Journal of Computing and Data Science (JOCDAS)
Vol. 2, No. 1, 2024, pp. 32~37

ANALYZING THE APLICATION OF THE NAIVE BAYES METHOD IN EMAIL SPAM FILTERING

Riski Waloyojati*[1] , Yoga Pratama Kusendi[2]
[1,2]Universitas Islam Negeri Maulana Malik Ibrahim, Jalan Gajayana 50, malang, Indonesia
[1]200605110015@student.uin-malang.ac.id, [2]200605110084@student.uin-malang.ac.id

*Corresponding Author

Abstract

The utilization of email as a communication medium has adverse effects, one of which is the inundation of unsolicited emails in the inbox, commonly known as spam. In response to the prevalence of spam in email communications, research has been conducted to develop software capable of automatically classifying spam and non-spam emails. Spam filter developers often employ the Naive Bayes algorithm due to its simplicity and ease of implementation. To enhance accuracy and expedite the computational process, several measures must be undertaken. The development of a Bayesian filter involves three stages: building a spam database, training the Bayesian filter, and filtering.Keywords: Naïve Bayes, Email, Spam.

Keywords: Naïve Bayes, Email, Spam

INTRODUCTION

According to “Email Communication” [1], emails contain data and information, which can be classified into two types: confidential and non-confidential. Non-confidential data typically receives less attention. However, confidential data demands significant attention due to its importance to the intended recipients and the ease with which it can be reproduced [2].

Email has become a fundamental aspect of internet usage. The modern world, particularly the internet, heavily relies on email as a primary means of validation. This reliance has prompted irresponsible parties to exploit email for personal or economic gain, either directly or indirectly. The prevalence of unsolicited emails, also known as spam, diminishes the efficiency of email usage for critical tasks such as accessing educational and governmental institutions, particularly within corporate environments.

Spam emails are bulk messages sent to a large number of recipients without their consent. These emails often contain advertisements, scams, or malicious software that can harm the recipient's data and systems. The ever-increasing volume of spam emails overloads inboxes, making it difficult for users to distinguish important messages from unimportant ones.

The rising volume of spam emails coincides with a decline in worker productivity. A significant portion of a worker's time is wasted deleting spam emails, negatively impacting the efficiency of companies and institutions. Whether consciously or not, workers lose substantial amounts of time daily, weekly, and even annually, as they must juggle their responsibilities with the task of deleting unwanted emails. This ongoing issue poses a question: How much damage should a corporation or organization tolerate due to the need to manage spam emails? To address this issue, various strategies have been proposed. Here, we will explore the method of spam email filtering using the Naive Bayes classification algorithm.

Several illicit methods are often employed to gain unauthorized access to confidential information. Currently, no technique claims to be an optimal solution with 0% false positives and 0% false negatives in spam detection [3]. The anti-spam systems currently in use employ various machine learning approaches for content classification. For instance, SpamAssassin uses genetic programming to construct

Bayesian classifiers for each release. Spam can include commercial pitches, pornography, viruses, and other irrelevant content delivered to thousands of email users [4].

Spam emails cause issues such as increased storage requirements and wasted time for users who must delete unwanted messages. Manually screening spam is particularly challenging when dealing with large volumes of email. Consequently, spam filter software has been developed to automatically categorize spam and non-spam (ham) communications [5]. Various approaches can be utilized as classification functions; however, some algorithms have demonstrated superior performance in spam filter construction [6]. These approaches are renowned for their high accuracy in detecting spam, achieving significant accuracy rates of up to 99.9% [7].

Email communication can be encrypted to ensure message security. Typically, sending emails via the internet involves data transmission without protecting the content of the transmitted data. Thus, when interception occurs, the intercepted data can be read directly by the interceptor [8]. To address this vulnerability, email data can be encrypted using specific encoding methods, enhancing the security of the transmitted messages [9]. Confidentiality is a service that ensures information remains accessible only to those with the appropriate authority or secret keys to decrypt the encoded information [10].

Despite the high accuracy of current spam detection algorithms, the challenge of achieving 0% false positives and false negatives remains unresolved. Additionally, the integration of encryption techniques in conjunction with spam filtering to enhance overall email security is an area that has not been thoroughly explored. This research aims to investigate the development of a more robust spam filtering mechanism that incorporates advanced encryption methods to protect confidential information during transmission while maintaining high accuracy in spam detection.

