

“Spam Detection Using Machine Learning”

Submitted in partial fulfillment of the requirements

of the degree of

Bachelor of Technology

In

Electronics and Telecommunication Engineering

by

(1) Gunjal Shahu Yogesh (2) Kulkarni Ajinkya Anantrao (3) Balki Saurabh Vilas

2018BEC005

2018BEC023

2018BEC025

Supervisor (s):

Dr. A. V. Nandedkar



Department of Electronics and Telecommunication Engineering,

**Shri Guru Gobind Singhji Institute of Engineering & Technology, Vishnupuri, Nanded,
Maharashtra, India, 431606.**

2021-22

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Gunjal Shahu Yogesh
(Reg No. 2018BEC005)

Kulkarni Ajinkya Anantrao
(Reg No. 2018BEC023)

Balki Saurabh Vilas
(Reg No. 2018BEC025)

Date: _____

CERTIFICATE

This is to certify that the report entitled “**Spam Detection Using Machine Learning**” being submitted by **Gunjal Shahu Yogesh (Reg No. 2018BEC005), Kulkarni Ajinkya Anantrao (Reg No. 2018BEC023), Balki Saurabh Vilas (Reg No. 2018BEC025)** to **Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri, Nanded (M.S.), India**, as partial fulfillment for the award of the degree of **Bachelor of Technology in Electronics and Telecommunication Engineering**, is a record of bonafide work carried out by him under our supervision and guidance. The matter contained in this report has not been submitted to any other university for the award of any degree or diploma.

Dr. A. V. Nandedkar

HOD & Supervisor

**Elect. and Telecom. Engg.
Dept.**

APPROVAL SHEET

This report entitled "**Spam Detection Using Machine Learning**" by **Gunjal Shahu Yogesh (Reg No. 2018BEC005)**, **Kulkarni Ajinkya Anantrao (Reg No. 2018BEC023)**, **Balki Saurabh Vilas (Reg No. 2018BEC025)** is approved for the degree of Bachelor of Technology.

Examiners

Supervisor (s)

Date: _____

Place: _____

ACKNOWLEDGEMENT

In the present world of competition there is a race of existence in which those are having will to come forward succeed. Project is like bridge between theoretical and practical working. With this willing we joined this particular project. First of all, I would like to thank our supervisor for this project course Dr. A. V. Nandedkar (Ph.D.) who is obviously the one has guided us to work on this project. Without his enthusiasm this project would not become been reality. We are feeling oblige in taking the opportunity to sincerely thanks our supervisor and special thanks to Department of Computer Engineering, University of Maiduguri, Maiduguri, Nigeria and Universiti Sultan Zainal Abidin (UniSZA) from where we collected the necessary information for the project. At last, but not least we are thankful for ourselves for doing all this work, giving best, dedicating oneself for this project in this crucial condition.

We are happy to work on this project which is very important for the conditions we are living in. Working on Machine Learning topic has taught us lot of things and made us confident enough to work on any important projects in future. Thank you EXTC department for conducting such special course in the curriculum.

Thank you, Professor, all friends and fellow groupmates.

Gunjal Shahu Yogesh
(Reg No. 2018BEC005)

Kulkarni Ajinkya Anantrao
(Reg No. 2018BEC023)

Balki Saurabh Vilas
(Reg No. 2018BEC025)

ABSTRACT

With the implementation of Naïve Bayes Classifier which is a simple probabilistic classifier with strong assumptions of independence, we were able to create a spam detection system. The upsurge in the volume of unwanted emails called spam has created an intense need for the development of more dependable and robust antispam filters. Machine learning methods of recent are being used to successfully detect and filter spam emails. 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. The preliminary discussion in the study background examines the applications of machine learning techniques to the email spam filtering process of the leading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters. Discussion on general email spam filtering process, and the various efforts by different researchers in combating spam through the use machine learning techniques was done. Our review compares the strengths and drawbacks of existing machine learning approaches and the open 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.

Keywords:

1. Machine Learning
2. Naïve Bayes Classifier
3. Computer security
4. Computer privacy
5. Analysis of algorithms
6. Spam filtering
7. Natural Language Toolkit (NLTK)

CONTENTS

List of Figures	i
List of Tables	ii
Abbreviations, Notations and Nomenclature	iii
1. Introduction	1
1.1 Current Scenario	1
1.2 What is Machine Learning?	3
1.3 About Classification Algorithm in Machine Learning	4
2. Literature Survey	6
2.1 Supervised Machine Learning	6
2.2 Naïve Bayes Classifier Algorithm and Multinomial Naïve Bayes Classifier	7
2.3 Recent Researches and Related Work	9
2.3.1 Recent Researches	9
2.3.2 Related Work	12
3. Pre-Requirements of Projects	13
3.1 Dataset	13
3.2 Softwares Tools	14
3.2.1 Jupyter Notebook	14
3.2.2 Streamlit	15
3.2.3 PyCharm	15
3.3 Libraries	16

4. Proposed Method (Naïve Bayes Classifier)	18
4.1 Methodology	18
4.1.1 Flow of Research	18
4.1.2 Block Diagram	19
4.2 Data Cleaning	20
4.3 EDA (Exploratory Data Analysis)	21
4.4 Data Pre-processing	22
4.5 Model Building	23
4.5.1 Data Splitting	23
4.5.2 Training Data	23
5. Experimental Results	26
5.1 Performance evaluation measures	26
5.2 Analysis	26
5.3 Screenshots of Result	27
6. Conclusions and Future Scope	29
6.1 Conclusion	29
6.2 Future Scope	29
References	30

List Of Figures

Figure. 1. The volume of spam emails 4th quarter 2016 to 1st quarter 2018.

Figure. 2. Block diagram for the working of Machine Learning algorithm

Figure. 3. Classification Algorithm diagram for two classes

Figure. 4. The working of Supervised learning

Figure. 5. Data flow of spam detection machine learning

Figure. 6. Block Diagram

Figure. 7. Steps involved in Data Cleaning

Figure. 8. Training Data through all three classifiers

Figure 9. Training Data through different classifiers/models imported from sklearn

List Of Tables

Table 1. Summary of previous reviews in email spam filtering.

Table 2. Related Work

Abbreviations

1. ML – Machine Learning
2. NLP – Natural Language Processing
3. NLTK – Natural Language Toolkit
4. NB – Naïve Bayes
5. EDA – Exploratory Data Analysis

1. Introduction

1.1: Current Scenario

In recent times, 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.

According to report from Kaspersky lab, in 2015, the volume of spam emails being sent reduced to a 12-year low. Spam email volume fell below 50% for the first time since 2003. In June 2015, the volume of spam emails went down to 49.7% and in July 2015 the figures were further reduced to 46.4% according to anti-virus software developer Symantec. This decline was attributed to reduction in the number of major botnets responsible for sending spam emails in billions. Malicious spam email volume was reported to be constant in 2015. The figure of spam mails detected by Kaspersky Lab in 2015 was between 3 million and 6 million. Conversely, as the year was about to end, spam email volume escalated. Further report from Kaspersky Lab indicated that spam email messages having pernicious attachments such as malware, ransomware, malicious macros, and JavaScript started to increase in December 2015. That drift was sustained in 2016 and by March of that year spam email volume had quadrupled with respect to that witnessed in 2015. In March 2016, the volume of spam emails discovered by Kaspersky Lab is 22,890,956. By that time the volume of spam emails had skyrocketed to an average of 56.92% for the first quarter of 2016. Latest statistics shows that spam messages accounted for 56.87% of e-mail traffic worldwide and the most familiar types of spam emails were healthcare and dating spam. Spam results into

unproductive use of resources on Simple Mail Transfer Protocol (SMTP) servers since they have to process a substantial volume of unsolicited emails. The volume of spam emails containing malware and other malicious codes between the fourth quarter of 2016 and first quarter of 2018 is depicted in Fig. 1 below.

