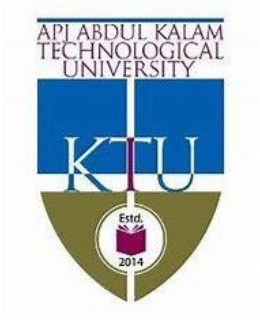
MESSAGE SPAM DETECTION IN CHAT ROOM USING MACHINELEARNING ALGORITHM

MINI PROJECT REPORT

Submitted by

ARAVINDH S	**SBC21CS021**
AROMAL B	**SBC21CS024**
SOORAJ SURESH	**SBC21CS068**
JEFFIN M JOHN	**SBC21CS043**



to

The APJ Abdul Kalam Technological University in partial fulfilment of the requirements for the award of the Degree Of

Bachelor of Technology

In

Computer Science& Engineering

Department of Computer Science& Engineering

SREE BUDDHA COLLEGE OF ENGINEERING

ALAPPUZHA - 690 529

MAY 2024

DECLARATION

We undersigned hereby declare that the Project Report " **MESSAGE SPAM DETECTION IN CHAT ROOM USING MACHINE LEARNING ALGORITHM " ,**submitted for partial fulfilment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me/us under supervision of f Ms. Supriya L P (Assistant Professor, CSE). This submission represents our ideas in our own words and where ideas or word s of others have been included, we have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Pattoor

Date: 07.05.2024

ARAVINDH S SBC21CS021

AROMAL B SBC21CS024

SOORAJ SURESH SBC21CS068

JEFFIN M JOHN SBC21CS043

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
SREE BUDDHA COLLEGE OF ENGINEERING PATTOOR P.O.,
ALAPPUZHA

CERTIFICATE

This is to certify that the seminar/project preliminary/project report entitled "**Message spam detection in chat room using machine learning algorithm"** Submitted by **ARAVINDH S(SBC21CS021),AROMALB(SBC21CS024) ,SOORAJ SURESH (SBC21CS068),JEFFIN M JOHN(SBC21CS043)**to the APJ Abdul Kalam Technological University in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Artificial Intelligence & Machine Learning Engineering is a bonafide record of the seminar/project work carried out by them under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Ms. SUPRIYA L P	**Ms. .DHANYA SREEDHARAN**	**Dr S V ANNLIN JEBA**
Internal Supervisor	**Project Coordinator**	**Head of Department**
(Assistant Professor)	(Associate Professor)	(Head Of Department)
Dept.of CSE	Dept.of CSE	Dept.of CSE

ACKNOWLEDGEMENT

We express our sincere gratitude to our respected Principal **Dr. K Krishnakumar**, for his valuable support and advice.

We express our sincere thanks to **Dr. S V Annlin Jeba**, Head of the Department, Department of Computer Science & Engineering, Sree Buddha College of Engineering Pattoor, for her valuable guidance and suggestions throughout the entire work.

We would like to thank our project coordinator, **Ms Dhanya Sreedharan**, Associate .Professor , Department of Computer Science & Engineering, Sree Buddha College of Engineering, Pattoor for the guidance and advice during the work.

We would like to thank with deep sense of gratitude and obligation to our project supervisor, **Ms. Supriya L P** Assistant .Prof, Department of Computer Science and Engineering, Sree Buddha College of Engineering, Pattoor for her valuable guidance and suggestions throughout the entire project work.

We express our sincere gratitude to all the members of the Department of Computer Science & Engineering, Sree Buddha College of Engineering, for their encouragement and valuable assistance.

In particular, we are thankful to all those who have helped us directly or indirectly in completing this work. Above all we thank Almighty for giving us the health and strength to complete the work on time.

ARAVINDH S
AROMAL B
SOORAJ SURESH
JEFFIN M JOHN

ABSTRACT

ML is termed as Machine Learning ; it is the study of computer algorithms which automatically improves through experience. The usage of mobile phones is growing popular in our everyday life. Messages are viewed as most generally applied correspondence administration which is less costly. In any case, this has prompted an increment in cell phones assaults like message Spam. Here, Naive Bayes algorithm is used in order to differentiate between spam and ham messages. Spam is the unnecessary fraud messages received whereas ham is legitimate message. The algorithm used here is machine learning classification algorithm, and it is implemented here and can be used in differentiating between spam and ham messages with the help of message spam collection data set provided. We train the machine by providing that data set such that it learns from that data and will be able to draw conclusions on its own. Now a days it is very much crucial to identify the spam messages to reduce many frauds happening around the globe. This algorithm can classify with an accuracy of 98 % .

Contents

:

LIST OF FIGURES

LIST OF ABBREVATIONS

ACRONYM	EXPANSION
MI	Machine learning
SMS	Short Message Service
NB	Naive Bayes
NLP	Natural Language Processing
AI	Artificial Intelligence
API	Application Programming Interface
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DFD	Data Flow Diagram
SVM	Support Vector Machine
RF	Random Forests
KNN	K-nearest neighbours
DT	Decision Tree
LR	Logistic Regression
TN	True Negative Rate
TP	True Positive Rate
FN	False Negative

CHAPTER 1
INTRODUCTION

In the developing period of the Internet, individuals are involving increasingly in free online services. Individuals tend to share their data on different sites, though that data is imparted to different organizations that spam individuals to offer their services.

Message Spamming in extremely disappointing for the clients: numerous critical and valuable messages can get lost because of spam messages, Spam messages are additionally used to trap individuals, or bait them into purchasing services. As overall utilization of cell phones has grown, another road for e-junk mail has been opened for notorious advertisers. These publicists use instant messages (SMS) to target probable purchasers with undesirable publicizing known as Message spam. This sort of spam is especially bothersome since, not at all like email spam, numerous PDA clients pay an expense for each SMS got .Building up a classification algorithm that channels Message spam would give a helpful apparatus for mobile phone suppliers.

Since naïve Bayes has been utilized effectively for spam detection, it appears to be expected that it could likewise be used to build Message spam classifier . With respect to email spam, message spam represents extra difficulties for automated channels. SMS texts are regularly restricted to 160 characters, lessening the measure of content that can be utilized to distinguish whether a message is a ham or spam. People have also regularly started using shorthand notations and slang which further makes it difficult to distinguish between ham and spam. We will test how well a simple naïve Bayes classifier manages these difficulties.

Spam messages can be classified as redundant messages sent to large number of people at once. The rise of spam messages are based on the following factors:

1) The accessibility to cheap bulk SMS - plans.

2) Dependability (since the message comes to the cell phone client).

3) Low possibility of accepting reactions from some unaware recipients.

4) The message can be customized.

5) Free services.

1.1 BACKGROUND STUDY

To construct the naïve Bayes classifier , we will use information and data collected from the Message Spam collection which is available openly and consists of about 5574 records .

This dataset incorporates the content of messages alongside a label signifying if the message is a ham or a spam. Junk messages are marked as spam, while true blue messages are marked as ham.

A few cases of ham (Table 1.1) and spam (Table 1.2) are illustrated in the following illustration:

1.HAM MESSAGES

Draft a reasonable one. And I will see if something can happen.
Okay I can try , but cannot commit.
I am good too . Yes weekdays are busy , all thanks to office.

TABLE 1.1 Ham messages

As watched these messages are the everybody messages that individuals trade with each other, these are not junk messages and the client ought to get these messages with the spam filter not separating them through.

2. SPAM MESSAGES

Post Diwali offer ! Get 30 % off + Free Cloudbar with select LED.Buy with your pre-approved loan.
Hi , good credit score makes you eligible for top loans & credit cards. Get your score in 3 minutes.
Want chocolate ? Get a whole-some Chocolate Shake free on orders above Rs.2000 .

TABLE 1.2 Spam messages

Taking a gander at the former specimen messages, we see some recognizing qualities or some repeated patterns of spam messages. One remarkable identification is that two of the three spam messages use the word "free", yet the same word (free) does not show up in any of the ham messages. Then again, two of the ham messages refer to particular days of week, at the point when contrasted with zero junk messages.

Our classifiers will exploit such examples in the word recurrence to decide if the SMS messages appear to better fit the profile of spam or ham. While it's not incomprehensible that "free" would show up outside of a spam SMS, a ham message is probably going to give extra words giving setting.

For example, a ham message may state " are you free on Saturday? " , while a spam message may utilize the expression "free melodies and ringtones . " The classifier will figure the likelihood of spam and ham given the confirmation gave by every one of the words in the message.

We have a total of 5574 records, out of which 4827 messages are ham and 747 messages are spam .

PIE CHART OFHAMV/S SPAM

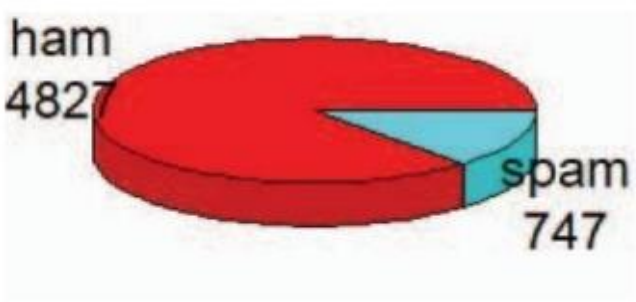


FIG1.1 Ham v/s Spam

1.2 PROBLEM DEFINITION

Spam recognition is considered as a NLP grouping issue utilizing AI calculations. Spam recognition is taken under a grouping issue, during which for a given short instant message, the goal is to characterize as spam or ham.

