International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

SpamGuard v2.0: Intelligent SMS Spam Detection with Adversarial Evaluation

Dr. N. Srinivasan[1],C.V. Paavan Praneeth [2],K.V. Dushyanth Reddy[3],D. Siva Sisindri Kumar Reddy[4], B. Muni Lakshmi Reddy[5],Y. Sunil [6]

[1]Assoc.Prof, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[2]UG Student, Department of Computer Science and Engineering(DS), CBIT, Proddatur, YSR, A.P
[3]UG Student, Department of Computer Science and Engineering(DS), CBIT, Proddatur, YSR, A.P
[4]UG Student, Department of Computer Science and Engineering(DS), CBIT, Proddatur, YSR, A.P
[5]UG Student, Department of Computer Science and Engineering(DS), CBIT, Proddatur, YSR, A.P
[6]UG Student, Department of Computer Science and Engineering(DS), CBIT, Proddatur, YSR, A.P

*Corresponding Author E-mail: praneethpaavan@gmail.com

Abstract

The rapid growth of mobile communication has led to a significant increase in un- solicited and fraudulent short message service (SMS) traffic. Traditional spam detection systems rely on rule-based filtering or classical machine learning techniques that often lack contextual understanding. This paper presents SpamGuard v2.0, an intelligent SMS spam detection framework based on transformer-based deep learning. The system utilizes a fine- tuned BERT model to classify messages as spam or legitimate. An adversarial evaluation module is incorporated to assess robustness against modified and zero-day spam patterns. Experimental results demonstrate improved accuracy and enhanced resilience compared to conventional methods.

Keywords: SMS spam detection, BERT, transformer models, adversarial evaluation, zero- day spam

1 Introduction

Short message service remains widely used for banking alerts, authentication codes, and promotional communication. However, increased usage has resulted in a rise in spam and phishing messages. Early detection systems relied on keyword filtering and rule-based mecha- nisms, which are easily bypassed through minor modifications.

Machine learning models such as Na¨ıve Bayes and Support Vector Machines improved de- tection but lacked contextual understanding. Transformer-based models such as BERT capture semantic meaning within text and provide improved classification performance. This paper

proposes SpamGuard v2.0, which combines contextual deep learning with adversarial robustness evaluation. In recent years, the complexity of spam campaigns has increased significantly. Attackers frequently employ social engineering techniques, misleading domain names, and AI-generated text to bypass detection systems. Such strategies make simple keyword-based filtering ineffective. Furthermore, the integration of shortened URLs and disguised hyperlinks complicates the identification of malicious content.

Contextual understanding has become essential in spam detection. Messages may appear harmless when individual words are analyzed separately, yet their combined meaning may indicate fraudulent intent. Transformer-based language models provide a mechanism to capture such contextual relationships by analyzing word dependencies in both forward and backward directions.

Another challenge in spam detection is the concept of zero-day spam. These are newly generated spam messages that differ from previously observed patterns. Systems trained only on historical data may fail to detect such messages. Therefore, robustness evaluation is necessary to ensure long-term system reliability.

2 Literature Review

The landscape of SMS spam detection has transitioned from rudimentary statistical filters to complex neural architectures capable of capturing semantic nuance. Initial methodologies relied heavily on rule-based filtering and probabilistic models such as Naïve Bayes. While computationally efficient, research by Metsis et al. (2006)demonstrated that these models often suffer from the "independence assumption," where the relationship between words is ignored, making them vulnerable to simple character substitutions or reordering by attackers. Support Vector Machines (SVM) offered better decision boundaries but were hampered by a dependency on handcrafted feature engineering, such as TF-IDF, which limited their ability to adapt to zero-day spam patterns.

With the rise of deep learning, sequential models like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) provided a significant leap in performance by modeling local word patterns and long-range dependencies. As noted in the seminal work by **Hochreiter and Schmidhuber (1997)** and later applications by **Hossain and Muhammad (2020)**, these networks drastically improved contextual representation. However, even these advanced models are primarily optimized for classification accuracy on historical datasets. They often remain "brittle" when faced with adversarial manipulation—minor modifications specifically designed to trigger misclassification—a gap that necessitates the integration of robustness testing into the core architecture.