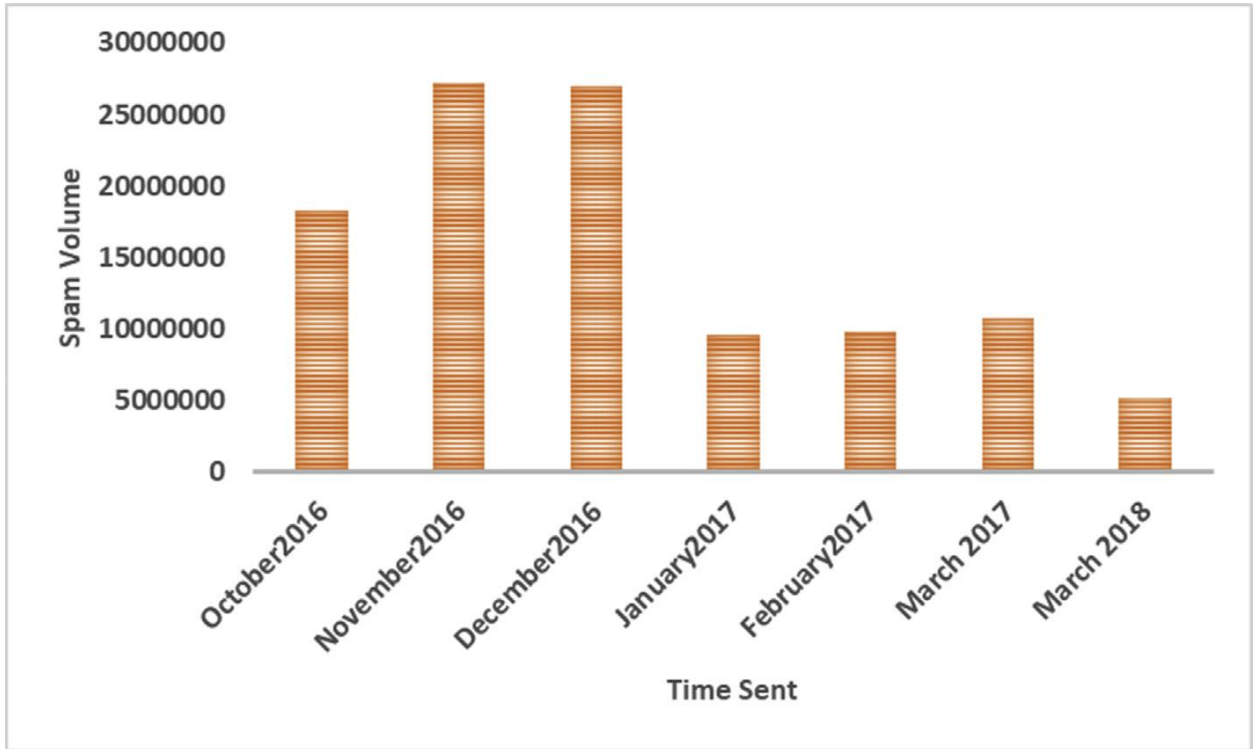


Fig. 1. The volume of spam emails 4th quarter 2016 to 1st quarter 2018.

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 analyzing 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 analyzed 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 paper. We explained below the different categories of spam filtering techniques that have been widely applied to overcome the problem of email spam.

1.2: What is Machine Learning?

Machine Learning^[1] is said as a subset of **Artificial Intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in 1959. We can define it in a summarized way as:

“Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.”

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

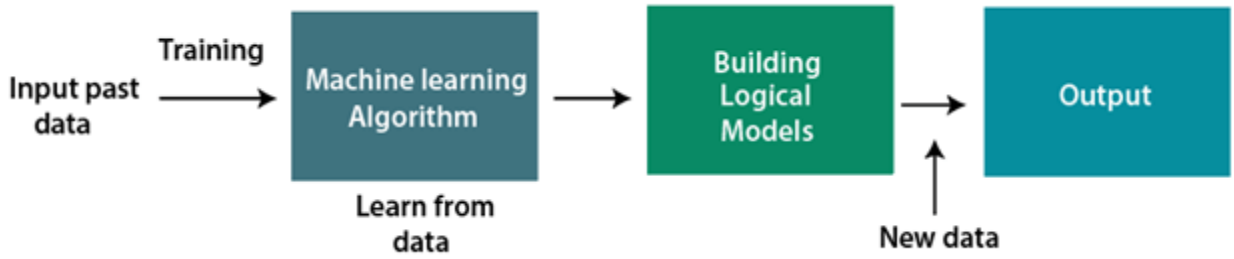


Fig. 2. Block diagram for the working of Machine Learning algorithm

1.3: About Classification Algorithm in Machine Learning

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning

technique, hence it takes labelled input data, which means it contains input with the corresponding output.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

$$y=f(x), \text{ where } y = \text{categorical output}$$

The best example of an ML classification algorithm is our project **Email/SMS Spam Detector**.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.

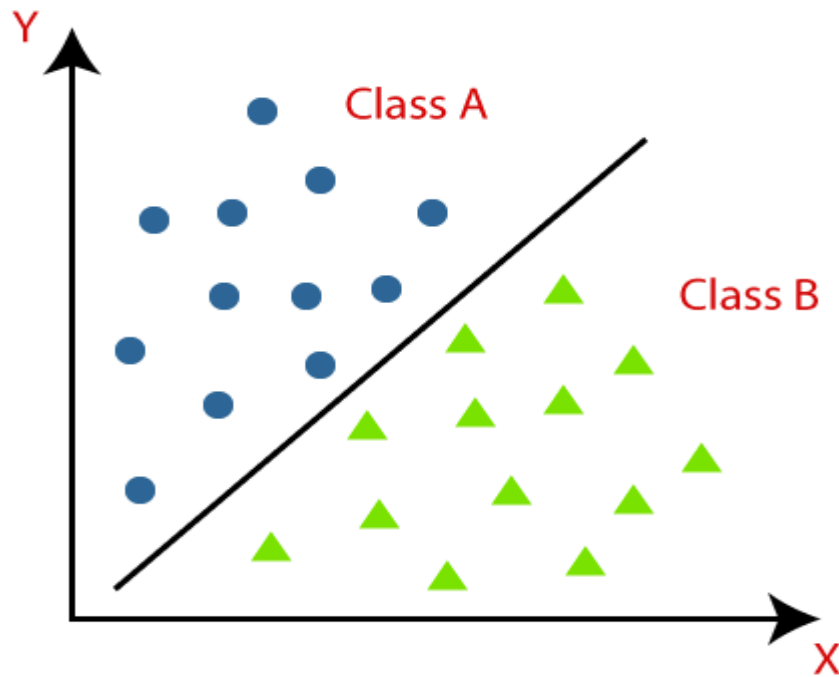


Fig. 3. Classification Algorithm diagram for two classes

2. Literature Survey

2.1: Supervised Machine Learning

Supervised learning ^[2] is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly.

