Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Comparative Performance Analysis of BERT and RoBERTa for Email Spam Classification

Purwadi [1]*, Hafizh Dzaky Ahya Gemilang[2]

[1]*Master of Computer Science Program, Faculty of Engineering and Informatics, Universitas Amikom Purwokerto*
[2]*Informatics Study Program, Faculty of Engineering and Informatics, Universitas Amikom Purwokerto*
purwadi@amikompurwokerto.ac.id[1*], hdag304001@gmail.com[2]

Abstract

The rapid advancement of information technology has increased the use of email as a primary digital communication medium, while also contributing to the growing volume of spam emails that threaten productivity and information security through phishing and malware. An accurate and adaptive email spam classification system is therefore required. This study aims to analyze and compare the performance of BERT and RoBERTa transformer models for email spam classification. An experimental research approach was employed using an email dataset consisting of spam and non-spam (ham) classes. The research process includes data collection, text preprocessing, model fine-tuning, and performance evaluation using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results show that both BERT and RoBERTa achieve high classification performance. However, RoBERTa demonstrates superior results, particularly in terms of spam recall and overall accuracy, indicating a stronger ability to detect spam emails. This advantage is attributed to RoBERTa's optimized pre-training strategy, which improves contextual semantic understanding of email content. In conclusion, RoBERTa is more effective than BERT for email spam classification and can serve as a reliable model for developing robust transformer-based spam detection systems.

Keywords: *BERT; Email Spam Classification; RoBERTa; Text Classification; Transformer Models*

1. Introduction

Email, or electronic mail, is a digital communication medium that allows users to send messages and documents through internet networks. This medium is widely used for written communication between individuals who possess email addresses. In general, an email consists of several main components, including the sender's address, the recipient's address, the subject, and the message body. In addition to text-based messages, email also supports the transmission of various digital content such as images, videos, and documents in multiple formats, making it a flexible and widely used communication tool [1].

The utilization of email in daily communication has increased productivity and the effectiveness of information exchange due to its ease of access and documentation capabilities. In recent years, the popularity of email has continued to grow alongside the widespread use of devices such as computers, laptops, and smartphones. As a communication medium, email combines flexibility with rapid information transmission, making it one of the primary tools in modern digital communication [2].

However, the increasing use of email has also been accompanied by various security issues, one of which is email spam. Email spam refers to messages sent without the recipient's consent and are generally distributed in bulk. Often referred to as unsolicited commercial email or unsolicited bulk email, spam disrupts daily communication activities. In 2010, nearly 90% of circulating emails were estimated to be spam, leading to significant resource consumption. The negative impacts of spam include wasted network bandwidth and storage space, as well as the dissemination of inappropriate content such as gambling and pornographic advertisements. Furthermore, spam emails may carry serious threats, including malware and phishing attacks, emphasizing the need for reliable spam detection systems to ensure secure and trustworthy email communication [3].

In the context of cybersecurity, spam remains a significant threat as attackers continuously develop increasingly sophisticated strategies to deceive users and steal sensitive information. Although many methods have been proposed for spam classification, existing approaches often struggle to balance accuracy, efficiency, and adaptability. These challenges are further intensified by the dynamic nature of spam, the massive volume of emails, and data imbalance in training datasets. In addition, conventional methods generally require long training times and are sensitive to noise, which highlights the need for more comprehensive, efficient, and scalable spam detection solutions [4].

Various approaches have been developed for email spam detection, ranging from conventional techniques to the use of large language models (LLMs). Recent studies indicate that models such as GPT-4 and Mixtral of Experts offer new opportunities through promising few-

shot learning capabilities. Meanwhile, deep learning–based approaches, particularly transformer models such as BERT and RoBERTa and their variants, have achieved significant advances in linguistic context understanding. These models are capable of capturing complex semantic meanings and effectively distinguishing between spam, phishing, and ham (legitimate emails). Moreover, the transfer learning capability of transformer models enables better adaptation to diverse types of spam across different languages and contexts, thereby improving the accuracy and flexibility of modern spam detection systems [5].

