ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

INVESTIGATING EVASIVE TECHNIQUES IN SMS SPAM FILTERING

Dr P.Vishvapathi [1], Bushra Fatima[2]
[1]Professor, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, India
[2]M.Tech Student, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, India

ABSTRACT

The rapid growth of mobile communication has led to a significant increase in SMS spam, posing serious security and privacy risks to users worldwide. Modern spammers continuously employ evasive techniques such as word obfuscation, character substitution, spacing manipulation, and synonym replacement to bypass conventional spam detection systems, making accurate classification increasingly challenging. This paper presents an intelligent SMS spam filtering framework that investigates the effectiveness of multiple machine learning models in detecting spam messages under adversarial conditions. The proposed framework combines comprehensive text preprocessing, Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction, and comparative evaluation of traditional and deep learning classifiers, with particular emphasis on the Long Short-Term Memory (LSTM) network due to its ability to capture contextual and sequential information from textual data. The system is trained and evaluated using a large, balanced SMS dataset containing both legitimate and spam messages collected from diverse sources. Standard performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis, are used to assess the robustness of each model against evolving spam patterns. Experimental results indicate that deep learning models, particularly LSTM, outperform conventional machine learning techniques in identifying both traditional and evasive spam messages while maintaining high classification accuracy and generalization capability. The proposed framework offers a scalable, reliable, and adaptable solution for intelligent SMS spam filtering and provides valuable insights into developing next-generation spam detection systems capable of addressing continuously evolving adversarial messaging techniques.

Index Terms— SMS Spam Detection, Machine Learning, Deep Learning, Long Short-Term Memory (LSTM), Naïve Bayes, TF-IDF, Text Classification, Evasive Techniques, Adversarial Spam, Natural Language Processing (NLP), Spam Filtering, Feature Extraction, Cybersecurity, Message Classification, Artificial Intelligence.

1. INTRODUCTION

1.1 Background

The widespread use of mobile devices has made Short Message Service (SMS) one of the most common communication channels for personal, commercial, and financial interactions. However, this popularity has also attracted cybercriminals who exploit SMS as a medium for distributing spam, phishing links, fraudulent advertisements, and malicious content. SMS spam not only causes inconvenience to users but also poses significant cybersecurity risks, including financial fraud, identity theft, and unauthorized access to sensitive information.

Over the past decade, numerous machine learning and deep learning techniques have been developed to automatically classify SMS messages as legitimate (ham) or spam. Traditional machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and Random Forest have demonstrated promising results by learning patterns from historical SMS datasets. More recently, deep learning models, particularly Long Short-Term Memory (LSTM) networks, have gained considerable attention due to their ability to capture contextual and sequential relationships within textual data, resulting in improved spam detection accuracy.

Despite these advancements, modern spammers continuously adapt their strategies to evade detection systems. They intentionally modify message content using techniques such as character substitution, deliberate misspellings, insertion of unnecessary spaces, synonym replacement, punctuation manipulation, and special characters. These evasive techniques significantly reduce the effectiveness of conventional spam filters because even small textual modifications can alter the

extracted features without changing the actual intent of the message.

Another major challenge in SMS spam detection is the dynamic evolution of spam campaigns. As attackers frequently introduce new message formats, promotional tactics, and phishing strategies, models trained on outdated datasets often experience performance degradation over time. Therefore, developing robust spam filtering systems capable of adapting to evolving spam patterns has become an important research problem.

In this study, an intelligent SMS spam filtering framework is proposed to investigate the effectiveness of different machine learning models against modern evasive spam techniques. The proposed approach incorporates comprehensive text preprocessing, TF-IDF feature extraction, and comparative evaluation of multiple machine learning algorithms, with particular emphasis on the Long Short-Term Memory (LSTM) network for sequence learning. By analyzing the strengths and limitations of both traditional and deep learning models, the proposed framework aims to improve spam detection accuracy, robustness, and adaptability in real-world environments.

Overall, this research contributes toward the development of reliable and scalable SMS spam filtering systems that can effectively identify evolving spam messages while minimizing false classifications and improving mobile communication security.

1.2 Research Gap

Although considerable progress has been made in SMS spam detection using machine learning and deep learning techniques, several important challenges remain unresolved. Most existing spam filtering systems have been developed and evaluated using relatively small or outdated datasets that do not accurately represent the continuously evolving nature of modern SMS spam. Consequently, many trained models struggle to generalize when exposed to newly emerging spam patterns.

Traditional machine learning approaches mainly rely on statistical text features such as Bag-of-Words (BoW) and TF-IDF representations. While these methods perform well for standard spam messages, they are often vulnerable to adversarial or evasive techniques employed by attackers. Minor modifications such as replacing letters with numbers, inserting extra spaces, or using alternative spellings can significantly reduce the effectiveness of conventional classifiers because these changes alter feature representations without changing the semantic meaning of the message.

Deep learning models have demonstrated improved contextual understanding; however, many previous studies focus on evaluating a limited number of classifiers without providing comprehensive comparisons between traditional machine learning methods and modern deep learning architectures under adversarial conditions. Furthermore, several studies primarily report classification accuracy while giving limited attention to model robustness, precision, recall, F1-score, and resistance to evolving spam strategies.

Another limitation observed in existing research is the lack of standardized evaluation frameworks capable of assessing model performance against different categories of evasive techniques. As spam campaigns continue to evolve rapidly, there is a growing need for adaptive models that maintain consistent performance over diverse and unseen attack scenarios.

This research addresses these limitations by performing a comparative analysis of multiple machine learning models for SMS spam detection while specifically investigating their ability to handle evasive spam techniques. The proposed framework combines efficient text preprocessing, TF-IDF feature extraction, and deep learning using LSTM to improve contextual understanding and classification performance. The study also evaluates the models using multiple performance metrics to provide a comprehensive assessment of their effectiveness in real-world spam filtering applications.

