandomForestClassifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

• AdaBoostClassifier

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

• BaggingClassifier

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

• ExtraTreesClassifier

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

• GradientBoostingClassifier

GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n_classes_ regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.

• XGBClassifier

The XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting.

Making a group of algorithms so that we can compare and analyse their performance output throughout at a time. There are multiple algorithms we can use for our classification That can help us in different textual input so that the accuracy and precision of the model determine automatically. Hence, we customized our model with the help of new data frame that can grade the algorithm through accuracy and precision with respect to time. By this feature we always have a preference of the algorithm according to our data.

For training the data here using train classifier. train_test_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train_test_split will make random partitions for the two subsets.

Evaluation:

The proper output now needs to be assessed one final time before being classified or filtered. We already have a data set on which the algorithm has been trained. Upon retrieving the user's emails, we also have a real-time result. The dataset results and the real-time results need to be compared and contrasted to optimize accuracy. The algorithm is further retrained with the same parameters. This enables us to analyse the uncertainty of the output obtained and further improvise the level and accuracy of filtering and prioritization. Now we get our results according to the all algorithms and the top performing are naïve bayes, Kneighbors , Random forest & ExtraTrees . For more details refer

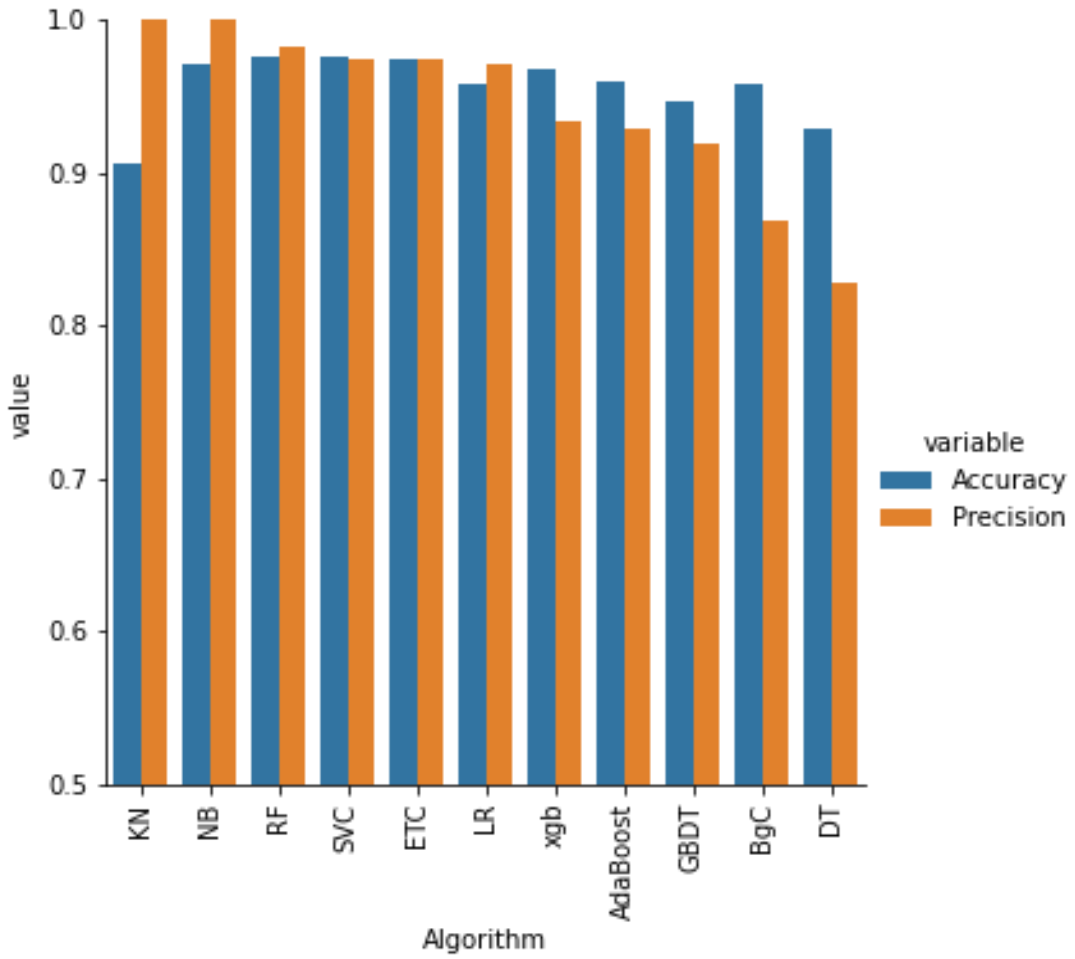


Fig.9: Algorithms with Accuracy and Precision

Improvements:

In TF IDF vectorization there is a feature limitation named as Max features which limits the count of word, ie. All Frequent words will be compared with the range of 3000. Using the voting classifier and stacking with the combination of good performing algorithms with low accuracy VOTING CLASSIFIER - A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. STACKING - Stacking is an ensemble machine learning algorithm that learns how to best combine the predictions from multiple well-performing machine learning models. The scikit-learn library provides a standard implementation of the stacking ensemble in Python.