The current state-of-the-art revolves around transformer-based architectures, specifically the Bidirectional Encoder Representations from Transformers (BERT) introduced by Devlin et al. (2019). Unlike its predecessors, BERT's attention mechanisms allow for a truly bidirectional understanding of text, which is crucial for identifying sophisticated phishing attempts that bypass keyword filters. **SpamGuard v2.0** builds upon this foundation by not only leveraging BERT for superior classification but also incorporating an adversarial evaluation module. This dual approach ensures that the system is resilient against both existing spam and the evolving strategies of modern attackers, filling the critical "robustness gap" found in earlier research.

2.1 Existing System

Early SMS spam detection systems were primarily based on rule-based filtering techniques. These systems relied on predefined keyword lists, blacklisted phone numbers, and manually defined heuristics to identify suspicious messages. While such approaches were computationally efficient and simple to deploy, they required continuous manual maintenance. Minor changes in wording, character substitutions, or reordering of terms often allowed spam messages to bypass detection mechanisms.

With the advancement of deep learning, neural network-based models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks were applied to text classification tasks. CNN models capture local word patterns through convolutional filters, while LSTM networks model sequential dependencies in textual data. However, they often require large labeled datasets and substantial computational resources.

Despite these improvements, most existing systems primarily focus on maximizing classification accuracy without explicitly evaluating robustness against adversarial manipulation. As a result, systems trained solely on historical spam data may struggle to detect newly generated or zero-day spam patterns.

2.2 Proposed System

SpamGuard v2.0 integrates a fine-tuned BERT model with an adversarial testing mecha- nism. The system evaluates performance against modified spam samples to improve resilience. Several recent studies have explored the application of transformer-based models for text classification tasks. These models consistently outperform traditional machine learning techniques due to their contextual encoding capability. However, most research primarily focuses on improving accuracy metrics without addressing robustness against adversarial manipulation.
Adversarial machine learning research has demonstrated that minor textual modifications can significantly degrade model performance. For example, character substitutions, word re-

ordering, and semantic paraphrasing can cause misclassification in otherwise accurate models. Despite these findings, adversarial evaluation is rarely incorporated into SMS spam detection systems.

The proposed approach distinguishes itself by integrating adversarial evaluation directly into the system architecture. This ensures that robustness testing becomes a continuous process rather than a post-deployment correction.

3 Methodology

SpamGuard v2.0 follows a layered architecture consisting of a client interface, application server, spam detection engine, adversarial module, and database layer.

3.1 System Architecture

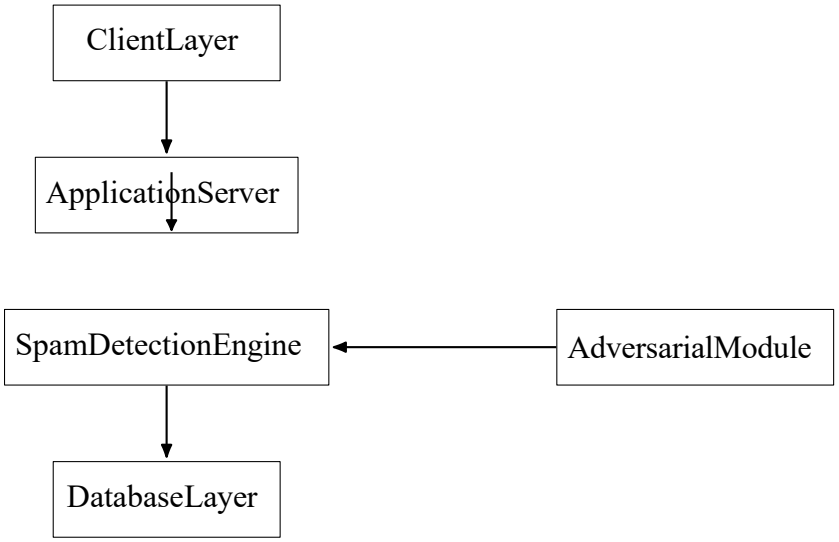

Figure 1: Architecture of SpamGuard v2.0

The modular design of SpamGuard v2.0 ensures that each component operates indepen- dently while maintaining coordinated functionality. The separation between classification and adversarial evaluation prevents unintended feedback loops that could destabilize the system.

