electronics

MDPI

Article

Filtering and Detection of Real-Time Spam Mail Based on a Bayesian Approach in University Networks

Maksim Sharabov [1], Georgi Tsochev [1,*], Veska Gancheva [2] and Antoniya Tasheva [3]

[1] Department of Information Technologies in Industry, Faculty of Computer Systems and Technology, Technical University of Sofia, 1000 Sofia, Bulgaria; msharabov@tu-sofia.bg
[2] Department of Programming and Computer Technologies, Faculty of Computer Systems and Technology, Technical University of Sofia, 1000 Sofia, Bulgaria; vgan@tu-sofia.bg
[3] Department of Computer Systems, Faculty of Computer Systems and Technology, Technical University of Sofia, 1000 Sofia, Bulgaria; atasheva@tu-sofia.bg
* Correspondence: gtsochev@tu-sofia.bg

Abstract: With the advent of digital technologies as an integral part of today's everyday life, the risk of information security breaches is increasing. Email spam, commonly known as junk email, continues to pose a significant challenge in the digital realm, inundating inboxes with unsolicited and often irrelevant messages. This relentless influx of spam not only disrupts user productivity but also raises security concerns, as it frequently serves as a vehicle for phishing attempts, malware distribution, and other cyber threats. The prevalence of spam is fueled by its low-cost dissemination and its ability to reach a wide audience, exploiting vulnerabilities in email systems. This paper marks the inception of an in-depth investigation into the viability and potential implementation of a robust spam filtering and prevention system tailored explicitly to university networks. With the escalating threat of email-based hacking attacks and the incessant deluge of spam, the need for a comprehensive and effective defense mechanism within academic institutions becomes increasingly imperative. In exploring potential solutions, this study delves into the applicability and efficacy of Bayesian filters, a class of probabilistic classifiers renowned for their aptitude in distinguishing between legitimate emails and spam messages. Bayesian filters utilize statistical algorithms to analyze email content, learning patterns and features to accurately categorize incoming emails.

Keywords: spam; Bayesian network; university network; phishing email detection; email security; spam filtering

Citation: Sharabov, M.; Tsochev, G.; Gancheva, V.; Tasheva, A. Filtering and Detection of Real-Time Spam Mail Based on a Bayesian Approach in University Networks. *Electronics* **2024**, *13*, 374. https://doi.org/10.3390/electronics13020374

Academic Editor: Hung-Yu Chien

Received: 25 November 2023
Revised: 13 January 2024
Accepted: 14 January 2024
Published: 16 January 2024

1. Introduction

In today's world, digitalization is perhaps the fastest growing process in social and business environments, which is related to the boom in new technology in recent years, as well as the drive for greater efficiency in the management of social, personal, and business life.

With the emergence of digital technologies as an integral part of today's everyday life, the risk of information security breaches is also increasing. Security breaches can be due to viruses or physical theft of equipment. The rapid pace at which dangers spread makes the security solutions used until recently insufficiently advanced and reliable. Nevertheless, the human factor remains the entity that accounts for the largest share of risk for information security breaches, so it is necessary to ensure technology users are aware of the dangers that can lead to a security breach: electronic identity theft or hackers running a home computer.

Spam messages are unwanted, unsolicited or unpleasant messages sent via email, text messages, social networks, or other means of communication. They may be advertisements for goods or services, scams, false notifications, or other types of unsolicited information.

Email is one of the main targets for spam messages. Spam often contains links to websites that sell goods or services, request personal information, or can be used for fraud. In addition, spam is often used to send malware that can infect the recipient's computer.

To protect our networks from spam messages, spam filters in e-mail and other technologies are used to detect and block unwanted messages. However, spam can be spoofed or shifted to evade such filters, so continued development of methods to counter spam is necessary.

Internationally, spam for the last 7 years (for 2023 is till end of September) is shown in Figure 1 [1–4].

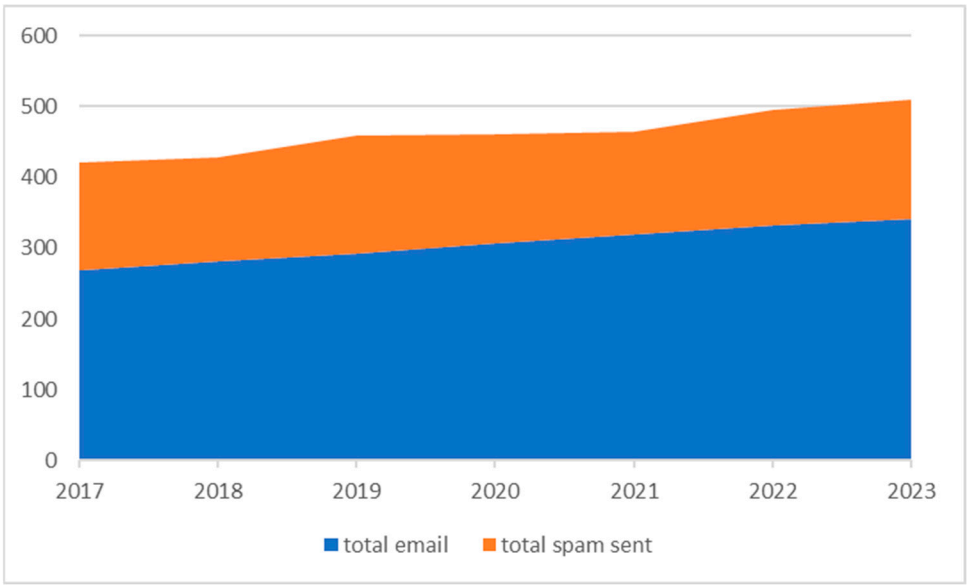


Figure 1. Send mails and spam.

Over the past year, spam has continued to be a serious problem in the online environment, affecting millions of users worldwide. In line with analytics and statistics, the following are some of the major trends in spam over the past year:

- The volume of spam messages: Despite various efforts to reduce junk mail, the volume of spam messages has remained significant. Over the last year, there has been a steady increase in the number of spam messages sent.
- Types of spam: Messages about unwanted goods and services continue to be the main type of spam. In addition, spam is changing and adapting to include social networks, mobile apps, and other platforms.
- Malware in spam: Malware in spam messages is becoming more sophisticated. It can hide in attachments, links to websites, or masquerade as a legitimate message.
- Focus on social engineering scams: Spam not only tries to sell goods and services, but also focuses on social engineering scams. It often involves fraudulent schemes, fake notifications, or messages that prompt users to share personal information.
- Geographical spread: Spam spreads globally, affecting users in different parts of the world. Different regions are exposed to different types of spam messages, depending on local trends and characteristics.

And despite efforts to fight spam through filters and spam detection technologies, it remains a serious problem. Its dynamics and volatility require constant updating of countermeasures to protect users and their data more effectively in the online environment.

This paper initiates an investigation into the feasibility of a university-based spam filtering and prevention system for countering email-based hacking attacks. The increasing frequency and sophistication of email-based hacking attacks pose significant security threats to academic institutions. Therefore, this research aims to evaluate the efficacy of a robust spam filtering mechanism tailored specifically to the university environment.

The study delves into the intricacies of email security within university networks, acknowledging the diverse communication needs and high volumes of email traffic prevalent in academic settings. It scrutinizes the vulnerabilities that leave university systems susceptible to spam attacks and explores potential avenues to fortify the email infrastructure against such incursions.

Furthermore, this research scrutinizes the potential of Bayesian filters as a pragmatic solution to curtail the incessant influx of spam messages. Leveraging machine learning techniques, particularly Bayesian classifiers, shows promise in accurately distinguishing between legitimate emails and unsolicited, potentially malicious spam.

By assessing the feasibility of Bayesian filters, this study endeavors to propose an efficient and adaptable framework capable of mitigating the risks posed by spam messages and bolstering the university's email security infrastructure. Ultimately, the aim is to establish a proactive defense mechanism, ensuring a secure communication environment conducive to academic pursuits.

