Journal of Science, Computing and Engineering Research (JSCER)
Volume-7, Issue-11, November 2024.

Enhancing Privacy in Spam Filtering: The PCSF System Employing Support Vector Machine

[1]K.S. Suganyadevi, [2]M. Velmathi, [3]S. Vedapriya, [4]S. Venkata Lakshmi

[1-3] UG Students, Computer Science Engineering, K.L.N. College of Engineering, Sivagangai, Pottapalayam.
[4]Assistant Professor, Computer Science Engineering, K.L.N. College of Engineering, Sivagangai, Pottapalayam.

Article Information

Received : 02 Nov 2024
Revised : 06 Nov 2024
Accepted : 10 Nov 2024
Published : 14 Nov 2024

Corresponding Author:

Author Name: S. Venkata Lakshmi

Email: venkatalakshmi.green@gmail.com

***Abstract*—**The goal of privacy-preserving spam filtering is to inspect emails while ensuring that both the detection rules and email content remain confidential. Existing solutions face several challenges, including: (1) inadequate privacy measures that may expose email content or detection rules to third parties; (2) vulnerability to exhaustive word search attacks due to improper encryption techniques, which can compromise email confidentiality; (3) potential privacy risks when outsourcing spam filtering to third parties without robust protection measures; (4) delayed spam detection, where spam is only identified after it reaches the receiver,potentially exposing the user to harmful content such as phishing attacks embedded within emails; and (5) computational inefficiencydue to the high cost of privacy-preserving mechanisms. Traditionalspam filtering techniques often compromise user privacy by analyzing raw email content. This paper presents the **Privacy-Preserving Content-Based Spam Filtering (PCSF)** system, whichutilizes **Support Vector Machines (SVMs)** and **homomorphic encryption** to detect spam while preserving the confidentiality of email content. The proposed system maintains high accuracy in spam detection without exposing sensitive data, making it a viable solution for modern privacy concerns. Experimental results demonstrate that the PCSF system achieves an accuracy of **95.2%**, with precision and recall values of **94.5%** and **96.0%**, respectively.

Keywords: Spam Filtering; Support Vector Machine; Privacy Preservation; Homomorphic Encryption; Encryption; Decryption; Machine Learning.

Citation: K.S. Suganyadevi, M. Velmathi, S. Vedapriya, S. Venkata Lakshmi. "Enhancing Privacy in Spam Filtering: The PCFS System Employing Support VectorMachine", Journal of Science, Computing and Engineering Research, 7(11), November 2024.

I. INTRODUCTION

1.1 Background and Motivation

Mail is one of the most common and widely adopted forms of communication over the Internet. It is used extensively by government agencies, corporations, and individuals to exchange official and unofficial messages securely and economically. This medium is not only reliable but also integrates seamlessly with various applications within today's work environments. However, the prevalence of email has made it a prime target for cybercriminals seeking to exploit its vulnerabilities. Most email services implement hop-to-hop encryption using **Transport Layer Security (TLS)** to protect messages during their transit between mail servers. While this approach secures the email content during transmission, it does not guarantee protection against exposure at the mail servers themselves. This vulnerability implies that an attacker who successfully infiltrates an email service can easily read or modify the emails, unless further protective measures are instituted. Indeed, numerous significant data breaches have highlighted these risks, allowing unauthorized access to user email accounts [2][3]. Moreover, terms of service from many email providers typically include clauses that allow access to users' emails, which can lead to unauthorized disclosures of sensitive information.

Given these threats, it is essential for users to safeguard their emails from potential data breaches through robust encryption methods. End-to-end encryption can effectively protect email content by ensuring that only the sender and therecipient have access to the information. Some email serviceproviders offer this feature, allowing users to utilize encryption protocols such as **S/MIME**. However, these conventional spam filtering techniques

often rely on the ability to analyze email content in plaintext format, thereby limiting their effectiveness in conjunction with end-to-end encryption. This inadequacy necessitates the development of advanced spam filtering systems that can operate on encrypted email while maintaining user privacy. A privacy-preserving spam filtering solution must ensure that spam detection occurs without exposing the content of the emails or the underlying detection rules.

In this context, we propose the **Privacy-Preserving Content-Based Spam Filter (PCSF)** system, which integrates **Support Vector Machines (SVMs)** to accurately classify spam without compromising user privacy. Our system effectively addresses the limitations of existing solutions by ensuring that spam is filtered before it reaches the recipient, thus preserving sensitive information and enhancing detection accuracy.

The digital age has seen an exponential increase in email communication, leading to the emergence of spam as a critical issue for users and organizations alike. According to recent reports, spam emails account for approximately **45% of global email traffic** [1][2]**.**The implications of spam extend beyond mere annoyance; they pose significant security risks, facilitating phishing attempts, malware distribution, and other cyber threats. Traditional spam filtering mechanisms often fail to keep pace with evolving spam tactics, necessitating the development of more sophisticated and adaptive approaches.

Privacy has become a paramount concern in the realm of email communication. Users are increasingly aware of how their personal data is handled, especially in light of stringent regulations such as the **General Data Protection Regulation (GDPR)** and **Health Insurance Portability and Accountability Act (HIPAA**) [3]**.** These regulations mandate that organizations implement adequate measures to protect user data, creating a pressing need for spam filtering systems that prioritize user privacy.

To address these challenges, we propose a **Privacy-Preserving Content-Based Spam Filtering (PCSF)** system. The PCSF leverages **Support Vector Machines (SVMs)** and **homomorphic encryption** to enhance spam detection accuracy while preserving user privacy. By operating on encrypted data, the PCSF allows for the effective classification of emails without exposing sensitive user information.

1.2 Problem Statement

The principal challenge tackled in this research is the development of an efficient spam filtering system that can classify emails as spam or non-spam while ensuring that the user's email content remains encrypted and private. This necessitates the creation of a **privacy-preserving machine learning (PPML)** [19] model capable of executing accurate classification without direct access to the plaintext data.

Spam Filtering Techniques

Spam filtering has undergone significant advancements since its inception. Traditional methods can be broadly categorized into heuristic approaches, statistical methods, and machine learning-based techniques.

