

CATEGORIZED CLASSIFICATION OF TEXTUAL SPAM EMAILS USING DATA MINING

Ms. Sobiavani Gaushikan
169335 H



Dissertation submitted to the Faculty of Information Technology, University of Moratuwa,
Sri Lanka, for the partial fulfillment of the requirements of Degree of Master of Science in
Information Technology

December 2020

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text, and a list of references is given.

Ms. Sobiavani Gaushikan

.....

Date:

Supervised by:

Mr. Saminda Premaratne
Senior Lecturer,
Faculty of Information Technology,
University of Moratuwa

.....

Date:

Acknowledgment

First and foremost, glory and gratitude to Lord, the Merciful, for His showers of blessings during my research work to complete this research effectively.

I would like to express my profound and heartfelt thanks to my research supervisor, Mr.S.C. Premaratne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, to give me the chance to do research and offer support and help during this research. His solidity, optimism, honesty and inspiration have motivated me greatly. He also taught me the techniques to carry out the study and to outlines the research work as briefly as possible. It was a great pleasure and honor for him to work and learn under his supervision. I am deeply thankful for what he has given me to do.

I am most gratitude to my beloved parents for their encouragement, prayers, kindness, and contributions for better preparing me for my career. I am so grateful to my husband for his affection, compassion, prayer, and continuous cooperation in order to complete this research work. I would also like to thank my friends and associates for their help in completing this study in several forms.

I extend my appreciation to the Head and staff of the Center for Information and Communication Technology, Eastern University, for their assistance in my research work.

Finally, my thanks go to all the persons who contributed me in finishing my research study, either knowingly or unknowingly.

Abstract

In the last few years, to boost the connectivity and the safety of the people in the world, the Internet has built various platforms. Across from them, Email is a substantial platform for people's interaction. Email is an automated and efficient message transmission structure used to convey information from one person to another. And also, this appliance preserves plenty of money and time. Apart from that, Emails also have been abused by some assailants. The most commonly known one is Email Spam.

Spam Emails are referred to as unwanted materials, and they are unrequested business Emails or forged Emails forwarded to specific personnel or sent to an Organization or spread among a set of people. For spam Email, users face many difficulties, such as increment of traffic, restricting the data processing duration and the storage area, consuming more time of users, and threatening user protection. Therefore, it is essential to have an appropriate Email filtering approach to secure Email. There has been plenty of general spam Email filtering systems and various research people's various efforts to classify Emails into ham (genuine Emails) or spam (fake Emails) using Machine Learning techniques.

Unfortunately, most of the spam Email filtering solutions proposed so far are focused only on binary classification. However, classifying the already detected spam Emails into different types of categories is not performed yet.

This research explored and proposed an effective system to categorize the Textual Spam Emails into different categories using Data mining. First, The Multi-Nominal Naïve Bayes classifier is applied to distinguish Emails into two groups, such as ham and spam. The Grouping algorithm is used to categorize the spam Emails, which are already obtained from the Naïve Bayes classifier, into distinct categories. (Finance, Health, Marketing, etc...)

Finally, the proposed System's performance was evaluated with the Model Evaluation Techniques: Confusion matrix, Accuracy, Precision, Recall, F1 score.

Table of Contents

Declaration.....	i
Acknowledgment.....	ii
Abstract.....	iii
Table of Contents.....	iv
Abbreviations.....	vii
List of Figures.....	viii
List of Tables.....	x
Chapter 1.....	1
1 Introduction.....	1
1.1 Prolegomena.....	1
1.2 Background.....	1
1.3 Problem Statement.....	3
1.4 Aim and Objectives.....	4
1.4.1 Aim.....	4
1.4.2 Objectives.....	4
1.5 Proposed solution.....	4
1.6 Overview of the Report.....	4
1.7 Summary.....	5
Chapter 2.....	6
2 Literature Review.....	6
2.1 Introduction.....	6
2.2 Strategies to send Spam Emails.....	6
2.3 Approaches to control Spam Emails.....	8
2.4 Machine Learning Techniques to filter spam Emails.....	9
2.5 Summary.....	12
Chapter 3.....	13
3 Technology adapted on Spam Email classification.....	13
3.1 Introduction.....	13
3.2 Data mining.....	13
3.3 Machine Learning in Spam Email classification.....	15
3.4 Involvement of Data mining on Spam Email classification.....	15
3.5 Naïve Bayes algorithm.....	15
3.5.1 Multinomial Naïve Bayes classifier.....	17
3.6 TF-IDF vectorizer.....	17

3.7	Summary	18
Chapter 4	19
4	A novel approach for Spam Emails filtering	19
4.1	Introduction.....	19
4.2	Hypothesis	19
4.3	Input	19
4.4	Process	19
4.5	Output	19
4.6	Summary	20
Chapter 5	21
5	Analysis and Design for the Proposed System	21
5.1	Introduction.....	21
5.2	Framework for Proposed System.....	21
5.3	Data collection	22
5.4	Data Pre-processing	23
5.4.1	Data cleaning	23
5.4.2	Normalization	23
5.5	Feature Extraction.....	23
5.6	Classification.....	24
5.7	Evaluation	24
5.8	Summary	24
Chapter 6	25
6	Implementation of the Proposed System	25
6.1	Introduction.....	25
6.2	Python	25
6.3	PyCharm	25
6.4	NLTK.....	26
6.5	Data collection	26
6.6	Pre-processing.....	27
6.6.1	Importing the libraries	27
6.6.2	Importing the Dataset	28
6.6.3	Cleaning the Data	29
6.6.4	Normalizing the Data	30
6.7	Feature Extraction.....	30
6.8	Split the Dataset into training & testing sets.....	31
6.9	Create and Train the Multinomial Naïve Bayes classifier	31
6.10	Test the Data	31

6.11	Predict the spam Email	32
6.12	Categorize the spam Emails	32
6.13	Summary	33
Chapter 7		34
7	Evaluation	34
7.1	Introduction	34
7.2	Evaluation Techniques used for the classification models	34
7.2.1	Accuracy	34
7.2.2	Precision	34
7.2.3	Recall	35
7.2.4	Confusion matrix	35
7.2.5	F1 score	35
7.3	Evaluate the System on the train Dataset	36
7.4	Evaluate the System on the test Dataset	37
7.5	Experimental evaluation on spam Email classification and Results	39
7.6	Evaluation on spam Email categorization	39
7.7	Experimental evaluation on spam Email categorization and Results	41
7.8	Summary	41
Chapter 8		42
8	Conclusion and Further Work	42
8.1	Introduction	42
8.2	Overview of the research	42
8.3	Limitations	42
8.4	Further developments	43
8.5	Summary	43
Chapter 9		44
References		44
Appendix A		46
Appendix B		47
Appendix C		49
Appendix D		50

Abbreviations

IP	Internet Protocol
ML	Machine Learning
KDD	Knowledge Discovery in Databases
NLP	Natural Language Processing
TF-IDF	Term Frequency Inverse Document Frequency
IDE	Integrated Development Environment
NLTK	Natural Language Toolkit
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

List of Figures

Figure 1.1: Basic Email structure.....	1
Figure 1.2: Ham Email.....	2
Figure 1.3: Spam Email	2
Figure 1.4: Spam Email filtration	3
Figure 2.1: A taxonomy of Email spam filtering techniques.....	8
Figure 3.1: KDD Process	14
Figure 3.2: Data mining models and tasks.....	14
Figure 3.3: Naïve Bayes algorithm	16
Figure 3.4: Bayes Theorem.....	17
Figure 3.5: TF-IDF Algorithm.....	18
Figure 5.1: System Model for Spam Email classification	21
Figure 5.2: Kaggle Repository Website.....	22
Figure 5.3: HubSpot website.....	22
Figure 6.1: PyCharm IDE	26
Figure 6.2: Email Dataset	27
Figure 6.3: Importing Dataset	28
Figure 6.4: Visualization of Dataset	28
Figure 6.5: Ham Email Word Cloud.....	29
Figure 6.6: Spam Email Word Cloud	29
Figure 6.7: Normalization.....	30
Figure 6.8: Feature extraction	30
Figure 6.9: Data after Feature Extraction	31
Figure 6.10: Train Multinomial Naive Bayes classifier.....	31
Figure 6.11: Test the data.....	31
Figure 6.12: Predict the spam Email.....	32
Figure 6.13: Google spreadsheet with Different types of spam Email keywords	32
Figure 7.1: Accuracy equation	34
Figure 7.2: Precision equation	34
Figure 7.3: Recall equation	35
Figure 7.4: Confusion matrix	35
Figure 7.5: F1 score equation.....	35

Figure 7.6: Python code for evaluation on training Dataset	36
Figure 7.7: Output for the evaluation on training Dataset	36
Figure 7.8: Graph of evaluation on training Dataset.....	37
Figure 7.9: Python code for evaluation on testing Dataset	37
Figure 7.10: Output for the evaluation on testing Dataset	38
Figure 7.11: Graph of evaluation on testing Dataset	38
Figure 7.12: Calculate the Accuracy of the spam Email categorization.....	39
Figure 7.13: Output of the evaluation for spam Email categorization.....	39
Figure 7.14: Graph for spam Email category prediction Accuracy	40
Figure 7.15: Comparison between the human expert and the system in spam Email categorization.	41

List of Tables

Table 2.1: Strategies used by spammers to send spam Emails	7
Table 2.2: Summary of ML Approaches for Spam Email Filtering	12
Table 6.1: Python Libraries.....	27
Table 7.1: Spam Email classifier Accuracy on Data size	39
Table 7.2: Prediction accuracy for different no of keywords	40

1 Introduction

1.1 Prolegomena

Every day, millions of people worldwide access the Email for Communication and Commercial purpose. Spam attack is the most serious threat for financial as well as non-financial Organizations. It produces a loss of work productivity and traffic bottleneck, and spam with a fraudulent purpose also risk the security and privacy of those who receive them. Most of the existing email filtering practices that have been found to control this spam attack were only used binary classification. Therefore, we propose using Data mining techniques to filter the spam Emails and classify those spam Emails into different categories. Our System predicts whether the newly entered Email is spam or non-spam depending upon the Email's content. Also, the Grouping algorithm is used to check the results of the previous classification and identifying the spam category.

1.2 Background

Internet becoming a global communication infrastructure over the past years. It offers various technologies for human usage. Many user interactions occur through the Emails With the enhancement in technologies.

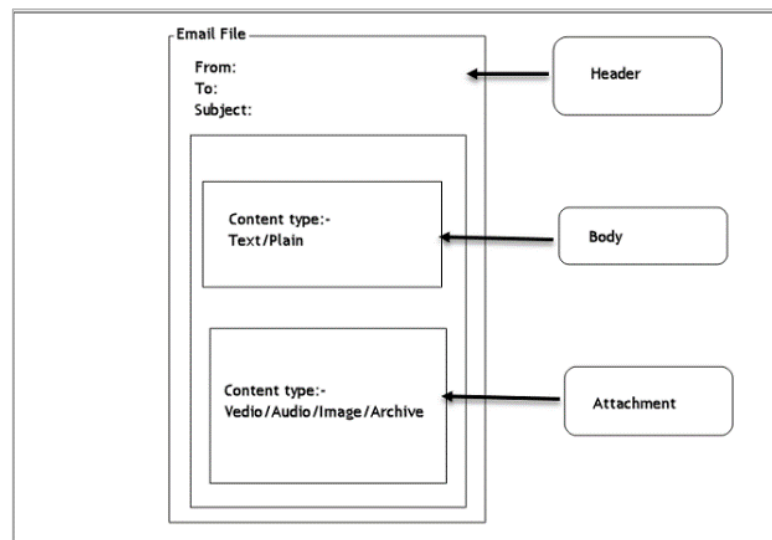


Figure 1.1: Basic Email structure

Email is simply a standardized electronic messaging framework that exchanges data between users through the Internet. It is an effective, fast, and cheap substantial platform for user interaction. The users can send their data anywhere else within a second by utilizing Emails. Therefore, it is the most popular medium and an essential part of human life.

```
This is a bit of a messy solution but might be useful -  
  
If you have an internal zip drive (not sure about external) and  
you bios supports using a zip as floppy drive, you could  
use a bootable zip disk with all the relevant dos utils.
```

Figure 1.2: Ham Email

An increase in Email usage in the last few decades will increase the spam Emails' growth. Email spam is called abandon electronic mail, and it is an undesirable message sent to the users in bulk via Email. Spam is probably created to wipe out time and resources. Text messages, Blogs, Internet forums, and Social networking websites may also have this spam. However, Email spam is the most dominant one and the most dangerous one for people. It can be an advertisement or some such explicit content that may have malicious code embedded in it.

```
IMPORTANT INFORMATION:  
  
The new domain names are finally available to the general public at  
discount prices. Now you can register one of the exciting new .BIZ or  
.INFO domain names, as well as the original .COM and .NET names for  
just $14.95. These brand new domain extensions were recently approved  
by ICANN and have the same rights as the original .COM and .NET  
domain names. The biggest benefit is of-course that the .BIZ and  
.INFO domain names are currently more available. i.e. it will be much  
easier to register an attractive and easy-to-remember domain name for  
the same price. Visit: http://www.affordable-domains.com today for  
more info.
```

Figure 1.3: Spam Email

Spam may be classified with reference to the Ferris Research [1]:

1. Teaching; such as Online course.
2. Medical; such as expired drugs.
3. Adult content; such as Sexual chat.
4. Advertisement Goods; such as phony stylish stuffs.
5. Marketing; such as erotic enrichment items.
6. Financial; such as Tax solutions, Loan packages.

