**Spam analysis and classification of the dynamic message using a vectorizing technique using NLP**

### A PROJECT REPORT

***Submitted by***

**Tushar Gupta [RA2011028010149] Sahil Pandey [RA2011028010134]** *Under the Guidance of*

## Dr. Thenmalar S

(Associate Professor, Department of Networking and Communications)

***in partial fulfillment of the requirements for the degree of***

## BACHELOR OF TECHNOLOGY

**in**

## COMPUTER SCIENCE ENGINEERING

**with specialization in Cloud Computing**

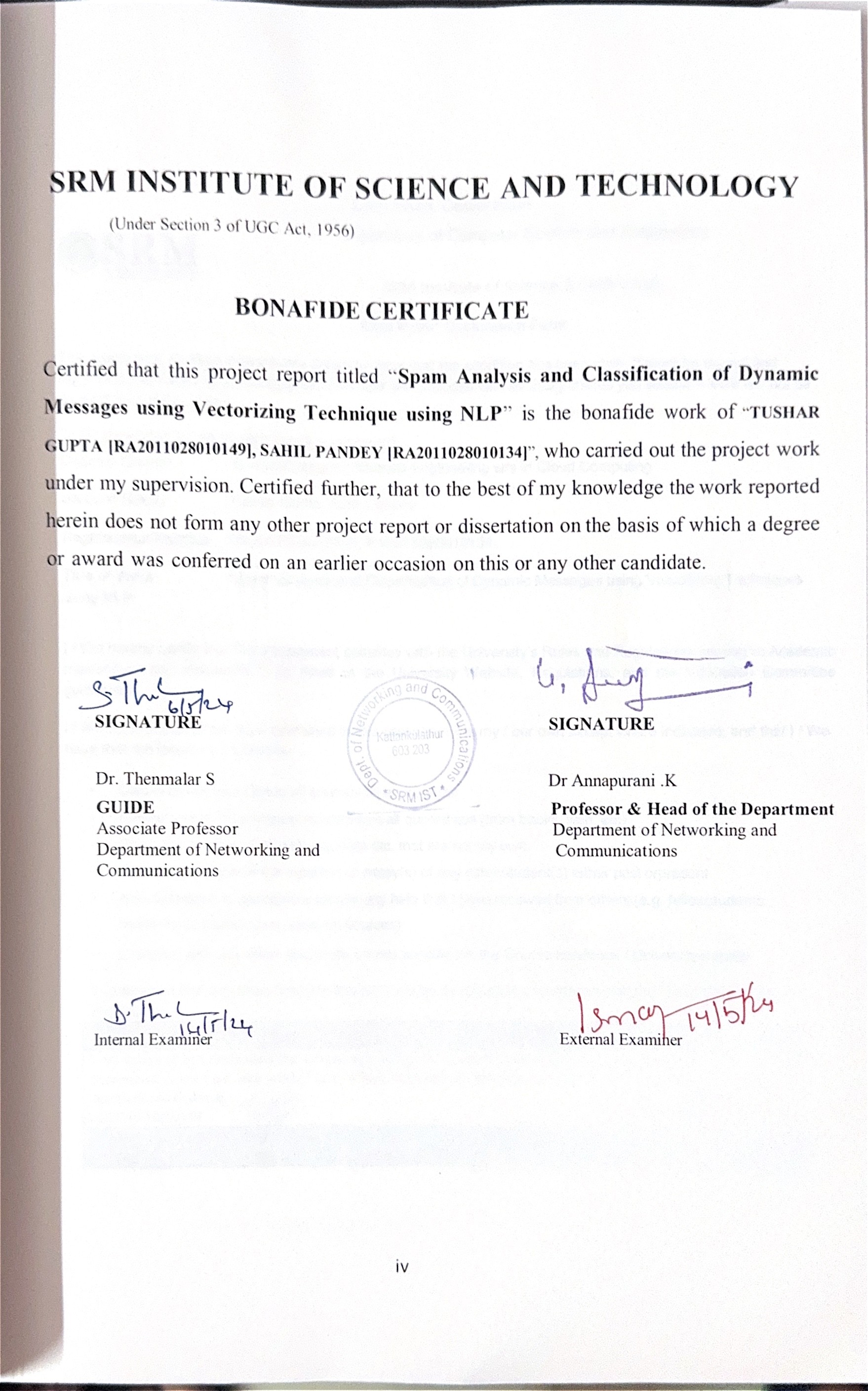


## DEPARTMENT OF NETWORKING AND COMMUNICATIONS

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

KATTANKULATHUR- 603 203

MAY 2024





# ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr. Sandeep Sancheti**, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr .C. Muthamizhchelvan**, Director, Faculty of Engineering and Technology, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank **Dr. Annapurani .K**, Professor & Head of Department, Department of Networking and Communications, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughoutthe period of the project work.

We are extremely grateful to our Academic Advisor **Dr. Sarvanan M**, Professor, Department of Networking and Communications, SRM Instituteof Science and Technology, for their great support at all the stages of project work.

We would like to convey our thanks to our Panel Head, **Dr. S. Thenmalar**, Associate professor , Depart- ment of Networking and Communications, SRM Institute of Science and Technology, for her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisor, **Dr. Senthamarai N**, Assistant Professor, and **Dr. T.Y.J. Naga Malleswari** , Associate Professor , Department of Networking and Communications, SRMInstitute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to our guide, **Dr. S. Thenmalar**, Associate professor , Department of Networking and Communications, SRM Institute of Science and Technology, for providing usan opportunity to pursue my project under her mentorship. She provided us the freedom and support to explore the research topics of our interest.Her passion for solving the real problems and making a difference in the world has always been inspiring.

We sincerely thank staff and students of the Computer Science and Engineering Department, SRM Institute of Science and Technology, for their help during my research. Finally, we would like to thank my parents, our family members and our friends for their unconditional love, constant support and encouragement.

**Tushar Gupta (RA2011028010149) & Sahil Pandey (RA2011028010134)**

# ABSTRACT

Spam messages pose a significant threat to communication platforms, necessitating advanced techniques for analysis and classification. This study focuses on the development of a dynamic message classification system using vectorizing techniques in conjunction with a multi-model machine learning algorithm. By leveraging the inherent characteristics of spam messages, this approach aims to accurately identify and categorize them in real-time. The vectorizing technique transforms text data from messages into a numerical format, enabling the algorithm to process and analyze vast amounts of information efficiently. Through the utilization of a multi-model machine learning algorithm, the system can adapt to the diverse and evolving nature of spam messages, enhancing its classification accuracy. By combining these methods into a unified framework, this research seeks to provide a comprehensive solution for dynamic spam message identification. The proposed system not only offers robust protection against spam but also establishes a foundation for continuous improvement and refinement in combating spam messages across various communication platforms. Vectorizing and multi-model machine learning methods for spam message dynamic analysis and categorization are novel features of this study. This research vectorizes spam communications into high-dimensional representations to better grasp their content and context than static feature extraction methods. The system can adapt to different spam messages and increase its classification accuracy by employing a range of machine learning methods, including decision trees and neural networks, and ensemble approaches.