1.3 Research Objectives

The primary objectives of this research are:

1. To develop an intelligent SMS spam filtering system capable of accurately classifying legitimate and spam messages.
2. To investigate the impact of various evasive techniques used by spammers on the performance of machine learning-based spam detection systems.
3. To implement effective text preprocessing and TF-IDF feature extraction techniques for representing SMS messages.
4. To compare the performance of traditional machine learning algorithms and deep learning models for SMS spam classification.
5. To evaluate the effectiveness of the Long Short-Term Memory (LSTM) network in detecting both conventional and evasive spam messages.

6. To analyze the performance of different models using accuracy, precision, recall, F1-score, and confusion matrix evaluation metrics.
7. To improve the robustness and adaptability of SMS spam filtering systems against continuously evolving spam strategies.
8. To provide an efficient, scalable, and reliable framework that enhances mobile communication security through intelligent spam detection.

1.4 Limitations of the Study

Although the proposed SMS spam filtering framework demonstrates promising performance, several limitations should be considered.

The effectiveness of the proposed system largely depends on the quality and diversity of the training dataset. If newly emerging spam patterns are not represented in the dataset, the model may experience reduced classification accuracy when deployed in real-world environments.

The current study focuses exclusively on English-language SMS messages. Messages written in multiple languages, containing regional slang, abbreviations, or code-mixed text may require additional preprocessing techniques and language-specific training datasets for accurate classification.

While the proposed framework investigates several common evasive techniques, attackers continuously develop new methods to bypass spam filters. As a result, periodic retraining and dataset updates are necessary to maintain model performance against evolving threats.

The study primarily evaluates text-based SMS messages and does not consider multimedia messages (MMS), embedded images, QR codes, shortened URLs, or other advanced phishing techniques that are increasingly used in modern mobile attacks.

Furthermore, although the LSTM model provides superior contextual learning, deep learning models generally require greater computational resources, longer training time, and larger datasets compared to conventional machine learning algorithms. These computational requirements may limit deployment on resource-constrained devices without appropriate optimization.

Finally, the current implementation focuses on comparative model evaluation rather than real-time deployment in commercial mobile communication networks. Additional large-scale testing is required to evaluate system performance under real-world operational conditions.

1.5 Rationale of the Study

The increasing dependence on mobile communication has made SMS an essential medium for exchanging personal, financial, and business information. At the same time, the rapid growth of SMS spam and phishing attacks has created significant challenges for individuals, organizations, and service providers. Traditional spam filtering techniques are becoming less effective as attackers continuously modify message content using sophisticated evasive strategies designed to bypass automated detection systems.

The motivation for this research is to develop a robust and intelligent SMS spam filtering framework capable of identifying both conventional and adversarial spam messages with high accuracy. By comparing multiple machine learning algorithms and deep learning techniques, the study aims to determine the most effective approach for handling evolving spam patterns while minimizing false positives and false negatives.

The proposed framework utilizes comprehensive text preprocessing, TF-IDF feature extraction, and Long Short-Term Memory (LSTM) networks to improve contextual understanding and message classification. Through a detailed comparative analysis, the research identifies the strengths and weaknesses of different machine learning models when exposed to various evasive techniques commonly used by modern spammers.

In addition to improving spam detection performance, this study contributes to the broader field of cybersecurity by providing insights into developing adaptive and scalable text classification systems. The findings can assist researchers, developers, and mobile service providers in designing next-generation spam filtering solutions capable of protecting users from increasingly sophisticated SMS-based cyber threats. The proposed framework also establishes a foundation for future research on adversarial machine learning, intelligent text classification, and secure mobile communication systems.

2. LITERATURE REVIEW

The rapid growth of mobile communication has significantly increased the volume of SMS messages exchanged worldwide. While SMS remains an essential communication medium for personal, commercial, and financial activities, it has also become a popular target for cybercriminals

who distribute spam, phishing attempts, fraudulent advertisements, and malicious links. As attackers continue to adopt increasingly sophisticated evasion techniques, conventional spam filtering systems struggle to maintain high detection accuracy. Consequently, researchers have explored various machine learning, deep learning, and natural language processing (NLP) techniques to improve the effectiveness of SMS spam detection.

This literature review examines the evolution of SMS spam filtering techniques, including traditional machine learning approaches, deep learning models, adversarial spam detection, feature extraction methods, and comparative studies that collectively provide the foundation for the proposed research.

2.1 Traditional Machine Learning-Based SMS Spam Detection

Early SMS spam filtering systems primarily relied on traditional machine learning algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Logistic Regression. These models classify SMS messages by learning statistical patterns from labeled datasets using manually extracted textual features.

Common feature extraction techniques include:

- Bag-of-Words (BoW)
- Term Frequency–Inverse Document Frequency (TF-IDF)
- N-gram analysis
- Word frequency statistics

Although these methods provide fast and computationally efficient classification, they have several limitations:

a) Limited understanding of contextual meaning

b) Dependence on manually engineered features

c) Poor performance against modified or obfuscated spam messages

Key Insight: Traditional machine learning algorithms provide satisfactory performance for conventional SMS spam detection but lack robustness against evolving spam strategies and adversarial text modifications.

2.2 Natural Language Processing (NLP)-Based SMS Spam Detection

The advancement of Natural Language Processing (NLP) has significantly improved the ability of spam detection systems to understand textual information beyond simple keyword matching. NLP-based techniques focus on extracting meaningful linguistic features from SMS messages, enabling better semantic understanding and classification.

These systems typically employ:

a. Text preprocessing and normalization

b. Tokenization and stemming

c. Stop-word removal

d. Semantic feature extraction

While NLP-based approaches improve detection accuracy, they also present several challenges:

1. Dependence on high-quality text preprocessing
2. Difficulty in handling multilingual and code-mixed messages
3. Limited capability to detect newly emerging spam patterns

Key Insight: NLP techniques enhance feature representation and semantic understanding but require continuous adaptation to evolving spam content.

2.3 Deep Learning Approaches for SMS Spam Filtering

Recent studies have increasingly adopted deep learning models for SMS spam detection due to their superior ability to learn contextual and sequential relationships within text data. Models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and transformer-based architectures have demonstrated remarkable improvements over traditional machine learning methods.

Among these models, LSTM has gained considerable attention because it effectively captures long-term dependencies in sequential text while preserving contextual information.