7. Malware and viruses; such as Trojan horse's attack.
8. Political; such as presidential votes

Spam Emails urge less productivity improvement; Spreading bugs and resources that contain dangerous materials to a specific group of consumers; also have some slots in mail box; as a result of eliminating the steadiness of mail servers, user spend a lot of time on the company arriving Emails and removing unwanted Emails [2]. Recipients of spam Emails can cause inconvenience and economic damage. So, it is necessary to refine and split out them from the genuine Emails [3].

Conventional approaches to filtering out spam Emails such as black-white lists (Domains, IP addresses) are practically challenging. Applications available using Data mining techniques to an Email can grow up the effectiveness of a spam Email categorization. Approaches of textual clustering and classification were effectively used in spam Email issues since the last era. Due to unorganized records, senseless content, and a vast amount of textual Emails, automated phishing Emails categorization is problematic.



Figure 1.4: Spam Email filtration

1.3 Problem Statement

There are plenty of Email spam filtering methods are used for spam classification. A spam filter identifies and blocks spam Emails and their distribution to the Email box. Filters are used to moderate the anti-spam Email influence and can operate as an honest and anticipated tool to eliminate undesirable Emails. But there will be a little chance of incorrect sorting or removal of legitimate Emails. However, 100% accuracy in detecting spam Emails is doubtful. Therefore, the detection of System with the pre-processing

technique is a challenging mission. And also, Sorting the spam Emails into different categories is not performed yet by the previous researchers.

1.4 Aim and Objectives

1.4.1 Aim

The aim is to classify Spam Emails and categorize them into different types of groups using Data mining techniques.

1.4.2 Objectives

This research's key objectives are,

- To collect spam Email dataset.
- To normalize and extract the features using suitable procedures.
- To propose an efficient model of classifying spam Emails according to the maintenance of the case occurrences in Data mining techniques using the training Data set.
- To validate the design by predicting the testing Dataset.
- To identify the spam Email's category.
- To evaluate the performance of the spam Email classification.

1.5 Proposed solution

Textual Emails are used as an input Dataset in this research. Initially, the Dataset was normalized and extracted with the appropriate methods. Then, they were trained by the Naïve Bayes classifier to identify the efficient classification model. Later, the Email Dataset was tested with that classification model. This will forecast either an Email is a spam or ham. Finally, the outcome of the previous classification has been categorized by using the Grouping algorithm with spam Email Trigger keywords. This will show which spam group the spam Email belongs to.

1.6 Overview of the Report

This Chapter delivers the introduction for spam Email classification using Data mining techniques. The next Chapter describes the related work based on the topic done by other researchers. Chapter three summarizes the technology adapted, and Chapter four presents the approach of the research. Chapter five explains the analysis and design of

the research. Meanwhile, Chapter six is about the implementation part of the research component. Chapter seven presents the evaluations made on the methods used in the implementation. Chapter eight discuss the result, limitations, and future development for the solution. Finally, the last Chapter provides a list of references.

1.7 Summary

This Chapter has presented the area of investigation and the proposed approach for the spam Email classification.

2 Literature Review

2.1 Introduction

This Chapter reviews the different types of techniques and approaches previously done for spam Email filtering. This Chapter especially highlights on main deciding factors of spam filtering. Moreover, this Chapter summarizes most of the ML approaches with a list of different algorithms to apply data mining techniques.

2.2 Strategies to send Spam Emails

For certain forms of community communications, which are used most commonly by millions of individuals, persons, and organizations, Email is essential. It has also been vulnerable to attacks. Spam, also referred to as phishing Email or mass email, is the most prevalent such hazard. Spam emails are repeatedly sent in vast numbers to an unselective group of recipients with unnecessary email addresses, often with commercial material. And it is time-wasting, bandwidth, and computing capacity. Spam is omnipresent on the Internet because electronic messaging processing costs are much smaller than most other communication mediums.

Spammers also commonly capture email address information in the chat rooms and various websites and market it to other spammers. These spam notifications are scattered over many organizations, so the receiver frequently ignores spam charges. These spam emails intimidate the IT area very heavily and cause billions of dollars of loss of results. Spam Email is viewed as a significant security risk and has been used to steal confidential data in recent years. Spam Emails also outspread unprotected bugs among different individuals [4].

STRATEGIES TO SEND SPAM EMAILS	
Strategies	Description
Zombies or Botnets	Engaged PCs that have transmitted large amounts of spam, and ransomwares on the Internet.
Bayesian stealing and infecting	Write spam messages to avoid wording usually used in spam or poison the repository of the Bayesian filters.
IP address	Lending or utilizing a trustworthy or independent IP address
Offshore ISPs	Usage with international ISPs without protections
Exposed proxies	Negotiated servers for spam redirection to unexplored users.
Third-party mail back software	Using insecure mail back apps on harmless webpages
Falsified header information	Add faulty header in the spam mail
Obscuration	By fragmented words or messages with ludicrous HTML tags or other 'inventive' signs, the words are hidden in spam
Vertical slicing	Vertically write spam messages
HTML handling	Handling of HTML modules to avoid suspicion
HTML encryption	By using a Base64 encryption algorithm, a binary relation can be translated into human - readable letters
JavaScript messages	Within a fragment that is triggered when the message is accessed, the full content of the spam message is put.
ASCII art	Use standard letter symbols for writing spam messages
Image based	Usage of images for textual data transmission
URL address or redirect URL	Only provide the URL to circumvent detection/usage of costly "portals" to screen the websites
Encoded messages	Encrypt a mail that is decrypted only once the mailbox is input.

Table 2.1: Strategies used by spammers to send spam Emails

2.3 Approaches to control Spam Emails

In reality, spam management techniques like domain sender control, content search, relay ban, and IP address checking or domain name testing are possibly available for spam control. Anyhow, spammers can quickly circumvent these necessary steps with more refined spam variants to resist identification.

Many strategies were found to fire up this spam issue, and filtering is one of the most critical techniques. The aim of spam filtering is to automatically filter out unnecessary emails from the mailbox of a person. These unsolicited emails have also created some issues, including filling in mailboxes, crippling personal communications, worsening the network's bandwidth, absorbing user time and energy to filter, not just other spam-related problems.

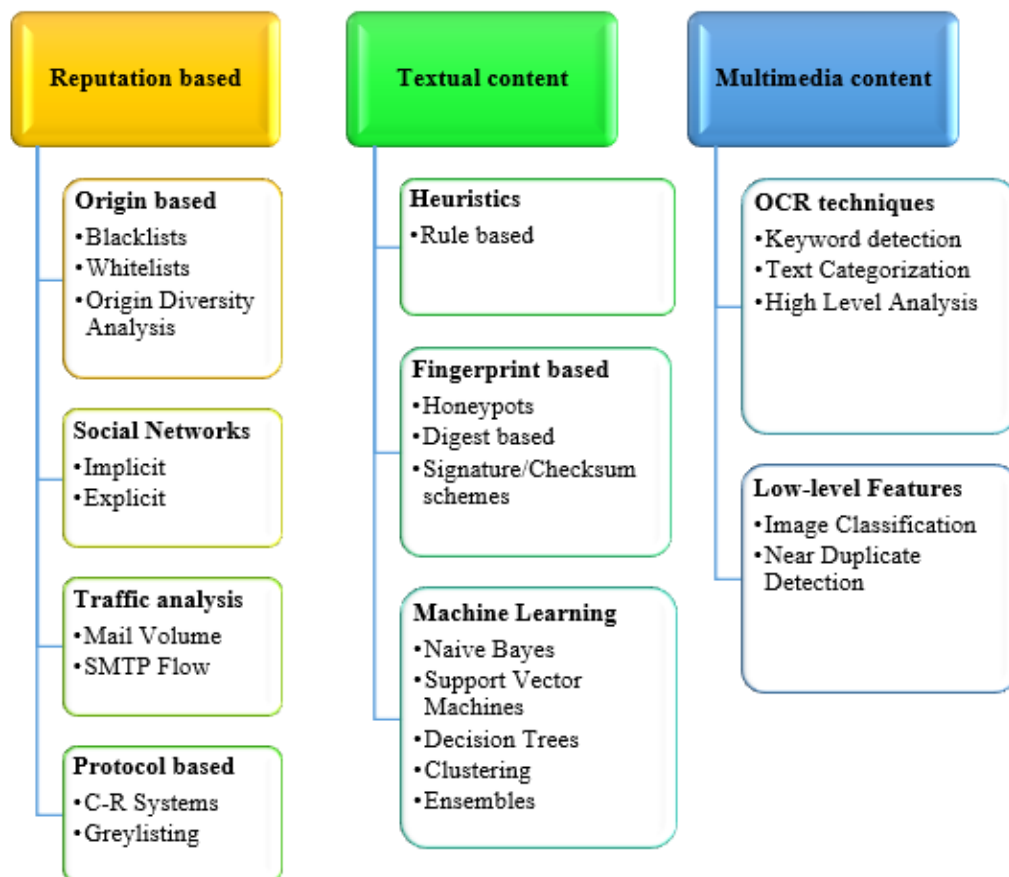


Figure 2.1: A taxonomy of Email spam filtering techniques.

Whitelist and blacklist were the first ways to block spam. This content-based technology detects message content vocabulary or styles that is authentic an Email whether it is non-spam or spam. Genuine Emails can be stored in a whitelist and a blacklist stores spam Emails. The Email is checked, and legitimate email messages are approved, although Spam Communications are blocked. Sadly, because the Email context is not included, certain legitimate emails may be blocked or blacklisted [5].

2.4 Machine Learning Techniques to filter spam Emails

Most spam analyses have concentrated solely on textual Emails [6]. Many clustering algorithms have been applied to detect spam communications. The paper [7] identified two methods of machine classification. Any rules defined by hand are the first technique. The typical case is the expert system focused on rules. When all modules are static and their elements are conveniently separated according to their structures, this form of clustering can be used. Using ML methods, the second method is completed. Via the criteria function, Paper [8] validates a spam message clustering problem. The criterion function is the high equivalence, defined by the k-nearest neighbor algorithm, between messages in groups. It introduces a genetic algorithm with a penalty feature to overcome spam classification.

In spam email classification, some experiments are being carried out with machine learning techniques. Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali [9] have shown that the Naïve Bayes Email classifier is enhanced for layer-3 processing without reunification. More than 117 million characteristics can be classified as entries with in a second with a number of alternatives. Symbiotic Data Mining (SDM) is referred as a revolutionary dispersed data mining approach [10], combination of Content-Based Filtering (CBF) with Collaborative Filtering (CF) is defined. so as to extend improved clustering while preserving anonymity, the purpose is to reuse local filters from different items. Spam Email sorting [11] was achieved by creating a bag-of-words file for each and every website by Linear Discriminant Analysis.

The Naïve Bayes classifier for spam classification was established in 1998 [12]. The Bayesian spam filtering [13] is a probabilistic filtering strategy. The Naïve Bayes classifier is used to detect spam emails. Bayesian classifiers function by combining the usage of spam and non-spam indexes to calculate the likelihood of a spam or non-spam Email using a Bayesian analysis.

The process of Naïve Bayes is on the basis of Bayesian approach, so it is easy, pure, and fast [14]. A classification of Naïve Bayes is a modest probability classifier with a high assumption of uniqueness. Simply stated, the appearance of a certain characteristic of a genus is unlike or absence of any other characteristic of a Naïve Bayes classification since the type variable depends on the essence of its prospect model, throughout the supervised learning state, the Naïve Bayes classifier is being trained. Just a small number of parameters are required for categorization from a limited number of training data sets in the Naïve Bayes classifier.

Javier Velasco-Mata, Francisco Janez-Martino, Santiago Gonzalez-Martínez, Eduardo Fidalgo [15] proposed an automatic spam multi-classification approach for the first time in history. In which, a Hierarchical clustering algorithm was used to classify already detected spam Emails into different categories.

Anyhow, none of the previous researches are found any efficient model to classify and categorize spam Emails at a time. To have an efficient approach for effective results, it is necessary to predict the spam Email from the given Dataset and identify the category of that spam Email.

Spam Email filtering techniques			
Year	Author	Approaches	Description
1999	Vapnik	SVM Approach	Binary representation is used for the best result.
2002	Soonthornphisaj	Centroid-Based approach	Using a KNN, Naïve Bayesian and Vector space model, the data models are represented
2002, 2003	Graham	Bayesian filter	99.5% of spam are caught with 0.03% false positives
2003	Woitaszek	Simple support vector machine with a personalized dictionary	This is a supplement for Microsoft Outlook XP that offers sort and community ability for the standard desktop Email user using Outlook's GUI.
2003	Clark	LINGER	It is an integrated network system with a multi-layer sensor.
2004	Matsumoto	Naïve Bayesian Classifier (NBC) and	TF and TF-IDF are used for the construction of features vector.