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** |  |
| **ABSTRACT** | **i** |
| **ACKNOWLEDGEMENTS** | **Ii** |
| **LIST OF FIGURES** | **vi** |
| **ABBREVIATIONS** | **vii** |
| 1. **INTRODUCTION**    1. Motivation | **1** |
| 2 |
| 1.2 Objective | 3 |
| 1.3 Problem Statement | 4 |
| 1.4 Challenges | 4 |
| **2 LITERATURE SURVEY** | **6** |
| 2.1 Types of Techniques | 6 |
| 2.1.1 Naive Bayes | 6 |
| 2.1.2 TF-IDF | 7 |
| 2.1.3 LSTM | 7 |
| 2.1.4 Bi-Directional LSTM | 8 |
| 2.2 Inference from Survey | 8 |
| **3 SYSTEM DESIGN** | **10** |
| 3.1 Architecture Overview | 10 |
| 3.2 System Modules | 11 |
| 3.2.1 Data Cleaning and Text Preprocessing | 11 |
| 3.2.2 Vectorization | 13 |
| 3.2.3 Classification | 15 |

|  |  |
| --- | --- |
| **4 SYSTEM REQUIREMENTS** | **16** |
| 4.1 Hardware Requirements | 16 |
| 4.2 Software Requirements | 16 |
| **5 EXPERIMENT AND RESULTS** | **17** |
| **6 CONCLUSION** | **26** |
| **7 FUTURE ENHANCEMENT** | **27** |
| **REFERENCES** | **28** |
| **APPENDIX** |  |
| **A CODING** | **30** |
| **B CONFERENCE PUBLICATION** | **34** |
| **C PLAGIARISM REPORT** | **37** |

# LIST OF FIGURES

#### 3.1 Overview of Spam Analysis and Classification 10

#### Accuracy for Spam Analysis and Classification 21

#### Training Loss for Spam Analysis and Classification 22

#### Confusion Matrix for Spam Analysis and Classification 22

#### Receiver Operating Characteristic for Spam Analysis and Classification 23

#### Email Classified as Ham 24

#### Email Classified as Spam 24

**LIST OF ABBREVIATIONS**

NLP - Natural Language Processing ML - Machine Learning

SVM - Support Vector Machine DT - Decision Tree

NB - Naive Bayes RF - Random Forest

CNN - Convolutional Neural Network LSTM - Long Short-Term Memory GAN - Generative Adversarial Network NE - Named Entity

PCA - Principal Component Analysis KNN - K-Nearest Neighbors

LDA - Latent Dirichlet Allocation Word2Vec - Word to Vector

TF-IDF - Term Frequency-Inverse Document Frequency

# CHAPTER 1 INTRODUCTION

Spam analysis is a crucial task in the field of cybersecurity aimed at efficiently detecting and categorizing unwanted, unsolicited messages that can pose potential threats to users. Using a multi- model machine learning approach in conjunction with a vectorizing technique, the dynamic nature of spam messages can be effectively classified, enhancing the overall accuracy of the spam filtering process. Through the vectorizing technique, textual data from spam messages is transformed into numerical representations, allowing for the extraction of meaningful patterns and features that aid in distinguishing between legitimate and malicious communications. The multi-model machine learning approach further refines this process by utilizing a combination of diverse algorithms to capture a wide range of characteristics exhibited by spam messages, thus boosting the robustness and reliability of the classification system. By integrating these advanced methodologies, the analysis of spam messages can be significantly improved, enabling cybersecurity professionals to detect and manage spam content more effectively, ultimately safeguarding users from potential security risks and preserving the integrity of digital communication environments.Spam analysis and classification of dynamic messages have become critical in the realms of information security and data analysis. With the escalating volume of unsolicited and malicious messages, developing efficient methods for spam detection is imperative. Advanced techniques such as vectorizing and multi-model machine learning algorithms are crucial in addressing this challenge. The concept of vectorizing involves converting text-based data into numerical representations for processing by machine learning algorithms. This technique, proven effective in natural language processing tasks, enables the extraction of important patterns and characteristics. Vectorizing allows the inclusion of both message content and metadata, providing a comprehensive analysis of dynamic messages. Metadata, encompassing sender details, time of sending, and message length, enhances the ability to distinguish between legitimate and spam messages. By incorporating content and metadata, a more reliable classification model can be constructed. The sheer volume and evolving nature of spam necessitate sophisticated methods. Traditional approaches often struggle to adapt to new tactics employed by spammers, leading to false positives or negatives. Additionally, the reliance on single-model algorithms may limit the overall effectiveness of the system. Addressing

these challenges is paramount for ensuring robust spam detection capabilities. The workflow of the proposed system involves the initial step of vectorizing dynamic messages. This includes converting text data into numerical representations, incorporating both message content and metadata. The vectorized data is then utilized in training multi-model machine learning algorithms. These algorithms, leveraging diverse features such as statistical measures, keyword frequencies, and linguistic patterns, collectively contribute to accurate spam classification. Evaluation of the system will involve assessing its performance on a dataset containing a mix of legitimate and spam messages. Metrics such as precision, recall, andF1 score will be employed to quantify the system's effectiveness. Vectorizing techniques play a crucial role in spam analysis and classification of dynamic messages, particularly when combined with multi-model machine learning algorithms. These techniques involve converting text data into numerical vectors to enable machine learning models to process and analyze the information effectively. A frequently used method for vectorizing is the Bag-of-Words approach, where a document is represented as a set of words, disregarding their sequence. Another widely used method is Term Frequency-Inverse text Frequency (TF-IDF), which assigns weights to words based on their frequency in a text compared to their frequency in the total corpus. Furthermore, Word Embeddings, such as Word2Vec and GloVe, encode semantic connections between words by modeling them as compact vectors in a continuous vector space. By utilizing these vectorization approaches alongside multi-model machine learning algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines, spam analysis systems may accurately detect and categorize dynamic communications with great precision. These algorithms leverage the numerical representations of text data provided by vectorizing techniques to learn patterns and relationships, enabling them to generalize well on unseen data and make robust predictions. The combination of advanced vectorizing techniques and multi-model machine learning algorithms offers a powerful solution for tackling the challenges of spam detection and classification in the context of dynamic messages.

# Motivation

Dynamic messages in the context of spam analysis can be classified into two main categories: content-based dynamic messages and behavior-based dynamic messages. Content-based dynamic messages are those that use constantly changing text or images to evade traditional spam filters by

altering the content of the message with each iteration. On the other hand, behavior-based dynamic messages are those that utilize dynamic links or attachments that change once the message is opened, making it difficult to identify and classify them as spam based on static features alone. To classify dynamic messages for spam analysis, a vectorizing technique can be employed to convert the text, images, and other components of the message into numerical vectors that can be used as input features for machine learning algorithms. One approach is to use a multi-model machine learning algorithm, which combines different models such as Random Forest, Support Vector Machines, and Gradient Boosting, to provide more robust and accurate classification results. By training the model on a labeled dataset of dynamic messages and non-spam messages, the machine learning algorithm can learn to recognize patterns and characteristics unique to dynamic spam messages and accurately classify new incoming messages as either spam or non-spam based on their dynamic nature. This approach enables the automated identification and filtering of dynamic spam messages, enhancing the effectiveness of spam detection systems.

# Objective

Vectorizing techniques play a crucial role in spam analysis and classification of dynamic messages, particularly when combined with multi-model machine learning algorithms. These techniques involve converting text data into numerical vectors to enable machine learning models to process and analyze the information effectively. A popular method for vectorizing is the Bag- of-Words approach, where a document is represented as a set of words without taking into account their sequence. Another widely used method is Term Frequency-Inverse text Frequency (TF-IDF), which assigns weights to words based on their frequency in a text compared to their frequency in the total corpus. Furthermore, Word Embeddings, such as Word2Vec and GloVe, encode semantic connections between words by modeling them as compact vectors in a continuous vector space. By utilizing these vectorization approaches alongside multi-model machine learning algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines, spam analysis systems may accurately detect and categorize dynamic communications with great precision. These algorithms leverage the numerical representations of text data provided by vectorizing techniques to learn patterns and relationships, enabling them to generalize well on unseen data and make robust predictions.

