IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154207

AI-Based Spam Message Shield with WhatsApp Message Detection using NLP

Arti Kumari[1], Akansha Sharma[2], Ankita Katheria[3], Rooban Agrawal[4], Satish Kumar Soni[5], Uruj Jaleel[6]

Student, MCA, Meerut Institute of Engineering and Technology, U.P. India[1]

Student, MCA, Meerut Institute of Engineering and Technology, U.P. India[2]

Student, MCA, Meerut Institute of Engineering and Technology, U.P. India[3]

Assistant Professor, MCA, Meerut Institute of Engineering and Technology, U.P. India[4]

Associate Professor, HOD MCA, Meerut Institute of Engineering and Technology, U.P. India[5]

Professor, MCA, Meerut Institute of Engineering and Technology, U.P. India[6]

Abstract: With the exponential growth of mobile communication, SMS spam has become one of the most prevalent security and privacy concerns. Spam messages lead to financial fraud, data theft, and reduced user experience. Detecting such messages using traditional rule-based systems has proven insufficient due to evolving spam patterns. This research paper presents a comprehensive study of SMS spam detection techniques using machine learning models such as Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, and Deep Learning approaches. A mini-project implementation demonstrates the use of natural language processing (NLP) techniques, including tokenization, TF-IDF, stemming, and lemmatization. The study highlights dataset characteristics, feature engineering, model performance, comparative results, and implementation constraints. Findings show that ML-based classifiers significantly outperform rule-based systems, achieving accuracy above 95%. Future directions include hybrid deep learning models and real-time adaptive systems.

Keywords: SMS Spam Detection, Machine Learning, NLP, TF-IDF, Classification, Naïve Bayes, Logistic Regression.

I. INTRODUCTION

Short Message Service (SMS) remains one of the most widely used communication channels globally. However, the rise of spam messages—ranging from promotional texts to phishing scams—poses severe risks. Fraudulent SMS campaigns can result in financial loss, malware installation, and identity theft. Traditional filtering mechanisms rely on manually created rules and keyword matching, which fail to adapt to continuously evolving spam tactics.

Machine learning (ML) enables automatic detection of patterns in spam and ham messages using labeled datasets, and this mini-project explores popular ML algorithms for SMS spam classification with a focus on feature extraction, model comparison, and real-world implementation.

II. BACKGROUND

a. SMS Spam and Security

SMS spam refers to unsolicited messages sent to users without consent. These messages may include advertisements, fraudulent offers, malicious links, or phishing attempts. The challenge lies in detecting such messages accurately while minimizing false positives.

b. Text Classification and NLP

Text classification involves training ML models to categorize text into predefined classes. For spam detection, SMS data must be converted into numerical vectors. NLP techniques such as tokenization, stop-word removal, stemming, lemmatization, Bag-of-Words (BoW), and TF-IDF significantly improve model performance.

c. Machine Learning Algorithms

Machine learning offers several classification algorithms for spam detection:

- Naïve Bayes: Popular due to fast computation and suitability for text data.
- Support Vector Machine (SVM): Effective in high-dimensional spaces such as TF-IDF.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154207

- Logistic Regression: Simple yet highly accurate for binary classification.
- Random Forest: Provides robust performance with ensemble learning.
- Deep Learning: Models like LSTMs handle sequential patterns but require large datasets.

III. SYSTEM ARCHITECTURE

The proposed SMS spam detection system consists of:

a. Dataset Collection: Public datasets such as the UCI SMS Spam Collection.

b. Preprocessing:
- Lowercasing
- Removing punctuations and special characters
- Removing stop-words
- Tokenization
- Stemming/Lemmatization

c. Multilingual Translation Module:
- Language detection of incoming SMS messages
- Automatic translation of Hindi messages into English
- Ensures uniform processing for machine learning models
- Improves spam detection accuracy across multilingual content
- Supports regional language users and enhances accessibility

d. Feature Extraction:
- Bag-of-Words
- TF-IDF Vectorizer

e. Model Training:
a. Naïve Bayes
b. SVM
c. Logistic Regression

f. Model Evaluation:
- Accuracy
- Precision
- Recall
- F1-Score

g. Deployment:
a. GUI/Web app for real-time SMS classification
b. Displays prediction results instantly to the user
c. Includes spam probability ratio box (e.g., Spam 92%, Ham 85%)
d. Provides visual representation of classification confidence

