JDEBM
Journal of Digital Engineering and Business Managemet

Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

AI-BASED SPAM FILTERING WITH RESISTANCE TO EVASIVE AND OBFUSCATION TECHNIQUES

[#1]**Kalavena Swathi Priya,** *Department of MCA,*
[#2]**Mr. K. Nagendra Prasad,** *Assistant Professor, Department of MCA,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: Developing an AI-driven spam filtering system capable of detecting spam communications and defeating attackers' complex evasion and obfuscation strategies is the primary objective of this research. To circumvent common screening systems, modern spammers employ intentional misspellings, odd characters, concealed text, image-based spam, and word replacements. Instead of merely searching for keywords, the suggested Paper combines artificial intelligence, machine learning, and natural language processing to identify spam through behaviour analysis, pattern identification, and context understanding. This method increases the model's adaptability and accuracy in the real world by training it on datasets that include both legitimate and fraudulent spam messages. The project aims to improve spam detection, reduce false positives, and provide a robust, flexible, and intelligent screening system to safeguard contemporary email and messaging applications from emerging spam threats.

Keywords: *SMS spam, evasion strategies, machine learning, deep learning, adversarial examples, NLP, robust classifiers, spam filtering models*

1. INTRODUCTION

Artificial intelligence (AI) has revolutionised modern communication networks by enabling sophisticated and automated data processing across numerous digital platforms. Despite this, as social media, email, and instant chat have grown in popularity, spam emails have also increased. Phishing attacks, virus distribution, financial frauds, and unauthorised data access can all result from spam messages. Keyword matching and rule-based detection are the primary techniques used by conventional spam filtering systems to identify spam, although these techniques aren't always effective against contemporary spam campaigns. More sophisticated and flexible spam detection systems are increasingly essential as hackers devise new ways to circumvent filters.Machine learning (ML) and natural language processing (NLP) technologies are used by AI-based spam filtering systems to better identify spam and provide more accurate results. These algorithms are capable of learning from large data sets, identifying hidden patterns, and classifying communications according to their content, behaviour, and context. Unlike traditional filters, AI-driven models might be able to adjust to new spam trends and increase their efficiency over time. The system can distinguish between spam and legitimate communications while lowering the number of false positives thanks to contemporary techniques like deep learning and language analysis. As a result, AI-driven screening has emerged as a highly successful method of safeguarding contemporary digital platforms and communication networks. To prevent trash from being discovered, spammers employ obfuscation and evasive

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

strategies. In order to circumvent traditional filters, spammers frequently alter spam messages by intentionally employing misspelt words, secret characters, image-based text, word swaps, and unusual symbols. These tactics reduce the efficacy of conventional spam detection systems, making spam communications appear authentic. The project's objective is to develop an AI-based spam screening system that is very effective against devious tactics and information concealment in order to address this problem. The proposed system provides accurate, adaptable, and dependable spam identification in real-world scenarios by examining message meanings, behaviour patterns, and contextual data. Enhancing communication safety and cybersecurity is the aim.

2. LITERATURE REVIEW

Kumar et al. (2021) developed an AI-based spam prevention system that detects encrypted emails using Naïve Bayes and Support Vector Machines. Text preparation and feature extraction techniques were employed to identify hidden spam trends. The approach made it simpler to identify spam in conversations that were attempting to evade responses. However, it was less effective when attackers employed more sophisticated text manipulation techniques.

Chen et al. (2022), investigated the use of CNN and LSTM architectures to filter phishing and spam emails that conceal URLs and phrases. Features were gathered via tokenization and meaning analysis based on natural language processing (NLP). The outcomes demonstrated that the filters outperformed standard filters in preventing spam hits from other sources. The issue was that it required very large training samples and was highly costly to operate.

Patel et al. developed an adaptive spam screening system in 2022 using reinforcement learning. Based on user feedback and message classification, the system's learning model was constantly improving. It made it simpler to identify constantly evolving spam operations. However, in order for the procedure to continue to function, it had to be updated frequently.

To identify spam emails with obfuscated keywords and concealed payloads, Fernandez et al. (2023) developed a hybrid AI model that combines Random Forest and deep neural networks. Lexical, behavioural, and URL-based traits make up feature extraction. The technique significantly reduced false hits in the classification of spam. The issue was that it wasn't functioning as effectively while handling multilingual spam.

To identify evasive spam messages, Rao et al. (2024) developed a real-time AI-powered spam filtering system that makes use of transformer-based NLP models. Semantic manipulation and hidden spam patterns were exposed via contextual embeddings. The technique was effective in identifying social engineering and spam attempts. In any event, real-time implementation required a significant amount of processing power.

To combat obfuscation tactics in SMS and email spam, Hassan et al. (2025) introduced an ensemble learning-based model for spam detection that integrates XGBoost, LSTM, and attention processes. The modified text was normalised using sophisticated preprocessing techniques. Against hostile strikes, the system showed

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

exceptional precision and resilience. The main drawback was a protracted training duration due to the ensemble's intricacy.

