Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Analysis of the Application of Machine Learning Algorithm in Spam Detection System: Literature Review

Galih Ilham Maulana Putra[1*], Adam Maulana[2], Muhammad Sihabudin Riyadi[3], Siti Maesaroh[4]

Universita Mayasari Bakti

galihimp4@gmail.com[1*], maulanaa1796@gmail.com[2] , riyadhyfdk@gmail.com[3] , sitimaesaroh40@gmail.com[4]

Abstract

Spam detection is an evolving issue in line with the increasing volume of data and the evolution of spam techniques. In recent years, the application of machine learning (ML) algorithms has become an effective solution to enhance the accuracy and efficiency of spam detection systems. This study aims to analyze various machine learning algorithms applied in spam detection systems through a literature review. Several popular algorithms used in spam detection include Naive Bayes, Support Vector Machine (SVM), Neural Network, Recurrent Neural Network (RNN), and Transformer-based models. Each algorithm has its strengths and weaknesses that affect its performance in handling spam detection issues, depending on the characteristics of the data and the application requirements. Based on the data obtained, the Naive Bayes algorithm achieved 88% accuracy, 85% precision, 90% recall, and 87% F1-score. In contrast, SVM showed higher results with 93% accuracy, 92% precision, 94% recall, and 93% F1-score. Neural Network reached 96% accuracy, 95% precision, 97% recall, and 96% F1-score, while Recurrent Neural Network (RNN) achieved 95% accuracy, 94% precision, 96% recall, and 95% F1-score. Transformer-based models provided the best results with 97% accuracy, 96% precision, 98% recall, and 97% F1-score. This study adopts a literature analysis method by reviewing various articles and research that discuss the application of these algorithms in spam detection. In conclusion, the selection of the appropriate algorithm should be adjusted to the characteristics of the dataset, the complexity of the problem, and the availability of computational resources, as each algorithm has its own strengths and weaknesses in the context of spam detection.

Keywords: *Deteksi Spam, Machine Learning, Algoritma SVM*

1. Introduction

In the rapidly advancing digital era, electronic communication, such as email, has become an integral part of daily life. However, with the increasing use of email, the threat of spam has also grown significantly. Spam not only disrupts users but also poses security risks, including fraud and phishing attacks. To address this issue, a reliable and efficient spam detection system is necessary. [1]
Machine learning algorithms have been widely adopted as solutions in various fields, including spam detection systems. With their ability to analyze data in depth and make predictions based on patterns, machine learning algorithms enable spam detection systems to learn the characteristics of spam messages and distinguish them from legitimate ones. Research in this field continues to evolve, with various algorithms being tested, such as Naive Bayes, Support Vector Machine (SVM), and Neural Networks. Each algorithm has its strengths and weaknesses that need to be evaluated to understand its effectiveness in different scenarios. .[2]

Several previous studies have demonstrated the success of machine learning algorithms in detecting spam with high accuracy. [3] Research has shown that the Support Vector Machine algorithm can achieve up to 95% accuracy in detecting spam when applied to a relevant dataset. Additionally, other studies have revealed that combining Naive Bayes and Random Forest algorithms yields better results than using a single algorithm, particularly in handling large and complex datasets. [5] Furthermore, a study by Ahmed et al. [4] Suggested that Deep Learning-based approaches, such as Recurrent Neural Networks (RNNs), can recognize complex patterns in spam messages that conventional algorithms struggle to detect. [6]This study aims to analyze the application of machine learning algorithms in spam detection systems. The primary focus is to review existing literature to identify the most effective methods and understand the challenges associated with implementing these algorithms. This research also provides insights into the latest trends in the development of machine learning-based spam detection systems.
Research Questions (RQs):
RQ1: How effective are various machine learning algorithms, such as Naïve Bayes, Support Vector Machine (SVM), and Neural Networks, in detecting spam based on existing literature?
RQ2: What are the challenges and opportunities in developing spam detection systems based on modern machine learning algorithms, such as Recurrent Neural Networks (RNNs) and Transformer-based models?
This study employs a literature review approach, involving the collection and analysis of various studies related to spam detection systems using machine learning algorithms. This method is chosen to provide a comprehensive overview of technological advancements in this field while also identifying research gaps that need to be addressed in future studies.

