Sri Lanka Institute of Information Technology

Email Spam Classification Using Machine Learning

[Audit report]

Individual Assignment

Secure Software Systems - IE3042

IT20618872- Thisitha K.L.D

Date of submission

25-05-2023

1. INTRODUCTION

In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world. Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam. No one wants to receive emails not related to their interest because they waste receivers' time and resources. Besides, these emails can have malicious content hidden in the form of attachments or URLs that may lead to the host system's security breaches. Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. So, it requires an immense demand for the security of the email system. Spam emails may carry viruses, rats, and Trojans. Attackers mostly use this technique for luring users towards online services. They may send spam emails that contain attachments with the multiple-file extension, packed URLs that lead the user to malicious and spamming websites and end up with some sort of data or financial fraud and identify theft. Many email providers allow their users to make keywords base rules that automatically filter emails. Still, this approach is not very useful because it is difficult, and users do not want to customize their emails, due to which spammers attack their email accounts.

The purpose of this audit report is to evaluate the security vulnerabilities in the email spam classification system developed using machine learning techniques. The audit follows the guidelines outlined in the Open Web Application Security Project (OWASP) framework.

2. Methodology

When conducting an audit using the OWASP framework for Email Spam Classification using Machine Learning, the following steps and considerations can be taken into account:

1. OWASP Application Security Verification Standard (ASVS): Refer to the ASVS guidelines provided by OWASP, which outline the security requirements and best practices specific to the application being audited. Adapt these guidelines to the context of email spam classification.
2. Threat Modeling: Perform a comprehensive analysis of the email spam classification system's architecture, components, and potential threats. Consider threats such as malicious input data, evasion techniques, and attacks targeting the machine learning model itself.
3. Vulnerability Assessment: Evaluate the system for potential vulnerabilities related to email spam classification. This may include:

a. Input Validation: Verify that the system effectively validates and sanitizes input data, ensuring that no malicious content is processed by the machine learning model.

b. Model Training Security: Assess the security of the model training process, ensuring that the training data is reliable, free from biases, and properly protected to prevent unauthorized modifications.

c. Authentication and Access Controls: Review the authentication mechanisms and access controls in place to protect the email spam classification system from unauthorized access or manipulation.

d. Error Handling: Evaluate the system's error handling mechanisms to ensure that errors and exceptions are handled securely without exposing sensitive information or creating vulnerabilities.

e. Data Protection: Assess the security measures in place to protect sensitive data used for email spam classification, such as encryption and secure storage practices.

4. Penetration Testing: Conduct penetration tests to identify potential vulnerabilities and weaknesses in the system. Test for common attack vectors such as injection attacks, evasion techniques, and unauthorized access attempts.
5. Security Controls Evaluation: Evaluate the effectiveness of the implemented security controls in mitigating the identified vulnerabilities and ensuring the security of the email spam classification system. This includes evaluating the strength of authentication mechanisms, access controls, and encryption techniques used to protect sensitive data.
6. Risk Analysis: Analyze the identified vulnerabilities in terms of their potential impact on the email spam classification system. Assess the likelihood of exploitation and the potential consequences to prioritize remediation efforts.

3. Findings

In Email Spam Classification using Machine Learning, there are several security vulnerabilities that can arise. Here are some common vulnerabilities to be aware of:

1. Lack of Input Validation:
   - Failure to properly validate and sanitize incoming email content and metadata can lead to various attacks such as SQL injection, cross-site scripting (XSS), and command injection.
2. Biased or Malicious Training Data:
   - The use of biased or malicious training data can result in incorrect classification of spam emails or allow the evasion of spam detection. Adversaries can manipulate the

training data to exploit vulnerabilities in the classification model.

