Securing SMS Spam Classifier Web Applications Using Zero Trust Architecture for Enhanced Data Privacy

Muhammad Furqan Ishaq, Mohammad Zaid
Roll No.: 2022 R/2021-CS-199, 2022 R/2021-CS-214
Section: C
Email: sheikhzaid768@gmail.com
Department of Computer Science, University of Engineering and Technology
Lahore, Pakistan

***Abstract*—In today's digital era, spam messages pose a significant threat to both individual privacy and organizational security. This paper presents a secure web-based SMS spam classification system fortified with Zero Trust Architecture (ZTA) principles. The system combines a machine learning classifier with a robust authentication and monitoring framework, ensuring secure user access, encrypted communication, and real-time threat detection. The integration of Zero Trust Security mechanisms—including multi-factor authentication, identity-aware policies, and behavioral analysis—ensures that only verified and contextually safe actions are permitted. This approach not only improves spam detection accuracy but also safeguards sensitive data from misuse and breaches.**

***Index Terms*—Zero Trust Security, Spam Classification, Machine Learning, Access Control, Multi-Factor Authentication, Data Privacy, Web Application Security**

I. Introduction

SMS remains one of the most widely used communication methods, especially in regions with limited internet penetration. However, the simplicity and ubiquity of SMS make it a common vector for spam, scams, and phishing attacks. As a response to this growing concern, SMS spam classifiers—typically powered by machine learning algorithms—are increasingly adopted to filter and block unwanted content.

While these systems can effectively detect and flag spam messages, many fail to address deeper security concerns. In traditional setups, access to these tools and their datasets is often inadequately protected, exposing sensitive communication to risks such as data breaches, manipulation, or unauthorized analysis. To address this gap, we propose an enhanced SMS spam classification web application that adopts the Zero Trust Security (ZTS) framework.

The Zero Trust model operates on the principle of "never trust, always verify," assuming that every request—whether internal or external—is potentially malicious. Instead of relying on perimeter defenses, ZTS enforces granular, context-based access policies, verifies user identity at every interaction, and continuously monitors activity. When applied to a spam classifier, this approach ensures not only accurate message filtering but also secure and privacy-aware handling of user data.

This project combines a logistic regression-based spam classifier with secure frontend-backend communication, strict role-based access, encrypted message logging, and real-time anomaly detection. The result is a privacy-focused solution ideal for deployment in environments like healthcare, banking, and enterprise messaging systems, where secure message classification is mission-critical.

II. Related Work

Spam filtering has been a significant area of research in natural language processing (NLP) and cybersecurity. Numerous machine learning models such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression have shown success in classifying SMS messages into spam or ham categories. Datasets like the UCI SMS Spam Collection have served as benchmarks for evaluating classifier performance [1].

Deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) have also been explored to enhance classification performance. However, these often require substantial computational resources and large datasets, making them less feasible for lightweight, real-time applications like mobile web apps.

On the security side, Zero Trust Architecture has gained traction for its ability to prevent data leaks, lateral movement, and credential abuse. ZTS emphasizes identity verification, micro-segmentation, and behavior-based policy enforcement. However, few works have integrated ZTS into smaller-scale, domain-specific tools like spam classifiers. Most ZTS applications are found in enterprise networks and cloud infrastructure [2].

Recent advancements attempt to bridge this gap. For example, Sharma et al. [3] proposed ZTS-based models for healthcare APIs, but did not extend it to content classification systems. Other works focus on Zero Trust applied to cloud-native applications or hybrid infrastructures but ignore personal communication platforms. Our work aims to fill this gap

by implementing a Zero Trust-enabled SMS spam classifier, balancing model accuracy with secure system operation.

III. Methodology

Our system design follows a modular and security-first approach. The architecture integrates spam classification with secure authentication, encrypted communication, access control, and monitoring. The major components include data preprocessing, machine learning model training, frontend interaction, and backend API enforcement under Zero Trust principles.

A. Dataset and Preprocessing

The dataset used is the SMS Spam Collection from UCI, which includes 5,574 messages labeled as "ham" or "spam." Messages undergo preprocessing steps such as lowercasing, removal of punctuation, stopword elimination, tokenization, stemming using the Porter Stemmer, and vectorization using Term Frequency-Inverse Document Frequency (TF-IDF). These transformations help convert raw text into a structured format suitable for machine learning.

