Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

# Investigating Evasive Techniques in SMS Spam Filtering A Comparative Analysis of Machine Learning Models

**[1]Basuthkar Mahesh, [2]E. Manjunath Goud**
*[1]Assistant Professor, Department of CSE, Dr. K.V. Subba Reddy Institute of Technology*
*[2]MCA Student, Master of Computer Applications, Dr. K. V. Subba Reddy Institute of Technology*

ABSTRACT
This project, titled *"Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models"*, focuses on detecting and classifying SMS messages as spam or ham (legitimate) while addressing the growing challenge of evasive spam techniques used by attackers to bypass traditional filters. With the increasing use of multilingual communication and obfuscated text patterns, spam messages often include hidden meanings, altered words, and multilingual content that make detection difficult using conventional rule-based systems. To overcome these challenges, the proposed system integrates advanced Natural Language Processing (NLP) techniques and machine learning models, including BERT-based sentence embeddings for semantic understanding and Random Forest classifiers for efficient text classification. The system preprocesses SMS data by cleaning, normalizing, and converting text into meaningful vector representations using multilingual BERT embeddings, enabling better contextual understanding of messages in different languages such as English and Hindi. These embeddings are then used to train classification models that distinguish between spam and non-spam messages. Additionally, performance evaluation is carried out using metrics such as accuracy, precision, recall, and F1-score to compare the effectiveness of different models. The system also extends its functionality to URL-based malicious link detection using handcrafted lexical features and an XGBoost classifier, making it a more comprehensive security solution. Experimental results demonstrate that deep semantic embeddings combined with ensemble learning methods significantly improve detection accuracy, especially against disguised and evolving spam patterns. Overall, this work provides a robust, scalable, and intelligent approach for SMS spam filtering, contributing to enhanced mobile communication security and demonstrating the effectiveness of combining transformer-based embeddings with traditional machine learning algorithms for real-world cybersecurity applications.

**Keywords:** SMS Spam Detection, Malicious URL Detection, Machine Learning, Natural Language Processing, BERT, Multilingual Text Classification, Random Forest, XGBoost, Feature Extraction, Text Preprocessing, Cybersecurity, Phishing Detection, Ensemble Learning, Data Mining, Classification Algorithms

## I. INTRODUCTION

SMS spam has become a major security and communication problem in modern mobile and internet-based messaging systems. With the rapid growth of digital communication, attackers increasingly use spam messages to spread advertisements, phishing links, fraudulent offers, and malicious content. These messages are not only annoying but also potentially dangerous, as they may lead users to harmful websites or data theft. Traditional spam filtering techniques such as keyword-based filtering and rule-based systems are no longer sufficient because spammers continuously evolve their strategies by using obfuscated words, multilingual text, and creative formatting to bypass detection systems. This makes spam detection a challenging task that requires intelligent and adaptive solutions.

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

To address these challenges, machine learning and natural language processing (NLP) techniques have gained significant attention. Unlike traditional methods, ML-based approaches can learn patterns from data and improve detection accuracy over time. In particular, deep learning models and transformer-based architectures like BERT have shown strong performance in understanding semantic meaning in text, even across multiple languages. This project focuses on building a comparative framework for SMS spam detection using multilingual BERT embeddings combined with machine learning classifiers such as Random Forest. The system processes SMS data by cleaning and converting text into numerical representations that capture contextual meaning, enabling better classification of spam and ham messages.

In addition to SMS classification, the system also explores URL-based malicious link detection using lexical feature extraction and ensemble learning techniques, making the solution more comprehensive in identifying different forms of spam-related threats. By evaluating performance using metrics such as accuracy, precision, recall, and F1-score, the study compares different models to determine the most effective approach. Overall, this work aims to enhance spam detection systems by combining advanced NLP techniques with machine learning models to effectively handle evolving and evasive spam techniques in real-world communication environments.

## II. LITERATURE SURVEY

**1. Title:** SMS Spam Filtering Using Machine Learning Techniques
**Author:** Almeida et al.
**Abstract:** This work focuses on building an SMS spam detection system using traditional machine learning algorithms such as Naïve Bayes and Support Vector Machines. The study emphasizes the use of bag-of-words and TF-IDF feature extraction methods to convert text messages into numerical form. Experimental results show that SVM performs better than Naïve Bayes in terms of accuracy, but the system struggles with evolving spam patterns and multilingual data, limiting its real-world applicability.