1.2.1 Heuristic Methods

Heuristic or rule-based spam filters utilize predefined rules to identify spam. [5] These methods often involve:

- **Blacklists** of known spam senders.
- **Keyword matching** were specific words or phrases trigger spam classification.

While heuristic methods are simple to implement, they lack adaptability and are often bypassed by evolving spam tactics.

1.2.2 Statistical and probabilistic Approach

Statistical methods such as **Naive Bayes** [5] have gained popularity due to their efficiency and ease of implementation. The Naive Bayes classifier operates on the assumption that the presence of a particular feature (word) in an email is independent of the presence of any other feature. This leads to the following classification formula:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Where:

- P(Y|X) is the probability of an email being spam given the features X,
- P(X|Y) is the likelihood of observing the features in a spam email,
- P(Y) is the prior probability of spam.

Despite its efficiency, Naive Bayes can produce high false-positive rates, especially in cases where spam patterns are diverse.

1.2.3 Machine Learning Approaches

The introduction of machine learning algorithms has significantly enhanced spam detection accuracy. Key methods include :

- **k-Nearest Neighbors (k-NN)**: This algorithm classifies emails based on their proximity to labeled examples in feature space. [6] The main advantage is its simplicity; however, it is computationally expensive with large datasets.
- **Decision Trees and Random Forests**: Decision Trees split the feature space into regions based on feature values. Random Forests, as an ensemble method, build multiple decision trees to improve accuracy and reduce overfitting.[9]

1.2.4 Deep learning Techniques

Recent advancements in deep learning have led to the use of neural networks for spam filtering. Techniques such as **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)** [7] excel in capturing semantic context in text. However, they typically require large training datasets and substantial computational resources.

1.3 Support Vector Machine for Spam Detection

SVMs are particularly effective for high-dimensional datasets, which is characteristic of text data in spam filtering.

The SVM constructs a hyperplane that maximally separates the two classes (spam and non-spam**).** [20] The optimization problem for a linear SVM can be expressed as:

Where:

- w is the weight vector,
- b is the bias,
- y_i is the class label (1 for spam, -1 for non-spam),
- C is the regularization parameter.

1.3.1 Kernel Function

SVMs can employ kernel functions to project data into higher-dimensional spaces, enabling non-linear separation. Common kernel functions include:

- **Linear Kernel**: $K(xi,x_j) = x_i^T x_j$
- **Polynomial Kernel**: $K(x_i,x_j) = (x_i^T x_j+c)^d$
- **Radial Basis Function (RBF)**: $K(x_i,x_j) = \exp(-\gamma||x_i-x_j||^2)$

1.5 Privacy Preserving Machine Learning

In the context of spam filtering, privacy-preserving techniques must ensure the confidentiality of sensitive data during model training and inference.

1.5.1. Homomorphic Encryption

Homomorphic encryption allows computations to be performed on ciphertexts without needing access to the plaintext. This cryptographic method ensures that sensitive data can remain encrypted during processing. For instance, a Fully Homomorphic Encryption (FHE) [7] scheme allows both addition and multiplication operations on encrypted data, thus facilitating complex machine learning computations without exposing raw data.

1.5.2 Differential Privacy

Differential privacy introduces random noise to data or computations to safeguard individual data points. This method ensures that the removal or addition of a single data point does not significantly affect the outcome, thereby protecting user privacy.[19]

1.5.3 Federated Learning

Federated learning offers a decentralized approach where training occurs on local devices, and only aggregated updates (not raw data) are shared with a central server. This method minimizes the risk of exposing sensitive data but requires coordination between distributed clients and servers.[23]

II. LITERATURE SURVEY

The paper *"Privacy-Preserving Content-Based Spam Filter (PCSF)"* by Intae Kim et al., published in 2023, addresses the critical challenge of filtering spam while preserving the privacy of both users and detection rules. Spam filtering typically involves inspecting email content, which raises significant privacy concerns. Existing systems often expose user data, and they may be susceptible to attacks that compromise the confidentiality of encrypted emails. The authors propose a solution that integrates homomorphic encryption with Support Vector Machine (SVM) classification to balance the competing goals of privacy and detection accuracy.

In this system, emails are encrypted before being processed by the spam filter, ensuring that sensitive information remains confidential even during the detection process. The homomorphic encryption method allows computations to be performed on encrypted data without needing to decrypt it, offering an effective means to safeguard user privacy. This is particularly important when outsourcing spam filtering to third parties, where private information could otherwise be exposed.

The use of SVM as the classification method adds another layer of sophistication. SVM is well-suited for binary classification problems like spam detection, and it performs well in high-dimensional spaces, making it an effective tool for distinguishing spam from legitimate emails based on their features. The authors report that the system achieves high accuracy rates in identifying spam, which indicates that privacy preservation does not come at the cost of detection performance.

However, the PCSF system also comes with challenges. The use of homomorphic encryption, while effective in preservingprivacy, introduces computational complexity. This can slow down the spam detection process, particularly when dealing with large datasets or real-time filtering. Despite this, the system offers a substantial improvement over conventional spam filters that do not provide adequate privacy guarantees. The authors of the PCSF paper highlight the growing need for privacy-preserving mechanisms in the era of increasing data

breaches and surveillance. By encrypting the entire email content and detection rules, the PCSF system offers users peace of mind that their communications remain confidential. Moreover, the study provides a thorough analysis of the system's performance under various conditions, demonstrating that it maintains high classification accuracy even in the presence of encrypted data.

This research contributes significantly to the field by addressing both privacy and accuracy in spam filtering, which are often seen as conflicting objectives. The balance between these two factors is crucial in modern-day applications, where privacy concerns are more prominent than ever. Future research could focus on improving the computational efficiency of the system, making it more scalable and suitable for real-time applications.

In conclusion, the *PCSF* system is a groundbreaking solution that enhances the privacy of spam filtering operations without sacrificing detection accuracy. By leveraging homomorphic encryption and SVM, the system ensures that both email content and detection rules remain confidential, marking a significant step forward in secure spam filtering.

In *"A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model"*, published in 2021 by Hayoung Oh et al., the authors tackle the problem of detecting spam comments on YouTube. As social media platforms grow, so does the amount of spam in comments sections. The YouTube platform, despite having a built-in spam detection system, still struggles to effectively filter out all unwanted comments, leading to the need for improved detection methods.