2. Literature Review

2.1. Spam

Spam refers to the practice of sending electronic messages that are not desired by the recipient and are transmitted without legitimate consent. This practice may violate privacy and legal regulations because it exploits personal data without proper authorization, potentially resulting in various forms of harm. According to the Truecaller Insights Report 2020, Indonesia was recorded as the country with the highest number of spam messages in Asia in 2020. The report also indicates that spam in Indonesia is dominated by financial services, accounting for 52% of total spam, followed by insurance at 25%, mobile operators at 11%, scams at 9%, and debt collection at 3% [6].

2.2. Email Spam

Email spam is a type of electronic message sent to email account owners without prior request or consent. Email spam appears in various forms, including advertisements, Nigerian spam, and phishing messages. Generally, spam emails are distributed in very large volumes, which can cause several issues such as excessive mailbox storage usage and increased difficulty in managing and maintaining email systems [7].

2.3. Text Mining

Text mining is a subfield of data mining that focuses on analyzing and processing text-based data. Textual data are typically obtained from various document sources, with the goal of identifying key words or terms that reflect the content of the documents and enable the analysis of relationships or similarities among them. The main objective of text mining is to extract useful knowledge, identify hidden patterns, and generate analyzable information from diverse text formats, such as documents, articles, tweets, short messages (SMS), emails, and other textual sources [8].

2.4. Natural Language Processing

According to Tarigan [9], Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on enabling computers to understand, process, and generate human language. Through the application of NLP, text obtained from various sources can be effectively processed to identify patterns, uncover opinions, and analyze sentiment contained within the text [10].

2.5. Transformer

Transformer is a deep learning model designed to process sequential data, particularly text, by utilizing a multi-head attention mechanism. This mechanism enables the model to capture and learn relationships among words within a sentence more effectively, without relying on sequential processing as in previous models. With this capability, the Transformer can understand textual context more comprehensively, including long-range dependencies between words, thereby producing more accurate contextual representations for various natural language processing tasks [11].

2.6. BERT

BERT is a natural language processing model based on the Transformer architecture that utilizes an encoder as its main component. The model employs a self-attention mechanism to learn relationships between words in a text, including long-range dependencies that are difficult for sequential models to capture. Unlike unidirectional approaches such as RNNs and LSTMs, BERT is bidirectional, as it processes contextual information from both left and right directions simultaneously, resulting in more contextualized word representations. Each token is analyzed by considering its semantic relationship with the entire sentence. BERT is first pre-trained on large and diverse text corpora, enabling the model to acquire broad linguistic knowledge. Subsequently, through fine-tuning on task-specific datasets, the model can achieve optimal performance even when trained on relatively limited data. In addition, the use of subword-based tokenization allows BERT to effectively handle rare or unknown words. Numerous studies have demonstrated that BERT provides competitive and stable performance across various NLP tasks, including text classification, information retrieval, and sentiment analysis [12].

2.7. RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is an extension of the BERT model that focuses on optimizing the training process. The model introduces improvements by leveraging larger training datasets, adjusting the number of optimization steps, and applying more carefully tuned hyperparameter settings. Through these enhancements, RoBERTa is able to perform pre-training in a deeper and more effective manner. As a model based on the Transformer architecture, RoBERTa demonstrates strong capability in understanding global sentence context and has achieved excellent performance across a wide range of natural language processing (NLP) tasks [13].

3. Research Method

3.1. Method

This study employs an experimental research method. Experimental research is an approach conducted through systematic testing or experimentation. This method belongs to quantitative research and aims to examine the effect of independent variables or treatments on dependent variables as outcomes, while controlling specific conditions to prevent the influence of variables outside the scope of the study [14].

