# Optimized Hybrid Deep Learning for Enhanced Spam Review Detection in E-Commerce Platforms

Abdulrahman Alghaligah, Ahmed Alotaibi, Qaisar Abbas, and Sarah Alhumoud*
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh 11432, Saudi Arabia

***Abstract*—Spam reviews represent a real danger to e-commerce platforms, steering consumers wrong and trashing the reputations of products. Conventional Machine learning (ML) methods are not capable of handling the complexity and scale of modern data. This study proposes the novel use of hybrid deep learning (DL) models for spam review detection and experiments with both CNN-LSTM and CNN-GRU architectures on the Amazon Product Review Dataset comprising 26.7 million reviews. One important finding is that 200k words vocabulary, with very little preprocessing improves the models a lot. Compared with other models, the CNN-LSTM model achieves the best performance with an accuracy of 92%, precision of 92.22%, recall of 91.73% and F1-score of 91.98%. This outcome emphasizes the effectiveness of using convolutional layers to extract local patterns and LSTM layers to capture long-term dependencies. The results also address how high constraints and hyperparameter search, as well as general-purpose represents such as BERT. Such advancements will help in creating more reliable and reliable spam detection systems to maintain consumer trust on e-commerce platforms.**

***Keywords*—*Spam review detection; CNN-LSTM; CNN-RNN; CNN-GRU; big data; deep learning; amazon product review dataset***

## I. Introduction

E-commerce platforms are becoming the primary marketplaces for almost every good, replacing traditional stores in many fields. Both the seller and the customer widely accept them because they reduce costs for the seller and allow the customer to access the goods faster. Platforms like Amazon, Alibaba, and Noon dominate the global retail landscape. However, consumers face a difficult challenge when shopping online. They are unable to assess products before purchase. They tend to check online reviews and base their buying decision on them Najada and Zhu [1]. This has given rise to the issue of deceptive content, known as spam reviews. It aims to misguide consumers for specific gains. According to a report from the Department of Business & Trade from the UK government, "At least 10% of all product reviews on third-party e-commerce platforms are likely to be fake"[2]. This underlined the importance of checking the trustworthiness of online reviews. Since online reviews are important in consumers' buying decisions, spam review detection is a priority. When traditional spam detection techniques began, the focus was on areas like email. Nowadays, the focus has shifted to review-based spam detection Li et al. [3]. Spam detection can be categorized into two techniques: content-based approach and user-behavior-based approach. Content-based approach analyzes the text content only and extracts semantic relations between words. On the other hand, the user-behavior approach focuses on the patterns of reviewer activities Li et al. [4].

Detecting spam reviews has several challenges. Spam reviews may look like genuine ones in content, making it difficult to distinguish spam from non-spam reviews. Also, spammers are always improving their techniques; they tend to use auto-generated tools to evade detection Bhuvaneshwari et al. [5]. ML started a revolution to protect consumers from scams. Traditional ML classifiers, such as Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), have good results in detecting spam reviews Rizali et al. [6] Additionally, with the increasing complexity of the reviews, ML classifiers were not enough to handle this matter. Hence, DL comes to replace ML by showing a better performance in detecting complex patterns. However, current studies have not used DL on large datasets such as the Amazon Product Reviews dataset Hussain et al. [7] which contains 26.7 million labeled reviews. This opens a research gap to leverage hybrid deep learning models' capabilities to process large datasets, achieving superior performance. This study aims to enhance the ability to detect spam reviews by using large datasets leveraging the latest DL technologies. CNN-LSTM, CNN-RNN, and CNN-GRU were applied to the Amazon Products Reviews dataset Hussain et al. [7], which offers a rich and diverse set of labeled data. To optimize the detection process, various text preprocessing techniques are evaluated, including tokenization, lowercasing, lemmatization, stop word removal, punctuation removal, and embedding. The proposed framework provides an efficient, and accurate spam reviews detection, providing valuable insights into the impact of preprocessing techniques on detection outcomes. To verify the effectiveness of the proposed classifiers and preprocessing steps, accuracy, precision, recall, and F1Score were applied. The contributions of the paper are as follows:

*1)* A novel and accurate spam review detection is presented in this paper in which the accuracy is 92% through CNN-LSTM model which is quite better than results acquired from traditional ML methods leading to new detection accuracy benchmark.

*2)* Using the Amazon Product Review Dataset (26.7 million reviews) for demonstration, the research provides evidence for how hybrid models can be utilized to efficiently process vast-scale, diverse data, which detects an essential scalability barrier.

*3)* In the same work, the authors state that less preprocessing obtains better performance, since it retains important information from the text, and that the common tendency to make more preprocessing does not favor the use of the model; these findings can provide important guidelines for the design of future models.

*4)* This study found that a large vocabulary size (200,000 max words) allows the model to better represent complex relationships between words, showing that careful vocabulary selection can improve spam detection effectiveness significantly.

*5)* This paper suggested a scalable and efficient spam detecting framework to better maintain consumer trust in the marketplace by excluding fake reviews and pretending user occurrence.

The rest of the paper is structured as follows: Section I and Section II discusses the background of spam detection and previous works in the field. Section III presents the proposed methodology. Section IV shows the result of the proposed work and discusses the findings in Section V. Finally, Section VI presents a summary of the study and future work.