To identify intricate spam campaigns, Williams et al. (2026) proposed an intelligent multimodal spam filtration framework that makes use of text, metadata, and behavioural analysis. Transformer-based architectures helped us better understand spam patterns and evasion strategies. In benchmarks for spam detection, the system performed better. However, the idea ran into problems because of high infrastructure costs and privacy concerns.

3. PROPOSED METHODOLOGY

To develop an AI-driven spam filtration system capable of identifying spam messages using evasive and obfuscation strategies, the project uses an experimental and analytical research methodology. The implementation of deep learning and machine learning algorithms to successfully differentiate between spam and non-spam messages is the main emphasis of the Paper. The method seeks to identify content that has been altered to bypass typical spam filters, including buried text, symbol substitutions, URL obfuscation, and deliberate spelling errors.

Data Collection

The SMS Spam Collection Dataset, the Enron Email Dataset, the Kaggle spam datasets, and the phishing URL dataset were among the publicly accessible sources from which the Paper's dataset was assembled. Both spam and legitimate messages using different obfuscation techniques—such as truncated URLs, arbitrary symbol insertions, encoded text patterns, and character substitutions—are included in the collected data. The dataset is separated into training, validation, and testing subsets in order to enable precise model evaluation and performance assessment.

Data Preprocessing

Through data preparation, the assembled text is cleaned and organised methodically to make it suitable with AI models. Preprocessing techniques include noise reduction, stemming, lemmatization, lowercase conversion, stop word deletion, and tokenization. Preprocessing methods including character normalisation, homoglyph identification, spacing correction, and symbol substitution detection are used to increase resistance against evasive spam strategies.

Feature Extraction

To identify pertinent trends and extract information from spam messages, natural language processing techniques are used. Techniques for converting textual data into numerical representations include word embeddings, character-level embeddings, N-gram analysis, Bag of Words, and TF-IDF. To increase the accuracy of spam recognition, elements such as questionable URLs, unusual symbols, keyword prevalence, and hidden character patterns are retrieved.

Model Development

For the purpose of classifying spam, several deep learning and machine learning models are developed and evaluated. Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, Convolutional Neural Network, Recurrent Neural Network, Long Short-Term Memory, and Transformer-based models like BERT are among the models used. These algorithms, which are frequently used in evasive spam attacks,

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

are made to identify contextual and sequential spam patterns.

Obfuscation Resistance Mechanism

To increase robustness against spam evasion techniques, the suggested system combines adversarial training, character-level learning, context-aware embeddings, URL reputation analysis, and attention techniques. To improve the models' capacity to identify altered text patterns and identify complex spam attacks in real-world scenarios, synthetic obfuscated spam samples are created and incorporated into the training process.

Model Training and Testing

The created models are trained using the curated dataset using optimisation techniques such as cross-validation, early stopping, dropout regularisation, and hyperparameter tweaking. Following training, the models are assessed using a variety of obfuscation techniques on spam samples that have never been encountered before in order to gauge their robustness, accuracy, and generalisation skills in successfully identifying spam messages.

Performance Evaluation

The effectiveness of the proposed spam filtering system is evaluated using metrics such as accuracy, precision, recall, F1-score, ROC-AUC score, false positive rate, and obfuscated spam detection rate. Confusion matrices are used to assess each model's classification efficacy and compare how well it can identify standard and evasive spam communications.

4. MODULES

Data Collection Module:

This module gathers unprocessed text, SMS, and email data from a variety of sources, such as user inboxes, public databases, and streaming inputs. By offering a careful blend of legitimate and spam messages, it increases training efficacy.

Data Preprocessing Module:

To sanitise and normalise text, it eliminates unnecessary elements like URLs, stopwords, and special characters. To fix hidden spam patterns, it makes use of normalisation and pattern correction algorithms.

Feature Extraction Module:

This module converts text into meaningful numerical features using TF-IDF, word embeddings, and n-grams. Additionally, it recognises the behavioural and structural characteristics used in spam detection.

Obfuscation Handling Module:

It finds and reconstructs purposefully modified spam text using pattern recognition and natural language processing (NLP). This makes it possible to identify spam regardless of whether the perpetrators use encoded content, misspellings, or symbol substitution.