Web application Deployment:

- For website architecture we are using streamlit library.
- Streamlit is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.
- Importing all the libraries and functions we are created in our model.
- For site structure we are making a heading, text area for writing email or SMS and button called as detect. After clicking button, it gives the result as spam or not spam. A shown in figure 10 below

Fig. 10 Web Application user Interface

Deploy

- Streamlit run webapp.py for launch local host.
- Also, for web deployment we are using heroku server.

CHAPTER 4

RESULT AND DISCUSSION:

4.1 Results:

We will start with a basic overview of the testing approaches used and the different machine learning algorithms used by each of the reviewed paper. Based on the information, we can obtain information on the nature of the approaches of the studies and the distribution of the different machine learning algorithms used to classify spam and ham.

High adoption of Supervised Based:

High adoption of the supervised learning technique can be seen in the distribution with 54 percent of the selected sample. The next most adopted framework is unsupervised approach. The majority of the researchers have adopted Supervised learning technique as their first choice. This clearly signifies that there is high degree of opportunities to expand and the availability for the research in the field of semi supervised and unsupervised Machine learning Approach.

Consistent and Higher accuracy in Supervised Based:

Approach The main objective of majority of the researchers are based on increasing a higher accuracy of spam detection from the developed systems. The accuracy of the systems using supervised learning model has a distribution of accuracy in higher level with tight close range with the minimum variation. This shows that, the outcomes are consistent with higher accuracy (average accuracy is above 90percent for supervised learning model).

From Above results, we can conclude that there is a huge opportunity to develop systems using semi-supervised and unsupervised frameworks for future researchers. Apart from that, the performance evaluation bar for the existing systems for these approaches are very low compared to supervised framework, meaning there is a higher flexibility in achieving higher percentage outcomes.

4.2 Discussions:

Algorithm Preference:

Majority of the developed system have used SVM machine learning algorithm. On our performance basis Naive Biased have given the best performance with highest precision

and accuracy. Some of the studies have used, their own advanced algorithms that have been derived from the main ML algorithm. These have been included in the original ML format for above analysis purpose. These can be further analysed in future studies in this section.

Using Single Algorithm Vs. Multi Algorithm Framework:

As we can see in the table, most of the systems (83 percent) have used a combination of different ML algorithms for their systems in-order to achieve higher results from their study. Out of twelve studies only two systems have used single ML algorithm as their approach. All the other studies have used two or more algorithms combined to achieve higher and better results from their studies.

Email Features Analysed using ML algorithms:

The following table demonstrates the different email features that are analysed in the reviewed articles. As we can see majority of the systems that are developed are focused on analysing for spam using the Body of the email and the BoW. BoW (bag of word) is the approach of analysing the email using key words, phrases and texts in the email. From this we can identify that, there is a higher opportunity to develop spam detection systems analysing the email features such as attachments in the email, structure of the email and spam hyperlinks added in the emails. These research areas have been approached only by a few current researchers leaving these as the higher opportunity areas to explore research areas for the future studies.

Analysing the Email Content:

Almost all the studies that have used in this have developed their systems based on analysing the content such as checking on keywords or phrases that are pre identified to be included in spam emails. This approach is reasonable and logical for a certain state for current context. However, as we have explained in the previous sector, the spammers are using new techniques day by day. Therefore, the spamming techniques are evolving. Hence, this traditional approach would not be adequate in the future to detect spam as spammers will find creative and efficient spamming techniques to go under this detecting method.

CHAPTER 5

SUMMARY & CONCLUSION:

After an in-depth analysis, the study results in several different observation especially in Machine Learning based proposition. High adoption rate for Supervised Machine Learning can be seen throughout the review. This approach is used mainly because it generates higher accuracy results with less variations giving high consistency for this approach. By highlighting the algorithm such as SVM and Navie Bayes are high in demand. We have also come into the conclusion that the potentiality of research into hybrid and multi-algorithm systems is quite promising. "Concept Drift" another important addressing area which would makes a system to perform optimally under gradual modification in spamming techniques. Now talking about current situations, the way of dealing spam emails of phishing nature is not most efficient, thus require more ways of innovative approach that will take into different angles of problem.

We will start with a basic overview of the testing approaches used and the different machine learning algorithms used by each of the reviewed paper. Based on the information in the table 2, we can obtain information on the nature of the approaches of the studies and the distribution of the different machine learning algorithms used to classify spam and ham.

Majority of the developed system have used SVM machine learning algorithm. Some of the studies have used, their own advanced algorithms that have been derived from the main ML algorithm. These have been included in the original ML format for above analysis purpose. These can be further analysed in future studies in this section. In review covers survey of the important concepts, attempts, efficiency and research trends in spam Filtering.

Our process starts with Database according to our classification.

The Database we have is in the raw format, so we are performing some following actions.

Data Cleaning- In this we are Removing unwanted columns, missing/duplicate values and converting data into numeric format.

Exploratory Data Analysis (EDA)- In this we are checking the percentage of ratio of spam & not spam mails and counting number of words, sentences, characters.