METHODS

The methodology employed in this research is the Naive Bayes methodology. Bayesian theory, named after its founder Thomas Bayes, emerged around the 1950s and is commonly used in statistical studies based on Bayes' theorem or rules. Bayes' theory specifies the probability of an occurrence (hypothesis) based on the condition of another event (evidence). Essentially, the theorem asserts that future events can be predicted if prior events have occurred.In general, Bayes' theory can be written as follows:

$$P(Ai|B) = \frac{P(B|Ai)P(Ai)}{\sum_i \quad P(B|Ai)P(Ai)P(Ai|B)} \quad 1$$

If {Ai} constitutes a partition of the event space, for each Ai in the partition. Bayes' theory statement:

$$P(A|B)P(B) = P(A_iB) = P(B|A)P(A)$$

Where P(A|B) is the combined probability of events A and B. Dividing both sides by P(B), we get:

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

The naive Bayes algorithm is used for spam filtering in emails. Spam Filter computations can also be accomplished using Naive Bayesian [11]. For example, each email is represented by a vector $X = (X_1,X_2,\ldots X_n)$, where $(X_1,X_2,\ldots X_n)$ represents the

attribute value of $X_1,X_2,\ldots X_n$. In an experiment, characteristics will correspond to words, i.e., each attribute will signal if a given word appears. Among the various qualities that will show, we can pick those attributes by calculating mutual information (MI) by denoting the category variable C:

$$MI(X;C) = \sum_{x \in \{0,1\},\, c \in \{spam,\ legitimate\}} P(X=x, C=c) \cdot \log \frac{P(X=x, C=c)}{P(X=x) \cdot P(C=c)}$$

The attributes we chose are the attributes with the biggest MI values.

Based on Bayes' theorem and Total Probability. Vector X = $(X_1,X_2,\ldots X_n)$, given from document *d*, the chance of *d* entering category c:

$$P(C=c \mid \vec{X}=\vec{x}) = \frac{P(C=c) \cdot P(\vec{X}=\vec{x} \mid C=c)}{\sum_{k \in \{spam,\ legitimate\}} P(C=k) \cdot P(\vec{X}=\vec{x} \mid C=k)}$$

In practice, the Probability P(Xi|C) cannot be properly determined due to the probable values of being greater than $\vec{X}$ which can generate data residue concerns. With the presence of a Naïve Bayesian classifier, it simplifies the assumptions further such that $X1,X2,\ldots,Xn$ are conditionally independent given the category/class c. Therefore:

$$P(C=c \mid \vec{X}=\vec{x}) = \frac{P(C=c) \cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=c)}{\sum_{k \in \{spam,\ legitimate\}} P(C=k) \cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=k)}$$

Where P(Xi|C) and P(C) can readily be approximated as the relative frequency of the trained corpus. Some studies have shown Naive Bayesian classification to be effective [12], even if the assumption of independence is generally oversimplified. Errors in filtering ham (categorizing it as spam) are safer than letting it pass (categorizing it as ham) via the filter. Suppose L -> S and S -> L signify two types of errors (error). Assume that transitioning from L->S is λ times more costly compared to transitioning from L->S. A message is categorized as spam if:

$$\frac{P(C=spam \mid \vec{X}=\vec{x})}{P(C=legitimate \mid \vec{X}=\vec{x})} > \lambda$$

The application of Bayes' theory in email filtering, whether the email falls to the spam category or not, can be stated to be quite accurate. Why? Because the features of spam will be reproduced in every client. These recurring traits create points for employing Bayes' theory as a tool in spam screening. The extensive usage of Bayesian filters in spam detection applications is because Bayesian filters have highly intimate filter levels on their objects, such as text corpus pairs, spam objects, and ham objects. This deep or intimate filtering is exhibited so that the Bayesian filter is accustomed to identifying its object first so that it may directly define what is spam or not spam. For example, if a message is broken into numerous pieces with specified characteristics, and these elements occur repeatedly in a message, then it may be claimed that the message is spam.

RESULTS AND DISCUSSION

There are numerous approaches to design a Bayesian filter; this time, the author takes a generic approach to the stages needed in building a Bayesian filter, such as:

- Building a Spam Database
- Training the Bayesian filter
- Filtering

1. Building a Spam Database

In this step, a database is constructed to recognize certain characteristics of spam, aiming for greater precision in filtering spam and minimizing errors in blocking ham. During the Building a Spam Database stage, the tasks include constructing a word probability database, building a ham database, and establishing a spam database.

The word probabilities database contains the probabilities of words or tokens, with each phrase or token assigned a probability based on how often it appears in spam. The ham database is primarily used by institutions or companies that rely on internet-based communication. The spam database is essential for classifying communications as spam.

To improve the accuracy of the Bayesian filter and enable faster identification, the spam database must contain a large number of spam samples and be continually updated using anti-spam software.

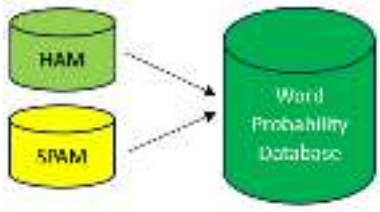

Figure 1. Spam Database Development

2. Training the Bayesian Filter

Training the Bayesian filter seeks to make it more adapted and always up-to-date in distinguishing spam or non-spam. In Bayesian filter training, there are numerous approaches that can be used: TEFT (Train Everything), TOE (Train Only Error), or TUNE (Train Until No Errors).