It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc. In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

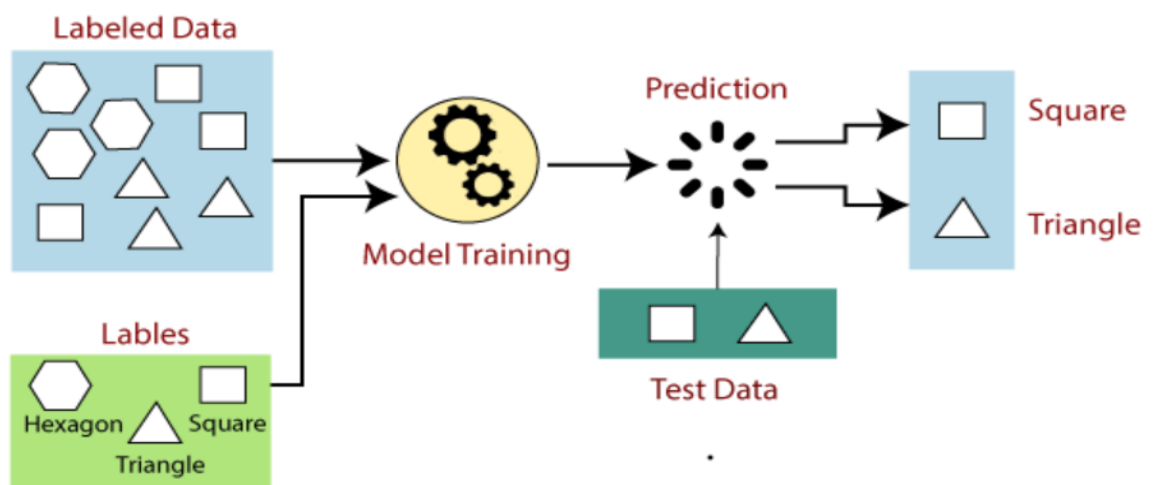


Fig. 4. The working of Supervised learning

2.2: Naïve Bayes Classifier Algorithm and Multinomial Naïve Bayes Classifier

Naïve Bayes Classifier Algorithm:

- Naïve Bayes algorithm ^[3] is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- 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.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

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

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

Multinomial Naïve Bayes:

There are thousands of Softwares or tools for the analysis of numerical data but there are very few for texts. Multinomial Naive Bayes ^[4] is one of the most popular supervised learning classifications that is used for the analysis of the categorical text data.

Text data classification is gaining popularity because there is an enormous amount of information available in email, documents, websites, etc. that needs to be analysed. Knowing the context around a certain type of text helps in finding the perception of a software or product to users who are going to use it.

What is Multinomial Naïve Bayes Algorithm?

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

Applications of Naïve Bayes Classifier:

- It is used for Credit Scoring.
- It is used in medical data classification.
- It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as Spam filtering and Sentiment analysis.

2.3: Recent Researches and Related Work

2.3.1: Recent Researches

There is a rapid increase in the interest being shown by the global research community on email spam filtering. In this section, we present similar reviews that have been presented in the literature in this domain. This method is followed so as to articulate the issues that are yet to be addressed and to highlight the differences with our current review. Lueg [17] presented a brief survey to explore the gaps in whether information filtering and information retrieval technology can be applied to postulate Email spam detection in a logical, theoretically grounded manner, in order to facilitate the introduction of spam filtering technique that could be operational in an efficient way. However, the survey did not present the details of the Machine_learning_algorithms, the simulation tools, the publicly available datasets and the architecture of the email spam environment. It also fails short of presenting the parameters used by previous researches in evaluating other proposed techniques. Wang [18] reviewed the different techniques used to filter out unsolicited spam emails. The paper also to categorized email spams into different hierarchical folders, and automatically regulate the tasks needed to response to an email message. However, some of the limitations of the review article are that; machine learning techniques, email spam architecture, comparative analysis of previous algorithms and the simulation environment were all not covered.

The paper titled “Spam filtering and email-mediated applications” chronicles the details of email spam filtering system. It then presented a framework for a new technique for linking multiple filters with an innovative filtering model using ensemble learning algorithm. The article also explained the notion of operable email (OE) in an email-mediated application. Furthermore, a demonstration was made of OE in executing an email assistant and other intelligent applications on the world social email network [19]. However, the survey paper did not cover recent articles as it was published more than a decade ago. Cormack [20] reviewed previously proposed spam filtering algorithms up to 2008 with specific emphasis on efficiency of the proposed systems. The main focus of the review is to explore the relationships between email spam filtering with other spam filtering systems in communication and storage media. The paper also scrutinized the characterization of email spams, including the user's information requirements and the function of

the spam sieve as a constituent of a huge and complex information system. However, certain important components of spam filters were not considered in the survey. These includes; the architecture of the system, the simulation environment and the comparative analysis of the performance of the reviewed filters.

Sanz, Hidalgo, and Pérez [21] detailed the research issues related to email spams, in what way it affects users, and by what means users and providers can reduce it effects. The paper also enumerates the legal, economic, and technical measures used to mediate the email spams. They pointed out that based on technical measures, content analysis filters have been extensively used and proved to have reasonable percentage of accuracy and precision as a result, the review focused more on them, detailing how they work. The research work explained the organization and the procedure of many machine learning approaches utilized for the purpose of filtering email spams. However, the review did not cover recent research articles in this area as it was published in 2008 and comparative analysis of the different content filters was also missing. A brief study on E-mail image spam filtering methods was presented by [22]. The study concentrated on email antispam filtering approaches used to transfer from text-based techniques to image-based methods. Spam and the spam filters premeditated to reducing it have spawned an upsurge in creativeness and inventions. However, the study did not cover machine learning techniques, simulation tools, dataset corpus and the architecture of email spam filtering techniques.

Bhowmick and Hazarika [23] presented a broad review of some of the popular content-based e-mail spam filtering methods. The paper focused mostly on machine learning algorithms for spam filtering. They surveyed the important concepts, efforts, effectiveness, and the trend in spam filtering. They discussed the fundamentals of e-mail spam filtering, the changing nature of spam, the tricks of spammers to evade spam filters of e-mail service providers (ESPs), and also examined the popular machine learning techniques used in combating the menace of spam. Laorden *et al.* [24] presented a detailed revision of the usefulness of anomaly discovery used for email spam filtering that decreases the requirement of classifying email spam messages and only works with the representation of single class of emails. The review contains a demonstration of the first anomaly based spam sieving method, an improvement of the method, which used a data minimization technique to the characterized dataset corpus to decrease processing phase while

retaining recognition rates and an investigation of the appropriateness of selecting non-spam emails or spam as a demonstration of normality.

This current review differed from the previous reviews presented in the preceding paragraph by focusing more on revisiting machine learning techniques used for email spam filtering. The review intends to cover the architecture of the email spam filtering systems, parameters used for comparative analysis, simulation tools and the dataset corpus. The period under review also includes all recent research articles that are found to be useful for the advancement of the email spam filtering methods s shown in Table 1 below.

Table. 1. Summary of previous reviews in email spam filtering.

Previous Reviews	Email Spam	Machine Learning	Comparative Analysis	Simulation Tool & Environment	Dataset Corpus	Architecture	Parameters	Period Covered
Lueg [17]	✓							2000-2005
Wang [18]	✓				✓		✓	1995-2005
Li et al. [19]	✓	✓	✓	✓	✓		✓	1997-2006
Gormack [20]	✓	✓			✓		✓	2000-2008
Sanz et al. [21]	✓	✓		✓	✓		✓	2000-2008
Dhanaraj and Karthikeyani [22]	✓						✓	1994-2013
Bhowmick and Hazarika [23]	✓	✓	✓		✓	✓	✓	2004-2013
Laorden et al. [24]	✓				✓		✓	2002-2014
Our Review	✓	✓	✓	✓	✓	✓	✓	2000-2018

2.3.2: Related Work

Most research has been conducted into detecting and filtering spam email using a variety of techniques. This includes techniques such as K-NN^[5], Bayesian^[6], and ANN^[7] as shown in Table 2.