## II. Literature Review

Spam refers to deceptive content intentionally made to mislead or manipulate users for a specific gain, usually a commercial gain Jakupov et al. [8]. In online reviews, spam is a false or misleading review designed to promote or demote a particular product, store, book, or goods and services. It can influence customers purchasing decisions and damage the reputation of products on e-commerce platforms like Amazon Fei et al. [9]. Most users refer to reviews to decide whether they buy a product, as they need physical access to assess it. Studies indicate that nearly 30% of online reviews are spam Farooq [10], highlighting the issue of reviews on famous platforms such as Amazon. Spam detection in online reviews has become crucial as most e-commerce platforms rely heavily on user input to guide consumer choices. In the past decade, spam detection focused more on traditional applications like spam Short Message Service (SMS). However, in the era of online stores, especially in advanced countries like China, attention has moved towards detecting deceptive reviews Li et al. [3]. Spam detection in online reviews can be classified into content-based or user behavior-based techniques. Content-based analyzes textual features of reviews, linguistic patterns, and sentiment analysis. User behavior-based focuses on patterns like reviewing users' behavior, metadata, and social connections between reviewers Li et al. [3], Ennaouri and Zellou [11]. Usually, these techniques are combined to improve detection accuracy. Identifying spam reviews is a sophisticated task due to several challenges. One of the primary challenges is distinguishing between genuine and fake reviews. While counterfeit reviews may imitate the style and content of genuine ones, specific patterns such as overuse of promotional language or usually high volumes of reviews in short periods can provide clues Li et al. [3]. Language variability also helps in detection, as spam reviews may appear in multiple languages or use specific accents Ennaouri and Zellou [11]. Moreover, spammers continuously improve their techniques, making relying on static detection methods difficult. They may use techniques such as duplicating legitimate reviews or using an automated system to generate spam, which can evade traditional methods Bhuvaneshwari et al. [5]. The large number of online reviews presents a scalability challenge as manual moderation becomes unfeasible for platforms like Amazon, which hosts millions of products with billions of reviews. ML and DL are powerful technologies for overcoming the challenges in spam detection. Traditional ML models, such as NB and SVM, have been used widely to classify reviews based on a content-based technique Saumya and Singh [12]. However, these models often struggle with the complexity of spam patterns in large datasets Kalaivani et al. [13], and this will be discussed in the next subsection based on related work. DL models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and their hybrid variants perform superiorly in detecting spam reviews Ghourabi et al. [14], Deshai and Rao [15], Shahariar et al. [16]. These models can learn complex patterns in text and capture long relations between words, making them more effective at detecting spam. Additionally, attention mechanisms, such as the one used in self-attention-based models, have successfully identified key features of spam reviews by focusing on specific text parts Bhuvaneshwari et al. [5]. The combination of DL and traditional ML technologies offers promising solutions to detect spam in online reviews. The following section will examine existing research on spam reviews using both ML and DL technologies.

### A. Spam Review Detection

Many studies have discussed spam detection methods. The following subsections will discuss these studies in detail, starting with traditional ML. After that, DL studies showed promising improvement in solving this problem. “Table I” summarizes all the ten related works, from both traditional ML and DL subsections.

*1) Traditional machine learning approach for spam review detection*: Traditional ML approaches were used to solve the problem of detecting Spam reviews using classifiers such as NB, SVM, and Random Forest (RF) Ahsan et al. [17], Tripathy et al. [18]. These classifiers are usually used alongside feature selection techniques to detect fake reviews. In Kalaivani et al. [13], two traditional ML models were used to detect spam reviews. The first algorithm was SVM, while the second was NB. The dataset that was used is from Kaggle, with 20k reviews. By preprocessing the data before training the above-mentioned models, SVM achieved 76%, while NB achieved 84%. In Etaiwi and Naymat [19], the author tried a bunch of traditional ML algorithms, which are Gradient Boosted Trees (GBT), NB, RF, DT, and SVM. All these algorithms were used to train models on the Hotel Reviews dataset of 1600 reviews named Deceptive Opinion Spam Corpus (DOSC). Among all these models, SVM achieved the best accuracy with 85.5%. In Saeed et al. [20], many spam review detection approaches were evaluated. The paper proposed four detection approaches: a rule-based classifier, machine learning classifiers, a majority voting classifier, and a stacking ensemble classifier. All these approaches were trained and tested on two datasets, DOSC and Hotel Arabic Reviews Dataset (HADR). The stacking ensemble approach clearly outperformed the other approaches with

95.25% accuracy on the DOSC dataset and 99.98% on the HARD dataset by combining a rule-based classifier with a k-means classifier.

In Ibrahim et al. [21] the authors investigated the use of ensemble learning techniques to enhance spam review detection accuracy. They explored three classifiers: NB, SVM, and Logistic Regression (LR). They combined these algorithms to form an ensemble classifier. They used the Amazon dataset and got the best accuracy of 88.09%. Etaiwi and Awajan [22] explored how different feature selection methods impact the performance of spam review detection. They applied four ML algorithms: NB, SVM, DT, and RF. They used the DOSC dataset and got the best accuracy results of 87.31% with NB.

*2) Deep learning approaches for spam review detection*: Deep learning techniques have succeeded in enhancing the accuracy of spam detection. They are more effective than traditional ML approaches, which rely more on manually engineered features. DL models can automatically learn from complex patterns, which makes them suitable for distinguishing between truthful and deceptive reviews. CNN, RNN, and hybrid approaches like CNN-RNN are widely used in the field Zhao et al. [23]. They have the potential to discover long relations between words. This section presents studies that have proposed DL techniques to detect spam reviews.

Shahariar et al. [16] presented a multi-layer perceptron (MLP) framework to detect spam reviews in the YELP and DOSC datasets. They compared the DL model with traditional ML, like NB and SVM. Their findings showed that Long Short-Term Memory (LSTM) outperformed all other algorithms with 96.75% accuracy. Deshai and Rao [15] Proposed two hybrid models integrating CNN and LSTM for fake reviews. Also, they presented LSTM-RNN for fake ratings detection. Their results showed that CNN-LSTM and LSTM-RNN methods are the most efficient, with 93.09% accuracy, using a subset of the Amazon Product Reviews dataset.

TABLE I. STATE-OF-THE-ART ALGORITHMS FOR SPAM REVIEW DETECTION

| Paper | Year | Algorithm | Dataset | Best Accuracy |
|---|---|---|---|---|
| [13] | 2023 | NB, SVM | Review dataset from kaggle | 84% NB |
| [19] | 2017 | GBT, NB, RF, DT, SVM | DOSC | 85.5% SVM |
| [20] | 2019 | rule-based classifier, machine learning classifiers, majority voting classifier, stacking ensemble classifier | DOSC, HARD | 99.98%<br>rule-based +<br>k-means |
| [16] | 2019 | MLP, CNN, LSTM | YELP., DOSC | 96.75% LSTM |
| [15] | 2023 | CNN-LSTM, LSTM-RNN<br>hybrid | Amazon Dataset | 93.09%<br>LSTM-RNN |
| [14] | 2020 | CNN-LSTM hybrid | UCI dataset | 98.37%<br>CNN-LSTM |
| [25] | 2023 | NB, KNN, SVM, CNN, LSTM | YELP<br>DOSC | 94.88% LSTM |
| [21] | 2017 | NB, SVM, LR | Amazon Dataset | 88.09% ensemble |
| [26] | 2020 | LSTM Autoencoder | YouTube | - |
| [22] | 2017 | NB, SVM, DT, RF | DOSC | 87.31% NB |

