AN ENHANCED JUNK EMAIL SPAM DETECTION USING MACHINE LEARNING BY SUPPORT VECTOR MACHINES OVER NAIVE BAYES.

C. Gnanendhra Reddy[1], S. Magesh Kumar[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The main aim of the study is to improve junk email detection by using machine learning algorithms with Novel Support Vector Machines and Naive Bayes.
Materials and Methods: In this work, we employed both new support vector machines and Naive Bayes to investigate their effectiveness in detecting spam emails. Using the G Power program, we calculated the sample size to be 10 in group, with pertest power of 2, The threshold of 50 & confidence interval of 95%. The results showed that the Novel Support Vector Machine outperformed the Naive Bayes method in detecting email spam, with an average accuracy of 93.52% compared to 82.35% for Naive Bayes. Significance value of a $p = 0.027$ indicates a significant difference between the two groups ($p < 0.05$).
Conclusion: In the conclusion, our findings suggest that Novel SVM Machine approach is more effective than Naive Bayes method in detecting spam emails.

Keywords: Spam Filtering, Spam, Machine Learning, Novel Support Vector Machine, Naive Bayes, Email Spam Detection.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and technical Science, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

1. Introduction

In this work, we address the problem of junk email spam detection using machine learning with Novel Support Vector Machines via Naive Bayes. With the low cost per sender and the ability to send millions of advertisements, email has become a popular medium for sending unsolicited emails, also known as junk email or spam. This poses a significant threat to society and the internet. The prevalence of spam emails has made it difficult for users to read and filter out spam text from their incoming emails. Despite this, email remains an important information exchange channel. This work applies Bayes' theorem using a simple Bayesian classifier to analyse the problem of spam detection. After reviewing the theory and methods of naive Bayesian classifiers, we provide two examples. The final section discusses machine learning techniques. Spam emails have become a major problem, causing huge financial losses. Spammers often modify the behaviour of spam emails to evade detection by spam filters, making it difficult for classifiers to accurately identify them. (Sahni 2021; Wei 2018; Peng and Chan 2013).

Email spam consumes bandwidth, wastes money on dial-up customers, and exposes children to inappropriate content. Many strategies have been developed in recent years to prevent email spam (Sahni 2021). An application that detects email spam is essential to any business. Not only does it keep spam out of your inbox, but it improves the quality of your business email by ensuring it works properly and is only used for its intended purpose. Spam often contains many useful terms that affect the perception of spam filters. The purpose of a hami word attack is to evade spam detection. Interrupting in learning process sees the attack as a causal attack. For example, the Focus attack removes spam filter training sets (Sahni 2021; Peng). Researchers are constantly improving the techniques and algorithms for removing this email spam. Email spam filtering uses text mining techniques to classify available email body content into spam and non-spam communications (Agarwal and Kumar 2018).Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

Unsolicited mass email, also known as "email spam," is a form of electronic spam in which the recipient's personal identification and context are not considered relevant because they apply equally to many other potential recipients. An electronic communication is considered spam if the recipient cannot be shown to have given intentional, explicit, and revocable permission for its transmission. The proliferation of spam emails has become a major issue for internet users, as it can be difficult to differentiate between important and spam communications. In addition, spam emails can decrease an organization's efficiency and productivity. Previous research has used various methods to address this problem, including the Naive Bayes algorithm. However, the present system's implementation of the Naive Bayes algorithm has been found to have poor accuracy in identifying the research gap. (Yang et al. 2015; Hossainetal.2019).

The study tries to improve classification accuracy by including a support vector machine and comparing performance to naive bayes. The proposed model improves classifiers accuracy of identifying email spam.

2. Materials and Methods

This common framework research was carried out at the Saveetha School of Engineering(sse), Saveetha Institute of Health Science and Engineering's Laboratory of Soft Computing. The search requires two samples, with Group 1 employing a svm and group 2 employing the Naive Bayes approach (Conway and White 2011). Samples were acquired from the machine and repeated 10 times to reach the desired accuracy, with a power G of 80%, a threshold of 0.05%, and a confidence interval of 95%. Kaggle provided me with a dataset for spam collection.

Support Vector Machine

Support vector machines are used to classify and distinguish between different types of input information. This help vector device is widely used in master devices to make predictions. Support Vector Machine is frequently used to detect spam in email. This has an effective impact on spam detection (Baig 2021)). Therefore, this system detects unsolicited mail.