Since the assertion is to create vigorous and dependable spam discovery model which can decide a given message as spam or ham. Spam email includes a pivotal financial effect in end clients and fix suppliers. The expanding significance of this issue has roused the occasion of various procedures to battle it, or at least giving some alleviation. These channels group the messages in to the class of Spam and Ham (non-Spam). The classifiers choose the classification of approaching message based on certain words in information part and order it. There are two sections, known as test information and preparing information, that fill in as the information base for the Spam classifier to order the messages and pro actives the spam sifting.

As the document classification tasks consists of unproductive data, so selecting most important, required features for improving the accuracy is one of the main objectives.

1.1 PROJECT SCOPE AND OBJECTIVES

The objective of this project is to classify and make analysis of spam and non-spam (ham) through using and utilize flash as it could be a web benefit advancement micro framework in python to form an API, such as multilayer perceptron and comparison of it with naive bayesian classifier.

The aim of this work is to concentrate on different classification techniques and to compare their performances on the domain of spam messages detection. A number of pre-classified SMS Spam detection messages were processed with the techniques to see which one is most successful and under which set of features .

1.2 APPLICATIONS

- It reduce Network Resource Costs.
- It reduces IT Administration Costs.
- It reduces Legal Liability Risk.
- It increases Security and Control.

1.3 LIMITATION OF WORK

The limitation of this project are:

1. This project can detect and calculate the accuracy of spam messages only.
2. It focus on filtering, analyzing, and classifying the messages.
3. Do not block the messages.

CHAPTER 2

LITERATURE SURVEY

[2.1] M. Rubin Julis et al proposed system to detection of spam in SMS using machine learning.

This system divided into some iterations. Every iteration has four phases: Inception, elaboration, construction and transition. In iteration it identify the idea of work. In elaboration International Journal of Scientific and Research Publications, Volume 11, Issue 7, July 2021 79 ISSN 2250-3153 This publication is licensed under Creative Commons Attribution CC BY. http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11510 www.ijsrp.org it design the architectural part. In construction implementation of code is done. And in transition validation of the developed part of system is done. They uses various algorithms such as logistic regression algorithm in that they use logistic function for calculate relation between the categorical dependent variable and independent variable. For training and testing data they uses Knearest neighbours algorithms. They also uses different classifiers that are Naïve Bayes Classifier , decision tree classifiers, support vector machines and compares all. It checks accuracy for all and support vector machines gives 98% accuracy which is good as compares to others**.**

[2.2] Heena Tamboli, Sambhaji Sarode et al proposed An Effective Spam and ham word Classification Using Naïve Bayes Classifier .

In this project they discuss about classification of emails to identify the spam and ham mails. For this purpose they are using Naïve Bayesian Classifier. Which is best technique to classify two objects in datasets. In that first it count of spam and ham words in dataset and print as target. In their dataset target are 33716. Then it shows separate count of spam and ham words for training that is in their data set spam words are 16545 and ham words

are 17071.Then it find Ground Truth value for each word that is that word is ham or spam. The spam messages are the messages which user don't need but that are receiving daily. Spam emails are message of anything it may any advertisement or may be any URL , or any king of virus. Naive Bayes classifier is very good technique for filtering spam emails. It is used to classify any two objects in datasets. Performance of Naïve Bayes algorithm is based on datasets that used. This method on any datasets for classification of any two objects. They filter email spam from dataset and calculate accuracy ,it gives 93% accuracy.

[2.3] Aditya Gupta et al proposed spam filter using Naive Bayesian Technique.

For identify spam emails they used supervised learning method. It identifies spam and non spam emails after receiving massages. Spam filter is used to find unwanted emails and prevents messages from reaching users inbox. Different python libraries are used that are NLTK, Word Cloud, Panda, Matplotlib for filtering emails and finds frequently used keywords. For processing NLTK libraries used ,for visualization Word Cloud and Matplotlib used and for loading data they used Pandas. They used dataset from Kaggle contains 5572 test cases of ham and spam messages. Data is split into trained and test dataset for testing the model **.**

[2.4] : Elly Firasari et al proposedComparation of K-Nearest Neighbour (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance.

In this case, the researcher uses data mining classification calculation by comparing 2 calculation methods, namely K-NN and Naive Bayes. The researchers use Rapid miner tools. The research stages are identification of problems, data collection,implementation K-NN, Implementation Naive bayes, data testing process to produce accuracy and compare the result. The results obtained are the accuracy of Naive Bayes higher than K-NN, namely Naive Bayes 89.04% and K-NN 87.67%. This figure is classified in the category of good classification. From the results of the study it can be concluded the Naive Bayes algorithm is suitable to be applied in the calculation of recipients of social assistance**.**

[2.5] Satyam Sagar et al proposed spam classification filter using Naïve Bayes classifier which is developed as a web application for classification of emails into spam and ham .

They use Python's Micro Flask Framework for developing web application in which input is new incoming emails and it predict output as spam and non spam emails. Their system contains main two parts first one is train the classifier and another is deploy the model. In train classifier it contains dataset of spam and ham emails and it generate classification model. In second part it deploy the model on server

2.1 EXISTING SYSTEM

Existing system was developed using RNN and the model detected spam messages using the given Dataset as input . The model is less accurate to find the exact spam and ham messages. Using the dataset as input we cannot find real time spam and ham messages using the existing RNN model.

DISADVANTAGES :

- The vanishing or exploding gradient problem .
- It cannot be stacked up .
- Slow and Complex training procedures .
- Difficult to process longer sequences .

2.2 PROPOSED SYSTEM

Our proposed system is developed using M l algorithm called Naïve Bayes Algorithm and our model is high accurate than the existing model and we can detect real-time messages sent by a user in a chat room. Our model is build using a dataset that contains spam and ham messages. Detection is done by the ML model using the dataset, that is when an user messages in the chat room the model classifies the message with the dataset and returns the output whether the entered message is spam or ham.

ADVANTAGES :

- It provides high security .
- It requires a small amount of training data to estimate the test data . So, the training period is less.
- It is easy to implement .

2.3 REQUIREMENT ANALYSIS

System requirement is needed in order to accomplish the project goals and objectives and to assist in development of the project that involves the usage of hardware and software. Each of these requirements is related to each other to make sure that system can be done smoothly.

2.3.1 SOFTWARE REQUIREMENTS

Operating system : Windows 10 .

Coding Language : Python 3.8/above,ML,HTML.etc…

Google chrome : Used to run web based system .

Microsoft Word : Creating and editing report .

Kaggle : Get dataset .

WinZip : Extract the data .

APIs : Numpy, Pandas ,Seaborn, Matplotlib etc..

2.3.1 HARDWARE REQUIREMENTS

Processor : intel core i3 or higher .

RAM : 4 GB or higher.

Hard Disk Drive : 20 GB (free).

Peripheral Devices : Monitor, Mouse and Keyboard.

It is better to use high performance processor to avoid any problem while doing this project. Machine learning project required a high speed processor for a better performance to train a large amount of data.

CHAPTER 3

SYSTEM STUDY

3.1 FEASIBILITY STUDY

The feasibility of the system has been studied from the various aspects like whether the system is feasible technically, operationally and economically. The present technology is found to be sufficient to meet the requirements of the system . This system is believed to work well when it is developed and installed. Hence , operational feasibility is achieved . Since the requirements for the project are easily available, ware headed with the intention to use the available resources to fulfill the system requirement . The detailed feasibility study is mentioned below:

3.1.1 TECHNICAL FEASIBILITY

The technology needed for the proposed system that we are going to develop is available. We can work for the project is done with current equipment existing tools like python. We can develop our system still using this technology if needed to upgrade. In future, if we want to use new technology like android app of our system it is possible. Hence, the system that we are going to develop will successfully satisfy the needs of the system for technical feasibility.

3.1.2 ECONOMIC FEASIBILITY

Since the system is developed as a part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it gives an indication that the system is economically possible for development. Economic justification is generally the “Bottom Line” consideration for most systems. The cost to

conduct a full system investigation is negotiable because required information is collected from internet. We can run our system in our normal hardware like desktop, laptop, mobiles and so on. This system won't require extra specific software to use it. Hence, the project that we are going to develop won't require enormous amount of Money to be developed so it will be economically feasible.

3.1.3 OPERATIONAL FEASIBILITY

The user interface will be user friendly and no training will be required to use the application. The solution proposed for our project is operationally workable and most likely convenient to solve the irrelevant document and fraud messages.

3.1.4 SCHEDULED FEASIBILITY

The project schedule took 11-12 weeks of time for study and analysis, data collection, implementation, testing and documentation.

CHAPTER 4

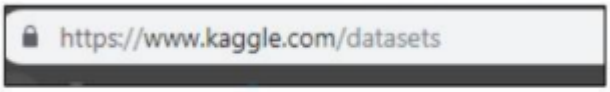
METHODOLOGY

This chapter will explain the specific details on the methodology being used in order to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering. So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself. This project is developed by using Python Language. It contains important function for preprocessing the dataset. Then, the dataset is going to be used to train and test either the model of the machine learning achieve the objectives of the project .

