

SPAM TEXT DETECTION BY USING ML ALGORITHM

BY

SUMIYA
ID: 191-15-2765

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S.M. Aminul Haque

professor

Department of CSE

Daffodil International University

Co-Supervised By

Amit Chakraborty Chhoton

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project/internship titled “Spam Text Detection By Using ML Algorithms”, submitted by Sumiya, ID No: 191-15-2765 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26 January,2024.

BOARD OF EXAMINERS

MIJ

Chairman

Dr. Md. Ismail Jabiullah(MIJ)

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

on behalf of,

MIJ

Internal Examiner

Saiful Islam (SI)

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Taslima Ferdous Shuva(TFS)

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

SMH

Dr. S. M. Hasan Mahmud (SMH)

Assistant Professor

Department of Computer Science
American International University-Bangladesh

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Professor and Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Dr. S.M. Aminul Haque

Professor

Department of CSE

Daffodil International University

Co-Supervised by:

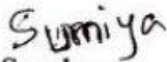
Amit Chakraborty Chhoton

Assistant professor

Department of CSE

Daffodil International University

Submitted by:



Sumiya

ID: 191-15-2765

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to Supervisor **Dr. S. M. Aminul Haque** ,**Professor & Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “**Machine learning**” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Nowadays, the number of mobile users is always increasing. SMS stands for "short messaging service," which lets the users send and receive short text messages on regular phones as well as smartphones.

The quantity of SMS texts increased significantly as a result. Additionally, Spam, or unsolicited messages, became more prevalent. Spammers send unsolicited emails with the intention of gaining business or money through things like buying lottery tickets, breaking into new markets, or disclosing credit card details. This directly leads to more attention being paid to sorting via spam. There exist numerous content based ML (machine learning) strategies that have been shown to be successful in removing spam from emails. Researchers of today have classified text messages as spam or ham by using certain stylistic elements. The actual existence of all well-known terms, phrases, acronyms, and idioms can have a great significant impact on SMS spam detection. This study compares various classification methods using various datasets gathered from earlier research projects. This paper proposed a powerful solution based on machine learning classification techniques. This paper developed and tested and evaluated this strategy utilizing five learning algorithms: Naive Bayes with 97% accuracy, k-Nearest Neighbors Algorithm with 89% accuracy, Decision tree learning algorithm with 96% accuracy, SVM algorithm with 94% accuracy, Random Forest algorithm with 95% accuracy level. The experimental data demonstrated that all of the proposed methods provide very high levels of accuracy for recognizing these data sets but Naive Bayes tops them all with 97% accuracy.

key words: SMS, spam and ham, machine learning, Naive Bayes, K-Nearest Neighbor, SVM, Random forest. AUC

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
 CHAPTER	
 CHAPTER 1: INTRODUCTION	 1-6
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	3
1.6 Project Management and Finance	3
1.7 Report Layout	4
 CHAPTER 2: BACKGROUND	
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	9
2.4 Scope of the Problem	11

2.5 Challenges	11
3.2 Data Collection Procedure/Dataset Utilized	12
3.3 Statistical Analysis	13
3.4 Proposed Methodology/Applied	13
3.5 Implementation Requirements	14

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup	16
4.2 Experimental Results & Analysis	17
4.3 Discussion	20

CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society	21
5.2 Impact on Environment	21
5.3 Ethical Aspects	21
5.4 Sustainability Plan	22

CHAPTER 6:SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1Summary of the Study	23
6.2Conclusions	23
6.3Implication for Further Study	23

REFERENCES

24-25

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Data plot pie chart	13
Figure 3.2: Proposed methodology of spam text detection algorithm	10
Figure 4.1: comparison of ML algorithm	17
Figure 4.2: the confusion matrix of naive bayes	18
Figure 4.3: the confusion matrix K-Nearest Neighbors algorithm	19
Figure 4.4:the confusion matrix of decision tree	19
Figure 4.5:the confusion matrix of Support Vector Machine (SVM)	20
Figure 4.6:the confusion matrix of random forest	20
Figure 4.7:presentation of AUC	21

LIST OF TABLES

FIGURES	PAGE NO
Figure 1.1: finances of project	4
Figure 2.1: Comparison analysis of different research	10
Figure 4.8: result table of all the algorithm	21

CHAPTER 1

Introduction

1.1 Introduction

The usage of text messages and email for communication has grown in recent years due to the sharp rise in the number of mobile phones. Among the most efficient and quick forms of communication is the Short Messaging Service (SMS). Worldwide, SMSs are sent and used for both professional and personal reasons. However, in addition to significant .We also get other irrelevant and fraudulent SMSs, which is really annoying for the subscribers. SMS spam is becoming a concern since a lot of fake messages are being sent for both business and personal purposes. Text analysis considers spam detection to be a major problem, and accurately recognizing spam SMS is a challenging and significant task. Advances in mobile device technology have made users more dependent on them and enabled them to retain private and sensitive data records, including emails, contact lists, bank account information, medical records, and other sensitive data records. This is because of the devices' mobility and broad use in daily life. Users can now store private information rapidly. Furthermore, the rapid advancement of mobile networks has enhanced daily operations by enabling real-time information distribution and communication. Sending brief text messages between devices is also known as Short Messaging Service or as SMS. Spam is a term that is typically used to describe messages or information. that is spam or not requested. Therefore, any garbage short message sent by text messaging to a mobile phone device can be classified as a spam SMS. Nowadays the datasets to train as well as test methods for Short Message Service spam detection are now currently few and modest in size despite the fact is there that different type of datasets are available for test email spam detection by ML algorithms. Furthermore, text messages are shorter than emails, which means that they include less statistically-differentiable information. As a result, fewer features are needed to identify spam SMS. Because informal type languages as regional terms, idioms,

phrases, and acronym greatly impact messages as well as email scam filtering techniques do not work well for text messages.