Pseudocode for Novel Support Vector Machine

```
Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.
Step4: Split the datasets for testing and training the dataset.
Step5: Apply SVM algorithm.
Step6: Predict the results.
```

Naive Bayes Algorithm

The probabilistic classifier, Naive Bayes is used to classify data. Is founded on probability theories that contain important independence presumptions. They are ultimately viewed as being naive (Priyoko and Yaqin 2019). Naive Bayes algorithm is a popular machine learning algorithm that is based upon Bayes' theorem. It is called "naive" because it assumes that the features or attributes of the data are independent of one another, which is often not the case in real-world data. Despite this assumption, the Naive Bayes algorithm has been shown to perform well in many classification tasks, including spam filtering and text classification. The algorithm works by calculating the probability of each class (e.g. spam or non-spam) given the features or attributes of the data. The class which is having highest probability is then chosen as the predicted classes. The Naive Bayes algorithm is often used as a baseline model in machine learning because of its simplicity and effectiveness.

Pseudocode for Naive Bayes Algorithm

```
Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.
Step4: Split the datasets for testing and training the dataset.
Step5: Apply the NBA algorithm.
Step 6: Predict the results.
```

Take into consideration that it testing setup includes both hardware and software factors. Hard drive, 64 bit os, Intel Core i5 5th generation 'CPU', 12GB of RAM, and an x86-based chipset are all features of the laptop. Currently written in Python and working with Windows 10, the programme. The prediction accuracy will show up when the application has finished. Process: Wi-Fi your laptop through it. Chrome's Google Collective Lookup Use Python to write the code. Launch the programme. Create a folder for the file, then upload it to the CD to save it. Use the message's ID to log in. Run the code to build a chart & indicate the accuracy.

Statistical Analysis

SPSS is one of a software tool using in statistical analysis. In the proposed system, 10 iterations were we performed for each group, and the predicted accuracy was recorded and analyzed. The t-test for independent samples was then performed to determine the significance between the two groups. The dependent variable in this case is number of words in white list, and independent variable is the number of words in black list. (Zhang, Zhu, and Yao 2004).

3. Results

Tables 1 and 2 shows the iterative accuracy values and statistical results of SVMand Naive Bayes algorithms, respectively. The results suggest that the SVM has an average accuracy of 93.52%, with a standard deviation of 1.77, while the Naive Bayes algorithm has an average accuracy of 82.35%, with a standard deviation of 4.33. The proposed support vector machine approach outperforms the naive Bayes method. The results of the independent samples t-tests for the two algorithms are shown in Table 3. The mean difference is 11.1, with a standard deviation of the error difference of 1.4. The p value of 0.027 ($p > 0.05$) indicates that the two groups are not statistically significant.

Figure 1 illustrates the correlation of the average accuracy of the Support Vector Machine and Naive Bayes algorithms in the form of a bar graph. The average accuracy of a Support Vector Machine (SVM) is 93.52%, while the average accuracy value of the Naive Bayes algorithm was 82.35%.

4. Discussion

This study, Support vector machine was found to have significantly higher accuracy in identifying junk email spam compared to Naive Bayes. The average accuracy of the support vector machine algorithm was 93.52%, while that of Naive Bayes was 82.35%. Furthermore, the results of the support vector machine were more consistent, with a lower standard deviation.

To evaluate the effectiveness of the Naive Bayes and the Novel Support Vector Machine(svm) on test emails based on different sizes of training emails, computational experiments were conducted. In this study, Multinomial Naive Bayes was used to classify spam emails with the Naive Bayes Classifier. The focus of our work is text categorization, which is a text mining method. To accomplish this, we employed a Bayesian classifier, a supervised learning technique based on machine learning. Using the knowledge obtained during the training phase, the Bayesian classifier is able to detect spam emails. Spam emails remain a pervasive problem on the internet, often containing no. of copies of the same message, advertisements, or other irrelevant content such as sexual material. Previous research has employed various filtering methods to identify these emails, including random forests, Naive Bayesian, Novel Support Vector Machines (SVMs), and neural networks. Email is a widely-used form of internet communication, benefiting lakhs of companies, organizations, and individuals on a daily basis. However, the proliferation of spam emails reduces productiviand can negatively impact network environments. Spam filtering is therefore crucial for maintaining a clean and efficient network environment. A drawback of this work is the inability to account for all the characteristic variables provided.

The probability estimate based on frequency is zero if there is no pair of occurrences of the class label and a particular attribute value. Large datasets are required to make accurate predictions about the probability of each class. Identifying junk email spam using class labels to reduce time complexity is a future area of planned study.

5. Conclusion

In this findings, the Support Vector Machine method significantly out-performed the Naive Bayes algorithm (82.35%) in detecting spam in spam with an accuracy of approximately 93.52%. The output from the svm appears to be more consistent and has a smaller standard deviation.

Declaration

Conflict of Interests

There are no conflicts of interest in this manuscript.