4.1 FRAMEWORK

4.1.1 DATA SOURCE

Collecting data is utterly difficult due to numerous constraints for instances the volume of data and the throughput required for proper and timely ingestion. The dataset that we've used in this project is the real existing data that can be downloaded from machine learning data repository site. There isone website that we've visit to get the dataset to be used in this project .



4.1.2 DATA SETS

The Message Spam Collection is a set of SMS tagged messages that have been collected for message Spam research . It contains one set of messages in English of 5,574 messages, tagged according being ham (legitimate) or spam .

The files contain one message per line . Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

4.1.3 FEATURE SELECTION

Feature selection is a very important task for the message Spam detection . Selected features should be correlated to the message type such that accuracy for detection of spam message can be increased. There is a length limit for message and it contains only text (i.e. no file attachments, graphics, etc.) while in the email, there is no text limit and it contains attachments, graphics, etc. message is usually of two types i.e. ham(legitimate) message and spam message. Spam and ham messages can be differentiated using various features. Identification of good feature that can efficiently filter spam messages is a challenging task. Moreover, we study the characteristics of spam messages in depth and find some features, which are useful in the efficient detection of spam messages . The features that we have extracted and evaluated for our proposed approach are summarized as follows :

- Presence of Mathematical Symbols - Spammers usually uses mathematical symbols for creating spam messages. For example, symbol + can be used for free service messages. Mathematical symbols that we have considered in our experiment are +,−, < , > , / and ^. The first feature is defined as S1 which could be 1 if any mathematical symbol is present in the message.

$$S1 = \begin{cases} \mathbf{1} & \text{MATHEMATICAL SYMBOL} \\ 0 & \text{NON MATHEMATICAL SYMBOL} \end{cases}$$

• Presence of URLs - We consider the presence of URLs as a feature since harmful spam message contains URLs and asks the user to visit those URLs to provide their personal details, debit/credit card information, password or to download some file (file containing the virus). The second feature S2 which could be 1 if any URL (http or www) is present in the message.

$$S2 = \begin{cases} 1 & \text{URL is present} \\ 0 & \text{No URL} \end{cases}$$

• Presence of Dots - The presence of dots seems to be good indicator for legitimate messages because people use dots while chatting. Moreover, People often use dots to separate the sentence, or words so that it becomes easy for the receiver to read the message. We define the presence of dots as feature S3, which will be 1 if the message contains dots.

$$S3 = \begin{cases} 1 & \text{Dot is present} \\ 0 & \text{No Dot} \end{cases}$$

• Presence of special symbols - The presence of special symbols usually refers to spam messages because spammers use special symbols for various reasons. E.g. special symbol "$" is being used to represent money in the dollar in fake award messages, similarly symbol "!" is used to the special attention of user like CONGRATULATIONS! WINNER!, etc. Special symbols that we have used in our approach are !, *, &, # and *. The fourth feature is defined as S4 which would be 1if any special symbol is present.

$$S4 = \begin{cases} 1 & \text{special symbol} \\ 0 & \text{No special symbol} \end{cases}$$

• Presence of emotions - The presence of emotion symbols seems to be a good indicator for legitimate messages because a person usually uses emotions while chatting. For example emotion :) is used for happy face, emotion :(is used for sad face, emotion -_- is used for angry face, etc. Emotion symbols that we have considered for our experiment are :), :(, -_-, :p, :v, :*, :o, B-) and :'(. We define the presence of emotions as feature S5.

$$S5 = \begin{cases} 1 & \text{Emotions} \\ 0 & \text{No Emotions} \end{cases}$$

• Lowercased words - Checks if the message contains lowercased words or not as all lowercased words in a message can be used to seek user's attention. The presence of lowercased words is given by feature S6 as:-

$$S6 = \begin{cases} 1 & \text{Lowercased words} \\ 0 & \text{No Lowercased words} \end{cases}$$

• Uppercased words - We consider the presence of uppercased words as a feature as spammers usually use uppercased words to seek user's attention. For example, WON, PRIZE, FREE, RINGTONE, ATTENTION, etc. The seventh feature is S7 given by rule:

$$S7 = \begin{cases} 1 & \text{Uppercased words} \\ 0 & \text{No Uppercased words} \end{cases}$$

• Presences of Mobile Number - We consider the presence of mobile number as a feature in order to identify spam messages because spammers usually give mobile number in a message. They ask the users to dial on the given number and when user dials on the given number, attacker on the other side ask for user's personal details, bank details, etc. For example, "you have won a $2,000 price! To claim, call 09050000301"we define the presence of mobile number as feature S8.

$$S8 = \begin{cases} 1 & \text{Mobile number} \\ 0 & \text{No Mobile number} \end{cases}$$

• Keyword specific - The presence of suspicious keywords like send, ringtone, free, accident, awards, dating, won, service, lottery, mins, video, visit, delivery, cash, Congrats, Please, claim, Prize, delivery, etc. are considered as spam keywords because these keywords are generally used to attract users in spam messages. We define ninth feature as S9 which will be 1 if message contains spam keywords otherwise it will be 0.

$$S9 = \begin{cases} 1 & \text{Presence of spam keywords} \\ 0 & \text{No spam keywords} \end{cases}$$

• Message Length - It includes the total length of the message including space, symbols, special characters, smileys, etc. The text limit of SMS messages is 160 characters only. We define tenth feature as S10 which counts the total length of each message.

FIG 4.1 shows that how each feature value is extracted from ham and spam messages.

Feature type	**Have you finished work Yet?** **(Ham message)**	**CONGRATULATIONS! Nokia 3650 video camera phone is your call 09066382422 calls cost 150 ppm ave call 3 min vary from mobiles 16 + close 300603 post BCM4284 Ldn WC1N3XX** **(Spam message)**
Presence of Mathematical symbols	0	1
Presence of URLs	0	0
Presence of dots	0	0
Presence of special Symbols	0	0
Presence of emotions	1	0
Lowercased words	1	1
Uppercased words	0	1
Presence of mobile number	0	1
Keyword specific	0	1
Message length	30	157

FIG 4.1: message feature value for ham and spam messages

4.1.4 PYTHON LIBRARY

Here are some details about some Libraries used in this project :

- flask is for creating the application server and pages.
- matplotlib, plotly, plotly-express are for data visualization.
- python- dotenv is a package for managing environment variables such as API keys and other configuration values.
- nltk is for natural language operations.
- numpy is for arrays computation.
- pandas is for manipulating and wrangling structured data..
- scikit-learn is a machine learning toolkit.
- word cloud is used to create word cloud images from text.

4.2 REQUIREMENT COLLECTION AND SPECIFICATION

The primary data were collected through our supervisor from the related corpus. The secondary data were collected from research papers and some test data were sampled data made by ourselves to check for the expected output.

- Corpus is maintained
- The user should be able to receive the genuine messages.

4.3 FUNCTIONAL REQUIREMENTS

The main function of this project is to classify the messages which is done by first taking out the feature vector extraction which involves first taking out whether the word is a spam or not by representing in the form of matrix.

4.4 NON FUNCTIONAL REQUIREMENTS

- Ensures high availability of message data from datasets.
- User should get the result as fast as possible.
- It should be easy to use i.e., user is just required to type the words and click then the result is displayed or user is just required to enter a pair of reasonable sentence.

Spam prevention is often neglected, although some simple measures can dramatically reduce the amount of spam that reaches your chat room. Before they are able to send you spam, spammer's obviously first need to obtain your username, which they can do through different routes. Our project is design to detect spam or unwanted messages.

4.5 DOCUMENT PREPROCESSING

4.5.1 TOKENIZATION

Tokenization is the process of breaking a stream of text up into words , phrases , symbols ,or other meaningful elements called tokens . The list of tokens becomes input for further processing such as parsing or text mining.

Typically , tokenization occurs at the word level. However, it is sometimes difficult to define what is meant by a "word". Often a tokenizer relies on simple heuristics, for example:

- All contiguous strings of alphabetic characters are part of one token; likewise with numbers.
- Tokens are separated by whitespace characters, such as a space or line break, or by punctuation characters.
- Punctuation and whitespace may or may not be included in the resulting list of tokens.

A token is an instance of a sequence of characters.

4.5.2 LEMMATIZATION

Lemmatization in linguistics, is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. In computational linguistics, lemmatization is the algorithmic process of determining the lemma for a given word. Since the process may involve complex tasks such as understanding context and determining the part of speech of a word in a sentence. It can be a hard task to implement a lemmatizer for a new language.

Lemmatization is also used in natural language processing and many other fields that deal with linguistics in general . It also provides a productive way to generate generic keywords for search engines or labels for concept maps.

Lemmatization is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. However , stemmers are typically easier to implement and run faster, and the reduced accuracy may not matter for some applications.

4.5.3 REMOVAL OF STOP WORD

Sometimes , the extremely common word which would appear to be of very little value in helping select documents matching user need are excluded from the vocabulary entirely.

These words are called stop words and the technique is called stop removal. The general strategy for determining a stop list is to sort the terms by collection frequency and then to make the most frequently terms , as a stop list , the members of which are discarded during indexing.