Table 2: Related Work

TITLE	TECHNIQUE	REMARK
Spam Filtering Using K-NN	K-nearest Neighbors	Computationally intensive, especially when the size of the training set grows.
A review of machine learning approaches to Spam filtering	Bayesian Filters	Require training period before it starts working well.
SVM-Based Spam Filter with Active and Online Learning	Support Vector Machines	Select the most useful example for labeling and add the labeled example to training set to retrain model.
A survey on spam detection techniques	Artificial Neural Network	Must be trained first to categorize emails into spam or non-spam starting from the particular data sets.

3. Pre-Requirements of Projects

3.1: Dataset

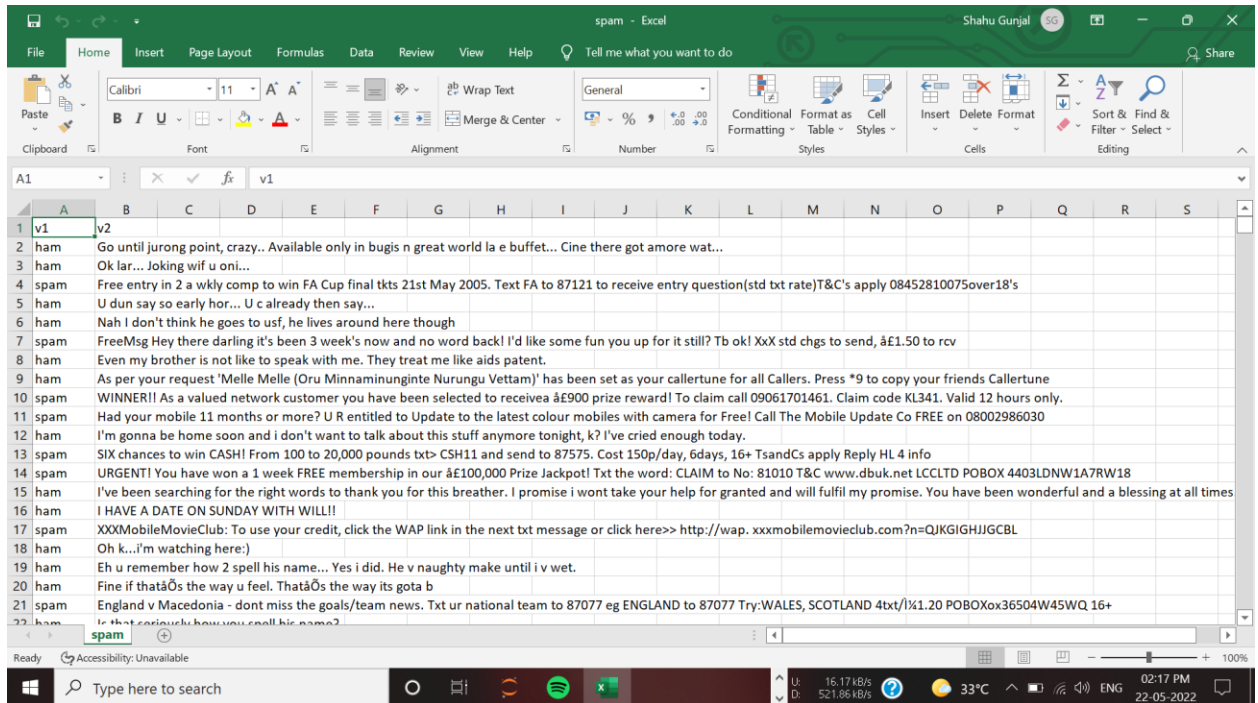
The SMS Spam Collection is a set of tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam. The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free for research sources at the Internet:

A collection of 425 SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages.



3.2: Softwares Tools

3.2.1: Jupyter Notebook

The Jupyter Notebook ^[8] is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The “notebook” term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context.

According to the official website of Jupyter, Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

Jupyter Book is an open-source project for building books and documents from computational material. It allows the user to construct the content in a mixture of Markdown, an extended version of Markdown called MyST, Maths & Equations using MathJax, Jupyter Notebooks, reStructuredText, the output of running Jupyter Notebooks at build time. Multiple output formats can be produced (currently single files, multipage HTML web pages and PDF files).

3.2.2: Streamlit

Streamlit makes it easy for you to visualize, mutate, and share data. The API reference is organized by activity type, like displaying data or optimizing performance. Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

Streamlit allows you to write an app the same way you write a python code. The streamlit has a distinctive data flow, any time something changes in your code or anything needs to be updated on the screen, streamlit reruns your python script entirely from the top to the bottom. This happens when the user interacts with the widgets like a select box or drop-down box or when the source code is changed.

3.2.3: PyCharm

PyCharm^[9] is an integrated development environment (IDE) used in computer programming, specifically for the Python programming language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

PyCharm is cross-platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License, and there is also an educational version, as well as

a Professional Edition with extra features (released under a subscription-funded proprietary license)

Features

- Coding assistance and analysis, with code completion, syntax and error highlighting, linter integration, and quick fixes
- Project and code navigation: specialized project views, file structure views and quick jumping between files, classes, methods and usages
- Python refactoring: includes rename, extract method, introduce variable, introduce constant, pull up, push down and others
- Support for web frameworks: Django, web2py and Flask [professional edition only]
- Integrated Python debugger
- Integrated unit testing, with line-by-line code coverage
- Google App Engine Python development [professional edition only]
- Version control integration: unified user interface for Mercurial, Git, Subversion, Perforce and CVS with change lists and merge
- Support for scientific tools like Matplotlib, NumPy and SciPy [professional edition only]

PyCharm provides an API so that developers can write their own plugins to extend PyCharm features. Several plugins from other JetBrains IDE also work with PyCharm. There are more than 1000 plugins which are compatible with PyCharm.

3.3: Libraries

This project requires you to have a good knowledge of Python and Natural Language Processing (NLP)^[10]. Modules required for this project are pandas, pickle, sklearn, numpy and nltk. You can install with using following command:

```
pip install pandas, pickle, sklearn, numpy, nltk
```

The versions which are used in this project for python and its corresponding modules are as follows:

- 1) python: 3.8.5
- 2) sklearn: 0.24.2
- 3) pickle: 4.0
- 4) numpy: 1.19.5
- 5) pandas: 1.1.5
- 6) nltk: 3.2.5

4. Proposed Method (Naïve Bayes Classifier)

4.1: Methodology

Of recent, spam mail classification is normally handled by machine learning (ML) algorithms intended to differentiate between spam and non-spam messages. Machine learning algorithms achieve this by using an automatic and adaptive technique. Rather than depending on handcoded rules that are susceptible to the perpetually varying characteristics of spam messages, ML methods have the capacity to obtain information from a set of messages provided, and then use the acquired information to classify new messages that it just received. According to, ML algorithms have the capacity to perform better based on their experience. In this section we will review machine learning method that have been applied to spam detection.

4.1.1: Flow of Research

The research has been started by studying different types Machine learning algorithms. Then different type Naïve Bayes Classifiers were studied. Multinomial Naïve Bayes was chosen to solve classification problem. We trained the data in above using three Naïve Bayes Algorithms and according to their accuracy and precision scores we found that **Multinomial Naïve Bayes** was best among all. Further, we implemented our model using MultinomialNB Classifier and deployed our project.