3.2. Research Stages

In conducting a research study, a researcher is required to develop an appropriate and systematic plan. Research design refers to a structured, detailed, and clearly defined framework that outlines the process of data analysis and interpretation. The research plan describes the stages to be carried out as well as the problems to be examined throughout the research process. With proper planning, researchers are able to determine appropriate actions and make well-informed decisions when encountering challenges during the research process [15].

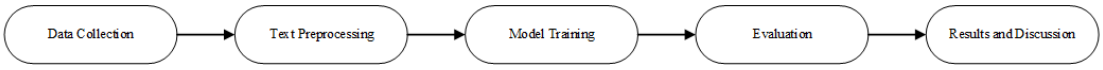

Fig. 1. Research Stages Diagram

The following is a description of the research stages applied in this study:

1. Data Collection: The selected data consist of an email spam dataset obtained from the Kaggle platform, which contains labeled email messages classified into spam and non-spam (ham) categories.
2. Preprocessing: The collected data are preprocessed by performing text cleaning, case normalization, tokenization using the built-in tokenizers of BERT and RoBERTa, and data balancing to address class imbalance between spam and non-spam emails.
3. Model Training: The pre-trained BERT and RoBERTa models are fine-tuned using a supervised learning approach to adapt the models to the characteristics of the email spam dataset.
4. Evaluation: The trained models are evaluated using accuracy, precision, recall, F1-score, and confusion matrix to measure classification performance.
5. Results and Discussion: The evaluation results are analyzed and discussed to compare the performance of BERT and RoBERTa in detecting spam and non-spam emails.

4. Results and Discussion

4.1. Dataset

The dataset used in this study was obtained from the Kaggle platform under the title *Spam Email Classification*. The dataset consists of 5,158 unique email records that have been labeled into two classes, namely spam and non-spam (ham). Each record contains the textual content of an email, which is used as input for the email spam classification process. The dataset represents real-world email conditions, where legitimate emails are more dominant than spam emails.

Table 1. Dataset Distribution

Class	Number of Samples
Ham	4,827
Spam	331
Total	**5,158**

To address the class imbalance problem in the dataset, a data balancing technique was applied during the training stage. This study employed an oversampling approach, in which samples from the minority class (spam) were duplicated to achieve a more balanced class distribution. The oversampling process was performed only on the training data to avoid data leakage and to preserve the original data distribution for evaluation purposes. By balancing the training data, the models are expected to learn representative patterns from both classes more effectively and reduce bias toward the majority class.

4.2. Email Dataset

The following table presents several examples of email data used in this study. Each record consists of email text and a class label indicating whether the email belongs to the spam or non-spam (ham) category. These examples are provided to illustrate the characteristics of the dataset utilized in the email classification process.

Table 2. Sample Email Data and Class Labels

No	Email Message	Label

1	Go until jurong point, crazy.. Available only in bugis n great world la ...	Ham
2	Ok lar... Joking wif u oni...	Ham
3	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May ...	Spam
4	U dun say so early hor... U c already then say...	Ham
5	Nah I don't think he goes to usf, he lives around here though	Ham
6	FreeMsg Hey there darling it's been 3 week's now and no word back! ...	Spam
7	Even my brother is not like to speak with me. They treat me like aids ...	Ham
8	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu ...	Ham
9	WINNER!! As a valued network customer you have been selected to ...	Spam
10	Had your mobile 11 months or more? U R entitled to Update to the ...	Spam

4.3. Classification Results Using BERT

The BERT model was applied to classify email spam through a fine-tuning process using the prepared dataset. Performance evaluation was conducted on the test data using accuracy, precision, recall, and F1-score metrics.

Table 3. Classification Report of the BERT Model

Class	Precision	Recall	F1-score	Support
Ham	0.9938	0.9990	0.9964	966
Spam	0.9931	0.9597	0.9761	149
Accuracy			**0.9937**	1115
Macro Avg	0.9934	0.9793	0.9862	1115
Weighted Avg	0.9937	0.9937	0.9937	1115

The confusion matrix is presented to provide a detailed illustration of the classification results produced by the BERT model, showing the distribution of correctly and incorrectly classified spam and non-spam emails.

