JURNAL INFORMATIKA, Vol. 11 No. 2 October 2025, Page 113-121
ISSN (Print) 2355-6579 | ISSN (Online) 2528-2247
DOI: https://doi.org/10.31294/inf.v12i2.25528

Comparative study of DistilBERT and ELECTRA-Small Models in Spam Email Classification

Ferdy Agusman

Ministry of Finance of the Republic of Indonesia
Dr. Wahidin Raya Street No. 1, Central Jakarta 10710, Indonesia

Correspondence e-mail: ferdy.agusman@kemenkeu.go.id

Submission:	Revision:	Acceptance:	Available Online:
14-03-2025	16-07-2025	30-07-2025	03-10-2025

Abstract

Spam email detection is one of the challenging tasks in cybersecurity due to the variability of spam content. These characteristics make it harder to identify spam, therefore researchers create different spam detection methods. Among these, Natural Language Processing (NLP) and machine learning techniques have shown outstanding results in classifying emails as spam or non-spam. Transformer-based models, such as BERT, have demonstrated pinpoint accuracy in text classification tasks. However, the computational requirements and resources are not practical in resource-limited environments. To mitigate this, smaller and more lightweight models, such as the DistilBERT and ELECTRA-Small, have been developed. This paper presents a comparative study of the DistilBERT and ELECTRA-Small models for spam email classification. The objective is to evaluate the performance and computational efficiency of these two compact transformer architectures. Both DistilBERT and ELECTRA-Small models were fine-tuned on an email dataset comprising 5728 samples. Our experimental results on the primary test set indicate that both models achieved an accuracy of almost 99%. However, when evaluated on a separate external validation set containing 10,000 emails, the ELECTRA-Small model achieved an accuracy of 86.53%, outperforming DistilBERT's 83.68%. Furthermore, ELECTRA-Small demonstrated superior computational efficiency with a training time of 00:02:00, compared to DistilBERT's 00:04:46. This study represents one of the few studies to directly compare the performance and computational efficiency of these two models in the context of spam email detection, highlighting their potential as lightweight and effective solutions for real-world applications.

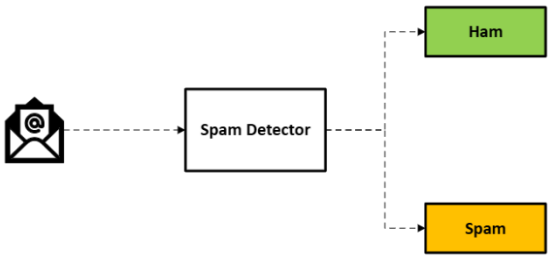
Keywords: Spam email, Machine learning, Transformer

1. Introduction

Email has become one of the most reliable and widely used communication mediums worldwide in the information technology era (AbdulNabi & Yaseen, 2021). However, due to its simplicity and accessibility, email is vulnerable to misuse (Ahmed et al., 2022). One of them is spam email; It is sent to many users at once, frequently containing cryptic messages, scams, or, most dangerously, phishing content (Sahmoud & Mikki, 2022). According to Statista (2023), approximately 347 billion emails were sent daily in 2023, and 160 billion of them were considered spam. The percentage of total email traffic that is identified as spam has consistently decreased (from 56.63% in 2017 to 45.6% in 2023). Despite the decrease, the actual number of spam emails continues to rise. Spam is not only annoying to the users but can also be categorized as a form of cybercrime, as it probably harms people, organizations, or governments (Jazzar et al., 2021).

There are multiple ways to detect spam, including blocklist and allowlist, heuristic, content, visual, artificial intelligence or machine learning, proactive, other techniques, and hybrid (Wood et al., 2022). Modern spam detection usually consists of a combination of all these methods. Each method complemented the other to create a robust system capable of handling spam email detection. Figure 1 illustrates how spam detectors work (adapted from Sahmoud & Mikki, 2022).

For non-machine learning base approach, spam detection was to manually construct document classifiers with rules compiled by domain experts. The problem with non-statistical approaches is that there is no learning component to admit messages whose content “look” legitimate and that leads to undetected spam emails, and therefore the detector accuracy will be low.