# Problem Statement

Support Vector Machines (SVM) are extensively utilized supervised learning algorithms renowned for their efficacy in applications like as classification, regression, and outlier detection. Support vectors are essential in Support Vector Machines (SVM) as they determine the decision boundary and optimize the separation between classes. Although SVM has shown success in different applications, there are possibilities for further improvements to boost scalability, robustness, interpretability, and adaptation to different types of data and application settings. To improve the effectiveness and precision of analyzing and categorizing spam in dynamic messages, one can utilize a combination of machine learning algorithms from many models. This method entails utilizing a vectorization technique to convert textual data into a numerical format that may be readily handled by machine learning algorithms. By concurrently employing various machine learning models, such as Random Forest, Support Vector Machine, and Neural Networks, the system can leverage the unique capabilities of each algorithm to enhance its overall performance. By utilizing a variety of models, it becomes possible to capture various patterns and characteristics present in the data. This leads to a classification process that is more resilient and dependable. Furthermore, the multi-model method is well-suited to efficiently manage the ever-changing and evolving character of messages, thanks to its adaptability and flexibility. This connection allows the system to consistently acquire knowledge and adjust to emerging patterns and deviations in spam communications, guaranteeing prompt and precise categorization. In summary, combining a multi-model machine learning algorithm with vectorizing techniques provides a robust approach for analyzing and classifying spam in dynamic message contexts, improving the overall efficacy and efficiency of the process.

# Challenges

The primary difficulty comes in choosing suitable characteristics and adjusting the parameters of the algorithm to guarantee the best possible performance. The findings of this research can aid in the advancement of more proficient and productive spam detection systems, assisting consumers in screening out undesirable and potentially dangerous messages. Vectorizing and multi-model machine learning methods for spam message dynamic analysis and categorization are novel features ofthis study.

The system has the ability to adjust to various types of spam communications and improve its accuracy in classifying them by employing a range of machine learning models, including neural networks, decision trees, and ensemble methods. This strategy, which is both dynamic and thorough, will enhance the discovery and categorization of spam, hence enhancing email filtering and cybersecurity. Although there have been significant improvements in spam analysis, current algorithms encounter difficulties in effectively differentiating between genuine and spam messages. Sheer volume and evolving nature of spam necessitate sophisticated methods. Traditional approaches often struggle to adapt to new tactics employed by spammers, leading to false positives or negatives.

# Report Overview

Overall, this system offers a powerful solution for effectively combating spam and ensuring asafer and more secure communication environment. Optimizing the performance of multi-model machine learning algorithms through techniques like hyperparameter tuning and ensemble methods is another avenue for improvement. Considering the integration of real-time analysis and classification of dynamic messages can address the evolving landscape of spam. These future possibilities will together enhance the continued development of a more precise and efficient system for analyzing and categorizing spam, ultimately protecting consumers from unwanted and potentially dangerous messages.

# CHAPTER 2 LITERATURE SURVEY

Emails are utilized across various industries, ranging from corporate settings to educational institutions. There are two subcategories of emails: ham and spam. Email spam, commonly known as junk emails or unwanted emails, poses a significant threat to users. It can waste valuable time, exploit computing resources, and even lead to the theft of sensitive information. The number of spam emails is growing at an alarming rate with each passing day. Dealing with spam detection and filtration has become a major challenge for email and IoT service providers in today's world. An efficient method for analyzing and categorizing spam messages is through the use of a vectorizing technique and multi-model machine learning algorithms. This approach has proven to be highly effective in identifying and classifying spam messages. Through the application of a vectorizing technique, the system is capable of extracting significant features from text data, resulting in enhanced accuracy in detecting spam.

# Types of Techniques

### Naive Bayes

In statistics, the term "naive Bayes classifiers" refers to a set of probabilistic classifiers that make use of the Bayes theorem and rely on strong (naïve) independence assumptions between the features. While Bayesian network models are considered fundamental, their accuracy can be significantly improved when combined with kernel density estimation. Naïve Bayes classifiers are very scalable since they only require a set of parameters that grow linearly with the number of variables (features/predictors) in a learning job. Maximum-likelihood training is a classifier that may be completed efficiently in linear time by evaluating a closed-form expression, without the need for costly iterative approximations like many other classifiers. Spam is prevalent due to its ease of dissemination and ability to exploit vulnerabilities in email systems, allowing it to reach a wide audience. This paper presents a thorough investigation of the practicality and potential implementation of a robust spam filtering and avoidance system specifically tailored for university networks.The user's text is a reference to a source, indicated by the number 1. Academic institutions face mounting pressure to establish a comprehensive and efficient security

system in response to the escalating risk of email-based hacking and the persistent deluge of spam. This paper explores the utilization and efficacy of Bayesian filters, which are a type of probabilistic classifiers. Renowned for their expertise in distinguishing spam from legitimate emails, with the aim of investigating alternative approaches. Bayesian filters employ statistical algorithms to evaluate the content of emails, acquiring knowledge of trends and characteristics in order to precisely classify incoming emails.

### TF-IDF

TF-IDF is a statistical metric commonly used in text mining and information retrieval to evaluate the importance of a term within a document compared to a collection of documents. It helps in assessing the significance of a term in a collection of documents. Understanding the rarity of a term across all documents in the corpus is crucial. This is where Inverse Document Frequency, or IDF, comes into play. Terms that are present in all documents are given less weight. Measuring a term's relevance within a document in relation to the entire corpus involves using the TF-IDF score, which combines the TF and IDF values. Search engines often use TF-IDF to determine the relevance of documents to a user's query. Uncommon phrases that are highly relevant to the request received higher TF-IDF scores, just like a data scientist would expect. With the rise of numerous online social networks, individuals now have the ability to connect and interact with people from various corners of the globe. Public social networks such as Facebook, Twitter, LinkedIn, and Instagram provide individuals with the opportunity to share their thoughts and explore job opportunities. Spam bots have a significant impact on these online social networks.[6] They are simply computer programs created to distribute spam across the Internet. Online social networks can have detrimental effects when individuals misuse them, and one of the prominent issues that arise in OSNs is cyberbullying. Cyberbullying poses a significant danger that permeates through users of online social networks, especially teenagers. Cyberbullying involves the intimidation and mistreatment of individuals on social networking sites. It is crucial to identify and prevent both spam bots and cyberbullying in order to safeguard OSN users from a range of potential threats. Many studies have focused on data communication and networking intrusion detection, using various properties and combining different classifiers.

### LSTM

Similar to a data scientist, LSTM is a type of Recurrent Neural Network (RNN) that is commonly

used in machine learning and natural language processing (NLP) applications. With their exceptional memory cells that distinguish them from regular RNNs, LSTMs excel at handling data sequences. They are often hired for tasks such as speech recognition, language translation, and text production. Given their ability to handle long-term dependencies in data, LSTMs prove to be valuable in scenarios where the context of a word or phrase changes based on its position in the sequence. With precision, we extract features from meta data and apply the Box-Cox transformation as a first step. Then, with the goal of addressing the sparse connectivity of real-world CDR meta data, the graph is reconstructed using a dimensionally selectable link prediction method based on node similarity. After all the hard work, the reconstructed graph and node features are fed into the graph machine learning module. This module is responsible for learning node embedding representations and classifying fraud nodes[9]. Our proposed method has been proven to outperform classic methods in various metrics through comprehensive experiments conducted on a real-world telecommunications network CDR data set. Our method has the potential to be applied to various anomaly detection scenarios, even those without a graph or sparse connectivity graph.