Some of the examples of stop-word are: a, an, the, and, are, as, at, be, for, from, has, he, in, is, it, its, of, on, that, to, was, were, will, with etc.

4.5.4 DOCUMENT REPRESENTATION

In order to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector which describes the contents of the document . A definition of a document is that it is made of a joint membership of terms which have various patterns of occurrence.

Vector spam model or term vector model is the popular algebraic model for representing text documents as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Using vector space model documents are represented using term frequency (tf) , inverse document frequency (idf) or tf-idf weighting scheme.

4.6 RELATED WORK

Most research has been conducted into detecting and filtering message spam using a variety of techniques.

Thiago S. Guzella et. Al (2009) has conducted “A Review of Machine Learning Approaches to Spam Filtering”. In their paper, they found that Bayesian Filters that are used to filter spam required a long training period for it to learn before it can completely well function.

S. Ananthi (2009) has conducted a research on “Spam Filtering using K-NN”. In this paper, she used KNN as it is one of the simplest algorithm. An object is classified by a majority vote of its neighbours where the class is typically small.

Anjali Sharma et. Al (2014) has conducted “A Survey on Spam Detection Techniques”. In this paper, they found that Artificial Neural Network (ANN) must be trained first to categorize emails into spam or non-spam starting from the particular data sets.

However, machine learning technique is chosen to overcome the disadvantages of other methods. For example, real-time black hole list techniques can generate false positive result while Bayesian filters requires a training period before it starts working well.

CHAPTER 5

SYSTEM STUDY

5.1 FLOWCHART

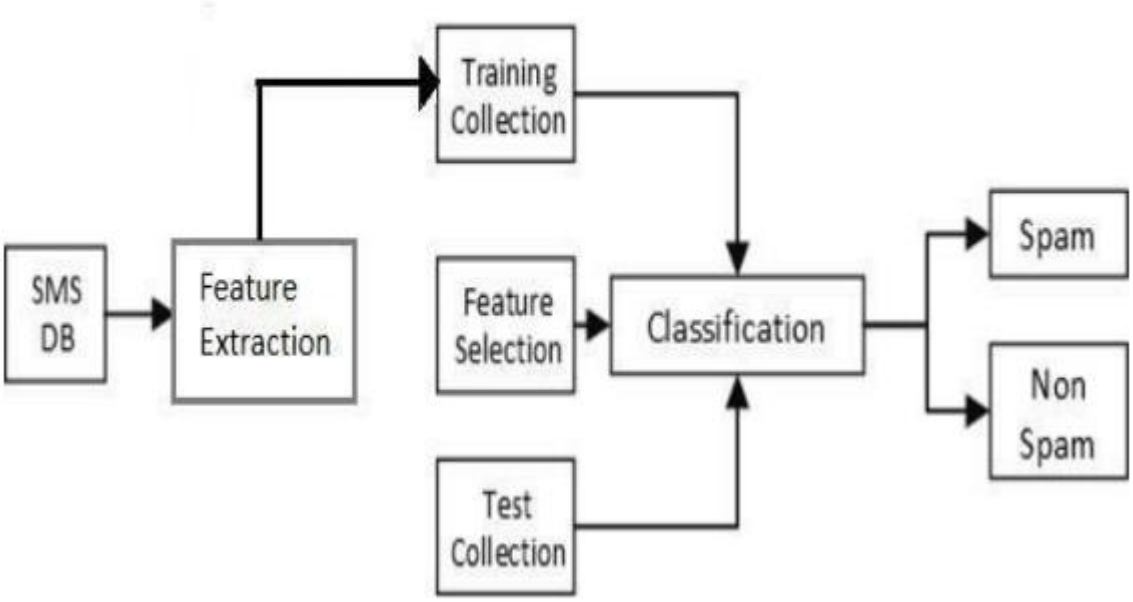


Fig 5.1 : flowchart of Message spam detection

5.2 DATA FLOW DIAGRAM

It is a directed graph where nodes represent processing activity and are represent data items transmitted between processing nodes.