2. Background of Spam

2.1. Types

Junk e-mail ("spam") is electronic mail that a user receives without having given permission for it to be sent and without having benefited from it, which at the same time involves the user in direct and indirect costs (in terms of time, Internet connection, etc.).

The term 'spam' gained popularity in the 1980s among online players of the network game MUD [5]. Back then, the word was used as a synonym for overwhelming one of the servers with too many messages or objects.

In the 1980s, the first cases of mass e-mails advertising mostly non-existent services were recorded. This is how the message entitled "MAKE MONEY FAST" sent on behalf of Dave Rhodes became known [6]. The idea he advertised was reverse-tracing a letter (which is not actually technically possible) and receiving a certain amount if the letter was forwarded to a given number of subscribers.

The term "spam" entered the Internet on 31 March 1993 within a message sent by Richard Depew [7]. It is interesting to note that the author was engaged in the development of a software product for the automatic moderation of electronic messages, in particular for the purpose of screening out advertising messages. The product was called Automated Retroactive Minimal Moderation (ARMM), but due to an error by the author, over 200 messages were sent to the news.admin.policy group, which handles USENET management. Joel Furr's research [8] is recognized for transferring the concept of "spam," initially conceived in the context of the MUD game, and extending its usage to the realm of the Internet.

Spam can be divided into different types depending on how it is sent and the type of information it contains. Herein are some of the main types of spam:

- Email spam: This is the most popular type of spam. It includes unsolicited messages sent via email that may be advertising, fraudulent, or contain malware or other types of unwanted information.
    - False advertising: this kind of spam offers of goods or services that are presented as profitable but are often false.
    - Phishing schemes: these are messages that are presented as legitimate by financial institutions, social networks, or others to solicit personal information from recipients.
    - Malware: these are emails that contain viruses, Trojans, ransomware, or other malware.
- Social media spam: This type of spam includes unsolicited messages, comments, or posts on social networks such as Facebook, Twitter, Instagram, and others. These messages may contain links to fake websites, fraudulent offers, or unwanted advertisements.

- Fake profiles and comments: this refers to the creation of fake profiles that post spam messages or comments with links to unwanted websites.
- Bribery for following: these are paid services used to increase numbers of followers or likes, often using fake profiles.

- SMS spam: Sending unsolicited text messages to mobile phones is also a type of spam. These messages can be promotional, fraudulent, or contain unwanted links or requests for personal information.
  - Promotional SMS: these are unsolicited promotional messages about products, services, or fraudulent schemes.
  - SMS with requests for personal information: these are fake SMS messages pretending to be official notifications from banks, companies, or other organizations in order to fraudulently obtain personal information.
- Forums and blogs (forum/blog spam): spam on forums and blogs includes unsolicited comments with links to websites, advertising messages or messages designed to attract visitors to certain services or products.
  - Comments with links: these are unsolicited comments that contain links to spam sites or content.
  - Automated posts: this refers to the grouping of mass sent messages for the purpose of advertising or promoting certain products.
- Botnet spam: this is a type of spam wherein a group of compromised computers (bots) are used to send spam messages en masse, often without the knowledge of the owners of those computers.
  - DDoS attacks: large networks of compromised computers are used to launch mass attacks to overwhelm websites or services.
  - Spam bots: this type of spam involves the mass sending of unsolicited messages from robotic accounts or compromised machines.
- Invitation spam: this type of spam includes unsolicited invitations to games, apps, or dating sent through various platforms or social networks.
  - Game invites: these are unsolicited invitations to play games that may often require personal information or payment.
  - Dating spam: these are unsolicited dating messages, often sent on online dating platforms.
- Fax spam: this term refers to the sending of unsolicited faxes with promotional material or other unsolicited messages, which is also a type of spam.
  - Promotional material: this involves unsolicited promotional faxes sent to fax machines that contain offers for services or goods.

These are just some of the main types of spam. Spam is constantly changing and evolving, with new methods and techniques being used to avoid filters and reach users.

2.2. Spam Filtering

Spam filtering in email is extremely important for separating junk mail from real and important messages. There are several methods and technologies used to filter spam, including the following:

1. Bayesian filters

These use Bayesian probability and analyze text characteristics to determine whether a message is spam or not. Their basic principle is to learn from multiple spam and non-spam messages to determine the probabilities of certain words, phrases, or characteristics occurring in these two types of messages.

2. Blacklist and whitelist filtering

Blacklists are lists of known spam source addresses or keywords that are used in spam messages. Whitelists are lists of known and trusted sources or addresses that are allowed and not considered spam. Mail servers use these lists to block or allow messages, respectively.

3. Keyword and phrase filtering

This method analyzes the message text and looks for certain keywords, phrases, or structures that are commonly associated with spam. Heuristics are used to detect such characteristics and decide whether the message is spam or not.

4. Machine learning and neural network filtering

Machine learning uses algorithms that learn from data to predict the classification of new data. Neural networks, such as deep learning models, can be trained to recognize spam characteristics and classify messages according to them.

5. Filtering by checking IP addresses and domains

This method analyzes the source IP addresses or domains of mail senders, and uses these data to determine their trust levels. Low-trust sources are often flagged as spam.

6. CAPTCHA filtering and human activity checking

Using CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) can help distinguish human activity from automated spam sending systems.

7. Filtering with SPF, DKIM, DMARC

These standards provide methods for authenticating mail senders that are used to fight against spoofing and message forgery.

Typically, modern mail systems and spam filters combine several of these methods to provide more effective detection and blocking of spam messages. They are constantly adapting and evolving to cope with new techniques and strategies used by spammers.

2.3. Background

The state-of-the-art in spam filtering using Bayesian classifiers involves several advancements and approaches that aim to enhance the effectiveness, accuracy, and adaptability of spam detection systems. Some notable developments include the following:

1. Improved Feature Representations:

Advancements in text representation techniques such as TF-IDF [9,10], word embeddings, and n-grams [11–16] have contributed to more robust feature representations of text data. These approaches help in capturing more nuanced information from messages, improving the classifier's ability to discern between spam and non-spam emails.

2. Advanced Bayesian Models [17–22]:

Researchers continue to explore and develop more sophisticated Bayesian models or variations of the naïve Bayes classifier. These efforts aim to address the limitations of the naïve Bayes assumption of feature independence, attempting to create more realistic models that better capture dependencies between features.

3. Hybrid Approaches [23–25]:

Integration of Bayesian classifiers with other machine learning techniques, such as support vector machines (SVM), neural networks, or ensemble methods, has shown promising results. Hybrid models leverage the strengths of multiple algorithms to enhance spam detection accuracy and reduce false positives.

4. Dynamic and Adaptive Filtering [26–29]:

Developments in adaptive spam filtering involve systems that can dynamically adapt to changing spamming techniques. These models continuously learn and update their parameters based on incoming data, allowing them to adapt to new forms of spam and maintain high accuracy over time.

5. Incorporating Behavioral Patterns [30–34]:

Some research focuses on incorporating behavioral patterns, user preferences, or contextual information into Bayesian classifiers. By considering user-specific behaviors or characteristics, these systems aim to personalize spam filtering, potentially reducing false positives.

6. Handling Imbalanced Datasets [35–39]:

Efforts have been made to address the issue of imbalanced datasets where the number of spam emails significantly outweighs non-spam emails. Techniques like oversampling, undersampling, or synthetic data generation aim to handle this imbalance and improve classifier performance.

The field of spam filtering is continually evolving with ongoing research and advancements.

3. Bayes Networks

Bayesian graphs [40] are statistical models that represent the probabilistic relationships between different variables by a graphical representation of these relationships. They are used to model probabilistic dependencies and analyze data in various fields such as machine learning, artificial intelligence, bioinformatics, medical applications, etc.