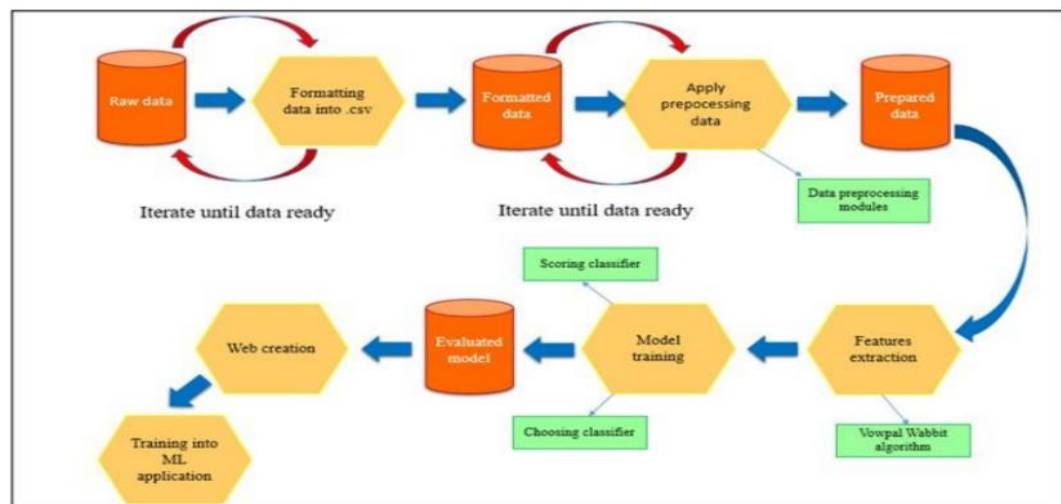


Fig. 5. Data flow of spam detection machine learning

4.1.2: Block Diagram

A block diagram is a short road map for that graphically represents how the data moves through the existing system. The block diagram shown in figure below has been used in design process. Data pre-processing followed by feature extraction has been done. Features have been extracted using a Natural Language Toolkit (NLTK) architecture explained in next sections. Model has been trained on given dataset. Text Message can be given using user interface and predicted output is shown.

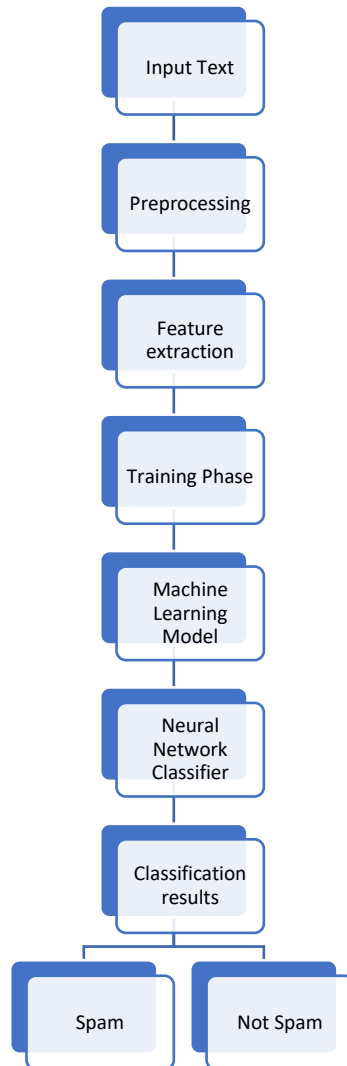


Fig. 6. Block Diagram

4.2: Data Cleaning

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning. Professional data scientists usually invest a very large portion of their time in this step because of the belief that **“Better data beats fancier algorithms”**.

If we have a well-cleaned dataset, there are chances that we can get achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large.

Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.



Fig. 7. Steps involved in Data Cleaning

Some data cleansing tools:

- Openrefine
- Trifacta Wrangler
- TIBCO Clarity
- Cloudfog
- IBM Infosphere Quality Stage

4.3: EDA (Exploratory Data Analysis)

Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. EDA focuses more narrowly on checking assumptions required for model fitting and hypothesis testing. It also checks while handling missing values and making transformations of variables as needed.

EDA build a robust understanding of the data, issues associated with either the info or process. it's a scientific approach to get the story of the data.

TYPES OF EXPLORATORY DATA ANALYSIS:

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

4.4: Data Pre-processing

Data pre-processing (also known as data cleaning, data wrangling or data munging) is the process by which the data is subjected to various checks and scrutiny in order to remedy issues of missing values, spelling errors, normalizing/standardizing values such that they are comparable, transforming data (e.g., logarithmic transformation), etc.

“Garbage in, Garbage out.”

— George Fuechsel

As the above quote suggests, the quality of data is going to exert a big impact on the quality of the generated model. Therefore, to achieve the highest model quality, significant effort should be spent in the data pre-processing phase. It is said that data pre-processing could easily account for 80% of the time spent on data science projects while the actual model building phase and subsequent post-model analysis account for the remaining 20%.

Whenever we have textual data, we need to apply several pre-processing steps to the data to transform words into numerical features that work with machine learning algorithms. The pre-processing steps for a problem depend mainly on the domain and the problem itself, hence, we don't need to apply all steps to every problem. We will be using the NLTK (Natural Language Toolkit) library here.

Steps involve in Data Pre-processing:

- Text lower case
- Tokenization
- Remove special characters
- Removing default Stopwords and Punctuations
- Stemming
- Visualization (ham and spam words)

4.5: Model Building

The initial step in building a machine learning model is to understand the need for it in our project. The machine learning development process can be resource intensive, so clear objectives should be agreed and set at the start. Clearly define the problem that a model needs to solve and what success looks like. A deployed model will bring much more value if it's fully aligned with the objectives of our project.