Ghourabi et al. [14] proposed a hybrid CNN-LSTM model for detecting mixed text messages written in Arabic and English. They designed a model to let CNN capture n-gram features while LSTM is used to retain long-term information. They achieved an accuracy of 98.37% using a dataset from UCI repository Almeida et al. [24]. Singh et al. [25] Explored the effectiveness of DL models, especially for CNN and LSTM, the authors benchmarked ML models with DL. They emphasized the superior performance of deep learning models over traditional approaches for handling textual data. The datasets used in their study are YELP and DOSC. They got the best results using LSTM model with an accuracy of 94.88%. Saumya and Singh [26] Presented an unsupervised model for detecting spam reviews without requiring labeled data. The authors used a combination of LSTM networks and autoencoders to learn patterns of true reviews, allowing the models to distinguish reviews anomalies. They used a YouTube dataset, which includes reviews of popular videos. The authors did not use accuracy metrics in their study. The studies showed that DL approaches perform superiorly in spam review detection, especially hybrid ones. In the next section, the gaps in existing studies will be discussed.

## B. *Gaps in Existing Research*

Spam review detection is a hot topic that is widely covered. However, several advancements in the field, particularly in the DL field, introduce gaps and challenges that open the door to resolving this problem by applying new technologies. This section aims to discuss these challenges. These gaps need to be addressed and apply new technologies to resolve them. One primary challenge is the need for a labeled dataset Hussain et al. [27].

Hussain et al. [7] addressed this issue and solved it with Spam Review Detection using Behavioral Method (SRD-BM). However, new ML and DL technologies are needed to help in labeling the dataset and help researchers work on it in the future. Another significant challenge is the use of hybrid DL architecture on large datasets. Ghourabi et al. [14], Deshai and Rao [15], Shahariar et al. [16], Wayal and Bhandari [28] applied a hybrid method in small datasets. Applying this kind of architecture on large datasets requires vast computational power. Fortunately, hybrid DL architecture has been applied to the Amazon Product Review Dataset, “Table II” shows the detailed distribution of the dataset. This study will discuss the proposed work in detail in the methodology section.

## III. Methodology

This study demonstrated the application of three hybrid DL models for spam review detection. Each model was selected based on its ability to recognize complex patterns in large datasets and understand long-range word relations. Also, the effects of applying extensive text preprocessing will be discussed. In the following subsections, start by introducing the development environment. Then, the dataset and data preprocessing steps that convert human sentences to a readable format for the machine will be presented. After that, the feature selection process, the three models' architecture, and how each model is trained on the dataset will be described respectively.

### *C. Development Environment*

In this study, all the preprocessing, training, and evaluation processes are conducted using Google Colab Pro. It is a solution on the cloud provided by Google to access high computational resources, utilizing the NVIDIA A100 graphics processing unit (GPU), which helped accelerate the training process across the hybrid DL models. Colab provides a high Random-Access Memory (RAM) up to 83GB and Virtual Random-Access Memory (VRAM) 40GB. It will allow the ability to load a large volume of data and preprocess it before feeding it into the neural network. Additionally, Python is the programming language used to implement the entire pipeline. Many libraries have been utilized to support this process, including Pandas and NumPy for data manipulation and preprocessing. Natural Language Toolkit (NLTK) is used to prepare text data. Scikit-learn for splitting train and test data, as well as for evaluation metrics. TensorFlow is used to train neural networks. Finally, Matplotlib and Seborn were used to visualize the performance analysis.

### *D. Dataset*

The dataset used for this study was acquired from Amazon Product Reviews Hussain et al. [7], which contains 26.7 million reviews written by 15.4 million reviewers on 3.1 million products. The dataset has six categories, shown in “Table II”. The reviews cover many product categories, such as electronics, home and kitchen, and more, to ensure that the spam detection models are exposed to various review patterns. The dataset's original source was unlabeled. However, Hussain et al. [7] did excellent work by labeling it using SRD-BM technique. The SRD-BM utilizes rich of behavioral features in the dataset to identify the spam and non-spam reviews, then labeling the data. In this study, the labeled dataset by the SRD-BM method was used, then preprocessing steps explained in the next section are applied.

### *E. Data Preprocessing*

Preprocessing is an essential step to minimize the noise of the data and transform the raw text into a format that can befed into the neural networks. However, applying the extensive preprocessing steps may upgrade or degrade the model performance HaCohen-Kerner et al. [29]. In this study, two different preprocessing steps were applied to hybrid DL models, one with extensive preprocessing steps that change the original text, another with a few steps that retain the original text. “Fig. 1.” shows the preprocessing steps used in this study. For the extensive preprocessing steps, the text is converted to lowercase to standardize the input and reduce the complexity caused by case differences. This step will convert words like “Product” to “product”, so they are treated as the same token. Additionally, all the punctuation marks are removed to avoid unnecessary symbols, which may add noise to the data. Another common concept in Natural Language Processing(NLP) was applied, which is Tokenization that converts text into numerical sequences to retain the frequent words in the dataset T. Limisiewicz et al. [30]. This will help reduce the noise in data and improve model efficiency. After that, Stop Words such as "is", "the", and "a" were removed. Also, it is essential to reduce every word to its root. Words like "playing" and "Played" will be reduced to "play" using a technique called Lemmatization. Also, to uniform the sequence length, Padding was used with configuration of 200 tokens in each sequence. Finally, an Embedding layer was added. It simply converts the integer sequences into dense vectors. This will allow the DL models to learn the relationship between words during the training Tegene et al. [31]. It is important to note that all these steps were applied to all the hybrid DL techniques to avoid biases in the benchmark. For the fewer preprocessing steps, only three steps were applied which are Tokenization, Padding, and Embedding. Also, for the max words parameter. Two configurations were applied, one with 10,000 and the other with 200,000 max words, this choice is due to the majority of words inside the dataset appearing very infrequently.