The main idea behind Bayesian graphs is to represent complex probabilistic models in an intuitive and easy-to-understand way. They use a graph structure that represents variables as nodes (nodes) in a graph, while the relationships between them are expressed by directed edges or links that show the probabilistic relationships or dependencies between variables.

Bayesian graphs consist of two main types of components:

- Vertices (nodes): these represent variables or events that are explored in the model. For example, in a medical context, vertices may represent diseases, symptoms, treatments, etc.
- Edges (relationships): These show probabilistic relationships or dependencies between variables. Edges can be directional or directionless, reflecting the direction of the interdependence between variables. For example, there may be direct relationships between a disease and a symptom, indicating the probability of a symptom being present for a particular disease.

Bayesian graphs use Bayesian probability theory, which allows probability estimates to be updated with the latest information. They can be used for statistical inference, forecasting, and decision making under uncertainty.

This type of modelling is used in a variety of applications where it is important to analyze probabilistic relationships between variables and make decisions based on these probabilities.

3.1. Bayesian Networks

Bayesian networks are probabilistic graphical models that use Bayesian probability theory to model probabilistic relationships between variables [40]. They are a graphical representation of probabilistic relationships in the form of an oriented acyclic graph (OAG).

The basic idea of Bayesian networks is to model the dependencies between variables by a graph, where the vertices of the graph represent random variables and the edges between them show the statistical dependencies. Bayesian networks allow the representation of complex probabilistic relationships between different variables in an intuitive and easy-to-understand way.

The main components of Bayesian networks are as follows:

1. Probabilistic vertices (random variables): Each variable or event being studied or modeled is represented as a vertex in the graph. For example, in a medical context, random variables can be diseases, symptoms, test results, etc.

2. Probabilistic relationships (edges): Edges between vertices reflect statistical dependencies or probabilistic relationships between variables. They show the interdependence between different variables and their probabilities.
3. Conditional probabilities: Bayesian networks use conditional probabilities, which indicate the probability of a variable given that other variables are present or known.

Bayesian networks are widely used in many fields including medical diagnosis, forecasting, machine learning, artificial intelligence, and others. They are a useful tool for modeling uncertainty and making decisions without complete information. Using Bayesian networks, users can analyze the probabilities of events occurring and make decisions based on those probabilities.

The distinction between Bayesian graphs and Bayesian networks is subtle, and the terms are often used interchangeably. However, they usually refer to the same concept of modeling probabilistic relationships, but with nuances in the technical details.

Bayesian graphs are a general term that encompasses all graphical models that use Bayesian probability to represent probabilistic relationships between variables. They include different types of graphs that can be oriented or unoriented and are composed of vertices (variables) and edges (probabilistic relationships).

Bayesian networks are a specific type of Bayesian graphs that are characterized by oriented acyclic graphs (OAGs), in which the vertices represent random variables and the edges show the statistical dependencies or probabilistic relationships between them. Bayesian networks have a rigorous structure that allows the representation of conditional dependencies between variables, which is important for analyzing probabilistic relationships in as compact a form as possible.

Thus, in the broader sense, Bayesian graphs encompass all types of graphical models, including Bayesian networks, which are a particular subtype of Bayesian graphs. Bayesian networks are a more specific case than Bayesian graphs, and are typically used to model conditional dependencies between variables, such as in fields like medical diagnosis, finance, artificial intelligence, and others where accurate modeling of these dependencies is essential.

3.2. Bayesian Networks for Spam Filtering

Bayesian spam filters [41] are extremely useful tools used to automatically detect and filter junk mail in email. These filtering methods are based on Bayesian probability, and are known for their effectiveness in various environments in which we face the problem of spam messages being sent.

The idea behind Bayesian spam filters is to recognize and classify whether an e-mail is spam or not. This is achieved by analyzing the characteristics of the text or other attributes of the message. The filter is trained using multiple spam and normal (non-spam) messages to learn the distinguishing features of each of these two message types.

The operation of Bayesian spam filters goes through several stages:

- Training the model: First, the filter is trained using a large amount of spam and normal messages. It analyzes these messages and learns what words, phrases, structures, or other characteristics are more associated with spam or non-spam messages.
- Probability calculation: Once trained, the filter calculates the probability of certain features (e.g., words, phrases, structures) occurring in spam and non-spam messages. These probabilities are used to evaluate whether a new message is spam or not.
- Message classification: When a new message arrives, the filter analyzes the text and its characteristics. According to the learned probabilities, it calculates the likelihood of the message being spam or non-spam. According to this probability, the message is classified as spam or not spam.
- Classification threshold: The filter uses a threshold (threshold) that determines when to classify a message as spam or non-spam. For example, if the probability of spam is greater than a certain threshold, the message is marked as spam.

- Model update: Bayesian filters can be updated with new data, allowing for continuous improvement of their classification accuracy.

These filters are used in email, social networks, online platforms, and other places where it is important to recognize and separate spam messages from normal communication. They are effective in catching multiple types of spam, such as sending advertisements, fraudulent messages, malware and more.

Bayesian-based spam filters provide high accuracy in recognizing spam messages, but can also produce some false positives or negatives that can lead to misclassified messages.

Indeed, Bayesian spam filters are based on Bayesian probability, and use formulas from this theory to classify e-mail messages as spam or non-spam. One of the main methods used in these filters is the Bayes rule formula (Bayes theorem). This formula is expressed as follows:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (1)$$

where

- (P(A | B)) is the probability of event A happening, given that event B has already happened;
- (P(B | A)) is the probability of event B happening, given that event A has already happened;
- (P(A)) and (P(B)) are the individual probabilities of events A and B occurring, respectively.

In the context of Bayesian spam filters, Formula (1) is used to calculate the probability of an e-mail message being spam or not. As mentioned above, the filter uses the probabilities of occurrence of certain features in spam and non-spam messages for the classification.

Let us represent that (A) is the event that a message is spam, and (B) is the feature of the message (e.g., the presence of a particular word or phrase). Then, applied to Bayes' formula, this can be expressed as follows:

$$P(spam|characteristics) = \frac{P(characteristic|spam) \times P(spam)}{P(characteristic)} \quad (2)$$

- (P({spam} | {characteristic})) is the probability of the message being spam, given that it has a certain characteristic;
- (P({characteristic} | {spam})) is the probability of that characteristic occurring in spam messages;
- (P({spam})) is the probability of the message being spam regardless of the characteristic;
- (P({feature})) is the overall probability of this feature occurring in all messages, whether spam or not.

The filter uses Formula (2) to estimate the probability of classifying a message as spam based on the presence or absence of certain characteristics that are associated with spam messages. Using this information, the filter decides whether to classify the message as spam.

4. Architecture

4.1. Methodology

This section provides a comprehensive overview of how the research was conducted, detailing each step from data collection to statistical analysis. It sets a clear path for replicating this study or building upon its findings.

1. Research design: this study employs a quantitative research design, focusing on the development and evaluation of a Bayesian-based spam filtering system within university networks.
2. Data collection: email traffic within the university's network was monitored to collect real-time data on spam infiltration and filtering effectiveness.
3. Bayesian filter implementation:
    - Training phase: The Bayesian filter was trained using the collected datasets, distinguishing between spam and non-spam emails.

- Testing and validation: the filter's performance was tested on real-time university email traffic.

4. Statistical analysis:
   - Performance metrics: statistical tools were used to analyze the effectiveness of the Bayesian filter, focusing on its rates of accuracy, false positives, and false negatives.
   - Comparative analysis: the Bayesian filter's performance was compared against other common spam-filtering methods.
5. Ethical considerations: All email data were anonymized to protect privacy. The study followed the university's data protection and privacy policies.
6. Limitations: During the study, the team encountered several case studies, one of which was the limitation that all the training of the presented model was carried out in real time with real data. This, in turn, had a negative effect, because different points of the programming implementation had to be constantly edited. Of course, the proposed architecture was left for 6 months without direct intervention so that the success rate of the model could be observed.