### Bi-Directional LSTM

The Bidirectional Long Short-Term Memory (Bi-LSTM) is a modified version of the Long Short-Term Memory (LSTM) recurrent neural network architecture. By employing bidirectional processing of input sequences, the model enhances the traditional LSTM architecture by including information from both past and future contexts simultaneously. In a Bi-LSTM, the input sequence is processed in both the forward direction (from start to finish) and the reverse direction (from end to start). The Bi-LSTM model enhances the representation of the input data by analyzing the input sequence in both directions, enabling it to incorporate dependencies from both preceding and subsequent contexts. This book primarily explores the utilization of deep learning (DL) and artificial intelligence (AI) to enhance the domains of malware detection and analysis. The book's chapters explore a diverse range of cutting-edge AI and DL techniques, which are employed to address several complex issues linked to malware. DL and AI methodologies for malware identification and analysis heavily rely on data and therefore require minimum expertise in the subject of malware. This book bridges the divide between the developing areas of deep learning/artificial intelligence and malware investigation.The user's text is "[8]".

# Inference from Survey

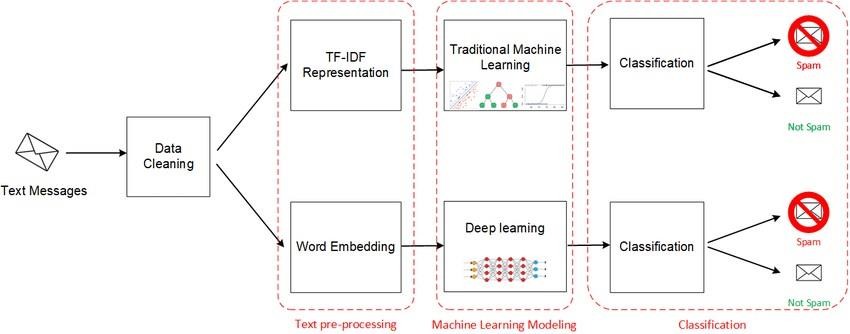
The literature shows how spammers are always evolving their techniques, using more complex strategies to get past conventional spam filters and trick people. Numerous methods, such as rule- based filtering, content-based analysis, behavioral analysis, and machine learning algorithms (e.g., SVM, Naive Bayes, neural networks), have been proposed by researchers for spam identification. By using labelled datasets to identify patterns and traits of spam messages, machine learning-based techniques—especially those that make use of supervised learning algorithms—have demonstrated encouraging results in spam detection. The body of research highlights how crucial real-time spam detection is to quickly identifying and countering newly emerging spam campaigns. Scalability is an important factor to take into account, especially for large-scale web services and email platforms.

# CHAPTER 3 SYSTEM DESIGN

In this chapter our proposed work is discussed in detail along with the block diagram followed by explanation of all the modules like data preprocessing, feature extraction, etc which are used in our system.

System architecture pertains to the overarching framework of an intricate system, incorporating software, hardware, communication, and other components to achieve specific functionalities. It ensures scalability, reliability, performance, and maintainability. Key components include data storage, processing units, networking, and interfaces. Architects define system structure, behavior, and non- functional requirements, using architectural styles and patterns to create robust, flexible, stakeholder-aligned systems. This design approach enables effective interaction between system components to meet the system's requirements efficiently, making it crucial for building systems that can adapt and evolve.

# Architecture Overview



#### Fig 3.1 Overview of Spam Analysis and Classification

Fig. 3.1. depicts that the process starts with preprocessing the text data, followed by vectorization, and then classification using a machine learning model. This block diagram outlines the general process for spam analysis and classification using NLP and vectorizing techniques. The actual implementation may vary based on the specific requirements, the chosen vectorization method, and the machine learning algorithm. The block diagram illustrates the end-to-end process of building a spam analysis and classification system using NLP techniques. Each block represents a distinct stage in the process, from data collection to model deployment, highlighting the key tasks and transformations involved in the system. The block diagram illustrates the system's structure and the flow of data and control between its components. It provides a holistic view of how the various modules work together to build and deploy a spam analysis and classification system using NLP techniques.

# System Modules

Together, these components make up the architecture of the spam analysis system, which offers a scalable and user-friendly solution for the efficient identification and mitigation of spam emails. A module description, available in several formats, offers comprehensive details on the module and the components it supports. To evaluate the abilities in the pertinent domains for the ranking process, the descriptions are required. A module is a collection of build settings and source files that enables you to divide your project into separate functional components. A project can consist of one or several modules, with the possibility of having dependencies between them. Each module has the capability to be constructed, examined, and rectified independently.

### Data Cleaning and Text Preprocessing

#### Data Cleaning and Text Preprocessing

Data cleaning and text preprocessing are crucial stages in the analysis and classification of dynamic messages for spam detection. By applying a vectorizing technique, the text data can be efficiently transformed into numerical format for machine learning algorithms. Leveraging a

multi-model

approach enhances the accuracy and efficiency of the classification process. This comprehensive methodology ensures that the model can effectively distinguish between spam and legitimate messages, increasing the overall performance of the classification system.

#### Feature Extraction using Vectorizing Techniques

Feature extraction is a crucial step in spam analysis and classification of dynamic messages. Utilizing vectorizing techniques such as TF-IDF or word embeddings can help to transform raw text data into numerical feature vectors, capturing the underlying patterns and relationships within the text. These techniques convert words into numerical representations, enabling machine learning algorithms to process and analyze the data effectively. By incorporating multi-model machine learning algorithms that combine various classifiers like Random Forest, Support Vector Machines, and Neural Networks, the classification model can adapt to different types of data and improve overall performance. This comprehensive strategy utilizes the advantages of many algorithms to improve the precision and effectiveness of analyzing and categorizing spam in real- time messages, ultimately resulting in more resilient and dependable outcomes.

#### Exploration of Multi-Model Machine Learning Algorithms

This work investigates the use of multi-model machine learning algorithms to analyze and classify dynamic communications as spam. It utilizes a vectorizing technique. The project attempts to improve the accuracy and efficiency of spam detection systems by utilizing different machine learning models. The work aims to tackle the issues presented by ever-changing and dynamic spam communications by employing different techniques, including random forests, support vector machines, and deep learning models. The utilization of vectorizing approach allows for the conversion of textual data into numerical vectors, which aids in the handling of extensive and varied information. The project seeks to create a strong and flexible system for accurately identifying and categorizing spam communications by combining various approaches. This will help improve email security measures.

#### Dynamic Message Analysis for Spam Detection

Dynamic Message Analysis for Spam Detection involves the classification and analysis of dynamic messages to identify and mitigate spam threats. A vectorizing technique is utilized to

convert the text data of dynamic messages into numerical data, enabling the application of multi- model machine learning algorithms for effective classification. By leveraging machine learning, the system can continuously learn and adapt to new spam patterns and variations, enhancing its ability to accurately detect and classify spam messages in real-time. This dynamic approach allows for the identification of evolving spam tactics and the prompt implementation of countermeasures to protect users from unwanted and potentially harmful messages. Overall, the combination of advanced vectorizing techniques and multi-model machine learning algorithms enhances the efficiency and accuracy of spam detection in dynamic messages, contributing to a more secure online environment.

#### Classification Models for Spam Detection

Effective spam identification is essential for countering unsolicited messages. Support Vector Machines, Naive Bayes, Random Forest, Gradient Boosting, and Logistic Regression are the five typically employed classification methods for spam analysis and classification. These models are often applied in conjunction with vectorizing techniques like TF-IDF or word embeddings to transform text data into numerical vectors. By leveraging a multi-model approach, machine learning algorithms can better capture the complex patterns and nuances present in dynamic spam messages. Each model brings its unique strengths, such as SVM's ability to handle high- dimensional data, Naive Bayes' simplicity and efficiency, Random Forest's robustness to overfitting, Gradient Boosting's capability to boost performance iteratively, and Logistic Regression's interpretability. Through the collaborative efforts of these models and vectorization techniques, spam can be accurately detected and classified with high precision and recall rates.