4.5.1: Data Splitting

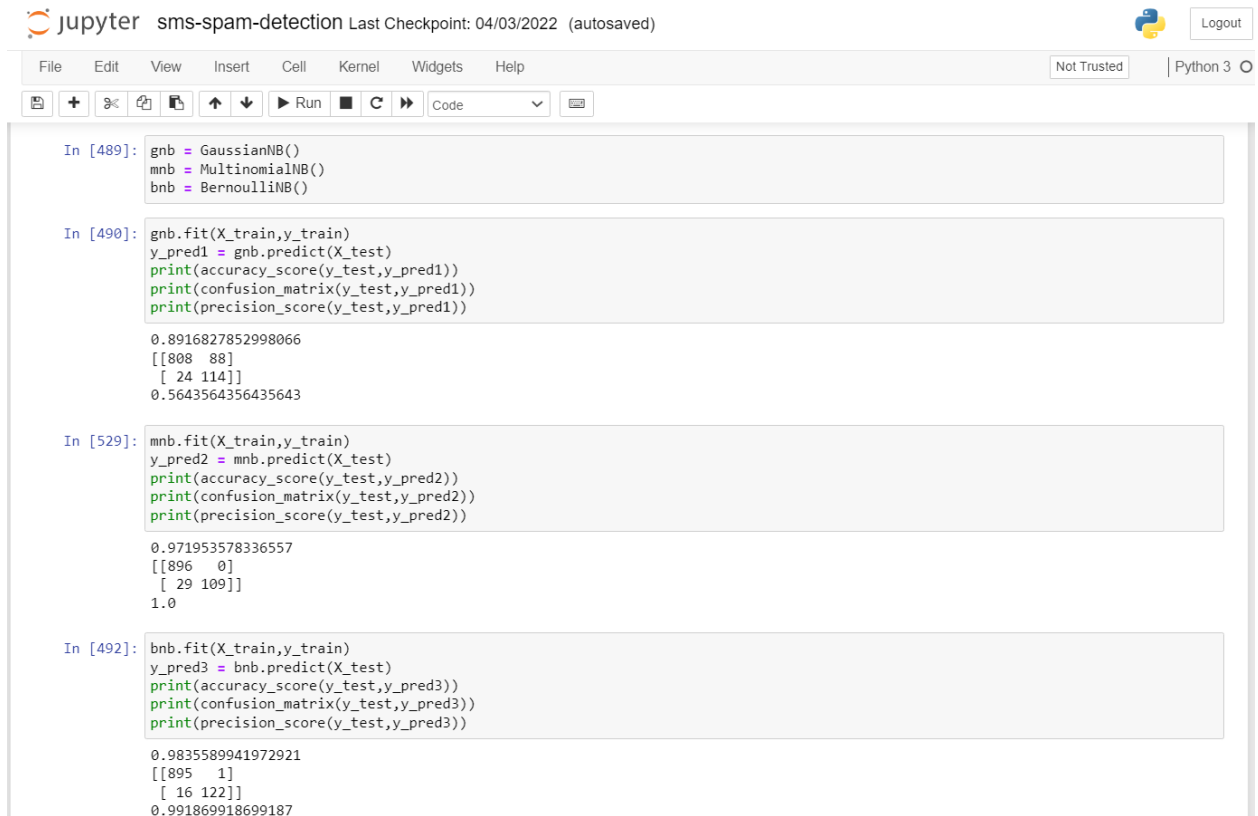
In the development of machine learning models, it is desirable that the trained model perform well on new, unseen data. In order to simulate the new, unseen data, the available data is subjected to data splitting whereby it is split to 2 portions (sometimes referred to as the train-test split). Particularly, the first portion is the larger data subset that is used as the training set (such as accounting for 80% of the original data) and the second is normally a smaller subset and used as the testing set (the remaining 20% of the data). It should be noted that such data split is performed once.

Next, the training set is used to build a predictive model and such trained model is then applied on the testing set (i.e., serving as the new, unseen data) to make predictions. Selection of the best model is made on the basis of the model's performance on the testing set and in efforts to obtain the best possible model, hyperparameter optimization may also be performed.

4.5.2: Training Data

We trained the data in all three Naïve Bayes Algorithms named GaussianNB, MultinomialNB and BernoulliNB and according to their accuracy and precision scores we found that **Multinomial Naïve Bayes** was best among all. After that we imported different algorithms (classifiers) from scikit-learn library and trained our data through these models as well. In a due

course, after comparing their precision and accuracy with each other, we still found MultinomialNB worth considering for further procedure.



```
In [489]: gnb = GaussianNB()
mnb = MultinomialNB()
bnb = BernoulliNB()

In [490]: gnb.fit(X_train,y_train)
y_pred1 = gnb.predict(X_test)
print(accuracy_score(y_test,y_pred1))
print(confusion_matrix(y_test,y_pred1))
print(precision_score(y_test,y_pred1))

0.8916827852998066
[[808  88]
 [ 24 114]]
0.5643564356435643

In [529]: mnb.fit(X_train,y_train)
y_pred2 = mnb.predict(X_test)
print(accuracy_score(y_test,y_pred2))
print(confusion_matrix(y_test,y_pred2))
print(precision_score(y_test,y_pred2))

0.971953578336557
[[896  0]
 [ 29 109]]
1.0

In [492]: bnb.fit(X_train,y_train)
y_pred3 = bnb.predict(X_test)
print(accuracy_score(y_test,y_pred3))
print(confusion_matrix(y_test,y_pred3))
print(precision_score(y_test,y_pred3))

0.9835589941972921
[[895  1]
 [ 16 122]]
0.991869918699187
```

Fig. 8. Training Data through all three classifiers

```

jupyter sms-spam-detection Last Checkpoint: 04/03/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3
In [494]: from sklearn.linear_model import LogisticRegression
          from sklearn.svm import SVC
          from sklearn.naive_bayes import MultinomialNB
          from sklearn.tree import DecisionTreeClassifier
          from sklearn.neighbors import KNeighborsClassifier
          from sklearn.ensemble import RandomForestClassifier
          from sklearn.ensemble import AdaBoostClassifier
          from sklearn.ensemble import BaggingClassifier
          from sklearn.ensemble import ExtraTreesClassifier
          from sklearn.ensemble import GradientBoostingClassifier
          from xgboost import XGBClassifier

In [495]: svc = SVC(kernel='sigmoid', gamma=1.0)
          knc = KNeighborsClassifier()
          mnb = MultinomialNB()
          dtc = DecisionTreeClassifier(max_depth=5)
          lrc = LogisticRegression(solver='liblinear', penalty='l1')
          rfc = RandomForestClassifier(n_estimators=50, random_state=2)
          abc = AdaBoostClassifier(n_estimators=50, random_state=2)
          bc = BaggingClassifier(n_estimators=50, random_state=2)
          etc = ExtraTreesClassifier(n_estimators=50, random_state=2)
          gbdt = GradientBoostingClassifier(n_estimators=50, random_state=2)
          xgb = XGBClassifier(n_estimators=50, random_state=2)
    
```