Data was split into training and test sets in a 70:30 ratio. Stratified sampling was used to maintain the proportion of spam/ham messages. Exploratory Data Analysis (EDA) was performed to visualize message lengths, keyword frequency, and class imbalance.

B. Model Training

A logistic regression model was selected due to its simplicity and high accuracy in binary classification tasks. The model was trained using scikit-learn's pipeline interface for consistent preprocessing and classification. Hyperparameters such as regularization strength and solver method were tuned using grid search.

The trained model achieved over 95

C. Web Interface

The frontend is developed using Streamlit, providing users with an intuitive way to enter SMS messages and view classification results. Streamlit also supports interactive widgets, authentication prompts, and user dashboards. Role-based access control is enforced—administrators can view logs, monitor access, and manage users, while standard users can only classify messages.

D. Zero Trust Components

1. Identity and Access Management: User identity is verified through secure login using bcrypt-hashed passwords and TOTP-based multi-factor authentication (e.g., Google Authenticator).

2. Least Privilege Access: Role-based access ensures users only access features necessary for their role. Admin panels, logs, and model retraining interfaces are hidden from standard users. Privileges are reviewed periodically.

3. Secure Communication: All traffic between the frontend and backend is encrypted using HTTPS and protected by JSON Web Tokens (JWTs). Tokens are short-lived and refreshed periodically to minimize the attack window [4].

4. Behavioral Monitoring: Each session is monitored for abnormal behavior such as repeated requests, excessive login attempts, or geolocation anomalies. These are logged using the ELK stack (Elasticsearch, Logstash, Kibana) and, if necessary, trigger automatic logout or re-authentication.

5. API Hardening: The backend APIs are secured with token-based validation, rate limiting, and input sanitization to protect against injection, spoofing, and brute-force attacks. API gateways enforce usage limits and deny malformed requests.

E. Testing and Evaluation

Security testing was conducted using simulated attacks including SQL injection, brute-force login, cross-site scripting (XSS), and JWT token tampering. Unit tests were written using pytest, and integration testing was performed via Postman.

The spam classification model was tested on unseen data, maintaining a high precision and recall. Confusion matrix and ROC-AUC analysis confirmed classifier robustness. The system's performance under load was tested using Apache JMeter.

IV. Discussion

The integration of Zero Trust with an SMS spam classifier provides layered security that goes beyond classification accuracy. While traditional spam filters excel at identifying malicious messages, they often ignore the wider ecosystem of security threats. Our approach strengthens this by enforcing identity verification, encrypting communication channels, and observing user behavior.

Furthermore, the system offers flexibility and scalability. It can be deployed as a cloud-hosted service or integrated into existing enterprise messaging systems. Since the architecture is modular, components like authentication, monitoring, or the classifier itself can be replaced or upgraded without major refactoring.

V. Conclusion and Future Work

This project demonstrates how combining Zero Trust Security principles with machine learning can produce a resilient and privacy-conscious SMS spam classification system. The integration of real-time verification, access control, and encrypted communication ensures that data is both accurate and securely processed.

Unlike traditional classifiers, which solely focus on message accuracy, our system prioritizes the entire security lifecycle of user interaction—authentication, classification, storage, and logging. The approach proves particularly relevant for deployment in sectors handling sensitive communication.

Future work includes:

- Integration with advanced NLP models like BERT for deeper semantic understanding.
- Real-time threat intelligence sharing with centralized security dashboards.

- Mobile application development for Android/iOS with fingerprint or biometric authentication.
- Expansion to multilingual spam detection and phishing link analysis.

REFERENCES

[1] Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A., "Contributions to the study of SMS spam filtering: new collection and results," *Proceedings of the 11th ACM Symposium on Document Engineering*, pp. 259–262, 2011.

[2] Rose, T., "Zero Trust Security: A Comprehensive Approach for Securing Modern Networks," *Journal of Cybersecurity and Privacy*, vol. 12, no. 3, pp. 50–65, 2020.

[3] Sharma, P., Soni, R., and Chauhan, D., "Implementing Zero Trust in Health IT Systems," *International Journal of Healthcare Information Systems*, vol. 28, no. 1, pp. 1–8, 2022.

[4] Basak, A., Bhattacharya, A., and Sadhukhan, D., "JWT Vulnerabilities: A Study of Attacks and Mitigation Techniques," *Proceedings of the 2021 International Conference on Computer and Communications Security*, pp. 1–6, 2021.