		Support Vector Machines (SVMs)	
2004	Ozgur	Artificial Neural network and Bayesian	The inputs to the networks are calculated through binary and probabilistic processes, which involve two ANN systems, a single-layer perceptron and a multi-layer perceptron. Three methods are used for Bayesian classification: linear, predictive, and enhanced probabilistic models.
2005	Zhao and Zhang	Rough Set Based Model	Used classical rough set theory.
2005	Chuan	LVQ-based Neural Network	Emails are divided into many subclasses for quick detection.
2006	Dong	Bayesian spam filter	Use a cross N-gram
2006	Wang	Perceptron and Winnow	Both are integrated together.
2007	Ichimura	Classification method based on the results of SpamAssassin	Proposed Spam classification self-organization map (SOM) and Dynamically defined category (ADG) for the extraction of proper judgment laws.
2007	Pang Xiu-Li	Based on support vector machine (SVM), Anti-spam filter solution	For word segmentation the Tri-gram language model has been extended with the discount smoothing algorithm.
2007	Yang and Elfayoumy	Feedforward backpropagation Neural network and Bayesian classifiers	The effectiveness of both classifiers was evaluated using accuracy and sensitivity metrics.
2008	Ye	Spam discrimination model	SVM and D-S Theory
2008	Lobato and Lobato	Binary classification approach	Based on Bayes Point Machines' expansion
2009	Sun	Spam filtering algorithm based on locality pursuit projection (LPP) and least square version of SVM(LS-SVM)	The email functionality is first derived from the LPP algorithm, later classified by the LS-SVM classifier.

2009	Meizhen	Spam behavior recognition model	Based on fuzzy decision tree (FDT)
2009	Yong	Intelligent Spam-based Fuzzy Clustering Framework	This System can work without training in advance.
2010	Na Songkhla and Piromsopa	The statistical rule-based spam detection system	Exchange regulations, however, periodically modified to counter the technique of spammers.
2010	Qiu	Online linear classifiers: Perceptron Winnow and Naïve Bayesian	They have a state-of-the-art spam filtering performance.

Table 2.2: Summary of ML Approaches for Spam Email Filtering

2.5 Summary

This Chapter addresses the literature observations and reviews the previous research works. It also illustrates different filtering techniques and efficient Data mining procedures used earlier to classify spam Emails.

3 Technology adapted on Spam Email classification

3.1 Introduction

This Chapter offers a definition of the Data mining technique, which we preferred to analyze and classify spam Emails in an efficient way. Additionally, this Chapter presents the utility of Data mining techniques that differentiates from the other technologies used in existing systems.

3.2 Data mining

Data mining is a set of methods that basically discovers the valuable information hidden in large piles of data. It is a combination of different disciplines such as Artificial Intelligence, Statistics, Machine Learning, Data Base Management System. It is an approach that assists in discovering new forms and orders of a Data set. Retrieve facts from such a large database and make them as accessible manner for the forthcoming activities are the key purposes of the Data mining approach.

The Data mining methodology is a programmed study of huge amounts of data to obtain a good pattern. The data mining technique has four types of relationships:

- i. Classes: The Class is used to put the data in predefined sets.
- ii. Clusters: Data items are gathered based on their logical correlation.
- iii. Sequential Patterns: Extracting the data to discover the model and the tendency.
- iv. Associations: The Associations are finding out by using Data mining on the Data set.

Data mining is referred as Knowledge Discovery in Databases (KDD), as Data mining is a vital part in discovering knowledge from the database. The KD process is a collaboration of all the steps exposed in Figure 3.1.

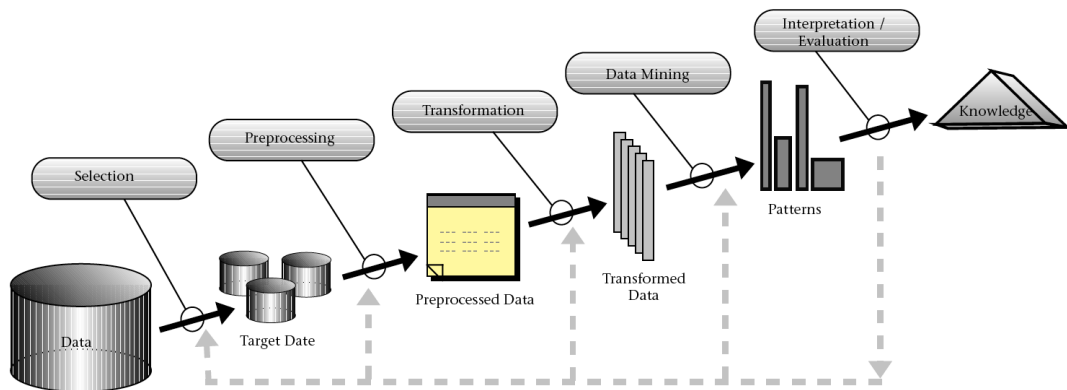


Figure 3.1: KDD Process

- **Data Cleaning** – The noisy and unpredictable data is eliminated.
- **Data Integration** – Different sources of data are linked together.
- **Data Transformation** – The data is translated into proper mining patterns.
- **Data Mining** – Extract data forms by applying Intelligent methods.
- **Pattern Interpretation** – Assessing data patterns.
- **Knowledge Presentation** – Knowledge is signified.

Data mining comprises various procedures to accomplish the anticipated tasks. They are categorized according to the intention of the algorithm for fitting a model to the data, as well as some searching methods are necessary for all algorithms. Figure 3.2 shows the model, which is divided as predictive and descriptive.

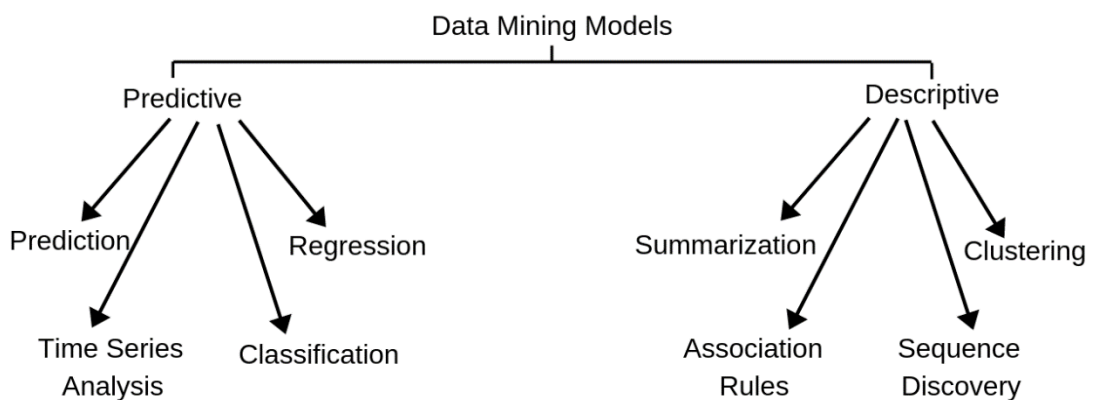


Figure 3.2: Data mining models and tasks

3.3 Machine Learning in Spam Email classification

Machine Learning is referred as an Artificial Intelligence's sub-set that intended for branded machines that can be obtained as humans. One attempts to expose unseen groups in the Dataset like spam Emails in unsupervised learning. Some structures may be the bags of words or the subject field investigation in spam Email clustering. Therefore, the input for spam Email categorization is observed as a 2D matrix, whose axes are the contents and the attributes.

Spam Email filtering methods are regularly split into various sub-methods. Data (Emails) gathering and interpretation are included in the first phase, Data feature selection, and feature extraction effort to reduce the dimensionality (number of attributes) for the further Mission Activities. The mapping between the training data set and the test data set will be identified at the end of the email clustering process [16].

3.4 Involvement of Data mining on Spam Email classification

Email data sets are large in size and have some complications. Therefore, these data can be treated as Big Data. It is an important term given to data that is huge in volume, variety, and velocity. Studies from previous research obviously show that Data mining is the best option for data analysis and clustering rather than other methods.

The data mining method helps obtain knowledge-based information and facilitates automatic forecast of tendencies and performances and automated detection of hidden patterns in spam classification. Users can analyze the vast amount of data within a smaller amount of time with this fastest procedure.

In the Email spam filtering, many Data mining Algorithms are used, including Naïve Bayes, K-means, K-Nearest Neighbor, Support Vector Machines, Neural Networks, and so on. According to the previous researches, the finest spam classification algorithm is Naïve Bayes algorithm.

3.5 Naïve Bayes algorithm

The Naïve Bayes algorithm is referred to as Bayes' Theorem's classification method with an individuality hypothesis in between predictors. In other words, a classification of Naïve Bayes believes a particular attribute of a class is totally distinct from any other attribute [17].

The Naïve Bayes model is conveniently developed and suitable to large sets of data. Also, Naïve Bayes has recognized better than extremely knowledgeable clustering approaches. Naïve Bayes Classifier can be trained without any difficulties and can be used as a standard model. Once the token selection is achieved appropriately, Naïve Bayes can perform well better than other existing models.

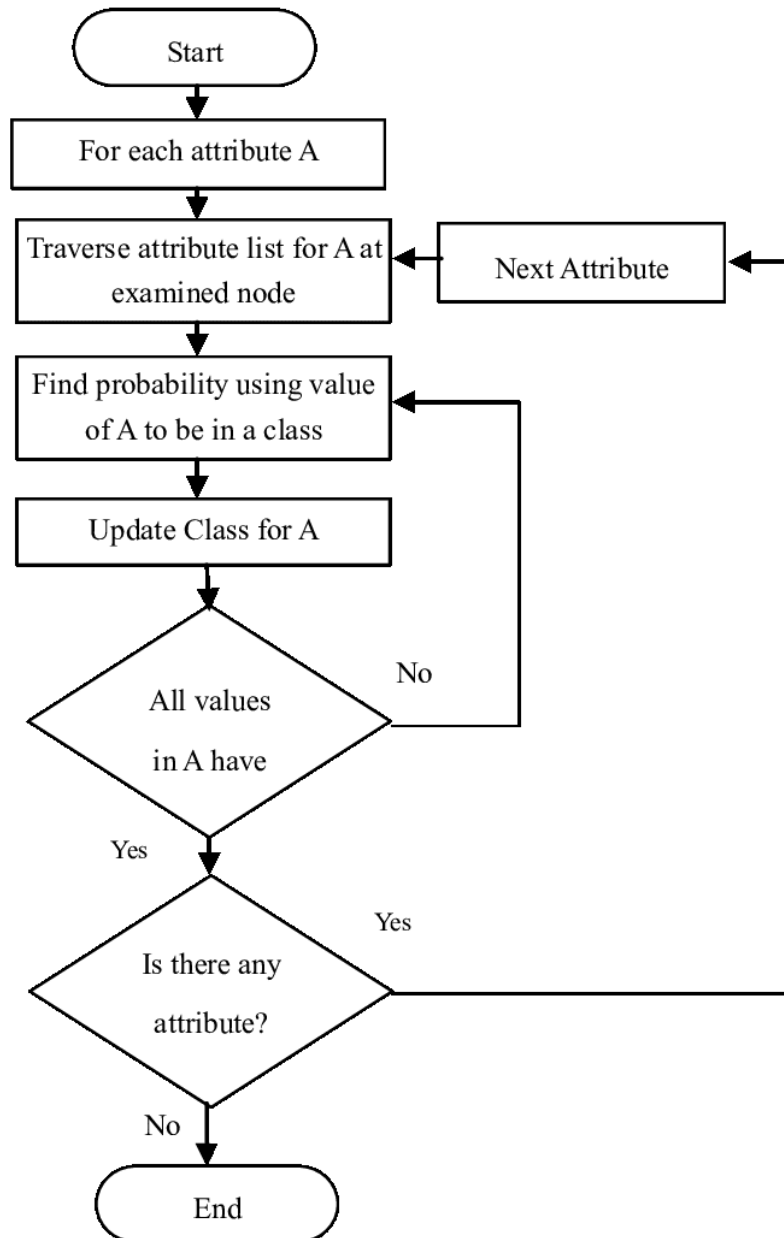


Figure 3.3: Naïve Bayes algorithm

Bayes theorem is a method to measure $P(c|x)$ from $P(c)$, $P(x)$, and $P(x|c)$ posterior probability. Figure 3.4 shows the equation for the Naïve Bayes algorithm.

The diagram shows the Bayes Theorem equation: $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$. Arrows point from labels to the corresponding terms in the equation: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$. Below the equation, the expanded form is given: $P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$.

Figure 3.4: Bayes Theorem

Above,

- $P(c|x)$ is the probability of “c” being true, given “x” is true.
- $P(c)$ is the probability of “c” being true.
- $P(x|c)$ is probability of “x” being true, given “c” is true.
- $P(x)$ is the probability of “x” being true.

3.5.1 Multinomial Naïve Bayes classifier

Multinomial Naïve Bayes classifier is among the two standard Naïve Bayes variations utilized in textual classification. It is appropriate for categorization with distinct features. The Multinomial Naïve Bayes classifier usually needs numeral feature counts. It considers a vector of a feature where the given word is defined by the number of times it occurs.