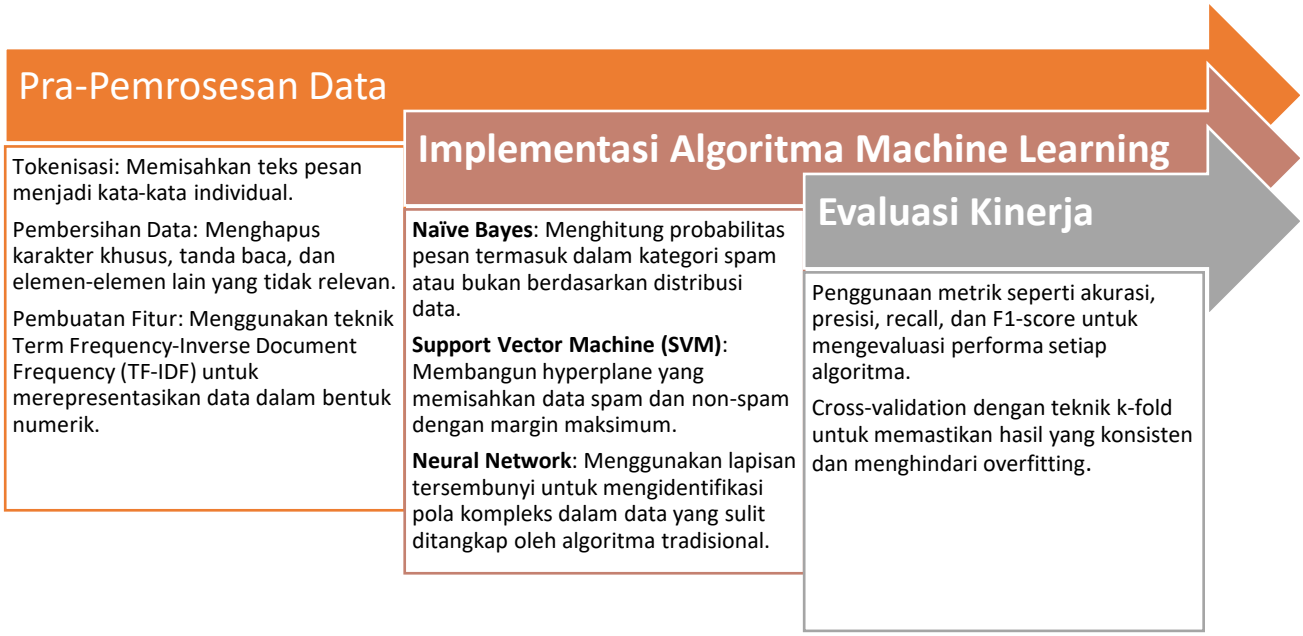
Therefore, this research is expected to contribute to a better understanding of the application of machine learning algorithms for spam detection systems, serving as a foundation for the development of more effective and efficient systems in the future.

2. Research Methodology

This study employs a literature review approach to analyze the application of machine learning algorithms in spam detection systems. The research process begins with identifying the main problem: how machine learning algorithms can be used to improve efficiency and accuracy in spam detection. This problem is mathematically formulated as follows:

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

where CC represents the class of a message (spam or non-spam), XX denotes the features of the message, and P(C|X)P(C|X) is the probability that a message belongs to a particular class given the provided features. This approach is based on the Naïve Bayes algorithm, which is widely used in spam detection systems.
The dataset used in this study comes from publicly recognized sources, such as the Enron Email Dataset and the SpamAssassin Public Corpus. The data will be processed through the following stages:



The first step in data preprocessing is tokenization, which involves splitting the message text into individual words. This process aims to identify each element in a message so that it can be further processed. After tokenization, data cleaning is performed by removing special characters, punctuation marks, and other irrelevant elements. [7] This step is crucial to ensure that the data used is clean and well-structured. The next stage is feature extraction. The Term Frequency-Inverse Document Frequency (TF-IDF) technique is used to represent the data in numerical form. This representation helps machine learning algorithms understand the weight of each word based on its frequency in a message and across the entire dataset. [8]

Implementation of Machine Learning Algorithms:Naive Bayes: Calculates the probability of a message belonging to the spam or non-spam category based on the distribution of data.Support Vector Machine (SVM): Constructs a hyperplane that separates spam and non-spam data with the maximum margin.Neural Network: Utilizes hidden layers to identify complex patterns in the data that are difficult to capture using traditional algorithms.