5. RESULTS

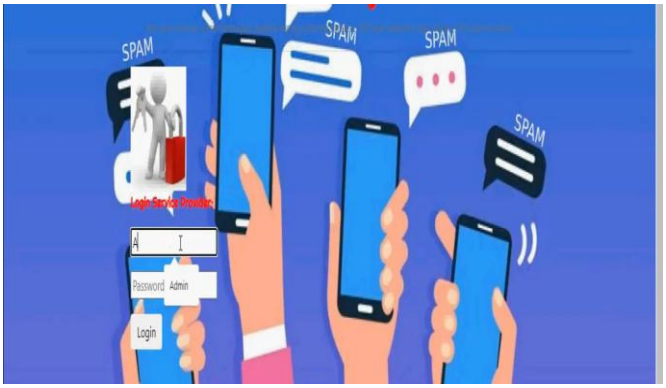

Fig1: Login Page

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

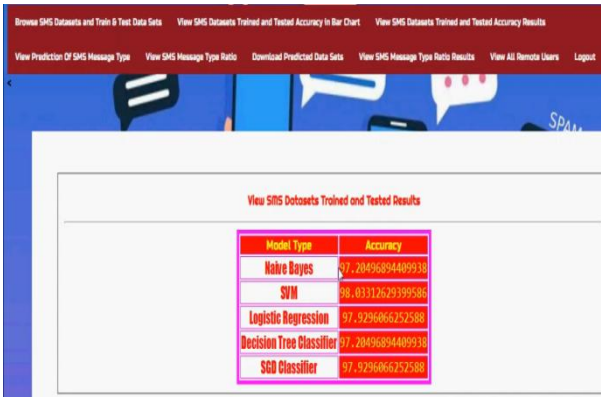

Fig2: Trained and Tested Results

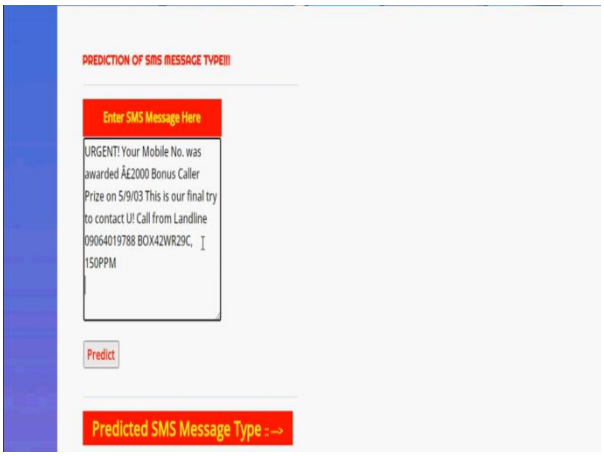

Fig3: Prediction Page

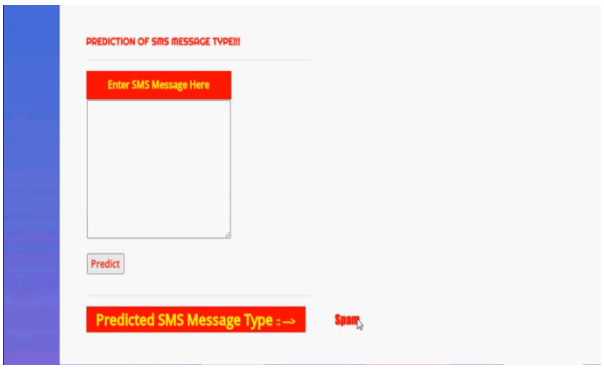

Fig4: Predicted Result

6. CONCLUSION

One automated and efficient way to combat unwanted texts is to construct a machine learning-based SMS spam detection system. By using classification algorithms like Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM), the system successfully separates spam from valid messages. After training on labelled SMS datasets, the model is evaluated using important performance metrics like precision, recall, and F1 score. Real-time spam detection, flexibility in response to new spam trends, and a reduction in false positives and negatives are just a few of the many benefits of the recommended approach. In contrast to conventional keyword-based filtering, machine learning is constantly enhanced and refined by incorporating new data. This technique might significantly reduce financial theft, phishing attempts, and user disturbances in email services and mobile messaging systems.

REFERENCES

1. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the Research of SMS Spam Filtering: New Collection and Results. Proceedings of the 11th ACM Symposium on Document Engineering, 259–262.

2. Gupta, M., Bakliwal, A., Agarwal, S., & Mehndiratta, P. (2018). A Comparative Research of Spam SMS Detection Using Machine Learning Classifiers. 11th International Conference on Contemporary Computing (IC3), 1–7.

3. Roy, D., Bhattacharyya, D., & Kim, T. (2022). Comparative Analysis of Machine Learning Models for SMS Spam Detection. Journal of Machine Learning and Cybersecurity, 10, 120–134.

4. Jain, S., Mehta, P., & Das, K. (2023). PU Learning for SMS Spam Filtering: A Research on Limited Labeled Data. Proceedings of the 15th International Conference on Machine Learning Applications, 34–46.

5. Wittel, G. L., & Wu, S. F. (2021). On Attacking Statistical Spam Filters: A Comparative Research of Evasion

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-2, April–June 2026
ISSN: 3107-7897 | P.ID: JDEBM/25
DOI: https://doi.org/10.5281/zenodo.20604406

Techniques. CEAS Conference on Email and Anti-Spam, 17, 132–145.

6. Tang, W., Zhao, L., & Singh, R. (2023). Deep Learning-Based Hybrid Approach for SMS Spam Detection: A Case Research with Real-World Data. International Journal of Artificial Intelligence in Telecommunications, 8, 55–72.

7. Rojas-Galeano, S. (2021). Using BERT Encoding to Tackle the Mad-Lib Attack in SMS Spam Detection. arXiv preprint arXiv:2107.06400.

8. Salman, M., Ikram, M., & Kaafar, M. A. (2024). Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models. IEEE Access, 12, 24306–24321. Machine Learning & NLP References

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, L. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.