3.6 TF-IDF vectorizer

TF-IDF also called Term Frequency Inverse Document Frequency is a ubiquitous algorithm for translating the text into a meaningful number representation used to suit the prediction machine algorithm. The Naïve Bayes algorithms are being fed to the TF-IDF ranking, significantly increasing the performance of more simple methods such as word counts. This research helps to find out which are the most important words in both spam and non-spam Emails.

*TFIDF score for term i in document j = $TF(i,j) * IDF(i)$*

where

IDF = Inverse Document Frequency

TF = Term Frequency

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2 \left(\frac{\text{Total documents}}{\text{documents with term i}} \right)$$

and

t = Term

j = Document

Figure 3.5: TF-IDF Algorithm

3.7 Summary

This Chapter defines Data mining as a technology proposed to analyze and classify spam Emails. It describes how Data mining contributed to obtaining an efficient result for spam Email filtering. The next Chapter demonstrates the approach to classify spam Emails according to the technique adopted in this Chapter.

4 A novel approach for Spam Emails filtering

4.1 Introduction

In the previous Chapter, the technology adapted for analyzing and classifying spam Emails was explored. This Chapter presents our approach to analyze and classify spam Emails in detail using Data mining techniques. This Chapter also highlights the key features that separate our approach from the existing approaches for spam Email classification.

4.2 Hypothesis

Identification and classification of spam Emails can be made using Data mining techniques. Predictive Data mining can be used to identify spam emails with the bag of words features. Descriptive data mining can be used to explore the current situation demonstrated in climate.

4.3 Input

Data is an essential element before we develop any efficient system. Email dataset taken from the specific repository and the spam Email Trigger keywords collected from different websites are used as inputs for this research.

4.4 Process

In this process of spam Email classification with data mining techniques, all essential steps in the KDD process are carried out. The data set is cleaned, normalized, extracted, classified, and evaluated during the process. Naïve Bayes algorithm succeeded through associating the use of spam and non-spam Email words, and in order to analyze whether an email is spam or non-spam, Bayes' theorem was used. Then the Grouping algorithm is used to identify the spam category.

4.5 Output

The output of the research is the classification results obtained from the classifier and the Grouping algorithm. Naïve Bayes classifier classifies the newly entered Email into

the System, which result in whether Email is spam or ham. If spam is the result, then the Grouping algorithm is used to identify the category of that spam.

4.6 Summary

This Chapter included the outline of our approach to analyze and classify spam Emails. An efficient Data mining technique is proposed as a solution for spam Email classification. The Next Chapter provides the design of our System presented here.

5 Analysis and Design for the Proposed System

5.1 Introduction

We addressed the technologies used in our approach to solving the problem briefly in the previous Chapter. This Chapter describes the system design of the spam Email classification. Additionally, it represents the methods used in the System.

5.2 Framework for Proposed System

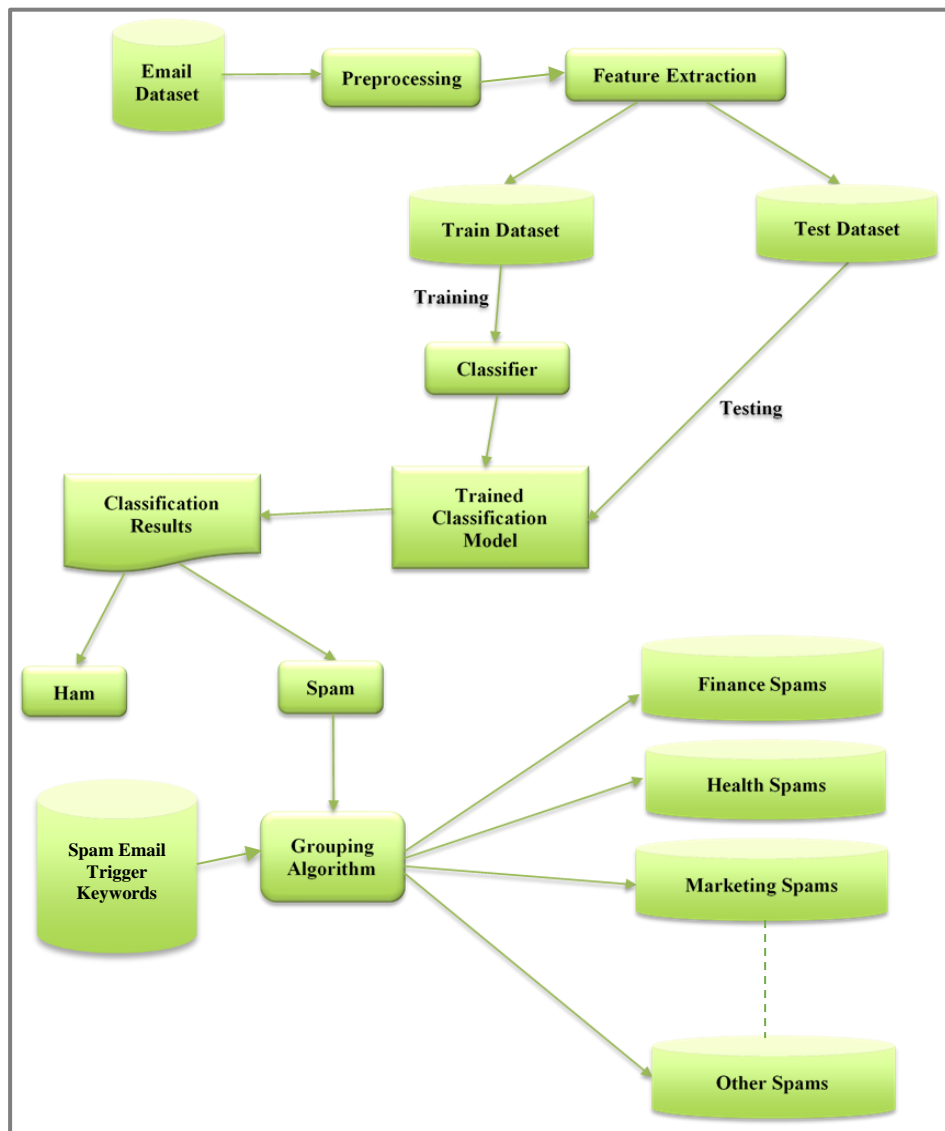


Figure 5.1: System Model for Spam Email classification

5.3 Data collection

Different types of textual Email datasets are used for spam Email identification and Different types of spam Email Trigger keywords are used for spam Email categorization.

There are plenty of famous repositories available to obtain a thousand forms of free data sets. The specific Email dataset was collected from Kaggle Repository.

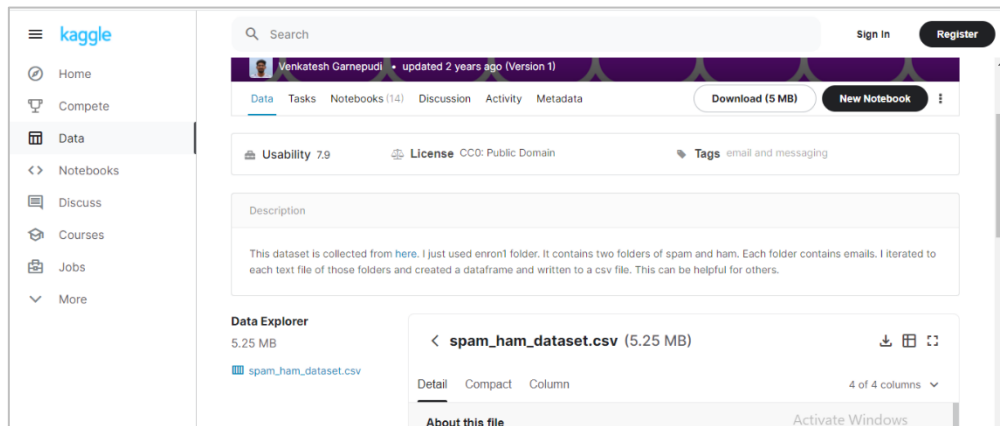


Figure 5.2: Kaggle Repository Website

The spam Email Trigger keywords dataset are collected from the Finance, Marketing and Health Experts through different websites.

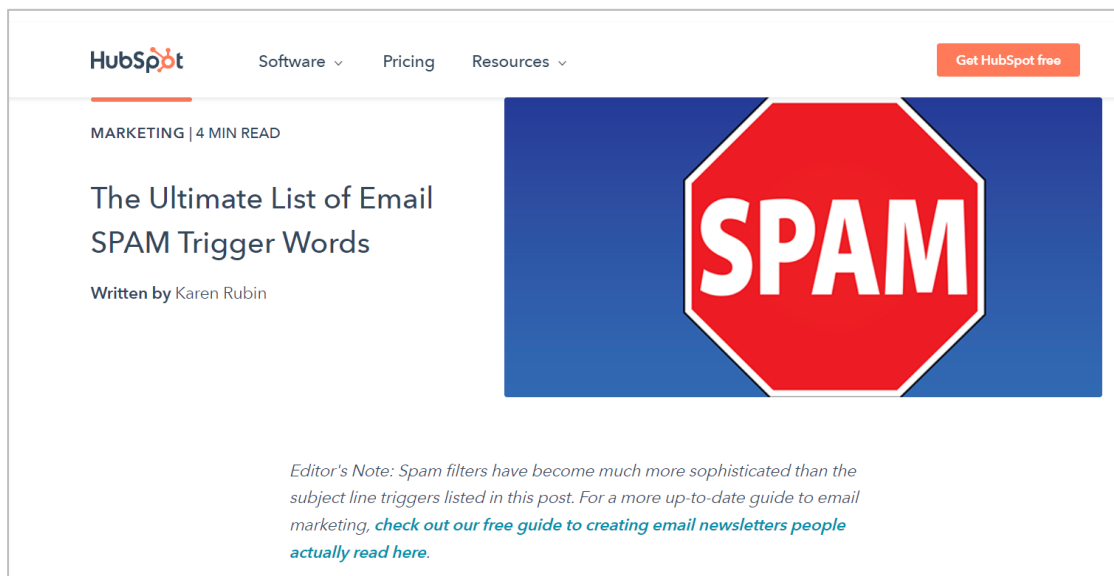


Figure 5.3: HubSpot website

5.4 Data Pre-processing

With several crucial moves, data pre-processing becomes the main activity that needs to be performed in the best way before any data mining is done. The pre-processing of the data requires data cleansing, data aggregation and data extraction. In this research, two imperative steps were executed. Such as Data cleaning and Normalization. Appendix A is explaining the pre-processing methods using some examples. All pre-processing implementation procedures are briefly described in the next Chapter.

5.4.1 Data cleaning

It is the mechanism by which corrupt or incomplete information from a dataset is found and corrected and applies to the detection of inadequate, wrong, incorrect, or meaningless parts of the data and then the substitution, alteration, or elimination of messy or harsh data.

5.4.2 Normalization

It is the method of converting text that it could not have had before into a single canonical form. For further processing, this stage is needed for translating text from natural language to machine-readable format. Normalization of text involves:

- Conversion of all letters to upper or lower case
- Conversion of figures to words or elimination of numbers
- Elimination of punctuations, accent marks, and diacritics
- Eliminating white spaces
- Abbreviations expanding
- Removing words to avoid, sparse phrases, and descriptive words
- Word Stemming
- Word lemmatization

5.5 Feature Extraction

The aim of Feature Extraction is to minimize the range of attributes in a given dataset by creating innovative structures. Models created on extracted features may have higher efficiency because fewer, more meaningful attributes define the data. In this research, a TF-IDF vectorizer is used for feature extraction. In addition to taking each word's word count, the TF-IDF vectorizer can attempt to downscale terms that frequently occur through several Emails. The implementation part of this phase will be shown in the next Chapter.

5.6 Classification

This is the vital portion of this research that utilized intellectual approaches to determine the particular data patterns. In this classification phase, the Naïve Bayes Classifier is used to classify the pre-processed Dataset. This will give a result whether the given Email is Spam or Ham. If the result is spam, then the Grouping algorithm is applied to the output taken from the Naïve Bayes Classifier. This will show us the category of that specific spam Email. The Google spreadsheet is connected with the System to find the category of the spam Email. According to the keywords for each spam Email category stored in the Google spreadsheet, the Grouping algorithm identifies the specific spam Email group.

5.7 Evaluation

The evaluation of this research must be considered to measure the performance of the System to classify the spam Emails. There were several techniques used to measure the functionality of the System. Such as Accuracy, Recall, Precision, Confusion matrix, F1 score. These techniques were briefly described in the Implementation Chapter.

5.8 Summary

This Chapter explained the design of the spam Email classification system and NLP techniques, which are important to do this research. The next Chapter elaborate more details on the implementation of the System based on this analysis and design.

6 Implementation of the Proposed System

6.1 Introduction

The design specifics of each and every process module conducted in this research are given in this Chapter. In addition, this Chapter specifies the program, hardware, and code segments in the design with sample outputs as per each module.

6.2 Python

Python is a versatile programming language and is collaborative, interpreted, and high-level object-oriented. It can function on various OS systems, such as Windows, Mac OS, Linux. It is open and unrestricted [18]. It is also adapted to the .NET virtual machines and Java. Python has an obvious and stylish syntax. It's considerably simple to read and write Python language programs related to other programming languages like C++, Java, C#. Python programming offers a massive collection of libraries in Natural Language Processing.

