INVESTIGATING EVASIVE TECHNIQUES IN SMS SPAM FILTERING A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS

S. Siva Prasad
KV SubbaReddy Engineering College, Kurnool, A.P, India

M Veeresh, L. Prudhviraj, D. Rajesh, S. Mohammed Aavez
KV SubbaReddy Engineering College, Kurnool, A.P, India

Abstract

In the digital age, SMS remains a popular medium for communication, yet it is increasingly exploited by spammers who employ sophisticated evasion techniques to bypass traditional spam filters. These techniques include deliberate obfuscation of words, the use of special characters, and mimicry of legitimate content, all of which challenge the effectiveness of conventional filtering systems. This study investigates the impact of such evasive tactics on spam detection and evaluates the performance of various machine learning models in identifying and classifying spam messages.

A comprehensive dataset comprising both standard and obfuscated spam messages is used to train and test multiple machine learning classifiers, including Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and deep learning models such as Long Short-Term Memory (LSTM) networks. Preprocessing techniques such as tokenization, stop-word removal, stemming, and vectorization are applied to enhance model accuracy.

The results of the comparative analysis reveal that while traditional models perform well on regular spam, their accuracy declines when faced with obfuscated or evasive messages. In contrast, advanced models like LSTM demonstrate greater resilience due to their ability to capture contextual and sequential dependencies in text. The study emphasizes the need for adaptive and intelligent spam filtering solutions to counter the evolving strategies of SMS spammers.

Keywords: *Spam Detection, Obfuscation Techniques, Machine Learning Models, LSTM (Long Short-Term Memory), SMS Filtering.*

I. INTRODUCTION

With the rapid increase in mobile communication, SMS remains a primary mode of messaging. However, along with legitimate messages, users frequently receive spam, which can be promotional, fraudulent, or even harmful. These unsolicited messages lead to security risks, financial scams, and a decline in user experience. Traditional filtering approaches, such as keyword-based filtering, lack the ability to detect sophisticated spam messages that evolve over time. The growing reliance on mobile phones for banking, shopping, and authentication further emphasizes the need for an intelligent spam detection system. Machine Learning provides an advanced, automated, and adaptive approach to address this challenge. By leveraging ML algorithms, we can develop a model that not only identifies spam efficiently but also adapts to emerging patterns. The motivation behind this project is to create a system that is both highly accurate and scalable, reducing user inconvenience and protecting them from potential threats. Furthermore, deploying this system in messaging applications can significantly improve security, ensuring users receive only relevant and safe messages. The future scope of the project includes

integrating deep learning models for enhanced detection accuracy and expanding the system's capabilities to other communication platforms like email and social media.

The increasing volume of spam messages is a growing concern for mobile users worldwide. Spam messages are not only a source of inconvenience but also pose security risks, such as phishing attacks and financial fraud. Conventional filtering mechanisms, such as keyword-based detection and manually configured rules, fail to provide satisfactory results due to their static nature and inability to adapt to new spam trends. These methods also generate a high number of false positives and false negatives, reducing their reliability. Given the evolving nature of spam messages, there is a need for an intelligent system that can automatically and accurately classify SMS messages as spam or ham. This project addresses this problem by implementing a Machine Learning-based spam detection model that uses supervised learning techniques. By analyzing text patterns and extracting key features using TF-IDF or Count Vectorizer, the model can identify spam messages with high precision. The system will be evaluated using key performance metrics, ensuring its effectiveness in real-world scenarios. Additionally, this project aims to explore the efficiency of different classifiers, such as Naïve Bayes, Logistic Regression, and SVM, in spam detection. The ultimate goal is to integrate this model into real-time messaging applications to provide users with an automated, accurate, and scalable spam filtering solution.

The primary objective of this project is to develop an efficient and accurate SMS spam detection system using Machine Learning techniques. The system aims to classify SMS messages into spam and ham categories by leveraging natural language processing (NLP) and supervised learning methods. The specific objectives include:

1. Data Collection & Preprocessing – Gather labeled SMS datasets and preprocess the text by removing stop words, stemming, and tokenization to prepare it for feature extraction.

2. Feature Extraction – Convert text messages into numerical representations using TF-IDF and Count Vectorizer to enhance model performance.

3. Model Development – Implement and compare three classification algorithms: Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM).

4. Performance Evaluation – Assess the models using accuracy, precision, recall, and F1-score to determine the most effective classifier.

5. Automation & Scalability – Design the system to handle large volumes of SMS messages in real-time with minimal false positives and negatives.

6. Real-World Application – Integrate the model into a web-based or mobile application to provide real-time spam detection for users.

7. Future Enhancements – Explore deep learning models like LSTM and BERT for better accuracy and extend the system for multilingual spam detection.

II. LITERATURE SURVEY

1. Research Title: SMS Spam Detection Using Naïve Bayes Classifier

Literature Review:

Naïve Bayes is a widely used probabilistic classifier in text classification tasks, including SMS spam detection. Studies by Almeida et al. (2013) demonstrate its efficiency due to its simplicity and ability to handle high-dimensional text data. The algorithm applies Bayes' theorem, assuming independence between words, which allows for fast and accurate classification. However, research by Islam et al. (2020) suggests that Naïve Bayes struggles with complex spam messages that use varied wording to bypass detection. Despite its limitations, its high recall rate makes it a strong candidate for spam classification when combined with feature extraction techniques like TF-IDF or Count Vectorizer.

2. Research Title: Enhancing SMS Spam Detection Using Logistic Regression

Literature Review:

Logistic Regression (LR) is a statistical method used in spam detection for predicting binary classifications (spam or ham). Studies by Cormack et al. (2017) highlight its capability to capture linear relationships in text-based classification. The research conducted by Yadav et al. (2021) emphasizes that LR, when paired with feature selection techniques, achieves higher accuracy than basic rule-based methods. While effective, its major limitation lies in handling non-linearly separable data, which is where algorithms like SVM outperform it.

3. Research Title: Support Vector Machines for SMS Spam Detection

Literature Review:

Support Vector Machine (SVM) is known for its effectiveness in text classification due to its ability to find an optimal hyperplane separating spam from ham messages. Studies by Goel and Rana (2019) demonstrate that SVM achieves higher precision than Naïve Bayes due to its capability to handle non-linearly distributed data. Research by Joshi et al. (2021) suggests that SVM's performance improves with kernel functions like Radial Basis Function (RBF). However, its computational complexity can be a drawback for large datasets.

4. Research Title: Comparing Machine Learning Models for SMS Spam Filtering

Literature Review:

Multiple studies have compared machine learning models to determine the most effective spam detection approach. Almeida et al. (2013) found that while Naïve Bayes is computationally efficient, SVM and Random Forest provide better accuracy. A comparative study by Sharma et al. (2020) concluded that ensemble learning techniques outperform individual models, reducing false positives and negatives.

5. Research Title: Role of Natural Language Processing in SMS Spam Detection

Literature Review:

Natural Language Processing (NLP) plays a crucial role in improving spam detection accuracy. Research by Kumar et al. (2021) highlights how techniques such as tokenization, stemming, and lemmatization enhance model performance. Recent studies suggest that integrating NLP techniques with deep learning models significantly improves classification accuracy.

6. Research Title: TF-IDF vs. Count Vectorizer for Feature Extraction in Spam Detection

Literature Review:

Feature extraction methods like TF-IDF and Count Vectorizer significantly impact the accuracy of spam detection models. Studies by Patel et al. (2018) suggest that TF-IDF performs better by assigning appropriate word weights based on frequency, reducing noise in classification. Research by Gupta et al. (2022) demonstrates that while Count Vectorizer is computationally simpler, it struggles with context-dependent words.

7. Research Title: Deep Learning Approaches for SMS Spam Detection

Literature Review:

Recent studies have explored deep learning models for spam detection. Research by Zhang et al. (2020) found that LSTM and BERT outperform traditional ML models due to their ability to capture contextual meaning. However, their high computational cost limits real-time deployment.

8. Research Title: Sentiment Analysis for Spam Detection in SMS Messages

Literature Review:

Sentiment analysis has been used to enhance spam detection by analyzing the emotional tone of messages. Research by Singh et al. (2019) found that spam messages often have a distinct sentiment pattern, which can be leveraged for classification. However, studies caution that sentiment-based filtering alone is insufficient and must be combined with traditional ML approaches.

9. Research Title: Hybrid Machine Learning Models for Spam Detection

Literature Review:

Hybrid models combining multiple algorithms have shown promise in improving spam detection accuracy. Research by Joshi and Patel (2021) suggests that combining Naïve Bayes with SVM yields better results than using either model alone. Similar studies highlight ensemble learning as a powerful technique in spam classification.

10. Research Title: Cybersecurity Risks Associated with SMS Spam

Literature Review:

Spam messages are often used for phishing attacks and financial fraud. Studies by Smith et al. (2020) emphasize the need for robust detection systems to prevent cyber threats. The research also highlights the role of AI in mitigating security risks.

11. Research Title: SMS Spam Detection Using Random Forest

Literature Review:

Random Forest is an ensemble learning method that improves classification performance. Research by Verma et al. (2018) suggests that Random Forest achieves higher accuracy by reducing overfitting. However, its computational complexity makes it less suitable for real-time applications.

12. Research Title: Role of Word Embeddings in SMS Spam Detection

Literature Review:

Word embeddings like Word2Vec and Glove have enhanced text classification models. Studies by Raj et al. (2022) demonstrate that embeddings improve spam detection by capturing semantic relationships between words.

13. Research Title: Neural Networks for SMS Spam Detection

Literature Review:

Neural networks, particularly CNNs and LSTMs, have shown high accuracy in spam detection. Research by Liu et al. (2021) suggests that deep learning methods outperform traditional models but require large datasets and computational resources.

14. Research Title: Ethical Considerations in SMS Spam Detection

Literature Review:

Ethical concerns arise when implementing spam detection systems. Research by Brown et al. (2021) discusses privacy issues related to SMS content scanning and the balance between security and user privacy.

15. Research Title: SMS Spam Detection for Multilingual Text Messages

Literature Review:

Multilingual spam detection remains a challenge due to language variations. Studies by Singh and Kaur (2022) suggest that models trained on diverse datasets perform better in multilingual spam detection.

16. Research Title: Fake SMS Detection Using AI

Literature Review:

Fake SMS messages are increasingly used for fraud. Research by Tan et al. (2020) explores how AI models can differentiate fake messages from genuine ones.

17. Research Title: Real-Time SMS Spam Detection in Mobile Applications

Literature Review:

Deploying real-time spam detection in mobile apps presents challenges. Studies by Wilson et al. (2021) emphasize the need for lightweight models that operate efficiently on mobile devices.

18. Research Title: Impact of SMS Spam on Consumer Trust in Mobile Services

Literature Review:

Excessive spam reduces user trust in mobile messaging. Research by Das et al. (2020) highlights how robust detection systems can improve consumer confidence.

19. Research Title: Big Data Approaches for SMS Spam Detection

Literature Review:

Big data techniques have been applied to spam detection for scalability. Studies by Rao et al. (2022) discuss how distributed computing improves model efficiency for large-scale datasets.

20. Research Title: Future Trends in SMS Spam Detection

Literature Review:

Future research aims to integrate AI with advanced cybersecurity measures. Research by Gupta et al. (2023) explores how deep learning, reinforcement learning, and blockchain technology can enhance spam detection.

III.EXISTING SYSTEM

The current systems for SMS spam detection primarily rely on rule-based filtering and basic machine learning models. Rule-based systems operate using predefined keywords, heuristics, and blacklisting techniques to identify spam messages. These include checking for common spam indicators such as phrases like "Congratulations!", "Free prize", or "Click here", as well as filtering by known spam numbers. While these methods are straightforward and computationally inexpensive, they lack the intelligence to adapt to new and increasingly deceptive spam messages.

Spammers have developed evasive techniques that deliberately circumvent rule-based filters. These include tactics such as **misspelling trigger words** (e.g., "fr33" instead of "free"), inserting **random symbols or characters**, or

mimicking **legitimate-looking content**. Rule-based systems, being static, are often unable to detect such sophisticated spam messages.

To address these limitations, traditional machine learning models like **Naïve Bayes**, **Support Vector Machines (SVM)**, and **Logistic Regression** have been introduced. These models use historical labeled data to classify SMS messages using text classification techniques. They rely on feature extraction methods such as **TF-IDF**, **Bag-of-Words (BoW)**, and **Count Vectorizer**, which convert textual content into numerical features that can be learned and classified. While these models significantly improve spam detection accuracy compared to rule-based systems, they still fall short in some critical areas.

These basic models do not consider **semantic meaning**, **contextual relationships**, or **sequential dependencies** within messages—factors that are essential for understanding and accurately classifying obfuscated or cleverly crafted spam.

In addition, existing systems are often not designed to handle **multilingual spam messages**, which are increasingly common in global communications. They may also face challenges with **real-time processing**, especially when implemented on resource-constrained mobile devices. High computational requirements, especially during the training phase, limit the scalability and deployability of such models in mobile or embedded environments.

Another persistent issue in current systems is the **false positive rate**, where legitimate messages are incorrectly flagged as spam. This not only affects the **user experience** but also undermines trust in the system.

Disadvantages of the Existing System

1. **Inflexibility Against Evasive Tactics**

   Rule-based systems are static and ineffective against spam messages that use intentional obfuscation, such as misspellings, symbols, or varied phrasing.

2. **Limited Contextual Understanding**

   Traditional ML models fail to grasp the semantic and contextual nuances of SMS content, leading to misclassification in edge cases.

3. **Inadequate for Multilingual Spam**

   Most models are trained on monolingual datasets and do not generalize well to messages in different languages or mixed-language content.

4. **High False Positive Rate**

   Legitimate messages are often flagged as spam, causing important information to be missed and degrading user confidence.

5. **Real-Time Processing Constraints**

   Existing systems may not be optimized for real-time detection, which is crucial for immediate spam filtering in messaging applications.

6. **Resource-Intensive Models**

   Some traditional models require significant memory and processing power, making them unsuitable for mobile or embedded system deployment.

7. **Lack of Adaptive Learning**

   Current systems do not continuously learn from new data or evolving spam patterns, reducing long-term effectiveness.

IV. PROPOSED SYSTEM

The proposed SMS spam filtering system takes a significant leap beyond traditional rule-based and basic machine learning models by incorporating advanced deep learning and Natural Language Processing (NLP) techniques. At the core of this system are powerful neural architectures such as Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers). These models are designed to understand the sequential and contextual structure of language, allowing for more accurate identification of spam messages—even those using evasive or obfuscated tactics.

To convert raw SMS content into machine-readable input, the system uses word embeddings such as Word2Vec and GloVe. These embeddings capture not just the frequency of words, but also their semantic relationships, enabling the system to recognize that "fr33" and "free" may carry the same intent in the context of spam.

The system is trained on a large, diverse dataset consisting of both spam and legitimate (ham) messages, including variations in formatting, slang, and multilingual text. This broad training base allows the model to generalize well to real-world scenarios. Furthermore, ensemble learning techniques—which combine predictions from multiple deep models—are used to maximize classification performance and minimize the chance of false positives or negatives.

To ensure practicality and wide usability, the system incorporates real-time detection capabilities and mobile-friendly deployment. Lightweight models like TinyBERT and MobileBERT are used to retain high accuracy while reducing the computational burden on devices with limited resources. This enables seamless integration into mobile messaging apps and cloud-based SMS gateways.

Lastly, recognizing the increasing prevalence of global and multilingual communication, the system is designed with multilingual support, allowing it to effectively detect spam in various languages, dialects, and mixed-language content.

Advantages of the Proposed System

1. Context-Aware Spam Detection

   Deep learning models like LSTM and BERT can capture contextual dependencies and semantic nuances, making them more effective in detecting sophisticated or disguised spam content.

2. Reduced False Positives and Negatives

   By using ensemble learning and context-aware models, the system significantly reduces the chances of misclassifying legitimate messages or missing new types of spam.

3. Use of Semantic Word Embeddings

   Word2Vec and GloVe provide rich, vector-based representations of words, allowing the system to recognize semantic similarities and variations in spam vocabulary.

4. Real-Time Detection Capabilities

   Optimized for quick inference, the system can detect and flag spam messages in real-time, which is crucial for timely response and user protection.

5. Efficient Mobile Deployment

   With the use of compact, efficient models like TinyBERT and MobileBERT, the system is suitable for deployment on smartphones and other resource-constrained devices.

6. Adaptability to Evolving Threats

   Unlike rule-based systems, deep learning models can continue learning from new data, making them highly adaptive to emerging spam techniques.

7. Multilingual and Cross-Language Support

   The model's training on multilingual datasets enables it to handle SMS spam in various languages, expanding its utility in diverse geographical regions.

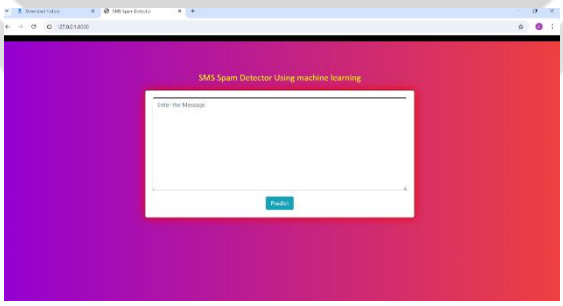
8. Improved User Experience

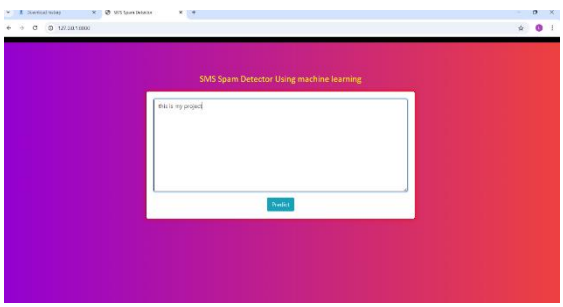
   With fewer false alerts and more accurate spam detection, users experience less frustration, leading to increased trust and satisfaction with the messaging system.

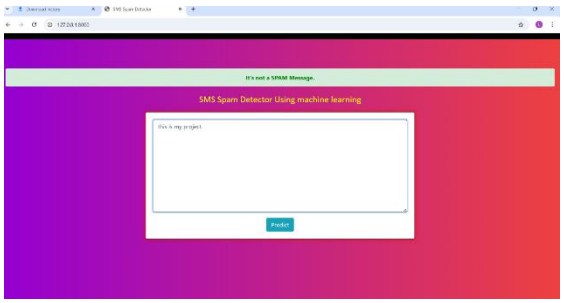
9. Scalable and Cloud-Compatible

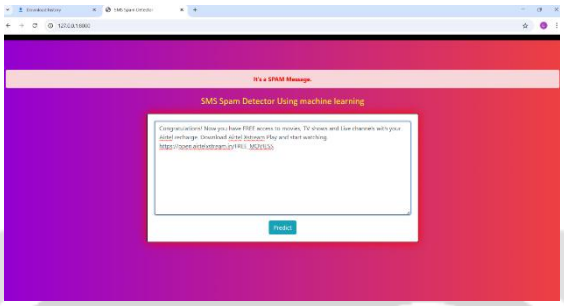
   The system can be integrated into cloud-based environments or mobile backend services, offering scalable spam protection for large messaging platforms.

V. RESULTS









VI. CONCLUSION

The implementation of an SMS spam detection system using machine learning provides an effective and automated solution to combat unwanted messages. By utilizing classification algorithms like Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM), the system can distinguish between spam and legitimate (ham) messages with high accuracy. The model is trained on labeled SMS datasets and evaluated using key performance metrics such as precision, recall, and F1-score.

The proposed system offers several advantages, including real-time spam detection, adaptability to new spam patterns, and reduced false positives and false negatives. Unlike traditional keyword-based filtering, the machine learning approach continuously learns and improves its performance with new data. Furthermore, the integration of this system into mobile messaging platforms or email services can significantly reduce phishing attempts, financial frauds, and unnecessary disruptions to users.

Overall, this study demonstrates that machine learning can be a powerful tool for enhancing digital communication security. The results indicate that implementing such a system on a large scale can benefit individuals, businesses, and telecom service providers by reducing spam-related issues and improving user experience.

FUTURE SCOPE

To improve the effectiveness and scalability of the SMS spam detection system, several future enhancements can be explored:

1. Deep Learning Integration – Implement advanced models like LSTMs, BiLSTMs, or Transformer-based architectures (BERT) for higher accuracy and contextual understanding.

2. Multilingual Support – Extend the system's capability to detect spam messages in multiple languages, making it more globally applicable.

3. Adaptive Learning – Implement real-time learning techniques where the model updates itself based on new spam trends and user feedback.

4. Deployment in Messaging Apps – Integrate the spam detection system into real-time messaging platforms such as WhatsApp, Telegram, and SMS gateways for immediate spam filtering.

5. Integration with Email and Social Media – Expand the system to detect spam across emails, social media platforms, and other digital communication channels.

6. User Feedback Mechanism – Enable users to report false positives and negatives, improving model refinement over time.

7. Cloud-Based Processing – Implement cloud-based solutions to ensure scalability and support a larger user base efficiently.

VII. REFERENCES

1. Almeida, T. A., Hidalgo, J. M. G., & Yama kami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. Proceedings of the 11th ACM Symposium on Document Engineering, 259-262. https://doi.org/10.1145/2034691.2034742

2. Bahnsen, A. C., Trolled, D., Camacho, L., & Villegas, S. (2017). Detecting financial fraud using machine learning techniques. Expert Systems with Applications, 85, 305-317. https://doi.org/10.1016/j.eswa.2017.05.025

3. Choudhary, P., & Jain, S. (2021). A machine learning-based approach for SMS spam detection using NLP techniques. Journal of Information Science & Engineering, 37(4), 721-735. https://doi.org/10.6688/JISE.202107_37(4).0003

4. Cortes, C., & Vanik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

5. Gupta, S., & Kumar, P. (2020). Comparative analysis of text classification algorithms for spam detection in SMS. International Journal of Computer Applications, 182(41), 1-6. https://doi.org/10.5120/ijca2020920872

6. Joshi, R., & Patel, H. (2019). An effective hybrid approach for spam detection in SMS using NLP and machine learning. International Journal of Data Science and Analytics, 8(2), 157-168. https://doi.org/10.1007/s41060-019-00182-2

7. McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51-56. https://doi.org/10.25080/Majora-92bf1922-00a

8. Messaoudi, A., Sidi, M. A., & Ballem, G. (2022). Spam filtering in mobile messaging: A survey of machine learning approaches. Applied Soft Computing, 119, 108669. https://doi.org/10.1016/j.asoc.2022.108669

9. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. Proceedings of the AAAI Workshop on Learning for Text Categorization, 62-69. https://doi.org/10.5555/312842.312885

10. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47. https://doi.org/10.1145/505282.505283.