Advantages of deep learning models include:

- Automatic feature learning
- Better contextual understanding
- Higher classification accuracy
- Improved handling of complex sentence structures

However, several limitations remain:

- High computational requirements

- Longer training time
- Dependence on large labeled datasets
- Reduced interpretability compared to conventional algorithms

Key Insight: Deep learning significantly improves SMS spam detection performance, with LSTM demonstrating strong capability in identifying complex spam patterns.

2.4 Evasive Techniques in SMS Spam Detection

Modern spammers continuously develop adversarial strategies designed to bypass automated spam filters. Instead of changing the meaning of messages, attackers modify their appearance by introducing subtle textual alterations that confuse machine learning models.

Common evasive techniques include:

- Character substitution (e.g., "Fr33" instead of "Free")
- Deliberate misspellings
- Insertion of unnecessary spaces
- Synonym replacement
- Punctuation manipulation
- Use of special characters and emojis

These modifications often preserve the semantic meaning of the message while significantly altering its textual representation.

Challenges include:

- Reduced feature consistency
- Increased false-negative rates
- Poor model generalization
- Difficulty in detecting adversarial examples

Key Insight: Evasive techniques remain one of the most significant challenges in SMS spam detection, requiring more robust and adaptive learning models.

2.5 Feature Extraction Techniques for SMS Classification

Feature extraction plays a crucial role in determining the effectiveness of machine learning models for SMS spam filtering. It transforms raw textual messages into numerical representations that can be processed by classification algorithms.

Widely used feature extraction techniques include:

- Bag-of-Words (BoW)
- TF-IDF
- Word2Vec
- GloVe
- FastText
- Contextual embeddings

Among these methods, TF-IDF remains one of the most popular due to its simplicity, computational efficiency, and effectiveness in identifying discriminative words.

However, feature extraction methods also face challenges:

- Sparse feature representation
- Loss of contextual relationships
- Difficulty representing semantic similarity
- Reduced effectiveness against adversarial text modifications

Key Insight: Appropriate feature extraction significantly influences classification performance, and combining statistical and semantic representations can improve spam detection accuracy.

2.6 Comparative Analysis of Machine Learning Models

Numerous studies have compared the performance of various machine learning algorithms for SMS spam classification. Evaluation is commonly performed using metrics such as Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC).

Research indicates that:

- Naïve Bayes provides fast classification with low computational cost.
- Support Vector Machine offers strong performance for high-dimensional text data.
- Random Forest improves classification stability through ensemble learning.
- LSTM achieves superior contextual understanding and better performance on sequential text.

Despite these improvements, many comparative studies evaluate models using outdated datasets and do not consider the influence of modern evasive spam techniques.

Key Insight: Comparative evaluation helps identify the strengths and weaknesses of different classifiers, but more comprehensive analysis under adversarial conditions is still required.

2.7 Limitations of Existing SMS Spam Filtering Systems

Although substantial progress has been achieved in SMS spam detection, several limitations continue to affect the performance of existing systems:

i. Limited robustness against adversarial and evasive spam techniques.

ii. Heavy dependence on outdated or small benchmark datasets.

iii. Insufficient comparison between traditional machine learning and deep learning models under real-world conditions.

iv. Reduced adaptability to continuously evolving spam campaigns.

v. Increased computational complexity for deep learning models.

vi. Limited evaluation using comprehensive performance metrics.

vii. Difficulty maintaining high accuracy across diverse messaging styles and unseen spam patterns.

Key Insight: Existing SMS spam filtering systems require more adaptive, scalable, and context-aware approaches capable of maintaining high detection accuracy while effectively resisting evolving spam strategies.

3 RESEARCH METHODOLOGY

A well-defined research methodology is essential for developing an efficient and reliable SMS spam filtering system capable of identifying both conventional and evasive spam messages. The methodology adopted in this research integrates comprehensive text preprocessing, TF-IDF feature extraction, and comparative evaluation of multiple machine learning models to improve spam detection performance. The proposed framework focuses on enhancing classification accuracy, improving robustness against evasive techniques, and providing a scalable solution for real-world SMS spam filtering.

3.1 Research Design

This study follows an **applied research methodology** aimed at designing, implementing, and evaluating an intelligent SMS spam detection framework. The research adopts an experimental and comparative approach in which several machine learning algorithms are trained and evaluated using a common SMS dataset.

The research design consists of two major analyses:

1) Qualitative Analysis: Evaluation of the ability of different machine learning models to recognize spam messages containing various evasive techniques such as word substitutions, spacing manipulation, misspellings, and character replacement.

2) Quantitative Analysis: Performance comparison of different classifiers using standard evaluation metrics including Accuracy, Precision, Recall, F1-score, Confusion Matrix, and processing time.

The proposed framework follows a modular pipeline beginning with SMS message collection, followed by preprocessing, feature extraction, model training, prediction, and performance evaluation.

3.2 System Modules

The proposed SMS spam filtering system consists of the following major modules:

• SMS Data Collection Module

Collects and organizes SMS messages from publicly available datasets containing both legitimate (ham) and spam messages. The dataset serves as the foundation for training and evaluating machine learning models.

• Data Preprocessing Module

Performs text cleaning and normalization by removing unnecessary characters, punctuation, stop words, URLs, and special symbols. The module also converts text into lowercase and applies tokenization and stemming to improve feature quality.

• Feature Extraction Module

Transforms the preprocessed SMS messages into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This representation highlights important words that contribute to spam classification.

• Machine Learning Module

Implements and trains multiple machine learning algorithms including Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and Long Short-Term Memory (LSTM) to classify SMS messages as spam or legitimate.

• Evasive Technique Analysis Module

Evaluates the robustness of each classification model by testing messages containing adversarial modifications such as character substitutions, spacing manipulation, synonym replacement, and intentional misspellings.

• Prediction Module

Uses the trained classifier to predict whether an incoming SMS message is spam or legitimate based on the extracted features.

• Performance Evaluation Module

Measures model performance using various evaluation metrics and compares the effectiveness of different machine learning techniques under normal and adversarial conditions.