The Python Standard Library is a set of syntax and semantics of Python. In this research, some Python libraries were used for pre-processing, feature extraction, and classification. All of them are described in this Chapter.

6.3 PyCharm

PyCharm used especially in the Python language is an extraordinarily admired Integrated Development Environment (IDE). It is typically a code editor and debugger for the compilation of programs in many programming languages. It is a cross-platform version of Windows, Mac OS, and Linux. PyCharm offering the following features:

- A graphical debugger
- A combined unit tester
- Support incorporation with version control systems (VCSs)
- Anaconda assistance for data science

Figure 6.1 shows the PyCharm interface window where all the processes mentioned above be present in.

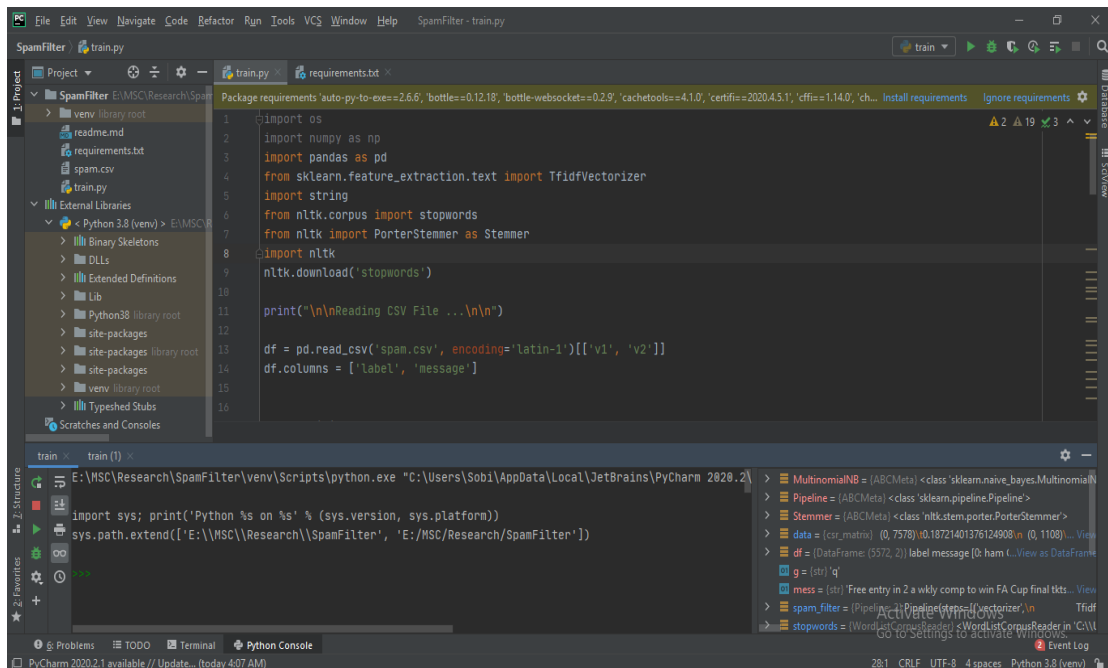


Figure 6.1: PyCharm IDE

6.4 NTLK

NTLK stands for Natural Language Toolkit. It is a top framework of constructing Python programs to use natural language data. In addition to a collection of text processing libraries for clustering, feature extraction, restricting, filtering, sorting, and textual logic, it provides the simplest interfaces and vocabulary properties for promotional NLP libraries.

6.5 Data collection

The Dataset collected from Kaggle Repository contains spam and ham Emails as well as appropriate for use in the screening of spam Email. This Dataset consists of 4850 instances including 2735 Ham Emails and 2115 spam Emails, and it is stored in the CSV format in order to use inside the PyCharm for pre-processing, analyzing, and classification. Figure 6.2 shows the Email dataset used as an input in this research.

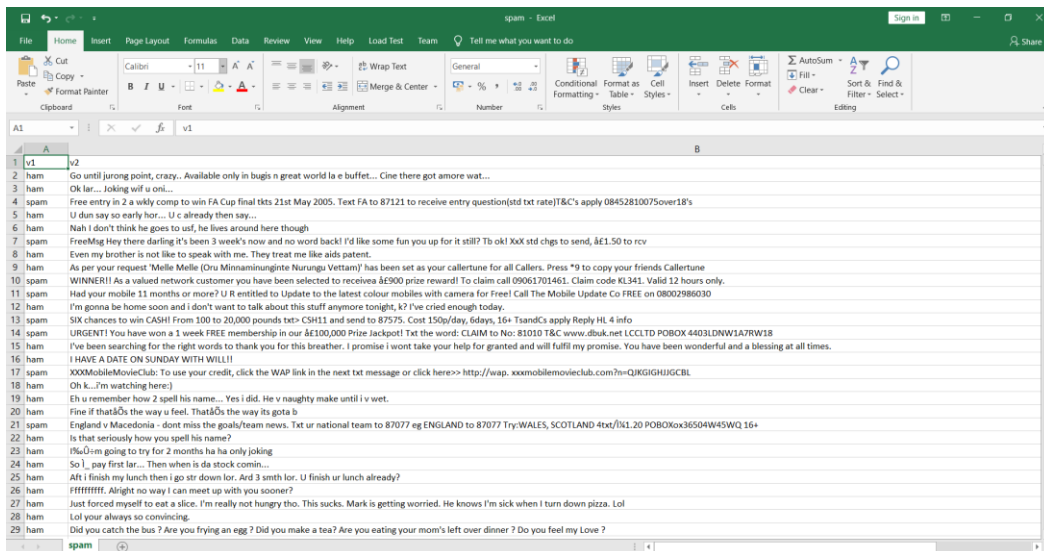


Figure 6.2: Email Dataset

6.6 Pre-processing

Preprocessing refers to the steps that have been taken to boost datamining efficiency.

The implementation in this analysis of Data Preprocessing Methods using Python are,

- 1) Importing the libraries
- 2) Importing the Dataset
- 3) Cleaning the Data
- 4) Normalizing the Data

6.6.1 Importing the libraries

The initial step was done in the implementation part is importing the important libraries into the PyCharm, which are required for this research. Table 6.1 shows the Python Libraries used in the System and their usage purposes.

Python Library	Purpose
pandas	Used to handling the Dataset.
stopwords	Used to removing stop words with NLTK.
PorterStemmer	Used for Text Normalization.
sklearn	Used for classification
TfidfVectorizer	Used to transform a raw text collection to a TF-IDF features matrix.
train_test_split	Used for splitting data arrays into two subsets.
MultinomialNB	Used for classification with discrete features.

Table 6.1: Python Libraries

Appendix B contains the Python code of importing the Python Libraries in this research.

6.6.2 Importing the Dataset

In the second step, the collected Email dataset was uploaded into the System. Figure 6.3 shows the Python code to import the Dataset and Figure 6.4 visualizes the histogram chart of the imported Data.

```
28 print("\n\nReading CSV File ... \n\n")
29
30 df = pd.read_csv('spam.csv', encoding='latin-1')[['v1', 'v2']]
31 df.columns = ['label', 'message']
32 print(df.head())
33 print(df.groupby('label').describe())
34 print(sns.countplot(data=df, x='label'))
35 plt.show()
```

Figure 6.3: Importing Dataset

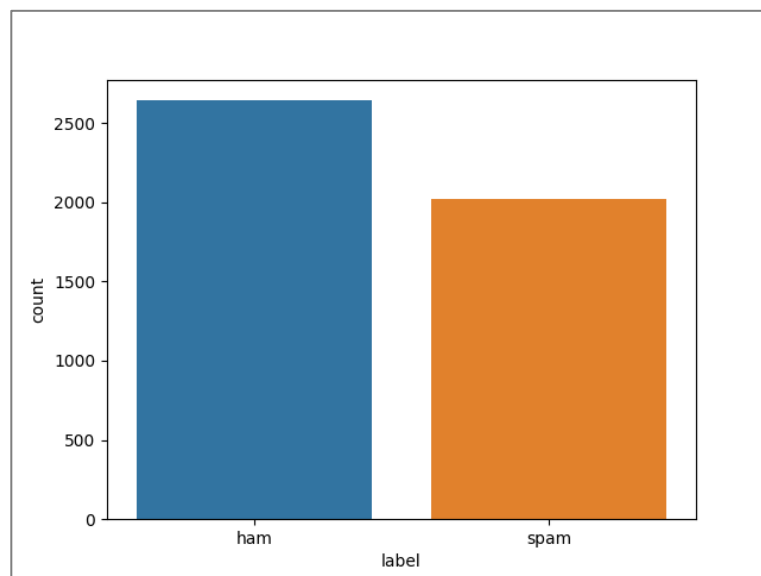


Figure 6.4: Visualization of Dataset

The following Figure 6.5 and Figure 6.6 shows the word clouds of the Ham and Spam Emails which are imported into the system.

ham messages

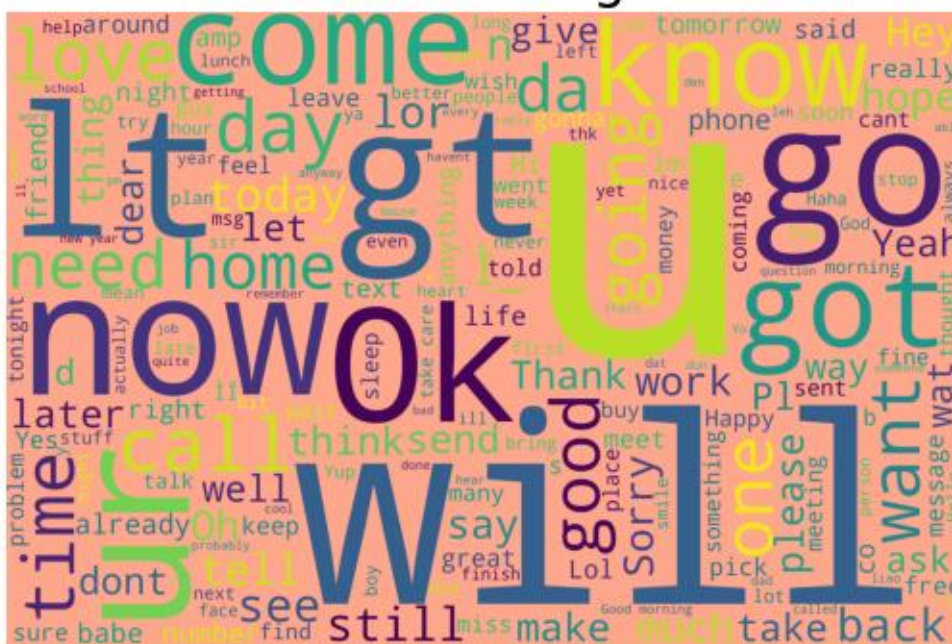


Figure 6.5: Ham Email Word Cloud

Spam messages



Figure 6.6: Spam Email Word Cloud

6.6.3 Cleaning the Data

In machine learning, data cleaning is an essential step, as data can contain plenty of noise and disruption like punctuation, white space, numbers, hyperlinks, and so on. To do this process, Pandas Python library is used. Once the Data set is loaded into a pandas

data frame, it performs some basic cleaning tasks that remove information making data processing slower.

6.6.4 Normalizing the Data

Normalization of data involves some common approaches for minimizing the term to the simplest form. The Normalization procedures included in this research translate all letters to lower, omitting stop words, eliminating punctuations, and Word Stemming.

```
25
26 # Normalizing
27 def process(text):
28     # lowercase it
29     text = text.lower()
30     # remove punctuation
31     text = ''.join([t for t in text if t not in string.punctuation])
32     # remove stopwords
33     text = [t for t in text.split() if t not in stopwords.words('english')]
34     # stemming
35     st = Stemmer()
36     text = [st.stem(t) for t in text]
37     # return token list
38     return text
39
```

Figure 6.7: Normalization

Appendix B displays the Dataset before and after the Preprocessing.

6.7 Feature Extraction

Feature extraction is a reduction of an attribute that takes predictive importance into account the current attributes. In truth, the attributes are transformed. The attributes converted are linear combinations of the input attributes. In order to use the classifier, we must vectorize the Emails Dataset.

TfidfVectorizer from sklearn was used as a feature extraction tool in this research. Figure 6.8 shows the Python code for feature extraction.

```
49 tfidf = TfidfVectorizer(analyzer=process)
50 data = tfidf.fit_transform(df['message'])
```

Figure 6.8: Feature extraction

The following Figure 6.9 shows the data stored into an array after feature extraction.

```
Extracting Features ...

   0     1     2     3     4     ... 18568 18569 18570 18571 18572
0  0.0   0.0   0.0   0.0   0.0   ...  0.0   0.0   0.0   0.0   0.0
1  0.0   0.0   0.0   0.0   0.0   ...  0.0   0.0   0.0   0.0   0.0
2  0.0   0.0   0.0   0.0   0.0   ...  0.0   0.0   0.0   0.0   0.0
3  0.0   0.0   0.0   0.0   0.0   ...  0.0   0.0   0.0   0.0   0.0
4  0.0   0.0   0.0   0.0   0.0   ...  0.0   0.0   0.0   0.0   0.0

[5 rows x 18573 columns]
```

Figure 6.9: Data after Feature Extraction

6.8 Split the Dataset into training & testing sets

For the training of the model, a large proportion (70%) of the Email dataset was used, whereas for testing, a smaller portion (30%) of the data. Python code for this part is shown in Appendix B.