Authors Contribution

Author CGR is responsible for data collection, data analysis, and paper writing, while Guide SMK is responsible for design, data validation, and critical assessment of the publication.

Acknowledgement

The authors would like to express their appreciation to Saveetha School of Engineering and Saveetha Institute of Medical and Technical Sciences (previously known as Saveetha College) for providing the facilities required to effectively conduct this study

Funding

We should appreciate the following organisations for ' financial assistance in helping us to finish this study.

1. VK technologies, Chennai,
2. Saveetha University,
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

5. References

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. “Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review.” *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2022.134596.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. “Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review.” *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2022.127234.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. “A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims.” *Soft Computing*. https://doi.org/10.1007/s00500-022-06931-1.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. “Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation.” *Journal of Thermal Analysis and Calorimetry*. https://doi.org/10.1007/s10973-022-11298-4.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. “Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm.” *Computational Intelligence and Neuroscience* 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. “Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands.” *Silicon*. https://doi.org/10.1007/s12633-022-01825-1.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. “A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies.” *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102155.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan

Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." *Energy*. https://doi.org/10.1016/j.energy.2022.123806 .

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." *Chemosphere* 301 (August): 134638.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." *Chemosphere* 299 (July): 134390.

Tables and Figures

Table 1. This table contains Accuracy Values for Support Vector Machine (SVM) and Naive Bayes Algorithm (NBA)

S.NO	SVM	NBA
1	96.80	89.56
2	94.59	75.30
3	93.30	77.69
4	92.66	86.25
5	91.10	82.65
6	92.20	80.92
7	95.00	81.45
8	94.32	79.39
9	91.30	86.32
10	94.00	84.00

Table 2. Group Statistics Results SVM has a mean (93.52%), std.deviation (1.77), whereas for NBA has mean (82.35%), std.deviation (4.33).

Group Statistics					
Accuracy	**Groups**	**N**	**Mean**	**Std deviation**	**Std. Error Mean**
	SVM	10	93.5270	1.77259	0.56054
	NBA	10	82.3530	4.33079	1.36952

Table 3. The significance value p=0.027 (p<0.05) shows that two groups are statistically significant.

		Independent Samples Test								
		Levene's Test for Equality of Variances					**T-test for Equality of Means**			
		F	**Sig**	**t**	**df**	**Sig(2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence Interval of the Difference**	
									Lower	**Upper**
Accuracy	**Equal variances assumed**	5.807	0.027	7.551	18	0.00	11.174	1.497	8.065	14.282
	Equal variances not assumed			7.551	11.933	0.00	11.174	1.497	7.9478	14.400

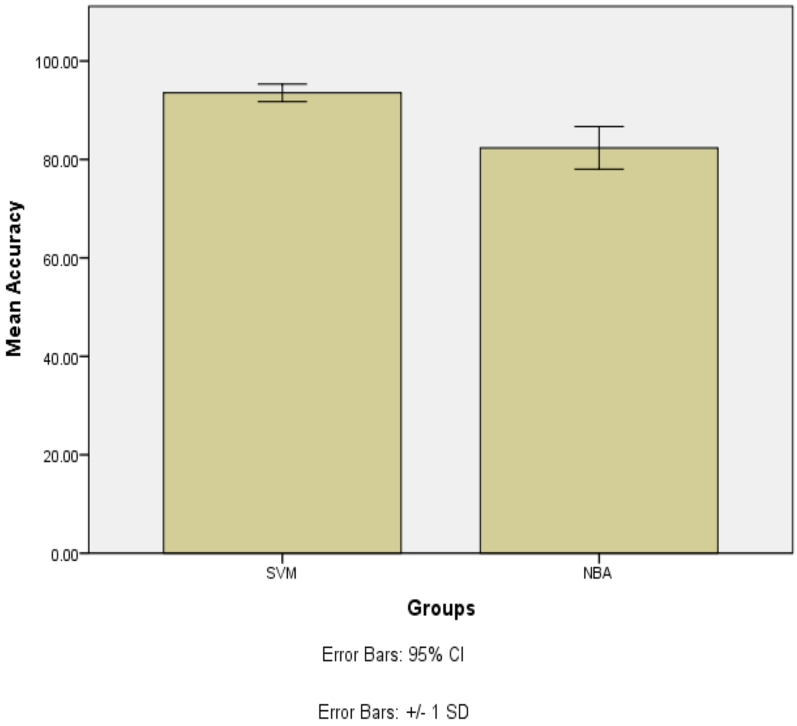

Fig. 1. Bar Graph Comparison on mean accuracy of Support vector machine (93.52%) and Naive Bayes algorithm (82.35%). X-axis is having SVM, NBA, Y-axis is having Mean Accuracy with ±1 SD.