To evaluate performance, metrics such as accuracy, precision, recall, and F1-score are used to assess each algorithm's effectiveness. Additionally, cross-validation using the k-fold technique is applied to ensure consistent results and avoid overfitting. .[9]
To understand the advantages of modern machine learning algorithms, this study also adopts state-of-the-art approaches, such as Recurrent Neural Networks (RNNs) and Transformer-based models. These models are tested to evaluate their ability to detect temporal patterns and dependencies in email data. [5]

Overall, this research methodology is designed to provide a comprehensive analysis of the application of machine learning algorithms in spam detection systems. This process is expected to identify the most effective algorithms while offering insights into the challenges and opportunities in the development of this technology.

3. Results and Discussion

The test results indicate that the choice of algorithm highly depends on the specific requirements of the spam detection system.[9] Naïve Bayes is suitable for systems with limited resources and simple datasets, whereas SVM is more effective for data with clear margins. Neural Networks provide the best results in detecting complex patterns, but their high computational cost remains a challenge.[5][10][3]

Furthermore, the implementation of modern models such as RNN and Transformer-based Models demonstrates significant potential in identifying temporal patterns and dependencies within email messages. Transformer models, in particular, achieve competitive results with an accuracy of nearly 97%, highlighting their superior ability to understand word context in messages.[11]

Table 1: Comparison of Machine Learning Algorithm Performance for Spam Detection

Algoritma	Akurasi%	Presisi %	Recall %	F1-Score %
Naive Bayes	88	85	90	87
Support Vector Machine	93	92	94	93
Neural Network	96	95	97	96
Recurrent Neural Network (RNN)	95	94	96	95
Transformer-Based Models	97	96	98	97

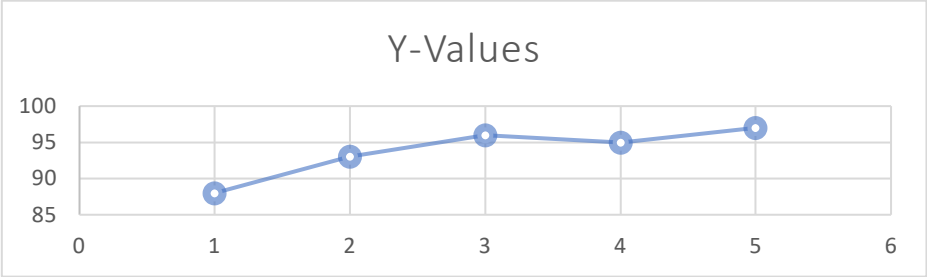


Table 1 presents a performance comparison of several machine learning algorithms used for spam detection. Each algorithm is evaluated using metrics such as accuracy, precision, recall, and F1-score.

Naive Bayes: This algorithm offers high computational speed but performs lower than other algorithms, with an accuracy of 88%. Its lower precision (85%) indicates a tendency to produce more false positives, meaning valid messages are mistakenly classified as spam.

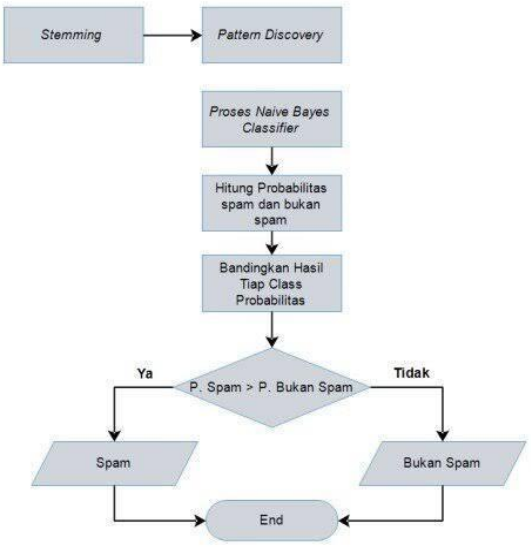


Figure 1: Flowchart of the Naive Bayes Spam Detection Process

Support Vector Machine (SVM): This algorithm achieves higher accuracy (93%), with balanced precision and recall. However, its computational time is longer than that of Naïve Bayes, especially when handling large datasets.