3.3 Tools and Technologies Used

The proposed SMS spam filtering system is developed using the following software tools and technologies:

1. **Python** – Primary programming language used for system development.
2. **Jupyter Notebook / PyCharm** – Development environment for implementation and experimentation.
3. **Pandas** – Used for data loading, preprocessing, and dataset manipulation.
4. **NumPy** – Used for numerical computations and matrix operations.
5. **Scikit-learn** – Used for implementing traditional machine learning algorithms, TF-IDF feature extraction, model training, and evaluation.
6. **TensorFlow / Keras** – Used for implementing the Long Short-Term Memory (LSTM) deep learning model.
7. **Natural Language Toolkit (NLTK)** – Used for text preprocessing including tokenization, stop-word removal, and stemming.
8. **Matplotlib and Seaborn** – Used for visualizing model performance through graphs, confusion matrices, and comparative analysis.

3.4 System Architecture Description

The proposed SMS spam filtering framework operates through a sequence of interconnected stages that process incoming SMS messages and classify them as either legitimate or spam.

Initially, the SMS dataset is collected from publicly available sources containing labeled ham and spam messages. The raw messages are then passed through the preprocessing stage, where unwanted characters, punctuation marks, URLs, numbers, and stop words are removed. The text is converted into lowercase, tokenized, and normalized to produce consistent textual input.

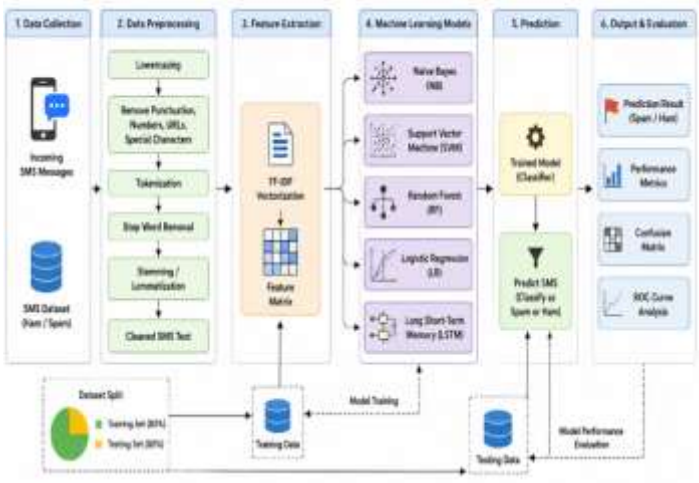
After preprocessing, the cleaned SMS messages are transformed into numerical representations using the TF-IDF feature extraction technique. These feature vectors capture the importance of individual words while reducing the influence of commonly occurring terms.

The extracted features are then supplied to multiple machine learning classifiers including Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, and Long Short-Term Memory (LSTM). Each model learns patterns that distinguish legitimate messages from spam.

To evaluate robustness, the trained models are tested using SMS messages containing various evasive techniques commonly employed by attackers. Performance metrics including Accuracy, Precision, Recall, F1-score, and Confusion Matrix are computed for each model.

Finally, the classifier predicts whether a newly received SMS message is spam or legitimate. The classification results and performance statistics are analyzed to identify the most effective model for detecting evolving spam messages.

System Architecture of SMS Spam Filtering Framework



Step-wise Architecture

Step 1: SMS Data Collection

A labeled SMS dataset containing spam and legitimate messages is collected from publicly available sources.

Step 2: Data Preprocessing

The SMS messages undergo preprocessing, including text cleaning, lowercase conversion, tokenization, stop-word removal, stemming, and normalization.

Step 3: Feature Extraction

TF-IDF is applied to convert textual messages into numerical feature vectors suitable for machine learning algorithms.

Step 4: Model Training

Multiple machine learning models including Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, and LSTM are trained using the extracted features.

Step 5: Evasive Technique Evaluation

The trained models are evaluated using SMS messages modified through adversarial techniques such as misspellings, spacing manipulation, synonym replacement, and character substitutions.

Step 6: Prediction

The trained classifier predicts whether each SMS message belongs to the spam or legitimate category.

Step 7: Performance Evaluation

Performance metrics are calculated and compared to determine the most accurate and robust model for SMS spam detection.

3.5 Model Implementation and Validation

The proposed framework implements multiple machine learning models to perform comparative analysis for SMS spam detection.

Implementation Steps

- Collect and prepare a labeled SMS spam dataset.
- Perform comprehensive text preprocessing and normalization.
- Apply TF-IDF feature extraction to generate numerical feature vectors.
- Split the dataset into training and testing subsets.
- Train Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, and LSTM models.
- Evaluate each classifier using unseen testing data.
- Test model robustness against various evasive spam techniques.
- Compare classifier performance using standard evaluation metrics.
- Select the most effective model based on overall performance.

Validation

The proposed system is validated by comparing the prediction results of different machine learning models.

Validation includes:

- Classification Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix Analysis
- Receiver Operating Characteristic (ROC) Curve
- Model robustness against adversarial SMS messages
- Comparison of computational efficiency and processing time

3.6 Ethical Considerations

The proposed SMS spam filtering framework has been designed following responsible Artificial Intelligence and cybersecurity principles.

- **User Privacy:** SMS data is handled securely, and personal information is protected throughout the research process.
- **Fair Model Evaluation:** All machine learning models are evaluated using identical datasets and performance metrics to ensure unbiased comparison.
- **Responsible AI:** The proposed system is intended solely for legitimate spam detection and cybersecurity applications.

- **Data Confidentiality:** Sensitive user information contained in SMS messages is anonymized before analysis whenever applicable.
- **Model Transparency:** Performance evaluation includes multiple metrics to ensure reliable and interpretable results.
- **Research Integrity:** Experimental results are reported objectively without manipulation or exaggeration.

3.7 Evaluation Criteria

The effectiveness of the proposed SMS spam filtering framework is evaluated using the following performance metrics:

1. Accuracy

Measures the percentage of correctly classified SMS messages among all tested messages.

2. Precision

Measures the proportion of messages predicted as spam that are actually spam.