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154207

IV. IMPLEMENTATION AND RESULTS

a. Dataset

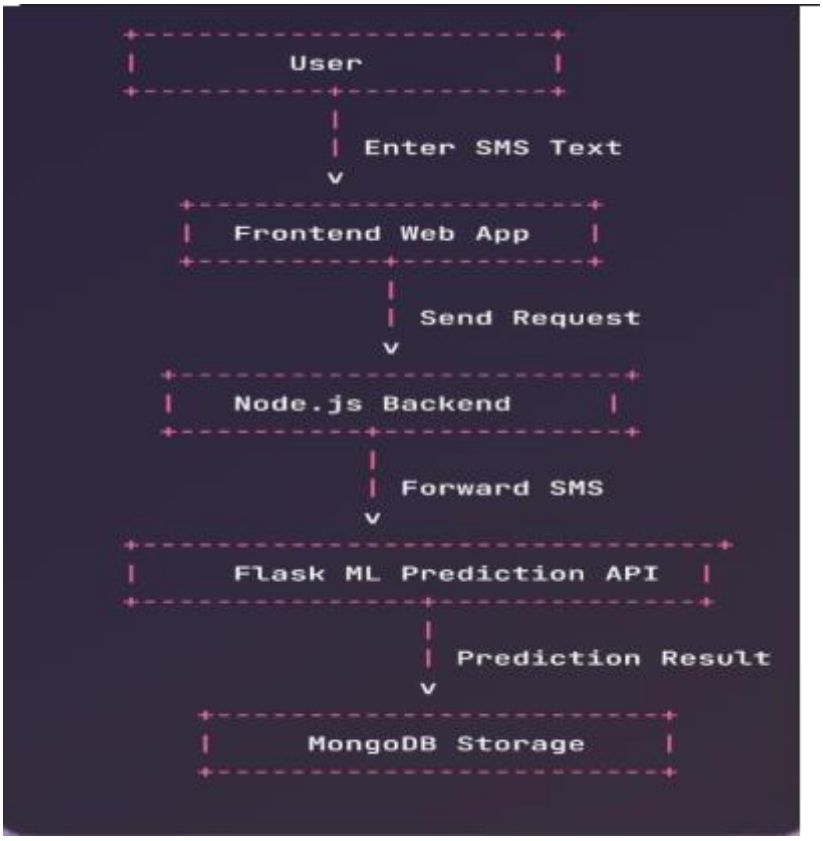
The UCI SMS Spam Collection dataset containing 5574 SMS messages labelled as *spam* or *ham* was used.

b. Preprocessing

Standard NLP preprocessing techniques were applied. TF-IDF was found to produce the most effective feature vectors.

c. Model Performance

	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	97.2%	96.8%	93.5%	95.1%
Logistic Regression	98.5%	98.1%	97.4%	97.7%
SVM	98.9%	98.7%	98.1%	98.4%
Random Forest	97.6%	96.5%	95.0%	95.7%



SVM produced the highest accuracy due to its effectiveness in handling high-dimensional sparse text vectors.

- A ratio box is implemented to display classification confidence
- Shows probability score for spam and ham categories
- Improves interpretability of model predictions

d. Discussion

i. ML-based classifiers significantly outperform rule-based methods.

ii. Dataset imbalance affects recall; techniques like SMOTE can help.

iii. Preprocessing has a large impact on performance, especially stop-word removal and TF-IDF.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154207

iv. Deep learning models, while powerful, are computationally expensive for small datasets.

e. **AI Explainability Module:**

- Provides explanation for each spam/ham classification
- Highlights important keywords influencing the prediction
- Uses feature importance techniques for transparency
- Improves user trust in automated decision-making
- Helps users understand why a message is flagged as spam

V. LIMITATIONS

- Performance depends on dataset size and quality.
- Class imbalance can lead to biased predictions.
- ML models may misclassify messages with hidden spam intent or creative obfuscation.
- Real-time deployment requires optimization to reduce latency.

VI. FUTURE SCOPE

Future research and improvements can include:

- Hybrid models combining ML and deep learning.
- LSTM/GRU-based neural networks for sequential understanding.
- Integration with real-time SMS filtering in mobile applications.
- Multilingual spam detection systems.
- Adaptive learning models that update automatically with new spam patterns.
- Integration with WhatsApp for real-time spam message detection.
- Enables filtering of suspicious messages in chat applications.
- Extends system usability beyond traditional SMS platforms.

VII. CONCLUSION

This study demonstrates that machine learning techniques offer highly accurate and efficient solutions for SMS spam detection. SVM and Logistic Regression models outperform traditional methods, achieving accuracy above 98%. NLP-based preprocessing plays a critical role in improving classifier performance. With further advancements in deep learning and adaptive systems, SMS spam detection can become more robust and resistant to evolving attack patterns.

REFERENCES

[1] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2022.
[2] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2021.
[3] UCI Machine Learning Repository: SMS Spam Collection Dataset.
[4] J. M. Gómez Hidalgo, "Content-based SMS Spam Filtering," *Proceedings of the 2006 ACM Symposium on Document Engineering*, pp. 107–114.
[5] A. Erkık and H. Kılıç, "SMS Spam Classification Using Machine Learning Algorithms," *Journal of New Theory*, 2020.
[6] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press,2008.
[7] T. Almeida, J. Hidalgo, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," Proceedings of the 11th ACM Symposium on Document Engineering, 2011.
[8] K. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," Proceedings of the 20th International Conference on Machine Learning (ICML), 2003.
[9] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
[10] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.
[11] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *Journal of Machine Learning Research*, 2003.