3. Inadequate Authentication and Access Controls:
   - Insufficient authentication mechanisms or weak access controls can lead to unauthorized access to the email spam classification system or compromise of sensitive data. Attackers may attempt to bypass authentication or exploit weak access controls to gain unauthorized privileges.
4. Inadequate Error Handling:
   - Improper error handling can inadvertently expose sensitive information, providing potential attackers with insights into the system's internal workings. Detailed error messages may disclose system vulnerabilities or expose sensitive data.
5. Data Leakage and Privacy Concerns:
   - If proper data protection measures are not implemented, there is a risk of data leakage or unauthorized access to classified emails or user information, compromising user privacy and potentially violating data protection regulations.
6. Adversarial Attacks:
   - Machine learning models used for email spam classification are susceptible to adversarial attacks, where malicious actors intentionally craft email content to bypass the classification system's detection algorithms. This can lead to the system misclassifying spam or failing to detect malicious emails.
7. Insecure Model Deployment:
   - Insecure deployment of the machine learning model can expose sensitive information or enable unauthorized access. Inadequate encryption, weak access controls, or improper configuration of the deployment infrastructure can create vulnerabilities.
8. Lack of Regular Updates and Monitoring:
   - Failing to keep the email spam classification system and its components up to date with security patches and updates can leave the system vulnerable to known exploits. Lack of monitoring and incident response capabilities can result in delayed detection and response to security incidents.

It's important to note that the specific vulnerabilities may vary depending on the implementation and the technologies used in the email spam classification system. Regular security assessments, code reviews, and penetration testing are recommended to identify and address any vulnerabilities in the system.

4. Recommendations

To address the vulnerabilities identified in Email Spam Classification using Machine Learning, the following recommendations can be considered:

1. Implement Robust Input Validation:
   - Apply strict input validation mechanisms to prevent common attacks such as SQL injection, cross-site scripting (XSS), and command injection.
   - Validate and sanitize incoming email content and metadata to prevent the processing of malicious or malformed data.
2. Enhance Model Training:
   - Improve the quality and diversity of training data to reduce bias and increase the model's ability to detect various types of spam emails.
   - Regularly update and retrain the model to adapt to evolving spamming techniques and patterns.
   - Implement techniques to detect and mitigate adversarial attacks, such as poisoning or evasion attempts.
3. Strengthen Authentication and Access Controls:
   - Enforce strong authentication mechanisms, such as multi-factor authentication (MFA), to ensure only authorized users can access the system.
   - Implement role-based access controls (RBAC) to limit access to sensitive data and system functionalities based on user roles and privileges.
   - Regularly review and audit user access permissions to ensure they are appropriate and aligned with the principle of least privilege.
4. Improve Error Handling:
   - Implement proper error handling mechanisms to avoid exposing sensitive information in error messages.
   - Ensure error messages provide limited information to prevent attackers from exploiting system vulnerabilities or gaining insights into the system's internal workings.
   - Monitor and log errors to enable effective incident response and analysis.
5. Implement Data Protection Measures:
   - Apply encryption techniques to protect sensitive data used for email spam classification both at rest and in transit.

    - Implement secure storage practices to safeguard sensitive data, ensuring it is not accessible to unauthorized parties.
    - Regularly review and update data protection measures to align with industry best practices and evolving security standards.
6. Conduct Regular Security Testing:
    - Perform periodic penetration testing and vulnerability scanning to proactively identify and address emerging security risks.
    - Consider engaging third-party security experts for independent security assessments and audits.
    - Stay updated with the latest security vulnerabilities and threats related to machine learning systems and apply patches or updates accordingly.
7. Provide Security Awareness Training:
    - Educate system users and administrators about common security risks, such as phishing attacks or social engineering techniques.
    - Foster a culture of security awareness to encourage responsible handling of sensitive data and adherence to security protocols.

It is important to note that these recommendations are general in nature and should be adapted to the specific context and requirements of your email spam classification system. Additionally, implementing a robust and comprehensive security program that includes ongoing monitoring, incident response capabilities, and regular updates is crucial for maintaining the security and resilience of the system.

5.Conclusion

This audit report provides an overview of the security vulnerabilities identified in the email spam classification system. By addressing the recommendations outlined above, the system can enhance its security posture and mitigate potential risks associated with email spam classification.

It is recommended to implement these recommendations in a timely manner to ensure the confidentiality, integrity, and availability of the system and its associated data.