4.2. Block Diagram of Mail Traffic

The mail flow starts from the mail colored in red in Figure 2. This is the unfiltered mail that comes from the internet. Three MX servers are configured: spam filters 1 and 2 are our primary focus, and the third filter is a backup from the ISP that holds unfiltered mails in its queue if spam filters 1 and 2 are unavailable for various reasons (most notably if there are power outages in the local test datacenter (TU-Sofia Datacenter)); when they are back online, it tries to relay them the queued mails. It chooses where to relay the mails via MX preference.

When mail reaches spam filter 1 or 2, first the sender server is filtered at the POSTSCREEN stage by our MTA (RBL, SPF (Sender Policy Framework)). When the stage is passed, mails are queued by our MTA. We have chosen "Postfix", because it is very versatile in terms of which milters can be implemented, and its architecture is modular, which makes it perfect for large volumes of mail traffic. In the future, if our traffic were even greater, we could upgrade to "Exim", but for now, "Postfix v.3.5" is an adequate solution.

When the mail is successfully queued, it is internally relayed to our milter. There the mail is tested via rule sets (for example, kam.sa-channels.mcgrail.com), and a score is given based on the internal Bayes learning whether the specific mail is spam or ham. Additionally, when we had enough data, we implemented the Bayesian filter for use at the stage where the spam collection module (an additional antivirus module) works; it gives a score based and how sure the Bayes is that specific emails are spam. With these two steps completed, a final score is given, and whether or not an email is spam or ham is determined. After that, if the score indicates non-spam, the mail becomes green (as shown in Figure 2), and it is relayed to the internal dovecot server where the users interact with the mail system.

4.3. Block Diagram of Spam Filtering

Filtering in our system (Figure 3) is carried out in two stages: in the post-screen stage of the postfix communication, and in the internal milter if the first is passed successfully.

The post-screen stage is simple; it consists of an RBL list and a test for SPF alignment. We initially thought of using graylisting, but our users complained of mail delays, so we chose to not to use this mail-filtering technique.

The second stage where the real analysis of spam happens is in the internal milter.

The first internal stage in the milter is to evaluate the mail with the configured rules. We chose the default ones (for example, kam.sa-channels.mcgrail.com) because most of the other configurations use them, and they receive regular updates. Based on those rules, it checks mail headers and performs DNS checks to further evaluate them.

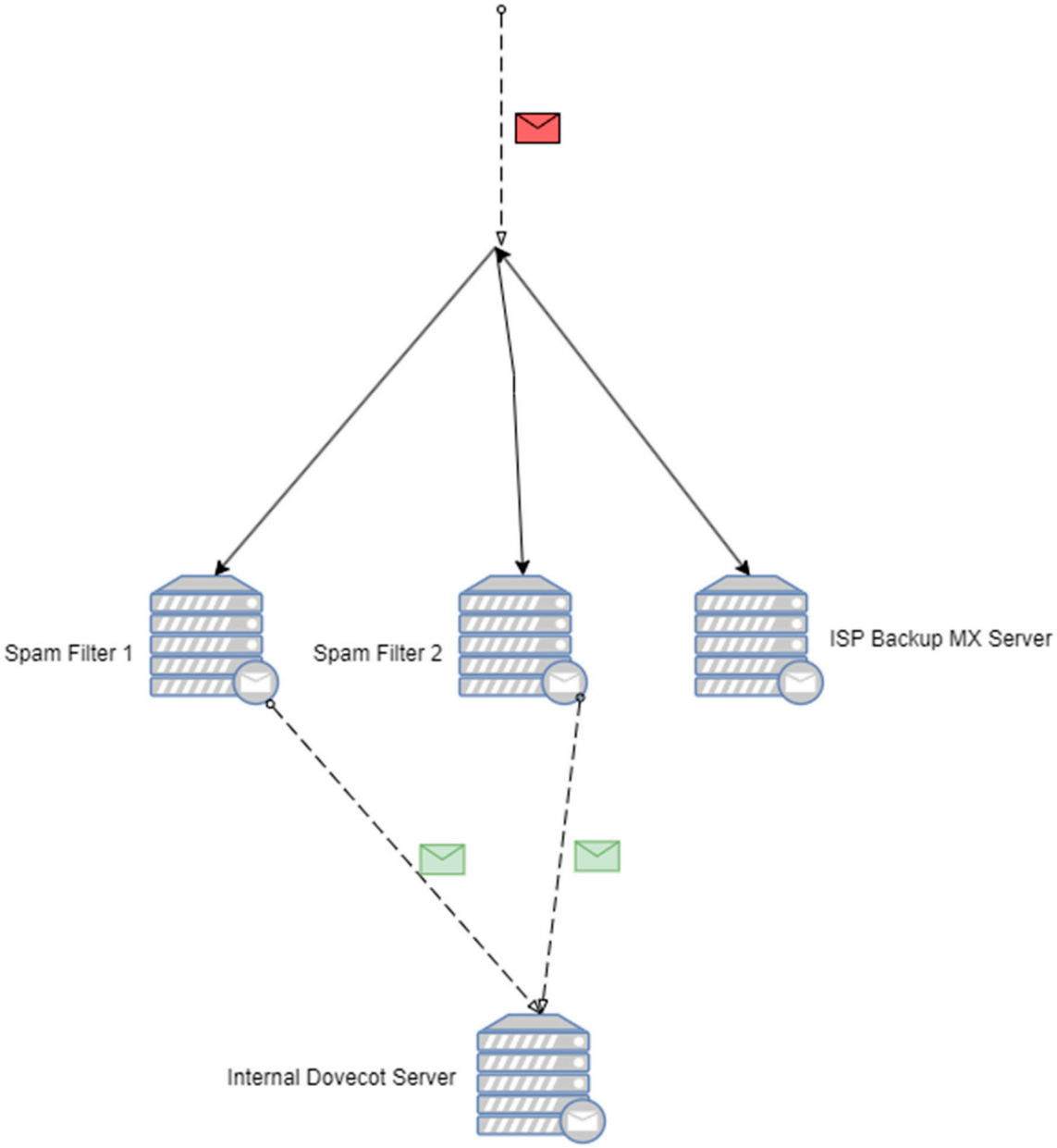


Figure 2. Block diagram of mail traffic.

We obtained enough data around the first month after we enabled the Bayes filter. The Bayes filter works in six stages:

1. Training phase: During the training phase, our module needs to be provided with a set of pre-labeled emails consisting of both spam and non-spam examples. These examples are used to build the initial Bayesian filter.
2. Tokenization: Each email in the training set is broken down into individual words or tokens. Frequently used words and punctuation are typically removed to focus on significant terms.
3. Creating the spam and non-spam token databases: Our module creates two databases: one for spam tokens, and another for non-spam tokens. For each token, the database keeps track of how frequently it appears in spam and non-spam emails.
4. Calculating token probabilities: The module calculates probabilities for each token in the databases using the following formulas:
    - P(Token | Spam): the probability of a token occurring in spam emails.

- P(Token | Non-Spam): the probability of a token occurring in non-spam emails.

These probabilities are calculated based on the token's frequency in the spam and non-spam databases.

5. Scoring new emails: When a new email arrives, the module tokenizes it and calculates a spam score based on the Bayesian filtering method. The score is determined by comparing the probabilities of the tokens found in the email with the probabilities in the spam and non-spam databases.
6. Thresholding: The module applies a threshold value to the spam score. If the score exceeds the threshold, the email is classified as spam; otherwise, it is considered non-spam.