LEVEL 0

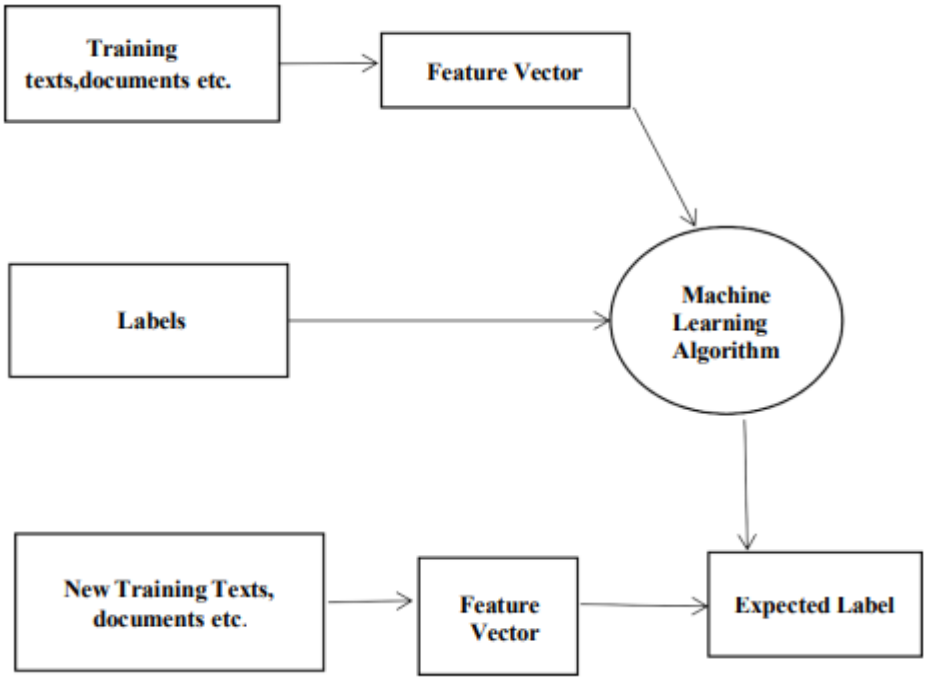


Fig 5.2 : DFD - level 0

LEVEL 1

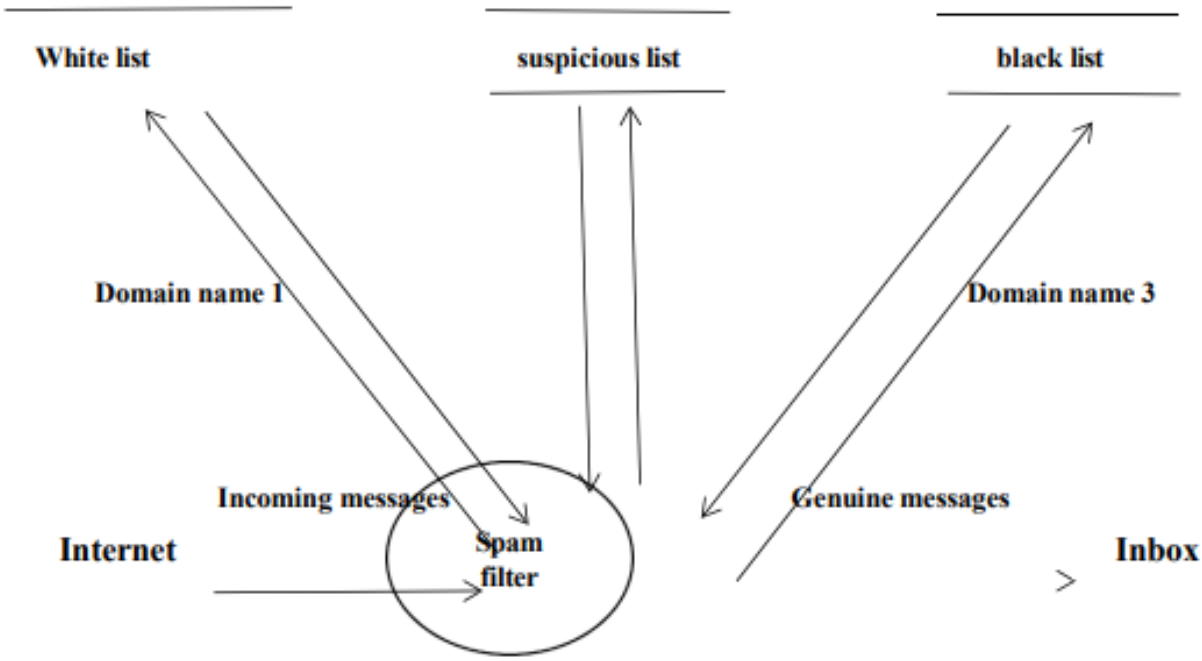


Fig 5.3 : DFD - level 1

LEVEL 2

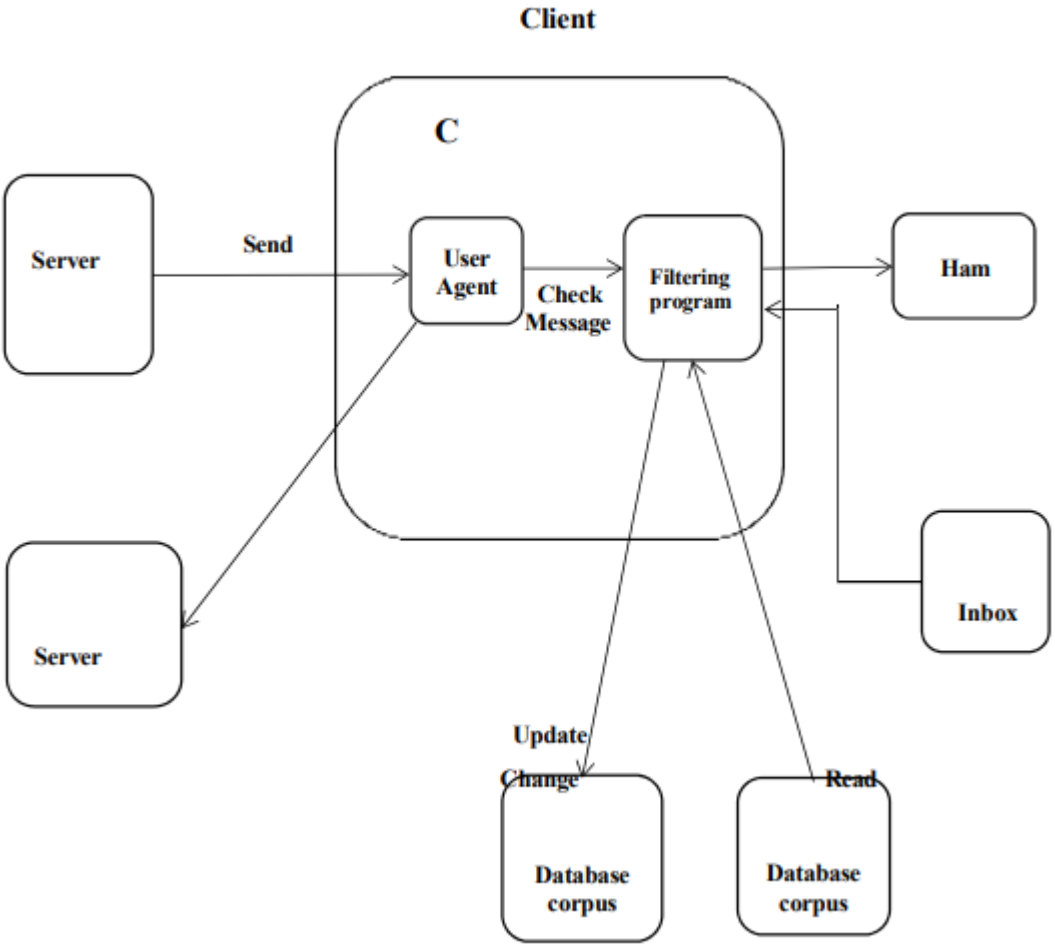


Fig 5.4 : DFD - level 2

5.3 SPAM FILTER ALGORITHM STEPS

- Handle Data : Load the corpus file and split it into training and test datasets .
- Summarize Data : summarize the properties in the training dataset so that we can calculate probabilities and make predictions .
- Make a Prediction : Use the summaries of the dataset to generate a single prediction .
- Make predictions : Generate predictions gives a test dataset and a summarized training dataset .
- Evaluate Accuracy : Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made .
- Tie it together : Use all of the code elements to present a complete and standalone implementation of the Naive Bayes algorithm .

5.4 VISUALIZATION USING WORDCLOUDS

Word Cloud is an approach to outwardly delineate the recurrence at which words show up in information . The cloud is comprised of words scattered fairly haphazardly around the figure.

Words seeming all the more regularly in the content are appeared in a bigger text style, while less normal terms are appeared in littler textual styles. This sort of figure has developed in fame as of late since it gives an approach to watch trending activities on social networking sites.

We compare the word clouds of ham and spam messages and see the difference between the frequently occurring terms in both the datasets.

Fig 5.5 : Wordcloud for Spam

As we observe the most frequent occurring terms in the spam messages are call, free, text, reply, claim etc . These are the words that we generally encounter in spam messages.

Contrasting the spam word cloud (Fig 5.5) and the ham word cloud (Fig 5.6) will give us a thought regarding the catchphrases that will be utilized by our classifiers in separating ham and spam. On the off chance that words present in the spam cloud likewise show up as often as possible in the ham cloud, our classifier would not have solid watchwords for correlation, while if the outcomes are distinctive, the models will have the capacity to separate among ham and spam well.

Fig 5.6 : Wordcloud for Ham

As we observe , with the words occurring in the ham word cloud being completely different from the spam word cloud . This difference suggests that our classifiers will have strong keywords to differentiate between ham and spam.

5.5 MACHINE LEARNING

Machine learning is a method of data analysis that automates analytical model building . Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is broadly categorized into two categories :

a) Supervised Learning

b) Unsupervised Learning

Main categories of machine learning : supervised learning : supervised learning also known as predictive modeling , is the process of making predictions using data .

Examples of Supervised learning are Classification and Regression . A Supervised learning Training data set is pre labelled for classification problems or function values are known in case of regression . After training is done and the model has a minimum cost function for the training data set, later switch for scoring where we can predict values for new data.

Classification: It identifies group membership. That means that if we have multiple events characterized by input parameters, which can be labelled differently, and we want our system to predict which label should be used .

Regression: Regression is a combination of multi-dimensional power supply and function interpolation. The regression problem is used to find the approximation of the function with a minimum error deviationor a cost function. In other words, the regression technique simply tries to predict numeric dependence , a function value, for example, of a data set .

Figure 5.7. Diagrammatically shows how supervised learning is to solve problems.

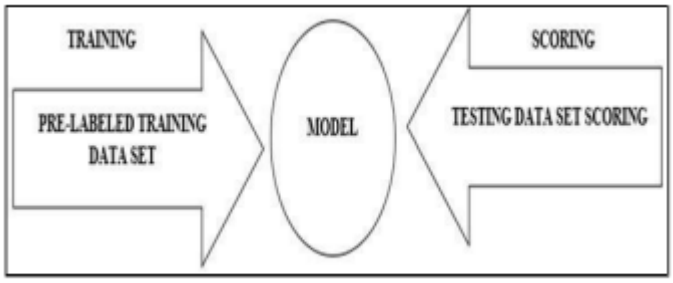


Fig 5.7 : Supervised Learning

Example of supervised learning , if a system has a data set which is a series of messages , supervised learning task is to predict whether each message is spam or non-spam(ham) . This is supervised learning because there is a specific outcome namely spam or ham. MESSAGE SPAM DETECTION IN CHAT ROOM USING MACHINE LEARNING ALGORITHM

Unsupervised Learning: Unsupervised Learning is the process of extracting structure from data or how to best represent data. Examples of Unsupervised Learning are Clustering (is partitioning a data set into meaningful similar sub classes called cluster) and Association (method for discovering relations between existing attributes within a data set or data base). In an un supervised learning situation, where the algorithm detects data features automatically, this depends on the purpose of the algorithm as well as the assumptions made on what the properties and observed values are.

Figure 5. 8. Describes how unsupervised learning is used to solve problems.

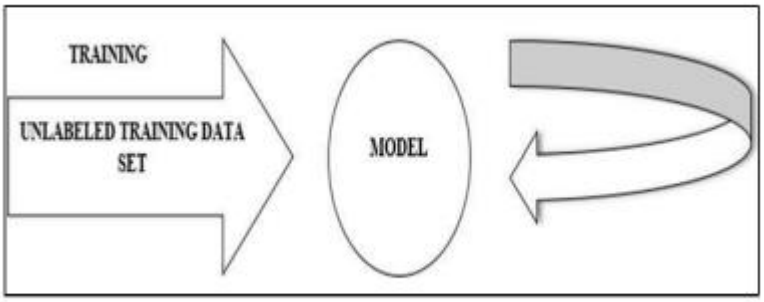


Fig. 5.8 :Unsupervised Learning

For example, if any data set was the characteristics and purchasing behavior of shoppers at grocery stores, the unsupervised learning task might be to segment the shoppers into groups or “clusters” that exhibit similar behaviors. Such learning methods might find that college students, parents with young children, and older adults have characteristic shopping behaviors that are similar within each group but dissimilar from the other. This is an unsupervised learning task because there is no right or wrong about how many clusters can be found in the data, which people belong in which cluster, or even how to describe each cluster.

5.6 MACHINE LEARNING ALGORITHMS

SUPPORT VECTOR MACHINE

Support vector machine is another simple algorithmic norm which every talented AI should have in their arsenal. Help vector machine is strongly favored by many, because it achieves critical precision with less computing power. Support Vector Machine, abbreviated as SVM, is used for every task of regression and classification. But, it's commonly used in goals for classification. The aim of the support vector machine algorithm is to search for a hyperplane in the associated N-dimensional space (N — the quantity of features) which separately classifies the information points. There are several potential hyperplanes which would be chosen to separate the 2 categories of information points. Our goal is to search for a plane with the utmost margin, that is, the utmost distance between the points of knowledge of each category. Increasing the margin gap should provide some reinforcement so that future data points can be identified with extra confidence.

