COMPREHENSIVE STUDY OF SMS SPAM DETECTION STRATEGIES BASED ON TCP PROTOCOLS

BY

MOSHREKUL ISLAM
ID: 193-15-3016

AND

ABDUR RAHMAN ALVI
ID: 193-15-3002

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

Supervised By

MR. AMIT CHAKRABORTY
Assistant Professor
Department of Computer Science and
EngineeringDaffodil International University

Co-Supervised By

MS. REFATH ARA HOSSAIN
Lecturer
Department of Computer Science and
EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2024

APPROVAL

This Project titled "**COMPREHENSIVE STUDY OF SMS SPAM DETECTION STRATEGIES BASED ON TCP PROTOCOLS**", submitted by **Md. Moshrekul Islam, ID: 193-15-3016** and **Abdur Rahman Alvi, ID: 193-15-3002** 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 **15 July, 2024**.

BOARD OF EXAMINERS

Taimur Ahad

Dr. Md. Taimur Ahad (MTA) — **Board Chairman**
Associate Professor & Associate Head
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

Shayla Sharmin

Ms. Shayla Sharmin (SS) — **Internal Examiner 1**
Senior Lecturer
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Mayen Uddin Mojumdar (MUM) — **Internal Examiner 2**
Senior Lecturer
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

M Islam

Dr. Md. Manowarul Islam (MI) — **External Examiner**
Associate Professor
Department of Computer Science & Engineering
Jagannath University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Amit Chakraborty, Assistant Professor, 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:

09.07.24

Mr. Amit Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Ms. Refath Ara Hossain
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Moshrekul Islam

Md. Moshrekul Islam
ID: 193-15-3016
Department of CSE
Daffodil International University

Abdur Rahman Alvi

Abdur Rahman Alvi
ID: 193-15-3002
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Mr. Amit Chakraborty**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in this field to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor **Dr. Sheak Rashed Haider Noori** and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member 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 discuss while completing the course work.

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

ABSTRACT

This study aims to enhance the accuracy of SMS spam detection, addressing the rising threat of unsolicited and potentially harmful messages in mobile communication. By leveraging both machine learning and deep learning techniques, we seek to identify the most effective model for distinguishing between spam and legitimate SMS messages. It uses a dataset consist of total 5572 number of data among 4825 non-spam and 747 was spam message. The dataset used consists of labeled SMS messages, which underwent preprocessing steps including cleaning, tokenization, and lowercasing. We implemented and evaluated several models: Logistic Regression, Multinomial Naive Bayes, Simple RNN, LSTM, and GRU. Among these, Logistic Regression outperformed all other models, achieving the highest accuracy. For machine learning, Multinomial Naive Bayes and Logistic Regression. Multinomial Naive Bayes has been trained in two ways. TF-IDF and CV (Count Vectorization). Among them, CV based MultinomialNB performed better than TF-IDF based MultinomialNB which is 97.68% compared to 95.85%. In deep learning, we choose RNN based models such as LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit) and SimpleRNN in order to achieve better accuracy. These deep learning models achieve accuracy of 97.78%, 97.78% and 97.97% respectively. Later Logistic Regression, another machine learning involves in this study outperformed all the models with an accuracy of 99.84% with precision of 99.69%. The findings highlight the robustness and efficiency of traditional ML models in handling the SMS spam detection task, while also providing insights into the performance of advanced DL models. This research contributes to the development of reliable and scalable spam detection solutions, enhancing user safety and the overall security of mobile communication systems.

Keywords*: Spam, Logistic Regression, Deep Learning, Recurrent Neural Network, Multinomial Naive Bayes, Machine learning, LSTM.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Spamming is the use of messaging systems to send multiple unsolicited messages to large numbers of recipients for the purpose of commercial advertising, non-commercial proselytizing, or any prohibited purpose, or simply repeatedly sending the same message to the same user. Spammers use many forms of communication to bulk-send their unwanted messages. Some of these are marketing messages peddling unsolicited goods. Other types of spam messages can spread malware, trick people into divulging personal information, or scare people into thinking they need to pay to get out of trouble. According to the experts, the annual overall loss resulting from spam is estimated to be tens of billions of Dollars. As a result, anti-spam protection is not only desirable, but an urgent necessity. If spammer activity is not restricted, email could easily become a thing of the past, eclipsed by the overwhelming volume of spam.

This study proposed various machine learning and deep learning approach to develop a more efficient model that can classify spam messages. The spam category selected for the study is SMS spam. It uses a dataset consist of total 5572 number of data among 4825 non-spam and 747 was spam message. Using this dataset, the study presents an approach that includes key components such as data preparation, model selection, and thorough model evaluation, aiming to create a robust AI model in detecting SMS spam messages.

For SMS spam detection, this study focusses on both machine learning and deep learning approach. For machine learning, we selected Multinomial Nave Bayes and Logistic Regression. Also three deep neural network architectures such as SimpleRNN, LSTM (long short term memory) and GRU (gated recurrent unit). The main objective of this project is to build and evaluate a text classification models capable of distinguishing between legitimate messages and spam messages. By training on a labeled dataset of SMS messages, the model learns patterns and characteristics of spam content, allowing it to make accurate predictions on new, unseen messages.

1.2 Motivation

The motivation behind this study stems from the growing prevalence of unsolicited and potentially harmful SMS messages, which pose significant risks to users by enabling phishing attacks, spreading malware, and compromising personal information. As mobile communication continues to be a vital part of daily life, effective spam detection has become increasingly important for ensuring user safety and maintaining trust in mobile networks. This study seeks to leverage both machine learning (ML) and deep learning (DL) techniques to identify the most accurate and efficient model for distinguishing spam from legitimate messages. By improving spam detection accuracy, the study aims to enhance user experience, reduce the incidence of fraudulent activities, and contribute to the overall security and reliability of mobile communication systems. Additionally, the findings of this research can provide valuable insights for developing robust, scalable, and real-time spam detection solutions that can be implemented in various real-world applications.

1.3 Research Question

i. How does the performance of machine learning and deep Learning models compare in detecting SMS spam?
ii. What features and preprocessing techniques contribute most significantly to improving SMS spam detection accuracy using machine learning and deep learning models?

1.4 Expected Outcome

The expected outcome of this study is to ensure an AI model that can detect spam message more accurately and achieve a result higher than any previous work. Figuring out the most effective model for this dataset through a comparative analysis of both machine learning and deep learning techniques. Also identifying potential future research avenues, such as exploring advanced techniques like ensemble methods, transfer learning, and hybrid models.

1.5 Report Layout

The readers will get a brief synopsis of the contents of the chapters from this report layout. There are five chapters in this thesis paper. Each one has a distinct thesis related theme. Let us get started now.

Chapter 1: This thesis's introduction provides background information on the research topic, justifies the investigation, walks the reader through the basic concepts of cloud technology, highlights importance of a comparative study, motivation behind the study, delineates particular research goals, and predicts the study's expected outcomes. Setting the scene, clarifying applicability, and outlining the study's goals—all of which will be further developed in later chapters—are crucial tasks for this chapter.

Chapter 2: This thesis's current chapter presents a thorough research roadmap that includes important components like case study selection, the integration of various research methodologies, procedural methodologies in detail, an extensive analysis of the current system, potential threats identification, the presentation of the proposed system, and a detailed explanation of the applied working methods. The chapter seeks to give readers a thorough grasp of the research process and related procedures by establishing this methodological framework, which lays the groundwork for later portions of the thesis.

Chapter 3: This chapter provides a thorough examination of how the research has been conducted and methodology behind the study. The chapter explains the process and framework on which the research has been conducted.

Chapter 4: In this chapter, we detail the experimental setup and results of our study on SMS spam detection. The dataset, comprising labeled SMS messages categorized as spam or ham, underwent preprocessing steps including cleaning, tokenization, and lowercasing. We implemented and evaluated several models: Logistic Regression, Multinomial Naive Bayes, Simple RNN, LSTM, and GRU. Logistic Regression outperformed all other models, achieving the highest accuracy. This chapter also discusses the methodologies and requirements used in conducting the study.

Chapter 5: In this chapter we have discuss the summary of the entire study. Conclusions and implication for future work etc.

CHAPTER 2

BACKGROUND STUDY

2.1 Related Works

SMS spam detection is crucial for ensuring text messaging systems' security and usability. Spam messages, also known as unwanted text messages, can be a nuisance and be used for phishing scams and other malicious activities. In this research, we analyze various SMS spam detection methods and their efficiency in recognizing spam messages. Three machine learning models- LSTM, a DNN, and a Bi-LSTM model have been used to examine the data. We found that our proposed methods are highly effective at detecting spam messages with high accuracy through our experiments on a dataset of SMS text messages. Of the three models, the DNN has produced the highest accuracy of 95.6522%, followed by Bi-LSTM with an accuracy of 93.9799% and LSTM with an accuracy of 92.9766% [1].

One of the most accessible ways to communicate via text is through a short message service. In recent years, profit-seeking people have taken advantage of the good features of this service to send large numbers of spam messages to random people for malicious purposes. In this respect, detecting spam messages is an important task. The unbalanced proportion of the spam and ham data and the extraction of efficient features from short messages have been the main challenges in the SMS spam detection problem. So far, various methods have been proposed to filter spam messages, whose accuracy still needs to be improved. In this study, we propose an ensemble learning method based on random forest and logistic regression algorithms to increase the accuracy of SMS spam detection. The proposed approach has been tested on two real datasets. The experimental evaluation based on accuracy and AUC shows the effectiveness of the proposed ensemble learning algorithm [2].

For performing SMS Spam Detection incorporating Support Vector Machine as well as logistic regression (LR) is done here. Depending on accuracy, analysis for SMS spam detection dataset associated with 5573 sentences. Classification of SMS spam detection is performed by a support vector machine whose sample takes (N=27) as well as Linear Regression whose sample takes (N=27), G-power generation is performed whose value takes 80%. SVM accuracy is 97.67%

whereas LR takes accuracy to be 97.02%. Significant value of accuracy is 0.02 ($p<0.05$). SVM performs better in tracking down precision when contrasted with Logistic Regression [3].