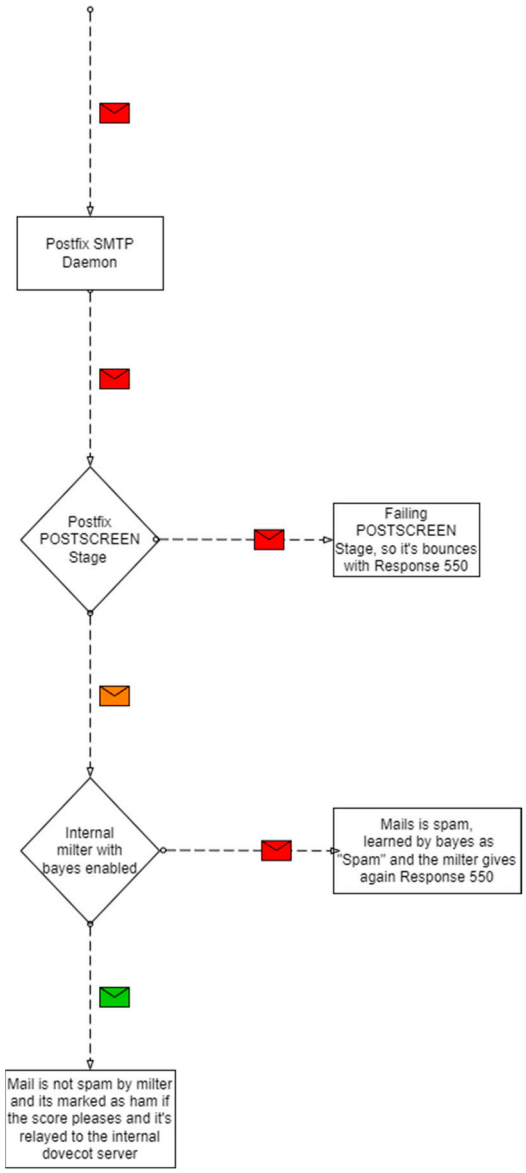
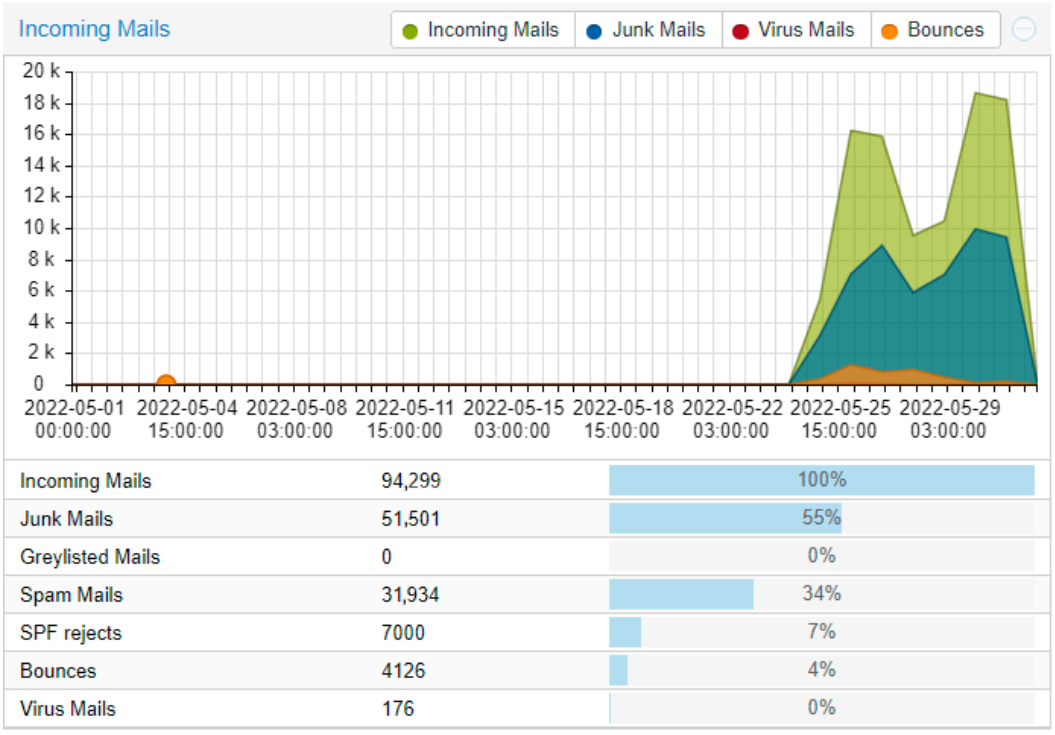


Figure 3. Block diagram of spam filtering.

5. Real-time Results and Discussion

5.1. Results

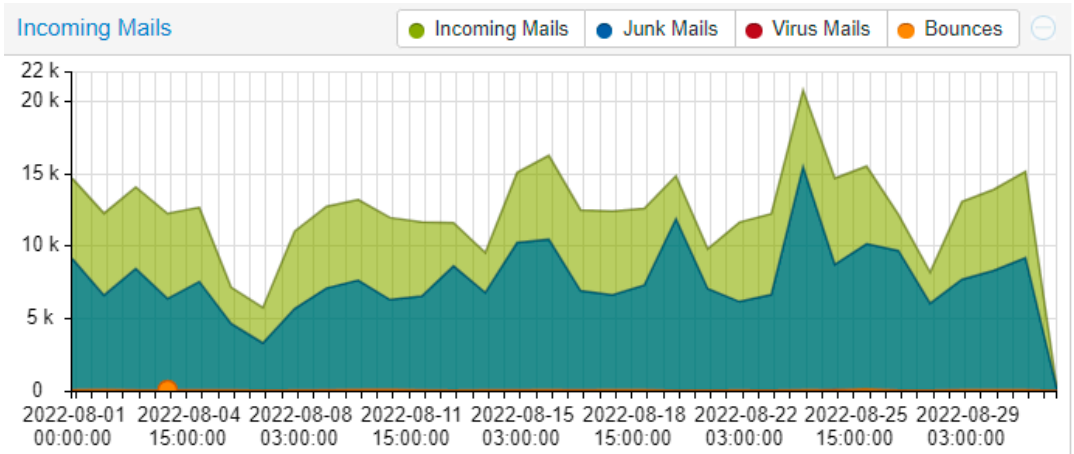
We installed and configured the servers in May 2022 (Figure 4). The data were collected over 1 year and 6 months (November 2023). We noticed that in our use case, implementation of Bayes did improve our mail filter.



Incoming Mails	94,299	100%
Junk Mails	51,501	55%
Greylisted Mails	0	0%
Spam Mails	31,934	34%
SPF rejects	7000	7%
Bounces	4126	4%
Virus Mails	176	0%

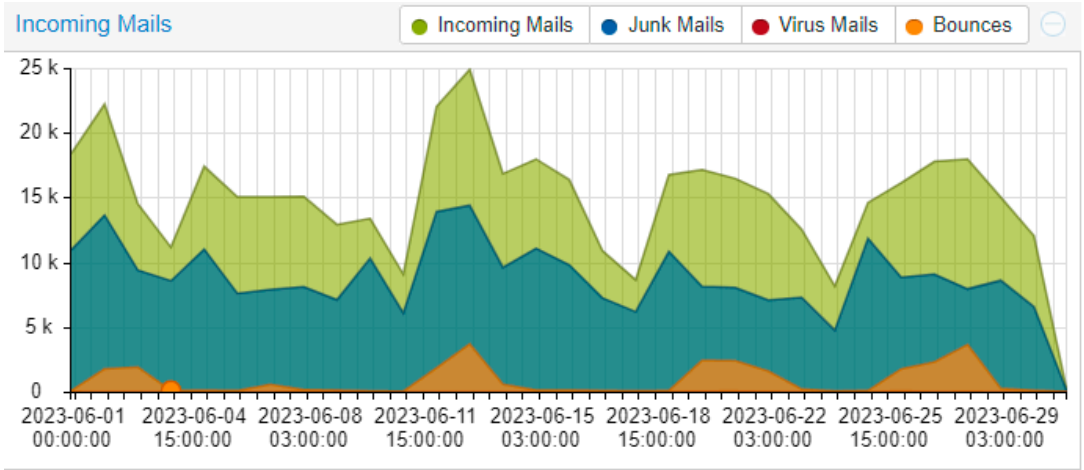
Figure 4. Data from May 2022 (starting around 25 May), with no Bayes.