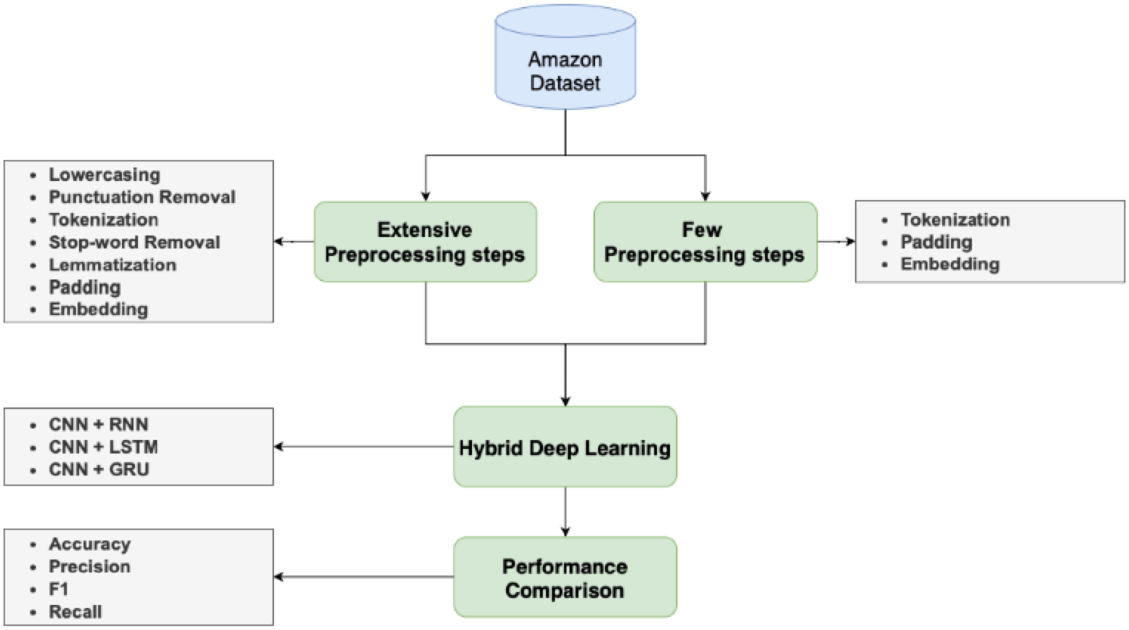


Fig. 1. Proposed system architecture.

### *F. Spam Review Detection*

The methodology for detecting spam reviews involves designing and evaluating hybrid deep learning models. These models aim to extract both local features and long-term dependencies from textual data, leveraging the strengths of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU).

Since many people rely on reviews before purchasing products, detecting spam reviews can improve customer experience by protecting them from being scammed. Hence, this study intends to compare multiple hybrid DL approaches to reach the best classifier in this matter which are: CNN-LSTM, CNN-RNN, and CNN-GRU. Combining CNN with LSTM, RNN, and Gated Recurrent Unit (GRU) will give the ability to extract features by learning the local patterns and n-grams using the CNN layer Zhou et al. [32]. For CNN-based hybrids, the CNN component identifies key local features. Combining CNN-LSTM will make the classifier capture long-term patterns and model the temporal dependencies using the LSTM component Greff et al. [33]. Similarly, GRU can also be combined with CNN (CNN-GRU) to use its efficient gating mechanisms that will help avoid vanishing gradient problems Cho et al. [34]. Another component that can be combined with CNN is RNN (CNN-RNN), which captures context and temporal relationships in the text to help detect subtle patterns. “Table III” presents hyperparameters configuration of the implemented models.

The dataset used in this study is the Amazon Product Review Dataset, containing 26.7 million reviews across six categories. Each review $x_i$ is labeled as spam ( $y_i = 1$) or non-spam ( $y_i = 0$ ). The raw text reviews are preprocessed to convert them into numerical representations. Two preprocessing strategies were compared: minimal and extensive preprocessing. Minimal preprocessing retained most of the text structure, while extensive preprocessing involved steps like lowercasing, stop word removal, punctuation removal, and lemmatization.

TABLE II. AMAZON PRODUCT REVIEW DATASET

| Category | Total Reviews | Total Reviewers | Total Products |
|---|---|---|---|
| Cell Phones and Accessories | 3,446,396 | 2,260,636 | 319,652 |
| Clothing, Shoes, and Jewellery | 5,748,260 | 3,116,944 | 1,135,948 |
| Electronics | 7,820,765 | 4,200,520 | 475,910 |
| Home and Kitchen | 4,252,723 | 2,511,106 | 410,221 |
| Sports and Outdoor | 3,267,538 | 1,989,985 | 478,846 |
| Toys and Games | 2,251,775 | 1,342,419 | 327,653 |
| **Total** | **26,787,457** | **15,421,610** | **3,148,230** |

The input dataset can be represented as:

$$D = \{(x_i, y_i) |\ xi\ \in R^d,\ y_i \in \{0,1\}, i = 1,2,..,n\} \qquad (1)$$

Where $d$ is the sequence length, and $n$ is the total number of reviews. Each review $xi$ is tokenized and padded to ensure uniform length. The tokens are then passed through an embedding layer, which maps them into dense vector representations:

$$e_i = embedding(x_i), e_i \in R^{d \times n} \qquad (2)$$

Where, the parameter $m$ is the embedding dimension, and $d$=200 is the maximum sequence length.

The CNN layer is applied to extract local patterns and n-grams from the embedding vectors. The convolution operation is defined as:

$$c_{i,j} = ReLU(W_{CONV} \times e_{i,j} + b_{Conv}) \quad (3)$$

Where, the parameter $W$conv and $b$conv are the convolutional filter weights and biases. In addition, the operator (×) represents the convolution operation. ReLU introduces non-linearity. The result, $ci$ is a feature map containing extracted local patterns. To capture long-term dependencies in the text, the feature map $ci$ is fed into an LSTM or GRU layer.

$$h_t = LSTM(C_t, h_{t-1}, C_{t-1}) \quad (4)$$

And for GRU

$$h_t = GRU(C_t, h_{t-1}) \quad (5)$$

Here, the $ht$ parameter represents the hidden state at time $t$, which encodes sequential dependencies. The final hidden state $h$ from the recurrent layer is passed through a fully connected layer to predict the probability of a review being spam:

$$p(yi = 1 \mid xi) = \sigma(Wfc \cdot hi + bfc) \quad (6)$$

Where, the parameter $\sigma$ is the sigmoid activation function. Also, the parameter $W$fc and $b$fc are the weights and biases of the fully connected layer. The binary cross-entropy loss function is used to optimize the model:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(p(\, yi = 1 \mid xi\,)\right) + (1 - yi) logp(\, 1 - yi = 1 \mid xi\,)\right] \quad (7)$$

### G. Evaluation Metrics

In our study, four evaluation metrics have been used to evaluate the models. These evaluation metrics depend on, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), defined as follows: TP is the number of correctly identified spam. TN is the number of reviews correctly identified as non-spam. FP is the number of non-spam reviews incorrectly identified as spam. FN is the number of spam reviews incorrectly identified as non-spam. All proposed models were evaluated using several key metrics, including Accuracy, Precision, Recall, and F1 score. Accuracy provides an overall measure of the model's performance in both positive and negative Sokolova and Lapalme [35].