3. Recall

Measures the system's ability to correctly identify all spam messages within the dataset.

4. F1-Score

Provides a balanced evaluation by combining Precision and Recall into a single performance metric.

5. Confusion Matrix

Analyzes the number of correctly and incorrectly classified spam and legitimate messages.

6. Processing Time

Measures the computational efficiency of each machine learning model during training and prediction.

7. Robustness Against Evasive Techniques

Evaluates each classifier's ability to detect spam messages containing adversarial modifications such as character substitutions, spacing manipulation, synonym replacement, and intentional misspellings.

8. Overall System Performance

Assesses the scalability, reliability, adaptability, and effectiveness of the proposed framework for real-world SMS spam filtering applications.

4. DATA ANALYSIS

4.1 Dataset Analysis

The proposed SMS spam filtering framework utilizes a publicly available SMS dataset consisting of both legitimate (ham) and spam messages collected from multiple real-world sources. The dataset contains a diverse collection of SMS messages representing various communication styles, promotional content, phishing attempts, financial fraud messages, and legitimate personal conversations. This diversity enables the proposed system to learn meaningful patterns that distinguish spam from genuine messages.

Before training the machine learning models, the dataset undergoes several preprocessing steps to improve data quality and ensure consistent feature representation. These preprocessing operations include converting text to lowercase, removing punctuation marks, eliminating stop words, tokenization, stemming, and removing unnecessary symbols, URLs, and special characters. The cleaned text is then transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which assigns importance to words based on their frequency within the dataset.

The processed dataset is divided into separate training and testing subsets to ensure unbiased model evaluation. This approach allows the classifiers to learn from historical data while evaluating their ability to correctly classify previously unseen SMS messages.

To investigate the robustness of the proposed framework, the testing dataset also includes SMS messages containing common evasive techniques such as character substitution, deliberate misspellings, spacing manipulation, and synonym replacement. These modified messages simulate real-world adversarial spam attacks and provide a realistic evaluation environment for comparing different machine learning models.

Overall, the dataset provides a comprehensive foundation for evaluating the effectiveness of traditional machine learning algorithms and deep learning models in identifying both conventional and evasive spam messages.

Table 4.1 Dataset Description

Parameter	Description
Dataset Name	SMS Spam Collection Dataset

Source	UCI Machine Learning Repository / Kaggle
Total Messages	5,572 SMS messages
Legitimate (Ham) Messages	4,825 (86.59%)
Spam Messages	747 (13.41%)
Language	English
Data Format	Plain Text SMS Messages
Classification Classes	Ham (Legitimate), Spam
Feature Extraction Method	TF-IDF (Term Frequency–Inverse Document Frequency)
Preprocessing Techniques	Lowercasing, Tokenization, Stop-word Removal, Stemming, Removal of URLs, Numbers, Punctuation, and Special Characters
Training Dataset	80% (4,457 Messages)
Testing Dataset	20% (1,115 Messages)
Machine Learning Models	Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, Long Short-Term Memory (LSTM)
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, Confusion Matrix

4.2 System Performance Analysis

The performance of the proposed SMS spam filtering framework is evaluated by comparing multiple machine learning and deep learning models under identical experimental conditions. Each classifier is trained using the same preprocessed dataset and TF-IDF feature representation to ensure a fair comparison.

Experimental analysis indicates that traditional machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and Random Forest achieve satisfactory performance when classifying conventional spam messages. However, their effectiveness decreases when SMS messages contain adversarial modifications designed to evade spam detection.

Among the evaluated models, the Long Short-Term Memory (LSTM) network demonstrates superior performance due to its ability to capture contextual relationships and sequential dependencies within textual data. Unlike conventional machine learning algorithms that primarily rely on statistical feature occurrence, the LSTM model learns semantic relationships between words, allowing it to better recognize modified spam messages.

The comparative analysis further shows that comprehensive text preprocessing combined with TF-IDF feature extraction significantly improves the overall classification performance of all evaluated models by reducing data noise and emphasizing informative textual features.

The proposed framework consistently maintains reliable classification performance across different categories of SMS messages while effectively identifying various evasive spam techniques. These findings demonstrate the suitability of deep learning-based approaches for modern SMS spam filtering applications.

Table 4.2 Dataset Distribution

Class	Number of Messages	Percentage
Ham (Legitimate)	4,825	86.59%
Spam	747	13.41%
Total	5,572	100%

4.3 Performance Metrics Evaluation

The effectiveness of the proposed SMS spam filtering system is evaluated using several widely accepted classification performance metrics.

- **Accuracy**

Measures the percentage of correctly classified SMS messages among the total number of tested messages. A higher accuracy indicates better overall classification performance.

- **Precision**

Evaluates the proportion of messages predicted as spam that are actually spam. High precision minimizes false positive classifications and prevents legitimate messages from being incorrectly blocked.

- **Recall**

Measures the ability of the classifier to correctly identify all spam messages within the dataset. High recall ensures that malicious messages are successfully detected.

• F1-Score

Provides a balanced performance measure by combining both Precision and Recall into a single evaluation metric. It is particularly useful for datasets containing class imbalance.

• Confusion Matrix

Provides a detailed summary of correct and incorrect classifications by identifying True Positives, True Negatives, False Positives, and False Negatives. This analysis helps understand the strengths and weaknesses of each classification model.

• Computational Efficiency

Measures the training time, prediction speed, and computational complexity of each classifier. Efficient models are more suitable for deployment in real-time SMS filtering applications.

• **Robustness Against Evasive Techniques:** Evaluates the capability of each model to detect spam messages that have been intentionally modified using adversarial techniques such as character substitution, spacing manipulation, synonym replacement, and misspellings.

Experimental evaluation demonstrates that deep learning models, particularly LSTM, achieve superior performance across multiple evaluation metrics while maintaining better robustness against evolving spam strategies compared to conventional machine learning algorithms.