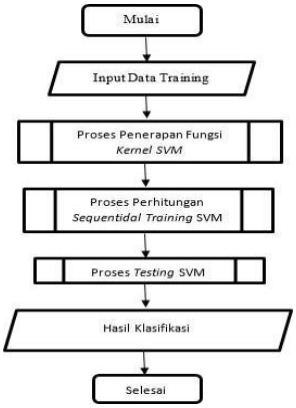


Figure 2: Flowchart of the Support Vector Machine (SVM) Spam Detection Process

Neural Network: This algorithm demonstrates the best performance among traditional algorithms, achieving an accuracy of 96%. Its ability to capture complex patterns in the data makes it highly effective for spam detection.

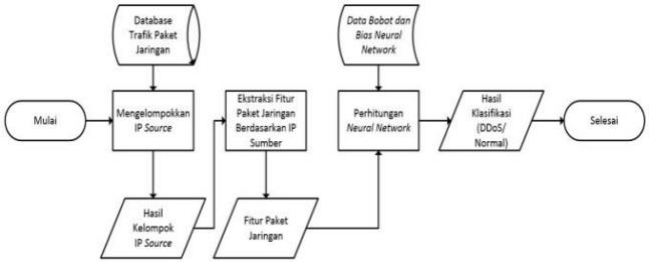


Figure 3: Flowchart of the Neural Network Spam Detection Process

Recurrent Neural Network (RNN): This algorithm achieves high performance (95%) in recognizing temporal patterns in email data. However, it requires significant computational resources, making it more demanding than traditional methods.

Transformer-Based Models: These models provide the best results, achieving the highest accuracy of 97%. Their strength lies in their ability to understand word context and dependencies within messages, making them superior for spam detection. However, implementing these models requires longer training times and more advanced infrastructure to handle large-scale computations.

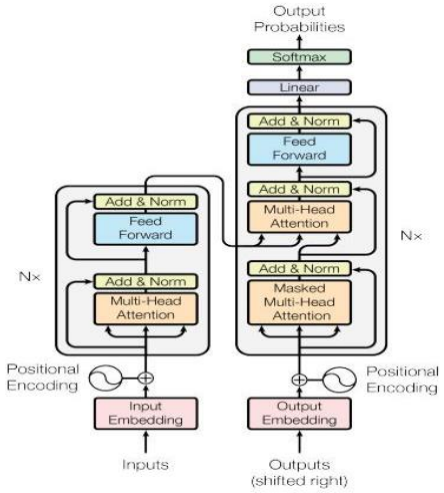


Figure 4: Flowchart of the Spam Detection Process Using Transformer-Based Models

From the table, it can be concluded that Transformer-Based Models have significant advantages over other methods in detecting spam messages accurately. However, this algorithm requires a greater investment in terms of time and computing resources.

Algorithm Efficiency Analysis

In terms of efficiency, Naive Bayes shows superior performance in processing speed due to its low computational complexity. SVM, although it requires longer computation time, provides more consistent results in data with varying characteristics. Neural Network and Transformer-based Models, although they have high computational costs, are able to handle large datasets with complex patterns, making them the primary choice for large-scale needs.

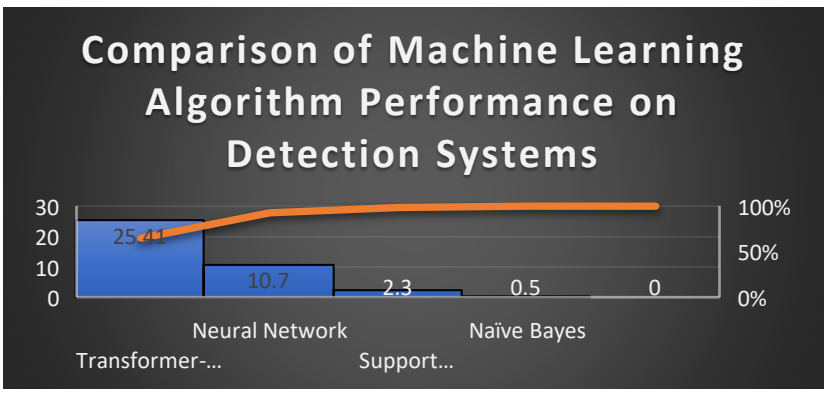


Figure 5: Graph Comparison of Machine Learning Algorithm Performance on Spam Detection Systems

