Research on Spam Filters using: SVM, Naïve Bayes, and KNN

Yiting Wang

New York University Tandon School of Engineering, Brooklyn NY 11201, USA
*yw7261@nyu.edu

Abstract. Email becomes a main way for people to communicate or send information to each other. However, spammers send people unwanted and harmful information using emails. Therefore, useful email filtering needs to be used for our email. This paper shows a comprehensive review and comparative concept of various spam filtering techniques by highlighting their strengths, weaknesses, and performance. The study focuses on three prominent approaches: K-Nearest Neighbors (KNN), Naïve Bayes, and Support Vector Machines (SVM). A large dataset of emails is used to determine how well each classifier performs. The testing set and the training set are two separate portions of the dataset. The computation of a number of performance metrics will be used. The performance metrics includes the precision, accuracy, f1-score, and recall of the specific filter. The analysis's findings show each technique's advantages and disadvantages. SVM exhibits great precision and accuracy but may be susceptible to parameter tuning and feature selection. KNN achieves competitive results with a straightforward implementation but can suffer from scalability issues. Naïve Bayes, despite its simplistic assumptions, performs well too.

Keywords: Spam filtering, Spam classification, SVM, KNN, Naïve Bayes.

1 Introduction

Emails are an important way for people to obtain information and communicate with the outside world in the information age. The daily emails sent and received was 333.2 billion in 2022 [1]. By the end of 2026, volume is anticipated to reach 392.5 billion [1]. It is now an essential component of both our personal and professional lives. There were around 4 billion active email users as of 2020 [2]. The 21st century has seen a significant increase in spam emails. It is impossible to determine with certainty who had the simple insight that, no matter what the proposition, if you send out a message to millions of people, at least one of them will respond [3]. These emails are often sent for malicious purposes, such as promoting a product or service, spreading malware, or sending false information. According to a recent FBI report, spam emails cost corporate email users USD 12.5 billion in losses in 2018 [4]. To solve this problem, organizations and email service providers have some effective

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_59

spam filtering techniques. An effective spam filtering system helps to protect users' privacy, ensure the integrity of communication channels, and improve the overall security of email. Significant enhancement has been made in this field so far.

This paper aims to provide a thorough analysis of spam filtering methods. This study investigates the efficacy and applicability of each strategy by analyzing the advantages and disadvantages of various approaches. The study divides spam filtering into three methods. They are rule-based filters, content-based filters, and machine learning-based filters.

Rule-based filters utilize predefined rules and heuristics to identify spam patterns, keywords, and suspicious email characteristics. Content-based filters analyze the content of emails, leveraging features such as text analysis, and Bayesian probability to classify messages as spam or legitimate. This paper will focus on the third method which is the machine learning-based approaches. Three classifiers are analyzed which are KNN, SVM, and Naive Bayes.

Several research studies have contributed to the advancements in spam filtering techniques. Ola Amayri and Nizar Bougulia [5] demonstrated that SVM has high accuracy and precision. SVMs have outperformed other learning algorithms due to their strong theoretical foundation, high generalization, global solution, number of tuning parameters, and global solution [6]. The analysis of the study's findings sheds light on the advantages and disadvantages of each technique. SVM demonstrates excellent precision and accuracy, making it a robust choice for spam filtering. However, SVM may be susceptible to parameter tuning and feature selection, which can impact its performance.

The K-Nearest Neighbor method (KNN), On the other hand, KNN achieves competitive results with its straightforward implementation. However, scalability can be a concern for KNN, especially when dealing with large datasets. L. Firte, C. Lemnaru, and R. Potolea [7] investigate the application of KNN in spam filtering, highlighting its simplicity, scalability, and competitive performance.

Furthermore, researchers have explored the effectiveness and performance of Naive Bayes classifiers in spam detection. Despite its simplistic assumptions, performs remarkably well in practice. It exhibits high efficiency and can handle high-dimensional data effectively. This makes Naïve Bayes a favorable option for email filtering applications. The studies conducted by Aditya Gupta, Khatri Mrunal Mohan, and Sushila Shidnal [8] showcase the efficacy of Naive Bayes in handling large-scale datasets and its simplicity in implementation.

In this paper, the SVM, KNN, and Naïve Bayes will be introduced. Then, a dataset will be used to test the performance of those three classifiers. Their performance will be analyzed separately, and their merit and demerit will also be introduced. Finally, according to this particular dataset, find the data set with the best performance.

2 Methods

2.1 Support Vector Machines (SVM)

SVMs are effective tools for classifying data. When using nonlinear classifiers or a higher dimensional feature space instead of the original input space of the issue, they classify two-category points by allocating them to one of two disjoint half spaces [6]. The support vector's kernels combined linearly to form the separation function are as follows:

$$f(x) = \sum_{jz \in S} \alpha_j y_j K(z_j, z) + b \quad (1)$$

Where S is the set of support vectors, $y_i \in (-1,1)$ is the associated class labels, and z is the training patterns.

The dual formulation yields is a follows:

$$\min(0 \le \alpha_i \le C)\, W = 0.5 \sum_{i,j} \alpha_i \alpha_j Q_{i,j} - \sum_i \alpha_i + b \sum_i y_i \alpha_i \quad (2)$$

Where $Q_{ij} = y_i y_j K(z_i, z_j)$ is a symmetric positive definite kernel matrix, α_i are the corresponding coefficients, b is the offset, and, in the inseparable situation, C is a value that is utilized to penalize mistake points.

2.2 K-nearest Neighbors (KNN)

KNN is frequently used to provide predictions or classifications regarding the grouping of individual data points. Before classifying the instances using the KNN approach, the classification module resamples the input data set to the ideal size and distribution. The KNN classification approach uses an instant-based machine learning technique to categorize objects based on the nearest feature space to the data set [9]. The core idea is to identify the category of a given query based on the categories of the K data that are closest to it, rather than just the dataset. The KNN method, where K = 1, is an example of the vector method. The set of x should be donated as S_x, which S_x is defined as:

$$S_x \subseteq D \text{ s.t. } |Sx| = k \text{ and } \forall(a', b') \in D \setminus S_x$$
$$dist(a, a') \ge \max_{(a'', b'') \in S_x} dis(a, a'') \quad (3)$$

In which the furthest point in S_x is at least as far away from a as any point in D that is not in S_x. First, vector for every data should in the training set; then, centroid vector should be made for each class; after that, similarity should be calculated between each dataset vector and class vector; finally, data belongs to the class should be maximum.

2.3 Naïve Bayesian

The most constrained variation of the feature dependence spectrum is a Naive Bayesian model. The effectiveness of spam filters has been studied in relation to allowing some degree of reliance between features [10]. The foundation of naive Bayesian classifiers is a statistical idea. In this experiment, whether a term appears in the training dataset affects how well a prediction performs. In other words, each processed term is assigned a probability that it belongs to a specific category. The

term is taken into account in the probability calculation from the training dataset. The equation to calculate the probability can be written as:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{4}$$

Where P(A) represents the probability that event A occurred, P(B) represents the probability that event B occurred, P(A|B) represents the probability that event A occurred under the assumption that event B occurred, and P(B|A) represents the probability that event B occurred under the assumption that event A occurred.

First, each word in the training dataset should be verified and stored in a vector. Next, the frequency of each word in the dataset should be calculated. Finally, the probability of each word in the dataset should be calculated.

3 Results

Those three best-known classifiers have been compared to find the best classifier for spam filtering. The datasets are trained with different classifiers. The data set has been made of 701 spam emails and 4398 ham emails. ROC and AUC have been used for each data set. The performance metrics has been applied for each dataset.

3.1 Experimental Result on Training Set

The experiments were first performed on the training set. The performance of each classifier is shown on the ROC graph. The experimental results indicate that AUC of the SVM is the highest throughout the SVM, KNN, and Naïve Bayers. On the other hand, SVM is able to achieve the highest detection rate.

Fig. 1 is the ROC curves of the training set. In Fig. 1, it can tell that both SVM and KNN have the best performance for spam filtering. In Table 1, it can be seen that the f1-score of KNN is relatively lower than the f1-score of SVM.

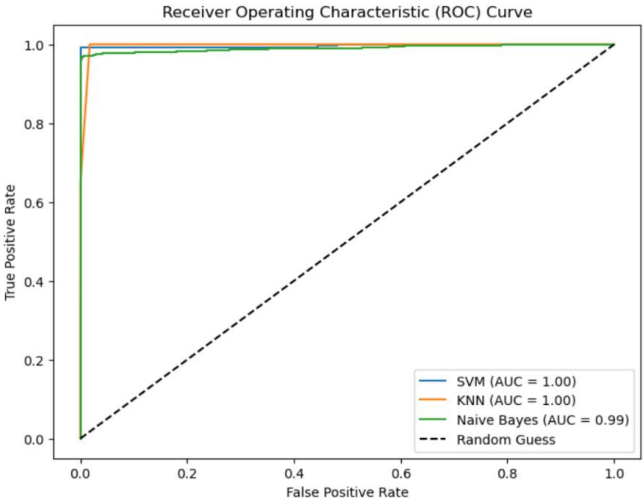

Fig. 1. ROC curves of the training set (Photo/Picture credit: Original)

Table 1. Performance metrics for training set

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.9958	1.0	0.9682	0.9838
KNN	0.9309	0.9961	0.4738	0.6421
Naïve Bayes	0.9939	0.9866	0.9663	0.9763

3.2 Experimental Result on Testing Set

The experiments were performed on the testing set. Same with the training set, the performance of each classifier is shown on the ROC graph. The experimental results shows that the AUC of the SVM is the highest throughout the SVM, KNN, and Naïve Bayes, and it was able to achieve the highest detection rate. Fig. 2 is ROC curves of the testing set. The Fig. 2 has shown that SVM is the one with the best performance. Table 2 is performance metrics for testing set. In Table 2, ii can be seen that the result of Naïve Bayes has the highest performance among all models.