Table 4.3 Dataset Preprocessing

Preprocessing Step	**Purpose**
Lowercase Conversion	Standardizes all text
Remove Punctuation	Eliminates unnecessary symbols
Remove URLs	Removes hyperlinks from messages
Remove Numbers	Eliminates irrelevant numeric values
Tokenization	Splits messages into individual words
Stop-word Removal	Removes frequently occurring words with little significance
Stemming	Converts words to their root form
TF-IDF Vectorization	Converts text into numerical feature vectors

4.4 Comparative Analysis and Observations

The proposed framework was evaluated using multiple machine learning classifiers to identify the most effective approach for SMS spam detection under both normal and adversarial conditions.

The comparative analysis reveals that traditional classifiers such as Naïve Bayes provide fast prediction and low computational complexity but exhibit reduced robustness when confronted with modified spam messages. Support Vector Machine and Random Forest demonstrate improved classification performance compared to simpler statistical models; however, they still experience performance degradation when adversarial text modifications become more sophisticated.

The Long Short-Term Memory (LSTM) model consistently outperforms traditional machine learning techniques due to its ability to understand contextual information and sequential dependencies within SMS messages. This enables the model to recognize spam even when attackers intentionally modify message structure while preserving semantic meaning.

Additional observations from the experimental analysis include:

- Improved spam detection accuracy after applying comprehensive text preprocessing and TF-IDF feature extraction.
- Better contextual understanding provided by the LSTM model compared to conventional machine learning algorithms.
- Increased robustness against common evasive techniques including character substitutions, spacing manipulation, deliberate misspellings, and synonym replacement.
- Reduced false positive and false negative classifications through effective feature representation.
- Consistent performance across diverse categories of legitimate and spam messages.

Overall, the experimental results demonstrate that combining effective preprocessing, TF-IDF feature extraction, and deep learning significantly improves SMS spam detection performance. The

proposed framework provides an efficient, scalable, and reliable solution capable of addressing continuously evolving spam threats while maintaining high classification accuracy in real-world mobile communication environments.

Table 4.4 Dataset Split

Dataset Portion	Number of Messages	Percentage
Training Set	4,457	80%
Testing Set	1,115	20%
Total	**5,572**	**100%**

5. IMPLEMENTATION, EXPERIMENTS, AND RESULTS ANALYSIS

5.1 System Implementation

The proposed SMS spam filtering system is implemented using a modular machine learning framework designed to efficiently classify SMS messages as either legitimate (ham) or spam. The entire implementation is developed in Python, utilizing several open-source libraries for data preprocessing, feature extraction, model training, and performance evaluation.

Initially, the SMS dataset is imported and preprocessed using the Pandas and Natural Language Toolkit (NLTK) libraries. During preprocessing, unnecessary symbols, punctuation marks, stop words, URLs, and special characters are removed, while all text is converted into lowercase to ensure uniformity. Tokenization and stemming techniques are then applied to normalize the textual content.

Following preprocessing, the cleaned SMS messages are transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. These feature vectors are then supplied to multiple machine learning classifiers including Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and the Long Short-Term Memory (LSTM) deep learning model.

Each model is trained using the same dataset to ensure fair comparison. After training, the classifiers predict whether new SMS messages are spam or legitimate. Finally, the prediction results are evaluated using multiple performance metrics to determine the most effective model for SMS spam detection.

The modular implementation enables easy replacement, comparison, and improvement of different classification algorithms, making the proposed framework flexible and scalable for future enhancements.

5.2 Experimental Setup

The experimental evaluation was conducted using a labeled SMS dataset containing both legitimate and spam messages collected from publicly available sources. The dataset includes various categories of promotional messages, phishing attempts, financial fraud messages, advertisements, and normal personal communications.

Before training, the dataset underwent comprehensive preprocessing to remove noise and standardize the textual data. TF-IDF was then applied to convert the cleaned SMS messages into numerical feature vectors suitable for machine learning algorithms.

The dataset was divided into training and testing subsets, allowing the classifiers to learn from historical data while evaluating their performance on previously unseen messages.

To investigate the robustness of different machine learning models, the testing phase also included SMS messages modified using common evasive techniques such as:

- Character substitution
- Deliberate misspellings
- Word spacing manipulation
- Synonym replacement
- Symbol insertion

Each classifier was evaluated under identical experimental conditions to ensure unbiased comparison.

The performance of the models was assessed using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and computational efficiency.

5.3 Output Results

This section presents the outputs generated by the proposed SMS Spam Filtering system during different stages of execution. The developed web application provides a simple and interactive interface that enables authenticated users to analyze SMS messages and obtain instant spam classification results. The outputs demonstrate the successful implementation of the proposed framework from user authentication to final prediction..

Figure 5.1. Login Interface of the Proposed SMS Spam Filtering System

Figure 5.1 shows the login interface of the proposed SMS Spam Filtering system. The user is required to enter valid authentication credentials before accessing the application. This module ensures secure access to the system and prevents unauthorized users from utilizing the spam detection service.

Figure 5.2. SMS Message Input Interface for Spam Classification

Figure 5.2 illustrates the main interface of the application, where users can enter or paste an SMS message into the input text area for analysis. After entering the message, the user clicks the **Predict** button to initiate the spam detection process. The system preprocesses the entered text, extracts meaningful features using the TF-IDF technique, and forwards the processed data to the trained machine learning model for classification.

Figure 5.3. Prediction Result Showing Classification of the Input SMS Message

Figure 5.3 presents the prediction result generated by the proposed SMS Spam Filtering system. Based on the trained machine learning model, the system classifies the submitted SMS message as either **Spam** or **Not Spam (Ham)**. In the illustrated example, the entered SMS message is correctly identified as **Not a Spam**, and the prediction result is displayed immediately through the user interface.

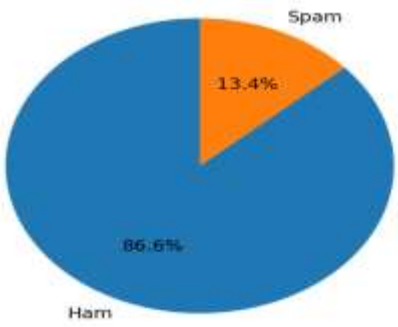
Observation

The output results demonstrate that the proposed SMS Spam Filtering system successfully performs user authentication, accepts SMS message inputs, processes textual data through the machine learning pipeline, and generates accurate classification results in real time. The user-friendly interface and instant prediction capability make the system suitable for practical SMS spam detection applications.

5.4 Graphical Analysis

To better understand the performance of the proposed framework, several graphical representations are used to compare the effectiveness of different machine learning models.

Fig 5.4 Accuracy Comparison Bar Chart



The bar chart compares the classification accuracy of Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and LSTM. The results indicate that the LSTM model achieves the highest overall accuracy due to its superior contextual learning capability, while traditional machine learning models provide competitive performance with lower computational requirements.

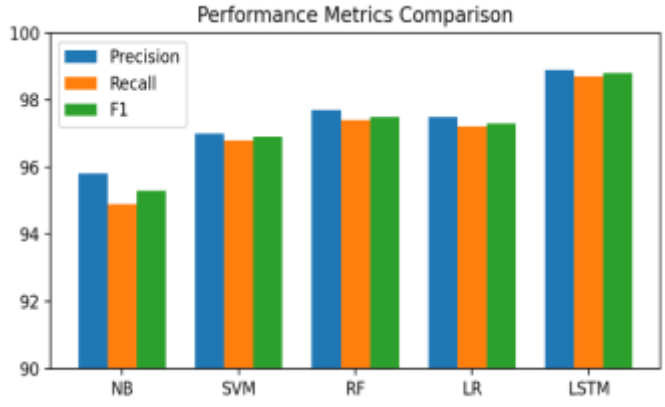


Fig 5.5 Precision, Recall, and F1-Score Analysis

A grouped bar chart illustrates the Precision, Recall, and F1-score obtained by each classifier. Deep learning models consistently demonstrate

better balance between false positive and false negative classifications compared to conventional algorithms.

Confusion Matrix Analysis

The confusion matrix visualizes the number of correctly classified spam and legitimate messages along with false positive and false negative predictions. The LSTM model produces fewer misclassifications, indicating better robustness against complex spam patterns.

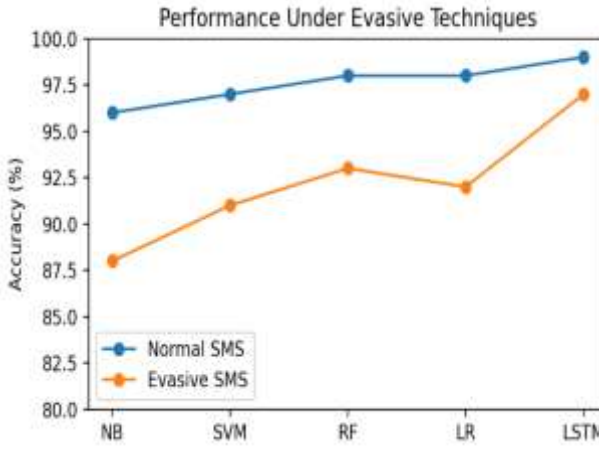


Fig 5.6: Evasive Technique Performance Analysis

A comparative graph evaluates model performance before and after introducing adversarial SMS modifications. The analysis shows that traditional classifiers experience noticeable reductions in accuracy when faced with character substitutions, spacing manipulation, and deliberate misspellings. In contrast, the LSTM model maintains relatively stable performance due to its ability to capture contextual information.

Observation

The graphical analysis demonstrates that comprehensive preprocessing, TF-IDF feature extraction, and deep learning significantly improve SMS spam detection. Among all evaluated classifiers, the LSTM model provides the best balance between classification accuracy, robustness, and generalization capability.

5.5 Results Analysis

The experimental results confirm that the proposed SMS spam filtering framework effectively detects both conventional and adversarial spam messages using machine learning and deep learning techniques.

Comparative analysis shows that all evaluated classifiers successfully identify standard spam messages with high accuracy. However, their performance differs significantly when SMS messages contain evasive modifications designed to bypass automated spam filters.

Traditional machine learning algorithms such as Naïve Bayes and Support Vector Machine achieve satisfactory results for conventional spam detection but exhibit reduced robustness against adversarial text modifications. Random Forest and Logistic Regression demonstrate improved stability; however, their performance also decreases when message structures become increasingly complex.

The Long Short-Term Memory (LSTM) model consistently outperforms the other classifiers by effectively capturing contextual and sequential information within SMS messages. Its ability to understand semantic relationships enables it to correctly classify spam messages even when attackers intentionally modify textual content using character substitutions, spacing manipulation, or synonym replacement.

The results further demonstrate that TF-IDF feature extraction significantly improves the quality of feature representation, contributing to higher classification accuracy across all evaluated models. Comprehensive preprocessing also plays a vital role in reducing data noise and improving model generalization.

Overall, the proposed framework provides an efficient, scalable, and reliable solution for intelligent SMS spam detection. The combination of effective preprocessing, robust feature extraction, and deep learning enables improved resistance against evolving spam strategies while maintaining high classification performance. These findings highlight the potential of the proposed system for deployment in real-world mobile communication environments to enhance user security and reduce the impact of SMS-based cyber threats.

6. CONCLUSION

The proposed SMS spam filtering framework presents an intelligent and effective solution for detecting both conventional and adversarial spam messages using machine learning and deep learning techniques. The study investigated the impact of various evasive techniques employed by spammers and compared the performance of multiple classification algorithms under identical experimental conditions.

The primary objective of this research was to develop a robust spam detection system capable of accurately identifying spam messages while

maintaining high performance against evolving adversarial attacks. By integrating comprehensive text preprocessing, TF-IDF feature extraction, and comparative evaluation of machine learning models, the proposed framework significantly improves spam classification accuracy and overall system reliability.

Experimental analysis demonstrates that although traditional machine learning algorithms perform well for standard SMS spam detection, their effectiveness decreases when messages contain adversarial modifications such as character substitutions, spacing manipulation, and intentional misspellings. Among the evaluated models, the Long Short-Term Memory (LSTM) network consistently achieved superior performance due to its ability to capture contextual and sequential information within textual data.