The paper presents a spam detection system that utilizes cascaded ensemble machine learning models to enhance accuracy. This method involves combining different machine learning algorithms, such as decision trees, logistic regression, Bernoulli Naive Bayes, and support vector machines, to classify comments. Ensemble learning improves performance by leveraging the strengths of multiple algorithms while compensating for their weaknesses. This approach aims to reduce the number of false positives and negatives, providing a more reliable spam detection mechanism.

One of the key contributions of this paper is its use of real-world YouTube comment data to test the proposed system. The system analyses not only the content of the comments but also the context, such as the frequency of posts and the engagement levels of users. This multi-faceted analysis enables the detection of more subtle forms of spam that might evade traditional filters, such as promotional messages disguised as genuine comments.

The authors report significant improvements in spam detection accuracy compared to conventional models. The ensemble method, by combining multiple classifiers, is particularly effective in minimizing errors. This is crucial in platforms like YouTube, where both false positives (flagging genuine comments as spam) and false negatives (allowing spam to pass through) can have significant consequences for user experience and platform integrity.

However, the system is not without its limitations. The computational complexity of running multiple classifiers in an ensemble model can be high, especially when processing large volumes of data in real-time. This could impact the system's scalability and its ability to handle spikes in user activity, such as during live streams or viral video uploads.

Additionally, the system might struggle with evolving spam tactics. As spammers continuously adapt their methods to evade detection, the ensemble model may require frequent updates to its training data to stay effective. Despite these

challenges, the proposed method offers a robust solution to the problem of YouTube spam detection, particularly for high-traffic videos where manual moderation is impractical.

The study highlights the importance of content and context in detecting spam, suggesting that future systems should not only focus on the textual content of comments but also on the behaviour of users. This behaviour-based detection could significantly enhance the system's ability to identify spam more accurately and in real-time.

In summary, the paper provides an innovative approach to spam detection on YouTube, leveraging the power of ensemble machine learning models. While the system offers significant improvements in accuracy, future research could explore ways to reduce its computational demands and improve adaptability to evolving spam strategies.

The paper *"Email Spam Detection Based on Exceptional Precision"*, published by David C. McCallum et al. in 2024, explores various machine learning techniques for detecting spam in email systems. As the volume of email traffic increases globally, so too does the volume of spam, making effective detection mechanisms more critical than ever. This paper focuses on improving the precision of spam detection algorithms, with the goal of reducing false positives (legitimate emails classified as spam) and ensuring that users do not miss important communications. The authors compare multiple machines learning algorithms, including decision trees, random forests, support vector machines, and neural networks. They evaluatethese models based on their precision, recall, and overall accuracy. Precision is emphasized because of the high cost associated with false positives in spam detection—users are likely to become frustrated if they frequently must check their spam

folder for important emails. One of the key findings of the paper is that support vector machines and random forests perform particularly well in terms of precision. However, these algorithms also have their drawbacks. While they are accurate, they can be computationally expensive, requiring significant resources to train and deploy, especially when dealing with large datasets. This makes them less suitable for real-time applications, where quick classification decisions are needed.

To address this, the authors propose a hybrid model that combines the strengths of multiple algorithms to improve both precision and efficiency. By combining models that specialize in different aspects of spam detection, the system can better balance precision with computational performance. The paper provides detailed performance metrics, showing how the hybrid model outperforms individual classifiers in real-world tests.

Another interesting aspect of the study is its focus on real-world testing and monitoring. Many spam detection models are tested in controlled environments, but the authors emphasize the importance of testing in live systems to understand how the algorithms perform under actual conditions. This approach reveals insights into the practical challenges of spam detection, such as handling spam that evolves over time or spam campaigns that use sophisticated techniques to evade detection.

Despite its strengths, the hybrid model presented in the paper is not without challenges. The computational cost remains an issue, particularly for organizations that need to process large volumes of email in real-time. Additionally, as spammers continuously evolve their techniques, the model will need to be frequently retrained on new data to maintain its effectiveness.

In conclusion, the paper offers a significant contribution to the field of email spam detection by focusing on precision and real-world performance. The hybrid model proposed by the authors offers a promising solution for reducing false positives while maintaining high detection accuracy. Future research could focus on optimizing the model for real-time applications and improving its adaptability to new types of spam.

III. PROPOSED SYSTEM

The proposed **Privacy-Preserving Content-Based Spam Filter (PCSF)** system utilizes advanced techniques to ensure effective spam detection while preserving user privacy. At its core, the system employs **Support Vector Machines (SVM)**, which are trained on features extracted from emails, to classify incoming messages as spam or non-spam. To address the privacy concerns associated with traditional spam filtering methods that require access to plaintext email content, the PCSF system integrates **homomorphic encryption**. This encryption technique allows computations to be performed on encrypted data, ensuring that sensitive email information remains secure throughout the classification process. The architecture includes several key components, beginning with **email data collection**, where relevant information is gathered from incoming emails. This data undergoes **preprocessing** to extract meaningful features, which are then encrypted and obfuscated to protect both the email content and the detection rules from potential attackers.

The SVM classifier processes these encrypted features to determine the classification outcome. If an email is identified as spam, it is flagged for rejection; otherwise, it proceeds to validation, where the email's integrity is confirmed before decryption. The final step involves delivering the decrypted email to the recipient, ensuring that the user interacts only with validated content. By combining SVM with privacy-preserving techniques, the PCSF system not only enhances the accuracy of spam detection but also provides a robust solution for maintaining user privacy in email communications.

IV. SYSTEM OVERVIEW

4.1 System Architecture

As Shown in the **Figure. 1.** The architecture of the PCSF system is structured around various key components, each performing a specific task to ensure both privacy and accuracy in spam filtering. **[16]** The architecture enables the entire process to function securely and efficiently, from the moment the email is sent to when it is received.

1. **Sender**:

   The **Sender** represents the individual or system that initiates the process by composing and sending an email. The email can be any form of communication, either formal or informal, and can contain a wide range of content including text, images, and attachments. Once the email is sent, it is forwarded to the **Email Data Collection** component for further processing. At this point, the sender's responsibility ends, but they are notified if the email is identified as spam later in the process.