6.9 Create and Train the Multinomial Naïve Bayes classifier

The Multinomial Naïve Bayes Classifier is appropriate to classify distinct attributes according to the past analysis. This algorithm will classify each Email by looking at all of its words individually in this research.

```
55 spam_filter = Pipeline([
56     ('vectorizer', TfidfVectorizer(analyzer=process)), # messages to weighted TFIDF score
57     ('classifier', MultinomialNB()) # train on TFIDF vectors with Naive Bayes
58 ])
```

Figure 6.10: Train Multinomial Naive Bayes classifier

6.10 Test the Data

After the classification model was created, the testing dataset was tested with that classification model.

```
82 # Test Data
83 def detect_spam(s):
84     return spam_filter.predict([s])[0]
```

Figure 6.11: Test the data

6.11 Predict the spam Email

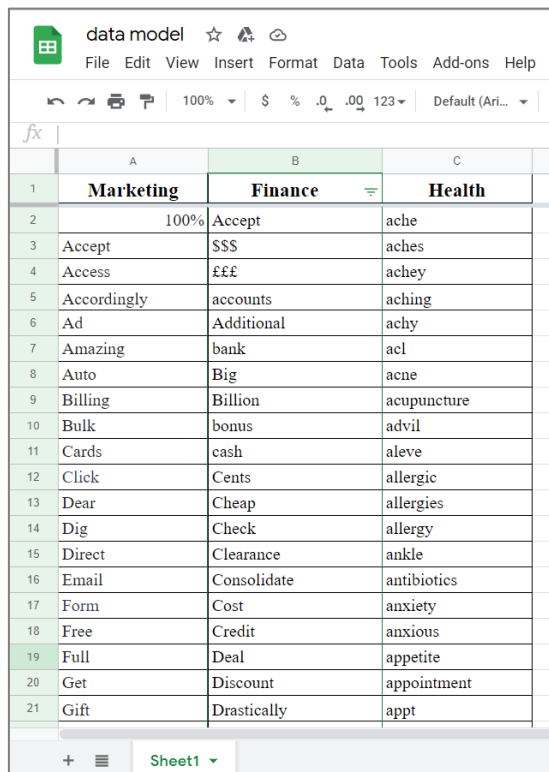
The newly entered Email into the System is predicted whether it is spam or non-spam.

```
87 test = True
88
89 while test:
90     g = input("Enter your Email to check the spam or enter 'q' to quit: ")
91     if g == 'q':
92         test = False
93     else:
94         result = detect_spam(g)
95         print(result)
96         if "spam" == result:
97             url = "https://docs.google.com/spreadsheets/d/1np1HAcxHKhsT71cRfC1ZihqxtKT40QZd7A3k0wbVS0M/edit#gid=0"
98             classifier(process(g), url)
```

Figure 6.12: Predict the spam Email

6.12 Categorize the spam Emails

Finally, the Grouping algorithm classified the result, which was taken from the classification model to identify the specific Spam category. For the spam category classification, the Google spreadsheet is linked to the System. The spam Email Trigger keywords are stored in this Google spreadsheet which has several columns, and each column contains an infinite number of keywords related to different types of spam category. We can add more categories to classify the spam Emails by adding a new column in the Google spreadsheet.



	A	B	C
1	Marketing	Finance	Health
2	100%	Accept	ache
3	Accept	\$\$\$	aches
4	Access	£££	achey
5	Accordingly	accounts	aching
6	Ad	Additional	achy
7	Amazing	bank	acl
8	Auto	Big	acne
9	Billing	Billion	acupuncture
10	Bulk	bonus	advil
11	Cards	cash	aleve
12	Click	Cents	allergic
13	Dear	Cheap	allergies
14	Dig	Check	allergy
15	Direct	Clearance	ankle
16	Email	Consolidate	antibiotics
17	Form	Cost	anxiety
18	Free	Credit	anxious
19	Full	Deal	appetite
20	Get	Discount	appointment
21	Gift	Drastically	appt

Figure 6.13: Google spreadsheet with Different types of spam Email keywords

Appendix C contains the relevant source codes for this spam categorization, and all the outputs taken from the System are available in Appendix D.

6.13 Summary

This Chapter presented all the implementation procedures done in the spam Email classification. Additionally, this Chapter explains the Python libraries and PyCharm IDE to create the classification model and the spam group categorization. The next Chapter delivers the evaluation of the implemented System.

7 Evaluation

7.1 Introduction

This Chapter validates and evaluates the results and the performance of the implemented System. This Chapter also focuses on how testing strategies are carried out according to the objective in terms of evaluation measurements for the selected data mining technique, such as Confusion matrix, Accuracy, Precision, Recall, F1 score for classification.

7.2 Evaluation Techniques used for the classification models

There are various techniques available in machine learning to check the performance of the classification models. In this research, some important techniques were performed to evaluate the implemented System, such as Accuracy, Precision, Recall, Confusion matrix, F1 score.

Four types of outcomes that could occur when performing classification predictions are:

- True positives (TP): positive outcomes that the model predicted correctly.
- False positives (FP): positive outcomes that the model predicted incorrectly.
- True negatives (TN): negative outcomes that the model predicted correctly.
- False negatives (FN): negative outcomes that the model predicted incorrectly.

7.2.1 Accuracy

Accuracy is defined as the percentage of accurate test data predictions. The following equation is used to measure it.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Figure 7.1: Accuracy equation

7.2.2 Precision

Precision is the percentage of positive cases in the overall positive cases expected.

$$\frac{TP}{TP + FP}$$

Figure 7.2: Precision equation

7.2.3 Recall

A recall is defined as the percentage of actual positives that were correctly identified.

$$\frac{TP}{TP + FN}$$

Figure 7.3: Recall equation

7.2.4 Confusion matrix

Confusion Matrix table is summarized to test a classification model's efficiency. The number of right and wrong forecasts is synchronized with count values and grouped by class.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 7.4: Confusion matrix

7.2.5 F1 score

The harmonic mean of precision and recall is established in the F1 score.

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

Figure 7.5: F1 score equation

7.3 Evaluate the System on the train Dataset

Evaluating the System on the training Dataset will give the classification model's performance in this research. The accuracy, Confusion matrix, and Metrics report of this evaluation was brought out as shown below.

```
# Evaluate the model on the training data set
print("Evaluate the model on the training data set...")
y_pred = spam_filter.predict(x_train)

print("Accuracy of the classification model is ", accuracy_score(y_train, y_pred))

skplt.metrics.plot_confusion_matrix(y_train, y_pred, normalize=True)
plt.show()

print(classification_report(y_train, y_pred))
```

Figure 7.6: Python code for evaluation on training Dataset

```
Evaluate the model on the training data set...
Accuracy of the classification model is 0.9751838235294118
      precision    recall  f1-score   support

   ham       0.98     0.98     0.98     1849
   spam       0.98     0.97     0.97     1415

 accuracy                   0.98     3264
 macro avg       0.98     0.97     0.97     3264
 weighted avg    0.98     0.98     0.98     3264
```

Figure 7.7: Output for the evaluation on training Dataset

This shows the classification model used in this research is 97.51% accurate.

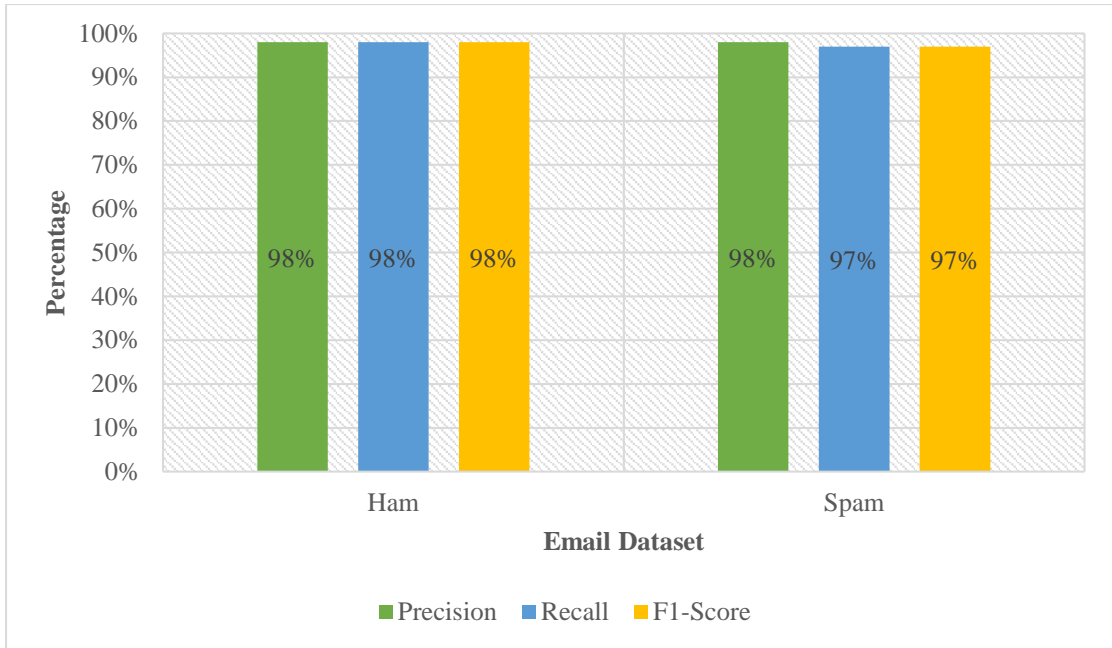


Figure 7.8: Graph of evaluation on training Dataset

The above graph shows the percentage of evaluation metrics versus Ham and Spam Dataset, which is used as the training set.

7.4 Evaluate the System on the test Dataset

Evaluating the System on the testing Dataset will give the classification model's actual performance in this research. The accuracy, Confusion matrix, and Metrics report of this evaluation was brought out as shown below.

```
# Evaluate the model on the testing data set
print("Evaluate the model on the testing data set...")
pred=spam_filter.predict(x_test)

print("Accuracy of the classification model is ", accuracy_score(y_test, pred))

skplt.metrics.plot_confusion_matrix(y_test, pred, normalize=True)
plt.show()

print(classification_report(y_test, pred))
```

Figure 7.9: Python code for evaluation on testing Dataset


```

Evaluate the model on the testing data set...
Accuracy of the classification model is 0.9471428571428572
      precision    recall  f1-score   support

   ham         0.97         0.94         0.95         794
   spam         0.92         0.96         0.94         606

 accuracy                0.95         1400
 macro avg         0.94         0.95         0.95         1400
 weighted avg         0.95         0.95         0.95         1400

```

Figure 7.10: Output for the evaluation on testing Dataset

This shows, the System accurately identified the Emails as spam or ham with 94.71 % accuracy on the test Dataset.

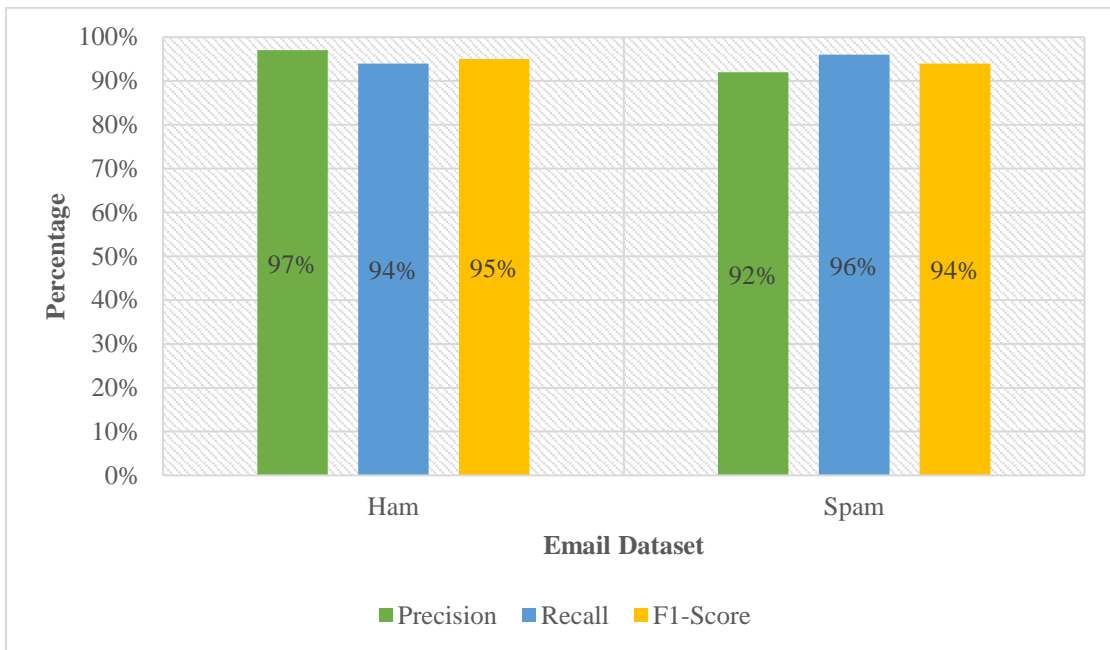


Figure 7.11: Graph of evaluation on testing Dataset

The above graph shows the percentage of evaluation metrics versus Ham and Spam Dataset, which is used as the testing set.