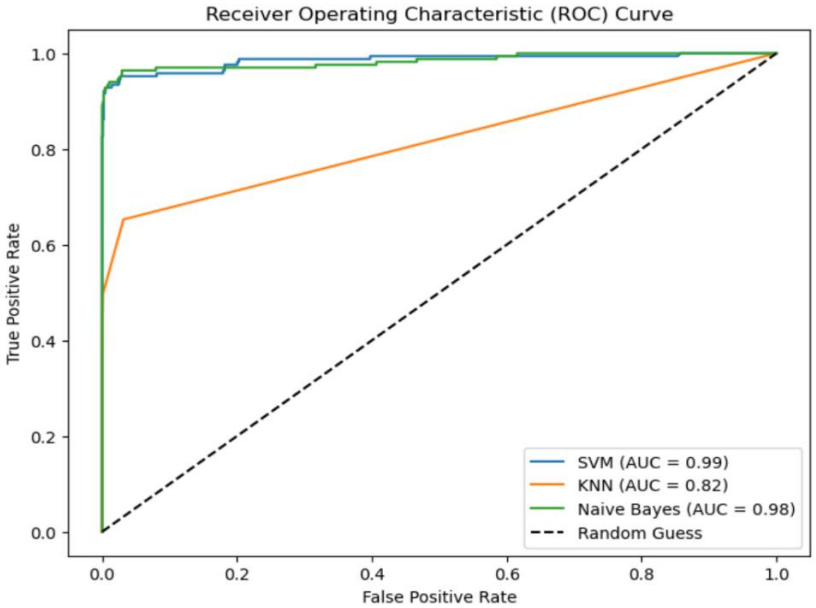

Fig. 2. ROC curves of the testing set (Photo/Picture credit: Original)

Table 2. Performance metrics for testing set

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.9765	0.9931	0.9423	0.9855
KNN	0.8980	1.0	0.3772	0.5478
Naïve Bayes	0.9823	0.9570	0.9341	0.9454

4 Conclusion

A thorough overview and analysis of spam filtering methods, with an emphasis on SVM, KNN, and Naive Bayes classifiers, have been provided in this paper. All of those three classifiers show their performance. However, the performance metrics shows that SVM performed best on the training and testing datasets. The ROC and AUC curves both showed that it had a high level of discrimination capacity. These findings help researchers and practitioners choose the best classifiers while also advancing our understanding of spam filtering methods. To further improve spam filtering systems, future research can concentrate on hybrid techniques or other machine learning algorithms.

References

1. OBERLO homepage, https://www.oberlo.com/statistics/how-many-emails-are-sent-per-day#:~:text=According%20to%20recent%20data%2C%20worldwide,in%202024%2C%20hitting%20361.6%20billion., last accessed (2023)

2. Statista Homepage, https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/, last accessed (2023)

3. Tretyakov, K.: Machine learning techniques in spam filtering. In: Data mining problem-oriented seminar, MTAT. Vol. 3. pp. 60-79. Citeseer (2004).

4. Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., Alazab, M.: A Comprehensive Survey for Intelligent Spam Email Detection. *IEEE Access* 7, 168261-168295 (2019).

5. Amayri, O., Bouguila, N.: A study of spam filtering using support vector machines. Artificial Intelligence Review 34, 73-108 (2010).

6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20, 273-297 (1995).

7. Firte, L., Lemnaru, C., Potolea, R.: Spam detection filter using KNN algorithm and resampling. In: Proceedings of the 2010 IEEE 6th international conference on intelligent computer communication and processing, pp. 27-33. IEEE (2010).

8. Gupta, A., Mohan, K. M., Shidnal, S.: Spam filter using Naïve Bayesian technique. International Journal of Computational Engineering Research (IJCER) 8(6), 26-32 (2018).

9. Gayathri, K., Marimuthu, A.: Text document pre-processing with the KNN for classification using the SVM. In: 2013 7th International Conference on Intelligent Systems and Control (ISCO), pp. 453-457. IEEE (2013).

10. Deshpande, V. P., Erbacher, R. F., Harris, C.: An evaluation of Naïve Bayesian

anti-spam filtering techniques. In: 2007 IEEE SMC Information Assurance and Security Workshop, pp. 333-340. IEEE (2007).