SMS spam has seen an exponential increase in the past few years. This is a pain for mobile users, since any important message often gets lost in the pile of spam messages. SMS spam detection is at a very nascent stage when it comes to implementing in user devices, although there has been research in the field for a long time. We have carried out a survey of the work that has been done in the field of SMS spam filtering, and we have identified certain features which give high accuracy when classifying SMS spam [4].

SMS spam is a pervasive issue that affects millions worldwide, leading to significant inconvenience, time wastage, and potential financial scams. Given the prevalence and potential harm, accurate and real-time detection of SMS spam is crucial. This paper proposes a novel approach to SMS spam detection involving five steps: preprocessing, feature extraction, feature fusion, feature selection, and classification. Our model is designed to simultaneously capture local, temporal, and global text message features using a hybrid deep learning model to enhance feature representation. We evaluated our model using the UCI dataset, comparing it with traditional and deep learning algorithms such as RF and BERT using cross-validation to ensure the robustness of our results. Our proposed method exhibited superior performance, achieving a good accuracy of 99.56%, surpassing other methods. The effectiveness of this method in SMS spam detection proved its potential for real-world implementation, where it could substantially mitigate the prevalence and impact of SMS spam [5].

SMS spam detection using Naive Bayes algorithm is a widely used technique in the field of text classification. The main aim of this approach is to classify the incoming messages into spam or ham categories. The Naive Bayes algorithm works by calculating the probability of a message belonging to a particular class, based on the occurrence of different words in the message. In this paper, we present an efficient and accurate approach for SMS spam detection using the Naive Bayes algorithm. The proposed approach utilizes a pre- processing step for feature extraction, which includes tokenization, stop-word removal, and stemming. The Naive Bayes algorithm is then trained on a dataset of labeled messages to learn the probability distributions of different words in spam and ham messages. Finally, the trained model is used to classify incoming messages into spam or ham categories. The results of our experiments show that the proposed approach achieves high accuracy in detecting SMS spam messages. Keywords: Naive Bayes Algorithm, Text Classification, Probability, Feature Extraction, Tokenization, StopWord Removal, Stemming, Labeled Dataset, Probability Distribution, Accuracy [6].

The aim of the study is to detect SMS spam using Support Vector Machine (SVM) and linear regression (LR). The dataset used in the study contains 5573 sentences, and accuracy is measured for SMS spam detection. The classification process is carried out using SVM and LR with sample sizes of N=27, which were obtained using a G-power value of 80%. The accuracy of SVM is found to be 97.67%, which is higher than LR with an accuracy of 92%. The p-value for the significant accuracy difference is 0.02 ($p<0.05$), indicating that SVM performs better than LR in achieving accuracy [7].

The term SMS (Short Message Service) refers to a popular text messaging service that is commonly used in telephone, internet, and mobile device systems. This service relies on standardized communication protocols that enable short text messages to be exchanged between mobile devices. The increase in SMS spam messages can be attributed to the higher limit of free SMS allowed by Internet Service Providers (ISPs) SMS spam detection relies heavily on the presence of known words, phrases, abbreviations, and idioms commonly used in spam messages. Studies have developed various datasets to train and test SMS spam detection models and have used different classification techniques to improve the accuracy and efficiency of these models. In the present study, various classification techniques for SMS spam detection have been explored such as Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forest, and Neural Networks. These techniques use different approaches to identify patterns and features in the messages that distinguish spam from legitimate messages. Among the various algorithms Naïve Bayes Classifier achieved a highest accuracy of 98.44% and Matthew Correlation Coefficients value of 0.93 for the dataset [8]. In recent times, there has been considerable interest among people to use short communication service (SMS) as one of the essential and straightforward dispatches services on mobile bias. The increased popularity of this service also increased the number of mobile bias attacks similar as SMS spam dispatches. SMS spam dispatches constitute a real problem to mobile subscribers. this worries telecommunication service providers as it disturbs their guests and causes them to lose business. to enhanced SMS spam filtering performance by combining two of data mining task association and bracket. FP- growth in association is employed for mining frequent pattern on SMS and Naive Bayes Classifier is used to classify whether SMS is spam or ham. Training data was using SMS spam collection from former exploration. The result of using collaboration of Naive Bayes and FP- Growth performs the highest average accuracy of 98,506 and 0.025% better than without using FP- Growth for dataset SMS Spam Collection v. 1, and improves the perfection score; therefore, the bracket result is more accurate [9].

With the growth of the use mobile phones, people have become increasingly interested in using Short Message Services (SMS) as the most suitable communications service. The popularity of SMS has also given rise to SMS spam, which refers to any unwanted message sent to a mobile phone as a text. Spam may cause many problems; such as traffic bottlenecks or stealing important users' information. This paper, presents a new model that extracts seven features from each message before applying a Multiple Linear Regression (MLR) to assign a weight to each of the extracted features. The message features are fed into the Extreme Learning Machine (ELM) to determine whether they are spam or ham. To evaluate the proposed model, the UCI benchmark dataset was used. The proposed model produced recall, precision, F-measure, and accuracy values of 98.7%, 93.3%, 95.9%, and 98.2%, respectively [10].

When it comes to SMS, the topic of spam is a big obstacle that requires urgent attention. The majority of SMS messages that fall under the category of spam are commercial communications, which are obviously illegal and unethical, followed by phishing scams and messages that may disturb the user's peace of mind. Users may find this to be extremely inconvenient as it consumes energy to erase those messages, slows down websites, and may even infect computers with viruses. We must therefore distinguish between SMS messages that are spam and those that are not in order to prevent this. We are recognizing this in this project utilizing machine learning algorithms. In this paper, we will talk about the algorithms we employ, compare them all to the dataset we use, and finally select the algorithm that will be most effective at determining which SMS messages fall under the spam category. Based on comparisons of precision and accuracy, we choose the optimum algorithm, then for the results of the research in our paper, we found that the best non-pretrained machine learning model was GRU with the best feature extractor being GloVe, with an accuracy of 91.9171% and precision of 91.9187%. However, the pretrained machine learning model BERT still outperformed the non-pretrained model with an accuracy of 99.0166% and precision of 99.0179% [11].

E-mail and SMS are the most popular communication tools used by businesses, organizations and educational institutions. Every day, people receive hundreds of messages which could be either spam or ham. Spam is any form of unsolicited, unwanted digital communication, usually sent out in bulk. Spam emails and SMS waste resources by unnecessarily flooding network lines and consuming storage space. Therefore, it is important to develop high accuracy spam detection models to effectively block spam messages, so as to optimize resources and protect users. Various word-embedding techniques such as Bag of Words (BOW), N-grams, TF-IDF, Word2Vec and Doc2Vec have been widely applied to NLP problems, however a comparative study of these techniques for SMS spam detection is currently lacking. Hence, in this paper, we provide a comparative analysis of these popular word embedding techniques for SMS spam detection by evaluating their performance on a publicly available ham and spam dataset. We investigate the performance of the word embedding techniques using 5 different machine learning classifiers i.e. Multinomial Naive Bayes (MNB), KNN, SVM, Random Forest and Extra Trees. Based on the dataset employed in the study, N-gram, BOW and TF-IDF with oversampling recorded the highest F1 scores of 0.99 for ham and 0.94 for spam [12].

The number of people using mobile devices increasing day by day. SMS (short message service) is a text message service available in smartphones as well as basic phones. So, the traffic of SMS increased drastically. The spam messages also increased. The spammers try to send spam messages for their financial or business benefits like market growth, lottery ticket information, credit card information, etc. So, spam classification has special attention. In this paper, we applied various machine learning and deep learning techniques for SMS spam detection. we used a dataset from UCI and build a spam detection model. Our experimental results have shown that our LSTM model outperforms previous models in spam detection with an accuracy of 98.5%. We used python for all implementations [13].

In recent times, the increment of mobile phone usage has resulted in a huge number of spam messages. Spammers continuously apply more and more new tricks that cause managing or preventing spam messages a challenging task. The aim of this study is to detect spam message to prevent different cybercrimes as spam messages have become a security threat nowadays. In this paper, studies on SMS spam problems to perform a better accuracy using several different techniques such as Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Random Forest, Logistic Regression and some more are performed. The result indicated that Support Vector

Machine achieved the highest accuracy of 99%, indicating it might be useful as an effective machine learning system for future research [14].

Social-media is a very common medium for spammers to unethically overwhelm normal users with unsolicited or false content via social-networking. Now a day most of the people are aware of social media. In this paper, we have used SMS Spam Collection dataset that is taken from Kaggle. The Bow with TF and TF-IDF weighing schemes features are used for feature selection. And we used chi-square matrix for features selection. We have done the comparison of state of the art and proposed model. The results show that our proposed model Multinomial Naïve Bayes gave highest accuracy. We offer the variable status of each feature so that it is easy to abolish the inappropriate features. Our results illustrated that our proposed model accomplished effectively high detection rates in terms of high accuracy, compared with other considered researches [15].

The ubiquity of Short Messaging Services (SMS) has been accompanied by an upsurge in spam messages, necessitating effective detection mechanisms. This paper delves into the intricate task of SMS spam detection using a Kaggle dataset of 5570 samples. As initial step, the challenge of class imbalance was addressed by applying oversampling technique ADASYN. Later, TF-IDF embedding technique was employed to capture semantic meaning and contextual information within SMS messages. A conventional ML classifier, Random Forest has achieved a good accuracy of 92.3%. Later, various RNN architectures, including Simple RNN, LSTM, Bi-LSTM and Gated Recurrent Unit (GRU), applied for SMS spam detection. LSTM variants demonstrated superior performance with 93% accuracy for RNN, 98% for LSTM and 98.2%, 98.3% for BI-LSTM and GRU. To further enhance the accuracy, a hybrid architecture combining Bi-LSTM and GRU layers was explored, resulting in improved accuracy of 99%. The proposed model outperformed traditional ML approaches for SMS spam detection. This paper sheds the efficacy of LSTM variants in addressing the SMS spam detection challenge. Emphasizing the significance of handling class imbalances and employing effective embedding techniques, the achieved accuracies hold promising implications [16].