$$Accuracy(ACC) = (TP + TN)/(TP + TN + FN + FP)(8)$$

Precision is the evaluated proportion of correctly predicted as positive and shows model capability to avoid FP J. Davis and Goadrich [36], T. Saito and Rehmsmeier [37].

$$Precision(PR) = TP/(TP + FP) \quad (9)$$

Recall or, in other words, sensitivity, is the ratio of correctly predicted positives to all the actual positives J. Davis and Goadrich [36], T. Saito and Rehmsmeier [37].

$$Recall(RE) = TP/(TP + FN) \quad (10)$$

F1 Score is the harmonic mean of precision and Recall; it balances the two and is widely used in scenarios where precision and Recall are important.

$$F1 - score = 2 \times (PR \times RE)/(\, PR + RE) \ (11)$$

These metrics are crucial to understanding the model's behavior across different situations. Therefore, they will all be used in the next section to benchmark different Hybrid DL models.

## IV. Results

In this section, results of the DL models with and without the text preprocessing step are discussed. The comparison is based on the four metrics mentioned earlier, accuracy, precision, recall and F1 score. Then, the section will present a discussion of the result.

### A. Results Analysis

The performance of hybrid DL models will be evaluated with and without some text preprocessing steps which are lowercasing, Stop Words removal, punctuation removal, and lemmatization. Also, two vocabulary sizes 10,000 and 200,000 max words will be used. "Table IV" shows all the results. Also, "Fig. 2." shows the models comparisons.

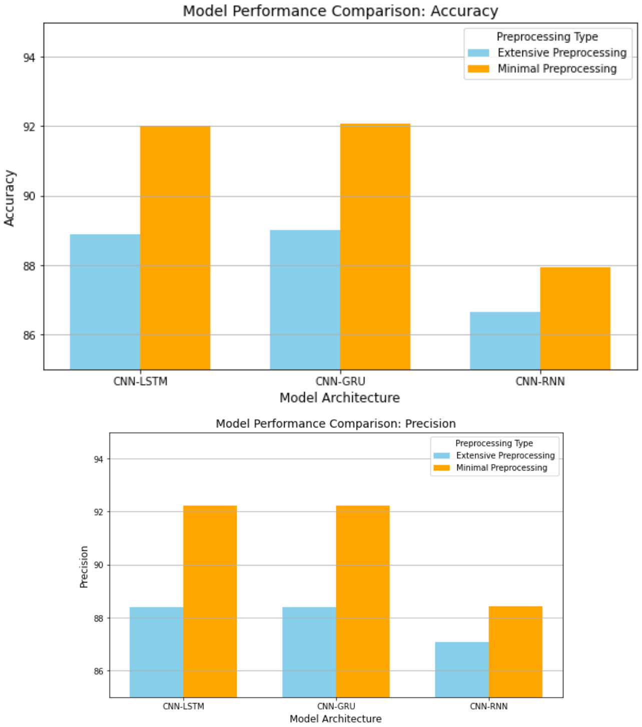
TABLE III. Models Configurations

| **Hyperparameters** | **CNN-LSTM** | **CNN-RNN** | **CNN-GRU** |
|---|---|---|---|
| Batch size | 128 | 128 | 128 |
| Dropout | 0.5 in each layer | 0.5 in each layer | 0.5 in each layer |
| Nodes | 128 in LSTM layer | 128 in RNN layer | 128 in GRU layer |
| Training split | 0.6 | 0.6 | 0.6 |
| Testing split | 0.2 | 0.2 | 0.2 |
| Validation split | 0.2 | 0.2 | 0.2 |
| Epoch | 10 | 10 | 10 |
| Optimizer | Adam | Adam | Adam |
| Loss function | Binary cross-entropy | Binary cross-entropy | Binary cross-entropy |
| Vector size | 128 | 128 | 128 |

Regarding the models trained with all the preprocessing steps, the CNN-LSTM achieved the highest performance with an accuracy of 88.88%, a precision of 88.39%, a recall of 89.52%, and an F1Score of 88.95%. The CNN-GRU model has a nearby result with an accuracy of 89%, a precision of 88.39%, a recall of 89.12%, and an F1Score of 88.75%. The CNN-RNN has the lowest performance with an accuracy of 86.64%, a precision of 87.07%, a recall of 86.05%, and an F1Score of 86.56%. All these results show that using a max word of 200,000 gives a better result than 10,000. When the models were trained by eliminating some preprocessing steps which are lowercasing, Stop Words removal, punctuation removal, and lemmatization, the CNN-LSTM also gave the best performance with an accuracy of 92%, a precision of 92.22%, a recall of 91.73%, and an F1Score of 91.98%. The CNN-GRU similarly performed well, accuracy of 92.08%, a precision of 92.22%, a recall of 91.19%, and an F1-score of 92.07%. While the CNN-RNN has the worst performance with an accuracy of 87.93%, a precision of 88.43%, a recall of 87.28%, and an F1Score of 87.85%.

The results showed that CNN-LSTM and CNN-GRU architectures give better performance without applying extensive preprocessing steps. Among all the models, CNN-LSTM with 200,000 max words has the best overall performance compared to all the other architectures. This shows that the combination of CNN and LSTM layers with a large vocabulary is very effective in detecting spam reviews, while keeping the original text. This explains the ability of CNN-LSTM to capture both local patterns and long-term dependencies, which gives importance to understanding relationships across multiple words in a text. This result is aligned with recent studies on text classification using CNN-LSTM architecture Bhuvaneshwari et al. [5] Sagnika et al. [38].

Etaiwi and Naymat [19] showed that using many preprocessing steps affects the overall performance of spam review classification when using ML models. As in this study, the result of applying many preprocessing steps on hybrid DL models will affect the performance. When comparing the models based on the vocabulary size, 200,000 max words consistently performed better. This indicates giving the hybrid DL models a larger vocabulary helps capture more complicated relationships between words, which leads to better performance. Although the study showed a promising result, some limitations were faced. One major constraint was the use of Google Colab Pro. It is a paid service that allows you to use high computational resources, such as A100 GPUs and a high mount of RAM, with a certain number of compute units. This restricted the ability to empiric many models and try different hyperparameters. For that reason, it is recommended that future studies explore additional hyperparameter tuning to further improve model performance, as well as experimenting with pre-trained models like BERT to improve the ability to capture both local and contextual word representation. It can also help in transferring knowledge from one domain to another.