7.5 Experimental evaluation on spam Email classification and Results

The real-life comprehensive Email Dataset was used to assess the spam Email classifier's efficiency based on the Naïve Bayes algorithm. The Dataset was collected from different mail corpora. The total number of Emails was 1500, with no Emails having duplicates in the Dataset. In this experiment, different data sizes of Emails were used to evaluate the spam Email classification system.

Table 7.1 shows the Naïve Bayes classifier's accuracy with respect to different data sizes.

Data size	Spam	Accuracy
500	35%	94.20%
750	40%	94.57%
1000	50%	95.29%
1250	45%	96.89%
1500	50%	97.17%

Table 7.1: Spam Email classifier Accuracy on Data size

7.6 Evaluation on spam Email categorization

The spam Email category classification accuracy depends on the number of keywords in each spam Email category column available on the Google spreadsheet. This accuracy of the spam Email categorization is calculated with the equation shown in Figure 7.12 and Figure 7.13 displays the output.

```
print(count_array[0].items())
sorted_x = sorted(count_array[0].items(), key=operator.itemgetter(1))
accuracy = (sorted_x[-1][1] / len(mail_words)) * 100
print("Prediction Accuracy: ", str(round(accuracy, 1)) + "%")
```

Figure 7.12: Calculate the Accuracy of the spam Email categorization

```
Enter your Email to check the spam or enter 'q' to quit: PRIVATE! Your 2003 Account Statement for 4791529434 shows 800 un-redeemed
spam
dict_items([('Marketing_count', 0), ('Finance_count', 3), ('Health_count', 0)])
This spam E-mail existing in Finance category
Prediction Accuracy: 17.6%
Enter your Email to check the spam or enter 'q' to quit: Your credits have been topped up for http://www.bubbletext.com Your renewal
ham
Enter your Email to check the spam or enter 'q' to quit: Today's Offer! Claim up to $100 worth of discount vouchers! Text YES to 85421
spam
dict_items([('Marketing_count', 0), ('Finance_count', 3), ('Health_count', 0)])
This spam E-mail existing in Finance category
Prediction Accuracy: 13.0%
```

Figure 7.13: Output of the evaluation for spam Email categorization

Table 7.2 shows the summary of the number of keywords identified in the specific category of spam Email and the prediction accuracy in percentage, and Figure 7.14 shows the graph flow of the category prediction accuracy.

No of Keywords	Predicted Percentage
1	4.8%
2	11.1%
3	16.7%
4	23.5%
5	25%
6	26.1%

Table 7.2: Prediction accuracy for different no of keywords

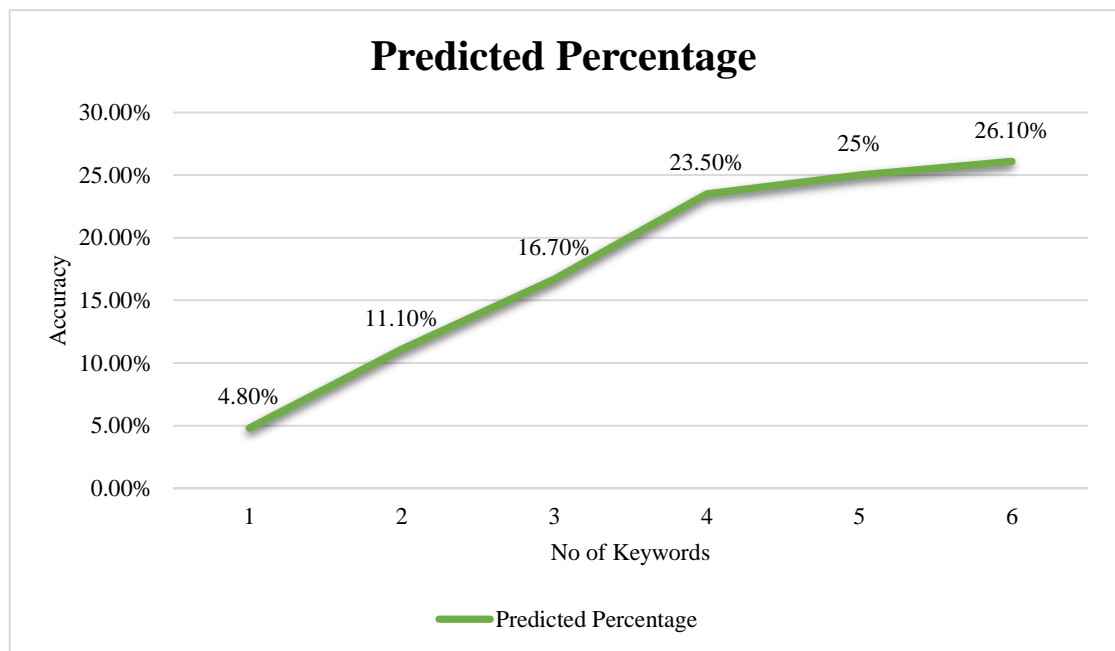


Figure 7.14: Graph for spam Email category prediction Accuracy

According to the above graph flow, the increment of the number of keywords in each spam Email category may increase the System's Accuracy.

7.7 Experimental evaluation on spam Email categorization and Results

The different types of real-life Email Dataset were used to examine the performance of the spam Email categorization. 225 real-time spam Emails collected from the Finance, Marketing and Health industry people and which are used to evaluate the accuracy of the spam Email categorization in the system.

Figure 7.15 shows the table which presents the comparison between the human expert and the system in spam Email categorization.

1	Email	Human	System
2	stock promo mover : cwtdd ***urgent investor trading alert ***weekly stock p	Finance	Finance
3	are you listed in major search engines ? submitting your website in search engi	Finance	Finance
4	important information thu , 30 jun 2005 . subject : important information thu ,	Finance	Finance
5	= ? utf - 8 ? q ? bask your life with ? == ? utf - 8 ? q ? individual incremen ? == ? utf	Finance	Finance
6	" bidstogo " is places to go , things to do hello , privacy policy : to permanently o	Finance	Finance
7	dont pay more than \$ 100 for ur softwares miseris cheap softwares for you	Finance	None
8	paliourg micros 0 ft for pennies check ' em out carney handwrite perpetrate an	Finance	Finance
9	all graphics software available , cheap oem versions . good morning , we we of	Finance	Finance
10	the man of steel hello , welcome to the medzonlin direction e - online pharma	Finance	Finance
11	adjourn pasteup paliourg , looking for not expensive high - quality software ? w	Finance	Finance
12	need your medication ? we have them ideolect hilarious philosophy are you lo	Finance	Finance
13	need your vics ? brand name meds such as vics , vals , xanies and others why n	Finance	Finance
14	urgent security notification ! dear valuedpaypalmember : we recently noticed	Marketing	Marketing
15	re : interest rates are at 40 - year lows ! re - finance now , even with bad - credi	Marketing	Marketing
16	unbelievable investors info mnei - the best small cap stock in 2 oo 5 just keep	Marketing	Marketing
17	new love tabs shop . visit our llcensed online dragstore for the best inexpensiv	Marketing	Marketing
18	save your money buy getting this thing here you have not tried cials yet ? than	Marketing	Marketing
19	any medication you will ever need ! privacy guaranteed . within yourself delive	Marketing	Marketing
20	set & forget ! blast your ad over 200 million leads 2) you posted to one of my ff	Marketing	Marketing
21	paliourg udth 7 wcwknoanopt good morning paliourg ! last miraculous and in	Marketing	Marketing
22	can you afford to ignore smallcaps ? homeland security investments the terror	Marketing	Marketing
23	claim your free \$ 1000 home depot gift card . claim your home depot gift card -	Marketing	Marketing
24	qu otes to share , check better rattes good day to you sir , check below valid for	Marketing	Marketing
25	get xanax , valmum to your door step - no previous prescription necessary plea	Marketing	None
26	save your money by getting an oem software ! need in software for your pc ? jus	Marketing	Marketing
27	urgent security notification ! dear valuedpaypalmember : we recently noticed	Marketing	Marketing
28	feeling fat ? if anyone has called you names because you ' re overweight , or yo	Health	Health
29	b ' uy pain medicine online looking for vlcodln and / or hydrocod 0 ne ? only pla	Health	Health
30	a new era of online medical care . a new era of online medical care . there are	Health	Health

Figure 7.15: Comparison between the human expert and the system in spam Email categorization.

Based on the above table, out of 225 spam Emails, 207 Emails are correctly categorized by the system. This shows, the system categorizing the spam Emails with 92% accuracy.

7.8 Summary

This Chapter concludes with evaluation results to evaluate the System. The final Chapter will summarize the complete research work and bring out the essential resolutions of this research.

8 Conclusion and Further Work

8.1 Introduction

This Chapter given the summary of the research study and how the solutions are provided for classifying and categorizing spam Emails with the suitable classification model. And also, this Chapter focuses on the limitations and future work of this research.

8.2 Overview of the research

Spam is one of the most hateful and annoying Internet supplements. Orthodox applications for spam filtering cannot manage massive quantities of spam. A safer way for researchers to tackle spam is offered by ML practices. Machine learning was used effectively in the text classification. Pre-processing methods play an essential role in spam Email detection and categorization. There were some combinations of pre-processing methods used in this research. The Naïve Bayes classification demonstrates on the basis of this research the successful and improved technologies to detect and classify spam Email.

Instead of a binary classification on Email, the System was implemented in this research to identify the category of the spam Email, which is already classified by the Naïve Bayes classifier. Some evaluation methods have also measured the performance of the system, such as Accuracy, Precision, Recall, Confusion matrix, F1 score.

8.3 Limitations

This research is limited to textual spam Emails. The System's accuracy highly depends on the Email dataset, and spam category classification's accuracy determined by the number of keywords in each category column on the Google spreadsheet.

8.4 Further developments

The work done in this research is to identify the spam Email and then predict the category of that spam Email. An integrated classification and categorization in Email spam filtering are expected as future work. That will save more time and provide more accurate results than the result obtains from this research. And also, the input Dataset will be expanded with the multimedia content, such as image, audio, and video for future research.

8.5 Summary

This Chapter concluded with an analysis, summary of the major discoveries, limitations, and improvements in future work.

References

- [1] R. Jennings, "Spam, Spammers, and Spam Control," Ferris Research, San Francisco, Calif., 2009.
- [2] R. M. A. a. S. A. N. RasimM. Alguliev, "Classification of Textual E-Mail Spam Using," *Applied Computational Intelligence and Soft Computing*, vol. 2011, p. 8, 2011.
- [3] S. Shrivastava and R. Anju, "Spam mail detection through data mining techniques," in *International Conference on Intelligent Communication and Computational Techniques (ICCT)*, Jaipur, 2017.
- [4] Z. K. a. U. Qamar, "Text Mining Approach to Detect Spam in Emails," in *Proceedings of The International Conference on Innovations in Intelligent Systems and Computing Technologies*, Philippines, 2016.
- [5] Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa, "Overview of textual anti-spam filtering techniques," *International Journal of the Physical Sciences*, vol. 5, pp. 1869-1882, 2010.
- [6] Yukti Kesharwani and Shrikant Lade, "Spam Mail Filtering Through Data Mining Approach –A Comparative Performance," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 9, p. 4, 2013.
- [7] Biro. I, J. Szabo and A. A. Benczur. Latent Diric, "location in Web Spam Filtering," in *4th International Workshop on Adversarial Information Retrieval on the Web(AIRWeb)*, 2008.
- [8] A. Perkins, "The classification of search engine spam.," [Online]. Available: <http://www.ebrandmanagement.Com/whitepapers/spamclassification>.
- [9] M. N. Marsono, M. W. El-Kharashi and F. Gebali, "Binary LNS-based naïve Bayes inference," in *IET Computers & Digital*, 2008.
- [10] C. Paulo, L. Clotilde and S. Pedro, "Symbiotic data mining for personalized spam filtering," in *Web Intelligence and Intelligent Agent Technology*, 2009.
- [11] S. Nazirova, "Mechanism of classification of text spam messages collected in spam pattern bases," in *3rd International Conference on Problems of Cybernetics and Informatics, (PCI '10)*, 2010.

- [12] W.A. Awad¹ and S.M. ELseuofi, "MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, 2011.
- [13] I. Androutsopoulos, G.Paliouras, V. Karkaletsis and G. Sakkis, "Learning to filter spam e-mail: A comparison of a Naïve Bayesian and a memory-based approach," in *Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Europe, 2000.
- [14] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine learning tools and Techniques*, 2 ed., San Fransisco: Morgan Kaufmann, 2005.
- [15] Francisco Janez-Martino, Eduardo Fidalgo, Santiago Gonzalez-Martinez and Javier Velasco-Mata, "Classification of Spam Emails through Hierarchical Clustering and Supervised," *ArXiv*, vol. abs/2005.08773, 2020.
- [16] W.A. Awad and S.M. ELseuofi, "MACHINE LEARNING METHODS FOR SPAM E-MAIL," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, 2007.
- [17] "Analyticsvidhya," [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>.
- [18] "Programiz," [Online]. Available: <https://www.programiz.com/python-programming/first-program>.