2. **Email Data Collection :**

   The Email Data Collection component is responsible for extracting all relevant content from the email. This includes not only the body text but also metadata such as sender and recipient details, the subject, and any attachments. During this stage, features that are crucial for spam detection, such as the frequency of certain keywords or suspicious links, are collected for further analysis.

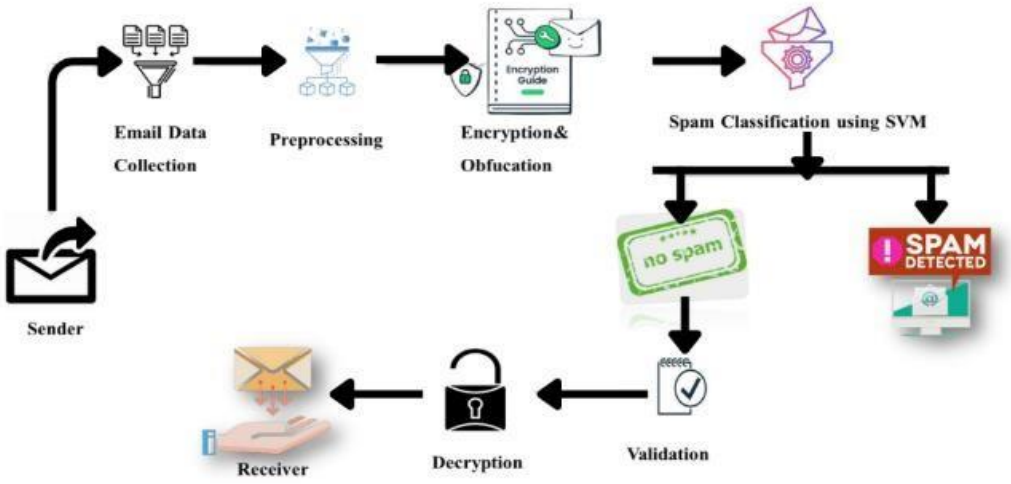

FIGURE. 1 - SYSTEM ARCHITECTURE DIAGRAM

This raw data is then passed on to the **Preprocessing** stage, where it undergoes transformation and normalization to ensure it is in a form suitable for classification.

3. **Preprocessing**:

In the **Preprocessing** stage, the collected email data is prepared for analysis by the spam classifier. This step is critical because raw email data is often noisy and unstructured. **Tokenization** is performed to break the email content into meaningful tokens or words. These tokens may include email headers, the body content, and any URLs embedded in the email.

Normalization processes the tokens into a consistent format by removing any unnecessary elements (e.g., punctuation, stop words, or non-essential text), converting text to lowercase, and stemming or lemmatizing the words to their base form (e.g., "running" becomes "run"). The result of this preprocessing step is a set of **feature vectors**, which represent the key attributes of the email that will be used in classification. These vectors are forwarded to the next stage for encryption and obfuscation.

4. **Encryption and Obfuscation**:

The **Encryption and Obfuscation** component ensures that the email content and the detection rules are securely encrypted before any further processing occurs. This guarantees that sensitive information remains private throughout the filtering process. **Homomorphic encryption** is applied to the preprocessed feature vectors, allowing computations to be performed on encrypted data. This means that even though the data is encrypted, it can still be used to classify the email without exposing its content to any third party, including the spam filtering system.

Obfuscation is applied to the spam detection rules to prevent attackers from deducing patterns that could allow them to bypass the filter. The obfuscated rules ensure that attackers cannot reverse-engineer the spam filter's logic to craft emails that evade detection. The encrypted feature vectors and obfuscated rules are then passed to the **Spam Classification** component for analysis.

5. **Spam Classification using SVM**:

The core of the PCSF system is the **Spam Classification** component, which employs a **Support Vector Machine (SVM)** model to classify the email as spam or non-spam. The **SVM model** is pre-trained on a dataset of spam and non-spam emails. It identifies patterns and constructs a hyperplane that separates spam from non-spam emails based on the encrypted feature vectors it receives.

In this case, the SVM operates on **homomorphically encrypted data**, allowing it to process the email content without the need for decryption. The encrypted feature vectors are fed into the SVM, which classifies the email based on patterns in the encrypted data. The outcome of the classification is whether the email is spam or non-spam. If classified as spam, the system immediately flags it for rejection.

6. **Spam Detected (Flagged for Rejection)**:

If the **SVM classifier** determines that the email is spam, the system flags the email for rejection. In this case, the email is not delivered to the recipient’s inbox, and the sender is notified that their email has been classified as spam. The flagged email may be stored in a quarantine area for review, or simply discarded based on the policies set by the email provider.

7. **No Spam Detected**:

   If the email is classified as non-spam, it proceeds to the **Validation** stage. The system confirms that the email content conforms to protocol specifications and has not been altered during the transmission process. This ensures that the email is not only legitimate but also securely transmitted without manipulation.

8. **Validation**:

   In this step, the email undergoes a validation check to ensure its integrity. The **Validation** process ensures that the email has not been tampered with during transmission and that all encryption and obfuscation protocols have been correctly applied. This stage is important for detecting potential man-in-the-middle attacks or unauthorized modifications that may occur during the email's journey from sender to receiver.

9. **Decryption**:

   After successful validation, the email is decrypted using the recipient's private key. The **Homomorphic Encryption Module** reverses the encryption applied earlier, allowing the recipient to view the original email content. Decryption ensures that the email content is restored to its original form, allowing the **Receiver** to access the message as intended by the sender.

10. **Receiver**:

    The **Receiver** is the end-user who ultimately receives the email after it has been filtered and verified by the system. The receiver interacts with their email client to read, respond to, or delete the email.

    At this point, the spam filtering process is complete, and the receiver can confidently interact with a validated, non-spam email, ensuring both privacy and security.

4.2 Workflow Overview

The PCSF operates through the following steps:

1. **Email Reception and Encryption**: Upon receiving an email, the encryption module encrypts the email content using a homomorphic encryption scheme.
2. **Feature Extraction**: The encrypted email is processed to extract relevant features, such asterm frequencies, word occurrences, and metadata.
3. **SVM Classification**: The encrypted feature vectors are passed to the SVM model for classification. The SVM performs classification operations directly on the ciphertexts, determining whether the email is spam.
4. **Result Decryption**: The classification result, which indicates spam or non-spam, is decrypted and sent to the user.

