WebMail Development and Implementation of Email Spam Detection using Convolutional Neural Network (CNN) Method

Muhammad Ihsan Ramadhan
School of Computing
Telkom University
Bandung, Indonesia
eberama@student.telkomuniversity.ac.id

Hilal H. Nuha
School of Computing
Telkom University
Bandung, Indonesia
hilalnuha@ieee.org

Niken Dwi Wahyu Cahyani
School of Computing
Telkom University
Bandung, Indonesia
nikencahyani@telkomuniversity.ac.id

Setyorini
School of Computing
Telkom University
Bandung, Indonesia
setyorini@telkomuniversity.ac.id

Mohd Arfian Bin Ismail
School of Computing
University Malaysia
Pahang Al-Sultan Abdullah, Malaysia
arfian@umpsa.edu.my

***Abstract*— Email spam poses a significant issue in the digital era, disrupting user experiences and potentially threatening security. This study investigates the performance and effectiveness of the Convolutional Neural Network (CNN) method in detecting email spam, leveraging its ability to extract and analyze patterns from textual data. The model's performance was evaluated through confusion matrix metrics, including precision, recall, and F1-score. in addition, a webmail application was developed to integrate the CNN spam detection model, providing users with essential email functionalities. This application allows users to send, receive, and manage emails via SMTP for email transmission and IMAP/POP3 for retrieval. Developed with a front-end using HTML, CSS, and Bootstrap, and a backend in PHP and JSON, the webmail system stores all email data in an SQL database. Python was utilized to implement and train the CNN model, which automatically filters incoming emails, classifying them as either spam or non-spam before they reach the user's inbox. The results demonstrate that the CNN model achieved a high accuracy of 99.80% with the Adam Optimizer, indicating its robust capability in accurately detecting spam emails.**

Keywords—Email Spam Detection, Convolutional Neural Network, Webmail, Deep Learning.

I. INTRODUCTION

Email has become an essential communication tool in the digital era, enabling fast and easy information exchange. However, this convenience is often exploited by malicious parties to spread spam emails, which can disrupt and endanger users. Spam emails may contain unwanted advertisements, phishing scams, malware, and other harmful content. Spam email has become one of the leading threats worldwide and has caused substantial financial losses. Despite ongoing updates to spam prevention techniques, the effectiveness of these strategies is still not fully optimal [1].

In Indonesia, spam email is a growing issue. According to data from Kaspersky Lab, Indonesia ranked first in Southeast Asia in terms of the number of spam emails received in 2022 [2]. This highlights that Indonesian email users are highly vulnerable to spam attacks. Spam emails not only waste users' time but can also compromise their security. For example, phishing emails are designed to trick users into revealing personal information or clicking on malicious links that can infect their devices with malware [3].

The problem of detecting spam emails has been the focus of various studies over the years, with researchers exploring numerous machine learning and deep learning methods to improve accuracy and efficiency. This section discusses several key studies that have explored different approaches for email spam detection.

Soni et al. [1] developed an advanced deep learning model for spam email detection using a modified Recurrent Convolutional Neural Network (RCCN). This model analyzes various elements of emails, including the header, body, character-level, and word-level features to improve detection accuracy. The introduction of staggered vectors and attention mechanisms enhanced the model's performance, resulting in an accuracy of 99.848% and a low false-positive rate (FPR) of 0.043%. However, the model's complexity requires significant computational resources for training.

Eckhardt and Bagui [9] conducted a comparative study between Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for phishing email classification. They experimented with different activation functions and optimizers (Adam and SGD) to optimize performance. Their findings showed that both models could effectively detect phishing emails, though the scope of comparison was limited to these two models, and the dataset used was relatively small.

Merugu et al. [3] focused on CNN for phishing detection, introducing preprocessing steps like feature engineering and transfer learning. Their model was specifically designed to recognize not only existing phishing templates but also zero-day phishing attacks. The results indicated high accuracy and generalizability, although the model's complexity made it resource-intensive to train.

Adebowale and Lwin [17] proposed a hybrid model combining CNN and LSTM to detect phishing websites. Their model analyzed website elements such as URLs, images, and frames, achieving an accuracy of 93.28%. This hybrid approach eliminates the need for manual feature

extraction, improving efficiency while demonstrating potential for further development.

