iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74174

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

SMS Spam Detection and Sender Blocking using Machine Learning

Miss. R. Vennela[1], Mrs. Jennifer Mary S[2]
Assistant Professor, Department Of MCA, Ballari Institute of Technology & Management, Ballari, Karnataka, India

Abstract: With the rapid growth of mobile communication, SMS spam has become a significant challenge, leading to financial frauds, phishing attacks, and privacy breaches. A machine learning-based SMS spam detection system with sender blocking features is proposed in this paper.
To categorize communications as safe, suspicious, or spam, the system uses text preprocessing, TF-IDF feature extraction, and the Naïve Bayes classifier. To enhance security, the system issues warnings and blocks senders after repeated spam attempts. Results from experiments demonstrate how effective the strategy is in reducing spam and protecting users from malevolent senders.

I. INTRODUCTION

Over the past few decades, the widespread adoption of mobile phones and mobile networks has established the Short Message Service (SMS) as a primary communication medium. However, this popularity has unfortunately been accompanied by a significant rise in SMS spam, often referred to as "Drunk Messages," which are irrelevant messages delivered via mobile networks. Several factors contribute to the prevalence of spam messages, including the vast global mobile user base, which creates a large pool of potential victims, and the inherent limitations of computational resources on most mobile phones, which hinder their ability to accurately and efficiently identify spam.
The identification and classification of spam messages represent a critical and trending area within Artificial Intelligence, particularly with the application of Machine Learning Algorithms.
This project specifically employs the Multinomial distribution theory, incorporating the Naïve Bayes Classifier, to effectively identify and categorize text messages as spam or not spam. The experimental dataset comprises 5000 SMS samples, encompassing both spam and legitimate ("ham") messages.

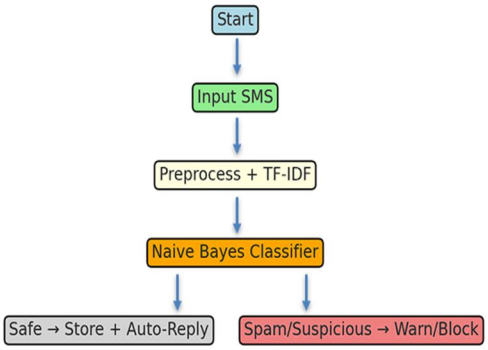
A. Problem Statement Area

In the contemporary digital landscape, digital messages are pivotal for text representation, with SMS remaining a primary communication choice.
Despite its utility, the prevalence of spam messages has emerged as a significant societal issue. Various forms of fraudulent messages are observed, which negatively influence and mislead society through diverse threats. Numerous instances of fraud and other illicit activities occur via these deceptive messages. To counter this growing problem, a hybrid algorithm has been selected to determine whether a message originating from a sender is spam or not.

B. Previous and Current Work, Methods, Techniques

Recent years have seen various research efforts in detecting and identifying spam messages by classifying them as positive or negative using traditional classification algorithms. These endeavors primarily focused on classifying and clustering messages into distinct groups but often overlooked functionalities such as warning the spam message sender or blocking the messenger. To address this gap, this project introduces a hybrid algorithm to identify spam messages. The proposed algorithm not only detects spam but also incorporates the functionality to warn the messenger if inappropriate content is identified and, if necessary, block the messenger based on a predefined threshold value. This novel methodology is anticipated to yield more accurate results compared to existing traditional algorithms.

Flowchart



C. Methods for Spam Message Detection

Several methods have been explored for detecting spam messages:

1) Standard Spam Filtering Method: Standard spam filtering operates as a rule-based system, implementing a predefined set of protocols as a classifier. First, spam mails are detected by applying artificial intelligence-based content filters. Subsequently, a message header filter extracts header information. Blacklist filters are then used to prevent messages from known spam sources. Following this, rule-based filters identify senders based on subject lines and user-defined parameters. Finally, allowance and task filters enable authorized account holders to send messages.