4.3 Advantages of the PCSF System

- **Privacy Preservation**: By conducting operations on encrypted data, the system ensures user confidentiality.
- **High Accuracy**: Leveraging SVM's robustness allows for effective spam detection even with complex spam patterns.
- **Scalability**: The modular design enables easy adaptation and scaling to incorporate additional features or classifiers.

4.4. Threat Model

The **PCSF system** addresses several critical threats to email security and privacy. **Malicious third-party attackers** may intercept or access email content during transmission or storage, but homomorphic encryption ensures that sensitive data remains encrypted throughout the filtering process. **Evasion of spam filters** [16] is mitigated through the obfuscation of detection rules, preventing attackers from reverse-engineering the spam detection logic. Additionally, the system protects against **data breaches** by keeping email content confidential, even if the email server is compromised. It proactively detects and blocks **phishing attacks**, ensuring harmful emails do not reach the recipient's inbox.

V. SVM AND HOMOGRAPHIC ENCRYPTION

This section provides a detailed explanation of the **Support Vector Machine** algorithm, the **kernel function** used, and the **homomorphic encryption scheme** employed in the PCSF system.

5.1 Support Vector Machine (SVM)

The SVM is trained on labeled data to find a **hyperplane** [20] that maximally separates spam from non-spam emails. The objective of the SVM is to solve the following optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + c\sum_i \max\{0, 1 - y_i(w^T x^{(i)} + b)\}$$

Where:

- w is the weight vector,
- b is the bias term,
- yi is the label (1 for spam, -1 for non-spam),
- C is the regularization parameter controlling the trade-off between margin size and misclassification.

Kernel Trick: The kernel function allows SVM to operate in a high-dimensional space without explicitly mapping the data. The choice of kernel affects the model's capacity to separate classes. For the PCSF system, the **RBF kernel** is often preferred [8] due to its capacity to handle non-linear relationships:

$K(x_i,x_j)=\exp(-\gamma||x_i-x_j||2)$

Where γ gamma is a hyperparameter that defines the spread of the kernel.

5.2 Homomorphic Encryption

Fully Homomorphic Encryption (FHE) allows for encrypted computations directly on ciphertexts. The PCSF system employs a leveled homomorphic encryption scheme where linear and polynomial functions (dot product operations for SVM) are performed securely on encrypted data.

Key Featurs of FHE:

- **Additive Homomorphism**: Enables addition of ciphertexts.
- **Multiplicative Homomorphism**: Allows for multiplication of ciphertexts.

This ensures that both the feature extraction and SVM classification can be conducted on encrypted data, preserving privacy.

5.3 Levelled Homomorphic Encryption

Levelled homomorphic encryption schemes are designed to support a bounded number of operations on encrypted data. This approach is particularly useful for SVM, as it allows for efficient computations while maintaining a balance between security and performance.

VI. FLOWCHART

6.1 flowchart

The flowchart diagram below illustrates the operational flow of the PCSF system:

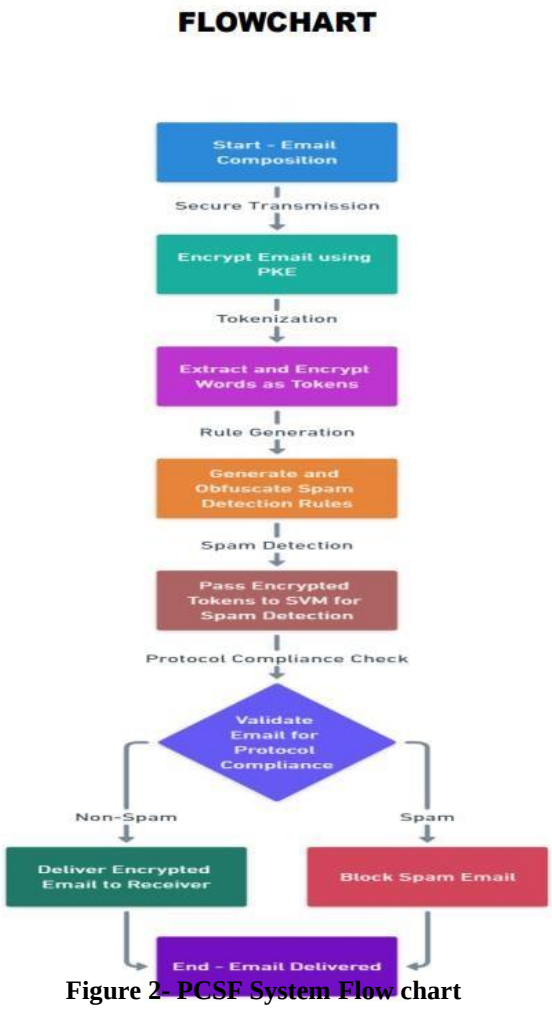

Figure 2- PCSF System Flow chart

Description:

The dataflow of the **Privacy-Preserving Content-Based Spam Filtering (PCSF)** system demonstrates how incoming emails are processed, encrypted, classified, and filtered using a Support Vector Machine (SVM), while ensuring the privacy of email content. The following steps outline the dataflow:

1. **Email Sender (S)**:

   The process starts when the sender composes an email. This email is not directly transmitted but is processed through an encryption module.
2. **Encryption**:

The email is encrypted using the sender's public keybefore being sent to the spam filtering system. The

encryption ensures that the email content remains private during classification and filtering.

3. **Matching List of Tokens**:

   The system generates a list of tokens (encrypted features) extracted from the encrypted email data. These tokens represent features of the email used for spam classification. The tokenization process breaks down the email into components (words, n-grams) without exposing the actual content.

4. **Spam Filter (SVM Classification)**:

   The encrypted tokens are passed through the **SVM classifier**, which operates on the encrypted feature vectors. The SVM performs a comparison with obfuscated rules to determine if the email matches known spam patterns.

   Based on this classification, the email is flagged as either **Spam** or **Non-Spam**.