Sefat and Ullah [8] combined CNN with Bidirectional Long Short-Term Memory (Bi-LSTM) to improve spam detection. The model achieved a balance between capturing word sequences and sentiment analysis within email content. This hybrid model demonstrated a training efficiency improvement compared to LSTM-only models, with accuracies ranging from 98% to 99%. Despite these advantages, the text-based extraction process remains computationally expensive.

Aiwan and Zhaofeng [4] tackled the challenge of image spam detection, which traditional spam filters struggle with due to their reliance on text analysis. They utilized CNN for real-time image spam detection, incorporating data augmentation through clustering analysis to improve training sample quality. The results showed a significant improvement of 7-11% in image spam recognition accuracy compared to traditional methods.

In another study, Reddy and Ahila employed both CNN and K-Nearest Neighbor (KNN) classifiers to classify spam emails [6]. Using datasets from Kaggle, their model achieved higher accuracy with CNN (91.18%) compared to KNN (87.05%). Although the CNN outperformed KNN, the study highlights the need for further model optimization to improve overall detection performance.

Zavrak and Yilmaz [10] introduced a Hierarchical Attentional Hybrid Neural Network (HAN) that combines CNN, Gated Recurrent Units (GRU), and attention mechanisms. Their approach outperformed previous models by better capturing important information in email texts, emphasizing structure rather than simple word frequency. The model also employed cross-dataset evaluation, reducing training data bias and enhancing performance.

Yang et al. [11] proposed a multi-modal spam detection model that fuses CNN and LSTM for analyzing both text and image content in emails. The model uses grid search optimization and K-fold cross-validation, resulting in accuracies between 92.64% and 98.48%. This fusion-based approach demonstrates significant potential for detecting modern spam that integrates text and images.

Lastly, Hidayat's research [14] employed the Naive Bayes method for spam email classification. While this probabilistic approach is simpler and easier to implement, it achieved only 60% accuracy on a small dataset, highlighting its limitations compared to more advanced methods like CNN and LSTM.

In summary, the reviewed literature demonstrates the progression of spam detection techniques from traditional machine learning models like Naive Bayes to advanced deep learning architectures such as CNN, LSTM, and their hybrids. Each study contributes unique insights, with attention mechanisms, hybrid models, and multi-modal approaches significantly improving detection performance. However, challenges remain in terms of model complexity, resource consumption, and the need for larger and more diverse datasets.

To address these issues, this study analyzes the problem using the Convolutional Neural Network (CNN) method. CNN is a type of artificial neural network highly effective in analyzing both image and text data [4]. CNN has proven capable of classifying data with high accuracy, making it ideal for spam email detection applications. CNN can learn patterns and characteristics in email text, such as specific words, sentence structures, and link patterns, to identify spam emails more accurately than traditional methods like keyword-based filters [5]. Several studies in Indonesia have implemented CNN for spam email detection. A notable case study conducted by Mercu Buana University in 2024 demonstrated CNN's significant effectiveness in classifying emails, achieving a high accuracy rate of 99.67% for 20% test data, 99.64% for 30% test data, and 99.63% for 40% test data. These results illustrate the strong potential of CNN in enhancing digital security [6].

CNN can be applied not only to analyze images and text within emails but also to detect phishing, malware, and other cyber threats in emails. Therefore, email spam detection has garnered considerable attention from academia and industry due to the data breaches and financial damages experienced by both private companies and government organizations caused by phishing attacks [7].

Another study found that CNN can accelerate training time on datasets and extract high-level text features [8]. While some research shows that CNN performs well for text classification tasks [9], many studies suggest that CNN has the potential to become a hybrid model when combined with other models, yielding superior results compared to previous models [10]. There is also research combining CNN with Long Short Term Memory (LSTM), achieving an accuracy range of 92.64% to 98.48% using a hybrid model called MMA-MF [11].

This study investigates the effectiveness of the CNN in email spam detection, but several limitations exist. First, the dataset used is specific to spam and non-spam emails, which may not represent the entire range of email types encountered in real-world scenarios. The model may face challenges when processing more complex spam techniques such as phishing or zero-day attacks, which were not fully addressed here. Second, the performance of the CNN model heavily depends on the quality and size of the training data. A lack of sufficient or balanced data can lead to issues such as overfitting or decreased generalization capability.

Additionally, the computational resources required for training CNN models are significant, which limits scalability for larger applications. Lastly, the focus of this research was primarily on spam detection and did not encompass other cyber threats such as phishing or malware, which could broaden the practical utility of the system.