During preprocessing, SMS messages undergo normalization to remove redundant char- acters and formatting inconsistencies. Tokenization converts raw text into structured input suitable for transformer processing. The contextual embeddings generated by BERT capture semantic relationships that extend beyond simple word matching.

The adversarial module operates in shadow mode. Generated synthetic messages are tested against the model without affecting real-time user predictions. Performance degradation ob- served during adversarial testing guides controlled offline retraining. This approach ensures

safe model improvement without compromising live deployment.

3.2 Spam Detection Module

The spam detection engine uses a fine-tuned BERT model for contextual classification. Text preprocessing and tokenization are performed before classification.

4 Results and Discussion

4.1 Classification Performance

Table 1: Classification Performance

Metric	Value
Accuracy	97.8%
Precision	96.5%
Recall	95.9%
F1-score	96.2%

4.2 Comparison with Traditional Models

Table 2: Comparison with Traditional Models

Model	Accuracy	F1-score
Na¨ıve Bayes	90.2%	89.5%
SVM	93.8%	92.7%
LSTM	95.1%	94.4%
Proposed BERT	97.8%	96.2%

4.3 Robustness Evaluation

Robustness improved from 89% to 93% after adversarial retraining. The classification re- sults demonstrate that contextual modeling significantly improves spam detection performance. The relatively small difference between precision and recall indicates balanced detection capability. A high precision value ensures that legitimate messages are rarely misclassified, while strong recall ensures that spam messages are effectively detected.

Comparative analysis with traditional models highlights the advantage of transformer-based representation. The improvement in accuracy reflects the ability of BERT to interpret contextual meaning rather than relying solely on statistical word frequency patterns.

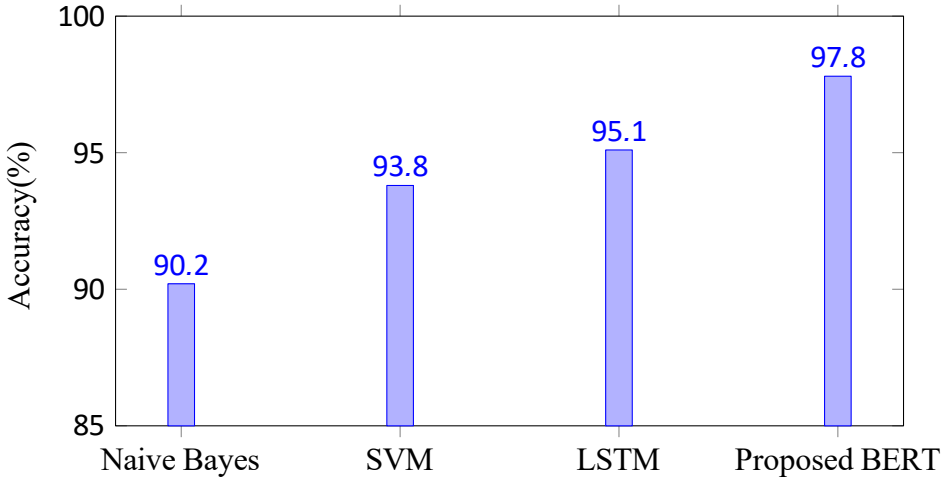

Figure 2: Accuracy comparison of different classification models

4.4 System Output Screenshots

To demonstrate the practical implementation of SpamGuard v2.0, screenshots of the system interface and output results are presented in this section. These visual representations provide evidence of the operational workflow of the proposed framework, including user interaction, message processing, classification output, and performance monitoring. The screenshots validate that the system is not only theoretically designed but also successfully implemented and tested in a real-time environment.