- TEFT (Train Everything) is utilized for every member of the text set by classifying the text and then recording its result (correct or incorrect), and then training the text into the database in the [1]proper category.
- TOE (Train Only Error) is similar to TEFT, except if the text is misclassified, it is still trained into the database in the right category.
- TUNE (Train Until No Errors) is different from TEFT and TOE since this method is utilized for every initial 500 messages by reclassifying them and then training the text if it is inaccurate. After that, the Bayesian filter is trained again if errors occur until no errors are identified.

3. Filtering

In the filtering stage, incoming emails are reviewed based on terms that meet preset criteria. The probability of an email being identified as spam or non-spam is then estimated based on these terms. An email is labeled as spam if its probability exceeds a pre-established tolerance limit; in that case, the email is prevented from accessing the client's mailbox. However, if the email does not fall into the spam category, it will be allowed into the client's mailbox.

CONCLUSION

The establishment of a spam database is crucial to ensuring that the Bayesian filter remains up-to-date in recognizing spam and non-spam emails. The advantage of the Bayesian filter lies in its accuracy and strong capability in identifying spam; however, this advantage can turn negative if error rates increase. Higher error rates occur when standards for identifying the values of words in an email are set too high. Training the Bayesian filter to become more accustomed to the terms in emails is essential for accurate identification. The better the training process, the more accurate the Bayesian filter will be in filtering spam.

ACKNOWLEDGMENT

With humility, on this occasion, I would like to convey my sincere gratitude to PT Teknologi Server Indonesia, notably to the X-Code team, and more specifically to Mr. Kurniawan and Ms. Helena for their important assistance and advice throughout my journey with the company. The excellent experience I have got from interacting with the entire X-Code team has increased my knowledge and skills in this subject tremendously. Thank you for this opportunity that has provided me with such a wonderful professional experience.

REFERENCES

[1] C. Dürscheid and C. Frehner, “Email communication,” *Pragmat. Comput. Commun.*, pp. 35–54, 2013, doi: 10.1515/9783110214468.35.
[2] I. Tafa, E. Koçi, R. Aliaj, and S. Muzhika, “Analysis of email phishing in session hijacking,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 19, no. 10, pp. 61–66, 2021.
[3] Z. Chuan, L. U. Xian-liang, Z. Xu, and H. O. U. Meng-shu, “An Improved Bayesian with Application to Anti-Spam Email,” vol. 3, no. 1, pp. 1–4, 2005.
[4] D. Berend and A. Kontorovich, “A finite sample analysis of the Naive Bayes classifier,” *J. Mach. Learn. Res.*, vol. 16, no. 1141, pp. 1519–1545, 2015.
[5] C. Lopes, P. Cortez, P. Sousa, M. Rocha, and M. Rio, “Symbiotic filtering for spam email detection,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9365–9372, 2011, doi: 10.1016/j.eswa.2011.01.174.
[6] B. Sonare, G. J. Dharmale, A. Renapure, H. Khandelwal, and S. Narharshettiwar, “E-mail Spam Detection Using Machine Learning,” *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, vol. 14, no. 03, pp. 677–683, 2023, doi: 10.1109/INCET57972.2023.10170187.
[7] T. A. Almeida, J. Almeida, and A. Yamakami, “Spam filtering: How the dimensionality reduction affects the accuracy of Naive Bayes classifiers,” *J. Internet Serv. Appl.*, vol. 1, no. 3, pp. 183–200, 2011, doi: 10.1007/s13174-010-0014-7.
[8] D. Nurani, “Perancangan Aplikasi Email Menggunakan Algoritma Caesar CIPHER dan Base64,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 2, no. 3, p. 175, 2018, doi: 10.14421/jiska.2018.23-07.
[9] S. Wibowo and Suprayogi, “Aplikasi Enkripsi Email Dengan Menggunakan Metode Blowfish Berbasis J2Se,” *Techno.COM*, vol. 13, no. 2, pp. 75–83, 2014.
[10] A. Ginting, R. R. Isnanto, and I. P. Windasari, “Implementasi Algoritma Kriptografi RSA untuk Enkripsi dan Dekripsi Email,” *J. Teknol. dan Sist. Komput.*, vol. 3, no. 2, p. 253, 2015, doi: 10.14710/jtsiskom.3.2.2015.253-258.
[11] A. K. Seewald, “An evaluation of Naive Bayes variants in content-based learning for spam filtering,” *Intell. Data Anal.*, vol. 11, no. 5, pp. 497–524, 2007, doi: 10.3233/ida-2007-11505.

[12] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, “Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, 2017, doi: 10.1088/1757-899X/226/1/012091.