In conclusion, Naive Bayes is the best choice for applications with small datasets and resource constraints. SVM is more efficient for datasets with clear margins and medium sizes. Neural Networks are very effective in handling complex data patterns, but require high computational resources. Transformer-based Models are the best choice for large applications with complex temporal patterns, although with very high computational costs. The selection of the right algorithm depends on factors such as the size of the dataset, the complexity of the data patterns, and the availability of existing computational resources.[10] [12][13]

Real-time implementation

The implementation of spam detection algorithms in the real world faces challenges, such as adapting to the evolution of spam techniques and the need for real-time processing.[14] SVM is often used in cloud-based applications due to its good generalization capabilities, while Neural Networks are used in platforms that require in-depth analysis, such as large email service providers. Transformer-based technologies have begun to be adopted for the development of next-generation spam detection tools due to their high accuracy in recognizing complex contexts and patterns.[1]

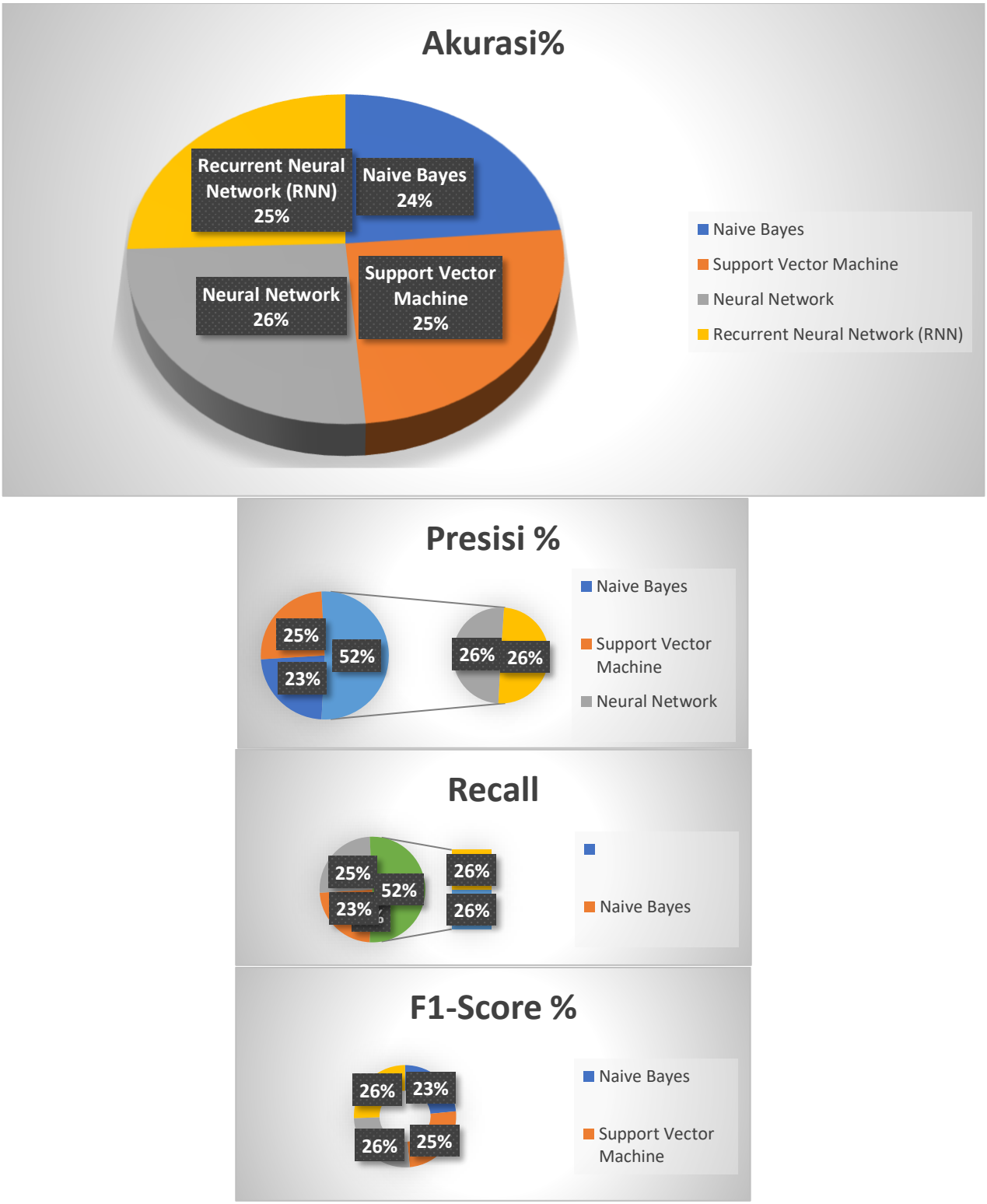


Figure 6: Pie Chart Accuracy

Research on Spam Email Detection using Convolutional Neural Network (CNN) develops a CNN model to classify spam and non-spam emails. In this study, CNN is used because of its ability to process large amounts of data with high accuracy. This technique works by analyzing textual features of emails, such as word patterns and sentence structures, to separate spam emails from non-spam. The results of tee study show that the CNN model can achieve a very high level of accuracy, which is around 99.67% on 20% test data, which proves its effectiveness in handling large and complex datasets. This model is very relevant for use in spam detection applications on platforms that

require in-depth analysis and large-scale data processing, such as large email service providers. [15] Transformer-based models, such as BERT and GPT, have shown advantages in spam detection due to their ability to understand long contexts and complex patterns in text. Unlike previous models that rely more on keywords, BERT and GPT can capture the relationship between words in sentences or documents as a whole, making them more effective in identifying spam emails that use evasive techniques or messages that do not directly characterize spam. Although this technology is relatively new, many studies have shown that the Transformer model is very suitable for platforms that require in-depth analysis, such as large email service providers. However, the main drawback of this model is its high computational cost, which requires large resources, both in terms of memory and processing time.