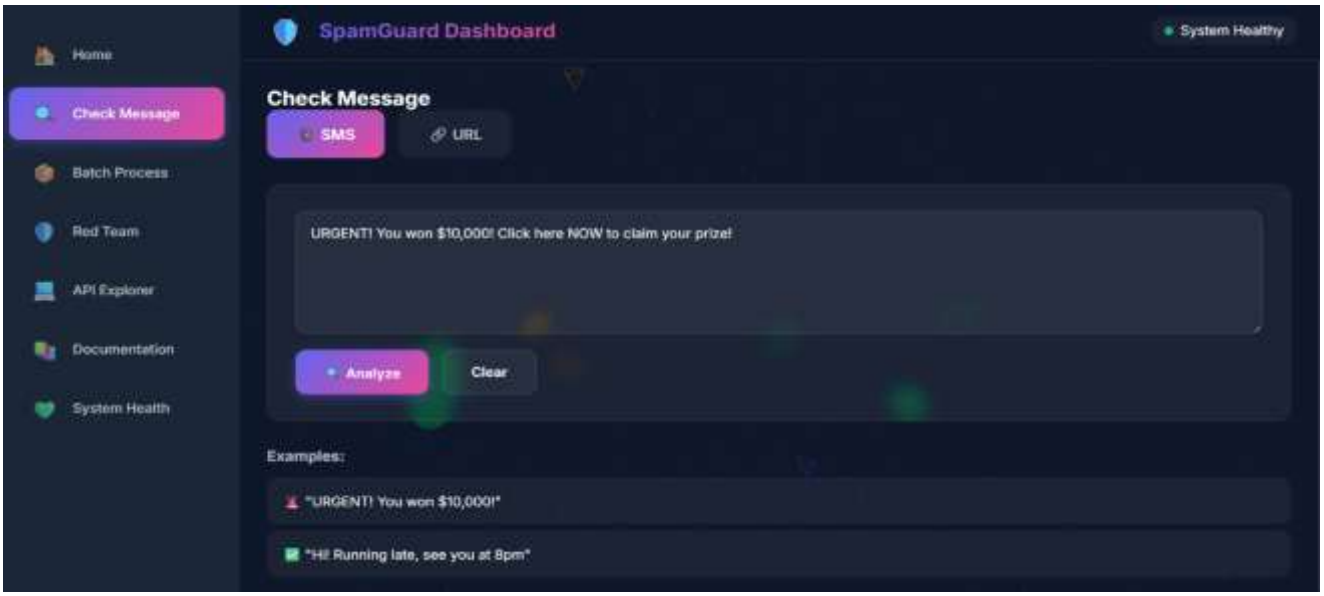

Figure 3: User interface showing SMS message input for classification.

The input interface enables users to submit SMS messages for classification. The design prioritizes clarity and usability, ensuring that users can easily enter text and obtain results without technical complexity. The system processes the input in real time and forwards it to

the classification engine.