The data collected are shown in Figures 5 and 6.



Incoming Mails	389,932	100%
Junk Mails	242,557	62%
Greylisted Mails	0	0%
Spam Mails	113,551	29%
SPF rejects	43,495	11%
Bounces	1680	0%
Virus Mails	314	0%

Figure 5. Data from July 2022 with Bayes after a few months of training.



Incoming Mails	461,618	100%
Junk Mails	271,912	59%
Greylisted Mails	0	0%
Spam Mails	101,118	22%
SPF rejects	10,536	2%
Bounces	26,160	6%
Virus Mails	257	0%

Figure 6. Data from June 2023 with around year of Bayes data.

Little clarification is needed for these data. One can observe that the Bayes did not have a great effect; this may be because as we said, the technology and pure Bayes are not enough anymore. Despite this, we can still see an improvement of around 5 to 10% depending on the month and the traffic. The important percentage within our results is that of "Junk Mails"; which combines Virus + Spam + Greylisted + SPF rejects + RBL rejects. A note on the Virus factor: it is not of great significance, because we used ClamAV. The results are not satisfactory, but we decided to include them in the whole package.

5.2. Discussion

Our solution cannot be fully compared to most other research works, because our measurement of the spam detection capabilities aims to report real data on real traffic, not just a preconfigured dataset (as many other works use, for example, the impressive work of Almeida and Yamaki [42]). We are aware that it may not be the best method from a purely scientific perspective, but our aim is to build a robust mail filter for real usage, and to prove that Bayes is still relevant in this field, regardless of current attitudes towards the method.

We can still make an interesting and important comparison. In our test, we used two Bayes classifiers, as mentioned; the naïve and hidden Markov methods are the default for most setups [43]. Our tests conclude interesting things when we compare both solutions found by the Bayes classifier. If we use them independently for a period of around six months, our spam-filtering method identified around 60% of the real-world email traffic. This means that 60% of this traffic is classified as spam, and further investigation proves that this score is correct.

When combing the data, we observed a massive increase in the accuracy of our classification. Our results show that in the two-month period (the period in which the filters worked together, daisy-chained), the proportion of spam within the total email traffic jumped from around 60% to 80% (Figure 7) with pure spam filtering on the main server, and identified an impressive 98.4% of spam with the backup filter, which has a lower mx record.

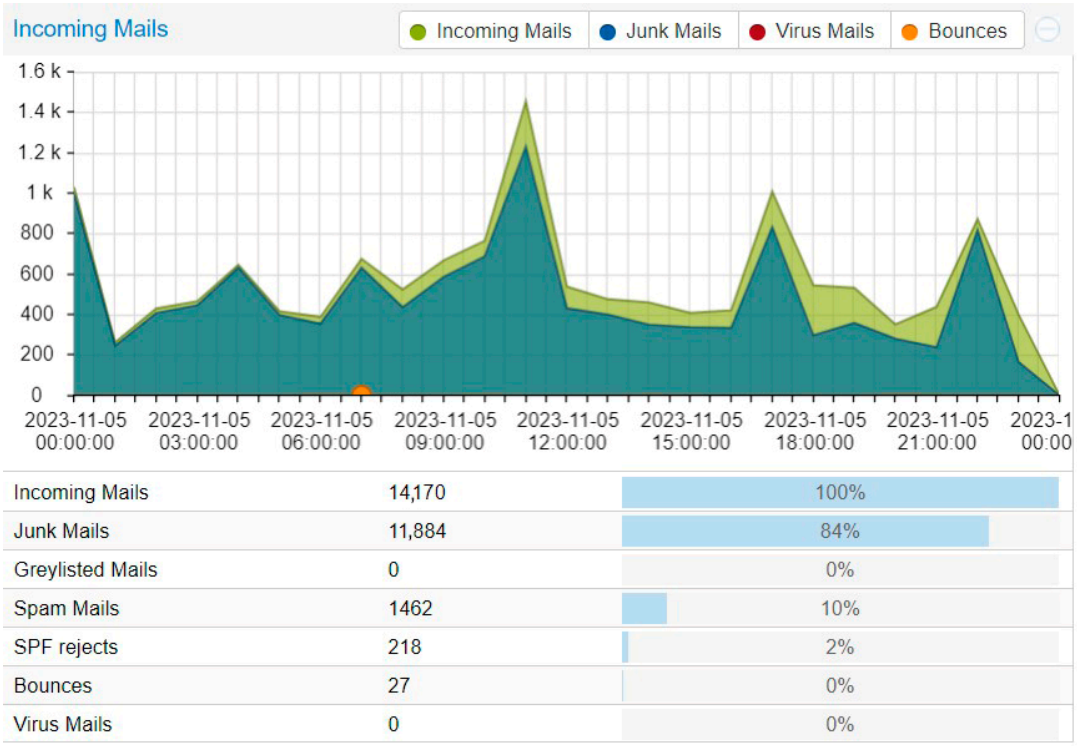


Figure 7. Data from November 2023 with the combined Bayes classifier on the main MX server.

We could've improved the naïve Bayes with support vector machines or with particle swarm optimization, because it is proven that they are efficient techniques, as shown in [19,44,45]; however, for our approach, we chose the hidden Markov method. This is because when we use it on real email traffic, it is crucial for us to undertake a sequential content analysis. Email correspondence is usually a sequence of events, and we aim to create an email filter that can detect spam patterns that can and will evolve over time; the HMM is well suited to this task.

Our results show that our proposed method, compared to the other two methods, showed a similar performance, with around a 98% success rate in spam detection on the secondary MX server (which mostly receives spam emails) that uses the same Bayes data. We also compared our data to a similar approach that combines naïve Bayes and custom intelligent text modification detection on real internet email traffic, which was developed by Huang, Jia, Ingram, and Peng [19]. Per Figure 8, we see that our approach to real email traffic had a success rate of 84% compared to the aforementioned study's 62%.

The result found with the secondary servers is likely because many spammers choose secondary servers because in some circumstances, they have a more tolerant spam filter, or no filter at all. Usually, they are backup servers provided by the hosting provider, for use only if something happens to the main server. So, most of the traffic to the secondary server is pure spam, and that is why we saw such a high score (in comparison to that of around 80% if the Bayes classifiers are used independently). In our work, the novelty lies in our combination of two different Bayesian classifiers. It is well known that naïve Bayes is very effective in email filtering because it is very accessible and computationally inexpensive, and it provides adequate basic protection against spam. Adding the hidden Markov classifiers amplifies the level of protection that the naïve Bayes classifier can provide, thanks to the advantages it offers. We used it to detect email anomalies like a directed unsolicited bulk spam attack, or to detect user anomalies; if a user account is compromised, it can block all spam emails coming from its mailbox. It is also effective against phishing emails, because it is effective in detecting phishing attempts based on an email's temporal characteristics.

