Email Spam Detection using Long Short-Term Memory (LSTM) Network Method

1st Agung Althaaf Emha Damanik
School of Computing
Telkom University
Bandung, Indonesia
agungalthaaf@student.telkomuniversity.ac.id

2st Hilal H. Nuha
School of Computing
Telkom University
Bandung, Indonesia
hilalnuha@ieee.org

3rd Niken Dwi Wahyu Cahyani
School of Computing
Telkom University
Bandung, Indonesia
nikencahyani@telkomuniversity.ac.id

4th Setyorini
School of Computing
Telkom University
Bandung, Indonesia
setyorini@telkomuniversity.ac.id

5th Mohd Arfian Bin Ismail
School of Computing
University Malaysia
Pahang Al-Sultan Abdullah, Malaysia
arfian@umpsa.edu.my

I. INTRODUCTION

Over the past few years, the use of email as a communication medium has grown rapidly. With over 2.6 billion active users and 4.6 billion registered email accounts, email has become one of the most significant and widely used internet communication media. Alongside the growth and expansion of email usage, there has also been an increase in the transmission of unsolicited messages, commonly known as spam, primarily for advertising purposes, which contributes to increasing the profitability of investments across various businesses. Out of approximately 12.5 million spam emails sent, only one response is received. This figure may seem small, but it becomes substantial when considering that over 14 billion spam messages are sent each day [1].

Email is very popular due to its speed and ease in sharing information with others. Spam is a major issue that involves sending large amounts of unwanted messages to recipients with the aim of stealing confidential information, spreading malware, and promoting various products. Although many machine learning algorithms have been designed to detect spam emails, none of these algorithms can predict spam emails with a high level of accuracy [2].

Conventional spam filtering systems that only analyze text have a high detection rate for text-based spam. However, spammers attempt to evade detection by hiding spam information within multi-modal elements of the email (such as images, links, and so forth). This situation renders conventional spam filtering systems less effective in addressing the spam problem [3].

Oni conducted research by implementing the Recurrent Neural Network (RNN) method with LSTM to perform sentiment analysis on tourist destinations in Yogyakarta using Twitter data. The system was able to classify with an accuracy rate of 95.98% through libraries and 70% through classification forms. However, the system was not yet able to preprocess abbreviated and slang words [7].

Various studies have explored the application for both machine learning and deep learning approaches for spam detection. Isik et al. analyzed the effectiveness of feature selection methods like Mutual Information (MI) and Weighted Mutual Information (WMI) combined with deep learning models such as Long Short-Term Memory (LSTM) to classify email spam in Turkish. This research achieved 100% accuracy using LSTM, indicating LSTM's suitability for spam detection in different languages [8].

Furthermore, by Wijaya et al. proposed a spam detection model based on LSTM for detecting spam emails. Using a dataset of over 5,000 entries, this study demonstrated that the LongSpam model could serve as an effective spam filter, outperforming traditional machine learning models [9]. LSTM, a type of recurrent neural network, is frequently used in spam detection due to its ability to capture long-term dependencies in sequential data.

Cahyadi et al. implemented LSTM to conduct sentiment analysis on Instagram comments, highlighting LSTM's potential adaptability for spam detection tasks by showcasing its capability to handle textual classification tasks [10]. Traditional methods, such as Support Vector Machine (SVM) and Naïve Bayes, have also been widely used for spam classification but lack the sequential data processing capabilities of LSTM, which can limit their effectiveness for complex text classification tasks.

Studies by Mubarikah compared LSTM and SVM, showing that LSTM consistently outperformed SVM in spam classification accuracy [11].

Effective spam detection models often rely on optimized feature selection and parameter tuning. Al Bataineh and Kaur introduced an immunocomputing-based approach to optimize LSTM architectures for text classification, underscoring the need for hyperparameter optimization to enhance LSTM's performance [12]. Despite LSTM's effectiveness, challenges remain in addressing issues like handling multi-modal spam and processing large, imbalanced datasets.

Studies such as Neha and Nair suggest integrating LSTM with attention mechanisms and inception layers to further improve its spam detection capabilities on platforms like Twitter, highlighting future directions that could apply similarly to email spam filtering [13].

In this study, the authors developed a spam detection classification system using LSTM, a method commonly used in deep learning applications to address text classification issues such as spam detection. LSTM represents a form of recurrent neural network architecture able to understanding long-term context in sequential data, such as text. The use of LSTM in spam detection enables the model to learn complex pattern and contextual relationships within text. However, it should be noted that effective use of LSTM requires an adequate amount of training data, along with proper data processing and model tuning to achieve optimal results [5].

The effectiveness of the LSTM model largely depends on the quality and volume of the dataset. A limited or imbalanced dataset may restrict the model's ability to accurately learn and detect spam patterns, potentially affecting overall accuracy. Training an LSTM model is computationally intensive, often requiring significant processing power and memory. Due to constraints on high-performance hardware, the model's performance and the speed of experimentation could be limited. This study primarily addresses text-based spam detection. However, spammers often embed spam content in multi-modal formats, such as images or links. The LSTM model used here may not efficiently detect such multi-modal spam without additional preprocessing or specialized models.

Achieving high accuracy with LSTM demand meticulous tuning of hyperparameters, including learning rate, batch size, and sequence length. Due to time constraints, it may not be feasible to exhaustively test all parameter combinations, which could limit the model's optimal effectiveness. The success of this model relies on effective text preprocessing. Variability in preprocessing methods-such as handling abbreviations, slang, or special characters-could impact the model's performance and consistency.