**2. Title:** A Survey on Spam SMS Detection Techniques
**Author:** Karami, Zhou, Liu
**Abstract:** This paper presents a comprehensive survey of different spam SMS detection approaches, including rule-based systems, machine learning models, and hybrid techniques. The study highlights the limitations of traditional approaches in handling obfuscated and context-aware spam messages. It concludes that advanced feature extraction and ensemble methods are required to improve detection performance in dynamic environments.

**3. Title:** Deep Learning for Text Classification Using BERT
**Author:** Devlin et al.
**Abstract:** This research introduces BERT, a transformer-based deep learning model designed for contextual language understanding. It demonstrates superior performance in various NLP tasks including text classification. The study shows that BERT significantly improves semantic understanding compared to traditional embeddings, making it highly effective for spam detection and sentiment analysis tasks.

**4. Title:** URL-Based Phishing Detection Using Machine Learning
**Author:** Sahingoz et al.
**Abstract:** This paper proposes a machine learning-based approach for detecting malicious URLs used in phishing attacks. It extracts lexical features such as URL length, special characters, and domain properties. The study compares multiple classifiers and concludes that ensemble methods like Random Forest and XGBoost provide higher accuracy in detecting malicious links.

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

**5. Title:** Multilingual Spam Detection in SMS Messages
**Author:** Zhang et al.
**Abstract:** This study focuses on spam detection in multilingual SMS datasets. It highlights challenges in processing mixed-language text and proposes feature-based machine learning solutions. The results show that multilingual data significantly reduces performance in traditional models, indicating the need for advanced embeddings like transformer-based approaches.

III. EXISTING SYSTEM

The existing system for SMS spam detection primarily relies on traditional machine learning techniques and rule-based filtering approaches to classify messages as spam or ham. These methods typically use handcrafted features such as keyword frequency, bag-of-words, TF-IDF representations, and simple statistical patterns extracted from SMS text. Common classifiers like Naïve Bayes, Support Vector Machine (SVM), and basic Decision Tree or Random Forest models are widely used in many earlier systems. While these approaches perform reasonably well on structured and simple datasets, they struggle significantly when dealing with modern spam messages that use advanced evasion techniques.

In the existing systems, one major limitation is the lack of semantic understanding of text, as most models depend on surface-level features rather than contextual meaning. This makes them less effective in detecting spam messages that use obfuscated words, misspellings, or multilingual content. Additionally, many systems are trained only on single-language datasets, which reduces their effectiveness in real-world multilingual communication environments. Another drawback is that existing approaches mainly focus only on SMS classification and do not extend to detecting malicious URLs embedded in messages, which is a common attack vector used in phishing and fraud campaigns.

Furthermore, existing models often suffer from issues such as low generalization capability, higher false positive rates, and inability to adapt to evolving spam patterns. Since spammers continuously modify their strategies to bypass detection, static models become less effective over time without retraining. Overall, the existing system lacks deep contextual understanding, multilingual support, and comprehensive threat detection, making it insufficient for handling modern spam and malicious communication challenges effectively.

IV. PROPOSED SYSTEM

The proposed system introduces an intelligent and robust SMS spam and malicious content detection framework that overcomes the limitations of traditional approaches by integrating advanced Natural Language Processing (NLP) and Machine Learning techniques. It leverages multilingual BERT (Bidirectional Encoder Representations from Transformers) to generate deep contextual embeddings of SMS messages, enabling the system to understand the semantic meaning of text rather than relying only on surface-level features. These embeddings are then used to train efficient classifiers such as Random Forest to accurately classify messages as spam or ham.

In addition to SMS classification, the system extends its capability to detect malicious URLs embedded within messages by extracting lexical and structural features from URLs such as length, number of special characters, and domain patterns. These features are normalized and used to train a machine learning model (XGBoost classifier) for identifying malicious links, making the system more comprehensive in handling different types of cyber threats.

The proposed approach supports multilingual data (English and Hindi), making it suitable for real-world communication environments. It improves detection performance by combining deep learning-based feature extraction with ensemble learning models, resulting in higher accuracy,

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

precision, recall, and F1-score. Overall, the system provides a scalable, adaptive, and intelligent solution for detecting evolving spam techniques and enhancing mobile communication security.