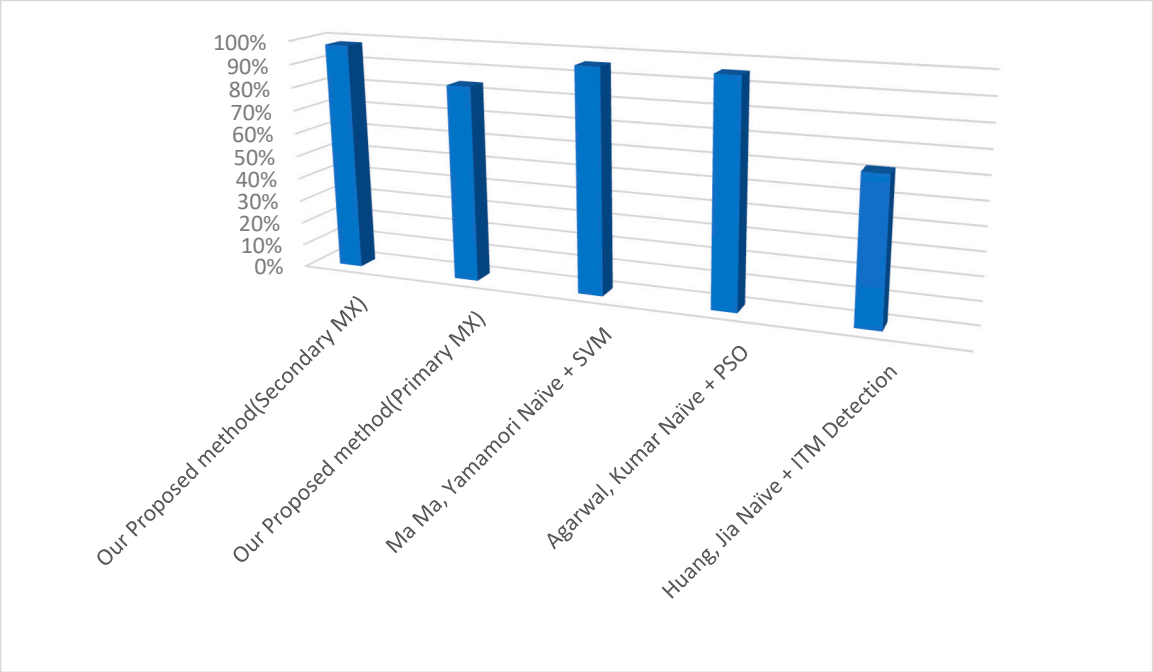


Figure 8. Comparison with similar studies.

There are advantages to our approach compared with some popular filtering techniques like rule-based filtering, content filtering or authentication methods like SPF or DKIM, but the latter is usually used with the mentioned approaches to email filtering. The advantages of Bayes compared to the other mentioned methods are as follows:

- It is not resource-intensive: This is a key advantage compared to other methods, because due to rising email traffic and "complicated" emails, processing time could become a problem for email delivery. So, the Bayes approach provides in-depth analysis of the email, because it looks for spam patterns rather than something specific in the email.
- It is an efficient learning mechanism: the Bayesian filter technique provides a learning mechanism so that the filter is constantly evolving and adjusting to the email traffic that it filters.
- It has configurable thresholds: Due to the real-world traffic moving through the system in a university network, thresholding also works on all emails, because if a mail is under a certain score for spam, it is marked as ham (a non-spam email) and passed automatically to the Bayesian filter, to be learned on a global level. This also goes for if an email has a high score; it is then considered spam and also passed to the filter as spam. The next time a similar email arrives, it will also be susceptible to being considered spam, making this method effective against zero-day attacks. Because of its efficient learning mechanism, it also provides a configurable threshold for each user. When a specific user wants a specific email, but for another, the email is considered spam, he may mark the email as "ham" (a non-spam email) and the email is processed again in the learning mechanism. This way, there are per-user preferences for email filtering that provide more accurate results according to their specific email flow.
- Zero-day attacks: thanks to the dynamic learning mechanism and the Bayes looking for spam patterns, this model is very efficient in identifying zero-day attacks.
- Scalable: scalability is easily achievable for a cluster of Bayesian filters thanks to the fact that its data can be stored in a Redis backend, easily accessible for every node.

6. Future Work

A primary focal point for improvement revolves around refining the underlying module and the methodologies governing Bayes' classification of incoming emails. Notably, a pivotal area of enhancement involves broadening language support. Presently, our module relies on a 'naïve (ASCII lowercase)' approach. However, a proposed enhancement aims to introduce comprehensive UTF-8 conversion/normalization techniques coupled with lemmatization. This upgrade seeks to embrace a more diverse linguistic spectrum, thereby augmenting the accuracy and relevance of spam classification. Additionally, the incorporation of considerably advanced checks beyond fundamental rules is envisioned to fortify the module's filtering capabilities. These upgrades are anticipated to furnish the Bayes classifier with more reliable and nuanced data, fostering enhanced accuracy and precision. As touched on in the discussion section, combing Bayes classifiers can be very effective.

In parallel to refining the Bayes classifier's methodologies, considerations extend to exploring alternative classifiers. While the current implementation centers on the 'naïve' approach, our attention is directed towards exploring the potential of hidden Markov models. This exploration seeks to leverage the strengths of different classifiers, aiming to ascertain which model yields the most promising results in our specific spam filtering context.

7. Conclusions

The results obtained from the Bayes approach exhibit positive outcomes, albeit falling short of reaching satisfactory levels. It is evident that substantial enhancements can significantly elevate the efficacy of our spam-filtering module through the implementation of various strategies.

In summary, the experimental findings affirm the relevance of Bayes graphs in the realm of spam filtering. However, it is evident that substantial enhancements are imperative. As such, our future endeavors will concentrate on refining the milter itself, incorporating new data insights, and rigorously assessing the performance of alternative classifiers. These collective efforts are poised to furnish a definitive solution that transcends current limitations, establishing a robust and adaptive spam-filtering mechanism for optimal email security.

Author Contributions: M.S., G.T., V.G. and A.T. were involved in the full process of producing this paper, including conceptualization, methodology, modeling, validation, visualization, and preparing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been accomplished with financial support from the European Regional Development Fund within the Operational Programme "Bulgarian national recovery and resilience plan", the procedure for direct provision of grants under the "Establishing of a network of research higher education institutions in Bulgaria", and under Project BG-RRP-2.004-0005 "Improving the research capacity anD quality to achieve intErnAtional recognition and reSilience of TU-Sofia (IDEAS)".

Data Availability Statement: The datasets presented in this article are not readily available, because some emails contain confidential information. The GDPR policy of the university restricts sharing this kind of information. Requests to access the datasets should be directed to the corresponding author, who will extend this question to the persons concerned.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Global Spam Volume as Percentage of Total E-Mail Traffic from 2011 to 2022. Available online: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/ (accessed on 14 September 2023).
2. Number of Sent and Received E-Mails per Day Worldwide from 2017 to 2026. Available online: https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/ (accessed on 14 September 2023).
3. 23 Email Spam Statistics to Know in 2023. Available online: https://www.mailmodo.com/guides/email-spam-statistics/ (accessed on 14 September 2023).