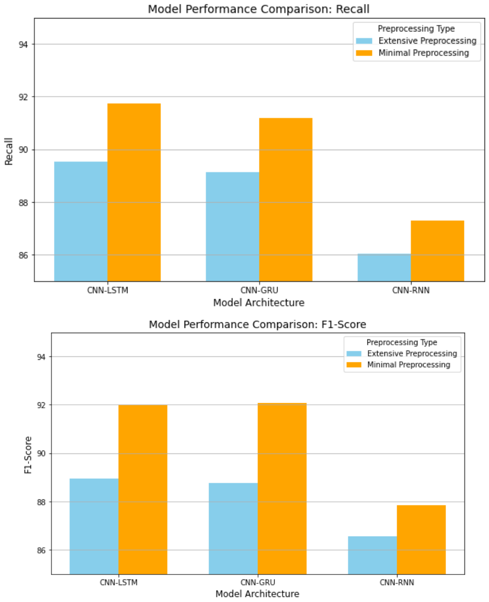


Fig. 2. Models comparisons.

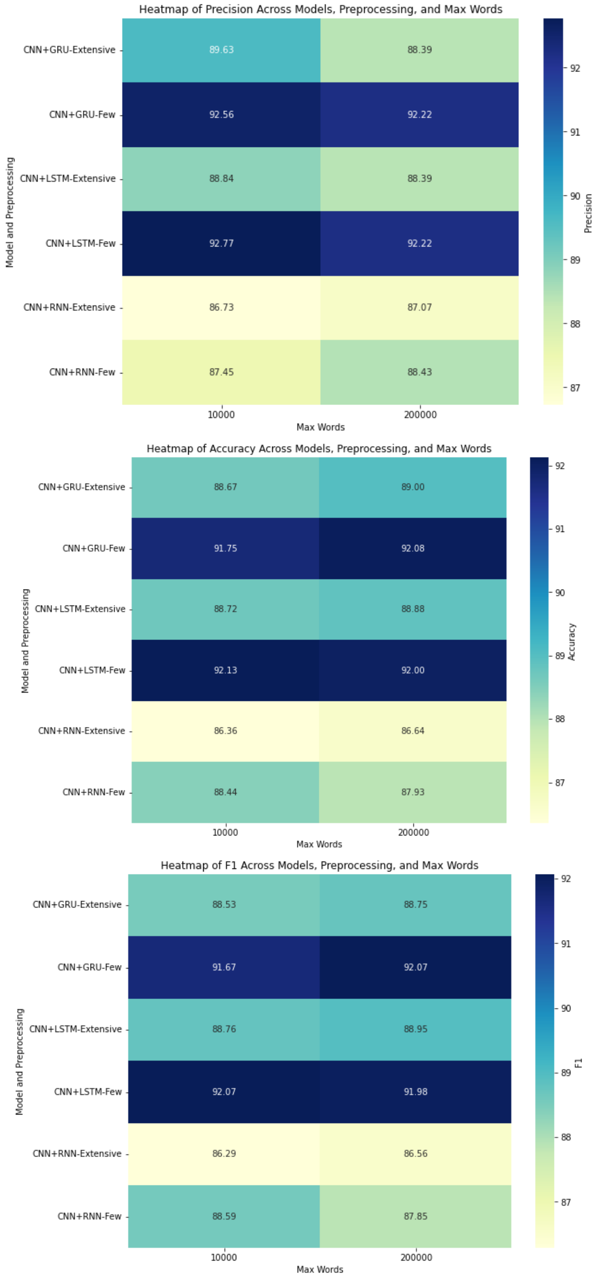
TABLE IV. MODELS RESULTS

| Model | Preprocessing | Max Words | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|---|---|
| CNN+LSTM | Few | 10,000 | 92.13 % | 92.77 % | 92.07 % | 91.38 % |
| CNN+RNN | Few | 10,000 | 88.44 % | 87.45 % | 88.59 % | 89.76 % |
| CNN+GRU | Few | 10,000 | 91.75 % | 92.56 % | 91.67 % | 90.80 % |
| CNN+GRU | Few | 200,000 | 92.08 % | 92.22 % | 92.07 % | 91.19 % |
| CNN+LSTM | Few | 200,000 | 92% | 92.22 % | 91.98 % | 91.73 % |
| CNN+RNN | Few | 200,000 | 87.93 % | 88.43 % | 87.85 % | 87.28 |
| CNN+GRU | Extensive | 10,000 | 88.67 % | 89.63 % | 88.53 % | 87.46 % |
| CNN+LSTM | Extensive | 10,000 | 88.72 % | 88.84 % | 88.76 % | 89.07 % |
| CNN+RNN | Extensive | 10,000 | 86.36 % | 86.73 % | 86.29 % | 85.85 % |
| CNN+GRU | Extensive | 200,000 | 89% | 88.39 % | 88.75 % | 89.12 % |
| CNN+LSTM | Extensive | 200,000 | 88.88 % | 88.39 % | 88.95 % | 89.52 % |
| CNN+RNN | Extensive | 200,000 | 86.64 % | 87.07 % | 86.56 % | 86.05 % |

## V. DISCUSSIONS

The experimental results revealed that hybrid deep learning models are effective in spam review detection, specifically for CNN-LSTM and CNN-GRU architectures. A simple CNN-LSTM model with very little preprocessing, and vocabulary of 200,000 outperformed others on a consistent basis. This shows that the model is capable of capturing rival patterns locally and the long-term dependence as observed by Bhuvaneshwari et al. Sequential models like LSTM: LSTMs have proven to be the backbone of text classification in various tasks Bhuvaneshwari et al. [5]. CNN-GRU model also provided competitive performance, which further confirms that the combined structures are proven efficient in these applications.

Our hybrid deep learning models have performed significantly higher than the traditional machine learning approaches (Support Vector Machines (SVM) and Naive Bayes (NB)). Studies, for example Etaiwi and Naymat [19], highlighted SVM performance with smaller datasets like DOSC, achieving high accuracies (even up to 85.5%). On the much larger dataset of the Amazon Product Review, our CNN-LSTM model reached 92% accuracy, while ML models underperformed dramatically. This large increase demonstrates that the merits of deep learning to effectively learn complex patterns in large-scale data.