The short messaging service (SMS) is one of the most popular and also most affordable telecommunication services. The popularity and affordability of SMS, however, have made it an ideal target for spamming. Spam is a major nuisance to mobile subscribers, but can also lead to security breaches or criminal activities. In this paper, we propose a novel lightweight deep neural model called Lightweight Gated Recurrent Unit (LGRU) for SMS spam detection. In addition, we incorporate enhancing semantics retrieved from external knowledge (WordNet) to augment the understanding of SMS text inputs for better classification. We compare the performance of our proposed model with that of more than 30 SMS spam classifiers that use various conventional machine learning and deep learning techniques. Experimental results show that our model outperforms these existing classifiers in terms of precision, recall and accuracy. In addition, our model requires fewer training parameters and incurs significantly less training time than state-of-the-art deep learning based classifiers [17].

2.2 Comparative Analysis

Prior researchers have analyzed numerous models to detect different diseases. The researchers have performed an investigation on multiple models and algorithms to evaluate their performance and choose the most reliable ones.

Table 2.2 Comparative Analysis of Previous Research Works.

Research Paper	**Models Used**	**Best Model**	**Accuracy**
Gandhi et al. (2023) "SMS Spam Detection Using Deep Learning Techniques" [1].	DNN, LSTM, Bi-LSTM	Bi-LSTM	98.4%
Hosseinpour & Shakibian (2023) "Ensemble Learning Approach for SMS Spam Detection" [2].	Ensemble Learning	Ensemble Learning	96.4%
K & N (2023) "Accurate SMS Spam Detection Using Support	SVM, Logistic Regression	SVM	97.5%

Vector Machine" [3].			
Manju et al. (2023) "SMS Spam Detection and Filtering of Transliterated Messages" [4].	Random Forest, Naive Bayes, SVM	Random Forest	96.3%
Al-Kabbi, H.A., Feizi-Derakhshi, M., & Pashazadeh, S. (2023) [18]	Multi-Type Feature Extraction, Early Fusion	BERT	99.5%
Gadde, S., Lakshmanarao, A., & Satyanarayana, S. (2021) [13].	LSTM	LSTM	98.5%
Jain, H., & Mahadev, M. (2022) [15].	MultinomialNB	MultinomialNB	97.6%

2.3 Scope of the Problem

Detecting SMS spam messages using machine learning and deep learning involves collecting and preprocessing datasets, exploring various models like traditional machine learning algorithms and deep learning models such as LSTM and GRU, evaluating performance metrics, addressing ethical considerations, and overcoming challenges like dataset imbalance and model generalization. This research aims to enhance cybersecurity in mobile communication by improving spam detection accuracy and efficiency.

2.4 Challenges

Several challenges arise in the domain of SMS spam message detection using machine learning and deep learning techniques. Addressing these challenges involves interdisciplinary approaches, combining expertise in machine learning, natural language processing, cybersecurity, and privacy protection. By overcoming these hurdles, researchers and practitioners can develop more robust and effective SMS spam detection systems that safeguard user experiences and privacy in mobile communication environments.

Imbalanced Datasets: SMS spam datasets often exhibit class imbalance, where the number of spam messages is significantly smaller than non-spam messages. This imbalance can lead to biased models that favor the majority class (ham), affecting detection accuracy for spam messages. The dataset we selected has this serious issue.

Lack of Data Availability: There are a huge shortage of spam dataset. So it is difficult to develop AI models that has an impactful application in countering spam detection. Obtaining large, variety of dataset is key to build successful Ai models and application.

Privacy Concerns: Balancing effective spam detection with user privacy is crucial. Models must analyze message content while respecting privacy regulations and user confidentiality, requiring careful handling of sensitive information within SMS messages.

Generalization: Ensuring that models generalize well to unseen data and new types of spam messages is essential. Overfitting to training data or specific types of spam can reduce the model's effectiveness in detecting emerging spam tactics or variations.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Proposed Methodology

The research experiments were carried out using Google Colab, a platform that supports a wide range of computational tasks. To implement machine learning models, the study leveraged the Keras library, which is renowned for its efficiency in developing deep learning models in Python. TensorFlow, another robust tool within the Python ecosystem, was utilized to execute the machine learning algorithms, providing a comprehensive environment for users to work with.

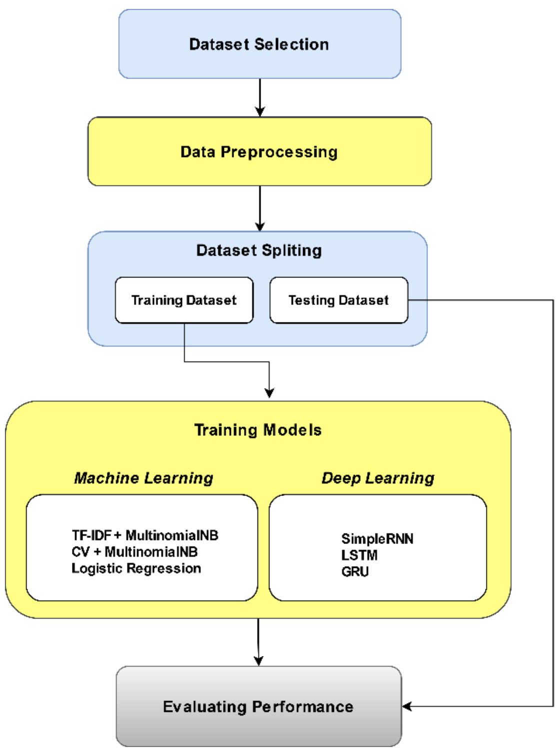

Figure 3.1 Proposed Methodology for this Study

For SMS spam detection, this study focusses on both machine learning and deep learning approach. For machine learning, we selected Multinomial Nave Bayes and Logistic Regression. Also three deep neural network architectures such as SimpleRNN, LSTM (long short term memory) and GRU (gated recurrent unit). The main objective of this project is to build and evaluate a text classification models capable of distinguishing between legitimate messages and spam messages. By training on a labeled dataset of SMS messages, the model learns patterns and characteristics of spam content, allowing it to make accurate predictions on new, unseen messages.

Data Preprocessing: Text preprocessing is an essential step in natural language processing that involves cleaning and transforming raw text data into a format suitable for analysis. In the provided code snippet, we perform the following preprocessing steps on the text data:

1. **Tokenization:** We tokenize the text into individual words, splitting it based on spaces and punctuation marks.
2. **Lowercasing:** All words are converted to lowercase to ensure consistent comparison and analysis.
3. **Stop word Removal:** We eliminate common words (stop words) that do not contribute significantly to the meaning of the text.

The processed words are then joined back together to form preprocessed text. This preprocessing is performed using the word tokenize function from the Natural Language Toolkit (NLTK) library and the stop words list provided by NLTK. By applying these preprocessing techniques, we create a cleaner and more manageable text dataset for further analysis and modeling. This process enhances the quality of text-based tasks such as classification, sentiment analysis, and topic modeling.

Model Selection: For this study, we selected both machine learning and deep learning approach in order to detect spam messages. This choice is based on the dataset's unique features, with attention given to the difficulties of spam dataset.

Multinomial Naive Bayes: Multinomial Naive Bayes is a classification algorithm tailored for discrete data, making it ideal for text classification tasks like spam detection. It works by calculating the probability of each class (e.g., spam or ham) based on the frequency of words, assuming conditional independence of features and a multinomial distribution of word occurrences. In this study, two Multinomial Naive Bayes models were used: one utilizing Term Frequency-Inverse Document Frequency (TF-IDF) to weigh words by their importance across documents, and another employing Count Vectorization (CV) to simply count word occurrences. This dual approach allowed for a comparative analysis of how different feature extraction methods impact the performance of the Multinomial Naive Bayes classifier.

SimpleRNN: The SimpleRNN model in this study was designed to process sequential text data. It consists of an embedding layer that converts text into dense vectors with 244,608 parameters, a SimpleRNN layer with 100 units and 13,300 parameters to capture dependencies, and a dense output layer with 101 parameters for classification. The model has a total of 258,009 trainable parameters. Built with Keras and executed using TensorFlow. This architecture is suitable for tasks involving sequential data, such as text classification.

LSTM (Long Short Term Memory): Long short-term memory (LSTM) is a type of recurrent neural network (RNN) aimed at dealing with the vanishing gradient problem present in traditional RNNs. LSTMs are designed to effectively capture long-range dependencies in sequential data by maintaining and updating a cell state over time. In this architecture, which consists of an embedding layer followed by an LSTM layer and a dense output layer, each component plays a crucial role. The embedding layer transforms input sequences into dense vectors (outputting shapes of (None, 70, 32)), preparing them for sequential processing. The LSTM layer with 100 units processes these embedding, preserving and updating information across sequences (outputting shapes of (None, 100)). Finally, the dense layer, with a single unit and sigmoid activation, produces binary classification outputs (output shape of (None, 1)). This architecture, totaling 297,909 trainable parameters.

GRU (Gated recurrent Unit): A simplified version of LSTMs. They use a single "update gate" to control the flow of information into the memory cell, rather than the three gates used in LSTMs. This makes GRUs easier to train and faster to run than LSTMs. This architecture, totaling 284,909 trainable parameters and same layer architecture as LSTM.

Logistic Regression: Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks, such as identifying whether an email is spam or not and diagnosing diseases by assessing the presence or absence of specific conditions based on patient test results. This model achieved highest accuracy among all previous model described for this study with an accuracy result of 99.84%.

3.2 Data Collection Procedure

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,573 messages, tagged according being ham (legitimate) or spam. A collection of datasets was utilized for SMS spam detection. This includes 425 SMS spam messages manually extracted from the Grumbletext website, a UK forum for reporting SMS spam [19]. Additionally, a subset of 3,375 ham messages was randomly chosen from the NUS SMS Corpus, a dataset of about 10,000 legitimate messages primarily from students at the National University of Singapore [20]. Furthermore, 450 ham messages were collected from Caroline Tag's PhD thesis [21]. Finally, the SMS Spam Corpus v.0.1 Big, which includes 1,002 ham messages and 322 spam messages, was incorporated and has been used in various academic researches [22]. Ethical concerns were crucial throughout the gathering procedure to ensure compliance with privacy and consent norms. The generated dataset provides a diverse and representative foundation for the construction and testing of deep learning and machine learning model for SMS spam detection.