4. Spam Statistics. Reports and Analysis. 2023. Available online: https://www.emailtooltester.com/en/blog/spam-statistics/ (accessed on 23 September 2023).
5. MUD1. Available online: https://en.wikipedia.org/wiki/MUD1 (accessed on 29 September 2023).
6. From Meat to Menace: The History of "Spam". Available online: https://medium.com/@GeorgeDarkow/from-meat-to-menace-the-history-of-spam-c1c0bc34d61e (accessed on 1 October 2023).
7. History of Spam. Available online: https://www.ocf.berkeley.edu/~angro/BA.html (accessed on 1 October 2023).
8. Mmmm, Chopped Pork Shoulder—31 March 1993: The Term "Spam" Coined. Available online: https://thedayintech.wordpress.com/tag/joel-furr/ (accessed on 1 October 2023).
9. Dar, M.; Iqbal, F.; Latif, R.; Altaf, A.; Jamail, N.S.M. Policy-Based Spam Detection of Tweets Dataset. *Electronics* **2023**, *12*, 2662. [CrossRef]
10. Yang, Y. Research and Realization of Internet Public Opinion Analysis Based on Improved TF—IDF Algorithm. In Proceedings of the 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), Anyang, China, 13–16 October 2017; pp. 80–83. [CrossRef]
11. Bozkir, A.S.; Sahin, E.; Aydos, M.; Sezer, E.A.; Orhan, F. Spam E-Mail Classification by Utilizing N-Gram Features of Hyperlink Texts. In Proceedings of the 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 20–22 September 2017; pp. 1–5. [CrossRef]
12. Mathew, N.V.; Bai, V.R. Analyzing the Effectiveness of N-gram Technique Based Feature Set in a Naive Bayesian Spam Filter. In Proceedings of the 2016 International Conference on Emerging Technological Trends (ICETT), Kollam, India, 21–22 October 2016; pp. 1–5. [CrossRef]
13. Xu, C.; Chen, Y.; Chiew, K. An Approach to Image Spam Filtering Based on Base64 Encoding and N-Gram Feature Extraction. In Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, Arras, France, 27–29 October 2010; pp. 171–177. [CrossRef]
14. Ashour, M.; Salama, C.; El-Kharashi, M.W. Detecting Spam Tweets Using Character N-Gram Features. In Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 18–19 December 2018; pp. 190–195. [CrossRef]
15. Siagian, A.H.A.M.; Aritsugi, M. Combining Word and Character N-Grams for Detecting Deceptive Opinions. In Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, Italy, 4–8 July 2017; pp. 828–833. [CrossRef]
16. Sahın, E.; Aydos, M.; Orhan, F. Spam/ham e-mail classification using machine learning methods based on bag of words technique. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4. [CrossRef]
17. Pajila, P.B.; Sheena, B.G.; Gayathri, A.; Aswini, J.; Nalini, M. A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications. In Proceedings of the 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 20–22 September 2023; pp. 1228–1234. [CrossRef]
18. Khamdamovich, K.R.; Elshod, H. Detecting spam messages using the naive Bayes algorithm of basic machine learning. In Proceedings of the 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 3–5 November 2021; pp. 1–3. [CrossRef]
19. Peng, W.; Huang, L.; Jia, J.; Ingram, E. Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection. In Proceedings of the 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), New York, NY, USA, 1–3 August 2018; pp. 849–854. [CrossRef]
20. Ji-Hui, F.; Xu-Yao, L.; Shao-Hua, T. Research on spam message recognition algorithm based on improved naive Bayes. In Proceedings of the 2022 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Hengyang, China, 26–27 March 2022; pp. 241–244. [CrossRef]
21. Fan, J.; Yuan, F. Research on spam message recognition algorithm based on improved naive Bayes. In Proceedings of the 2022 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 9–11 December 2022; pp. 1088–1091. [CrossRef]
22. Lv, T.; Yan, P.; Yuan, H.; He, W. Experiment Research on Spam Filter Classifier Based on Naive Bayesian Algorithm. In Proceedings of the 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), Nanjing, China, 25–27 June 2021; pp. 798–801. [CrossRef]
23. Wijaya, E.; Noveliora, G.; Utami, K.D.; Rojali; Nabiilah, G.Z. Spam Detection in Short Message Service (SMS) Using Naïve Bayes, SVM, LSTM, and CNN. In Proceedings of the 2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 31 August–1 September 2023; pp. 431–436. [CrossRef]
24. Hossain, M.S.; Zubair, M.; Rahman, M.O.; Patwary, M.K.H.; Rajib, M.G.S. A Modified Naïve Bayesian-based Spam Filter using Support Vector Machine. In Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019; pp. 1–7. [CrossRef]
25. Ashraf, M.S.; Rehman, F.; Sharif, H.; Aqeel, M.; Arslan, M.; Rida, A. Spam Consumer's Reviews Detection for E-Commerce Website using Linguistic Approach in Deep Learning. In Proceedings of the 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS), Karachi, Pakistan, 14–15 December 2022; pp. 1–7. [CrossRef]

26. Zhou, Y.; Mulekar, M.S.; Nerellapalli, P. Adaptive spam filtering using dynamic feature space. In Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), Hong Kong, China, 14–16 November 2005; pp. 627–646. [CrossRef]
27. Tian, X.; Tang, D. A multi-dimensional spam filtering framework based on threat intelligence. In Proceedings of the 2019 12th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 14–15 December 2019; pp. 158–162. [CrossRef]
28. Sonbhadra, S.K.; Agarwal, S.; Syafrullah, M.; Adiyarta, K. Email classification via intention-based segmentation. In Proceedings of the 2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI), Yogyakarta, Indonesia, 1–2 October 2020; pp. 38–44. [CrossRef]
29. Priya, S.; Uthra, R.A. An Effective Concept Drift Detection Technique with Kernel Extreme Learning Machine for Email Spam Filtering. In Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 3–5 December 2020; pp. 774–779. [CrossRef]
30. Jeong, S.; Lee, K.-H. Spam Classification Based on Signed Network Analysis. *Appl. Sci.* **2020**, *10*, 8952. [CrossRef]
31. Držík, D.; Magdin, M. A Comprehensive Analysis of the Success of Classification Algorithms for the Classification of Emotional States Based on the User's Behavioral Characteristics. *IEEE Access* **2023**, *11*, 24953–24970. [CrossRef]
32. Daisy, S.J.S.; Begum, A.R. Email Spam Behavioral Sieving Technique using Hybrid Algorithm. In Proceedings of the 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 11–13 October 2023; pp. 687–693. [CrossRef]
33. Kwong, A.; Muzamal, J.H.; Khan, Z. Privacy Pro: Spam Calls Detection Using Voice Signature Analysis and Behavior-Based Filtering. In Proceedings of the 2022 17th International Conference on Emerging Technologies (ICET), Swabi, Pakistan, 29–30 November 2022; pp. 184–189. [CrossRef]
34. Hussain, N.; Mirza, H.T.; Hussain, I.; Iqbal, F.; Memon, I. Spam Review Detection Using the Linguistic and Spammer Behavioral Methods. *IEEE Access* **2020**, *8*, 53801–53816. [CrossRef]
35. Zhao, C.; Xin, Y.; Li, X.; Yang, Y.; Chen, Y. A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data. *Appl. Sci.* **2020**, *10*, 936. [CrossRef]
36. Sethi, M.; Tyagi, N.; Kalsi, P.S.; Rao, P.A. Deep Learning-based Binary Classification for Spam Detection in SMS Data: Addressing Imbalanced Data with Sampling Techniques. In Proceedings of the 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 25–26 May 2023; pp. 1–9. [CrossRef]
37. Purwitasari, D.; Zaqiyah, A.A.; Fatichah, C. Word-Embedding Model for Evaluating Text Generation of Imbalanced Spam Reviews. In Proceedings of the 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 23–25 October 2021; pp. 1–6. [CrossRef]
38. Aich, P.; Venugopalan, M.; Gupta, D. Content Based Spam Detection in Short Text Messages with Emphasis on Dealing with Imbalanced Datasets. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–5. [CrossRef]
39. Rao, S.; Verma, A.K.; Bhatia, T. Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. *Expert Syst. Appl.* **2023**, *217*, 119594. [CrossRef]
40. Bayes Server Learning Center. Available online: https://www.bayesserver.com/docs/ (accessed on 1 October 2023).
41. Bayes' Theorem. Available online: https://plato.stanford.edu/entries/bayes-theorem/ (accessed on 1 October 2023).
42. Almeida, T.A.; Yamakami, A. Occam's razor-based spam filter. *J. Internet Serv. Appl.* **2012**, *3*, 245–253. [CrossRef]
43. Ghahramani, Z. An introduction to Hidden Markov Models and Bayesian Networks. *Int. J. Pattern Recognit. Artif. Intell.* **2001**, *15*, 9–42. [CrossRef]
44. Ma, T.M.; Yamamori, K.; Thida, A. A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification. In Proceedings of the 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan, 13–16 October 2020; pp. 324–326. [CrossRef]
45. Agarwal, K.; Kumar, T. Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization. In Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 14–15 June 2018; pp. 685–690. [CrossRef]