RANDOM FORESTS

Random forests (RF) are associated with the overall methodology for classification. The set could be a combination of decision trees made from a bootstrap sample of the coaching set. In addition, when designing a chosen tree, the split that is chosen when splitting a node is that the best partition is only between a random set of features. This may increase the bias of one model, but the average reduces the variance and may also make adjustments to increase the bias. As a result, a higher model is built. In this work, the creation of random forests in the scikit learn python library is employed, that averages the probabilistic predictions.

ADABOOST

Adaboost or adaptive boosting may be a technique that encourages ensemble building classifiers that are modified by previous classifiers in favor of misclassified instances. The classifiers that it uses will be as weak as only slightly higher than the random estimate, and will still improve the final model . This technique is also used to enhance the ultimate ensemble model, in combination with alternative strategies. Sure weights are added to the coaching samples in every iteration of Adaboost. Such units of area weights distributed evenly prior to initial iteration. Then, when each iteration increases the weights for incorrectly labeled labels by current model, and weights are shriveled for properly sorted samples. This means that the current predictor focuses on previous

classificator vulnerabilities. We tried Adaboost implementation using scikit-learn python library with decision trees.

KNN

The k-nearest neighbors (KNN) algorithmic rule could be a easy, easy to-implement supervised machine learning which will that may be wont to solve each classification and regression issues. A supervised machine learning algorithm rule (as unkind associate unsupervised machine learning algorithm) is one that depends on labeled input file to be told a function that produces associate applicable output once given new unlabelled information.K-nearest neighbor is applied as a basic instance based learning algorithmic rule to classification problems. The label for a check sample is expected in this method to support the majority vote of its K-nearest neighbors.

DECISION TREE DT

is a supervised learning algorithm which is normally preferred for classification tasks. The algorithm works well for both types of variables i.e., categorical and continuous. It starts with splitting the population into multiple homogeneous sets which is done on the basis of most significant attributes or independent variables. DT is non-parametric and hence the need for checking outlier existence or data linearity separation is not required.

LOGISTIC REGRESSION

It is considered as the go-to method for classification involving binary results. It is mainly used in estimating discrete values which are based on set of variables which are independent. In more relative terms, LR outputs the probability of an event by fitting it into a logistic function which helps in prediction. The logistic function which is mostly used is sigmoid .

ARTIFICIAL NEURAL NETWORK

ANNs are nonlinear statistical data modelling techniques defined over complex relationships between inputs and outputs. They have various advantages, but among them, learning by observing datasets is the most recognised one. It is considered as tool for random function approximation, which helps in estimating the most effective methods to come to solutions. One such network is CNN.

CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network (CNN) is a particular type of artificial neural network which uses perceptrons for supervised learning. The supervised learning is used to analyze data. There are a wide of range of applications involving CNNs. Traditionally used for image processing, CNNs are nowadays used natural language processing as well. A CNN in relative terms is known as a Conv Net. Similar to other ANNs, a CNN also has an input layer, some hidden layers and an output layer, but it is not fully connected. Some layers are convolutional, that use a mathematical model to pass on the results to layers ahead in the network.

RNN

As recurrent Neural Network (RNN) may be a kind of Neural Network where the output is taken from the preceding step is fed as input to the present step . Some keywords are used to distinguish message from ham and spam. First, we need to train the dataset using keywords and then we need to upload the file and then we get the output as expected and the accuracy as well. By using RNN the precision is greater as compared with various methods such as SVM, Navie bayes, Multi NB, KNN .

5.7 NAIVE BAYES ALGORITHM

o Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems .

o It is mainly used in text classification that includes a high-dimensional training dataset.

o Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

o It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

o Here are some of the common applications of Naive Bayes for real-life tasks:

Document classification: This algorithm can help you to determine to which category a given document belongs.

Spam filtering: Naive Bayes easily sorts out spam using keywords.

Sentimental analysis: Based on what emotions the words in a text express, Naive Bayes can calculate the probability of it being positive or negative.

Image classification: For personal and research purposes, it is easy to build a Naive Bayesian classifier. It can be trained to recognize hand-written digits or put images into categories through supervised machine learning.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

o Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of colour, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

o Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Types of Naive Bayes

1) **Gaussian Naive Bayes**

This type of Naive Bayes is used when variables are continuous in nature. It assumes that all the variables have a normal distribution. So if you have some variables which do not have this property, you might want to transform them to the features having distribution normal**.**

2) **Multinomial Naive Bayes**

This is used when the features represent the frequency. Suppose you have a text document and you extract all the unique words and create multiple features where

each feature represents the count of the word in the document. In such a case, we have a frequency as a feature. In such a scenario, we use multinomial Naive Bayes. It ignoresthe non-occurrence of the features. So, if you havefrequency 0 then the probability of occurrence of that feature will be 0 hence multinomial naive Bayes ignores that feature. It is known to work well with text classification problems.

3) **Bernoulli Naive Bayes**

This is used when features are binary. So, instead of using the frequency of the word, if you have discrete features in 1s and 0s that represent the presence or absence of a feature. In that case, the features will be binary and we will use Bernoulli Naive Bayes.

CHAPTER 6
INTRODUCION TO NAÏVE BAYES THEOREM

Bayes theorem is one of the earliest probabilistic inference algorithms developed by Reverend Bayes (which he used to try and infer the existence of God no less) and still performs extremely well for certain use cases.

It's best to understand this theorem using an example. Let's say you are a member of the Secret Service and you have been deployed to protect the Democratic presidential nominee during one of his/her campaign speeches. Being a public event that is open to all, your job is not easy and you have to be on the constant lookout for threats. So one place to start is to put a certain threat-factor for each person. So based on the features of an individual, like the age, sex, and other smaller factors like is the person carrying a bag?, does the person look nervous ? etc. you can make a judgement call as to if that person is viable threat.

If an individual ticks all the boxes up to a level where it crosses a threshold of doubt in your mind, you can take action and remove that person from the vicinity. The Bayes theorem works in the same way as we are computing the probability of an event(a person being a threat) based on the probabilities of certain related events(age, sex, presence of bag or not, nervousness etc. of the person). One thing to consider is the independence of these features amongst each other. For example if a child looks nervous at the event then the likelihood of that person being a threat is not as much as say if it was a grown man who was nervous. To break this down a bit further, here there are two features we are considering, age and nervousness.

Say we look at these features individually, we could design a model that flags all persons that are nervous as potential threats. However, it is likely that we will have a lot of false positives as there is a strong chance that minors present at the

event will be nervous. Hence by considering the age of a person along with the 'nervousness' feature we would definitely get a more accurate result as to who are potential threats and who aren't.

This is the 'Naive' bit of the theorem where it considers each feature to be independent of each other which may not always be the case and hence that can affect the final judgement.

In short, the Bayes theorem calculates the probability of a certain event happening (in our case, a message being spam) based on the joint probabilistic distributions of certain other events (in our case, a message being classified as spam) . We will dive into the workings of the Bayes theorem later in the mission, but first, let us understand the data we are going to work with.

6.1 UNDERSTANDING OUR DATASET

We will be using a dataset from the kaggle which has a very good collection of datasets for experimental research purposes.

Ham : Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...

Ham : Ok lar... Joking wif u oni...

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

Ham : U dun say so early hor... U c already then say... Ham : Nah I don't think he goes to usf, he lives around here though

Spam : Free Msg 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, å£1.50 to rcv

Ham : Even my brother is not like to speak with me. They treat me like aids patent.

Ham : As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your caller tune for all Callers. Press *9 to copy your friends Caller tune

Spam : WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.

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

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.

Spam : SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ Ts and Cs apply Reply HL 4 info

Spam : URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18

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.

Ham : I HAVE A DATE ON SUNDAY WITH WILL!!

Spam : XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap.

xxxmobilemovieclub.com?n=QJKGIGHJJGCBL Ham : Oh k...i'm watching here:)

The columns in the data set are currently not named and as you can see, there are 2 columns.

The first column takes two values, 'ham' which signifies that the message is not spam, and 'spam' which signifies that the message is spam.

The second column is the text content of the message that is being classified.

6.2 DATA PREPROCESSING

Now that we have a basic understanding of what our dataset looks like, lets convert our labels to binary variables, 0 to represent 'ham'(i.e. not spam) and 1 to represent 'spam' for ease of computation.

You might be wondering why do we need to do this step? The answer to this lies in how scikit-learn handles inputs. Scikit-learn only deals with numerical values and hence if we were to leave our label values as strings, scikit-learn would do the conversion internally(more specifically, the string labels will be cast to unknown float values).

Our model would still be able to make predictions if we left our labels as strings but we could have issues later when calculating performance metrics, for example when calculating our precision and recall scores. Hence, to avoid unexpected 'gotchas' later, it is good practice to have our categorical values be fed into our model as integers.

6.3 BAG OF WORDS

What we have here in our data set is a large collection of text data (5,572 rows of data) . vMost ML algorithms rely on numerical data to be fed into them as input, and email/sms messages are usually text heavy.

Here we'd like to introduce the Bag of Words (BOW) concept which is a term used to specify the problems that have a 'bag of words' or a collection of text data that needs to be worked with. The basic idea of BOW is to take a piece of text and count the frequency of

the words in that text. It is important to note that the BOW concept treats each word individually and the order in which the words occur does not matter.