1.2 Motivation

Machines have the power to handle their work and execute in the right way. For time saving and not to take a step in wrong ways we use machine learning. In this case of study, spam detection is highly used on the machine mechanism. In further study, a detection system can be very useful for what you need to do or not. Our confidential and valuable data is stored in devices and it could be a great loss for us if we could not protect this data. So we need to detect spam SMS. Spam SMS detection is similar to it. It can determine which of those are spam or ham. And as we live in a modern world now which controls AI, we need to apply better ML for a smooth lifelead.

1.3 Rationale of the Study

Nowadays the way of communication totally depends on modern technology. SMS is the great source of information passing and the easiest way to communicate. Through SMS confidential and valuable information pass to each other. If in a confidential case the spam text passes and the person or the company's take steps based on this text it will be a great cause of loose communication and would be a problem to take action. So we need to be secure about ham SMS passing and so we need to build a system that could detect spam text and can secure our way of communication and our data will be protected from danger.

1.4 Research Question

The most reliable question to purpose this research is:

- How to detect spam text and which kind of characteristic they carry the most?
- Where to find the dataset?
- Which kind of dataset should we use?
- How many algorithms need to perform for the best accuracy?

- Which framework and methodology need to be used.
- Evaluate and apply where the characteristic is needed.
- How to use this system for real life problems and to deal with financial cost?

1.5 Expected output

SMS is the firstest communication method. Globally SMS are used for business and personal purposes to communicate. But we face a problem if we receive a fraudulent text . So we need to propose a method to detect spam SMS and need to execute in a proper way. So from our proposed method we got some expected outcome. Those will be:

- Recognize the spam and ham SMS.
- Proposed system needs to be run smoothly.
- Need the highest spam text detection accuracy from this system.
- Make an awareness about fraudulent text and prevent data loss for spam SMS.
- Preventing the harm caused to humans by not recognizing spam SMS.

1.6 Project Management and Finance

This research is supposed to be done in a certain amount of time. And for better results from this project I need to set a bunch of goals . After finishing this project it needs to be maintained in the right way to achieve those goals. And of course those goals should be realistic and divided into more manageable, achievable spaced out over the project. This project needs to be clean design, modern and upgrade technology based and minimize distraction.

Finance: For establishing this project we need some tools. And those tools market value is affordable. So those project bear cost will be:

Table 1.1: finances of project

Used tools	Estimate cost
<ul style="list-style-type: none"> ● Physical device/ Hardware 	<ul style="list-style-type: none"> ● 60,000-70,000
<ul style="list-style-type: none"> ● Motherboard/ main circuit 	<ul style="list-style-type: none"> ● 12,000-17,000
<ul style="list-style-type: none"> ● Runnable tools 	<ul style="list-style-type: none"> ● 00-00

1.7 Report Layout

- This background study is in chapter two which includes the Preliminaries, Related Works, Compare and Analysis, Scope of the Problems and Challenge
- Research methodology has research and subject data collection, statistical analysis, proposed methodology and required instrumentation
- Experimental Result and Discussion Provides a proper explanation about Experimental Setup and Results & Analysis, Discussion
- It contains the impact on the society, the environment, the ethical aspect and the sustainability plan
- Summary, Conclusion and Future Analysis discussed in chapter six .
- Reference in the end

CHAPTER 2

Background Study

2.1 Preliminaries

Nowadays the internet is gradually gaining influence over physical objects. Starting with banking, educational systems, money earning, jobs, shopping, travel agencies, and salary payments, we cannot imagine our lives without the internet. The Internet has taken away all of our access. The usual spam SMS is sent by con artists posing as trustworthy firms, most often banks or credit card companies. These SMS are designed to trick you into disclosing your login credentials or financial credentials like peoples credit card or social security type numbers.

2.2 Related Works

Delany and colleagues [9] assembled their personal collected dataset of spam text from websites such as who can call to reach me and GrumbleText. They implemented content based clustering to evaluate most of the spam text and those were almost able to properly identify nine or ten of distinct groupings.

The method that Kim et.al. [4] employed was totally based on frequency calculation, which gauges the total speed and lightness of all filtering techniques. They took into account the J-48, logistic, and naive Bayes algorithms. Their suggested technique had capabilities comparable to those of others.

An Artificial Immune System (AIS) was created by Mahmoud and Mahfouz [5] for SMS classification. Trained dataset that contained phone numbers, spam phrases, and other information, it was then utilized to classify the sms . When categorized as ham or spam , the findings of this experiment outperformed the Naive Bayesian method in terms of accuracy and convergence speed and also obtained results that were more accurate than those obtained using other methods.

Yang and Elfayoumy research [6] assesses the efficacy of Bayesian classifiers based on feed-forward backpropagation neural networks in the classification of spam emails. Although the neural network technique required a somewhat longer training period, it demonstrated high accuracy and sensitivity, making it a strong rival for conventional classifiers. Conversely, the creation of Bayesian classifiers was shown to be remarkably simple, despite their lower accuracy.

Clark et al. [12] demonstrated how their backpropagation-based system surpassed numerous other algorithms in terms of classification performance. It was also investigated how different approaches, such as feature selection weighting, and normalization, affected all the results. The normalization technique at mailbox level using frequency as well as tf-idf weighting yielded the greatest results in email spam filtering.

Recurrent neural networks (RNNs) in the form of long and short-term Memory Networks (LSTM) were used by the authors of articles [5] and [6]. After experimenting with a number of machine learning models, including Random Forest, SVM, Naive Bayes, Decision Trees, KNNs, and Logistic Regression, they discovered that LSTM outperformed them all with a 98.5% accuracy rate..

The research report [7] employed BiLSTM as a deep learning technique to detect spam communications. The results of their investigation demonstrate that this model works better than other ML algorithms, such as Naive Bayes, Bayes Net,SVM,K nearest neighbor, J48, as well as decision trees, with an accuracy rate of 98.6%. The researchers tested their method on two datasets: the UCI datasets and ExAIS_SMS, an SMS spam corpus with an African-English context [8].