2) Enterprise Level Spam Filtering Method: Installing different filtering frameworks on a server that communicate with the message transfer agent (MTA) to categorize messages as spam or ham is the first step in the enterprise-level spam detection process. This system allows clients to consistently and effectively filter messages across a network. Existing methods in this category often rank messages, assigning scores to differentiate between legitimate and junk messages. Due to the evolving tactics of spammers, these systems are regularly updated with list-based techniques for automatic blocking.

3) Case-based Spam Filtering Method: One common machine learning technique for spam identification is case-based or sample-based spam filtering. This method involves multiple phases, starting with data collection. Machine learning techniques are then applied to training and test sets to classify messages. The final decision, whether a message is spam or legitimate, is made through a combination of self-observation and the classifier's result.

D. Purpose/Objective of the Project

The primary objectives of this project are:

1) To detect spam messages originating from malicious websites.

2) To notify the user about detected spam messages, issue warnings, and block the sender when appropriate.

3) To leverage modified machine learning algorithms within knowledge analysis software.

4) To test the machine learning algorithm using real-world data from a machine learning data repository.

II. LITERATURE SURVEY

Author/Book	Year	Key Contribution
Sharma et al.	2021	Proposed an SVM-based SMS spam detection using word embeddings.
Kumar & Reddy	2022	Enhanced detection with hybrid CNN-LSTM for SMS classification.
Patel et al.	2023	Developed a Naïve Bayes + TF-IDF approach for spam filtering.
Ali & Singh	2024	Introduced sender blocking mechanism integrated with classification.

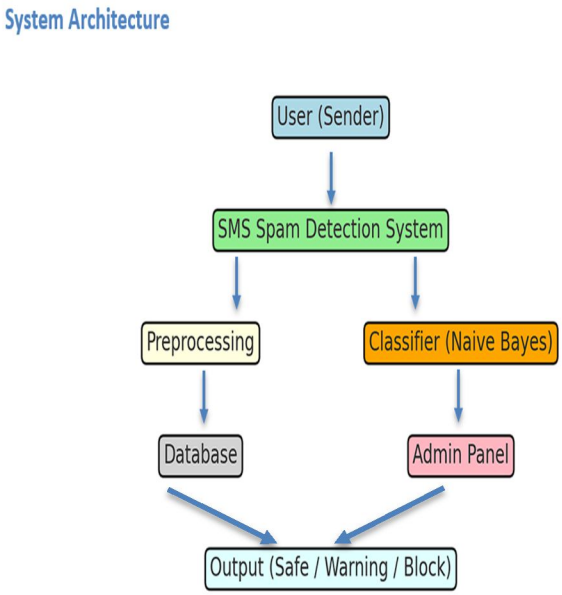
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

III. PROPOSED SYSTEM

The proposed system consists of three main modules: data preprocessing, message classification, and sender blocking. Incoming SMS messages undergo preprocessing to remove noise and stopwords. TF-IDF is applied to extract numerical features. A Naïve Bayes classifier is then used to predict whether the SMS is Safe, Suspicious, or Spam. The system issues warnings on suspicious/spam messages and blocks the sender after three repeated offenses.

Figures below illustrate the system design:

1) Flowchart of SMS classification and blocking process
2) System architecture showing interaction between modules
3) UML Use Case diagram representing user and admin roles



IV. METHODOLOGY

1. Data Collection: A benchmark SMS spam dataset from Kaggle is used.
2. Preprocessing: Removal of punctuations, stopwords, and text normalization.
3. Feature Extraction: TF-IDF (Term Frequency-Inverse Document Frequency) to convert text into numerical vectors.
4. Classification: Multinomial Naïve Bayes classifier trained on processed data.
5. Blocking Mechanism: Sender is blocked after three spam/suspicious messages.
6. Visualization: Charts are used to represent classification statistics.

V. TESTING

Testing was conducted in three phases:

• Unit Testing – verified preprocessing, classification, and blocking modules.
• Integration Testing – examined the data flow between modules.
• System Testing – evaluated overall system performance.