Table 4. Confusion Matrix of the BERT Model

	Predicted Ham	Predicted Spam
Actual Ham	965	1
Actual Spam	6	143

Based on the evaluation results, the BERT model demonstrates very strong performance with an accuracy of 99.37%. The high precision value for the spam class indicates that almost all emails predicted as spam are indeed spam. However, the recall value for the spam class is slightly lower, suggesting that some spam emails were not successfully detected by the model. This observation is supported by the confusion matrix, which shows that 6 spam emails were misclassified as non-spam, while misclassification of non-spam emails was minimal, with only 1 instance incorrectly classified.

4.4. Classification Results Using RoBERTa

The RoBERTa model was evaluated using the same dataset and evaluation scheme as the BERT model to ensure a fair and consistent performance comparison.

Table 5. Classification Results of the RoBERTa Model

Class	Precision	Recall	F1-score	Support
Ham	0.9979	0.9969	0.9974	966
Spam	0.9800	0.9866	0.9833	149
Accuracy			**0.9955**	1115
Macro Avg	0.9890	0.9917	0.9903	1115
Weighted Avg	0.9955	0.9955	0.9955	1115

The confusion matrix is used to further analyze the classification performance of the RoBERTa model by examining the number of correct and incorrect predictions for each email class.

Table 6. Confusion Matrix of the RoBERTa Model

	Predicted Ham	Predicted Spam
Actual Ham	963	3
Actual Spam	2	147

Based on the evaluation results, the RoBERTa model achieved an accuracy of 99.55%, which is slightly higher than that of the BERT model. The higher recall value for the spam class indicates that RoBERTa is more effective in detecting spam emails. According to the confusion matrix, only 2 spam emails were misclassified as non-spam, demonstrating improved spam detection capability. However, there was a small increase in misclassification of non-spam emails as spam, with 3 non-spam emails incorrectly classified, reflecting a minor trade-off between recall and precision for the spam class.

5. Conclusion

Based on the results of the research conducted on the comparison of BERT and RoBERTa models for email spam classification, it can be concluded that the RoBERTa model demonstrates more optimal performance than the BERT model. This conclusion is supported by the performance evaluation results, which show that RoBERTa achieved a higher accuracy of 99.55% compared to 99.37% obtained by BERT. In addition, RoBERTa exhibits a higher recall value for the spam class, indicating its superior ability to detect spam emails more effectively. The confusion matrix analysis further confirms this result, as RoBERTa misclassified only 2 spam emails as non-spam, whereas BERT misclassified 6 spam emails. Although RoBERTa shows a slight increase in the misclassification of non-spam emails as spam, this trade-off is acceptable considering its improved spam detection capability. Overall, these results demonstrate that RoBERTa is more effective for email spam classification, particularly in terms of detecting spam emails. This finding is expected to serve as a reference for the development of more accurate and reliable email spam detection systems based on transformer models, thereby enhancing email security and user protection against spam-related threats.

References

[1] M. B. M. Amin *et al.*, "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 6, pp. 1291–1302, Dec. 2024, doi: 10.25126/jtiik.2024118121.

[2] Y. R. Hutagaol and Y. Arifin, "KLASIFIKASI SPAM EMAIL BERBASIS SEMANTIK MENGGUNAKAN METODE BERT SEMANTIC-BASED EMAIL SPAM CLASSIFICATION USING BERT METHOD," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 7, no. 5, pp. 1823–1836, 2024, doi: https://doi.org/10.31539/intecoms.v7i5.12515.

[3] F. Y. Arini *et al.*, "Optimasi algoritma deteksi spam email dengan BERT-MI dan jaringan dense," *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 6, no. 2, pp. 319–328, 2025, doi: 10.37859/coscitech.v6i2.9460.