Fig. 9. Training Data through different classifiers/models imported from sklearn

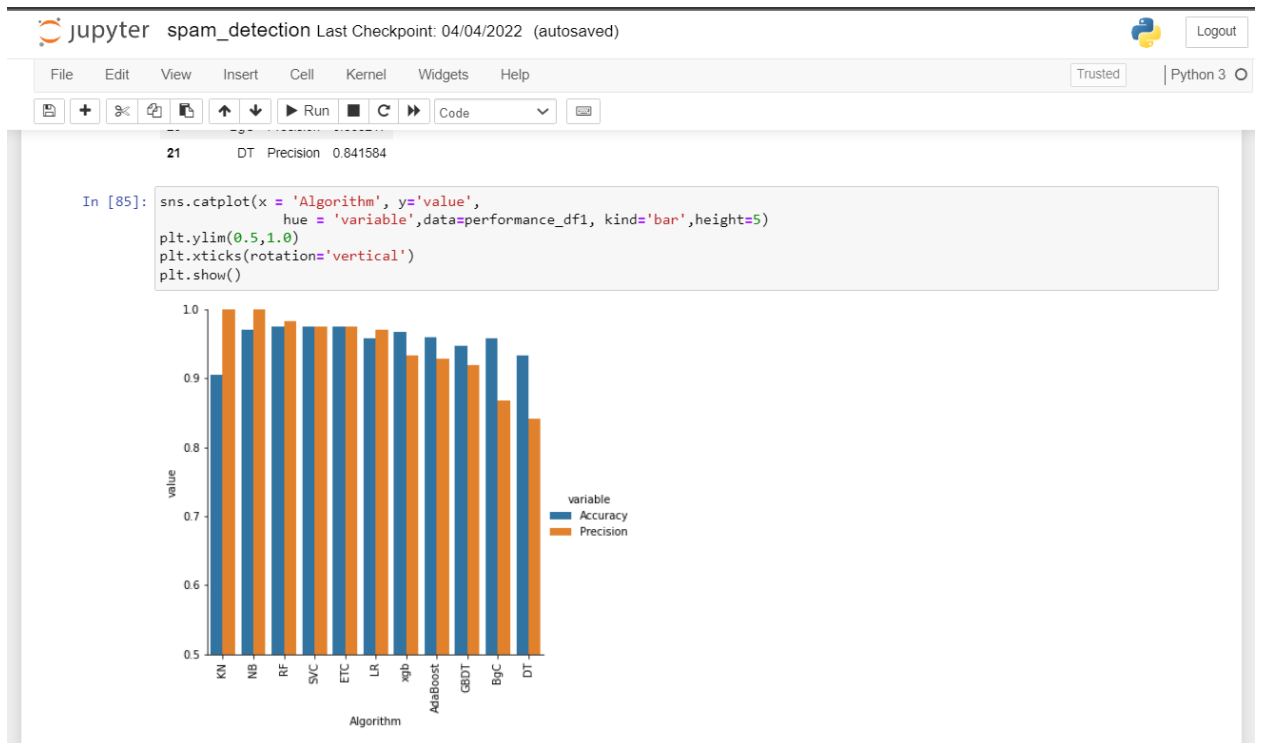


Fig. 10. Comparing precision and accuracy of different classifiers/models

5. Experimental Results

5.1: Performance evaluation measures

Spam filters are usually evaluated on large databases containing ham and spam messages that are publicly available to users. An example of the performance measures that are used is classification accuracy (Acc). It is the comparative number of messages rightly classified; the percentage of messages rightly classified is used as an added measure for evaluating performance of the filter. It has however been highlighted that using Accuracy as the only performance indices is not sufficient. Other performance metrics such as recall, precision and derived measures used in the field of information retrieval must be considered, so also is false positives and false negatives used in decision theory. This is very important because of the costs attached to misclassification. When a spam message is wrongly classified as ham, it gives rise to a somewhat insignificant problem, because the only thing the user need to do is to delete such message.

5.2: Analysis

We trained our data on MultinomialNB Classifier model and found accuracy of nearly 97%. After model building, we suspected that our user interface would give accuracy of more than 90%. For finding accuracy we analysed around 30 spam text messages by providing it to the user interface for detecting the accuracy of the spam detection system. Among these 30 spam text messages, 90% (27) of them came to be positive (spam detected).

Message	Spam (1/0)	Accuracy (%)
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&Cs apply 08452810075over18's	1	
WINNER!! As a valued network customer you have been selected to receive a \$900 prize reward! To claim call 09001701461. Claim code KL341. Valid 12 hours only.	1	
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	1	
SIX chances to win CASH! From 100 to 20,000 pounds txt- CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info	1	
England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try-WALES, SCOTLAND 4txt/1.20 POBoxOx36504W45WQ 16+	1	
Thanks for your subscription to Ringtone UK your mobile will be charged \$5/month Please confirm by replying YES or NO. If you reply NO you will not be charged	1	
SMS ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice Hockey. Correct or Incorrect? End? Reply END SPTV	0	
Congrats! 1 year special cinema pass for 2 is yours. call 09061209465 now! C Suprman V, Matrix3, StarWars3, etc all 4 FREE! bx420-ip4-5we. 150pm. Dont miss out!	1	
Please call our customer service representative on 0800 169 6031 between 10am-9pm as you have WON a guaranteed \$1000 cash or \$5000 prize!	1	
Your free ringtone is waiting to be collected. Simply text the password 'MIX*' to #5069 to verify. Get Usher and Britney. FML	0	
PRIVATE! Your 2004 Account Statement for 07742679699 shows 786 unredeemed Bonus Points. To claim call 08719180248 Identifier Code: 45239 Expires	1	
URGENT! Your Mobile No. was awarded \$2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM	1	
Want 2 get laid tonight? Want real Dogging locations sent direct 2 ur mob? Join the UK's largest Dogging Network bt Txtng GRAVEL to 69888! Nt. cc2a. 31p.msg@150p	1	
You'll not rcv any more msgs from the chat svc. For FREE Hardcore services text GO to: 69988 If u get nothing u must Age Verify with yr network & try again	1	
Congratulations ur awarded 500 of CD vouchers or 125gift guaranteed & Free entry 2 100 wkly draw txt MUSIC to 87066 Txts www.ldeco.com/wmi150ppm3age16	1	
We tried to contact you re reply to our offer of a Video Handset? 750 anytime networks mins? UNLIMITED TEXT? Camcorder? Reply or call 08000930705 NOW	1	27
December only! Had your mobile 11mths+? You are entitled to update to the latest colour camera mobile for Free! Call The Mobile Update Co FREE on 08002986906	1	90
4mths half price Orange line rental & latest camera phones 4 FREE. Had your phone 11mths? Call MobilesDirect free on 08000938767 to update now! or2stoptxt	1	
Loan for any purpose \$500 - \$75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1950609 or text back 'help'	1	
FREE MESSAGE Activate your 500 FREE Text Messages by replying to this message with the word FREE For terms & conditions, visit www.07781482378.com	1	
Someone has contacted our dating service and entered your phone because they fancy you! To find out who it is call from a landline 09111032124 . PoBox12n146f150p	1	
22 days to kick off! For Euro2004 U will be kept up to date with the latest news and results daily. To be removed send GET TXT STOP to 83222	1	
88800 and 89034 are premium phone services call 08718711108	1	
Do you realize that in about 40 years, we'll have thousands of old ladies running around with tattoos?	0	
Romantic Paris. 2 nights, 2 flights from \$79 Book now 4 next year. Call 08704439880T&Cs apply	1	
Orange customer, you may now claim your FREE CAMERA PHONE upgrade for your loyalty. Call now on 0207 153 9996. Offer ends 14thMarch. T&Cs apply. Opt-out availa	1	
FREE for 1st week! No1 Nokia tone 4 ur mobile every week just txt NOKIA to 8077 Get txting and tell ur mates. www.getzed.co.uk POBox 36504 W45WQ 16+ norm150p/tone	1	
Dear Voucher Holder, To claim this weeks offer, at you PC please go to http://www.e-tlp.co.uk/expressoffer T&Cs apply. To stop texts, txt STOP to 80002	1	
Got what it takes 2 take part in the WRC Rally in Oz? U can with Lucozade Energy! Text RALLY LE to 61200 (25p), see packs or lucozade.co.uk/wrc & itcould be u!	1	
As a registered optin subscriber u draw a \$100 gift voucher will be entered on receipt of a correct ans to 80002 Whats No1 in the BBC charts	1	