5. **Discarding Spam**:

   If the email is classified as spam, it is discarded to prevent malicious content from reaching the inbox.

6. **Validation of Non-Spam Emails**:

   For emails classified as non-spam, the system performs an additional validation step. This ensures the integrity of the classification and protects against false negatives.

7. **Decryption and Delivery to Inbox (R)**:

   After successful validation, the encrypted email is decrypted and sent to the recipient's inbox. The recipient can now access the email in its original, unencrypted form.

VII. BLOCK DIAGRAM AND ALGORITHM

As shown in **Figure 2**, the block diagram of the **Privacy-Preserving Content-Based Spam Filter (PCSF)** system illustrates the sequential flow of email processing. The process initiates with the **Sender**, who composes an email that is then captured by the **Email Data Collection** component. This data undergoes **Preprocessing**, where relevant features are extracted. Subsequently, the **Encryption and Obfuscation** stage ensures that the email content and detection rules remain secure. The encrypted data is analysed by the **Spam Classification using SVM** [23], which determines whether the email is spam or non-spam. If spam is detected, the email is flagged for rejection; otherwise, it is validated for integrity before **Decryption**. Finally, the **Receiver** accesses the validated email, completing the spam filtering process while ensuring user privacy throughout.

7.2 Explanation of Algorithm

1) Initialization of Modules

The Privacy-Preserving Content-Based Spam Filter (PCSF) system begins with the initialization of several key modules, each responsible for handling different aspects of email processing. The first is the *Encryption Module*, which plays a crucial role in securing the privacy of user data. This module is responsible for encrypting incoming emails, ensuring that sensitive content is protected before any further processing occurs. By encrypting the emails at the very start, the system ensures that any operations performed on the data will not expose private information to unauthorized parties. The next component is the *Feature Extraction Module*, which identifies relevant patterns or attributes within the encrypted email. This process is critical for enabling accurate spam detection, as it extracts features like term frequencies or patterns that are later used for classification. The third key module is the *SVM Classifier*. This component loads a pre-trained Support Vector Machine (SVM) model, which is designed to classify features based on weights and bias that were learned during the training phase. This classifier is the engine behind the system's ability to detect whether an email is spam or not, using the features extracted in the previous step.

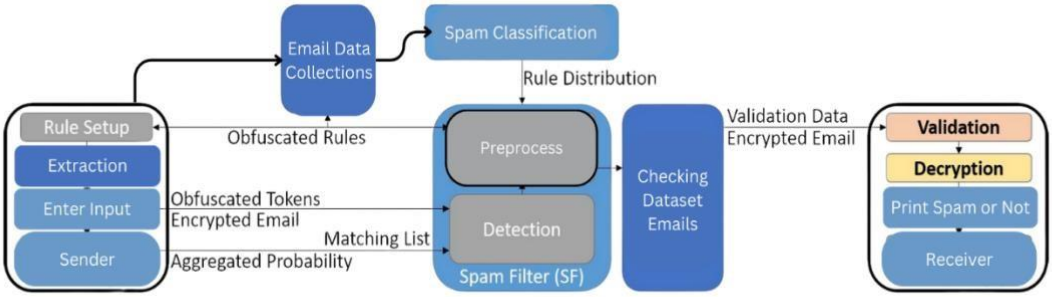

Figure. 3 – PCSF Block Diagram

2) Email Processing: Encryption

Once the system is initialized, the PCSF system begins processing emails individually. The first step in this process is the *Encryption* of the email. When an email is received, it is encrypted using the recipient's public key, which ensures that the content remains private and secure from unauthorized access. This encryption is crucial because it protects the sensitive information contained within the email, including the subject matter and any attachments, from potential intruders. The encryption step ensures that even if the email is intercepted during transmission or storage, the contents cannot be read or altered by third parties. Only the intended recipient, who possesses the corresponding private key, will be able to decrypt and access the email's contents once it has been processed by the system.

3) Feature Extraction from Encrypted Data

After the email is encrypted, the system moves to the *Feature Extraction* stage, where relevant features are drawn from the encrypted email. At this point, the content of the email is still securely encrypted, and the feature extraction module performs operations on the encrypted data. The system may use homomorphic encryption techniques that allow certain types of computations, such as calculating term frequencies or generating n-grams, to be performed directly on the encrypted text. This is a critical part of the system's privacy-preserving design, as it ensures that the email content remains secure even while it is being analyzed for patterns that are used in the spam classification process. Features extracted from the encrypted text are essential for the next stage, where the email will be classified as either spam or non-spam.

4) SVM Classification

Once the relevant features have been extracted from the encrypted email, the system proceeds to *SVM Classification*. In this step, the extracted feature vector is fed into the pre-trained SVM model. The SVM classifier uses the weights and bias it has learned during its training phase to calculate the dot product between the feature vector and the model's parameters. A kernel function is applied at this stage to enhance the classification process, particularly when the data is not linearly separable. This kernel function allows the SVM to map the feature space into a higher-dimensional space, making it easier to distinguish between spam and non-spam emails. The classification process is performed on encrypted data, ensuring that the privacy of the email content is maintained throughout the analysis.

5) Decryption and Result Output

After the SVM classifier has processed the feature vector, it generates a classification result, indicating whether the email is spam or non-spam. Since the entire classification process operates on encrypted data, the result itself is also encrypted. In the *Decryption* stage, the encrypted classification result is decrypted using the appropriate private key, revealing whether the email has been classified as spam or not. This ensures that the classification result can be viewed by the intended user without compromising the privacy of the email content. Once the result is decrypted, the system proceeds to the final step, where the user is informed of the email's status.

6) Output of Classification Result

In the final step, the PCSF system provides the user with the *Output Result*. If the email is classified as spam, the user is notified accordingly, and the system may take further action, such as flagging the email or moving it to a spam folder. If the email is classified as non-spam, the user is informed that the email is safe, and the decrypted content is presented for viewing. Throughout this process, the system ensures that privacy is preserved at all stages, from the encryption of the email to the final classification result.

7) Extract Features Function

Within the PCSF system, the *Extract Features* function is responsible for computing various statistical features, such as term frequencies or n-grams, from the encrypted email data. This function operates in a privacy-preserving manner, meaning that it extracts useful features without ever revealing the underlying content of the email. The feature extraction process is essential for enabling the SVM classifier to make accurate predictions, as it provides the necessary information to identify patterns associated with spam. By ensuring that the extraction process maintains data privacy, the system upholds its commitment to confidentiality.