Using a process which we will go through now, we can convert a collection of documents to a matrix, with each document being a row and each word(token) being the column, and the corresponding (row, column) values being the frequency of occurrence of each word or token in that document.

Data preprocessing with CountVectorizer()

Some of important parameters of Coun Vectorizer().

1. lowercase = True The lowercase parameter has a default value of True which converts all of our text to its lowercase form.

2. Token pattern = (?u)\b\w\w+\b The token pattern parameter has a default regular expression value of (?u)\b\w\w+\b which ignores all punctuation marks and treats them as delimiters, while accepting alphanumeric strings of length greater than or equal to 2, as individual tokens or words.

3. Stop words The stop words parameter, if set to english will remove all words from our document set that match a list of English stop words which is defined in scikit-learn. Considering the size of our dataset and the fact that we are dealing with SMS messages and not larger text sources like e- mail, we will not be setting this parameter value.

6.4 TRAINING AND TESTING SETS

Now that we have understood how to deal with the Bag of Words problem we can get back to our dataset and proceed with our analysis. Our first step in this regard would be to split our dataset into a training and testing set so we can test our model later.

Instructions: Split the dataset into a training and testing set by using the train_test_split method in sklearn. Split the data using the following variables:

• X train is our training data for the 'sms_message' column.

• Y_train is our training data for the 'label' column .

• X_test is our testing data for the 'sms_message' column.

• y_test is our testing data for the 'label' column Print out the number of rows we have in each our training and testing data.

6.5 APPLYING BAG OF WORDS PROCESSING TO OUR DATASET

Now that we have split the data, our next objective is to follow the steps from Bag of words and convert our data into the desired matrix format. To do this we will be using CountVectorizer(). There are two steps to consider here:

• Firstly, we have to fit our training data (X_train) into CountVectorizer() and return the matrix.

• Secondly, we have to transform our testing data (X_test) to return the matrix. Note that X_train is our training data for the 'sms_message' column in our dataset and we will be using this to train our model.

X_test is our testing data for the 'sms_message' column and this is the data we will be using(after transformation to a matrix) to make predictions on. We will then compare those predictions with y_test in a later step.

6.6 IMPLEMENTATION OF NAIVE BAYES ML ALGORITHM

we will be using sklearns sklearn.naive_bayes method to make predictions on our dataset for message Spam Detection.

Specifically, we will be using the multinomial Naive Bayes implementation. This particular classifier is suitable for classification with discrete features. It takes in integer word counts as its input.

```
from sklearn.naive_bayes import MultinomialNB

naive_bayes = MultinomialNB()

naive_bayes.fit(training_data,y_train)

MultinomialNB(alpha=1.0,
```

class_prior=None, fit_prior=True)

predictions = naive_bayes.predict(testing_data)

6.7 EVALUATING OUR MODEL

Now that we have made predictions on our test set, our next goal is to evaluate how well our model is doing. There are various mechanisms for doing so, but first let's do quick recap of them.

Accuracy measures how often the classifier makes the correct prediction. It’sthe ratio of the number of correct predictions to the total number of predictions (the number of test data points).

Precision tells us what proportion of messages we classified as spam, actually were spam. It is a ratio of true positives(words classified as spam, and which are actually spam) to all positives (all words classified as spam, irrespective of whether that was the correct classification) , in other wordsit is the ratio of

True Positives/(True Positives + False Positives)

Recall(sensitivity) tells us what proportion of messages that actually were spam were classified by us as spam. It is a ratio of true positives(words classified as spam, and which are actually spam) to all the words that were actually spam, in other words it is the ratio of

True Positives/(True Positives + False Negatives)

For classification problems that are skewed in their classification distributions like in our case, for example if we had a 100 text messages and only 2 were spam and the rest 98 weren't, accuracy by itself is not a very good metric. We could classify 90 messages as not spam (including the 2 that were spam but we classify them as not spam, hence they would be false negatives) and 10 as spam(all 10 false positives) and still get a reasonably good accuracy score. For such cases, precision and recall come in very handy. These two metrics can be combined to get the F1 score, which is weighted average of the precision and recall scores . This score can range from 0 to 1, with 1 being the best possible F1 score .

We will be using all 4 metrics to make sure our model does well. For all 4 metrics whose values can range from 0 to 1, having a score as close to 1 as possible is a good indicator of how well our model is doing.

The following table shows the evaluation measures for spam filters .

Evaluation Measure	**Evaluation Function**
Accuracy	Acc = TN+TP/(TP+FN+FP+TN)
Recall	R =TP/(TP+FN)
Precision	P=TP/TP+FP
F-Measure	F=2PR/(P+R)

Table 6.1 : Evaluation measures for spam filters

Where accuracy,recall,precision,F-measure,FP,FN,TP and TN are defined as follows :

Accuracy : Percentage of correctly identified spam and not spam message .

Recall : Percentage spam message manage to block . Precision : Percentage of correct message for spam message.

F-measure : Weighted average of precision and recall.

False Positive Rate (FP) : The number of misclassified non spam messages.

False Negative Rate (FN) : The number of misclassified spam messages.

True Positive (TP) : The number of spam messages are correctly classified as spam.

True Negative (TN) : The number of non-spam messages are correctly classified as non spam .

One of the major advantages that Naive Bayes has over other classification algorithms is its ability to handle an extremely large number of features. In our case, each word is treated as a feature and there are thousands of different words. Also, it performs well even with the presence of irrelevant features and is relatively unaffected by them. The other major advantage it has is its relative simplicity . Naive Bayes' works well right out of the box and tuning it's parameters is rarely ever necessary, except usually in cases where the distribution of the data is known. It rarely ever overfits the data. Another important

advantage is that its model training and prediction times are very fast for the amount of data it can handle. All in all, Naive Bayes' really is a gem of an algorithm.

6.8 THE GRAPHICAL MODEL

The naive Bayes classifier is a simple probabilistic model with strong independence assumptions . In particular , the underlying Bayesian network consists of a single node y (whether the message is spam , for instance) that determines a set of features F1,…., Fn independently (with Fi corresponding to the word at position i, for instance) . For example , if the first word is “ Free “ , the message is more likely to be spam than if it is “ Hey “ (or more accurately , spam is more likely to begin with “ Free “ than Hey) . The graphical model is shown below :

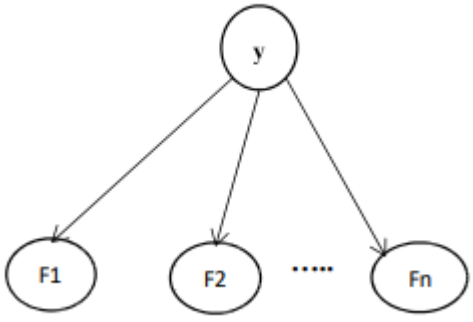


Fig 6.1 : Graphical model

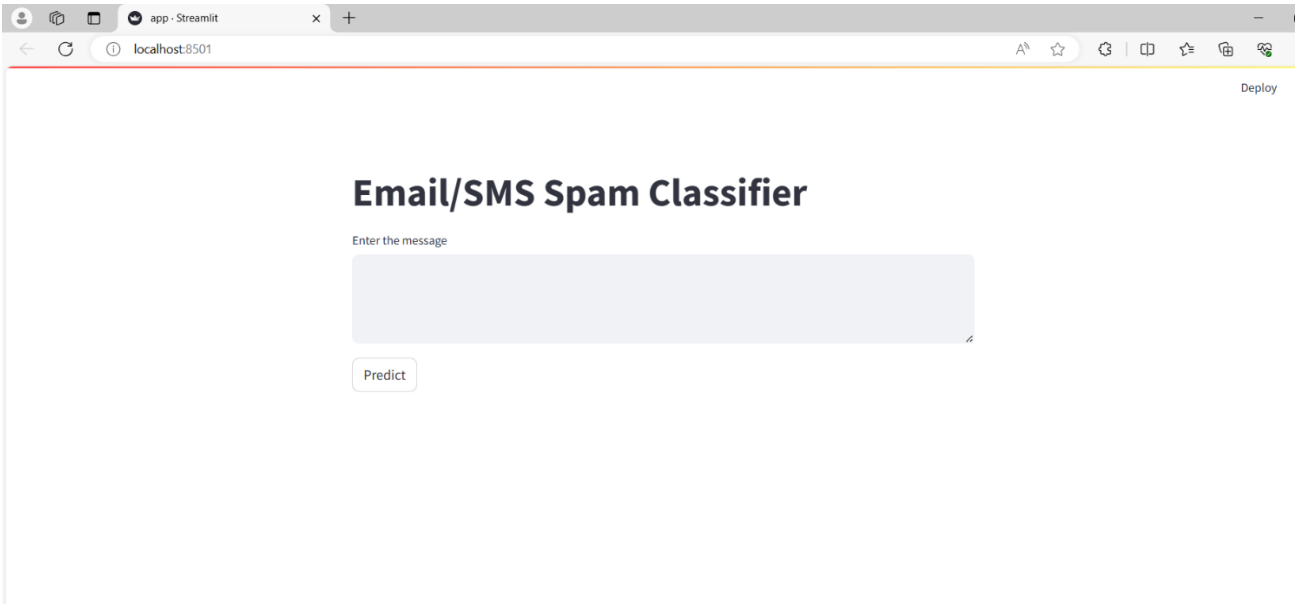
Based on this assumption , we know that the distribution of y given the features (words at various positions) is Pr(y|F1,…,Fn)∝Pr(y)∏iP(Fi|y)

(Recall that Pr(y|F1,….,Fn)=Pr(y,F1,….,Fn)Pr(F1,…,Fn) by Bayes rule). In other words ,to determine the distribution ,you simply compute Pr(y)∏iP(Fi|y) and normalize the final distribution . Remember, Pr(y|F1,….Fn) is a distribution where ,in the case of spam detection, the two possible values of y are " Spam " or " Ham " .