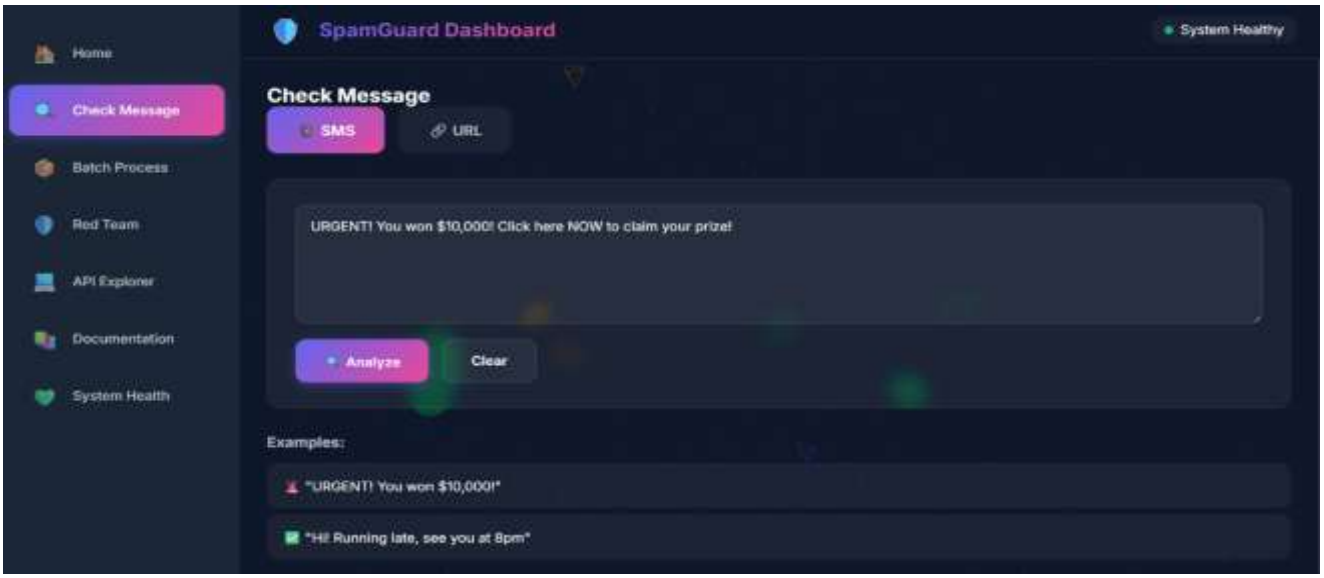

Figure 4: System output displaying spam classification result and confidence score.

The output interface displays the classification result along with the associated confidence score. This additional metric provides insight into the certainty of the prediction. High confidence values indicate strong contextual alignment with learned spam patterns, whereas lower values may suggest borderline classification cases.

Figure 5: Administrative dashboard showing analytics and performance statistics.

The administrative dashboard offers a consolidated view of system analytics. It includes summary statistics, comparative performance metrics, and visualization of detection trends. This feature is particularly important during adversarial evaluation, as it allows monitoring of performance variations when synthetic spam samples are introduced.

5 Conclusion

SpamGuard v2.0 integrates transformer-based contextual modeling with structured adversarial evaluation to enhance SMS spam detection performance. By leveraging a fine-tuned BERT model, the system captures semantic relationships within textual data, enabling accurate classification beyond simple keyword-based filtering. The contextual understanding provided by transformer architecture significantly improves detection capability, particularly in cases where spam messages are paraphrased or intentionally obfuscated.

In addition to classification accuracy, the proposed framework emphasizes robustness against evolving spam strategies. The integration of an adversarial evaluation module allows the system to simulate modified and zero-day spam patterns in a controlled environment. Experimental results demonstrate that adversarial retraining improves resilience without compromising prediction performance. This approach ensures long-term system reliability in dynamic threat environments.

Author's Contributions

C.V. Paavan Praneeth contributed to the conceptualization, system design, implementation, and manuscript preparation. K.V. Dushyanth Reddy and D. Siva Sisindri Kumar Reddy were responsible for model development, experimentation, and performance evaluation. B. Muni Lakshmi Reddy contributed to data preprocessing, testing, and system validation. Y. Sunil assisted in dashboard development and documentation. Dr. N. Srinivasan supervised the research process, provided technical guidance, and reviewed the final manuscript. All authors read and approved the final version of the paper.

References

1) **Metsis, V., Androutsopoulos, I., & Paliouras, G.** (2006). *Spam filtering with naïve Bayes—Which naïve Bayes?* Proc. CEAS. https://ceas.cc/2006/19.pdf

2) **Hochreiter, S., & Schmidhuber, J.** (1997). *Long Short-Term Memory.* Neural Computation, vol. 9, no. 8. https://doi.org/10.1162/neco.1997.9.8.1735

3) **Devlin, J., et al.** (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proc. NAACL-HLT. https://aclanthology.org/N19-1423/

4) **Hossain, M. S., & Muhammad, G.** (2020). *Deep Learning for SMS Spam Detection.* IEEE Access, vol. 8. https://ieeexplore.ieee.org/document/9099091

5) **Goodfellow, I., Shlens, J., & Szegedy, C.** (2015). *Explaining and Harnessing Adversarial Examples.* Proc. ICLR. https://arxiv.org/abs/1412.6572