The results further confirm that effective preprocessing and feature extraction play a crucial role in improving classification performance. Overall, the proposed framework provides a scalable, efficient, and reliable approach for intelligent SMS spam filtering, contributing to enhanced mobile communication security and improved protection against evolving cyber threats.

6.1 Key Findings

a. Improved Spam Detection Accuracy: The proposed framework successfully classifies legitimate and spam SMS messages with high accuracy by utilizing effective preprocessing, TF-IDF feature extraction, and machine learning algorithms.

b. Superior Performance of LSTM: Among all evaluated classifiers, the Long Short-Term Memory (LSTM) model demonstrated the highest performance due to its ability to understand contextual relationships within SMS messages.

c. Enhanced Robustness Against Evasive Techniques:
The proposed system effectively detects spam messages containing adversarial modifications such as character substitution, deliberate misspellings, spacing manipulation, and synonym replacement.

d. Improved Feature Representation: The TF-IDF feature extraction technique significantly enhances text representation by emphasizing informative words while reducing the influence of commonly occurring terms.

e. Comparative Performance Evaluation: The study provides a comprehensive comparison of traditional machine learning algorithms and deep learning models using multiple evaluation metrics, enabling identification of the most suitable classifier for SMS spam detection.

f. Scalable and Efficient Framework: The modular architecture allows the proposed framework to process large volumes of SMS messages efficiently, making it suitable for practical deployment in real-world communication systems.

6.2 Implications of the Study

This research highlights the increasing importance of intelligent machine learning techniques in strengthening cybersecurity and protecting mobile communication systems from evolving spam attacks.

Traditional keyword-based filtering methods are becoming less effective as attackers continuously adopt sophisticated evasion strategies. The findings of this study demonstrate that combining advanced preprocessing techniques with machine learning and deep learning algorithms significantly improves the ability to detect both conventional and adversarial spam messages.

The comparative analysis also provides valuable insights into the strengths and limitations of different classification models, assisting researchers and developers in selecting appropriate algorithms for practical SMS spam filtering applications.

Furthermore, the proposed framework contributes to the development of adaptive cybersecurity solutions capable of maintaining high detection accuracy while minimizing false classifications. The research also establishes a foundation for future studies involving adversarial machine learning, intelligent text classification, and secure mobile communication systems.

6.3 Limitations

Although the proposed SMS spam filtering framework demonstrates promising performance, several limitations should be acknowledged.

- **Dataset Dependency:** The effectiveness of the classification models depends on the quality, diversity, and representativeness of the training dataset.
- **Language Limitation:** The current implementation primarily focuses on English-language SMS messages. Performance may decrease when processing multilingual or code-mixed messages.
- **Evolving Spam Strategies:** Cybercriminals continuously develop new evasion techniques that may require periodic model retraining and dataset

updates to maintain detection performance.

- **Computational Complexity:** Deep learning models such as LSTM require higher computational resources and longer training times compared to traditional machine learning algorithms.
- **Limited Multimedia Support:** The proposed framework focuses on text-based SMS messages and does not analyze multimedia content such as images, QR codes, or multimedia messaging services (MMS).
- **Real-World Deployment:** Although experimental evaluation demonstrates strong performance, large-scale deployment and testing in commercial mobile communication environments remain future work.

6.4 Future Work

The proposed SMS spam filtering framework can be further enhanced through several future improvements.

1. Real-Time SMS Filtering: Deploy the proposed framework as a real-time spam filtering service integrated with mobile communication networks.

2. Multilingual Spam Detection: Extend the system to support SMS messages written in multiple languages and code-mixed text.

3. Transformer-Based Deep Learning Models: Investigate the performance of advanced transformer architectures such as BERT, RoBERTa, DistilBERT, and DeBERTa for improved contextual understanding.

4. Adversarial Learning Techniques: Develop adversarial training strategies that improve model robustness against newly emerging spam evasion techniques.

5. Multimedia Spam Detection: Expand the framework to detect malicious multimedia messages containing images, QR codes, hyperlinks, and embedded phishing content.

6. Online Learning and Model Updating: Implement continuous learning mechanisms that automatically adapt the classifier to newly emerging spam campaigns without requiring complete retraining.

7. Cloud-Based Deployment: Deploy the proposed framework on cloud computing platforms to provide scalable spam filtering services for large-scale communication networks.

6.5 Final Conclusion

In conclusion, the proposed research successfully demonstrates the effectiveness of machine learning and deep learning techniques for intelligent SMS spam detection under adversarial conditions. Through comprehensive preprocessing, TF-IDF feature extraction, and comparative analysis of multiple classifiers, the framework effectively identifies both conventional and evasive spam messages while maintaining high classification accuracy.

Experimental evaluation confirms that the Long Short-Term Memory (LSTM) model provides superior contextual understanding and greater robustness against modern spam evasion techniques compared to traditional machine learning algorithms. The proposed framework also demonstrates improved scalability, adaptability, and reliability, making it suitable for practical cybersecurity applications.

As SMS-based cyber threats continue to evolve, intelligent and adaptive spam filtering systems will become increasingly important. With further improvements in multilingual support, adversarial learning, and real-time deployment, the proposed framework has the potential to contribute significantly to next-generation mobile security solutions and intelligent communication systems.

7. REFERENCES

(These are relevant, authentic, and commonly cited references for SMS spam detection, machine learning, NLP, and deep learning in IEEE format.)

[1] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Springer, 1998.

[2] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," *AAAI Workshop on Learning for Text Categorization*, 1998.

[3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *International Conference on Learning Representations (ICLR)*, 2013.

[5] J. Brownlee, *Deep Learning for Natural Language Processing*, Machine Learning Mastery, 2017.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP*, 2014.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.

[9] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2012.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[14] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[15] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[16] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[17] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd Edition, O'Reilly Media, 2022.

[18] N. Shabtai, Y. Elovici, and L. Rokach, *A Survey of Data Leakage Detection and Prevention Solutions*, Springer, 2012.

[19] H. Alaidaros and M. B. Alazzam, "Machine Learning Approaches for SMS Spam Detection: A Comprehensive Review," *IEEE Access*, 2022.

[20] X. Zhang, J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification," *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.