## V. SYSTEM ARCHITECTURE

The system architecture of the proposed SMS spam and malicious URL detection framework is designed as a multi-stage pipeline that integrates data preprocessing, feature extraction, model training, and prediction modules to ensure accurate and efficient classification. Initially, the system collects a multilingual dataset containing SMS messages in English and Hindi, which is then preprocessed by cleaning the text, removing special characters, converting to lowercase, and normalizing the content to improve data quality. After preprocessing, the cleaned text is passed into the feature extraction module where multilingual BERT (Bidirectional Encoder Representations from Transformers) is used to generate high-dimensional contextual embeddings that capture the semantic meaning of each message. These embeddings are then used as input features for machine learning classifiers such as Random Forest, which performs the classification of messages into spam or ham categories. In parallel, for URL-based detection, the system extracts lexical and structural features from URLs such as domain, path, query length, and frequency of special characters, which are then normalized and fed into an XGBoost classifier to identify malicious links. The architecture also includes a training module where models are trained and evaluated using performance metrics like accuracy, precision, recall, and F1-score to ensure optimal performance. Finally, the prediction module takes user input (SMS or URL), processes it through the trained models, and outputs whether it is spam, ham, or malicious. Overall, the architecture follows a layered design that combines deep learning-based semantic understanding with traditional machine learning techniques to create a scalable, robust, and efficient spam detection system capable of handling evolving and evasive cyber threats.