The work's authors [9] tested many supervised and unsupervised machine learning models, Naive Bayes, Bayes Net,SVM,K nearest neighbor, J48, as well as decision trees including logistic regression using UCI datasets . Out of all the three combinations that were tested, Kmeans-NB, K Means-LR, and K Means-SVM, produced the best accuracy of 98.8%.

In the last few years, numerous research projects have focused on classification as well as detection of SMS spam. This research included a various number of machine learning

methods that include deep learning [9, 10], Naive Bayes [6–8], and the hidden markov model. pre-existing languages models that have been trained [12, 13], etc. In earlier studies [9,12], we put out models to identify spam communications that combine Arabic and English. In the first technique, a hybrid deep learning classifier is used to stack. The second method classified the collected texts using a pre-trained language model called "BERT" and an MLP network. The experimental evaluation revealed 98.37% accuracy with the first model and 99.45% accuracy with the second.

In [10], Roy et al. used deep learning algorithms for determination of which messages should be reported as spam and also which ones should we ignore . The main objective was to set up a system to distinguish between legitimate and spam SMS messages. A variety of ML techniques, such as random forest, Naive Bayes, gradient boosting, Logistic Regression and Stochastic Gradient Descent, were used to compare them with their methodology.

In order to identify Smishing messages and reduce false positives, the authors of [6] developed a technique they dubbed the "Smishing Detector". The proposed model was divided into four sections. The initial component of the system scanned the text it received. Communications are scanned for keywords and potentially harmful content using "Naive Bayes" categorization technology. The second step consisted of looking at the URLs in the mails. Thirdly, it involved going over the source code pertaining to messages on the website. The download detector was the last feature, and it determined whether a download URL was connected to a fake APK. By doing an empirical assessment, the authors determined that the model's accuracy was 96.29%.

A different model for SMiShing message identification was proposed by Joo et al. [7]. This model consists of four components: An analyst for message content analysis, a classification determinant, and a brief message service module for activity tracking and SMiShing text message blocking; all of that information is additionally stored in a database. The Naive Bayes classification method was used in this model.

The authors of [2] introduced a model they named "SmiDCA," which uses machine learning techniques to identify SMiShing messages. The Random Forest classifier demonstrated an accuracy rate of 96.4% in the trial data.

The author of [14] proposed a processing technique that standardizes textual communications in order to increase the effectiveness of text message classification systems. Semantic analysis, semantic dictionaries and disambiguation techniques served as the foundation for this process. The main goal was to improve the original content by adding new features and standardizing the language while reducing elements that could negatively impact performance, such as verbosity and ambiguity. Arifin et al. presented a method of SMS spam filtering using Naive Bayes and FP-growth in 2015. The Naive Bayes approach is used to classify the text messages once FP-growth has extracted the set of often recurring items, as well as stop sending unauthorized emails. An overall accuracy of 98.5% was obtained from the experimental investigation of this strategy.

The authors of [11] introduced a Hidden Markov model for spam message filtering that is based on a weighted feature method. Developing a system of word weighting that included the differences in word distribution between cheese and trash correspondence. This method was then applied to the development of an SMS word labeling system, yielding an HMM observation sequence. The trial results indicated that the accuracy of this model was 96.9%.

Liu et al. published a modified version of this technique for identifying SMS spam in [13], which was based on the Vanilla Transformer. Text messages were transformed into vector representations by using this model in conjunction with the GloVe technique. Tests carried out using this model revealed precision of 98.92%.

2.3 Comparative Analysis and Summary

Table 2.1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Yong Fang ¹ , Yunyun Zhang and Cheng Huang ¹ , et al.[1]	Light Gradient Boosting Machine algorithm, RF and GBM model.	LightGBM = 99%
2.	Ekta Gandotra and Deepak Gupta et al. [2]	RF, J48 and IB1	RF= 96.52%
3.	Mr. Thirunavukkarasu.M1 ,Achutha Nimisha ² , Adusumilli Jyothsna ³ et al. [3]	RF	Random Forest=99.93%
4.	Ala Mughaid ¹ , Shadi AlZu'bi ² , Adnan Hnaif, Salah Taamneh ¹ , Asma Alnajjar ¹ , Esraa Abu Elsoud ¹ et al [4]	Locally-deep support vector machine ,support vector machine ,boosted decision tree ,logistic regression ,averaged perceptron ,neural network ,decision forest	Decision tree and Neural Network.=97.05%
5.	Ozgur Koray Sahingoz a, Ebubekir Buber b, Onder Demir b, Banu Diri c et al. [5]	Naive bayes ,random forest, kNN (n = 3), adaboost, K-star ,SMO ,and decision tree	random forest= 97.98%

6.	Jibrilla Tanimu, Stavros Shiaeles et al. [6]	decision tree ,support vector machine , naive base, 2D and 3D Neural Networks,	—
7.	Anindita Khade , Dr. Subhash K Shinde et al. [7]	fuzzy logic also the RIPPER classification algorithm.	RIPPER algorithm = 85.4%
8	AliAljofey1,2, Qingshan Jiang1*, Abdur Rasool1,2, Hui Chen1,2, Wenyin Liu3, Qiang Qu1 &YangWang4 et al. [8]	LR, XGBoost, RF, NB, DNN, LSTM, CNN	XGBoos=98.28%
9	Pradheepan Raghavan, Neamat El Gayar et al. [9]	RBM, Autoencoders, Random Forest, CNN, SVM, KNN	RF= 86.05%
10	Mahdi Maktabdar Oghaz, Mohd Aizaini Maarof, Anazida Zainal et al. [10]	naive bayes ,multinomial naive bayes ,bernoulli ,support vector machine, stochastic Gradient Descent , logistic Regression ,voted Classifiers	Multinomial Naïve Bayes= 99.41%
11	This study	Naive Bayes, K-Nearest Neighbor, S V M, Random forest. AUC	Naive bayes=97%

2.4 Scope of the Problem

This research actually makes a specialty of efficiently locating in a manner to develop SMS detection devices for blind people who are not sufficiently aware about confidentiality and can not protect data from insecurities . This research aim is to construct a system for a highly maintained secure way of data and prevent data loss. Our project turned into a realistic way of applying and getting benefits in real life.

2.5 Challenges

Here the goal is to create a reliable system that has improved usability and accuracy so that we can produce one of these for the people who aren't concerned enough. In the process of improving this research, which is mostly based on, some of the challenging factor below there:

- Data collection: We have to face many difficulties in our efforts to customize our big amount of information and categorize them in spam and ham.
- Time Complexity: Project aims to achieve better results as soon as possible. But sometimes it occurs to not to run and finish this project in the right time. so it could be a great trouble.
- Hardware roadblocks: The execution speed of the dataset may be slowed by obstacles in our hardware. We need to add more powerful hardware, like a CPU, HDD, or GPU, to our architecture is an easy way to solve this problem.
- Data transfer: Sometimes we need this data from local people or many other sources. From collecting this data it might be a problem to transfer from one hand to another device. And it might be the cause of losing enough test and train data.
- Need to preprocess the big dataset and to eliminate the missing value. It could be the reason for trouble.

CHAPTER 3

Research methodology

3.1 Research Subject and Instrumentation

This study is based on an ML algorithm to detect SMS scamming as it is becoming a huge threat to our society that has cost us a huge financial loss lately. So this paper prototyped and evaluated this strategy using five machine learning classifiers: naive bayes ,K-Nearest ,decision tree ,SVM ,random Forest.

3.2 Data Collection Procedure/Dataset Utilized

For the purpose of this studyI need two kinds of data. One could be spam and the other will be ham. So I used two different kinds of dataset.

- Collecting data of spam SMS from online platform Kaggle
- Collecting data of ham text from online platform kaggle

Kaggle Data Store: Kaggle is a famous online website for collecting data. In the case of spam text, those are not much more like ham text. But if we want to gather those information about spam and ham that could be difficult and time consuming. It will take time to find out ham and spam text and gather them as a dataset. Therefore, Kaggle came up with the best option; they provide the biggest dataset with short time and low effort.

3.3 Statistical Analysis

In this research there are only two kinds of attributes in this dataset. Spam and ham related dataset is chosen for this work. As the valid and non valid text is in one dataset so need to store in a local file this dataset for use.

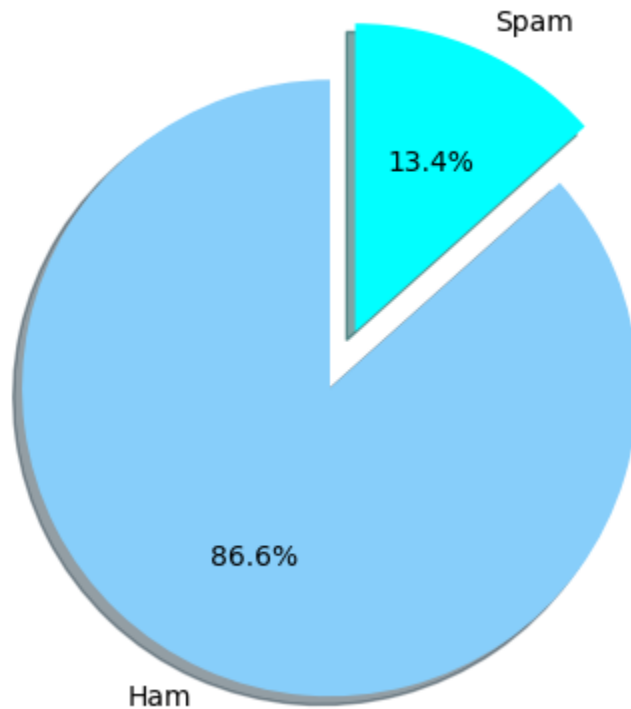


Figure 3.1: Data plot pie chart

For this analysis , data needs to be plotted and plotting was all about size, explode, and label . So data labeled was into ham and spam data. Here data sizes were counted from 0 to 1. As the ML model can not read the data when needed to send this work train and test data set so it will be counted as a numerical number and they differentiate the ham data 86.6% and spam 13.4% and that data will be used for further analysis and classification.

3.4 Proposed Methodology/Applied

Our thesis goal is to identify spam text for which we have used custom naive bayes , K-Nearest ,decision tree , SVM , random forest trained models to get our best accuracy. After data gets preprocessed we label and split it into two different sectors. Then the

train and test data go through into different algorithms and matrices and finally get our result and best accuracy. So the proposed methodology below there:

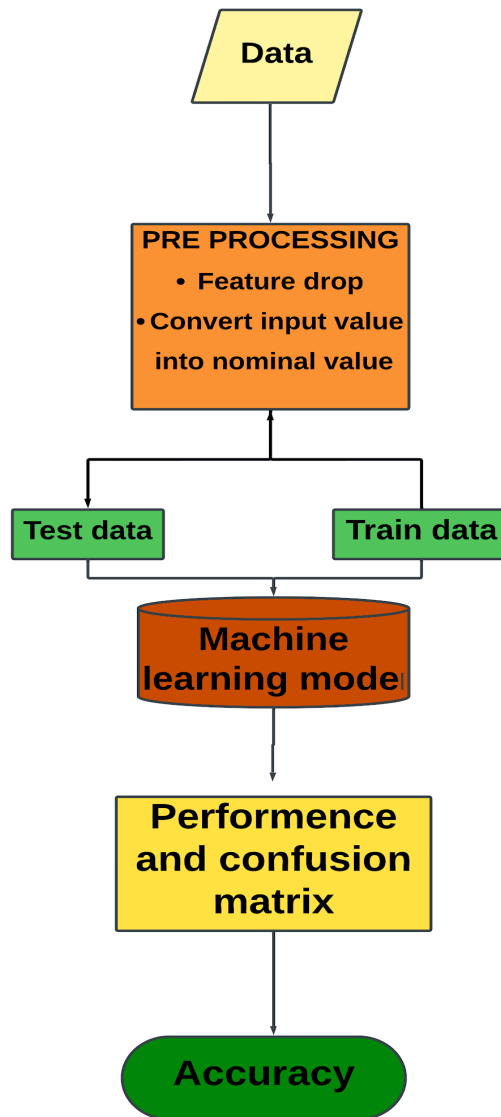


Figure 3.2: Proposed methodology of spam text detection algorithm

This paper proposed methodology in the first place we need a dataset. Then did the preprocessing and got train data and test data. And apply five different kinds of algorithms in this data and get the result . After that this paper got the result with the highest accuracy model and this work analyzed the best accuracy model efficiency with Auc and ROC.