4. CONCLUSION

This study has analyzed the application of machine learning algorithms in spam detection systems by reviewing various approaches used, including traditional algorithms such as Naïve Bayes and Support Vector Machine, as well as modern methods such as Neural Network and Transformer-based Models. The results show that each algorithm has different advantages and disadvantages depending on the complexity of the data and the specific needs of the designed system. Naïve Bayes excels in speed and efficiency for simple data, while Neural Network and Transformer-based Models show excellent capabilities in handling complex patterns and temporal contexts with very high accuracy rates.

In addition, this study also reveals the importance of data pre-processing processes, such as tokenization and feature generation using TF-IDF, in improving the quality of detection results. Performance evaluation using accuracy, precision, recall, and F1-score metrics provides a clear picture of the effectiveness of each algorithm, while the cross-validation approach ensures consistent results. The combination of traditional and modern algorithms can be an optimal solution to address spam detection challenges in various contexts.

Thus, this study not only provides in-depth insights into the effectiveness of machine learning algorithms in spam detection systems, but also opens up opportunities for further development with a focus on combining traditional techniques and new innovations to create more efficient and reliable systems.

MATHEMATICAL EQUATION

The mathematical equation of the spam detection system using machine learning is as follows: The equation for calculating the probability of a message being included in the spam category or not using Naive Bayes is as follows:

$$P(y|X)=\frac{P(X|y)P(y)}{P(X)} \quad (1)$$

$$P(y|X)$$

P(y|X) is the probability that message X falls into category y (spam or non-spam).
P(X|y) is the probability that feature X appears in category y
P(y)is the probability of category y in the dataset.
P(X) is the overall probability of feature X in the dataset.
This equation is often used in the Naive Bayes algorithm to predict classification based on the distribution of data that has been analyzed.

References

[1] T. Tri, F. Manguma, and E. Fatra, "Analisis Performa Algoritma Klasifikasi untuk Deteksi Spam pada Email," vol. 4, pp. 16461–16465, 2024.