Source: Adapted from Sahmoud & Mikki, 2022
Figure 1: Workflow of spam email detection

Despite the technological and security improvement in spam detection systems, spam detection remains a never-ending problem across the globe (Akinyelu, 2021). The limitations of rule-based and non-statistical approaches in spam detection became increasingly evident. As spam emails grew more sophisticated, static rule-based systems could no longer keep up with the constantly evolving patterns and tactics used by modern attacks (Khan & Ghafoor, 2024). These traditional methods lacked the adaptability needed to address new forms of spam, often resulting in high rates of false negatives or false positives. The inability to generalize beyond predefined rules highlighted the need for a more dynamic and intelligent approach to spam detection. This growing challenge, coupled with the exponential growth of email traffic, set the stage for machine learning to emerge as a powerful alternative. This study bridges this gap by evaluating the effectiveness of lightweight transformer-based models.

Machine learning has emerged as a prominent technique for spam detection, proving effective in identifying and filtering unwanted emails (Nallamothu & Khan, 2023). This approach was developed to address common issues in spam filtering, such as false positives that incorrectly categorize genuine emails as spam. Machine learning uses text classification to differentiate emails by teaching the computer to classify text into different categories, thus catching and labeling spam emails before they reach the user's inbox.

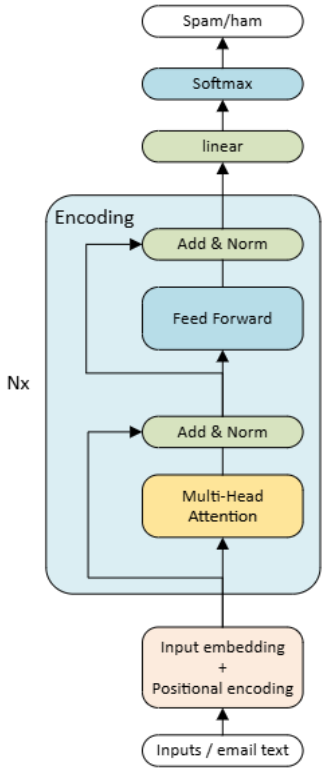
Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP often uses machine learning models, especially transformer-based models such as BERT, GPT, ELECTRA, or T5. These transformer-based models have revolutionized NLP by transforming input sequences into output sequences, thus leveraging self-attention techniques to understand context and relationships between different elements. Before the invention of transformer-based models, machine learning relied on simpler approaches in text classification tasks, with the raw text converted into numerical formats that models could process. While the methods of converting the text work as the foundation of machine learning, these approaches had notable problems, such as loss of context and fixed word embeddings. For example, the word “bank” would have the meaning of a financial institution even if the context of the texts referred to “bank” as the side of a river.

Unlike the traditional method, transformer-based models use different approaches to process text, the models recognize the entire sequence instead of sequentially processing every word. Transformer models are based on a self-attention mechanism that learns the relationships between elements of a sequence (S. Khan et al., 2022). By leveraging bidirectional attention, transformer models learn the context of a word based on both its preceding and succeeding words. This approach enables the model to understand semantic relationships within a sentence more effectively. Such contextual comprehension enhances performance across various natural language processing tasks, including text classification, machine translation, and question answering. This versatility extends to handling text classification in various linguistic contexts, including low-resource languages (Agbesi et al., 2023). These capabilities extend to complex real-world challenges such as offensive language identification in code-switched online content (Ranasinghe et al., 2020). The effectiveness of bidirectional transformers in spam classification tasks was also highlighted by (Guo et al., 2022), reinforcing our choice to adopt transformer-based architectures. Further research continues to improve these models, demonstrating their ongoing utility in diverse text classification applications (Tezgider et al., 2022). Figure 2 illustrates the general workflow of transformer models in a binary classification task, along with a corresponding explanation.

a. Inputs / Email Text: Represents the raw text input (emails in this case).
b. Input Embedding + Positional Encoding: The input text is tokenized into embeddings, and positional encoding is added to account for the sequence information.
c. Encoder Stack: This part consists of Multi-Head Attention to captures relationships between tokens. Add & Norm that applies normalization and residual connections. Feed Forward to applies transformations to enhance the representation. Nx is different from epochs in the training process. Nx is the number of layers in the model architecture, and both DistilBERT and ELECTRA-Small have 6 layers. A bigger N number means a deeper model, and it allows the model to learn more complex tasks. As an analogy, think of

stacked layers (Nx) as floors of a building**.** The building's height is fixed during construction. While epochs are like maintenance cycles for the building, where workers repeatedly do.

d. Linear Layer: Maps the output of the encoder to the desired output dimensions for classification.
e. Softmax: Converts the output scores into probabilities for each class.
f. Output (Spam/Ham): Final prediction of whether the email is spam or not spam.



Source: Research Data, 2024

Figure 2. General workflow of a transformer model for spam classification (Adapted from Tezgider et al., 2022)

However, the extensive computational and memory requirements of these models hinder their application in resource-limited scenarios (Yi & Xiao, 2024). Their large size limits their application on resource-constrained devices. To address this challenge, smaller, more efficient models have been developed. Two of these are DistilBERT and ELECTRA-Small. Both of these models offer a more balanced approach between efficiency and accuracy. Beyond these specific models, other architectural innovations, such as attentive convolutional transformers, also contribute to the development of efficient solutions for text classification (Li et al., 2021). Given the high performance-to-efficiency ratio demonstrated in prior studies such as (Akpatsa et al., 2022), our decision to fine-tune DistilBERT and ELECTRA-Small was informed by their demonstrated real-world applicability. For this reason, we decided to use DistilBERT and ELECTRA-Small as models for our spam detector. This study evaluates and compares both model's performance in text classification tasks, particularly spam detection. To determine the suitability of these models, we focus on metrics such as accuracy, precision, recall, and F-1 score.

DistilBERT is a compact version of BERT, created through a process of knowledge distillation. In this process, a larger BERT model (the "teacher") trains the smaller DistilBERT model (the "student") to mimic its behavior. (Silva Barbon and Akabane, 2022) demonstrated that DistilBERT preserves high accuracy (≈ 96%) while significantly reducing model size (≈ 40%) and training time (≈ 45%)—supporting the case for lightweight transformer models in multilingual text classification. resulting in a model that is smaller and faster while retaining over 95% of BERT's performance. On the other hand, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) uses a different pre-training approach. Instead of the Masked Language Model (MLM) objective used by BERT, ELECTRA is trained to identify which tokens in a sentence have been artificially "replaced" by a small generator network. This replaced token detection objective is computationally more efficient and allows ELECTRA-Small to achieve performance comparable to larger models. Jones (2023) evaluated the capabilities of fine-tuned ELECTRA for multi-class sarcasm detection, and while preliminary in nature, it highlights the potential adaptability of the model across nuanced language tasks. The applicability of these pre-trained language models also extends to tasks like emotion detection in various languages (Tepecik & Demir, 2024).

While both models have been used in various NLP tasks, their direct comparison, particularly within the context of spam email classification, remains limited. This study aims to fill that gap by comparatively evaluating the performance and computational efficiency of the DistilBERT and ELECTRA-Small models. We specifically highlight the trade-offs between training speed and accuracy, providing valuable insights into which model is best suited for spam detection in real-world applications.

2. Research Methods

The study employs a data-driven approach to classify spam emails using DistilBERT and ELECTRA-Small. The research methodology consists of several key steps from the dataset, text preprocessing, tokenization and training the models, classification process, and finally evaluation.

2.1 Datasets and Preprocessing

For this research, a dataset of 5728 emails, obtained from the Kaggle website, was used. This dataset was split into two subsets: 80% (4582 emails) was allocated for the training set, and the remaining 20% (1146 emails) was used as the test set. This division is a widely used heuristic in machine learning to ensure that the model has enough data for training while reserving a representative portion for unbiased performance evaluation (Bichri et al., 2024). Both models underwent identical training on the training set before being evaluated on the held-out test set. Additionally, a separate, large-scale external validation dataset, a subset of the ENRON Spam dataset containing 10,000 emails, was used for final model evaluation.

The text data from the emails underwent minimal pre-processing. The steps performed included the removal of double quotes ("), extra space, and emojis. We intentionally did not perform more advanced techniques such as lowercasing, stopword removal, or stemming. This decision was based on the consideration that modern transformer models are pre-trained on vast corpora and can extract contextual information from original text without aggressive pre-processing. This approach aims to preserve as much contextual information as possible from the original email text, which is crucial for distinguishing between spam and non-spam emails. However, it is noted that techniques such as text data augmentation have been explored to further enhance the performance of models like DistilBERT in text classification tasks (Nair et al., 2024).

Each dataset is divided into two columns, one is containing email text and the other containing label. The label is a binary indicator, where '1' indicates the email is spam and '0' indicates the email is not spam. Before the training process, datasets are normalized to match the requirements of the models. This process will ensure consistency of the data and minimize the probability of errors during the training process. Table 1 illustrates an example of the normalized dataset structure:

Table 1. Normalized dataset

Text	Label
Exclusive discount just for you	1
re: sms conference yes, i shall be glad to make a presentation. thanks for considering me.	0
Win a free iPhone by clicking here!	1

Source: Research Data, 2024

2.2 Model

In this study, we propose DistilBERT and ELECTRA-Small to leverage their unique strengths in natural language processing tasks, particularly text classification task.

2.2.1 DistilBERT

DistilBERT shares the same general architecture as BERT but optimized for efficiency and speed. It is a smaller, faster, cheaper, and lighter Transformer model trained using a process called knowledge distillation. In this process, the smaller model (DistilBERT) learns to mimic the behavior of a larger model (BERT). DistilBERT has 40% fewer parameters than BERT and runs 60% faster than BERT. Despite the reduction in size, it retains over 95% of BERT's performance across various natural language processing tasks. This combination of speed and accuracy makes DistilBERT highly suitable for resource-constrained environments (Sanh et al., 2020).

2.2.2 ELECTRA-Small

ELECTRA stands for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately". Unlike BERT, which uses a masked language model (MLM) for pretraining, ELECTRA introduces a unique replaced token detection (RTD) objective. Instead of masking words, ELECTRA replaces certain tokens with alternatives generated by a lightweight generator model. The discriminator (main model) then predicts whether each token is original or replaced. This approach is computationally efficient because it allows the model to learn from all input tokens, unlike MLM, where only masked tokens contribute to learning (Clark et al., 2020). ELECTRA-Small is the smaller version of ELECTRA, and particularly well-suited for tasks requiring lower computational costs, making it a strong competitor to models like BERT and DistilBERT. Its efficacy has been shown in various specialized text classification tasks, such as emotional classification of Chinese short comment text (Zhang et al., 2022).

2.3 Tokenization and Training the Models

The first step of our spam detection is tokenization using tokenizer. The tokenizer is a preprocessing tool that works by converting the texts into a format as input to the models. It breaks words into subword units in the training data based on their frequency. Table 2 shows an example of how a sentence is tokenized in a transformer-based model. This illustration was created to explain the presence of tokens like [CLS], [SEP], and subword fragments (e.g., "##ization").

"This text is for tokenization only"

The tokenizer split the sentence into these token

```
['[CLS]', 'this', 'text', 'is', 'for', 'token', '##ization', 'only', '[SEP]']
```

Then, every word converted into ID or encoding process

```
[101, 2023, 3793, 2003, 2005, 19204, 15128, 2069, 102]
```

Where the token IDs represented by table 2.

Table 2. Example of tokenization process in a transformer model (for explanation purposes)

Token ID	Token	Notes
2023	this	
3793	text	
2003	is	
2005	for	
19204	token	Subword, part of "tokenization"
15128	##ization	Subword, continuation of "tokenization"
2069	only	

Source: Research Data, 2024

101→[CLS]:represents the beginning of a sequence (used for classification tasks).
102→[SEP]:represents the end of the sequence.

By splitting into subwords, the tokenizer helps BERT and ELECTRA understand word parts and their meanings. It also splits rare or unknown words into subword components rather than discarding them.

After the tokenization process, we utilized pre-trained DistilBERT and ELECTRA-small models to initiate the training phase. To optimize the training workflow and ensure the models achieve efficient and accurate performance, a set of training arguments was configured. These arguments serve as guidelines for controlling various aspects of the training process, such as learning rate, batch sizes, number of epochs, evaluation strategy, and model saving behavior.

The learning rate, set at 5e-5, this number is scientific notation for 0.00005. Learning rate determines how quickly the model updates its weights during training. This value strikes a balance between making steady progress and avoiding overcorrection, which could destabilize the training process. Similarly, batch sizes for training and evaluation are specified as 16 per device, balancing memory efficiency and the ability to process more data at once for gradient calculations.

The training runs for a total of 3 epochs, a common choice for fine-tuning pre-trained models like DistilBERT and ELECTRA-small, allowing the models sufficient time to learn without overfitting to the data. Additionally, evaluation is conducted at the end of each epoch, providing insights into the model's performance at regular intervals and enabling adjustments if needed. Alongside evaluation, the best-performing model, determined based on accuracy, is saved to ensure only the most effective version is retained.

Other parameters, such as weight decay, set at 0.01, are included to regularize the model and improve generalization. Logging is also configured, with updates provided every 10 steps to monitor training progress. Furthermore, the save strategy ensures that model checkpoints are stored at the end of each epoch, safeguarding against data loss and enabling resumption from the last saved state if training is interrupted.

These configurations collectively ensure a well-structured and efficient training process, balancing computational demands and model accuracy while adapting to the requirements of the spam classification task.

All experiments were implemented using Python and taken on a computer equipped with an Apple M3 processor and 16 GB of memory. While training times averaged around two minutes per run, this was considered modest given the resource constraints and the use of transformer-based models.

2.4 Evaluation Metrics

To evaluate both models and make a comparison study, we need to calculate the accuracy, precision, recall, and F1 Score. These metrics defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

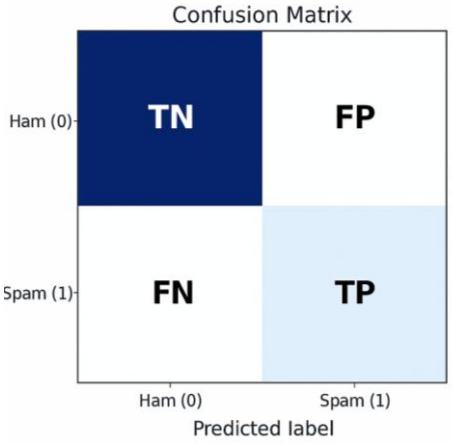
$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Where TP is the number of true positives where spam samples are predicted as spam, TN is the number of true negatives, or correctly predicted ham, FP is the number of false positives or ham incorrectly predicted as spam, and FN is the number of false negatives or spam incorrectly predicted as ham. Furthermore, Accuracy is the

percentage of total correct predictions (both spam and ham) out of all predictions. Precision is the percentage that was actually spam from all the predicted spam. While recall, is the percentage correctly identified as spam from all actual spam emails. Finally, F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. Figure 3 below is the explanation of confusion matrix often used in binary classification of machine learning. A confusion matrix is a fundamental tool for evaluating the performance of classification models, especially in binary classification problems (Fahmy Amin, M., 2022).



Source: Research Data, 2024

Figure 3. Confusion matrix explanation

From the confusion matrix, we can calculate four measures for each class which are accuracy, precision, recall and the f1-score. The closer these metrics are to 100%, the better the model's performance, as it indicates the model accurately predicting spam with minimum errors.

3. Results

Table 3 below shows the performance metrics of the trained models on the test set. While both models achieved high scores, DistilBERT exhibited a slightly higher mean F1-score and recall, while ELECTRA-Small achieved higher precision.

Table 3. Result Comparison

Model	Acc	Prec	Recall	F1 Score
DistilBERT	98.69%	98.93%	95.86%	97.37%
ELECTRA-Small	98.69%	99.64%	95.17%	97.35%

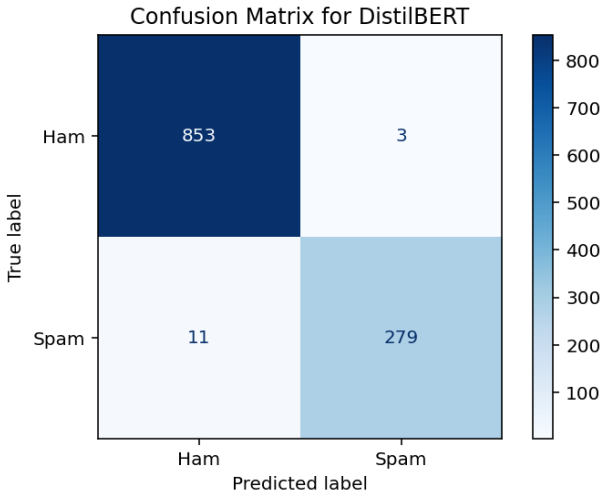
Source: Research Data, 2024

In addition to performance, computational efficiency is also a key factor in this comparison. While the training time for both models was relatively short, ELECTRA-Small demonstrated faster training, completing it in 2 minutes compared to 4 minutes and 46 seconds for DistilBERT. Model size and inference time are also critical considerations for real-world deployment. ELECTRA-Small has approximately 14 million parameters and a model size of around 55 MB, while DistilBERT contains 66 million parameters and is roughly 255 MB. These results confirm that ELECTRA-Small not only achieves competitive performance but also offers clear advantages in efficiency, making it a more attractive choice for deployment in resource-constrained environments.

3.1 DistilBERT Result

The DistilBERT model demonstrated strong performance in the spam email classification task, effectively distinguishing spam from non-spam emails with high accuracy on the test dataset and good overall efficiency. DistilBERT exhibited balanced precision and recall, showcasing its ability to minimize both false positives and false negatives effectively. A high F1 score further underscored the model's robustness in handling the classification task.

Overall, the results highlight DistilBERT as a lightweight yet powerful model, combining computational efficiency with strong classification performance, making it an excellent candidate for practical applications in spam detection systems. From the confusion matrix, we can see that DistilBERT correctly predicted 853 "Ham" emails (True Negatives) and 279 "Spam" emails (True Positives). The DistilBERT model incorrectly predicted 3 "Ham" emails as "Spam" (False Positives) and 11 "Spam" emails as "Ham" (False Negatives). The matrix shows that the model performed exceptionally, with a high number of correctly classified emails and a low number of errors.



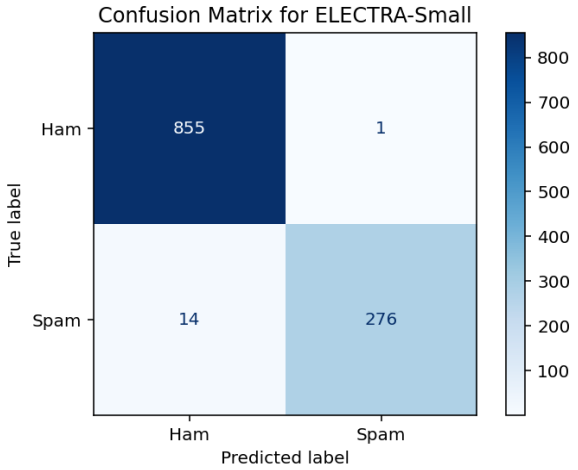
Source: Research Data, 2024

Figure 4. DistilBERT confusion matrix result

3.2 ELECTRA-Small Result

The ELECTRA model delivered competitive performance in the spam email classification task. Overall, ELECTRA-small proved to be a powerful model for spam classification, offering a robust and precise solution for handling complex email datasets. As the confusion matrix shows, ELECTRA-small correctly predicted 855 "Ham" emails (True Negatives) and 276 "Spam" emails (True Positives).

The model incorrectly predicted 1 "Ham" email as "Spam" (False Positive) and 14 "Spam" emails as "Ham" (False Negative). The matrix shows that compared to DistilBERT, the ELECTRA-small model performed exceptionally well, with a higher number of correctly classified "Ham" emails (True Negatives) and a lower number of "Ham" emails incorrectly classified as "Spam" (False Positives).



Source: Research Data, 2024

Figure 5. ELECTRA-Small confusion matrix result

3.3 External Validation Results

The robustness of these findings is further supported by the evaluation of an external validation set containing 10,000 emails. On this dataset, both models maintained high performance, with the results showing a similar pattern to our primary test set. This demonstrates that the models' comparable performance is not an artifact of a single test split and that they generalize effectively to a completely new and independent dataset.

Table 4. Performance on the External Validation Set (10,000 records)

Model	Accuracy	Precision	Recall	F1-Score
DistilBERT	82.91%	77.04%	96.71%	85.76%
ELECTRA -Small	83.82%	78.60%	95.64%	86.29%

Source: Research Data, 2024

3.4 Statistical Significance Analysis

To determine if the observed performance differences were statistically significant, a paired bootstrap resampling test was conducted with 10,000 iterations. The analysis was based on the F1-score as the primary metric. The results are summarized in Table 5.

Table 5. Bootstrap Resampling Results

Metric	ELECTRA-Small	DistilBERT
Mean F1-Score	0.9716	0.9788
Std Dev	0.0072	0.0062

Source: Research Data, 2024

A 95% confidence interval (CI) was calculated for the difference in F1-scores (F1_ELECTRA - F1_DistilBERT). The CI was found to be [-0.0195, 0.0052]. As this interval includes zero, the performance difference between ELECTRA-Small and DistilBERT is not considered statistically significant. This finding was further corroborated by a paired t-test, which yielded a p-value of 0.2484 ($p > 0.05$).

4. Discussion

4.1 Performance Trade-offs and Statistical Implications

The statistical analysis revealed a crucial insight: while ELECTRA-Small and DistilBERT exhibited a slight difference in mean F1-score on the test dataset (0.9716 vs. 0.9788), this disparity was not statistically significant. This suggests that for the task of spam detection on this dataset, both models offer comparable performance in terms of classification accuracy and F1-score.

This finding has important implications for real-world applications. Given that the performance is statistically equivalent, the decision of which model to use should then be based on other factors, such as computational efficiency (training time) and model size. As shown in our previous results, the ELECTRA-Small model demonstrated superior training time, making it a more attractive choice for scenarios where faster fine-tuning is required without sacrificing classification performance. The statistical test validates that choosing ELECTRA-Small does not come at the cost of significant performance degradation. These findings align with insights from Lu et al. (2022), suggesting that while transformer encoders offer superior accuracy, alternative architectures like Convolutional Neural Network (CNN) may be considered for balanced or resource-constrained scenarios.

5. Conclusion

In conclusion, this study evaluated the performance of DistilBERT and ELECTRA-Small in the task of email spam detection, focusing on their accuracy, precision, recall, F1 score, and computational efficiency. Both models demonstrated remarkable capabilities in identifying spam emails, showcasing the effectiveness of transformer-based architectures in text classification tasks.

While ELECTRA-Small and DistilBERT showed slightly different performance scores, a robust statistical analysis revealed that this difference was not statistically significant. Given their comparable classification performance, ELECTRA-Small's superior training time makes it the more practical and efficient choice for resource-constrained environments. This research underscores the potential of lightweight transformer models to provide effective and efficient solutions for spam detection, offering valuable insights for future advancements in cybersecurity and natural language processing.

Following recommendations by Khan and Ghafoor (2024), future work should explore ensemble defenses and adversarial training tailored for small transformer models deployed in constrained environments. Future research could explore fine-tuning the models with larger datasets to enhance performance. Investigating additional lightweight models, such as TinyBERT or MobileBERT, may offer more efficient solutions for resource-constrained systems. Real-time deployment and evaluation could test the models' practicality in dynamic environments, while integrating multimodal data, such as links or attachments, may further improve accuracy. Additionally, assessing adversarial robustness and comparing transformer-based approaches with traditional models could provide valuable insights. Lastly, optimizing the models for adaptability and efficiency would ensure long-term effectiveness and broader applicability.

Reference

AbdulNabi, I., & Yaseen, Q. (2021). Spam Email Detection Using Deep Learning Techniques. *Procedia Computer Science*, *184*, 853–858. https://doi.org/10.1016/j.procs.2021.03.107

Agbesi, V. K., Chen, W., Yussif, S. B., Hossin, M. A., Ukwuoma, C. C., Kuadey, N. A., ... & Al-antari, M. A. (2023). Pre-Trained Transformer-Based Models for Text Classification Using Low-Resourced Ewe Language. *Systems*, *12*(1), 1. https://doi.org/10.3390/systems12010001

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges. *Security and Communication Networks*, *2022*, 1–19. https://doi.org/10.1155/2022/1862888

Akinyelu, A. A. (2021). Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques. *Journal of Computer Security*, *29*(5),473–529. https://doi.org/10.3233/JCS-210022

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. *International Journal of Advanced Computer Science and Applications, 15*(2). 331-339.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* (No. arXiv:2003.10555). arXiv. https://doi.org/10.48550/arXiv.2003.10555

Fahmy Amin, M. (2022). Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial. *Journal of Engineering Research:* Vol. 6: Iss. 5 Article 1. https://digitalcommons.aaru.edu.jo/erjeng/vol6/iss5/1

Guo, Y., Mustafaoglu, Z., & Koundal, D. (2022). Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal of Computational and Cognitive Engineering*,*2*(1),59.https://doi.org/10.47852/bonviewJCCE2202192

Jazzar, M., F. Yousef, R., & Eleyan, D. (2021). Evaluation of Machine Learning Techniques for Email Spam Classification. *International Journal of Education and Management Engineering*, *11*(4),35–42. https://doi.org/10.5815/ijeme.2021.04.04

Jones, I. (2023). *Assessing the efficacy of the ELECTRA pre-trained language model for multi-class sarcasm subcategory classification* [Master's thesis, University of Bath]. Bath Research Portal.https://researchportal.bath.ac.uk/en/publications/assessing-the-efficacy-of-the-electra-pre-trained-language-model-

Khan, M., & Ghafoor, L. (2024). *Adversarial machine learning in the context of network security: Challenges and solutions. Journal of Computational Intelligence and Robotics*, *4*(1), 51-63.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kofi Akpatsa, S., Lei, H., Li, X., Kofi Setornyo Obeng, V.-H., Mensah Martey, E., Clement Addo, P., & Dodzi Fiawoo, D. (2022). Online

News Sentiment Classification Using DistilBERT. *Journal of Quantum Computing*, *4*(1), 1–11. https://doi.org/10.32604/jqc.2022.02665

Li, P., Zhong, P., Mao, K., Wang, D., Yang, X., Liu, Y., Yin, J., & See, S. (2021). ACT: An Attentive Convolutional Transformer for Efficient Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(15), 13261–13269. https://doi.org/10.1609/aaai.v35i15.17566

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, *22*(1), 181. https://doi.org/10.1186/s12874-022-01665-y

Nair, A. R., Singh, R. P., Gupta, D., & Kumar, P. (2024). Evaluating the Impact of Text Data Augmentation on Text Classification Tasks using DistilBERT. *Procedia Computer Science*, *235*, 102–111. https://doi.org/10.1016/j.procs.2024.04.013

Nallamothu, P. T., & Khan, M. S. (2023). Machine learning for SPAM detection. Asian Journal of Advances in Research, 6(1), 167-179. https://jasianresearch.com/index.php/AJOAIR/article/view/296

Ranasinghe, T., Gupte, S., Zampieri, M., & Nwogu, I. (2020). *WLV-RIT at HASOC-Dravidian-CodeMix-FIRE2020: Offensive language identification in code-switched YouTube comments* (arXiv:2011.00559). ArXiv.https://doi.org/10.48550/arXiv.2011.00559

Sahmoud, T., & Mikki, M. (2022). *Spam detection using BERT* (arXiv:2206.02443). arXiv. https://doi.org/10.48550/arXiv.2206.02443

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (No. arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, *22*(21), 8184. https://doi.org/10.3390/s22218184

Statista. (2023). Number of sent and received e-mails per day worldwide from 2017 to 2026. https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/

Tepecik, A., & Demir, E. (2024). *Emotion Detection with Pre-Trained Language Models BERT and ELECTRA Analysis of Turkish Data. Intelligent Methods In Engineering Sciences, 3*(1), 7-12. https://doi.org/10.58190/imiens.2024.82

Tezgider, M., Yildiz, B., & Aydin, G. (2022). Text classification using improved bidirectional transformer. *Concurrency and Computation: Practice and Experience*, *34*(9), e6486. https://doi.org/10.1002/cpe.6486

Wood, T., Basto-Fernandes, V., Boiten, E., & Yevseyeva, I. (2022). Systematic Literature Review: Anti-Phishing Defences and Their Application to Before-the-click Phishing Email Detection. *arXiv preprint arXiv:2204.13054.*

Zhang, S., Yu, H., & Zhu, G. (2022). An emotional classification method of Chinese short comment text based on ELECTRA. *Connection Science*, *34*(1), 254–273. https://doi.org/10.1080/09540091.2021.1985968