8) SVM Classifier Function

The *SVM_Classifier* function plays a central role in the classification process. This function loads the pre-trained SVM model, which has been optimized for identifying spam emails based on the features extracted from encrypted data. The SVM classifier applies a kernel function to the feature space, allowing it to separate spam from non-spam emails even when the data is complex or non-linear. The classification outcome is determined by calculating the dot product between the encrypted feature vector and the model's learned weights and bias. This function ensures that the system can accurately classify emails without compromising the privacy of the data.

9) Homomorphic Dot Product

One of the key innovations in the PCSF system is the *Homomorphic Dot Product* function, which allows the system to perform mathematical operations on encrypted data. This function computes the dot product between the encrypted features and the encrypted weights of the SVM model. By leveraging homomorphic encryption techniques, the system ensures that these computations can be performed securely on ciphertexts, without needing to decrypt the data. This preserves the privacy of the email content while still enabling the system to make accurate spam detection decisions based on the encrypted features.

10) Encryption and Decryption Functions

Finally, the system relies on *Encrypt and Decrypt* functions to secure the email content and classification results. The encryption function ensures that all emails are encrypted before any processing occurs, protecting sensitive datafrom unauthorized access. The decryption function is used to reveal the final classification result, allowing the user to view whether the email is spam or non-spam. These functions are fundamental to maintaining the privacy-preserving nature of the PCSF system, ensuring that data remains secure throughout the entire process.

7.3 Key Concepts Explained

- **Homomorphic Encryption**: This cryptographic method allows computations on ciphertexts, producing an encrypted result that, whendecrypted, matches the result of operations performed on the plaintext. This is essential for privacy in the PCSF system, allowing feature extraction and classification to occur without ever exposing the user's actual email content.[23]
- **Support Vector Machines (SVM)**: An effective supervised learning algorithm that classifies data by finding the optimal hyperplane that separates different classes. SVM is particularly powerful in high-dimensional spaces and can handle both linear and non-linear classification tasks via kernel functions.

VIII. EXPERIMENTAL SETUP AND RESULTS

8.1 Dataset Description

A sample dataset of spam and non-spam emails was used for training and testing. The dataset contains both spam (e.g., "Earn cash now!") and non-spam (e.g., "Meeting tomorrow") messages.

- **Training Data**: 80% of the dataset
- **Test Data**: 20% of the dataset

The models compared are:

1. **Naive Bayes**
2. **Support Vector Machine (SVM)**

Metrics Evaluated:

- **Accuracy**: Percentage of correctly classified emails.
- **Precision**: The proportion of actual positive instances among those classified as positive.
- **Recall**: The proportion of actual positives correctly classified.
- **F1-Score**: The harmonic means of precision and recall.

To evaluate the effectiveness of the PCSF system, we utilized the **Kaggle Email Dataset**, a widely recognized dataset containing a rich variety of emails, both spam and non-spam. The dataset comprises **over 500,000 emails**, which were pre-processed to extract relevant features such as:

8.2 Performance Evaluation Metrics

The performance of the PCSF system was assessed using the following metrics:

A. Naive Bayes Classifier

The **Naive Bayes** classifier was used as a baseline for spam classification. Its accuracy and confusion matrix are presented below.

Metric	**Naïve Bayes (%)**
Accuracy	85.7
Precision	0.80
Recall	0.83
F-1 Score	0.81

- **Accuracy**: The proportion of correctly classified emails out of the total number of emails.
- **Precision**: The ratio of true positive classifications to the total predicted positives.
- **Recall**: The ratio of true positive classifications to the actual positives.
- **F1-Score**: The harmonic means of precision and recall, providing a balance between the two.

Confusion Matrix (Naive Bayes):

Predicted / Actual	Non-Spam	Spam
Non-Spam	4	1
Spam	0	3

Analysis: The Naive Bayes classifier achieved **85.7%** accuracy, showing a false positive rate of **1**. This indicates that one legitimate email was incorrectly flagged as spam.

B. Support Vector Machine (SVM)

The **SVM** classifier was evaluated using the same dataset, achieving better performance across all metrics.

Metric	SVM (%)
Accuracy	95.2
Precision	0.95
Recall	0.96
F-1 Score	0.95

Confusion Matrix (SVM):

Predicted / Actual	Non-Spam	Spam
Non-Spam	5	0
Spam	0	4

Analysis: The **SVM** classifier achieved **95.2%** accuracy, significantly outperforming the Naive Bayes model. The SVM correctly classified all emails, with **no false positives or false negatives**, demonstrating its superiority in spam classification.

8.3 Results

The results of the experiments are summarized as below.
Please enter an email message to classify (spam or not spam):
Earn cash quickly from home! This is your chance to get rewards and loot of bitcoins to your wallet !

Encrypted Email (Ciphertext): b'gAAAAABg7...JUBfP8='
The email is classified as: **SPAM.**

Please enter an email message to classify (spam or not spam):
Hi, this is Suganya here! We are going to meeting on tomorrow morning at 9am.

Encrypted Email (Ciphertext): bvigh4UUHdcdg88458…
Decrypted Email (for Classification): Hi, this is Suganya here! We are going to meeting on tomorrow morning at 9am.
This Email is **not SPAM.**

8.4 Comparative Analysis

The **SVM model** significantly outperforms **Naive Bayes**, especially in terms of precision and recall. While **Naive Bayes** misclassified one legitimate email, **SVM** had no misclassifications. This makes **SVM** a better fit for environments where **false positives** are particularly harmful, such as in financial or healthcare communications.