- **Data:** For this thesis work collect data set from kaggle.com and this data set has five attributes . which was ham, spam and other three were null attributes. Ham and spam attributes carry all the valid and non valid messages and the other three attributes carry the null section.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	v1	v2																	
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...																	
3	ham	Ok lar... Joking wif u oni...																	
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's																	
5	ham	U dun say so early hor... U c already then say...																	
6	ham	Nah I don't think he goes to usf, he lives around here though																	
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, â€1.50 to rcv																	
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.																	
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune																	
10	spam	WINNER!! As a valued network customer you have been selected to receive a â€900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.																	
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030																	
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.																	
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info																	
14	spam	URGENT! You have won a 1 week FREE membership in our â€100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18																	
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times																	
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!																	
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap.xxxmobilemovieclub.com?n=QJGIGHJGCBL																	
18	ham	Oh k...i'm watching here:)																	
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.																	
20	ham	Fine if that's the way u feel. That's the way it's got to b.																	

Figure 3.2: Dataset

- **Pre-processing:** After collecting the data set this work needs to be preprocessed to get the data for train and test and applying the choosing algorithms. So in the preprocessing sector the dataset needs to feature drop and this drops the unnecessary null section. After dropping the null data there are spam and ham attributes and this input converts into numerical value for work purpose.
- **Train and test dataset:** After preprocessing this feature consumes 30% dataset for testing purpose and the rest of the dataset will count as a train dataset. This data set counting percentage will be 70%.

- **ML model:** After getting this train and test dataset applying the five different type of ML algorithms in the dataset. These five different algorithms are - Naive Bayes, K-Nearest, decision tree, support vector machine, random forest. This work is gonna apply all these different models in this dataset for better accuracy.
- **Result:** After applying these different kinds of models they showed their performance percentage. And every different kind of algorithm performs with a different accuracy and confusion matrix. This paper chose the best performance and the result will choose the algorithms.
- **Efficiency:** After choosing the best algorithm with highest accuracy this model needs to check the performance efficiency of their algorithms. So this work needs to take the best performance algorithms to the ROC and AUC curve and this will show the actual efficiency of these algorithms.

3.5 Implementation Requirements

After the earlier - Naive Bayes tasks are finished, I use the data set to verify accuracy. I separated my work into 10 main categories to make it easier to implement. To ensure the success of my project, the following steps must be taken.

- Collecting data from online data storage kaggle.
- Data Preparation and labeled the dataset
- splitting the labels and make the data separately
- Data visualization
- Plotting ham and spam data percentage in pie chart
- Splitting the test and train data
- Extracting N-grams from the text data

- Implement different kind of algorithms and matrix
- Compared the method
- Result
- Efficiency

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

This dataset is collected mainly from online data storage kaggle and prepares the dataset for labeling. Then split the data into spam and ham and need to visualize it. Then I need to plot a pie chart in spam and ham data and then split this data into a test and train data set. After splitting i got the test data 3733 and train data 1839. Then extract N-grams from the text data and apply 5 different kinds of algorithms. Those algorithms were:naive bayes ,K-Nearest ,decision tree ,SVM random forest. So after that we compared the method for better accuracy and got the result .

4.2 Experimental Results & Analysis

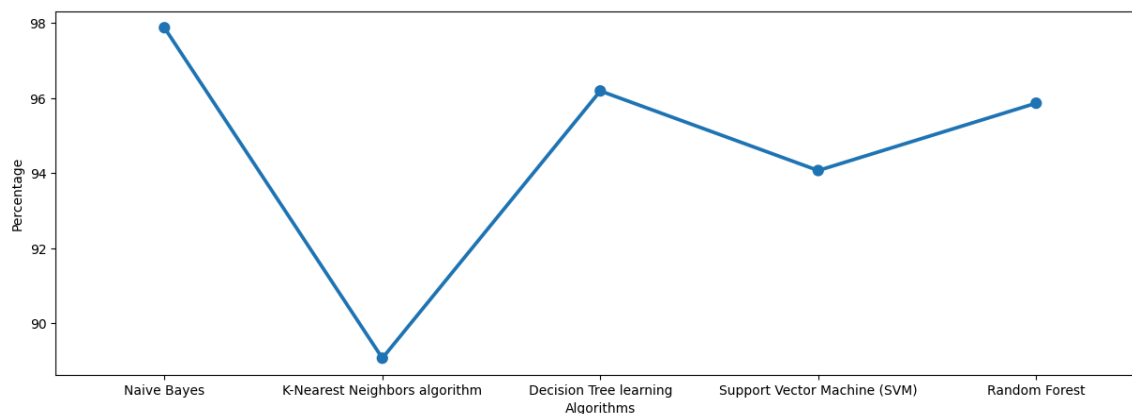


Figure 4.1: comparison of ML algorithm

Algorithms	Percentage	
0	Naive Bayes	97.879282
1	K-Nearest Neighbors algorithm	89.070147
2	Decision Tree learning	96.193583
3	Support Vector Machine	94.072866

Confusion matrix

Confusion matrix is basically a certain type table that is offered to describe any classification algorithm's performance. A confusion matrix properly visualizes and summarizes any classification methodologies actual performance precisely.

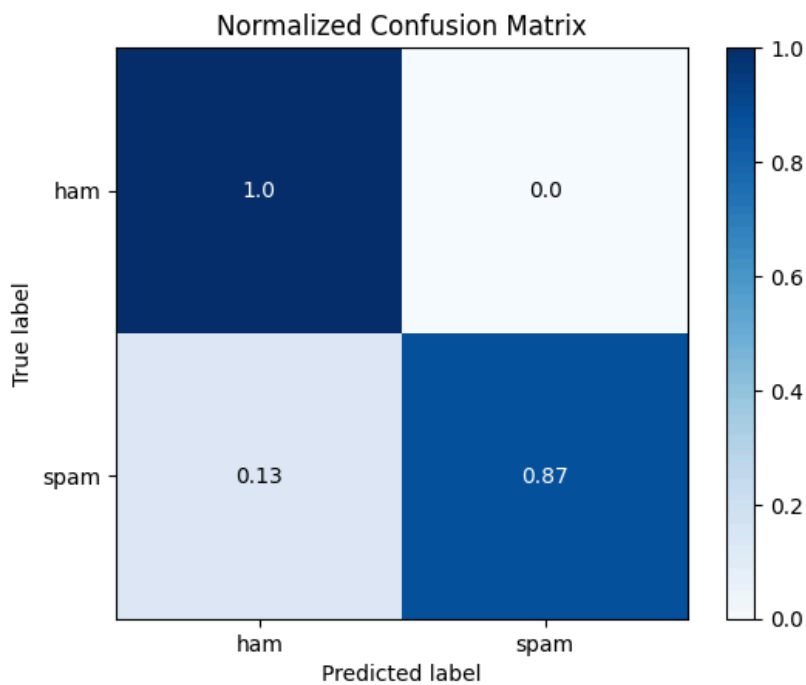


Figure 4.2: the confusion matrix of naive bayes

Confusion matrix of K-Nearest Neighbors algorithm:

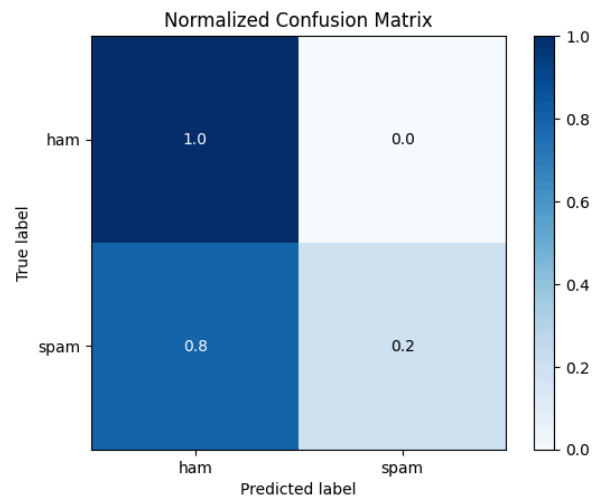


Figure 4.3 :the confusion matrix K-Nearest Neighbors algorithm

Confusion matrix of Decision Tree learning:

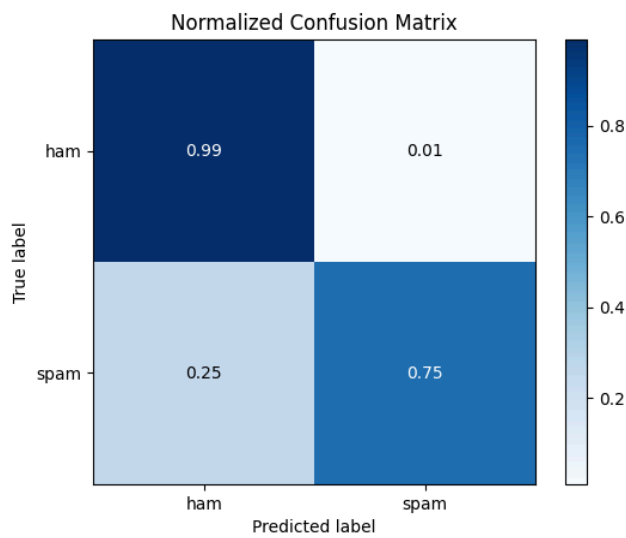


Figure 4.4:the confusion matrix of decision tree

The confusion matrix of Support Vector Machine (SVM):

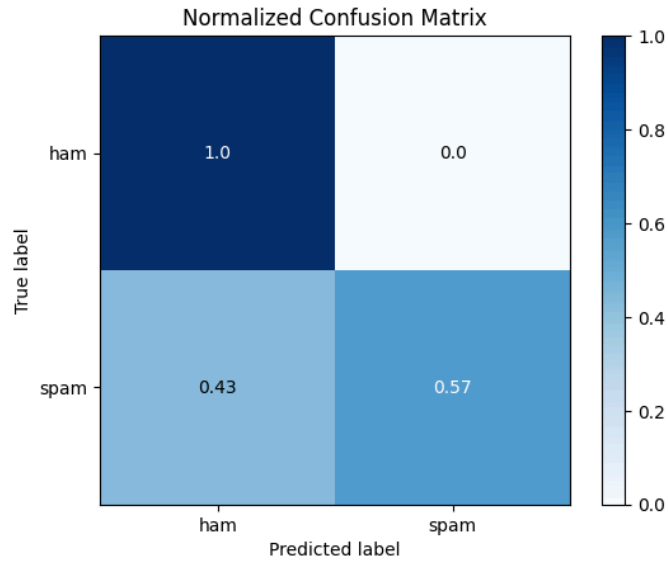


Figure 4.5: the confusion matrix of Support Vector Machine (SVM)

The confusion matrix of random forest:

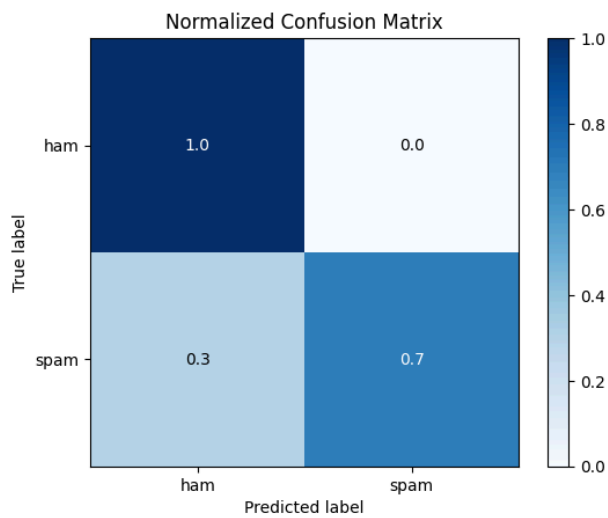


Figure 4.6: the confusion matrix of random forest

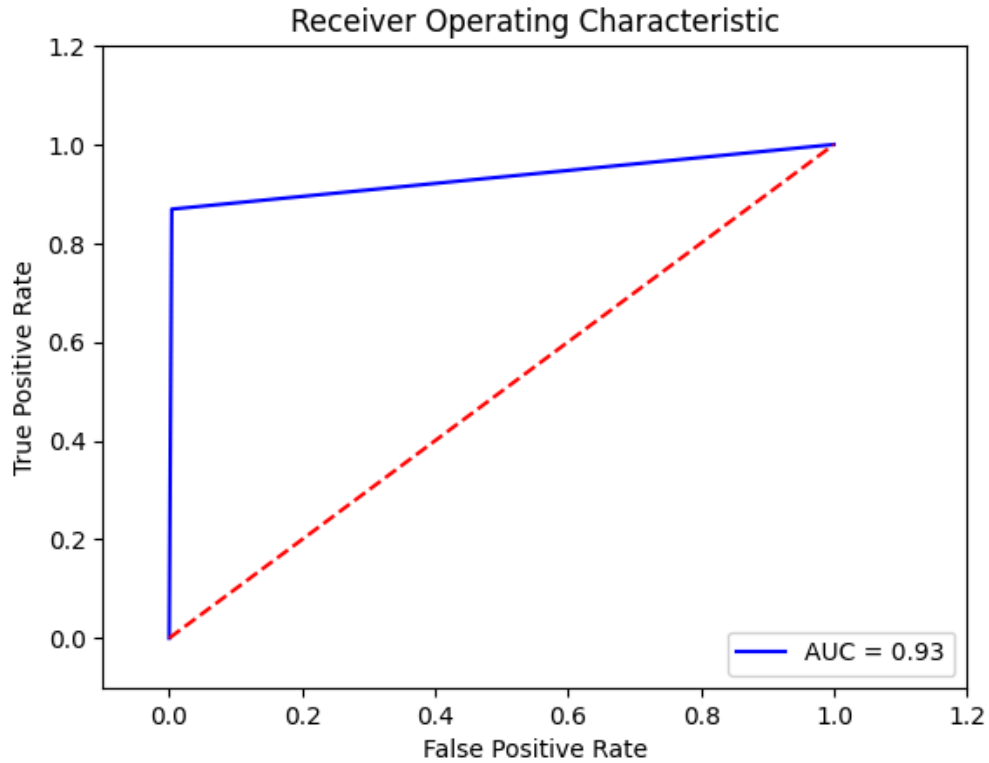


Figure 4.7: presentation of AUC

4.3 Discussion

Table 4.8: result table of all the algorithm

	Naive Bayes	K-Nearest neighbors algorithm	decision tree	support vector machine	random forest
Accuracy Rate	97.879282	89.070147	96.193583	94.072866	95.867319

In this paper after applying all those classifiers here naive Bayes showed 97.88% of accuracy, K-Nearest Neighbors algorithm showed 89.07% accuracy, Decision Tree learning accuracy was 96.13%, and Support Vector Machine (SVM) accuracy showed 94.07%. And after applying Random Forest that accuracy performance was 95.87%. After applying all of those classifiers here the best accuracy provider was Naive Bayes with the best percentage of accuracy.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Scam texts lately emerged as a significant Internet security problem. It has always posed a danger to internet users' privacy and security. Its a type of spoofing in which the attackers employ spoofing tactics such as bogus text that seem like the real thing to deceive users into disclosing private credentials as user id , password , credit card number also bank account details , among other things. We choose to apply this notion to detect fake sms and alert our visitors about them in order to save personal information and minimize financial loss because it is a serious concern.

5.2 Impact on Environment

When spam text enters our device, it's annoying, harmful and at the same time it is very dangerous to protect confidentiality. Although built-in spam detection filters are included in all SMS platforms, they are frequently overlooked. As an alternative, spending money on effective anti-spam software will reduce device carbon footprint and maintain an organized inbox. The more complex the text is the carbon footprint will rise so the more energy it will require and will impact on climate change. So if we can reduce the spam text the carbon footprint will decrease and the environment pollution rate will be more safe.

5.3 Ethical Aspects

Ethical aspects refers to passing info between two media which should be secured, transparent, and in a respected way. Our proposed method is fully committed to provide security, confidentiality and clear and precise information. They will carry integrity, compassion, loyalty, law abiding, and environmental concern. They are fully free of harassment and discrimination for business purposes. Any communication should maintain most three ethical concerns and they are much more into it.

5.4 Sustainability Plan

We are used to maintaining our life in a modern purpose, better and time saving communication way. Nowadays both sender and receiver accept sms for many important facts. Spammer's take advantage of this and they want to gain access like business info, market penetration, financial matters, credit card information and so many more. SMS spam is going to be increased as well as will increase business handle and sharing information through sms. People need to maintain a secure way of communication and need to protect confidential information. And it will increase day by day. People will continue their communication way through SMS, and spammer's will continue to send spam SMS to carry their business. So if people are not concerned about data loss and not gonna take action and follow the privacy issues they will suffer to lose their valuable data. So spam sms detection needs to be increased . In our project plan , we proposed a method to detect spam text . This paper aims to create a ML based model to identify spam SMS . So in the end of the technology communication will continue through SMS and this problem will be increased. So this plan is more valuable for protecting our data security than we thought.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

The ultimate goal of this research is to reliably identify scam language using multiple classification methods, and in the first step, we segregated our dataset into two primary components: training and testing. We initially sought to extract from the dataset some helpful properties that may be utilized to detect spam content. Data preprocessing is then used to clarify and prepare the data. Then we compare the outputs of the naive bayes , K-Nearest , decision tree , SVM , random forest . to get the optimal outcome.