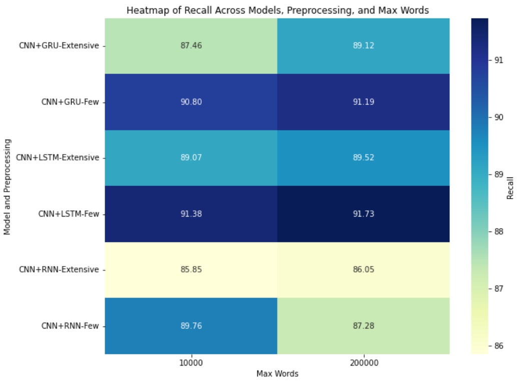


Fig. 3. Models Comparisons in terms of heatmap by different parameter sizes.

Fig. 3 represented the heatmaps for the performance of different models (CNN+LSTM, CNN+GRU, CNN+RNN) across key metrics: Accuracy, Precision, F1, and Recall. The heatmaps provide a clear comparison for the proposed hybrid system based on preprocessing type and vocabulary size, highlighting the effectiveness of minimal preprocessing and large vocabulary sizes.

Also, it is in accordance with the results of Ghourabi et al. [14] CNN text classification which has shown that CNN-LSTM hybrid could successfully learn n-gram features as well as long-term dependencies. However, they performed their study from a smaller dataset with a restricted size of vocabulary. This study used larger vocabularies than used in the previous study to train the models, and our results appear to expand on these findings, suggesting that larger vocabularies produce even better models, at least on some data sets, this further boost in size compares to a level of detail in relationships examinable within the text.

Fig. 4 bar chart highlights that CNN+LSTM with minimal preprocessing and a large vocabulary size (200k) achieves the highest accuracy, followed closely by CNN+GRU under similar conditions. Models with extensive preprocessing generally show lower accuracy. An interesting takeaway is the effect of preprocessing on model performance. As also mentioned by HaCohen-Kerner et al. [29] among others, aggressive preprocessing like lemmatization and stop word removal limited the models' ability to pick up any useful signals. On the other hand, less preprocessing helps the models keep the richness of the original text, which performed better. This is an important finding because it counters the widespread assumption that more preprocessing is always a good thing for model performance.

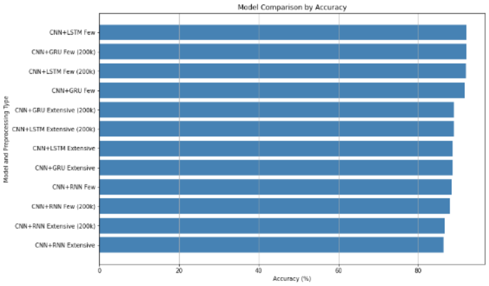


Fig. 4. A bar chart comparing the accuracy of different models and preprocessing configurations.

CNN-LSTM model has shown the best performance with less preprocessing steps and using large vocabulary size thus providing a very practical way for spam review detection. The benefit of preserving its original text is that this model can utilize its full potential to find hard to detect spam. This method is especially beneficial when the datasets are large, and context retention is essential for efficient classification. Moreover, this study shows that the proposed model is deployable efficiently under limited resources on cloud platforms such as Google Colab Pro.

Despite these optimistic results, there were several limitations to the study. However, the need for labeled datasets presents a critical issue because manual labeling is often a laborious, inconsistent process. Semi-supervised or Active Learning based techniques may also be a direction for future research to automate the labeling process as well as further explore the state-of-the-art discussed in the previous section for a better initial core for semi-supervised learning. For even better performance, pretrained models like BERT can be utilized, as they are able to learn much more complex relationships, as shown recently in the field of NLP.

Therefore, this study offers a solid foundation for hybrid deep learning models to detect spam reviews. The solution proposed not just increases detection accuracy but also provides a scalable approach with the use of dataset used in e-commerce enabling a better legitimate and deterring e-commerce platforms.

## VI. Conclusion and Future Work

This research has underlined the importance of combining many deep learning architectures to achieve optimal results in detecting spam reviews. It showed the capabilities of CNN and the sequential learning strength of LSTM and GRU. This contribution will help e-commerce platforms to build consumer trust by detecting spam reviews effectively. Another superior contribution of this paper is the impact of preprocessing steps on hybrid DL models' performance. Interestingly, when eliminating some of preprocessing steps the models performed better than those trained with all preprocessing steps. The combination of large datasets with hybrid DL models showed promising results in spam detection. However, the study identified a key limitation, the need for new labeled datasets for online spam reviews. As spamming techniques evolve, addressing this limitation in future work will encourage researchers to keep datasets updated for recent spam behaviors. Also, exploring recent ML techniques to automate the task of labeling the datasets is important. Methods such as semi-supervised learning or active learning could be implemented to get accurate datasets, reducing the dependence on manually labeled datasets. Furthermore, empiric hyperparameters and optimizers may further improve the performance of the models. Finally, the findings of this study indicate that the CNN-LSTM model using 200,000 max words outperformed other Hybrid DL models with an accuracy of 92%, a precision of 92.22%, a recall of 91.73%, and an F1Score of 91.98%.

## Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University for funding and supporting this work through Graduate Students Research Support Program.

## References

[1] H. A. Najada and X. Zhu, "iSRD: Spam review detection with imbalanced data distributions," in Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, USA: IEEE, Aug. 2014, pp. 553–560. doi: 10.1109/IRI.2014.7051938.

[2] Department for Business and Trade (DBT), "FAKE ONLINE REVIEWS RESEARCH," UK Government, London, UK, 2023. Accessed: Oct. 12, 2024. [Online]. Available: https://assets.publishing.service.gov.uk/media/6447c00c529eda000c3b03c5/fake-online-reviews-research.pdf

[3] Y. Li, Y. Liu, and C. Liu, "Research on Spam Review Detection: A Survey," in 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Harbin, China: IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/ICNC-FSKD59587.2023.10281054.

[4] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, "Unsupervised User Behavior Representation for Fraud Review Detection with Cold-Start Problem," in Advances in Knowledge Discovery and Data Mining, vol. 11439, Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, and S.-J. Huang, Eds., in Lecture Notes in Computer Science, vol. 11439. , Cham: Springer International Publishing, 2019, pp. 222–236. doi: 10.1007/978-3-030-16148-4_18.

[5] P. Bhuvaneshwari, A. N. Rao, and Y. H. Robinson, "Spam review detection using self attention based CNN and bi-directional LSTM," Multimed. Tools Appl., vol. 80, no. 12, pp. 18107–18124, May 2021, doi: 10.1007/s11042-021-10602-y.