[4] M. Rustam, A. Brotokuncoro, and R. Roestam, "Deteksi Email Spam dengan Continuous Bag-Of-Words dan Random Forest," *Ranah Research Journal (R2J)*, vol. 6, no. 4, pp. 758–765, 2024, doi: 10.38035/rrj.v6i4.

[5] D. Tejo Arum and A. Ichsan Pradana, "IMPLEMENTASI BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) UNTUK KLASIFIKASI SPAM PADA EMAIL," *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 9, no. 2, pp. 2491–2496, 2025, doi: https://doi.org/10.36040/jati.v9i2.13114.

[6] M. A. Sofyan, N. Rahaningsih, and R. D. Dana, "DETEKSI SMS SPAM BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE," *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 8, no. 3, pp. 3071–3079, 2024, doi: https://doi.org/10.36040/jati.v8i3.9532.

[7] I. Fauzi *et al.*, "COMPARATIVE STUDY OF SPAM EMAIL CLASSFICATION DECISION TREE BETWEEN USING CART AND J48," *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 9, no. 3, pp. 4032–4036, 2025, doi: https://doi.org/10.36040/jati.v9i3.13533.

[8] I. Fitriyanto, T. Radillah, L. Tambunan, and A. Fauziyyah, "IMPLEMENTASI METODE RANDOM FOREST PADA TEXT MINING UNTUK KLASIFIKASI SMS SPAM MENGGUNAKAN PYTHON," *INFORMATIKA: Jurnal Informatika, Manajemen dan Komputer*, vol. 17, no. 1, pp. 2580–3042, 2025, doi: http://dx.doi.org/10.36723/juri.v17i1.742.

[9] H. P. Tarigan, "Integrasi Chatbot Berbasis NLP pada Sistem Layanan Akademik Universitas," *Jurnal Komputer*, vol. 3, no. 1, pp. 13–18, 2024, doi: https://doi.org/10.70963/jk.v3i1.110.

[10] R. Merdiansah and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, vol. 7, no. 1, pp. 221–228, 2024, doi: https://doi.org/10.55338/jikomsi.v7i1.2895.

[11] N. Sofa, F. S. Utomo, and R. E. Saputro, "Eksplorasi Model Hybrid Transformer-Latent Semantic Analysis (LSA) Untuk Pemahaman Konteks Teks Berita Berbahasa Indonesia," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 5, pp. 1239–1252, May 2025, doi: 10.52436/1.jpti.662.

[12] A. Surahman Sulaeman, A. Sujjada, and I. Lucia Kharisma, "Penerapan Algoritma Cerdas Bidirectional Encoder Refresentations From Transformers Dalam Menganalisis Opini Publik Terhadap Produk Yang Mengalami Boikot," *Jurnal Inovtek Polbeng – Seri Informatika*, vol. 9, no. 1, pp. 460–473, 2024, doi: https://doi.org/10.35314/isi.v9i1.4252.

[13] F. N. Budiman, W. Witanti, and P. Nurul Sabrina, "Analisis Sentimen Ulasan Aplikasi CapCut Menggunakan Model RoBERTa Dengan Fitur Ekstraksi Word2vec," *Jurnal Algoritma*, vol. 22, no. 2, pp. 358–369, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2480.

[14] I. Maulana, "Pengaruh Penggunaan Code.org Sebagai Media Pengenalan Coding Dalam Mata Pelajaran Informatika MTS Nurul Huda Jubang Kelas VII," *JIEP: Jurnal Inovasi dan Evaluasi Pembelajaran*, vol. 1, no. 1, pp. 32–42, 2024, [Online]. Available: https://jeip.ipbcirebon.ac.id/

[15] Nur Saida and Muhammad Yasin, "Implementasi Metode Learning Vector Quantization (LVQ) untuk Klasifikasi Jumlah Penduduk Menurut Jenis Kelamin dan Kabupaten di Sumatera Utara," *Jurnal Teknik Informatika dan Teknologi Informasi*, vol. 5, no. 3, pp. 29–43, Oct. 2025, doi: 10.55606/jutiti.v5i3.6083.