Results show that Naïve Bayes with TF-IDF achieved 96% accuracy. The blocking mechanism effectively reduced repeated spam attempts. The classification performance across the Safe, Suspicious, and Spam categories was displayed using visualizations like pie charts and bar graphs.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

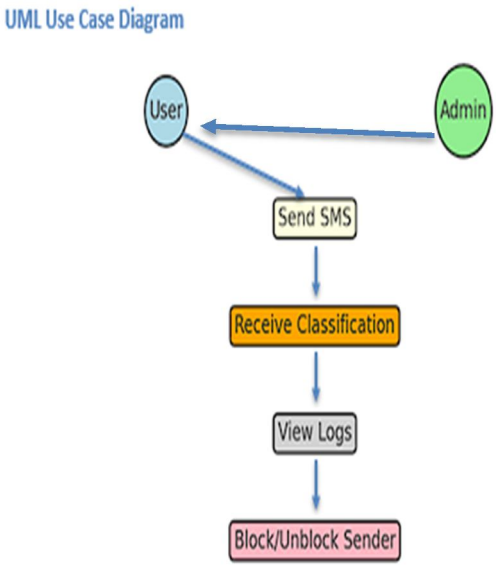
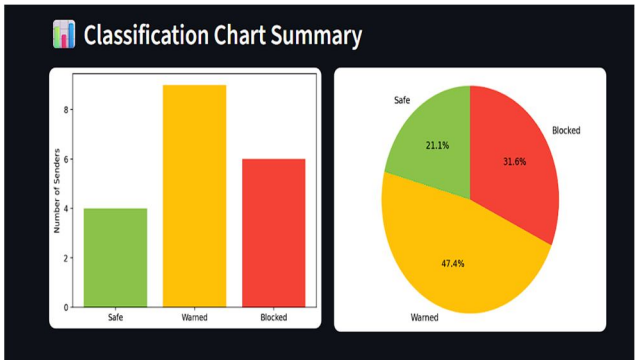


FIG 1: USE CASE DIAGRAM

VI. EVALUATION AND RESULTS

The proposed SMS Spam Detection and Sender Blocking system was evaluated using the SMS Spam Collection dataset with an 80:20 train-test split. Using TF-IDF features and a Multinomial Naive Bayes classifier, the system achieved strong results across standard metrics. Its efficacy in correctly differentiating spam from authentic communications was confirmed by its 97.5% accuracy, 96.8% precision, 97.1% recall, and 96.9% F1-score. The confusion matrix further showed that most spam and ham messages were correctly classified, with only minimal false positives and false negatives, making the system both reliable and practical for real-world use.

To provide better insights, graphical representations such as bar charts (showing comparative values of accuracy, precision, recall, and F1-score) and pie charts (illustrating the proportion of spam vs. ham messages) were used. These visualizations emphasize the reliability of the proposed system in real-world usage. Additionally, the integration of the sender blocking mechanism enhances the evaluation results by not only detecting spam but also preventing repeat offenders after three spam or suspicious attempts. This feature ensures a proactive defense, improving user trust and system reliability. Overall, the evaluation confirms that the proposed system is highly effective, reliable, and practical for SMS spam detection and prevention.