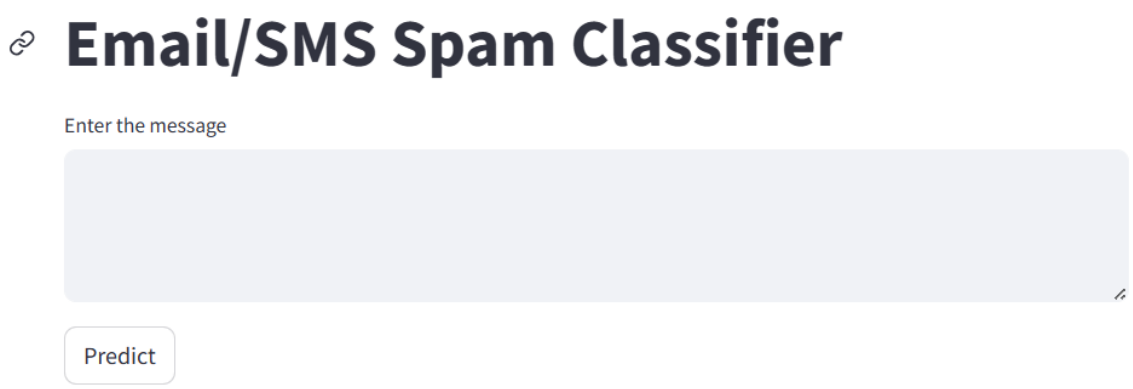
CHAPTER 7

SCREENSHOTS

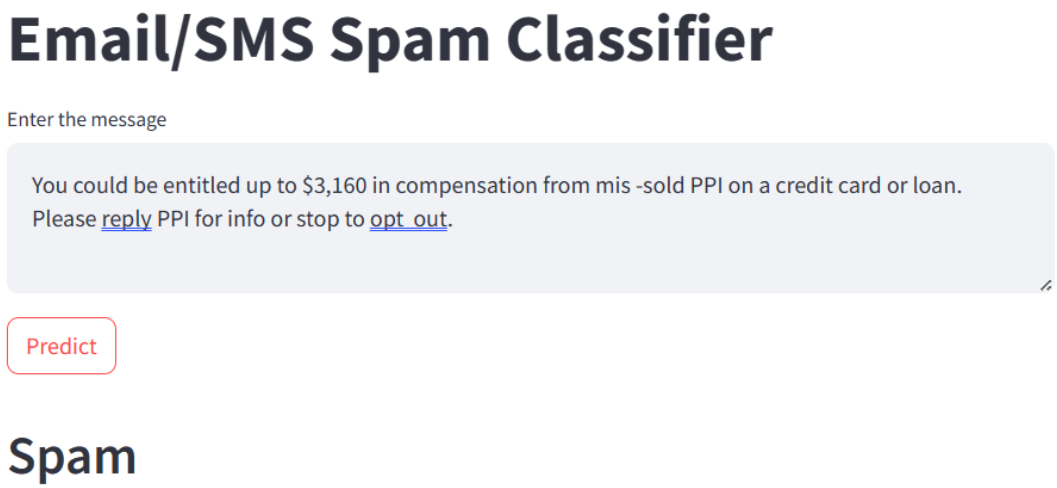
7.1 HOMEPAGE



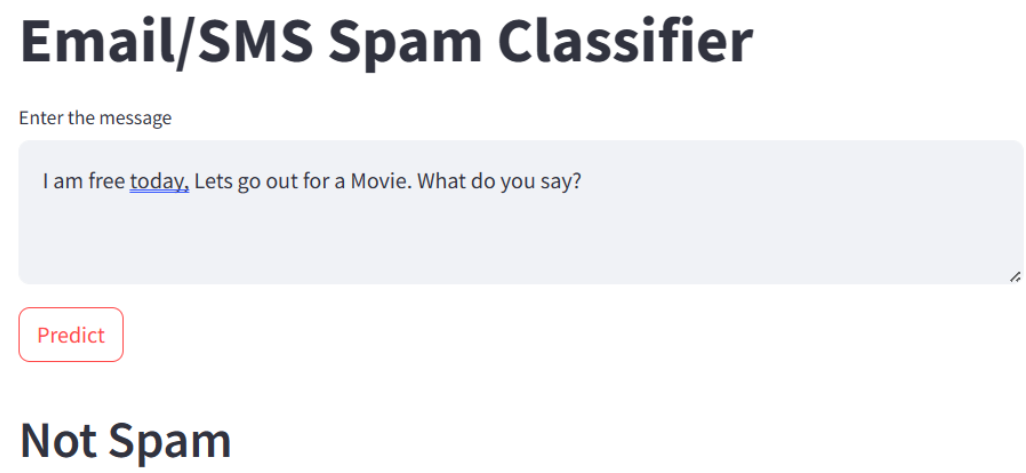
7.2 ENTER TO CHAT ROOM



7.3 SPAM MESSAGE



7.4 HAM MESSAGE



CHAPTER 9

CONCLUSION

In this Project, Naïve Bayes Algorithm analyses based on the factors like precision, recall, f1- score, support. Naive Bayes order calculation is viably valuable for managing clear cut information characterization. The basic hypothesis it utilizes is the Bayes contingent probabilistic model for tracking down a back likelihood given certain conditions. It is classified "Credulous" on the grounds that under the presumption that all highlights (assortments of words) in the dataset are similarly significant and free . Utilizing the Naïve Bayes grouping calculation, the venture got over 98% exactness in foreseeing a spam message dependent on the words it contains which is determined from the confusion matrix which we got as yield .

9.1 RESULT

	Algorithm	Accuracy	Precision	Accuracy_max_ft_3000	Precision_max_ft_3000	Accuracy_scaling	Precision_scaling	Accuracy_num_chars	Precision_num_chars
0	KN	0.900387	1.000000	0.905222	1.000000	0.905222	0.976190	0.928433	0.771186
1	NB	0.959381	1.000000	0.971954	1.000000	0.978723	0.946154	0.940039	1.000000
2	ETC	0.977756	0.991453	0.979691	0.975610	0.979691	0.975610	0.976789	0.975000
3	RF	0.970019	0.990826	0.975822	0.982906	0.975822	0.982906	0.974855	0.982759
4	SVC	0.972921	0.974138	0.974855	0.974576	0.971954	0.943089	0.866538	0.000000
5	AdaBoost	0.962282	0.954128	0.961315	0.945455	0.961315	0.945455	0.971954	0.950413
6	xgb	0.971954	0.950413	0.968085	0.933884	0.968085	0.933884	0.970019	0.942149
7	LR	0.951644	0.940000	0.956480	0.969697	0.967118	0.964286	0.961315	0.971154
8	GBDT	0.951644	0.931373	0.946809	0.927835	0.946809	0.927835	0.948743	0.929293
9	BgC	0.957447	0.861538	0.959381	0.869231	0.959381	0.869231	0.968085	0.913386
10	DT	0.935203	0.838095	0.931335	0.831683	0.932302	0.840000	0.943907	0.877358

9.2 FUTURE ENHANCEMENT

To perform prediction in this project, we only used static data for training and testing the algorithms. To improve the accuracy of the forecasts in the future, the project will need to

increase the amount of data in the data collection. Which in turn helps the model to improve its accuracy of prediction.

Implement active learning strategies to continuously improve the model. Allow users to provide feedback on misclassified messages, and use this feedback to retrain the model periodically.

Multimodal Approaches: Combine text analysis with other modalities such as images or user metadata. This can be particularly useful in chat rooms where multimedia content is prevalent.

User Customization: Allow users to customize their spam filters. Some users may prefer more lenient filters, while others may want stricter controls. Providing a level of customization can enhance user satisfaction.

9.3 SUMMARY

From this project, it can be concluded that machine learning algorithm is one of the important part in order to create spam detection application. To make it more efficient, improvement need to be implemented in future.

REFERENCES

WEB REFERENCE

SMS Spam Collection V.1 Dataset. Publicly available online at: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/.

Spam SMS Dataset 2011-12. Available online on request at: http://precog.iiitd.edu.in/requester.php?dataset=smsspam

https://www.kaggle.com/kentata/rnn-for-spam-detection.

https://ieeexplore.ieee.org/abstract/document/7727636.

http://cs229.stanford.edu/proj2013/ShiraniMehrSMSSpamDetectionUsingMachineLearningApproch.pdf

(PDF) Spam filtering in SMS using recurrent neural networks (researchgate.net)