Text Pre-processing- Converting the data into Lowercase, done the tokenization, removing special characters, stop words, punctuation, and stemmed words.

Model Building- In this, we are performing the most used and frequent Navie Bayes Algorithm. This Algorithm requires numeric input, so we are using TFIDF to convert the text.

According to all algorithm the top performing results are Navie Bayes, Kneighors, Random Forest, ExtraTrees. But According to Time and Performance required we choose navie bayes algorithm.

CHAPTER 6

REFERENCES:

RAZA, M., Jayasinghe, N. and Muslam, M., 2021. A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms. 2021 International Conference on Information Networking (ICOIN),.

Karim, A., Azam, S., Shanmugam, B. and Kannoorpatti, K., 2020. Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework. IEEE Access, 8, pp.154759-154788.

Reis, J., Correia, A., Murai, F., Veloso, A. and Benevenuto, F., 2019. Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), pp.76-81.

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K. and Alazab, M., 2019. A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, pp.168261-168295.

Al-Rubaie, M. and Chang, J., 2019. Privacy-Preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy, 17(2), pp.49-58.

Smadi, S., Aslam, N. and Zhang, L., 2018. Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. Decision Support Systems, 107, pp.88-102.

International Journal of Recent Trends in Engineering and Research, 2018. PERFORMANCE OF MACHINE LEARNING TECHNIQUES FOR EMAIL SPAM FILTERING. pp.245-248.

Alurkar, A., Ranade, S., Joshi, S., Ranade, S., Sonewar, P., Mahalle, P. and Deshpande, A., 2017. A proposed data science approach for email spam classification using machine learning techniques. 2017 Internet of Things Business Models, Users, and Networks.

Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake News Detection on social media. ACM SIGKDD Explorations Newsletter, 19(1), pp.22-36.

M., H. and H., M., 2017. A Survey of Email Service; Attacks, Security Methods and Protocols. International Journal of Computer Applications, 162(11), pp.31-40.

Kaspersky Lab Reports Significant Increase in Malicious Spam Emails in Q1 2016

https://www.kaspersky.co.in/about/press-releases/2016_kaspersk

y-lab-reports-significant-increase-in-malicious-spam-emails-in-q1-2016

Conroy, N., Rubin, V. and Chen, Y., 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.

Explorations in Computing, 2014. - Spam, Spam, Spam, Mail, and Spam: A machine learning approach to filtering junk mail. pp.164-207.

Zhang, L., Zhu, J. and Yao, T., 2004. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing, 3(4), pp.243-269.

Abbreviations

BVN - Bank Verification Number

CAN-SPAM - Controlling the Assault of Non-Solicited Pornography and Marketing Act

ANN - Artificial Neutral Network

SVM - Support Vector Machine

KNN - K-nearest Neighbour machine learning algorithm

SMTP - Simple Mail Transfer Protocol

MTAs - Message Transport Agents

IMAP - Internet Message Access Protocol

POP3 - Post Office Protocol 3

EDA - Exploratory Data Analysis

SMS - Short Message Service

TFIDF - Term Frequency–Inverse Document Frequency

NB - Naïve Bayes

KN – K Neighbors

DT – Decision Tree

LR - Logistic Regression

RF – Random Forest

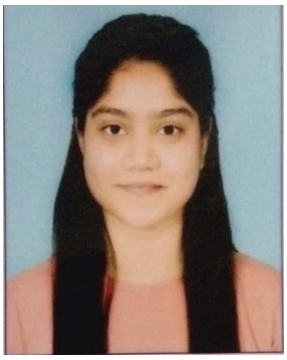
BgC – Bagging Classifier

ETC – Extra Trees Classifier

GBDT – Gradient Boosting Classifier

XGB - eXtreme Gradient Boosting

AUTHOR'S NOTE



Name: Ayushi Ganpati Umredkar

Email Id: ayushiumredkar2000@gmail.com

Permanent Address: Radhika,plot no 101/c,Ramna Maruti nagar,near Nandanwan ,Nagpur.

Contact No: 9405906790, 7498111280

Name: Dewal Manoj Atkare

Email Id: dewalatkare1115@gmail.com

Permanent Address: Plot No 349, Ashirwad Nagar, Hudkeshwar Road, Nagpur 440024

Contact No: 8329155029

Name: Sakshi Vinod Mandurkar

Email Id: sakshimandurkar1227@gmail.com

Permanent Address: Plot No 17, Samata Nagar, Near Gayatri Mandir, Bhandara

Contact No: 8805327929

Name: Shashank Santosh Pawseakr

Email Id: shashankpawsekar@gmail.com

Permanent Address: Plot No. 39, Vanrai Nagar, Besa Road, Manewada, Nagpur - 440034

Contact No: 8208614386

Name: Shweta Raju Tayde

Email Id: taydeshweta2000@gmail.com

Permanent Address: G-004, Greenfield 1,

Hingna road, Wanadongri, Nagpur,441110

Contact No: 9075561297

Guide: Prof. Alok Chauhan

Email Id: alok.chauhan@rgcer.edu.in

Contact No: 7038914679