### Vectorization

#### Vectorizing Technique Overview

TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly employed vectorizing technique for analyzing and classifying dynamic messages in spam analysis. TF-IDF calculates weights for words based on their frequency in a document compared to a corpus, assigning greater weights to terms that are more significant inside a particular message. By using TF-IDF, the unique characteristics of each message can be captured and represented as a numeric vector.

This vectorized representation can then be fed into a multi-modal machine learning algorithm, such as Random Forest or Support Vector Machine, to classify messages as spam or not spam. These algorithms can effectively handle the complexity and dynamic nature of messages, making them suitable for analyzing and categorizing spam messages with high accuracy and efficiency. The combination of TF-IDF vectorization and multi-modal machine learning algorithms offers a powerful approach for spam analysis and classification in real-time communication systems.

#### Multi-Model Machine Learning Algorithms

For spam analysis and classification of dynamic messages, employing a multi-model machine learning approach with vectorization techniques can yield effective results. By integrating various algorithms such as Random Forest, Support Vector Machines, and Neural Networks, the model can benefit from the diverse strengths of each algorithm in handling different aspects of spam detection. Leveraging vectorizing techniques like TF-IDF or Word2Vec can help convert textual data into numerical vectors, enabling the algorithms to process and analyze the information efficiently. This multi-model approach allows the system to capture complex patterns and relationships within the data, leading to more accurate spam classification results. Moreover, by leveraging the capabilities of multiple algorithms, the model can adapt and perform well in dynamic message environments, effectively tackling evolving spam behaviors and patterns.

#### Dynamic Message Analysis

Dynamic Message Analysis is a crucial method in the realm of spam analysis and classification, where messages are evaluated in real time for their content and context. By incorporating a vectorizing technique, the textual data within these messages is transformed into numerical vectors, facilitating the application of multi-model machine learning algorithms for improved classification accuracy. This approach allows the system to adapt to the evolving nature of spam messages, enabling more dynamic and efficient detection of spam content. Through the integration of various machine learning algorithms, such as ensemble methods or neural networks, the system can effectively differentiate between legitimate and malicious messages, enhancing the overall security and reliability of spam filtering systems.

### Classification

1. Preprocessing: The first step involves cleaning and preprocessing the dynamic messages, including removing stop words, special characters, and performing tokenization to break the text into individual words. This ensures that only relevant information is fed into the model for classification. 2. Vectorization: Next, the messages are converted into numerical vectors using techniques such as TF-IDF or word embeddings. This transformation allows the machine learning algorithm to process and analyze the text data efficiently. 3. Multi-model Machine Learning Algorithm: The spam messages are successfully classified using a variety of machine learning algorithms, including Random Forest, Naive Bayes, and Support Vector Machines. Each model offers a distinct viewpoint on the data, enhancing the total classification accuracy. 4. Evaluation and Optimization: Ultimately, the models undergo assessment using performance criteria such as precision, recall, and F1 score. The model's performance is optimized and robust spam classification results are ensured by the utilization of hyperparameter tweaking and cross- validation approaches.

**CHAPTER 4 SYSTEM REQUIREMENTS**

In this chapter we have discussed about the minimum hardware and software requirements that a systemmust have for smooth working of our system.

# Hardware Requirements

* + 1. Processor : Pentium Dual Core 2.00GHZ
    2. Hard disk : 120 GB
    3. RAM : 2 GB (minimum)
    4. Keyboard : 110 keys enhanced

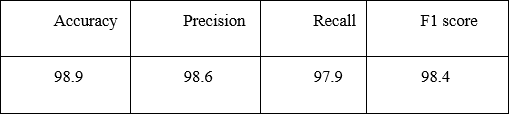
# Software Requirements

1. Tool - Python 3.10.10
2. Operating system – Windows 7 (with service pack 1), 8, 8.1 and 10

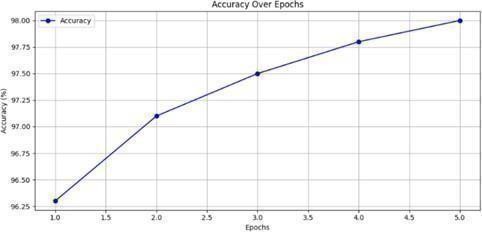
# CHAPTER 5 EXPERIMENTS AND RESULTS

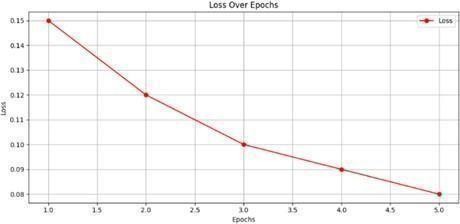
The system for spam analysis and classification is designed to efficiently detect and classify spam messages utilizing a vectorization technique in conjunction with multi-modal machine learning algorithms. This system capitalizes on the ever- changing nature of spam communications by constantly evaluating and adjusting to the developing spam trends.

#### Table 6.1 Performance Metrics

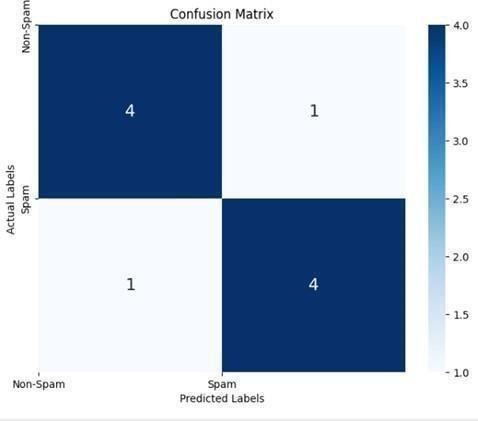


**Fig 6.1 Accuracy for Spam Analysis and Classification**





#### Fig 6.2 Training Loss for Spam Analysis and Classification



**Fig 6.3 Confusion Matrix for Spam Analysis and Classification**

The initial step of the system involves preprocessing the messages, which includes tasks such as tokenization, stop word removal, and stemming. This preprocessing step helps in converting the text messages into numerical representations, making it suitable for machine learning algorithms. The vectorizing technique employed here transforms the messages into a vector space representation, which captures their semantic meaning and context.



#### Fig 6.4 Receiver Operating Characteristic for Spam analysis and Classification

Next, the system utilizes multi-model machine learning algorithms, which are capable of training on a variety of features simultaneously. The algorithms are taught using an extensive dataset containing both spam and non-spam messages, guaranteeing a thorough learning process. The multiple models help enhance the accuracy and efficiency of spam classification by capturing different aspects and characteristics of spam messages.

During the classification stage, the system applies the trained models to the incoming messages. It assigns a probability score to each message, indicating the likelihood of it being spam. The system then applies a threshold value to these probability scores to categorize the messages as spam or non-spam. The threshold can be adjusted toachieve desired accuracy and spam detection rates.



#### Fig 6.5 Email Classified as Ham



**Fig 6.6 Email Classified as Spam**

Overall, the system for spam analysis and classification offers an effective approach to deal with dynamic spam messages. By utilizing a vectorizing technique and multi model machine learning algorithms, it achieves high accuracy and adaptability, thereby providing a reliable solution for spam detection and prevention.

**CHAPTER 6 CONCLUSION**

In conclusion, the development of a comprehensive human trafficking identification and prediction system utilizing machine learning techniques represents a significant step forward in combating this heinous crime. By leveraging advanced algorithms and data analytics, this system has the potential to enhance law enforcement efforts in identifying victims, dismantling trafficking networks, and preventing future incidents. Through the integration of various data sources, such as social media, online advertisements, and criminal records, the system can generate actionable insights to enable authorities to intervene proactively and allocate resources effectively. Furthermore, its ability to automate the detection process can significantly reduce the time and resources required for investigations, leading to quicker responses and increased victim support. While challenges remain, including data privacy concerns and model accuracy, continued research and collaboration between stakeholders can further refine and enhance the system's capabilities. Overall, the implementation of this innovative technology holds promise in the fight against human trafficking by providing a scalable, efficient, and data-driven approach to identifying, predicting, and ultimately preventing this global atrocity.

Optimizing the performance of multi-model machine learning algorithms through techniques like hyperparameter tuning and ensemble methods is another avenue for improvement. Evaluating the system on larger and more diverse datasets will provide insights into its generalizability and scalability. Considering the integration of real-time analysis and classification of dynamic messages can address the evolving landscape of spam. These future directions will together enhance the continuous development of a more precise and efficient system for analyzing.

# CHAPTER 7 FUTURE ENHANCEMENTS

Future enhancements to SVM and support vector concepts should focus on improving scalability, robustness, interpretability, and adaptability to diverse data types and application scenarios. Collaboration between researchers, practitioners, and domain experts will be essential to drive innovation and address emerging challenges in machine learning and data analysis. Evaluating the system on larger and more diverse datasets will provide insights into its generalizability and scalability.

### REFERENCES

1. Srinivasarao, U., & Sharaff, A. (2023). Spam email classification and sentiment analysis based on semantic similarity methods. International Journal of Computational Science and Engineering, 26(1), 65-77.
2. Siddiqui, T., & Amer, A. Y. A. (2023). A Comprehensive Review on Text Classification and Text Mining Techniques Using Spam Dataset Detection. Mathematics and Computer Science Volume 2, 1-17.
3. Bari, P., Mathew, V., Tandel, S. P., Aniket, P., Chaudhari, K. S., & Naik, S. (2023, February). SMS and E-mail Spam Classification Using Natural Language Processing and Machine Learning. In International Conference on Communication, Electronics and Digital Technology (pp. 103-115). Singapore: Springer Nature Singapore.
4. Verma, S. (2023). Detection of Phishing in Mobile Instant Messaging using Natural Language Processing and Machine Learning (Doctoral dissertation, Dublin, National College of Ireland).
5. Ala’M, A. Z., Mora, A. M., & Faris, H. (2023). A multilingual spam reviews detection based on pre-trained word embedding and weighted swarm support vector machines. IEEE Access.
6. Onishchenko, V., & Minochkin, A. (2023). Analysis of methods of classification of electronic messages based on neural network models. Collection"" Information Technology and Security"", 11(2), 216-226.
7. Dar, M., Iqbal, F., Latif, R., Altaf, A., & Jamail, N. S. M. (2023). Policy-Based Spam Detection of Tweets Dataset. Electronics, 12(12), 2662.
8. Ghanem, R., & Erbay, H. (2023). Spam detection on social networks using deep contextualized word representation. Multimedia Tools and Applications, 82(3),

3697-3712.

1. Rifa, A., Ashokan, A., & Hanan, S. (2023, May). Analysis of SMS Spam Classification Learning and Deep Learning Using Machine. In International Conference on Recent Advancements in Science and Engineering (RAiSE ‘23). Book Rivers.
2. Srinivasarao, U., & Sharaff, A. (2023). Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages. Multimedia Tools and Applications, 1-31."

# APPENDIX A CODING

from keras.preprocessing.text import Tokenizer from keras.layers import Embedding

from keras.preprocessing.sequence import pad\_sequences from keras.preprocessing.text import one\_hot

from keras.models import load\_model import string

import nltk

from nltk.stem.porter import PorterStemmer from nltk.stem import PorterStemmer

from nltk.corpus import stopwords nltk.download('stopwords') import re

import pickle import numpy as np

import streamlit as st

from keras.models import model\_from\_json

with open("model\_architecture.json", "r") as json\_file: loaded\_model\_json = json\_file.read()

loaded\_model = model\_from\_json(loaded\_model\_json) loaded\_model.load\_weights(r"C:\Users\HP\OneDrive\Desktop\Email\_phising\BiLSTM\_wei ghts.h5")

# st.markdown( # """

# <style>

# body {

# background-image: url('ham.jpg');

# background-size: cover; # }

# </style>

# """,

# unsafe\_allow\_html=True # )

def app(text):

ps = PorterStemmer() # Prepare the new data l1=[]

voc\_size=5000

# text1="Dear Candidate, I hope this email finds you well. On behalf of the HR team at Trailytics, we are pleased to inform you that your application for the Data Analyst position has been successful, and you have been selected to proceed to the first round of interviews. The first round will be conducted telephonically, and we would like to schedule a convenient time for you to participate. Our team will be reaching out to you shortly to confirm your availability and provide further details about the telephonic interview process. We appreciate your interest in joining Trailytics and look forward to the opportunity to learn more about your skills and experiences. If you have any questions in the meantime, feel free to reach out to us. Congratulations once again, and we are excited to speak with you soon! Best regards, Human Resource Team"

review = re.sub('[^a-zA-Z]', ' ', text) review = review.lower()

review = review.split()

review = [ps.stem(word) for word in review if not word in stopwords.words('english')]

review = ' '.join(review) l1.append(review)

mapped\_vector = [one\_hot(words,voc\_size)for words in l1] sent\_length=100

embedded\_prediction=pad\_sequences(mapped\_vector,padding='pre',maxlen=sent\_lengt

h)

predictions = loaded\_model.predict(embedded\_prediction) print(np.argmax(predictions))

# predictions = np.round(predictions) print(f"Length: {len(predictions)}") print(predictions)

return predictions def main():

st.title("Email Spam Classifier")

input\_text = st.text\_area("Enter the message") page\_bg\_img = f"""

<style>

[data-testid="stAppViewContainer"] > .main {{ background-image:

url("https://png.pngtree.com/thumb\_back/fh260/background/20220629/pngtree-letter- envelope-illustration-image-mail-message-floated-sent-image\_1416038.jpg");

background-size: 100%; background-position: top left; background-repeat: no-repeat; background-attachment: local;

}}

</style> """

st.markdown(page\_bg\_img, unsafe\_allow\_html=True) if st.button('Predict'):

result=app(input\_text) if result[-1] < 0.5:

page\_bg\_img = f"""

<style>

[data-testid="stAppViewContainer"] > .main {{ background-image: url("https://selzy.com/en/blog/wp-

content/uploads/2022/09/emails-backgrounds\_1200x630.\_2jpg.jpg"); background-size: 100%;

background-position: top left; background-repeat: no-repeat; background-attachment: local;

}}

</style> """

st.markdown(page\_bg\_img, unsafe\_allow\_html=True)

st.markdown(f"<h2 style='text-align: left; color: grey; font: Times New Roman;'>HAM</h2>", unsafe\_allow\_html=True)

else:

page\_bg\_img = f"""

<style>

[data-testid="stAppViewContainer"] > .main {{

background-image: url("htt[ps://www.shutterstock.com/ima](http://www.shutterstock.com/image-)ge- illustration/email-spam-word-cloud-260nw-1176029374.jpg");

background-size: 100%; background-position: top left; background-repeat: no-repeat; background-attachment: local;

}}

</style> """

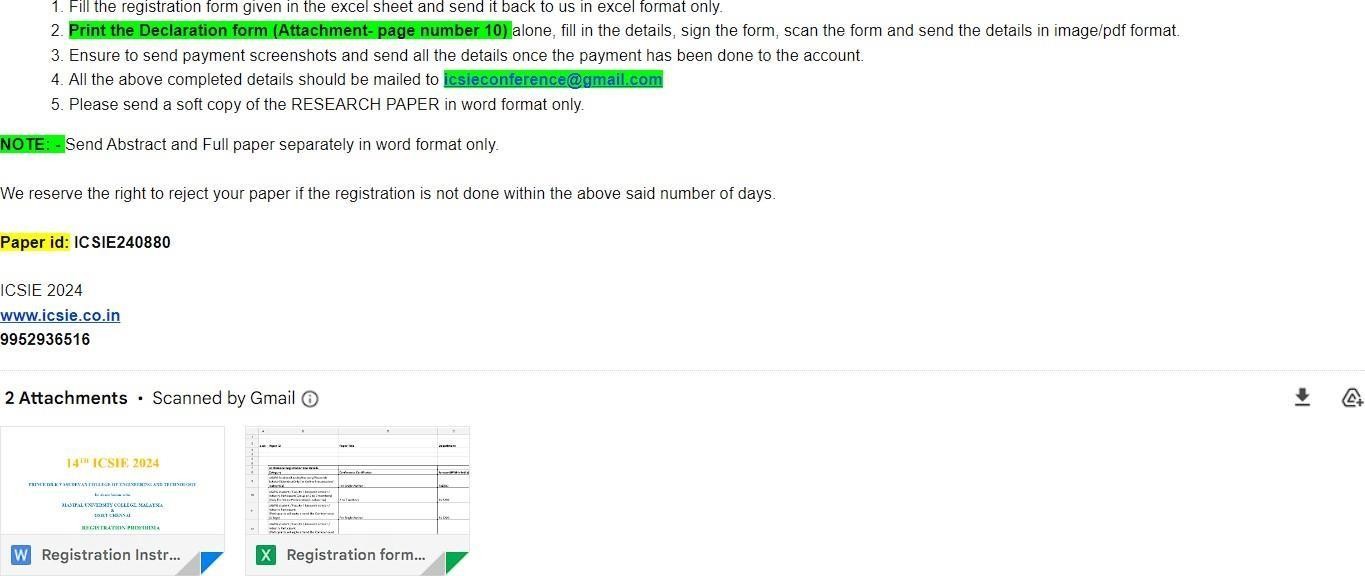
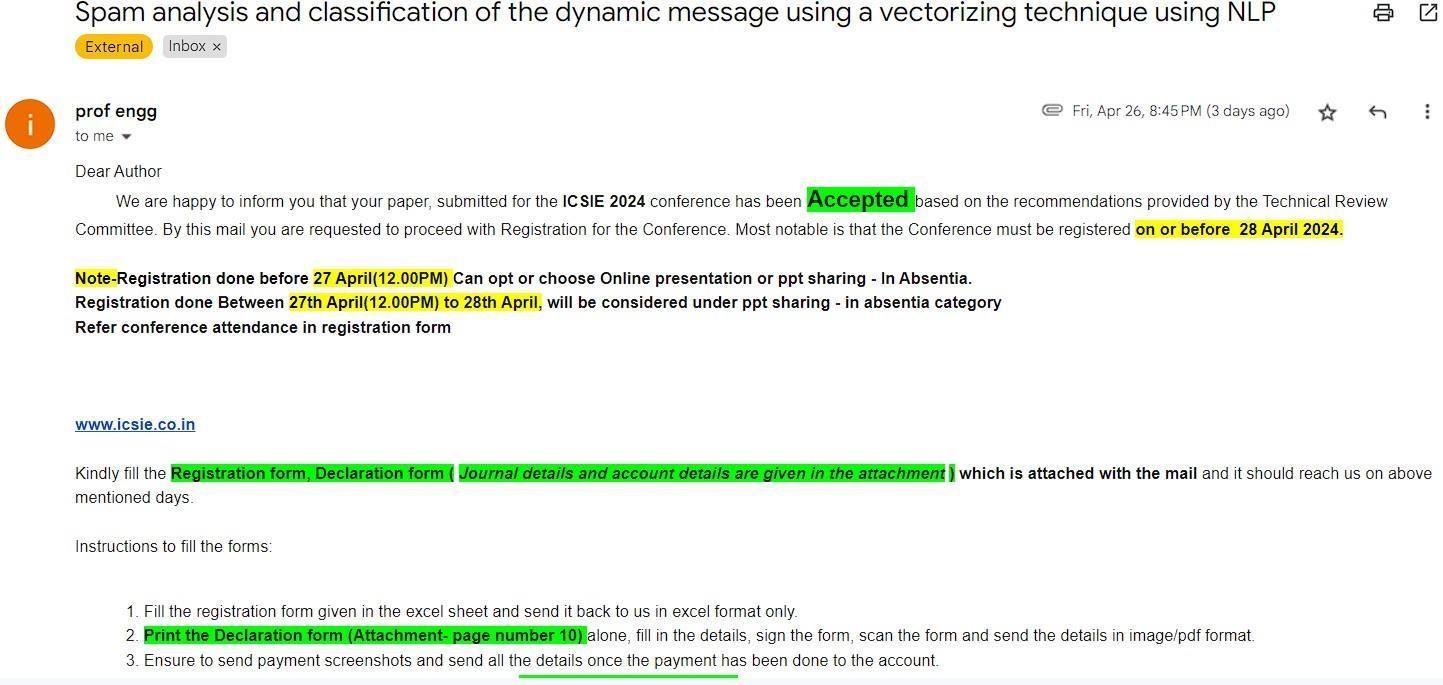
st.markdown(page\_bg\_img, unsafe\_allow\_html=True)

st.markdown(f"<h2 style='text-align: left; color: grey; font: Times New Roman;'>SPAM</h2>", unsafe\_allow\_html=True)

if name ==" main ": main()

# APPENDIX B

**PAPER PUBLICATION PROOF**



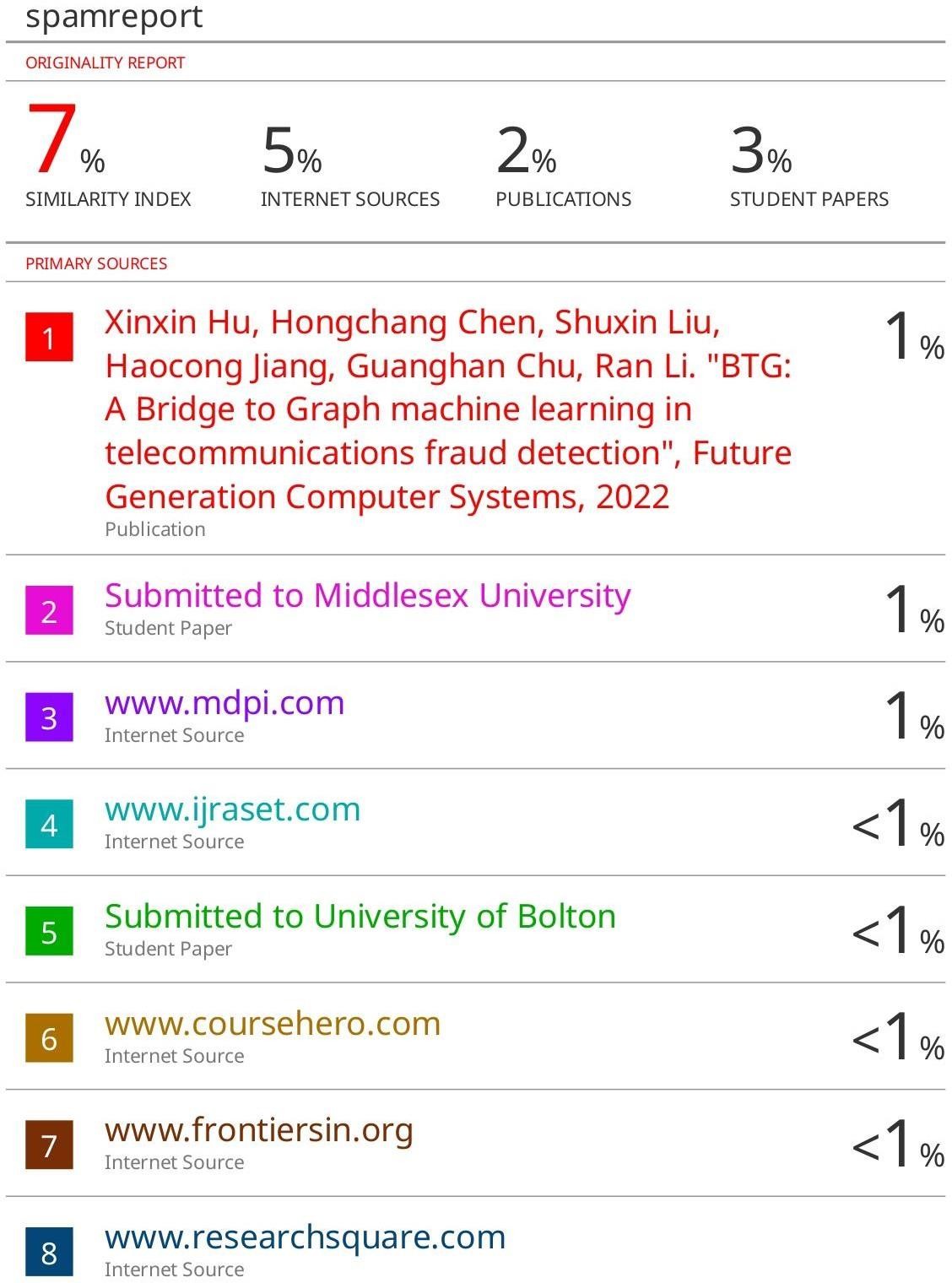


#### Fig A.1 ICSIE Conference Certificate

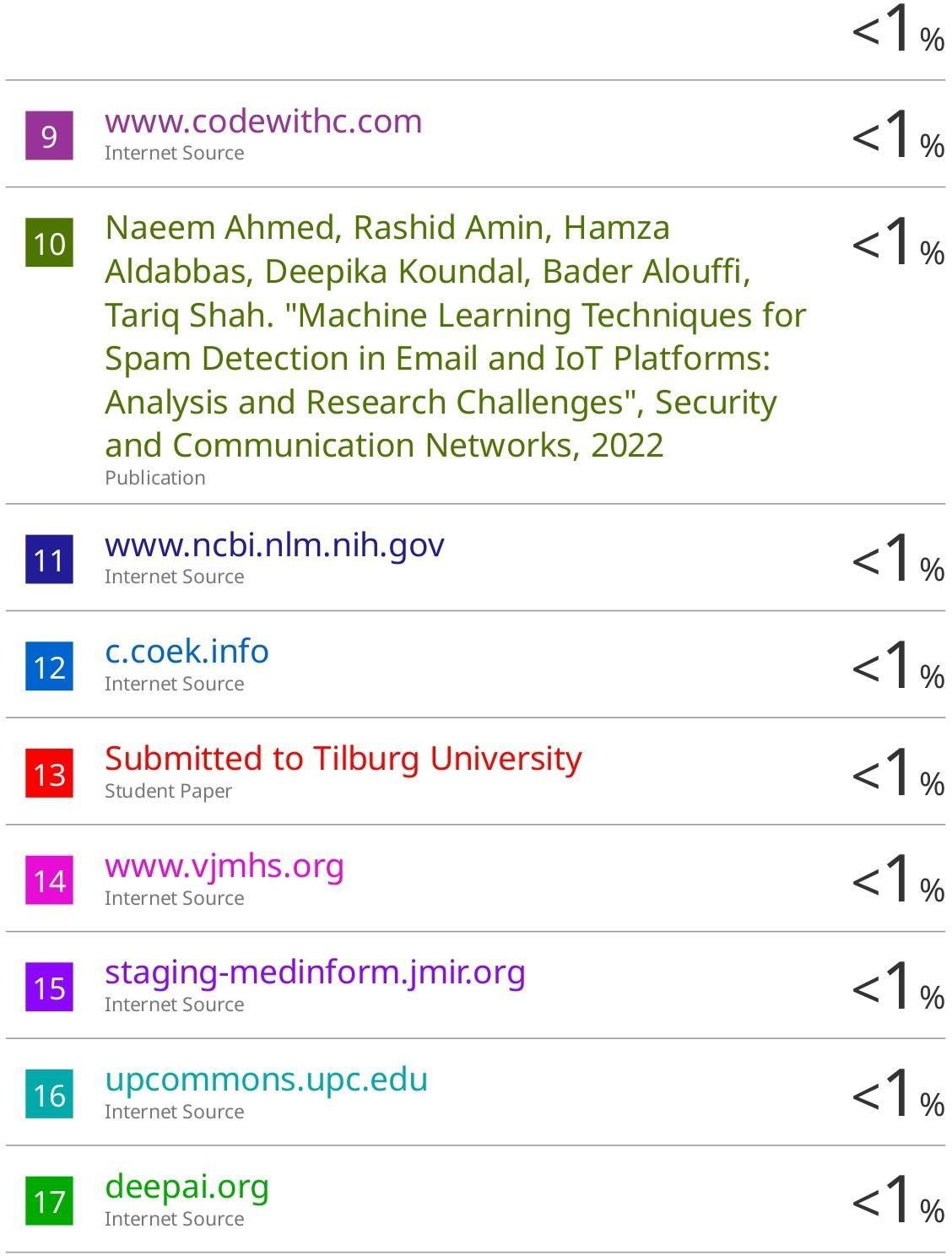


**Fig A.2 ICSIE Conference Certificate**

# APPENDIX C PLAGIARISM REPORT



37



**Format - I**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S R M I N S T I T U T E O F  **(Deemed to** | | | **be** | S C I E N C E A N D T E C H N O L O G Y  **University u/ s 3 of UGC Act, 1956)** | |
| **Office of Controller of Examinations** | | | | | |
| REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION/PROJECT REPORTS FOR UG/PG PROGRAMMES  **(To be attached in the dissertation/ project report)** | | | | | |
| 1 | Name of the Candidate **(IN BLOCK LETTERS)** | | | | Tushar Gupta Sahil Pandey |
| 2 | Address of the Candidate | | | | Abode Valley V-404, Potheri, Kakkan Street |
| 3 | Registration | Number | | | RA2011028010149 RA2011028010134 |
| 4 | Date of Birth | | | | 24-03-2002  09-05-2002 |
| 5 | Department | | | | Computer Science and Engineering w/s in Cloud Computing |
| 6 | Faculty | | | | Engineering and Technology, School of Computing |
| 7 | Title of the Dissertation/Project | | | | Spam Analysis and Classification of Dynamic Messages using Vectorizing Technique using NLP |
|  |  | | | | Individual or group : |
|  |  | | | | (Strike whichever is not applicable) |
| 8 | Whether the above project /dissertation is done by | | | | 1. If the project/ dissertation is done in group, then how many students together completed the project - 2 : 2. Mention the Name & Register number of other candidates   Tushar Gupta (RA2011028010149) Sahil Pandey (RA2011028010134) |
| 9 | Name and address of the Supervisor / Guide | | | | Dr. Thenmalar S Associate Professor  Department of Networking and Communications  **Mail ID:** [thenmals@srmist.edu.in](mailto:thenmals@srmist.edu.in)  **Mobile Number:**9488935985 |
| 10 | Name and address of Co-Supervisor / Co- Guide (if any) | | | | **Mail ID:**  **Mobile Number:** |