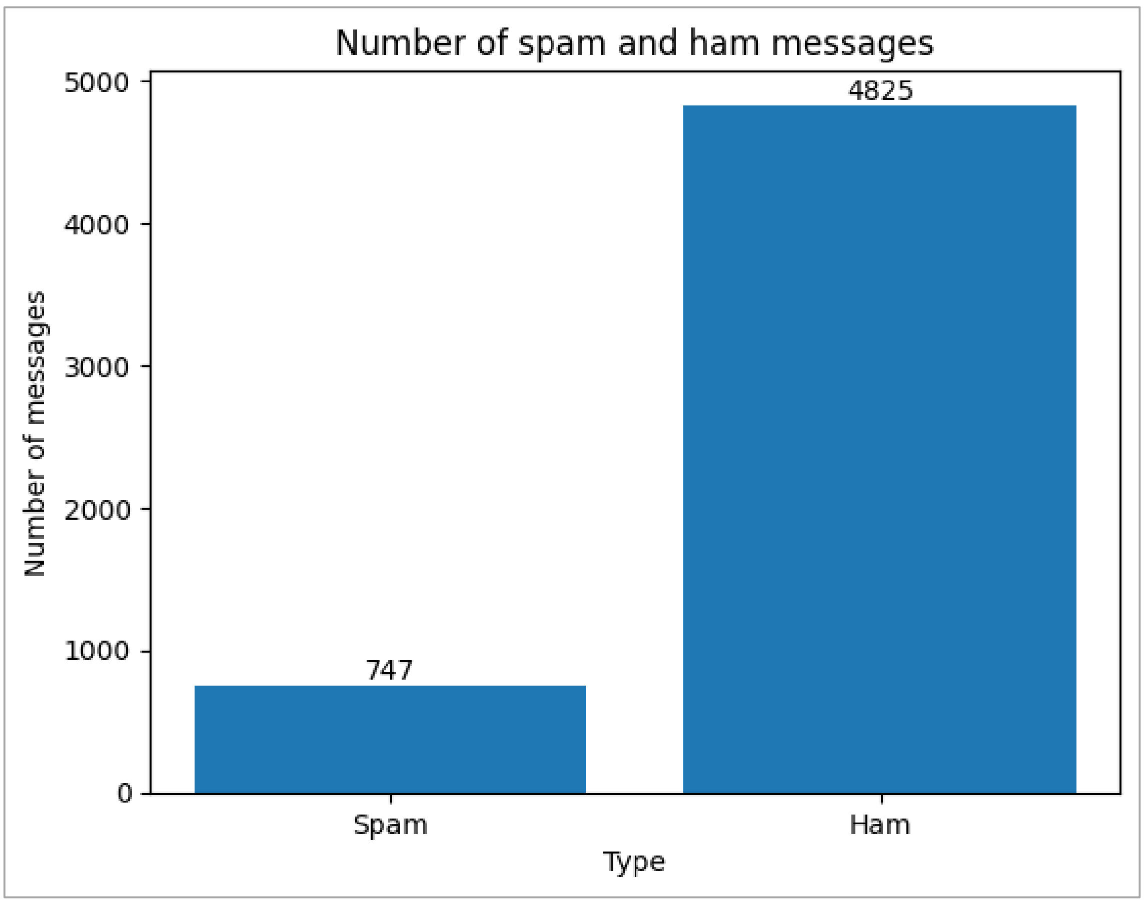

Figure 3.2 Class Distribution: Spam vs. Ham Messages

Figure 3.2 shows a pie chart that describes the number of each class present in the dataset. Where total number of spam and not spam data are 747 and 4825 respectively.

3.3 Implementation Requirement

The research was conducted using Google CoLab, which supports various computational tasks. The Keras library was utilized for implementing machine learning models due to its effectiveness in developing deep learning models in Python. TensorFlow was employed to facilitate the execution of machine learning algorithms, providing a comprehensive environment for building and training the models. Additionally, standard libraries such as Pandas, NumPy, and Matplotlib were used for data manipulation, numerical computations, and data visualization, respectively. Access to the specified datasets and an internet connection for downloading any necessary dependencies were also required.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

The SMS spam Collection is a set of SMS tagged messages that have been collected for SMS spam research. It contains one set of SMS messages in English of 5,573 messages, tagged according being ham (legitimate) or spam. Multinomial Naive Bayes, Logistic Regression, LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit) and SimpleRNN are the models selected for the study. This study uses both machine learning and deep learning in order to achieve better prediction result. Multinomial Naive Bayes and Logistic regression are the machine learning model selected and LSTM, GRU, SimpleRNN are the deep learning models in this study. Where Logistic Regression outperformed all the selected models for this study with an accuracy of 99.84%. We implemented Multinomial Naive Bayes in two ways, one for TF-IDF (Term Frequency-Inverse Document Frequency) and another for CV (Count Vectorization). The experiment was carried on Google Colab which is managed Jupiter based notebook service used for data science and AI. We utilized a range of libraries, such as Scikit-learn, TensorFlow, and others, to preprocess the dataset and construct and train our AI models. The Pandas library was utilized to manipulate and convert data, ensuring its quality and consistency. The models were carefully trained, and parameters were tuned based on validation results. The last evaluation was performed on an independent test set. After testing the models, we analyzed and visualize the results of the selected models.

4.2 Experimental Results

The experimental shows that the selected models give different results. Where Logistic regression outperformed all the models including deep learning models. Among two Multinomial Naïve Bayes model, the one with CV (Count Vectorization) gives better accuracy which is 97.68% where TF-IDF (Term Frequency-Inverse Document Frequency) based Multinomial Naive Bayes gives only 95.84%. Deep learning models GRU, LSTM and SimpleRNN given almost same result with accuracy such as 97.78%, 97.78% and 97.97% respectively. We evaluated the Accuracy, Precision, Recall, and F-1 Score of the Confusion Matrices for our proposed methods.

Accuracy: The ratio of the number of correctly predicted instances to the total number of instances. It measures the overall effectiveness of the model in making correct predictions.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision: The ratio of the number of correctly predicted positive instances to the total number of instances predicted as positive. It measures the accuracy of the positive predictions.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

F1 Score: The harmonic mean of precision and recall. It provides a single metric that balances both the precision and recall, especially useful when you need a balance between the two or when dealing with imbalanced datasets.

$$F - 1\, Score = 2 * \frac{Recall * \ Precision}{Recall + Precision}$$

Recall: The ratio of the number of correctly predicted positive instances to the total number of actual positive instances. It measures the ability of the model to find all the relevant positive cases.

$$Recall = \frac{TruePositive}{TruePositive + \ FalseNegative}$$

The result of classification model is compared on the basis of Accuracy, Precision, Recall, F1 Score in below table of 4.2:

Table 4.2: Performance Evaluation

Model Name	Accuracy	Precision	Recall	F1
TF-IDF + MultinomialNB	95.84%	100%	70.34%	82.59%
CV + MultinomialNB	97.68%	94.16%	88.97%	91.49%
LSTM	97.78%	93.57%	90.34%	91,93%
GRU	97.78%	95.21%	87.57%	91.70%
SimpleRNN	97.97%	96.27%	92.47%	92.47%
Logistic Regression	99.84%	99.69%	100%	99.84%

Table 4.2 shows traditional methods, TD-IDF + MultinomialNB achieves high precision (100%) but lower recall (70.34%) and F1 score (82.59%). CV + MultinomialNB improves with higher accuracy (97.68%), balanced precision (94.16%), and recall (88.97%), resulting in a higher F1 score (91.49%). LSTM, GRU, and SimpleRNN all achieve strong accuracies around 97-98%. LSTM shows balanced precision (93.57%) and recall (90.34%), with an F1 score of 91.93%. GRU excels in precision (95.21%) but has lower recall (87.57%), resulting in an F1 score of 91.70%. Logistic regression outperforms all with the highest accuracy (99.84%), near-perfect precision (99.69%), recall (100%), and F1 score (99.84%).

4.3 Accuracy

The accuracy results indicate the importance of machine learning and deep learning models in detecting spam messages. Logistic Regression achieved the highest accuracy rate of 99.84%. The figure 4.3 shows the accuracy comparison of the different model:

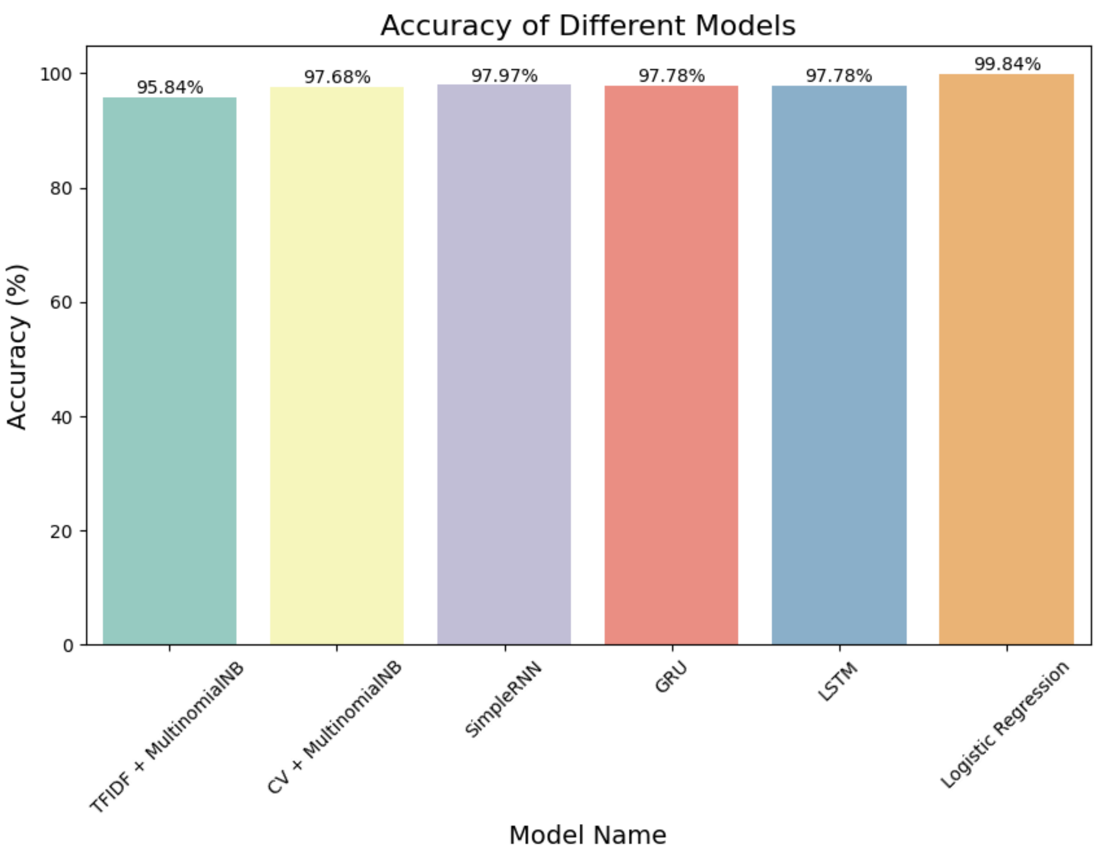

Figure 4.3: Accuracy Comparison of Deep Learning and Machine Learning Models

4.4 Performance Analysis

TFIDF + Multinomial Naive Bayes:

Table 4.4.1 Performance Evaluation (TFIDF + Multinomial Naive Bayes)

	Precision	Recall	F1-Score	Support
ham	0.95	1.00	0.98	889
spam	1.00	0.70	0.83	145
Accuracy			0.98	1034
Macro avg	0.98	0.85	0.90	1034
Weighted avg	0.96	0.96	0.96	1034

Table 4.4.1 shows that the TFIDF + Multinomial Naive Bayes model achieves a high overall accuracy of 98% and balanced F1-scores for spam detection. It performs excellently for "Ham" cases, but there are some inconsistencies in detecting "Spam" due to lower recall.

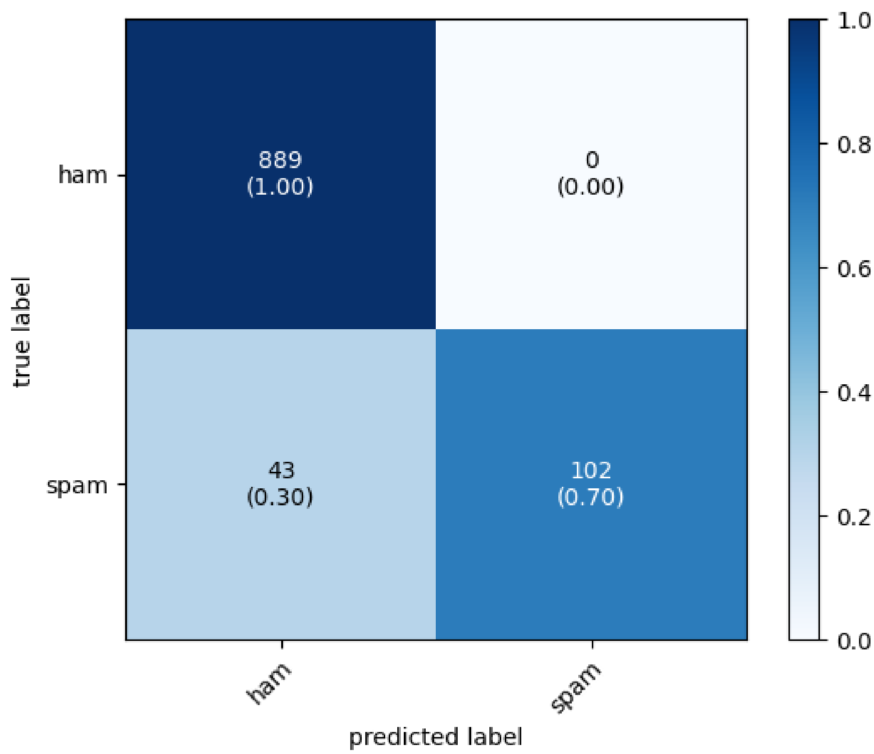

Figure 4.4.1 Confusion Matrix for TFIDF + Multinomial Naive Bayes

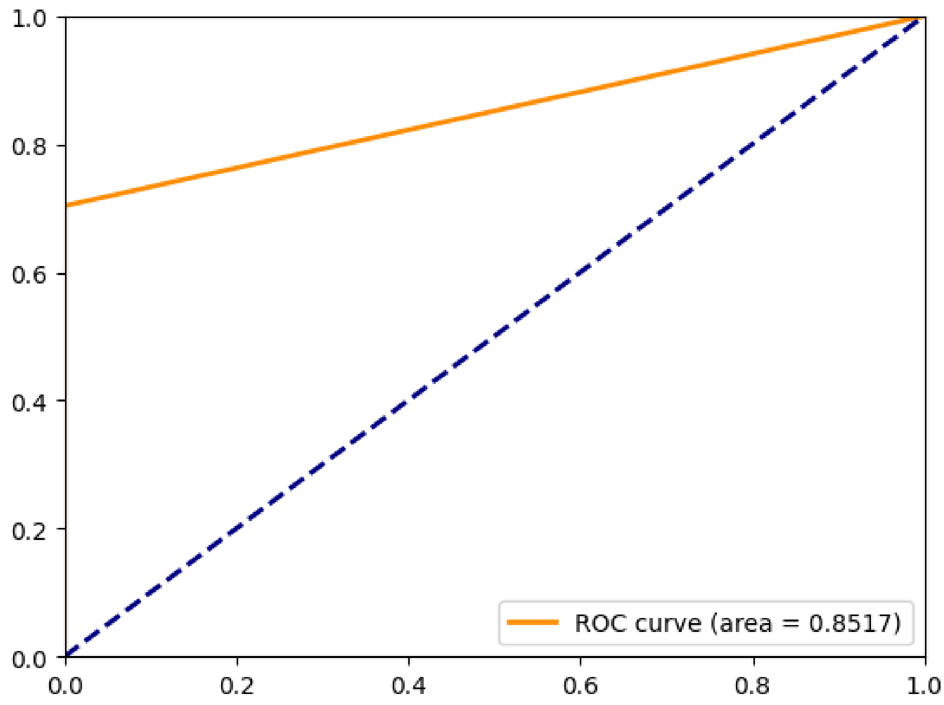

Figure 4.4.2 ROC Curve for TFIDF + Multinomial Naive Bayes

CV + Multinomial Naive Bayes:

Table 4.4.2 Performance Evaluation (CV + Multinomial Naive Bayes)

	Precision	Recall	F1-Score	Support
ham	0.98	0.99	0.99	889
spam	0.94	0.89	0.91	145
Accuracy			0.98	1034
Macro avg	0.96	0.94	0.95	1034
Weighted avg	0.98	0.98	0.98	1034

Table 4.4.2 shows high precision (98% for "ham", 94% for "spam") and recall (99% for "ham", 89% for "spam"), with an overall accuracy of 98%. The F1-scores are 99% for "ham" and 91% for "spam", indicating strong performance in classifying messages as spam or non-spam.