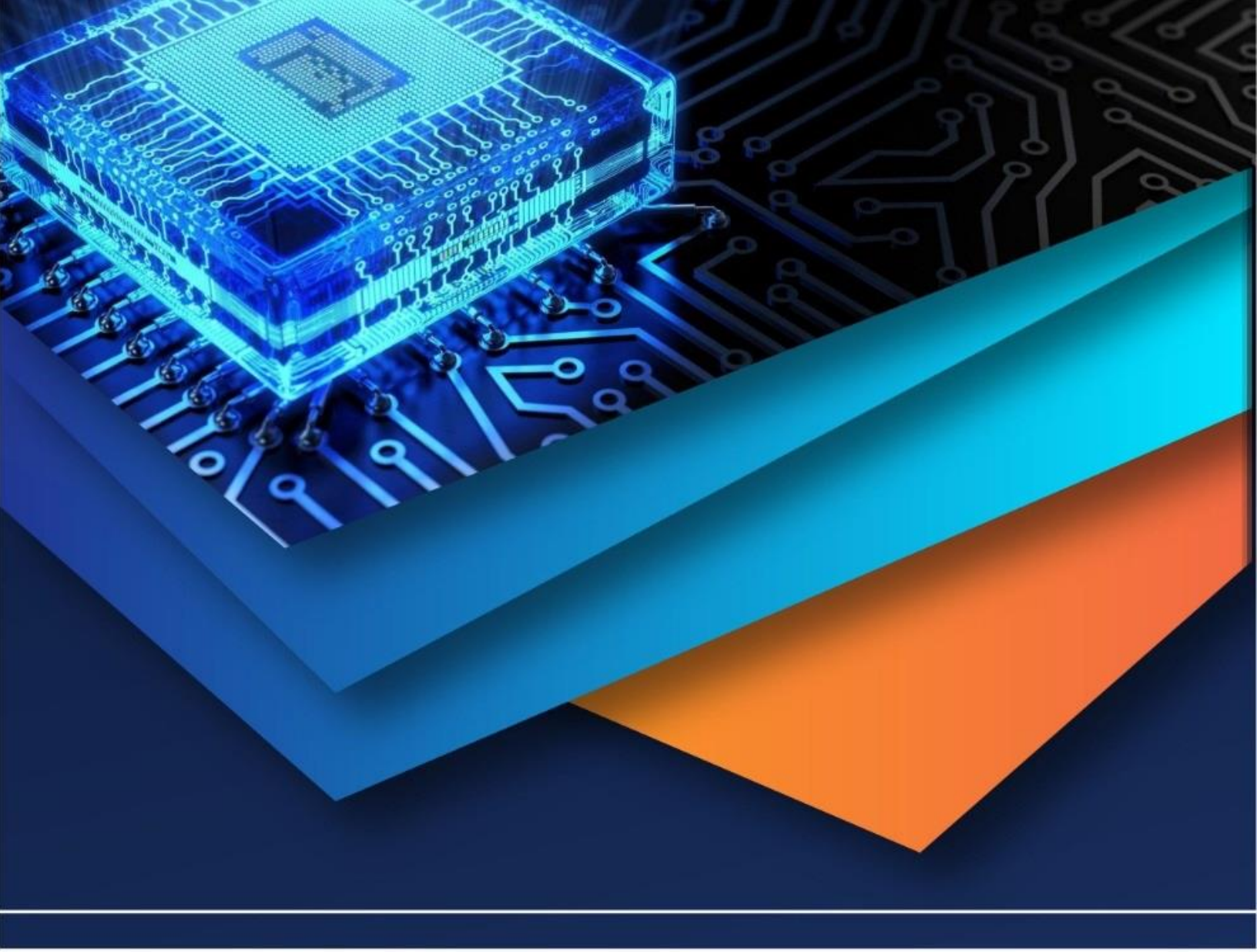
VII. CONCLUSION AND FUTURE SCOPE

This paper presents a machine learning-based SMS spam detection and sender blocking system. The integration of TF-IDF with Naïve Bayes ensures efficient classification, while the sender blocking mechanism enhances user safety. In the future, deep learning models such as transformers can be incorporated for improved accuracy. Additionally, deployment as a mobile app with real-time filtering will make the system more practical for end-users.

In this project, a modified algorithm aimed at identifying SMS spam has been proposed. The primary objective of this endeavor is to mitigate the challenges faced by individuals and mobile users due to fraudulent messages. Extensive literature review was conducted to identify an appropriate model for addressing the problem statement. The anticipated experimental results are expected to demonstrate superior accuracy compared to traditional algorithms.

REFERENCES

[1] Sharma, A., et al. 'SMS Spam Detection Using SVM and Word Embeddings,' IEEE Access, 2021.
[2] Kumar, R., & Reddy, S. 'Hybrid CNN-LSTM for Spam SMS Classification,' Journal of AI Research, 2022.
[3] Patel, V., et al. 'Naïve Bayes and TF-IDF for SMS Spam Detection,' Springer, 2023.
[4] Ali, M., & Singh, P. 'Integrated Spam Detection and Sender Blocking System,' Elsevier, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY