A SMART APPROACH FOR DETECTING EMAIL SPAM TEXT DETECTION

BY

FARHANA AKTER SMRITEE
ID: 191-15-12467

AHMED ZUBAYER SUNNY
ID: 191-15-12960

SAYEM AL MAHADI
ID: 191-15-12949

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

Supervised By

Mr. Narayan Ranjan Chakraborty
Associate Professor
Department of CSE
Daffodil International University

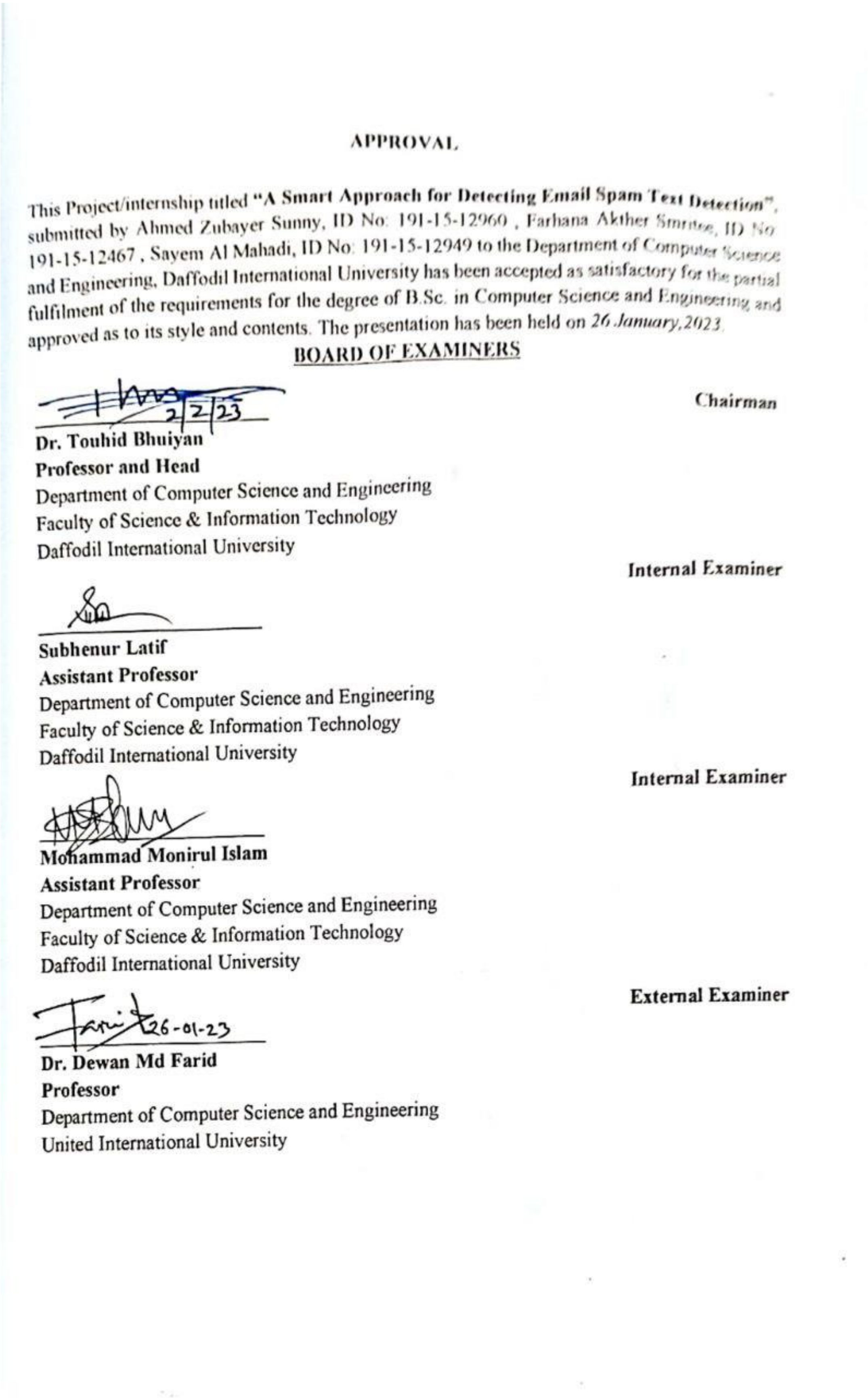
Co-Supervised By

Mr. Md. Tarek Habib
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023



APPROVAL

This Project/internship titled **"A Smart Approach for Detecting Email Spam Text Detection"**, submitted by Ahmed Zubayer Sunny, ID No: 191-15-12960 , Farhana Akther Smritee, ID No 191-15-12467 , Sayem Al Mahadi, ID No: 191-15-12949 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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,2023.*

BOARD OF EXAMINERS

2/2/23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Subhenur Latif
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

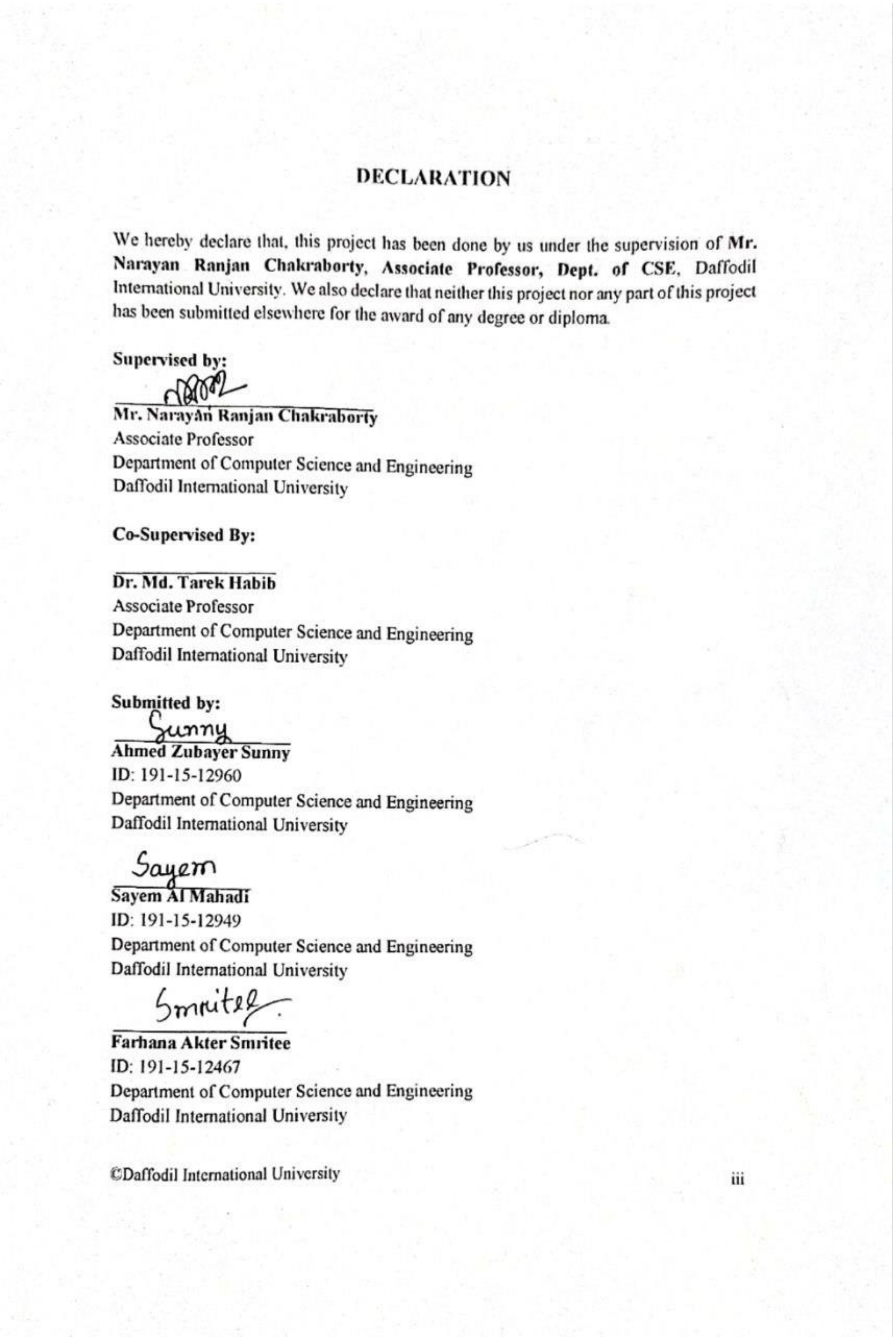
Internal Examiner

Mohammad Monirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

26-01-23

Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University



DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Narayan Ranjan Chakraborty, Associate Professor, Dept. of CSE**, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Narayan Ranjan Chakraborty
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By:

Dr. Md. Tarek Habib
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Sunny
Ahmed Zubayer Sunny
ID: 191-15-12960
Department of Computer Science and Engineering
Daffodil International University

Sayem
Sayem Al Mahadi
ID: 191-15-12949
Department of Computer Science and Engineering
Daffodil International University

Smritee
Farhana Akter Smritee
ID: 191-15-12467
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

And foremost, we offer our heartfelt appreciation and gratitude to Almighty God for His divine gift, which has enabled us to successfully finish the final year proposal.

We are really grateful and wish our profound indebtedness to **Mr. Narayan Ranjan Chakraborty**, **Assistant Professor**, Department of CSE,Daffodil International University, Dhaka. Our supervisor has extensive knowledge and a great interest in the subject of Deep Knowledge & keen interest of my supervisor in the field of "Deep Learning, Machine Learning" to carry out this paper. His unending patience, scholarly guidance, constant encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages, and reading many inferior drafts and correcting them at all stages enabled us to complete this project.

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

We would like to thank everyone of our Daffodil International University classmates who participated in this discussion while completing their course work.

ABSTRACT

In practically every industry today, from business to education, emails are used. Ham and spam are the two subcategories of emails. Email spam, often known as junk email or unwelcome email, is a kind of email that can be used to hurt any user by sapping their time and computing resources and stealing important data. Spam email volume is rising quickly day by day. Today's email and IoT service providers face huge and massive challenges with spam identification and filtration. Email filtering is one of the most important and well-known methods among all the methods created for identifying and preventing spam. The amount of unwanted emails has increased due to the increased usage of social media globally, making the implementation of a reliable system to filter out such issues necessary. On the internet, spam emails are the most prevalent issue. Sending an email with spam messages is a straightforward process for spammers. Spammers are capable of stealing crucial data from our devices, including contacts and files. In recent years, numerous deep learning-based word embedding techniques have been created. This study provides an overview of various machine learning techniques (MLTs) for email spam filtering, including Naive Bayes, K-Nearest Neighbor, Logistic Regression, Gradient Boosting Classifier, and Random Forest. However, in this article, we classify, assess, and compare various email spam filtering systems and provide a summary of the overall situation with relation to the accuracy rate of various currently used methods. I got the best accuracy that was 98% with the help of The Random Forest Classifier algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Even though the recipient does not immediately pick up the letter, it is still waiting to be opened in the mailbox. A single email can also be sent simultaneously to several recipients. As a result, email is quick and economical. A user's mailbox may occasionally contain unsolicited emails in addition to the many benefits that emails can offer. Spam emails on the Internet have long been a problem. The email server's memory is negatively impacted by its huge volume on the Internet. Furthermore, because of the deception of malicious individuals, spam emails could cause financial damage. Additionally, since readers of unwanted mail must read the entire message to determine whether it is such, they will be inconvenienced and spend time. The accuracy of the detection achieved by the system is around 98%.

The internet has recently developed several platforms that increase the security of human existence. Email is a significant venue for user contact among these. Email is merely referred to as an electronic communications infrastructure that allows users to send messages to one another [1]. Due to its numerous branches, including Yahoo mail [2], Gmail [3], Outlook [5], and others, which are all entirely free for all online users by adhering to some administration [6, 7], email has today become a common medium [2]. Given its many uses today, email is regarded as a secure global communication tool. However, some "Spam Emails" can make emailing more dangerous.

Because of the aforementioned problems, it is difficult for academics to create an effective filtering mechanism that can detect spam e-mails. Two popular techniques for filtering unwanted emails are wisdom engineering and machine learning. A set of guidelines are necessary for knowledge engineering. Due to the requirement for ongoing rule set updates, it is a weak technique.

Machine learning is effective for this since a rule set is not necessary. A collection of training and test data is used in machine learning. Training data is made up of emails that have already been flagged as spam or junk mail. Whether an email is unsolicited or not can be determined in large part by NLP techniques.

Spam messages are those that the beneficiary has mentioned not to get. To much email beneficiaries, much duplicates of a similar message are sent. Offering our email clue on an illicit or untrustworthy site now and again brings about spam. Spam has a large number of adverse consequences. countless inept messages fill our inboxes. significantly decreases the speed of our Web. takes crucial data from your contact list, like our information. change the query items you get from any PC programming.

Email filtering has been the subject of numerous research projects, some of which have yielded promising results and others that are still in progress. The practice of sorting emails based on specific criteria is known as email filtering, according to the researcher's perspective. Inbound and outbound filtering are two of the many methods for email filtering that are available. Outbound filtering reads messages from local users, while inbound filtering reads messages from internet addresses. Moreover, spam filtering, which works through antispam technology, is the most efficient and practical email filtering. Because spammers have proactive personalities and employ dynamic spam structures that are constantly evolving to thwart anti-spam measures, spam filtering is a difficult undertaking [9, 10].

1.2 Motivation

The goal of this project is to reveal a system or model that can use historical data to forecast if mail is spam or not. Various researchers have attempted to solve this problem in a variety of ways to see which strategy works and produce the best results.

- It provides useful information.

- Be wary about forwarding such an article to others.

- True emails are revealed.

- Preventing the occurrence of fictitious crises

1.3 Rationale of the Study

We read research papers from different sites. After reading lots of papers, then we decided the title and start working for paper. Actually, few days ago my friends had sent an email and said that if I log into the facebook from this link which was sent, then I see all friend requests that have sent from my facebook account. That was spam email and this types of hacking is called fishing attack. I think, this types of hacking must be prevented by machine learning.

1.4 Research Questions

- What is Email?
- What is spam Email?
- Why is it a problem?
- Will we detect spam?
- Which algorithms will give the best accuracy?

1.5 Expected Output

- Good knowledge about algorithms.
- Know about the spam detection.
- Can secure email from spam text.

1.6 Project Management and Finance

Here, we are three members and working together. Then, we collect data from online news portal and different sites. Then pre-processing the datasets. After pre-processing, we work in google colab and total datasets divided two ways training datasets and testing datasets.

80% for training and 20% for testing datasets. Then, we used various library functions and get output.

1.7 Report Layout

This report varied in a total of six different chapters. Which are capable of extending the understanding of “Mail Spam Detection” more briefly. In the first chapter, we’ll mention introduction, motivation, research questions and the last one is the expected outcome. In the second chapter, we’ll brief about some related works, which types of challenges that we had faced and about the research summary. In the third chapter, we’ll talk about our research subject and instrumentation, workflow of the model. In the fourth chapter, we’ll talk about the result that we got, the detecting way of spam. In the fifth chapter, we’ll describe its impression on our society, impression on our environment and sustainability. In the sixth chapter, which is our last chapter, we’ll mention the conclusion and our future works.

CHAPTER 2

Background Study

2.1 Introduction

The system is a Web application that assists users in identifying bogus news. We've provided a text box where the user may paste the message or the URL link to the news or another message, and it will then display the truth about it. All data provided by the user to the detector may be saved for future usage to update the model's state and conduct data analysis. We also assist users by providing instructions on how to avoid such bogus events and how to stop them from spreading.

Spam filtering is a technique that looks for unwanted email and blocks it from reaching users' inboxes. Various systems are available. Create an anti-spam strategy to stop unsolicited bulk emails. Most anti-spam techniques exhibit some inconsistency between false negatives and false positives which serves as a barrier for the majority of systems to create effective anti-spam systems. Web users, therefore, have the greatest need for an intelligent and effective spam-filtering solution. This existing technology can assist us in employing machine learning to train our model.

In the realm of quickly expanding innovation, data sharing has turned into a simple assignment. There is no question that the web has made our lives more straightforward and given us admittance to loads of data. This is an advancement in mankind's set of experiences, and yet, it unfocussed the line between spam messages. A collection of protocols are used in the email spam filtering process to identify whether a message is a spam or not. There are several spam filtering methods available right now. The Standard Spam Filtering Process is one of them and adheres to a set of rules and procedures while serving as a classifier. The figure demonstrates how an ordinary spam filtering procedure carried out the analysis by doing a few things [14]. The first one is content filters, which use a variety of machine-learning algorithms to identify spam messages [8, 10, 15–18]. Various models are utilized to give a precision extent of 60-75%. which incorporates the Guileless Bayes classifier, semantic highlights based, limited option tree model, SVM, and others. The boundaries that are thought about don't yield high precision. This undertaking intends to expand the precision of identifying spam more than the current outcomes that are accessible. Manufacturing this new model, which will pass judgment on fake news stories based on specific standards like spelling botches, muddled sentences, accentuation mistakes, and words utilized.

2.1.1 Need for a new system:

Many individuals are now using the internet as a central platform to gather information about the world's reality, and this trend must continue. As I previously stated, we will develop an email spam detection program that will determine the truth of the news and message.

Users of our website can see the most up-to-date information about the major sources or phrases that are receiving the phoniest news and messages, as well as a map marked up with a chart. After all, everyone wants to know how to avoid this, so we're providing some helpful hints for avoiding spam text that spreads rumors throughout the world.

2.2 Related work:

Creators of the paper [5] announced digital assaults. Email administrations are regularly utilized by phishers and vindictive aggressors to convey counterfeit correspondences that can make target clients lose cash and their social adjustable. These lead to the robbery of private data, including passwords, Visa numbers, and other confidential data. The creators of this paper utilized Bayesian classifiers. Ponder each and every letter you get. continually advances to manage new spam sampling.

In a different study, [2] authors suggested a unique approach for Twitter spam identification based on deep learning (DL) techniques. The author uses both the content of tweets and users' meta-data to identify spammers (i.e. age of the account, number of followers, and so on). The authors of [3] compare standard machine learning methods for review classification. The authors of this work suggested a strategy based on attention and bidirectional LSTM for retrieving semantic information. In [4]. The authors suggested a novel method of differentiating between real and fraudulent texts. Horse herd meta-heuristic Optimization Algorithm is the method employed. Continuous HOA is used to build the discrete algorithm.

Rubin concentrated on the qualification between the items in genuine and comic news by means of multilingual elements, in light of a piece of near news (The Onion and The Beaverton) and genuine news (The Toronto Star and The New York Times) in four areas of common, science, exchange, and normal news. She got the best presentation in identifying counterfeit news with a bunch of elements including irrelevant, checking, and syntax.

The proposed approach in the paper[4] really tries to apply AI procedures to recognize an example of repeating catchphrases that are sorted as spam. The framework likewise recommends classifying messages in view of extra factors tracked down in their construction, like the space, header, and Cc/Bcc fields. Applying every boundary to the AI calculation would regard it as an element. The pre-prepared AI model will have an input system to separate between a right result and an uncertain result. This approach offers a substitute design for the execution of a spam channel. This paper likewise considers the email body, which might contain generally utilized words and punctuation. Spam recognition, in some measure in the space of spam discovery [7], utilizes factual AI procedures to characterize text (e.g., tweets [8] or messages) as spam or legitimate. These strategies include preprocessing of the message, highlight extraction (i.e., the sack of words), and component determination in light of which elements lead to the best presentation on a test dataset. When these elements are acquired, they can be characterized utilizing Gullible Bayes, Backing Vector Machines, TF-IDF, or K-closest neighbors classifiers. These classifiers are normal for regulated AI, implying that they require a piece of named information to get familiar with the capability where m is the message to be ordered and is a vector of boundaries and Spam and Cleg are, separately, spam and genuine messages.

Cspam and Cleg are spam and legitimate messages, respectively, and are parameter vectors. In that, they try to separate samples of genuine content from examples of illegitimate, ill-intended material, the challenge of detecting fake news is similar and almost analogous to the task of detecting spam.

There are two categories of important research in the automatic classification of real and fake news up to this point:

In the subsequent class, semantic methodologies and reality-thought procedures are utilized at a down-to-earth level to look at genuine and counterfeit items. Etymological methodologies attempt to recognize text highlights like composing styles and content that can assist in recognizing spamming text distinctions. The fundamental thought behind this method is that etymological ways of behaving like utilizing marks, picking different kinds of words, or adding names for parts of a talk are fairly unexpected, so they are past the creator's consideration. Subsequently, a proper instinct and assessment of utilizing etymological methods can uncover confident outcomes in identifying counterfeit news.

Authors in [8] developed a method for changing the email classification problem into a graph classification problem. This project doesn't demand that the email text be transformed into a vector format. In contrast, this approach uses a graph neural network to classify spam emails by converting the email's content into a graph (GNN). In [9], creators fostered a few strategies, including the B-TransE mode, to distinguish misleading news in view of information content and information charts. The creator gave different new ways to deal with recognizing counterfeit news in light of deficient and defective information diagrams, utilizing the current TransE model and the recently introduced B-TransE model.Authors in [5] focused on approaches to proficiently sharpen SMS spam. The Nave Bayes, Inclination Lift Calculated Relapse, SGD classifier, and Profound learning-based models like CNN and LSTM were among the AI-based classifiers that were tried. As per their discoveries, the CNN model, which had a precision of 99.44% on haphazardly created ten times cross-approval information, performed best for screening genuine instant messages. The methodology was in any case obliged by the way that it was absolutely reliant upon messages distributed in English. The Credulous Bayes (NB) technique and a computational knowledge procedure in view of Molecule Multitude Enhancement are consolidated in [6] utilizing a coordinated system (PSO). The Credulous Bayes calculation is utilized to learn and arrange email content in sequential requests.

Banday et al. [25] [2008] go over how statistical spam filters are created by integrating Bayes Additive Regression Tree, Naive Bayes, KNN, SVM, and. Here, the methods' precision, recall, accuracy, and other attributes are evaluated. Although all machine learning classifiers are efficient, this approach claims that CBART and NB classifiers have superior spam filtering abilities. According to this method, false positive calculations during spam filtering are more expensive than false negative computations.

2.3 Comparative Analysis and Summary

In the realm of quickly expanding innovation, data sharing has turned into a simple errand. There is no question that the web has formed our lives more straightforward and given us admittance to bunches of data. This is an advancement in mankind's set of experiences, and yet, it unfocuses the line between evident media and malevolently fashioned media. Today, anybody can distribute content - trustworthy or not - that can be consumed by the internet. Tragically, email spam collects a lot of consideration across the web, particularly via virtual entertainment. Individuals get bamboozled and don't think long and hard about flowing such misinformative parts of the world. This kind of information disappears, however not without inflicting the damage it was planned to cause. media locales like Facebook, Twitter, and Whatsapp assume a significant part in providing this bogus news. Numerous researchers accept that issues encompassing falsified news might be tended to through AI and man-made brainpower. Different models are utilized to give an exactness extent of 60-75%. which incorporates the Innocent Bayes classifier, semantic elements based, limited choice tree model, SVM, and others. The boundaries that are thought about don't yield high precision. The intention of this undertaking is to expand the exactness of distinguishing counterfeit news more than the current outcomes that are accessible. Manufacturing this new model, which will pass judgment on the fake news stories based on specific measures like spelling botches, confused sentences, accentuation blunders, and words utilized. The spam arrangement framework is created in this framework to recognize spam and nonspam to address the spam issue. Since spammers might send spam messages over and again, it is trying to identify it each time physically. Thusly, we will utilize a portion of the spam discovery procedures in our proposed framework. As well as distinguishing the spam term, the proposed arrangement likewise recognizes the IP address of the framework used to send the spam message. Along these lines, whenever the spam message is conveyed from a similar framework, our recommended framework will promptly remember it as spam in light of the IP address.

2.4 Scope of the Problem:

Shloka gilda introduced an idea roughly the way that NLP is pertinent to stagger on counterfeit data. They have utilized time span recurrence reverse record recurrence (TFIDF) of bi-grams and probabilistic setting free punctuation (PCFG) location. They have inspected their dataset over more than one class calculation to figure out the incredible model. They find that TF-IDF of bi-grams took care of squarely into a stochastic inclination plummet model recognizes non-believable assets with a precision of 77%.

Mykhailo Granik proposed a basic strategy for email spam text recognition: the use of gullible Bayes classifiers. They utilized BuzzFeed news for getting to be aware and giving a shot at the credulous Bayes classifier. The dataset is taken from Facebook news distribution and finished precision up to 74% on the test set.

2.5 Challenges:

The most difficult challenge for us is to collect data. We’ve no idea how it’ll happen. We took data from online which were publicly used. On the other hand, we didn’t know how to do pre-processed data, how to tokenize, or how to remove other words & punctuation. Moreover, we had no knowledge of the LSTM process. We had a little knowledge of Python but that was not enough. We practiced more and more on python, CNN, and LSTM algorithms. As it was totally new and unknown so it became a big challenge for us. We have considered the slant analysis based on voyager inputs in regard to carrier organizations in this study. Our suggested method revealed that both element determination and over-inspecting methods are equally important in improving our results. Using highlight-choosing algorithms, we were able to recover the best selection of highlights while also reducing the number of calculations required to create our classifiers. It has, however, reduced the skewed appropriation of classes observed in several of our smaller datasets without creating overfitting. Our findings show that the suggested model has a high level of grouping precision when it comes to predicting how the six classes would be structured. Managing Bengali text and processing it for model training was also a difficult challenge. As can be observed, several of the applied classifiers have outperformed the others.

CHAPTER 3

Research Methodology

3.1 Introduction

Here, I will rapidly portray the means I took to achieve our review project. At the point when a progression of news stories is introduced to the recommended framework, the new articles are named valid or misleading in view of the current information. This identification is made by taking a gander at how the words in the article are connected with each other. The recommended framework incorporates a Word2Vec model for deciding the connection among words, and the new articles are named phony or genuine news in light of the data gathered from existing connections. The spam grouping framework is created in this framework to recognize spam and non-spam to address the spam issue. Since spammers might communicate spam messages over and again, it is trying to identify it each time physically. Accordingly, we will utilize a portion of the spam discovery procedures in our proposed framework. As well as recognizing the spam term, the proposed arrangement likewise distinguishes the IP address of the framework used to send the spam message. Along these lines, whenever the spam message is conveyed from a similar framework, our proposed framework will promptly remember it as spam in view of the IP address.

3.2 Data Collection Procedure:

Table 3.2.1: Setup for my Project

Mandatory	Optional
IDS	Graphing tool
Capture Tool	Secondary Capture tool

Database	TCP Viewer
Data Miner	Database GUI

The spam arrangement framework is created in this framework to recognize spam and nonspam to address the spam issue. Since spammers might communicate spam messages over and over, it is trying to identify it each time physically. Accordingly, we will utilize a portion of the spam location methods in our proposed framework. As well as recognizing the spam term, the proposed arrangement additionally distinguishes the IP address of the framework used to send the spam message. Along these lines, whenever the spam message is conveyed from a similar framework, our proposed framework will quickly remember it as spam in view of the IP address.

3.3 Statistical Analysis

Data is gathered from a variety of sources, including newspapers and social media, and kept in datasets. Datasets will be used to feed the system. The datasets are subjected to tests.

It is preprocessed, and any extraneous information is deleted, as well as the data types of the columns if necessary. The above step makes use of a Jupyter notebook and Python libraries. In the first step, the count vectorizer approach is utilized. We must use a dataset to train the machine to recognize bogus news. Before diving into the identification of false news, there are a few things to keep in mind.

The total dataset is parted into two sections. The excess 20% is used for testing, and the leftover 80% is utilized for preparing. The calculated relapse calculation is utilized to prepare the model utilizing the preparation dataset during preparation. The test dataset is utilized as the contribution for testing, and the result is anticipated. Following the testing period, the normal and real results are thought about utilizing the disarray grid. On account of real spam text, the disarray framework gives data on the quantity of right and wrong expectations. Condition No. of Right Forecasts/Absolute Test Dataset Information Size is utilized to ascertain the precision.

3.4 Proposed Methodology:

For the coding part I took some steps:

- Data Collection
- Data Pre-processing
- Model Selection & Evaluation
- Get the best accuracy
- Result
- Testing

3.3.2 Flow Chart of my project:

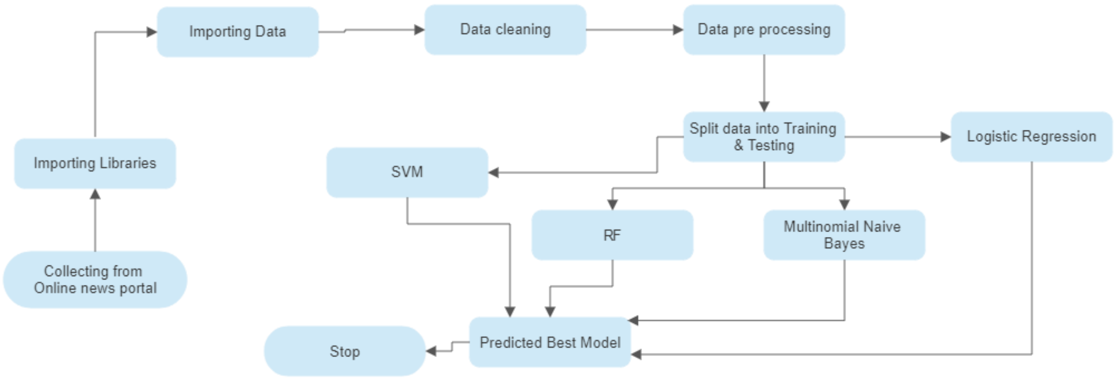
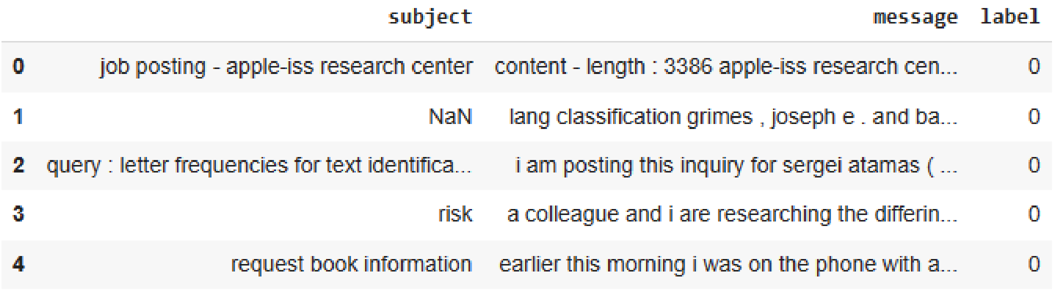


Figure 3.3.2.1: Flow chart of my project

3.3.3 Proposed Model:

In this research, we attempt to create a flexible user interface with visual concepts joint by a browser interface. Our aim is to use a machine learning model to classify master card fraud using data obtained from Kaggle as accurately as possible. Once we had done our initial research, we had a tendency to know that the Random Forest would provide the most accurate results.

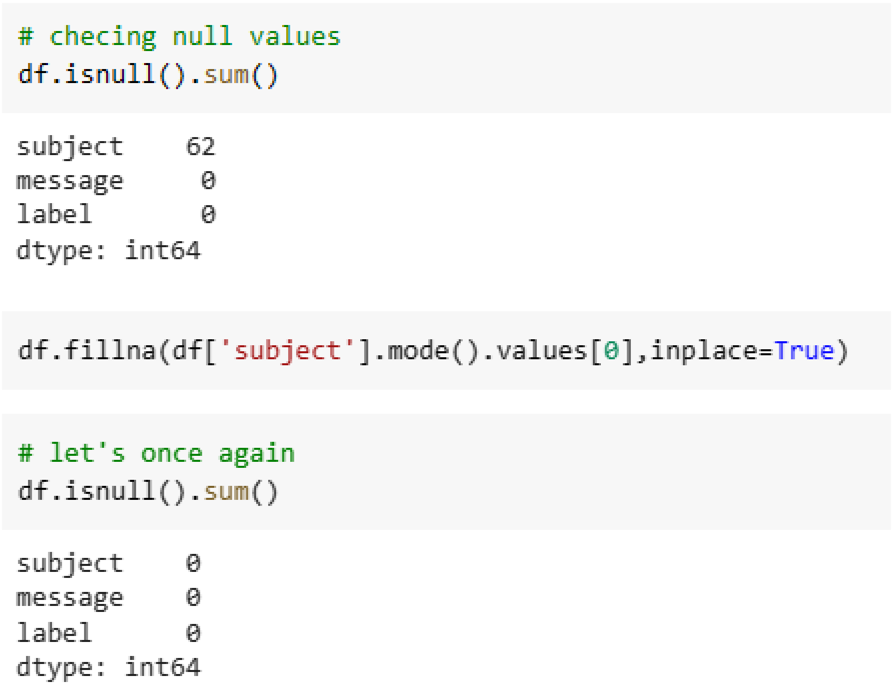
- Data Collection: I took the data from an online source that was publicly usable. Here they collect the data in a google form. They arranged 4 questions. After getting the data, they convert it into CSV format. It was very easy for me



	subject	message	label
0	job posting - apple-iss research center	content - length : 3386 apple-iss research cen...	0
1	NaN	lang classification grimes , joseph e . and ba...	0
2	query : letter frequencies for text identifica...	i am posting this inquiry for sergei atamas (...	0
3	risk	a colleague and i are researching the differin...	0
4	request book information	earlier this morning i was on the phone with a...	0

Figure 3.3.3.1: Head part of my Project

- Data Pre-processing: In this part, I cleaned the data. Missing values in the collected data could result in discrepancies. Preprocessing of the data is necessary to improve outcomes and the algorithm's efficiency. I must transform the variables and remove the outliers. To overcome these concerns, we use the chart function.



```
# checing null values
df.isnull().sum()
```

```
subject    62
message     0
label       0
dtype: int64
```

```
df.fillna(df['subject'].mode().values[0],inplace=True)
```

```
# let's once again
df.isnull().sum()
```

```
subject    0
message    0
label      0
dtype: int64
```

Figure 3.3.3.2: Null Values

3.5 Implementation Requirements

A subtype of artificial intelligence called machine learning teaches machines to think and act like humans without being explicitly taught. We employ supervised techniques in this paper. For the prediction of Android applications, five machine-learning classification models have been applied. The models can be found in free source Python software. Below are brief descriptions of each model.

- **Naive Bayes [13]:** The Naive Bayes technique is typically employed when a huge dataset needs to be predicted. Conditional Probability is utilized. The probability of event A happening given that an earlier event B has already happened is known as conditional probability. The most typical application of this algorithm is the screening of spam emails in your email account. For instance, you recently received new mail. The model employs the Naive Bayes method to predict whether or not the mail received is spam by looking through your previous spam mail records [13].

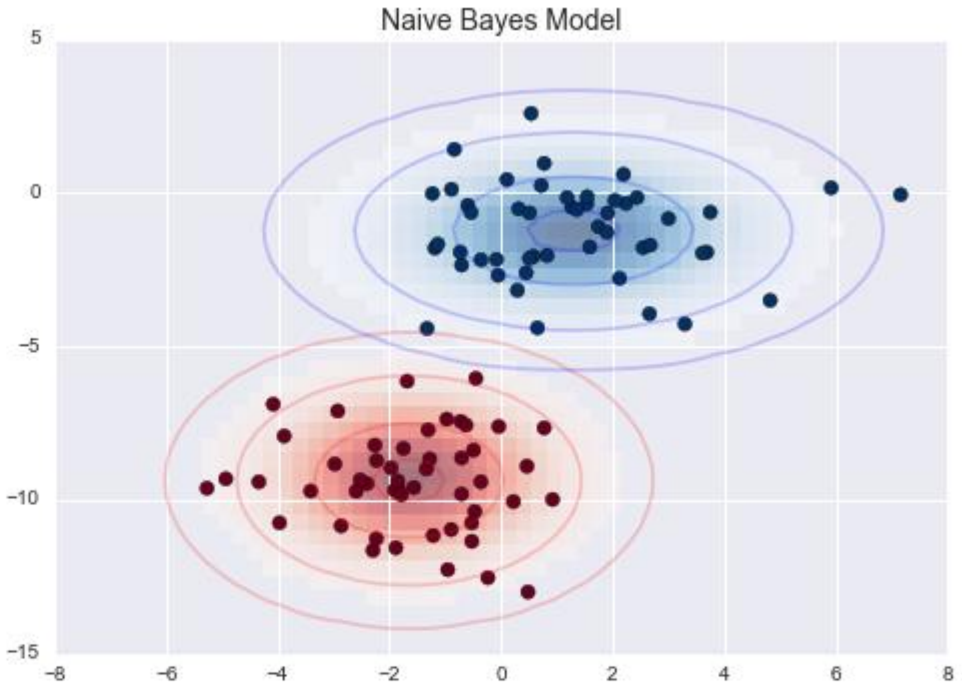


Figure 3.4.1: Naive Bayes [14]

- **SVM [15]:** Support Vector Machine is a Supervised Machine Learning algorithm that is used for regression and/or classification. Although it is occasionally quite helpful for regression, classification is where it is most often used. In essence, SVM identifies a hyper-plane that establishes a distinction between the various types of data [15]. This hyper-plane is just a line in two-dimensional space. Each dataset item is plotted in an N-dimensional space using SVM, where N is the total number of features and attributes in the dataset. The best hyperplane should then be found to divide the data. You must have realized by now that SVM can only perform binary classification by nature. For multi-class problems, there are numerous techniques to use [15].

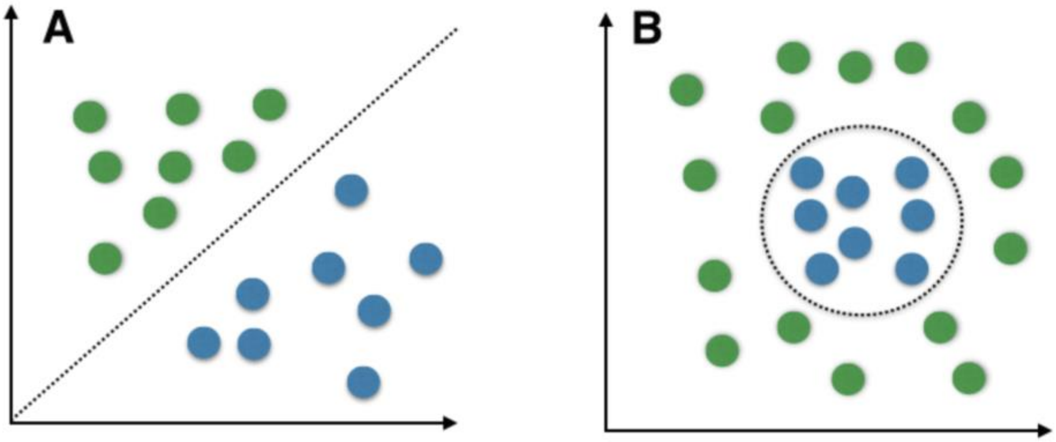


Figure 3.4.2: SVM[16]

- **Random Forest [17]:** The bagging method is extended by the random forest algorithm, which uses feature randomness in addition to bagging to produce an uncorrelated forest of decision trees. The random subspace method, also known as feature bagging, creates a random subset of features that guarantees a low correlation between decision trees. The main distinction between decision trees and random forests is this. Random forests merely choose a portion of those feature splits, whereas decision trees take into account all possible feature splits [17].

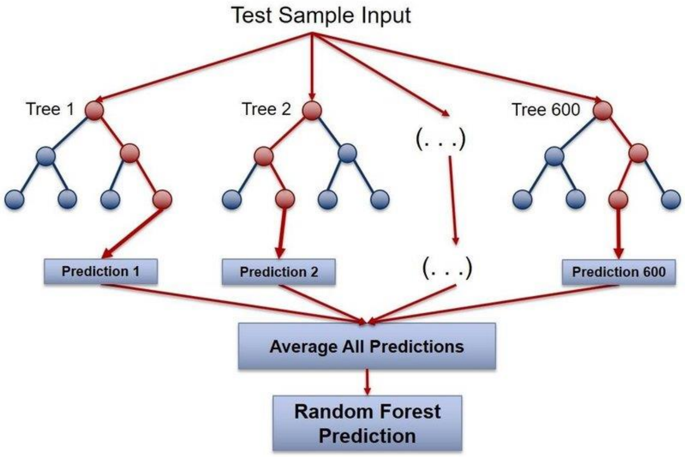


Figure 3.4.3: Random Forest[18]

- **Logistic Regression:** The logistic function, also versed as the sigmoid function, was created by statisticians to characterize the properties of population expansion in ecology, such as how quickly it grows and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1, but never exactly within those two limitations.

$$\text{sigmoid } (Z) = 1 / (1 + e\text{^}{-z})$$

$$\text{Hypothesis} \Rightarrow Z = WX + B$$

$$h\Theta(x) = \text{sigmoid } (Z)$$

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

I used a Colab notebook for my coding part. My useable language was python. For getting accuracy I uploaded some libraries. This project may be run on standard computer hardware. We conducted an Intel I5 processor with 8 GB of RAM and a 2 GB Nvidia graphics processor. It also has two cores that run at 1.7 GHz and 2.1 GHz, respectively. The first half of the process is training, which takes about 10-15 minutes, and the second part is testing, which takes only a few seconds to make seven predictions and calculate accuracy.

4.2 Experimental Results & Analysis

The model has to be tested after it has been trained. The model is evaluated using the data that we divided during the test-trained module. Confusion metrics, precision, recall, accuracy, and F1 score techniques are mostly used in utilized to assess the classification issue.

4.2.1.1 True Positive:

Figure 4.2.1.1.1: True Positive

Algorithm	TP
Naive Bayes	11
Logistic Regression	40
Random Forest	43
KNN	46
Gradient Boosting Classifier	40

4.2.1.2 False Positive:

Figure 4.2.1.2.1: False Positive

Algorithm	FP
Naive Bayes	37
Logistic Regression	8
Random Forest	5
KNN	2
Gradient Boosting Classifier	8

4.2.1.3 False Negative:

Figure 4.2.1.3.1: False Negative

Algorithm	FN
Naive Bayes	0
Logistic Regression	0
Random Forest	0
KNN	11
Gradient Boosting Classifier	1

4.2.1.4 True Negative:

Figure 4.2.1.4.1: True Negative

Algorithm	TN
Naive Bayes	242
Logistic Regression	242
Random Forest	242
Gradient Boosting Classifier	232
KNN	231

4.2.2 Accuracy:

Table 4.2.2.1: Accuracy

Algorithm	Accuracy(%)
Naive Bayes	87
Logistic Regression	96
Random Forest	98
KNN	93
Gradient Boosting Classifier	96

4.2.3 Recall:

Table 4.2.3.1: Recall

Algorithm	Recall
Naive Bayes	0.97
Logistic Regression	0.97
Random Forest	0.98
Gradient Boosting Classifier	0.96
KNN	0.96

4.2.4 Precision

Figure 4.2.4.1: Precision

Algorithm	Precision
Naive Bayes	0.87
Logistic Regression	0.98
Random Forest	0.98
Gradient Boosting Classifier	0.97
KNN	0.97

4.3 Result Discussion

With the help of the Random Forest I got the best accuracy which was 98%. With certainty, it can be said that the Random Forest model is quite effective and produces better results than other models. Data is gathered from a variety of sources, including newspapers and social media, and kept in datasets. Datasets will be used to feed the system. The datasets are subjected to tests.

It is preprocessed, and any extraneous information is deleted, as well as the data types of the columns if necessary. The above step makes use of a Jupyter notebook and Python libraries. In the first step, the count vectorizer approach is utilized. We must use a dataset to train the machine to recognize bogus news. Before diving into the identification of spam text there are a few things to keep in mind.

The complete dataset is split into two parts. The remaining 20% is utilized for testing, and the remaining 80% is used for training. The KNN, Gradient Boosting Classifier, RF, Logistic Regression, Naïve Bayes are used to train the model using the training dataset during training. The test dataset is used as the input for testing, and the outcome is predicted. Following the testing period, the expected and actual outputs are compared using the confusion matrix. In the case of actual and fake news, the confusion matrix provides information on the number of correct and incorrect predictions. The equation No. of Correct Predictions/Total Test Dataset Input Size is used to calculate the exactness.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Every human feeling may be linked to the words we view on a daily basis on various online platforms in the digital world. In this case, it is critical for these platforms to have a mechanism in place to discern which are genuine emotions and which are pre-programmed aggressiveness. This is why I've decided to focus on one of the most fascinating genres of all time,by doing so, we can expect to create a more definitive and diverse digital era.

5.2 Impact on Environment

Due to the complexity of the network system of openness, sharing of resources, system, linking the variety, the uneven distribution of the terminal, network agnostic, and other barriers, computer networks continue to exhibit their distinctive benefits. Computer's cause. The biggest issue is security, which is one of the numerous issues brought on by the network. Data is gathered from a variety of sources, including newspapers and social media, and kept in datasets. Datasets will be used to feed the system. The datasets are subjected to tests. Since the beginning of email, there has always been email spam. Spam is by definition an undesired or pointless message. We refer to it as junk mail. Spam can include a variety of material and topic matter, including promotions, adverts, requests for miracle cures, pornography, and a wide range of other things.

The majority of email traffic that businesses get, according to the email security experts at Gatefy, is spam.We're talking about billions of spam emails being sent and received everyday if we consider that, according to figures from the research firm Radicati Group, more than 360 billion emails will be issued and received daily by 2024.Spam is problematic since it significantly hinders a company's ability to

It is preprocessed, and any extraneous information is deleted, as well as the data types of the columns if necessary. The above step makes use of a Jupyter notebook and Python libraries. In the first step, the count vectorizer approach is utilized. We must use a dataset to train the machine to recognize bogus news. Before diving into the identification of false news, there are a few things to keep in mind.Everybody thinks it is a normal issue. But it is not. So that’s why I decided to work on it.

5.3 Ethical Aspects

Loans account for a large portion of bank profits. Despite the fact that many people are looking for loans. Finding a legitimate applicant who will return the loan is difficult. Choosing a real applicant may be difficult if the process is done manually. As a result, we are creating a machine learning-based spam detection system that will choose the qualified applicants on its own. Both the applicant and the bank staff will benefit from this. There will be a significant reduction in the loan sanctioning period of time. In this research.

5.4 Sustainability

- There are over 2.3 billion active internet-based life clients worldwide.
- At least two internet-based life cycles are present in 91 percent of large business brands.
- When they can't access their online life profiles, 65 percent of individuals feel uneasy and uncomfortable.
- It will be a helping hand for the researcher.
- Able to gain more knowledge about spam text detection methods.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication For Future Research

6.1 Summary of the Study

The purpose of this study was How can we unearth the Spam text. That means whether the text is fake or real. This work implements function extraction and data processing for customer basic attribute data and downloads transaction data based on the scenario of a bank credit application. Then, to increase the accuracy of bankruptcy assessment and achieve local optimization, a linear regression model with the penalty and a neural network prediction model are presented. By doing this, the implicit risk detection is control. The system is a Web application that assists users in identifying bogus news. We've provided a text box where the user may paste the message or the URL link to the news or another message, and it will then display the truth about it. All data provided by the user to the detector may be saved for future usage in order to update the model's state and conduct data analysis. We also assist users by providing instructions on how to avoid such bogus events and how to stop them from spreading.

To raise the level of risk management for banks, the most suitable penalty linear regression prophecy algorithm is chosen based on the characteristics of the sample data that was collected.

6.2 Conclusion

This essay explores many methods for categorizing spam and junk email.

Diverse machine and deep learning classifiers are conducted in experiments. Results indicate that Bi LSTM has an F1-Measure 0 of 96% and a maximum accuracy of 98.5%. Future versions of the current work should be improved by expanding it to a number of industries, including e-commerce, employment profile-based websites, and other locations where fake news is common, as well as by developing an app that enables users to quickly identify false information using their smartphone. The job can also be prolonged by using a real-time classifier.

This survey study describes many existing spam filtering systems using machine learning techniques by investigating various approaches, evaluating the effectiveness of various proposed approaches with reference to various parameters, and concluding the overview of various spam filtering approaches. Additionally, all currently used techniques for email spam filtering are efficient. Some have successful results, while others are attempting to use a different technique to improve their accuracy rate. Researchers are working to develop the next generation of spam filtering systems, which will be able to take into account a huge amount of multimedia data and filter spam email more effectively, even if all of the current spam filtering systems are effective.

6.3 Implication for Further Study

- It will be a contribution.
- Easier.
- More flexible.
- User-friendly.

REFERENCES

[1] I. AbdulNabi and Q. Yaseen, “Spam email detection using deep learning techniques,” in Procedia Computer Science, 2021, vol. 184, pp. 853–858. doi: 10.1016/j.procs.2021.03.107. Electronic copy available at: https://ssrn.com/abstract=4145123

[2] S. Madisetty and M. S. Desarkar, “A Neural Network-Based Ensemble Approach for Spam Detection in Twitter,” IEEE Trans. Comput. Soc. Syst., vol. 5, no. 4, pp. 973–984, Dec. 2018, doi: 10.1109/TCSS.2018.2878852.

[3] A. Salunkhe, “Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification,” Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.14789

[4] A. Hosseinalipour and R. Ghanbarzadeh, “A novel approach for spam detection using horse herd optimization algorithm,” Neural Comput. Appl., Mar. 2022, doi: 10.1007/s00521-022- 07148-x.

[5] P. K. Roy, J. P. Singh, and S. Banerjee, “Deep learning to filter SMS Spam,” Futur. Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020, doi: 10.1016/j.future.2019.09.001.

[6] K Agarwal abd T Kumar, "Email Spam Detection using Integrated approach of Naïve Bayes and Particle Spam Optimation, "IEEE Electron Devices Society, Institute of Electrical and Electronics Engineers, and Vaigai College of Engineering, Proceeding of the 2018 International Conference on Intelligent Computing and Control Systems (ICICCS) : June 14-15, 2018.

[7] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, “Efficient email classification approach based on semantic methods,” Ain Shams Eng. J., vol. 9, no. 4, pp. 3259–3269, Dec. 2018, doi: 10.1016/j.asej.2018.06.001.

[8] W. Pan et al., “Semantic Graph Neural Network: A Conversion from Spam Email Classification to Graph Classification,” Sci. Program., vol. 2022, 2022, doi: 10.1155/2022/6737080.

[9] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, “Content based fake news detection using knowledge graphs,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11136 LNCS, pp. 669– 683. doi: 10.1007/978-3-030-00671-6_39.

[10] [41] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Technologies, 31(2), e3767.

[11] B. K. Dedeturk and B. Akay, “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm,” Appl. Soft Comput. J., vol. 91, Jun. 2020, doi: 10.1016/j.asoc.2020.106229.

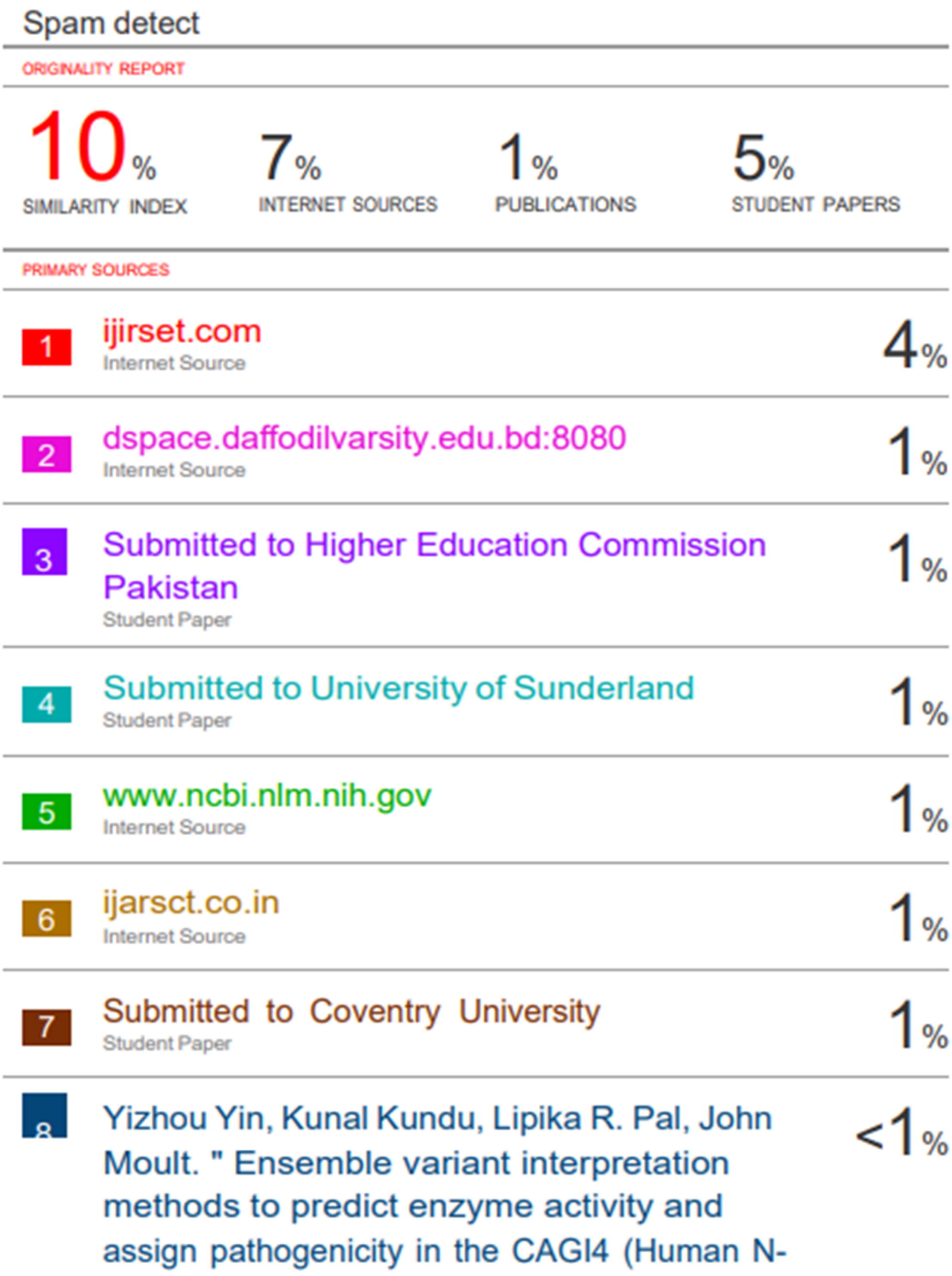
[12] W. Feng, “2016 IEEE 35th International Performance Computing and Communications Conference, IPCCC 2016,” 2016 IEEE 35th Int. Perform. Comput. Commun. Conf. IPCCC 2016, 2017.

[13] Amanoul, S. V., Abdulazeez, A. M., Zeebare, D. Q., & Ahmed, F. Y. (2021, June). Intrusion Detection Systems Based on Machine Learning Algorithms. In 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS) (pp. 282-287). IEEE.

[14] T. Verma and D. C. Rana, “Data Mining Techniques for the Knowledge Discovery,” Int. J. Eng. Technol., vol. 9, no. 3S, pp. 351–354, Jul. 2017, doi: 10.21817/ijet/2017/v9i3/170903s054.

[15] D. Tang, B. Qin, and T. Liu, “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification,” Association for Computational Linguistics, 2015. [Online]. Available: http://ir.hit.edu.cn/

[16] A. Ishaq et al., “Extensive hotel reviews classification using long short term memory,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 10, pp. 9375–9385, Oct. 2021, doi: 10.1007/s12652- 020-02654-z. Electronic copy available at: https://ss-15



Spam detect

ORIGINALITY REPORT

10%	7%	1%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijirset.com Internet Source	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Submitted to Higher Education Commission Pakistan Student Paper	1%
4	Submitted to University of Sunderland Student Paper	1%
5	www.ncbi.nlm.nih.gov Internet Source	1%
6	ijarsct.co.in Internet Source	1%
7	Submitted to Coventry University Student Paper	1%
8	Yizhou Yin, Kunal Kundu, Lipika R. Pal, John Moult. " Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 (Human N-	<1%