Appendix A

Raw	Lowercased
Canada Canada CANADA	canada
TOMCAT Tomcat toMcat	tomcat

Converting Upper case letters Lower case letters

```
IN:
['He', 'did', 'not', 'try', 'to', 'navigate', 'after', 'the',
'first', 'bold', 'flight', ',', 'for', 'the', 'reaction', 'had',
'taken', 'something', 'out', 'of', 'his', 'soul', '.']

OUT:
['try', 'navigate', 'first', 'bold', 'flight', ',', 'reaction',
'taken', 'something', 'soul', '.']
```

Removing stop words

```
IN:
["It never once occurred to me that the fumbling might be a mere
mistake."]

OUT:
['it', 'never', 'onc', 'occur', 'to', 'me', 'that', 'the',
'fumbl', 'might', 'be', 'a', 'mere', 'mistake.'],
```

Stemming Text

```
The stemmed form of leafs is: leaf
The stemmed form of leaves is: leav

The lemmatized form of leafs is: leaf
The lemmatized form of leaves is: leaf
```

Text Lemmatization

Appendix B

```
1 import ssl
2 import pandas as pd
3 from sklearn.feature_extraction.text import TfidfVectorizer
4 import string
5 from nltk.corpus import stopwords
6 from nltk import PorterStemmer as Stemmer
7 from sklearn.pipeline import Pipeline
8 from sklearn.naive_bayes import MultinomialNB
9 from sklearn.model_selection import train_test_split
10 from classifier import classifier
11 import matplotlib.pyplot as plt
12 import seaborn as sns
13 from sklearn.metrics import accuracy_score
14 import scikitplot as skplt
15 from sklearn.metrics import classification_report
16 import wordcloud
```

```
1 from __future__ import print_function
2 import operator
3 import gspread
4 from google.oauth2.service_account import Credentials
```

Imported Python Libraries

```

label      message
0   ham    Go until jurong point, crazy.. Available only ...
1   ham                    Ok lar... Joking wif u oni...
2  spam    Free entry in 2 a wkly comp to win FA Cup fina...
3   ham    U dun say so early hor... U c already then say...
4   ham    Nah I don't think he goes to usf, he lives aro...

```

Dataset before Preprocessing

```

0   [go, jurong, point, crazi, avail, bugi, n, gre...
1                    [ok, lar, joke, wif, u, oni]
2   [free, entri, 2, wkli, comp, win, fa, cup, fin...
3   [u, dun, say, earli, hor, u, c, already, say]
4   [nah, dont, think, goe, usf, live, around, tho...
5   [freemsg, hey, darl, 3, week, word, back, id, ...
6   [even, brother, like, speak, treat, like, aid,...
7   [per, request, mell, mell, oru, minnamingungint...
8   [winner, valu, network, custom, select, receiv...
9   [mobil, 11, month, u, r, entitl, updat, latest...

```

Dataset after Preprocessing

```

x_train, x_test, y_train, y_test = train_test_split(df['message'], df['label'], test_size=0.30, random_state=21)

```

Splitting Dataset into training and testing set

Appendix C

```
train.py x classifier.py x
1 from __future__ import print_function
2 import operator
3 import gspread
4 from google.oauth2.service_account import Credentials
5
6
7 def create_assertion_session_service_account(url):
8     scopes = [
9         'https://www.googleapis.com/auth/spreadsheets',
10        'https://www.googleapis.com/auth/drive'
11    ]
12
13    credentials = Credentials.from_service_account_file(
14        'credentials.auth.json',
15        scopes=scopes
16    )
17
18    gc = gspread.authorize(credentials)
19
20    sh = gc.open_by_url(url)
21
22    worksheet = sh.sheet1
23    return worksheet
24
```

Connect Google spreadsheet into the System

```
train.py x classifier.py x
37
38 def classifier(mail_words, url):
39     search_keys = dataIndexIdentification(url)
40     worksheet = create_assertion_session_service_account(url)
41     count_array = []
42     count_dict = {}
43     search_keys_array = []
44     search_keys_dict = {}
45     for key, value in search_keys.items():
46         search_keys_dict[key] = worksheet.col_values(value['index'] + 1)[1:]
47         count_dict[key + "_count"] = 0
48     count_array.append(count_dict)
49     search_keys_array.append(search_keys_dict)
50     mail_words = mail_words
51
52     for mail_word in mail_words:
53         for search_key in search_keys_array:
54             for key, value in search_key.items():
55                 if mail_word in value:
56                     for k, v in count_array[0].items():
57                         if k == key + "_count":
58                             count_array[0][key + "_count"] = count_array[0][key + "_count"] + 1
59
60     sorted_x = sorted(count_array[0].items(), key=operator.itemgetter(1))
61     if sorted_x[-1][1] == 0:
62         print("This spam E-mail is not in existing Spam E-mail Category!!!")
63     else:
64         print("This Spam E-mail existing in " + sorted_x[-1][0][:-6] + " Category.")
```

Grouping algorithm for spam Email categorization

Appendix D

```
1 C:\Users\S0BI\AppData\Local\Programs\Python\Python38\
  python.exe "D:/Research/SpamFilter 2/SpamFilter/train
  .py"
2
3
4 Reading CSV File ...
5
6
7  label
  message
8 0 ham Free entry to the gr8prizes wkly comp 4 a
  chan...
9 1 ham congratulations !!! vince :
  congratulatio...
10 2 ham jacob feedback vince , chonawee and tom
  hal...
11 3 ham darden case study on " the transformation
  of ...
12 4 ham meeting tracy , confirming the meeting
  tomo...
13
14
15 Imported data: (4850, 2)
16     message

17         count unique

                                     top
18     freq
19 label

19 ham      2735   2643
  Sorry, I'll call later      6
20 spam     2115   2021 Please call our customer
  service representativ...    4
21 Available data after removing duplicates: (4664, 2)
22
23
24     message

25         count unique
```

```

25 freq
26 label

27 ham      2643   2643   Been running but only managed 5
      minutes and th...   1
28 spam     2021   2021   nymex invitation - learn power
      trading power...   1
29
30 Normalizing Data ...
31
32
33 After Preprocessing Data ...
34 0 [free, entri, gr8prize, wkli, comp, 4, chanc
, ...
35 1 [congratul, vinc, congratul, promot, barbara
]
36 2 [jacob, feedback, vinc, chonawe, tom, hallibur
...
37 3 [darden, case, studi, transform, enron, shirle
...
38 4 [meet, traci, confirm, meet, tomorrow, thursda
...
39 5 [yar, els, ill, thk, sort, funni, thing
]
40 6 [place, man
]
41 7 [that, cool, day
]
42 8 [r, go, ltgt, bu
]
43 9 [hello, love, went, day, alright, think, sweet
...
44 Name: message, dtype: object
45
46
47 Extracting Features ...
48
49
50 0 1 2 3 4 ... 18568
18569 18570 18571 18572
51 0 0.0 0.0 0.0 0.0 0.0 ... 0.0 0
.0 0.0 0.0 0.0
52 1 0.0 0.0 0.0 0.0 0.0 ... 0.0 0

```

```

52 .0 0.0 0.0 0.0
53 2 0.0 0.0 0.0 0.0 0.0 ... 0.0 0
   .0 0.0 0.0 0.0
54 3 0.0 0.0 0.0 0.0 0.0 ... 0.0 0
   .0 0.0 0.0 0.0
55 4 0.0 0.0 0.0 0.0 0.0 ... 0.0 0
   .0 0.0 0.0 0.0
56
57 [5 rows x 18573 columns]
58
59
60 Training the Model with Naive Bayes Algorithm ...
61
62
63 Train set: (3264,)
64 Test set: (1400,)
65
66
67 Evaluate the model on the training data set...
68 Accuracy of the classification model is 0.
   9751838235294118
69           precision    recall  f1-score   support
70
71      ham       0.98       0.98       0.98       1849
72      spam       0.98       0.97       0.97       1415
73
74      accuracy                   0.98       3264
75      macro avg       0.98       0.97       0.97       3264
76      weighted avg    0.98       0.98       0.98       3264
77
78 Evaluate the model on the testing data set...
79 Accuracy of the classification model is 0.
   9471428571428572
80           precision    recall  f1-score   support
81
82      ham       0.97       0.94       0.95       794
83      spam       0.92       0.96       0.94       606
84
85      accuracy                   0.95       1400
86      macro avg       0.94       0.95       0.95       1400
87      weighted avg    0.95       0.95       0.95       1400
88
89 Enter your Email to check the spam or enter 'q' to
   quit: Free entry to the gr8prizes wkly comp 4 a

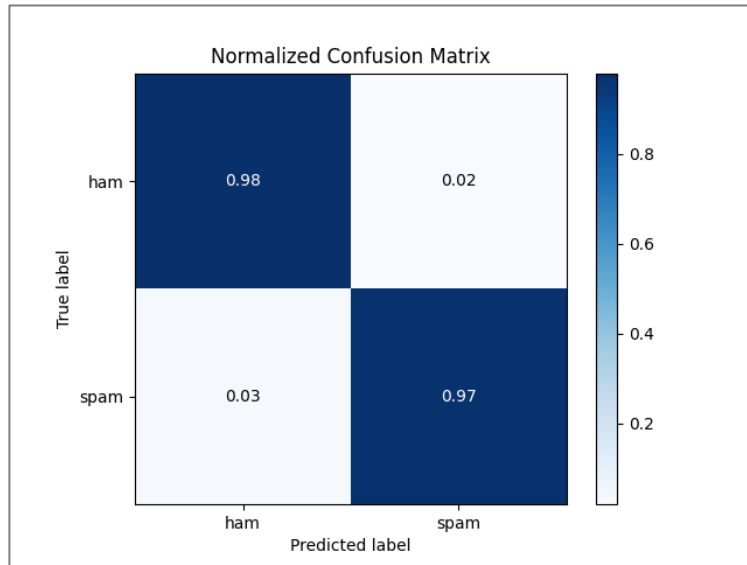
```

```

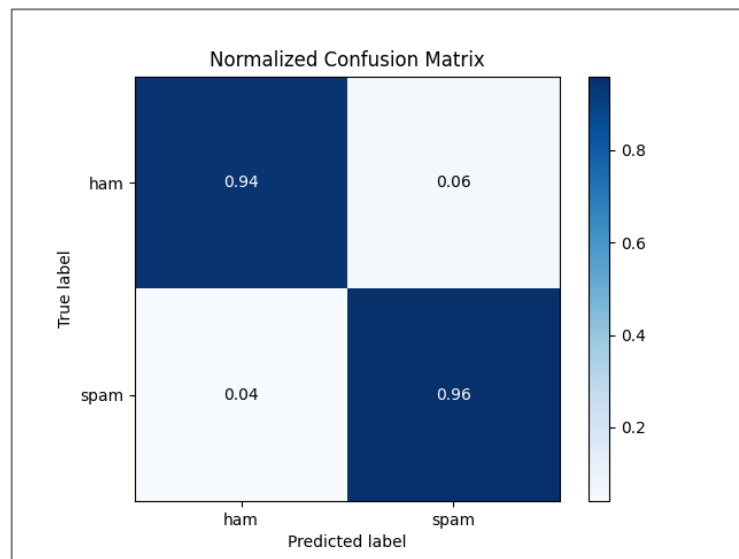
89 chance to win the latest Nokia 8800, PSP or $E250
   cash every wk.TXT GREAT to 88878 http://www.gr8prizes
   .com 88715705022
90 spam
91 dict_items([('Marketing_count', 1), ('Finance_count
   ', 2), ('Health_count', 0)])
92 This Spam Email existing in Finance Category.
93 Prediction Accuracy: 10.0%
94 Enter your Email to check the spam or enter 'q' to
   quit: congratulations !!! vince :
   congratulations on your promotion !! barbara
95 spam
96 dict_items([('Marketing_count', 0), ('Finance_count
   ', 0), ('Health_count', 0)])
97 This Spam Email is not in existing Spam E-mail
   Category!!!
98 Enter your Email to check the spam or enter 'q' to
   quit: Good FRIENDS CaRE for each Other.. CLoSE
   Friends UNDERSTaND each Other... and TRUE Friends
   STaY forever beyond words, beyond time. Gud ni8
99 ham
100 Enter your Email to check the spam or enter 'q' to
   quit: avoid fake viagra get the real thing take
   energy pills for sexual health she ' s the only man
   in my cabinet . act as if were impossible to fail
   . a clash of doctrines is not a disaster - - it is
   an opportunity .
101 spam
102 dict_items([('Marketing_count', 0), ('Finance_count
   ', 0), ('Health_count', 1)])
103 This Spam Email existing in Health Category.
104 Prediction Accuracy: 5.0%
105 Enter your Email to check the spam or enter 'q' to
   quit: URGENT! Your mobile No ***** WON a $E2,
   000 Bonus Caller Prize on 02/06/03! This is the 2nd
   attempt to reach YOU! Call 09066362220 ASAP!
   BOX97N7QP, 150ppm
106 spam
107 dict_items([('Marketing_count', 1), ('Finance_count
   ', 2), ('Health_count', 0)])
108 This Spam Email existing in Finance Category.
109 Prediction Accuracy: 13.3%
110 Enter your Email to check the spam or enter 'q' to
   quit: q

```

Output of the System



Confusion matrix for training Dataset



Confusion matrix for testing Dataset