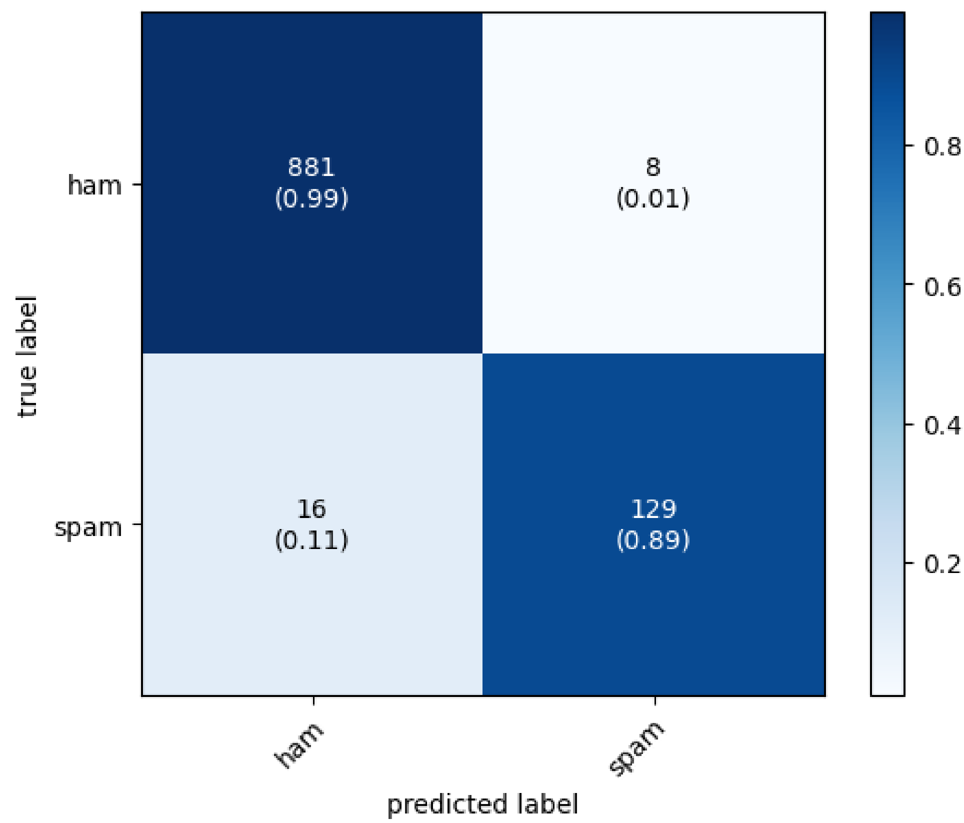

Figure 4.4.3 Confusion Matrix for CV + Multinomial Naive Bayes

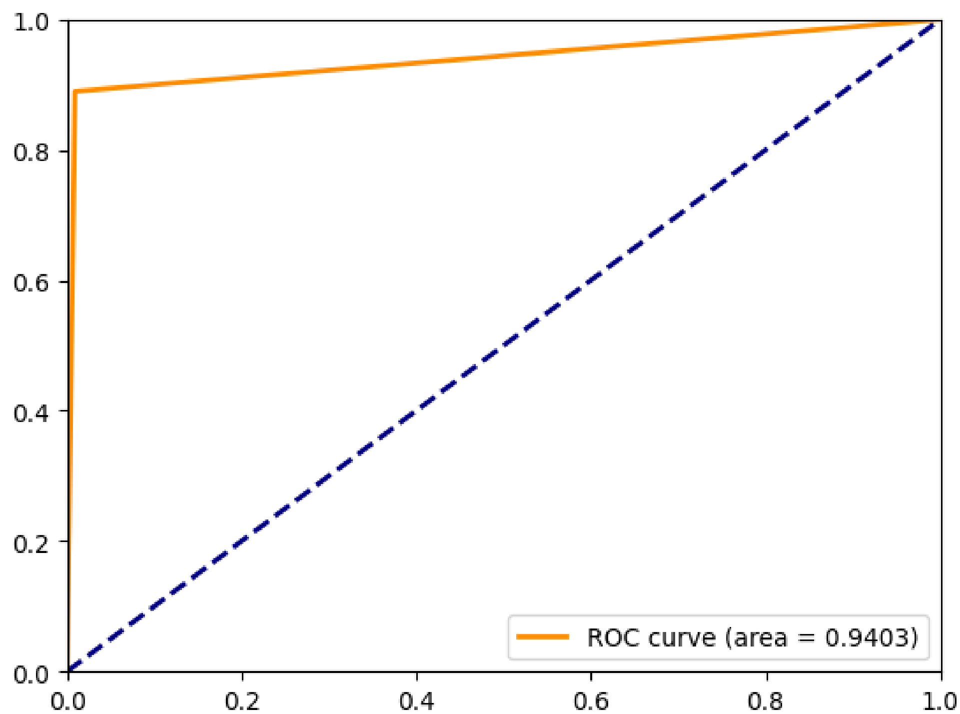

Figure 4.4.4 ROC Curve for CV + Multinomial Naive Bayes

GRU (Gated Recurrent Unit):

Table 4.4.3 Performance Evaluation (GRU)

	Precision	Recall	F1-Score	Support
ham	0.98	0.99	0.99	889
spam	0.96	0.88	0.92	145
Accuracy			0.98	1034
Macro avg	0.97	0.94	0.95	1034
Weighted avg	0.98	0.98	0.98	1034

Table 4.4.3 shows strong precision (98% for "ham", 96% for "spam") and recall (99% for "ham", 88% for "spam"), resulting in high F1-scores (99% for "ham", 92% for "spam"). The model achieves an overall accuracy of 98%, demonstrating robust performance in classifying messages as either "ham" or "spam".

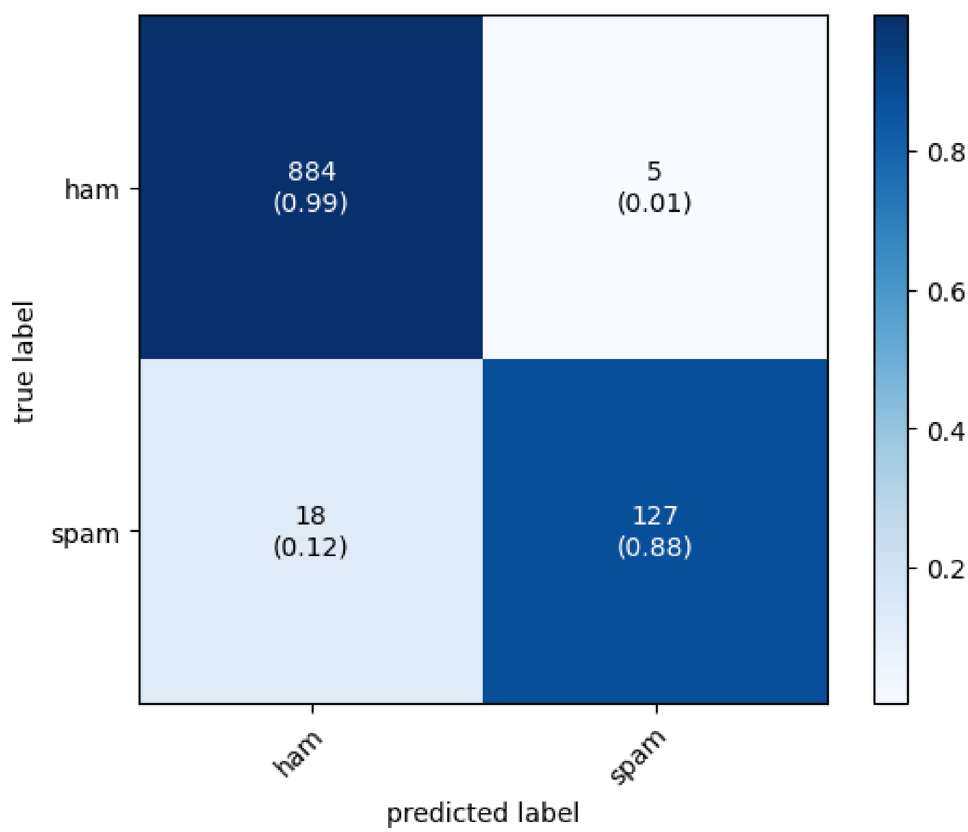

Figure 4.4.5 Confusion Matrix for GRU

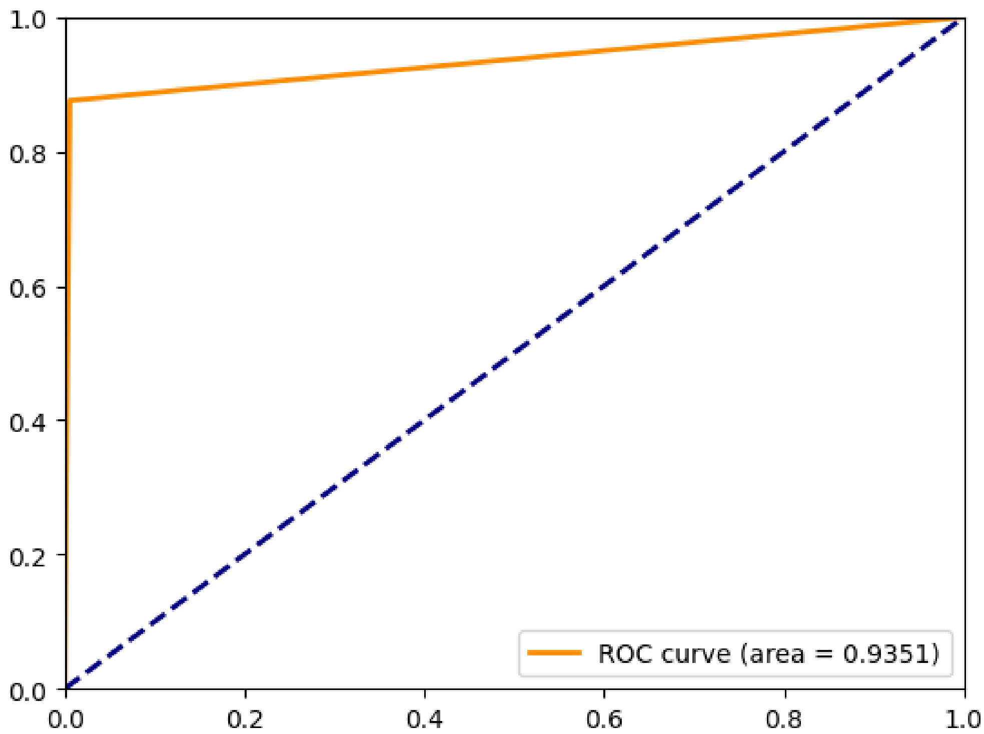

Figure 4.4.6 ROC Curve for GRU

LSTM (Long Short Term Memory):

Table 4.4.4 Performance Evaluation (LSTM)

	Precision	Recall	F1-Score	Support
ham	0.98	0.99	0.99	889
spam	0.94	0.90	0.92	145
Accuracy			0.98	1034
Macro avg	0.96	0.95	0.95	1034
Weighted avg	0.98	0.98	0.98	1034

Table 4.4.4 indicates high precision (98% for "ham", 94% for "spam") and recall (99% for "ham", 90% for "spam"), resulting in strong F1-scores (99% for "ham", 92% for "spam"). The model achieves an overall accuracy of 98%, showing consistent and effective performance in classifying messages as either "ham" or "spam".