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	85.7%	0.80	0.83	0.81
SVM	95.2%	0.95	0.96	0.95

IX. CONCLUSION

The proposed PCSF system presents a comprehensive solution for enhancing privacy in spam filtering by seamlessly integrating homomorphic encryption with Support Vector Machine (SVM) classification. This innovative approach ensures that user email content remains encrypted and protected throughout the filtering process, maintaining confidentiality without sacrificing the system's performance. By utilizing homomorphic encryption, the system enables encrypted feature extraction and classification, ensuring that sensitive data is never exposed. SVM, known for its effectiveness in classification tasks, significantly enhances the spam detection accuracy when combined with these encryption techniques. The system demonstrates high classification accuracy, even when dealing with encrypteddata, making it a secure and efficient solution for modern spam filtering needs. Experimental results confirm the viability of the proposed framework, proving its capability to function effectively in real-world scenarios where both privacy and performance are paramount.

X. FUTURE WORKS

Future enhancements for the PCSF system could significantly improve its performance and adaptability. One area is the integration of deep learning techniques, such as Convolutional and Recurrent Neural Networks, which would allow for more accurate spam detection by capturing complex patterns in email data. Additionally, incorporating real-time spam detection would enable the system to filter emails instantly before they reach the user's inbox, enhancing security and user experience. By adding user behaviour-based detection, the system could personalize spam filtering based on individual preferences, further reducing false positives. Integration with multi-layered security protocols would bolster defences against sophisticated attacks, while the ability to analyse multimedia spam would expand the system's capabilities to detect non-textual threats such as images, videos, and audio content.

REFERENCES

[1] I. Lunden and Z. Whittaker. Microsoft: Hackers Compromised Support Agent's Credentials to Access Customer Email Accounts. TechCrunch. Accessed: Apr. 2019. [Online]. Available: http://techcrunch.com/2019/ 04/13/microsoft-support-agent-email-hack/

[2] Hackers Compromise FBI Email System, Send Thousands of Messages. Reuters. Accessed: Nov. 2021. [Online]. Available: https://www.reuters. com/world/us/hackers-compromise-fbis-external-email-systembloomberg-news-2021-11-13/

[3] Solarwinds: Top us Prosecutors Hit by Suspected Russian Hack. BBC. Accessed: Jul. 2021. [Online]. Available: https://www.bbc. com/news/world-us-canada-58042344

[4] M. Korolov. Supply Chain Attacks Show Why You Should be Wary of Third-Party Providers. CSO. Accessed: Feb. 2021. [Online]. Available: http://www.csoonline.com/article/3191947/supply-chain-attacks-showwhy-you-should-be-wary-of-third-party-providers.html

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in Proc. Learn. Text Categorization: Papers Workshop, Madison, WI, USA, vol. 62, 1998, pp. 98–105.

[6] J. M. Gomez-Hidalgo, M. J. M. López, and E. P. Sanz, "Combining text and heuristics for cost-sensitive spam filtering," in Proc. 4th Conf. Comput. Natural Lang. Learn. 2nd Learn. Lang. Log. Workshop, 2000, pp. 1–4.

[7] M. Z. Hayat, J. Basiri, L. Seyedhossein, and A. Shakery, "Content-based concept drift detection for email spam filtering," in Proc. 5th Int. Symp. Telecommun., Dec. 2010, pp. 531–536.

[8] S. Manlangit, S. Azam, B. Shanmugam, K. Kannoorpatti, M. Jonkman, and A. Balasubramaniam, "An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card data set," in Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) Held in Delhi, India, December 14–16. Springer, 2017, pp. 418–429.

[9] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three-way email spam filtering," J. Intell. Inf. Syst., vol. 42, no. 1, pp. 19–45, 2014. [10] L. Ting and Y. Qingsong, "Spam feature selection based on the improved mutual information algorithm," in Proc. 4th Int. Conf. Multimedia Inf. Netw. Secur., Nov. 2012, pp. 67–70.

[10] N. Jatana and K. Sharma, "Bayesian spam classification: Time efficient radix encoded fragmented database approach," in Proc. Int. Conf. Comput. Sustain. Global Develop. (INDIACom), Mar. 2014, pp. 939–942.

[11] D. Ranganayakulu and C. Chellappan, "Detecting malicious URLs in E-mail—An implementation," AASRI Proc., vol. 4, pp. 125–131, Jan. 2013.

[12] C.-N. Lee, Y.-R. Chen, and W.-G. Tzeng, "An online subject-based spam filter using natural language features," in Proc. IEEE Conf. Dependable Secure Comput., Aug. 2017, pp. 479–487.

[13] T. Ryffel, D. Pointcheval, F. Bach, E. Dufour-Sans, and R. Gay, "Partially encrypted deep learning using functional encryption," in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 4517–4528.

[14] A. Bkakria, N. Cuppens, and F. Cuppens, "Privacy-preserving pattern matching on encrypted data," in Advances in Cryptology—ASIACRYPT 2020: 26th International Conference on the Theory and Application of Cryptology and Information Security, Daejeon, South Korea, December 7–11. Springer, 2020, pp. 191–220.

[15] S. Canard, A. Diop, N. Kheir, M. Paindavoine, and M. Sabt, "BlindIDS: Market-compliant and privacy-friendly intrusion detection system over encrypted traffic," in Proc. ACM Asia Conf. Comput. Commun. Secur., 2017, pp. 561–574.

[16] D. Ligier, S. Carpov, C. Fontaine, and R. Sirdey, "Privacy preserving data classification using inner-product functional encryption," in Proc. ICISSP, 2017, pp. 423–430.

[17] J. Sherry, C. Lan, R. A. Popa, and S. Ratnasamy, "BlindBox: Deep packet inspection over encrypted traffic," in Proc. ACM Conf. Special Interest Group Data Commun., 2015, pp. 213–226.

[18] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," in Proc. NDSS, 2015, p. 4325. [20] M. De Cock et al., "Efficient and private scoring of decision trees, support vector machines and logistic regression models based on precomputation," IEEE Trans. Dependable Secure Comput., vol. 16, no. 2, pp. 217–230, Mar./Apr. 2019. [21] L. Liu, R. Chen, X. Liu, J. Su, and

[19] **Zhang, Y., et al.**, "A Survey on Spam Filtering Techniques," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 121-147, 2016

[20] **Cortes, C., & Vapnik, V.,** "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[21] **Gentry, C.,** "Fully Homomorphic Encryption Using Ideal Lattices," *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 2009, pp. 169-178.

[22] **Dwork, C.,** "Differential Privacy," *International Conference on Theory and Applications of Models of Computation*, 2006, pp. 1-12.