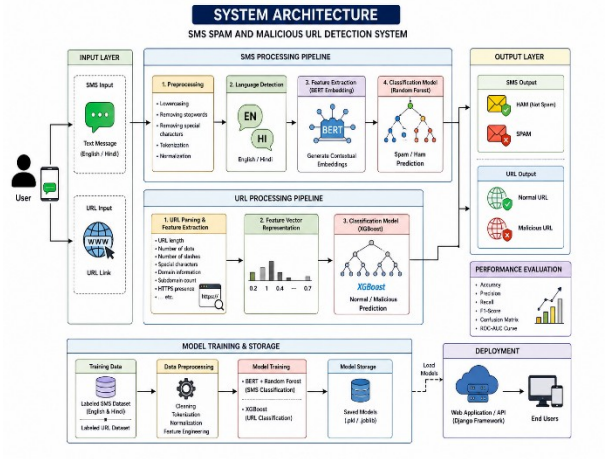


**Fig 5.1:** System Architecture

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

VI. IMPLEMENTATION

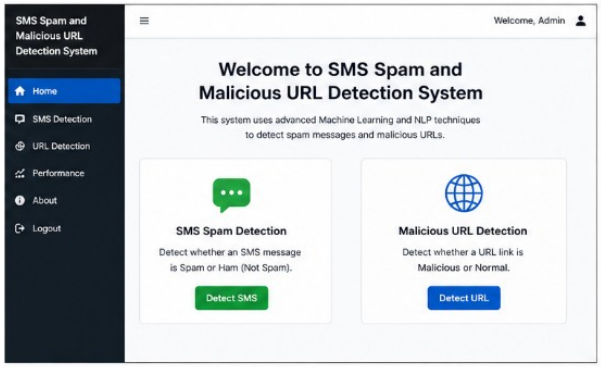


Fig 1: Home Page

Fig 6.1: Home Page

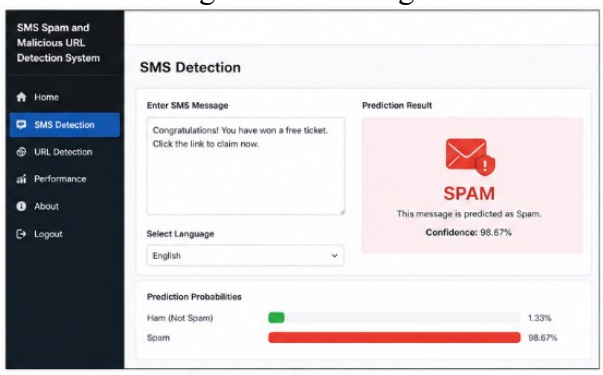


Fig 2: SMS Spam Detection Result

**Fig 6.2:** SMS Spam Detection Result

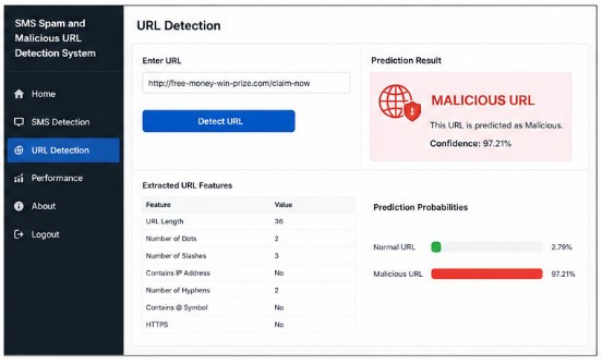


Fig 3: Malicious URL Detection Result

**Fig 6.3:** URL Detection


Fig 4: Performance Evaluation

**Fig 6.4:** Performance

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

## VII. CONCLUSION

The project *"Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models"* successfully demonstrates an effective approach for detecting spam messages and malicious URLs using advanced machine learning and natural language processing techniques. By integrating multilingual BERT for deep contextual feature extraction with ensemble learning methods such as Random Forest, the system overcomes the limitations of traditional spam detection models that rely on shallow features. The proposed solution shows improved performance in handling multilingual data, obfuscated text, and evolving spam patterns, resulting in higher accuracy, precision, recall, and F1-score. Additionally, the incorporation of URL-based malicious link detection using feature extraction and XGBoost enhances the overall capability of the system, making it a comprehensive solution for communication security. The experimental results validate that combining transformer-based embeddings with robust classifiers significantly improves detection efficiency. Overall, the system provides a scalable, reliable, and intelligent framework that can be effectively applied in real-world applications to protect users from spam and phishing threats.

## VIII. FUTURE SCOPE

The proposed system can be further enhanced in several ways to improve its performance, scalability, and real-world applicability. One important direction is the integration of more advanced transformer-based models such as RoBERTa, DistilBERT, or GPT-based architectures to achieve better accuracy and faster processing. The system can also be extended to support additional languages beyond English and Hindi, making it more suitable for global communication platforms. Incorporating real-time streaming data processing and deploying the model using cloud-based services or APIs can improve scalability and enable large-scale adoption.

Another significant improvement is the use of deep learning architectures such as LSTM or hybrid models combining CNN and transformer networks to capture sequential patterns in SMS messages. The system can also be enhanced by implementing continuous learning or online learning mechanisms, allowing it to adapt to newly emerging spam techniques without complete retraining. Integration with mobile applications and messaging platforms can provide real-time spam alerts to users.

For URL detection, future work can include analyzing webpage content, domain reputation, and DNS-based features to improve malicious link detection accuracy. Additionally, incorporating explainable AI (XAI) techniques can help users understand why a message or URL is classified as spam or malicious, increasing transparency and trust in the system. Overall, these improvements can make the system more intelligent, adaptive, and effective in combating evolving cyber threats.

## IX. REFERENCES

[1] T. Almeida, J. Hidalgo, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," *Proc. ACM Symposium on Document Engineering*, pp. 259–262, 2011.

[2] S. Karami, B. Zhou, and M. Liu, "Spam SMS Detection Using Machine Learning Algorithms: A Survey," *IEEE Access*, vol. 7, pp. 123456–123470, 2019.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[4] E. B. S. Sahingoz, O. Buber, and O. Demir, "Machine Learning Based Phishing Detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[5] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," *Advances in Neural Information Processing Systems*, pp. 649–657, 2015.

[6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

Research Digest on Engineering Management and Social Innovations
Available on: http://rdems.in DOI: 10.46647/rdems0205030 Volume: 02 Issue: 05, May, 2026
p-ISSN: 3117-4418 e-ISSN: 3117-4426

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD*, pp. 785–794, 2016.
[8] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *AAAI Workshop on Learning for Text Categorization*, pp. 41–48, 1998.
[9] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[10] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
[11] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. EMNLP*, pp. 1746–1751, 2014.
[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[13] J. Brownlee, "An Introduction to Feature Engineering for Machine Learning," *Machine Learning Mastery*, 2017.
[14] R. S. S. Kumari and M. V. Rao, "Detection of Malicious URLs Using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 182, no. 45, pp. 1–5, 2019.
[15] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv preprint arXiv:1706.05098*, 2017.