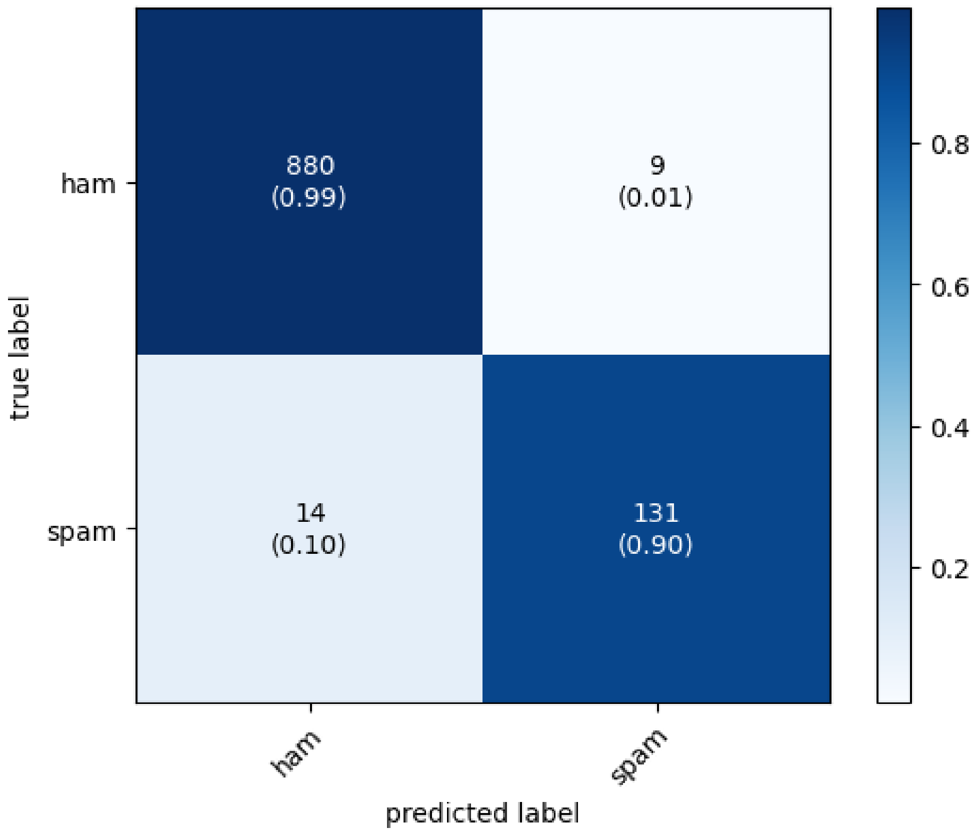

Figure 4.4.7 Confusion Matrix for LTSM

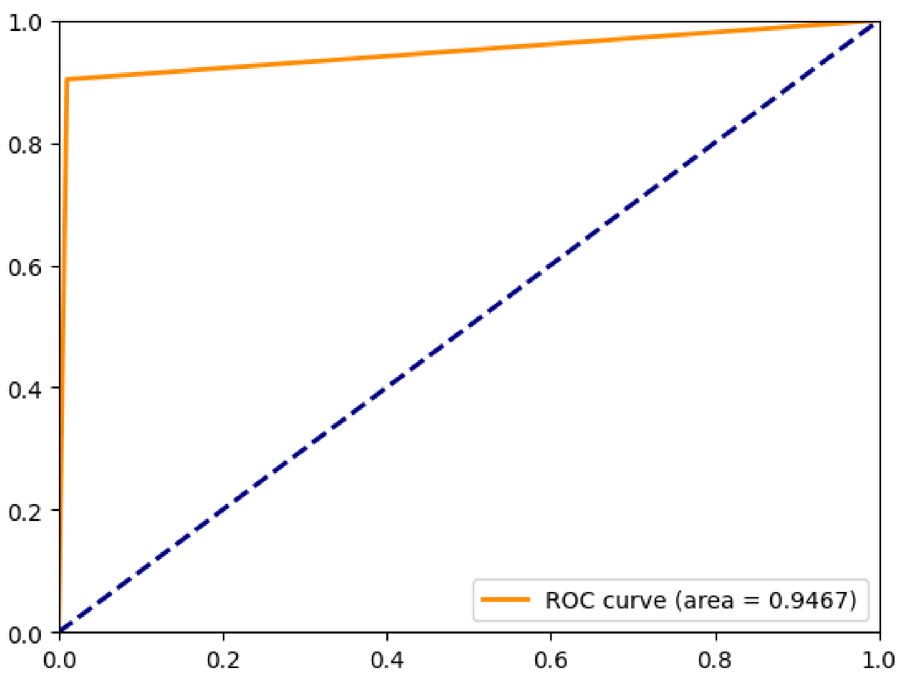

Figure 4.4.8 ROC Curve for LSTM

SimpleRNN:

Table 4.4.5 Performance Evaluation (SimpleRNN)

	Precision	Recall	F1-Score	Support
ham	0.98	0.99	0.99	889
spam	0.96	0.89	0.92	145
Accuracy			0.98	1034
Macro avg	0.97	0.94	0.96	1034
Weighted avg	0.98	0.98	0.98	1034

Table 4.4.5 shows SimpleRNN model achieves high precision (98% for "ham", 96% for "spam") and recall (99% for "ham", 89% for "spam"), resulting in strong F1-scores (99% for "ham", 92% for "spam"). The model demonstrates an overall accuracy of 98%, showing effective performance in distinguishing between "ham" and "spam" messages.

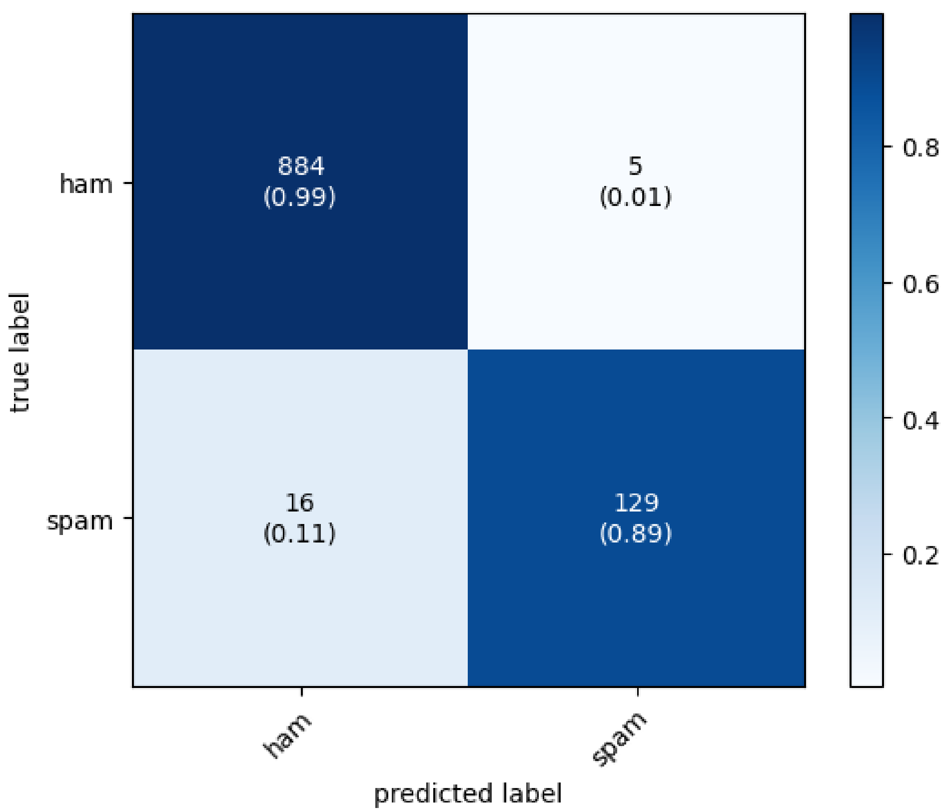

Figure 4.4.9 Confusion Matrix for SimpleRNN

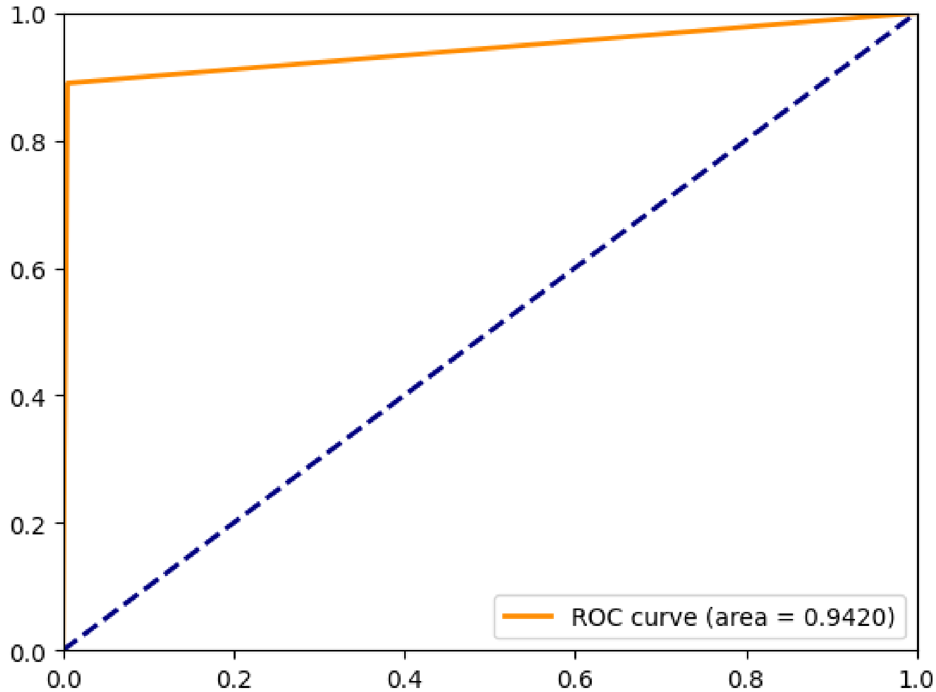

Figure 4.4.10 ROC Curve for SimpleRNN

Logistic Regression:

Table 4.4.6 Performance Evaluation (Logistic Regression)

	Precision	Recall	F1-Score	Support
ham	1.00	1.00	1.00	960
spam	1.00	1.00	1.00	970
Accuracy			1.00	1930
Macro avg	1.00	1.00	1.00	1930
Weighted avg	1.00	1.00	1.00	1930

Table 4.4.6 shows that Logistic Regression perfect precision (100%) and recall (100%) for both "ham" and "spam" classes, resulting in flawless F1-scores (100%). The model achieves an outstanding accuracy of 100%, demonstrating impeccable performance in classifying messages as either "ham" or "spam". Both macro average and weighted average metrics confirm the model's

flawless performance across the entire dataset, emphasizing its exceptional accuracy and reliability.

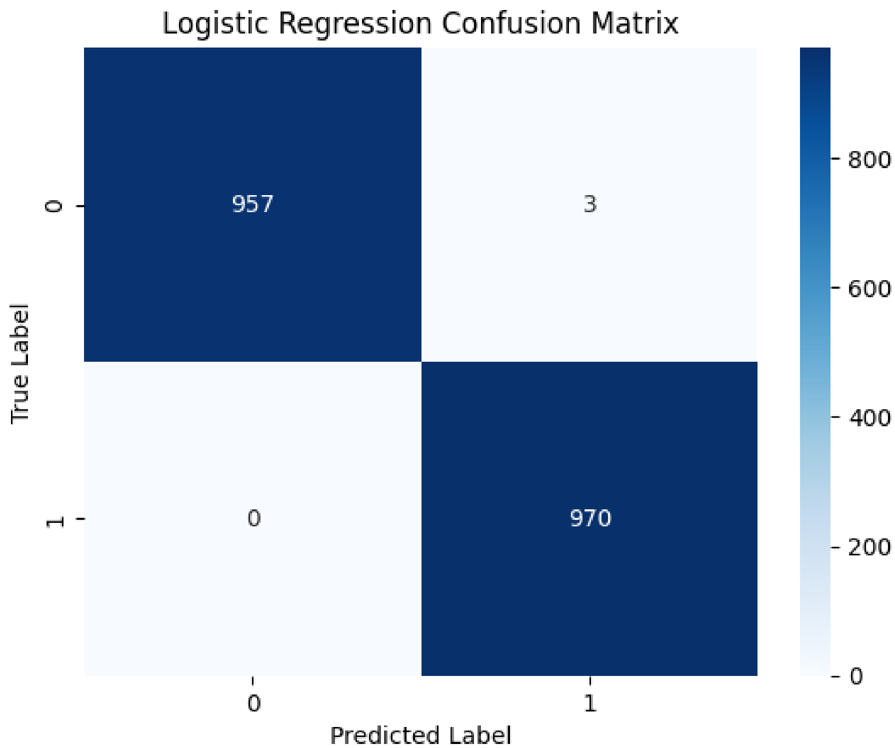

Figure 4.4.11 Confusion Matrix for Logistic Regression

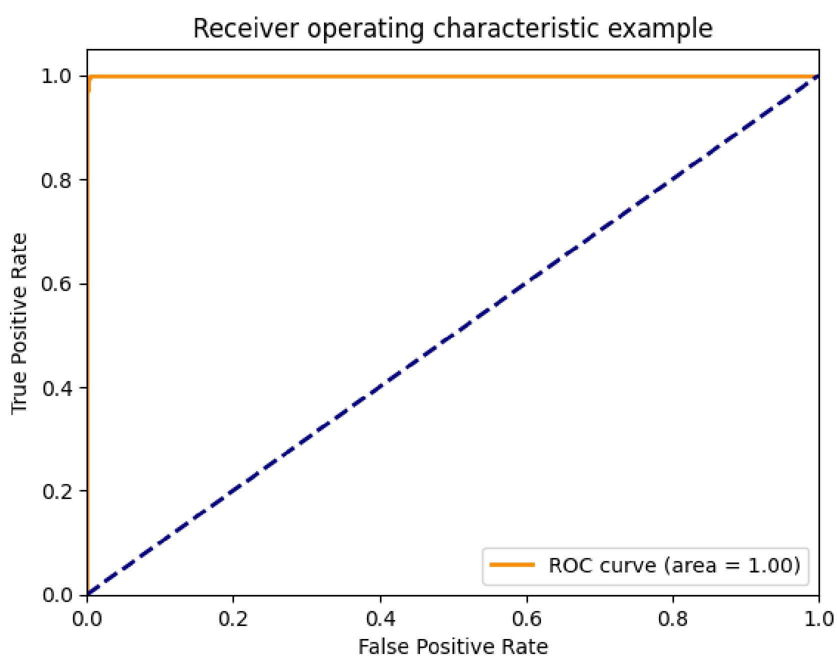

Figure 4.4.12 ROC Curve for Logistic Regression

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary of the Study

This study focuses on finding the best AI model that can detect SMS spam with highest accuracy possible. Both machine learning and deep learning is being used to achieve better accuracy in detecting SMS spam. It uses a dataset consist of total 5572 number of data among 4825 non-spam and 747 was spam message. Before training and evaluating models, preprocess, feature extraction has been done according to the target result. For machine learning, Multinomial Naive Bayes and Logistic Regression. Multinomial Naive Bayes has been trained in two ways. TF-IDF and CV (Count Vectorization). Among them, CV based MultinomialNB performed better than TF-IDF based MultinomialNB which is 97.68% compared to 95.85%. In deep learning, we choose RNN based models such as LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit) and SimpleRNN in order to achieve better accuracy. These deep learning models achieve accuracy of 97.78%, 97.78% and 97.97% respectively. Later Logistic Regression, another machine learning involves in this study outperformed all the models with an accuracy of 99.84% with precision of 99.69%.

5.2 Conclusion

In this project, we explored various machine learning (ML) and deep learning (DL) techniques to detect SMS spam messages. The goal was to identify the most effective model for accurately classifying SMS messages as spam or ham. We implemented and evaluated two ML models: Logistic Regression and Multinomial Naive Bayes. Among these, Logistic Regression demonstrated superior performance, outclassing Multinomial Naive Bayes. This can be attributed to Logistic Regression's capability to find the optimal decision boundary for classifying SMS messages. We also experimented with deep learning models such as Simple RNN, LSTM, and GRU models. Each of these models has its strengths in handling sequential data. However, despite their advanced architectures and ability to capture temporal dependencies, they did not outperform the Logistic Regression model. This result might be due to the complexity and size of our dataset, where traditional ML models like Logistic Regression can achieve high performance without the need for extensive training and computational resources.

5.3 Future Work

The future of SMS spam detection using Machine learning and deep learning looks promising with advancements aimed at enhancing accuracy, real-time detection capabilities, and adaptability to evolving spam tactics. These technologies are poised to improve significantly, leveraging larger datasets and sophisticated algorithms to achieve higher detection rates. Integration with mobile networks could streamline spam filtering processes, while multilingual support ensures effectiveness across diverse linguistic contexts. Privacy considerations and user experience enhancements are also focal points, aiming to mitigate spam without compromising user data or satisfaction. Overall, the future holds potential for more robust and efficient SMS spam detection systems that can better safeguard users worldwide.

REFERENCES

[1] Gandhi, C., Kumar Sarangi, P., Saxena, M., & Sahoo, A.K. (2023). SMS Spam Detection Using Deep Learning Techniques: A Comparative Analysis of DNN Vs LSTM Vs Bi-LSTM. 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 189-194.

[2] Hosseinpour, S., & Shakibian, H. (2023). An Ensemble Learning Approach for SMS Spam Detection. 2023 9th International Conference on Web Research (ICWR), 125-128.

[3] K, R.B., & N, D. (2023). Accurate SMS Spam Detection Using Support Vector Machine in Comparison with Logistic Regression. 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1-5.

[4] Manju, G., S., D.N., Mondal, S.J., & Sthapak, H. (2023). SMS Spam Detection and Filtering of Transliterated Messages. 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), 1-9.

[5] Al-Kabbi, H.A., Feizi-Derakhshi, M., & Pashazadeh, S. (2023). Multi-Type Feature Extraction and Early Fusion Framework for SMS Spam Detection. IEEE Access, 11, 123756-123765.

[6] Manikkannan, P.D. (2023). SMS Spam Detection using Machine Learning. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT.

[7] K, R.B., & N, D. (2023). Accurate SMS Spam Detection Using Support Vector Machine In Comparison With Linear Regression. 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-4.

[8] Ahluwalia, K., GururajHL1, □., & Lin, H. (2023). Comparative Analysis of Various SMS Spam Detection Methods using Machine Learning. Proceedings of 2023 the 13th International Workshop on Computer Science and Engineering.

[9] Sankar, E. (2023). SMS SPAM DETECTION USING MACHINE LEARNING. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT.

[10] Ali, Z.H., Salman, H.M., & Harif, A.H. (2023). SMS Spam Detection Using Multiple Linear Regression and Extreme Learning Machines. Iraqi Journal of Science.

[11] Siagian, W., Setiadi, M.R., & Prasetyo, S.Y. (2023). Improving SMS Spam Detection Through Machine Learning: An Investigation of Feature Extraction and Model Selection Techniques. 2023 International Conference on Information Management and Technology (ICIMTech), 288-293.

[12] Joseph, P., & Yerima, S.Y. (2022). A comparative study of word embedding techniques for SMS spam detection. 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 149-155.

[13] Gadde, S., Lakshmanarao, A., & Satyanarayana, S. (2021). SMS Spam Detection using Machine Learning and Deep Learning Techniques. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 1, 358-362.

[14] Aliza, H.Y., Nagary, K.A., Ahmed, E., Puspita, K.M., Rimi, K.A., Khater, A., & Faisal, F. (2022). A Comparative Analysis of SMS Spam Detection employing Machine Learning Methods. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 916-922.

[15] Jain, H., & Mahadev, M. (2022). An Analysis of SMS Spam Detection using Machine Learning Model. 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 151-156.

[16] Kalyani, V.V., Siva, P., Prasad, S., Dr.M.V.Rama, Professor, S., Aiml, I., P.PattabhıRama, Neelima, S., & A.Lakshmanarao (2024). SMS Spam Detection using NLP and Deep Learning Recurrent Neural Network Variants. 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), 92-96.

[17] Wei, F., & Nguyen, U.T. (2020). A Lightweight Deep Neural Model for SMS Spam Detection. 2020 International Symposium on Networks, Computers and Communications (ISNCC), 1-6.

[18] Al-Kabbi, H.A., Feizi-Derakhshi, M., & Pashazadeh, S. (2023). Multi-Type Feature Extraction and Early Fusion Framework for SMS Spam Detection. IEEE Access, 11, 123756-123765.

[19] The Grumbletext Web http://www.grumbletext.co.uk/

[20] The NUS SMS Corpus is avalaible at http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/

[21] http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

[22] http://www.esp.uem.es/jmgomez/smsspamcorpus/

COMPREHENSIVE STUDY OF SMS SPAM DETECTION STRATEGIES BASED ON TCP PROTOCOLS

ORIGINALITY REPORT

22%	18%	9%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	restpublisher.com Internet Source	1%
4	Submitted to Southampton Solent University Student Paper	1%
5	encyclopedia.kaspersky.com Internet Source	1%
6	wikimili.com Internet Source	1%
7	www.coursehero.com Internet Source	1%
8	Submitted to Aspen University Student Paper	1%
9	Submitted to University of Central Lancashire Student Paper	1%

10	en.wikipedia.org Internet Source	1 %
11	spectrum.library.concordia.ca Internet Source	1 %
12	Submitted to University of Leeds Student Paper	1 %
13	isset.shiksha.com Internet Source	1 %
14	smartcontent.top Internet Source	1 %
15	Submitted to Nanyang Technological University Student Paper	<1 %
16	www.mdpi.com Internet Source	<1 %
17	Submitted to October University for Modern Sciences and Arts (MSA) Student Paper	<1 %
18	www.semanticscholar.org Internet Source	<1 %
19	Submitted to La Trobe University Student Paper	<1 %
20	dblp.dagstuhl.de Internet Source	<1 %