6.2 Conclusions

The primary focus of this research was the evaluation and discussion of all machine learning methods for spam and ham SMS identification. This theory conducted comparisons between 5 distinct ml classifiers such as naive bayes , K-Nearest , decision tree , SVM , random Forest.

6.3 Implication for Further Study

We have completed our research and acquired unprocessed data. With this data, we have been able to recognise spam SMS. We consider future intentions for this very research. also The future is with us right now . This paper aims to improve the accuracy of the algorithms and collect new dataset and is going to apply this method for better performance. We will collect more data for better work and our results via an online application that is considerably more user-friendly. A graphic presentation of the data analysis will be made. This effort has produced important findings that allow the research to be applied in the real world to the detection of spam SMS.

Reference:

- [1] Gadde, S., Lakshmanarao, A. and Satyanarayana, S., 2021, March. SMS spam detection using machine learning and deep learning techniques. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 358-362). IEEE.
- [2] Shirani-Mehr, H., 2013. SMS spam detection using machine learning approach. *unpublished* <http://cs229.stanford.edu/proj2013/ShiraniMeh>
r-SMSSpamDetectionUsingMachineLearningApproach.pdf.
- [3] Jain, T., Garg, P., Chalil, N., Sinha, A., Verma, V.K. and Gupta, R., 2022, January. SMS spam classification using machine learning techniques. In *2022 12th international conference on cloud computing, data science & engineering (confluence)* (pp. 273-279). IEEE.
- [4] Gupta, M., Bakliwal, A., Agarwal, S. and Mehndiratta, P., 2018, August. A comparative study of spam SMS detection using machine learning classifiers. In *2018 eleventh international conference on contemporary computing (IC3)* (pp. 1-7). IEEE.Reference
- [4] Ghourabi, A. and Alohal, M., 2023. Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning. *Sensors*, 23(8), p.3861.
- [5] Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q. and Wang, Y., 2022. An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, 12(1), p.8842.
- [6] Raghavan, P. and El Gayar, N., 2019, December. Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 334-339). IEEE.
- [7] Fang, Y., Zhang, Y. and Huang, C., 2019. Credit Card Fraud Detection Based on Machine Learning. *Computers, Materials & Continua*, 61(1).
- [8] Maktabar, M., Zainal, A., Maarof, M.A. and Kassim, M.N., 2018. Content based fraudulent website detection using supervised machine learning techniques. In *Hybrid Intelligent Systems: 17th International Conference on Hybrid Intelligent Systems (HIS 2017) held in Delhi, India, December 14-16, 2017* (pp. 294-304). Springer International Publishing.
- [9] Maktabar, M., Zainal, A., Maarof, M.A. and Kassim, M.N., 2018. Content based fraudulent website detection using supervised machine learning techniques. In *Hybrid Intelligent Systems:*

17th International Conference on Hybrid Intelligent Systems (HIS 2017) held in Delhi, India, December 14-16, 2017 (pp. 294-304). Springer International Publishing.

[10] Gandotra, E. and Gupta, D., 2021. An efficient approach for phishing detection using machine learning. *Multimedia Security: Algorithm Development, Analysis and Applications*, pp.239-253.

[11] Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE.

[12] Rashid, J., Mahmood, T., Nisar, M.W. and Nazir, T., 2020, November. Phishing detection using machine learning technique. In *2020 first international conference of smart systems and emerging technologies (SMARTTECH)* (pp. 43-46). IEEE.

[13] Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A. and Elsoud, E.A., 2022. An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, 25(6), pp.3819-3828.

[14] Amiri, I.S., Akanbi, O.A. and Fazeldehkordi, E., 2014. *A machine-learning approach to phishing detection and defense*. Syngress.

[15] Kanchana, M., Chavan, P. and Johari, A., 2020. *Detecting Banking Phishing Websites Using Data Mining Classifiers*.

[16] Dong, X., Clark, J.A. and Jacob, J.L., 2008, October. User behavior based phishing websites detection. In *2008 International Multiconference on Computer Science and Information Technology* (pp. 783-790). IEEE.

[17] Aburrous, M., Hossain, M.A., Dahal, K. and Thabtah, F., 2010, April. Predicting phishing websites using classification mining techniques with experimental case studies. In *2010 seventh international conference on information technology: New generations* (pp. 176-181). IEEE.

SPAM TEXT DETECTION BY USING ML ALGORITHM

ORIGINALITY REPORT

16%

SIMILARITY INDEX

15%

INTERNET SOURCES

3%

PUBLICATIONS

11%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	3%
3	eprints.utar.edu.my Internet Source	1%
4	www.jetir.org Internet Source	1%
5	Submitted to The Robert Gordon University Student Paper	1%
6	Submitted to TechKnowledge Student Paper	<1%
7	M. Michael Gromiha, Shandar Ahmad, Makiko Suwa. "Neural network based prediction of protein structure and Function: Comparison with other machine learning methods", 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008 Publication	<1%

8

link.springer.com

Internet Source

<1 %

9

Submitted to Asia Pacific University College of Technology and Innovation (UCTI)

Student Paper

<1 %

10

www.researchgate.net

Internet Source

<1 %

11

Jibrilla Tanimu, Stavros Shiaeles. "Phishing Detection Using Machine Learning Algorithm", 2022 IEEE International Conference on Cyber Security and Resilience (CSR), 2022

Publication

<1 %

12

Shweta Dasharath Shirsat. "Demonstrating Different Phishing Attacks Using Fuzzy Logic", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018

Publication

<1 %

13

Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal, Pulkit Mehndiratta. "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers", 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018

Publication

<1 %

14

journals.plos.org

Internet Source

<1 %

15 patents.google.com
Internet Source

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off