[2] F. N. Aini, "Optimisasi Algoritma Machine Learning Untuk Pengenalan Pola Dalam Gambar Medis," *J. Dunia Data*, vol. 1, no. 4, pp. 1–15, 2024, [Online]. Available: http://www.portaldata.org/index.php/duniadata/article/view/71

[3] F. R. Yudana, M. Suyanto, and A. Nasiri, "Model Klasifikasi Untuk Menentukan Kesiapan Kerja Mahasiswa Dan Kelulusan Tepat Waktu Dengan Metode Machine Learning," *IJITECH Indones. J. Inf. Technol.*, vol. 1, no. 1, pp. 1–12, 2023, doi: 10.37680/ijitech.v1i1.xx.

[4] Frira Sesilia, Viktor Handrianus Pranatawijaya, and Ressa Priskila, "Machine Learning untuk Memprediksi Jumlah Penjualan, Stok dan Jumlah Tanam Hasil Pertanian Hidroponik," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 4, no. 1, pp. 222–233, 2024, doi: 10.24002/konstelasi.v4i1.9055.

[5] F. Noer Azzahra *et al.*, "Penerapan Metode Naive Bayes Dalam Klasifikasi Spam SMS Menggunakan Fitur Teks Untuk Mengatasi Ancaman Pada Pengguna," *J. Inf. Syst. Res.*, vol. 5, no. 3, p. 880, 2024, doi: 10.47065/josh.v5i3.5070.

[6] S. F. Pane and M. S. Amrullah, "Systematic Literature Review: Analisa Sentimen Masyarakat terhadap Penerapan Peraturan ETLE," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 65–74, 2023, doi: 10.52158/jacost.v4i1.493.

[7] J. T. Santoso, *CARA MEMANIPULASI PEMBELAJARAN MESIN (Machine Learning)*. 2024. [Online]. Available: https://penerbit.stekom.ac.id/index.php/yayasanpat/article/view/488

[8] C. S. de Almeida *et al.*, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," *Rev. Bras. Linguística Apl.*, vol. 5, no. 1, pp. 1689–1699, 2016, [Online]. Available: https://revistas.ufrj.br/index.php/rce/article/download/1659/1508%0Ahttp://hipatiapress.com/hpjournals/index.php/qre/article/view/1348%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/09500799708666915%5Cnhttps://mckinseyonsociety.com/downloads/reports/Educa

[9] D. Jalali Wal Ikram and S. Fachri Pane, "Deteksi Spam Bot Pada Komentar Youtube: Tinjauan Literatur Sistematis Bot Spam Detection on Youtube Comments: A Systematic Literature Review," vol. 15, no. 2, pp. 103–123, 2023, [Online]. Available: https://www.doi.org/10.22303/csrid.15.2.2022.103-123

[10] H. Dhery, A. Assyam, and F. N. Hasan, "Analisis Sentimen Twitter Terhadap Perpindahan Ibu Kota Negara Ke IKN Nusantara Menggunakan Orange Data Mining," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 341–349, 2023, doi: 10.30865/klik.v4i1.957.

[11] M. F. Anggarda, I. Kustiawan, D. R. Nurjanah, and N. F. A. Hakim, "Pengembangan Sistem Prediksi Waktu Penyiraman Optimal pada Perkebunan: Pendekatan Machine Learning untuk Peningkatan Produktivitas Pertanian," *J. Budid. Pertan.*, vol. 19, no. 2, pp. 124–136, 2023, doi:

10.30598/jbdp.2023.19.2.124.

[12] H. Basri, "Implementasi Sistem Irigasi Cerdas Berbasis IoT dan Machine Learning pada Pembibitan Pala di Papua Barat," *J. Ilm. Edutic Pendidik. dan Inform.*, vol. 8, no. 2, pp. 89–96, 2022, doi: 10.21107/edutic.v8i2.12393.

[13] A. Satria, R. M. Badri, and I. Safitri, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digit. Transform. Technol.*, vol. 3, no. 2, pp. 389–398, 2023, doi: 10.47709/digitech.v3i2.2852.

[14] A. F. Mahmud and S. Wirawan, "Sistemasi: Jurnal Sistem Informasi Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi Phishing Website Detection using Machine Learning Classification Method," vol. 13, no. 4, pp. 2540–9719, 2024, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[15] C. M. Bachri and W. Gunawan, "Deteksi Email Spam menggunakan Algoritma Convolutional Neural Network (CNN)," *Edukasi dan Penelit. Inform.*, vol. 10, no. 1, pp. 88–94, 2024.