5.3: Screenshots of Result

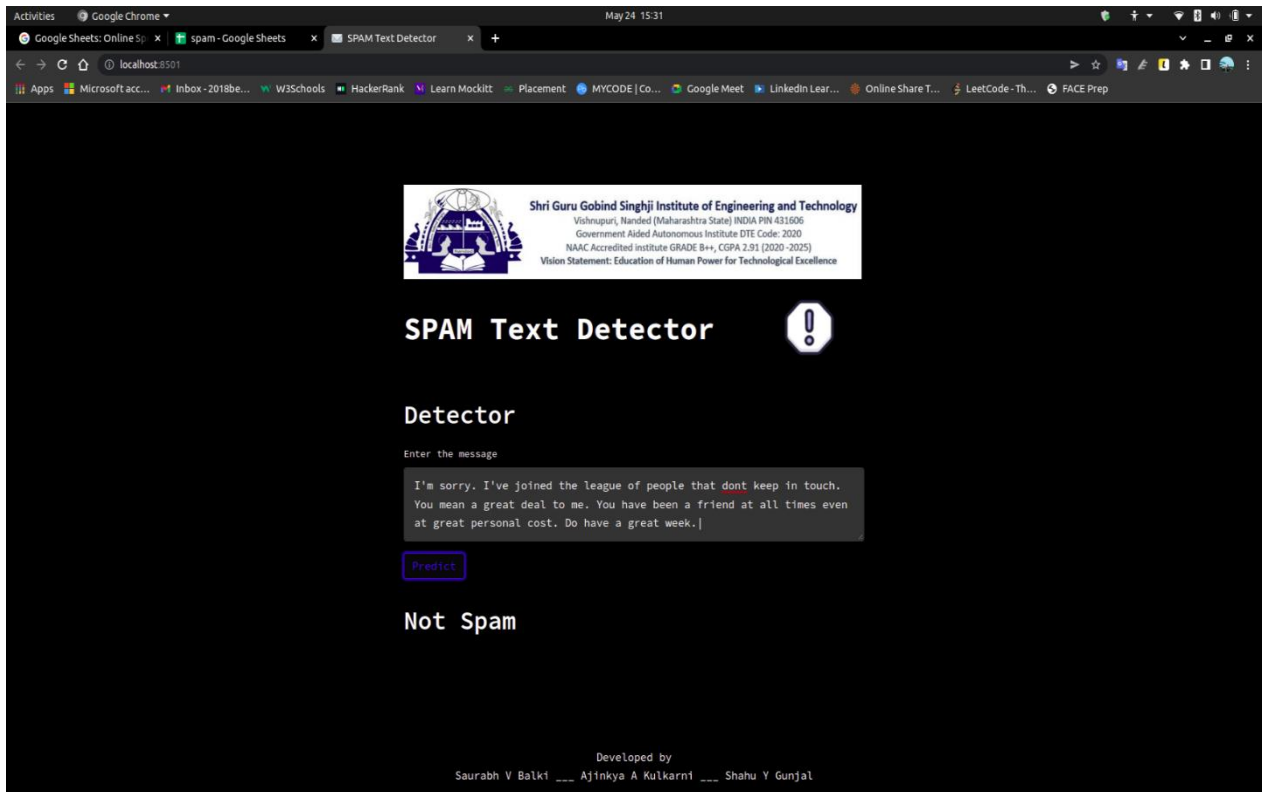
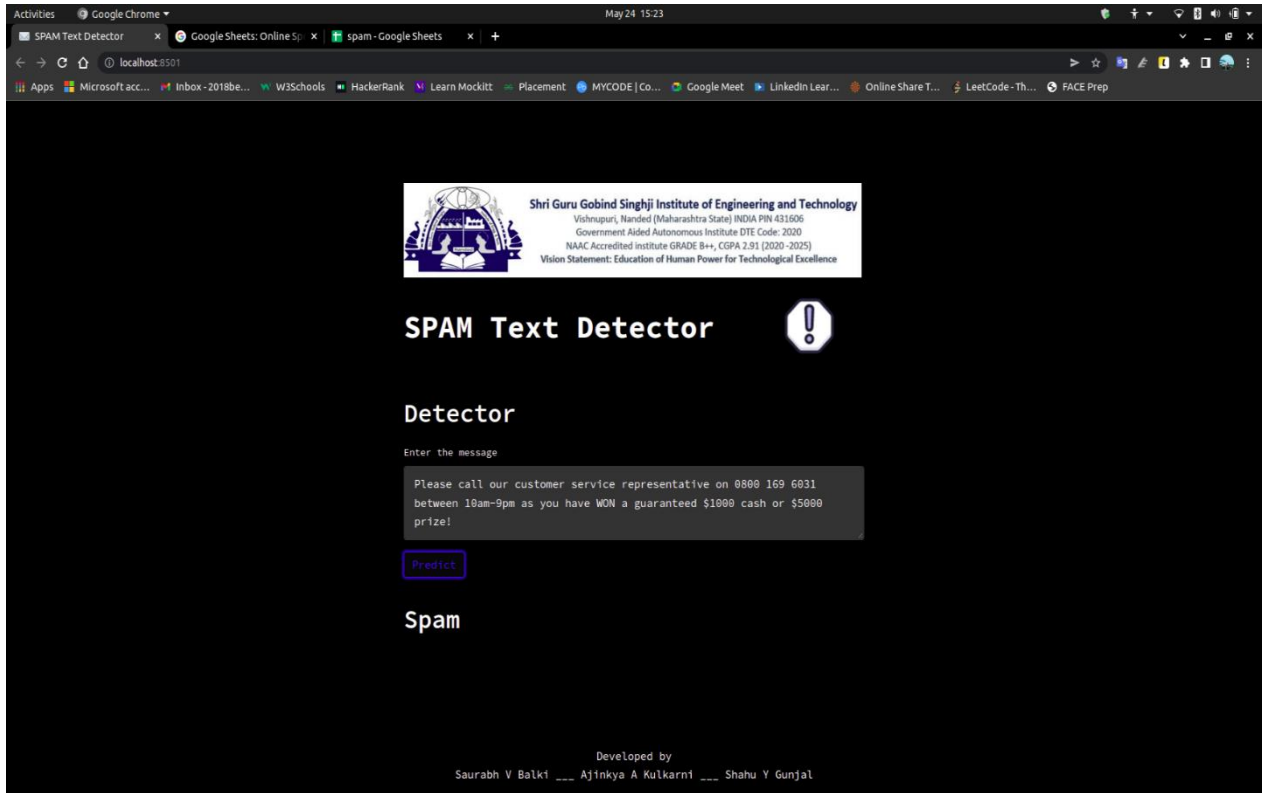
Shri Guru Gobind Singhji Institute of Engineering and Technology
 Vishnupuri, Haridwar (Maharashtra State) INDIA PIN 431506
 Government Aided Autonomous Institute DTE Code: 2000
 NAAC Accredited institute GRADE B++, CGPA 2.91 (2020-2025)
 Vision Statement: Education of Human Power for Technological Excellence

SPAM Text Detector

Enter the message

Predict

Developed by
 Saurabh V Balki ___ Ajinkya A Kulkarni ___ Shahu Y Gunjal



6. Conclusions and Future Scope

6.1: Conclusion

In this study, we reviewed machine learning approaches and their application to the field of spam filtering. Naïve Bayes algorithms been applied for classification of messages as either spam or ham. The attempts made by different researchers to solving the problem of spam through the use of machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of the machine learning algorithms in efficiently handling the menace of spam were pointed out and comparative studies of the machine learning technics available in literature was done. We also revealed some open research problems associated with spam filters. In general, the figure and volume of literature we reviewed shows that significant progress have been made and will still be made in this field. Having discussed the open problems in spam filtering, further research to enhance the effectiveness of spam filters need to be done. This will make the development of spam filters to continue to be an active research field for academician and industry practitioners researching machine learning techniques for effective spam filtering.

6.2: Future Scope

Review spam detection is essential since it can ensure justice for the sellers and retain the trust of the buyer on the online stores. The algorithms developed so far have not been able to remove the requirement of manual checking of the reviews. Hence there is scope for complete automation of spam detection systems with maximum efficiency. With growing popularity of online stores, the competition also increases. The spammers get smarter day by day and spam reviews become untraceable. It is necessary to identify the spamming techniques in order to produce counter algorithms

References

- [1] <https://www.javatpoint.com/machine-learning>
- [2] <https://www.javatpoint.com/supervised-machine-learning>
- [3] <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
- [4] <https://www.upgrad.com/blog/multinomial-naive-bayes-explained/>
- [5] <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>
- [6] <https://www.sciencedirect.com/topics/computer-science/bayesian-classifier>
- [7] <https://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/>
- [8] <https://jupyter.org/>
- [9] <https://www.jetbrains.com/pycharm/>
- [10] [https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20\(NLP\)%20refers,same%20way%20human%20beings%20can.](https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.)