[6] M. N. Rizali, M. M. Rosli, and N. A. S. Abdullah, "Spam Review Detection in E-Commerce Using Machine Learning," in 2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia: IEEE, Aug. 2024, pp. 189–193. doi: 10.1109/IICAIET62352.2024.10730100.

[7] N. Hussain, H. Turab Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam Review Detection Using the Linguistic and Spammer Behavioral Methods," IEEE Access, vol. 8, pp. 53801–53816, 2020, doi: 10.1109/ACCESS.2020.2979226.

[8] A. Jakupov, J. Longhi, and B. Zeddini, "The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses," Languages, vol. 9, no. 1, p. 10, Dec. 2023, doi: 10.3390/languages9010010.

[9] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," Proc. Int. AAAI Conf. Web Soc. Media, vol. 7, no. 1, pp. 175–184, Aug. 2021, doi: 10.1609/icwsm.v7i1.14400.

[10] M. S. Farooq, "Spam Review Detection:A Systematic Literature Review," Sep. 17, 2020. doi: 10.36227/techrxiv.12951077.v1.

[11] M. Ennaouri and A. Zellou, "Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review," J. Web Eng., Dec. 2023, doi: 10.13052/jwe1540-9589.2254.

[12] S. Saumya and J. P. Singh, "Detection of spam reviews: a sentiment analysis approach," CSI Trans. ICT, vol. 6, no. 2, pp. 137–148, Jun. 2018, doi: 10.1007/s40012-018-0193-0.

[13] P. Kalaivani, V. D. Raj, R. Madhavan, and A. P. Naveen Kumar, "Fake Review Detection using Naive Bayesian Classifier," in 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India: IEEE, Jun. 2023, pp. 705–709. doi: 10.1109/ICSCSS57650.2023.10169838.

[14] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A Hybrid CNN-LSTM Model for SMS Spam Detection in Arabic and English Messages," Future Internet, vol. 12, no. 9, p. 156, Sep. 2020, doi: 10.3390/fi12090156.

[15] "Deep Learning Hybrid Approaches to Detect Fake Reviews and Ratings," J. Sci. Ind. Res., vol. 82, no. 01, Jan. 2023, doi: 10.56042/jsir.v82i1.69937.

[16] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. B. Hassan, "Spam Review Detection Using Deep Learning," in 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Oct. 2019, pp. 0027–0033. doi: 10.1109/IEMCON.2019.8936148.

[17] M. N. I. Ahsan, T. Nahian, A. A. Kafi, Md. I. Hossain, and F. M. Shah, "An ensemble approach to detect review spam using hybrid machine learning technique," in 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh: IEEE, Dec. 2016, pp. 388–394. doi: 10.1109/ICCITECHN.2016.7860229.

[18] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Syst. Appl., vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.

[19] W. Etaiwi and G. Naymat, "The Impact of applying Different Preprocessing Steps on Review Spam Detection," Procedia Comput. Sci., vol. 113, pp. 273–279, 2017, doi: 10.1016/j.procs.2017.08.368.

[20] R. M. K. Saeed, S. Rady, and T. F. Gharib, "An ensemble approach for spam detection in Arabic opinion texts," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 1, pp. 1407–1416, Jan. 2022, doi: 10.1016/j.jksuci.2019.10.002.

[21] A. J. Ibrahim, M. M. Siraj, and M. M. Din, "Ensemble classifiers for spam review detection," in 2017 IEEE Conference on Application, Information and Network Security (AINS), Miri: IEEE, Nov. 2017, pp. 130–134. doi: 10.1109/AINS.2017.8270437.

[22] W. Etaiwi and A. Awajan, "The Effects of Features Selection Methods on Spam Review Detection Performance," in 2017 International Conference on New Trends in Computing Sciences (ICTCS), Amman: IEEE, Oct. 2017, pp. 116–120. doi: 10.1109/ICTCS.2017.50.

[23] S. Zhao, Z. Xu, L. Liu, and M. Guo, "Towards Accurate Deceptive Opinion Spam Detection based on Word Order-preserving CNN," Mar. 19, 2018, arXiv: arXiv:1711.09181. Accessed: Oct. 19, 2024. [Online]. Available: http://arxiv.org/abs/1711.09181

[24] T. A. Almeida, T. P. Silva, I. Santos, and J. M. Gómez Hidalgo, "Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering," Knowl.-Based Syst., vol. 108, pp. 25–32, Sep. 2016, doi: 10.1016/j.knosys.2016.05.001.

[25] D. Singh, M. Memoria, and R. Kumar, "Deep Learning Based Model for Fake Review Detection," in 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India: IEEE, May 2023, pp. 92–95. doi: 10.1109/InCACCT57535.2023.10141826.

[26] S. Saumya and J. P. Singh, "Spam review detection using LSTM autoencoder: an unsupervised approach," Electron. Commer. Res., vol. 22, no. 1, pp. 113–133, Mar. 2022, doi: 10.1007/s10660-020-09413-4.

[27] N. Hussain, H. Turab Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam Review Detection Techniques: A Systematic Literature Review," Appl. Sci., vol. 9, no. 5, p. 987, Mar. 2019, doi: 10.3390/app9050987.

[28] G. Wayal and V. Bhandari, "Enhancing Review Spam Detection with a Hybrid Approach Integrating Association Rule Mining and Convolutional Neural Networks," in 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India: IEEE, Sep. 2024, pp. 1–8. doi: 10.1109/ACROSET62108.2024.10743414.

[29] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," PLOS ONE, vol. 15, no. 5, p. e0232525, May 2020, doi: 10.1371/journal.pone.0232525.

[30] T. Limisiewicz, J. Balhar, and D. Marecˇek, "Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages".

[31] A. Tegene, Q. Liu, Y. Gan, T. Dai, H. Leka, and M. Ayenew, "Deep Learning and Embedding Based Latent Factor Model for Collaborative Recommender Systems," Appl. Sci., vol. 13, no. 2, p. 726, Jan. 2023, doi: 10.3390/app13020726.

[32] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," 2015, arXiv. doi: 10.48550/ARXIV.1511.08630.

[33] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," 2015, doi: 10.48550/ARXIV.1503.04069.

[34] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014, arXiv. doi: 10.48550/ARXIV.1406.1078.

[35] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[36] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd international conference on Machine learning - ICML '06, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[37] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[38] S. Sagnika, B. S. P. Mishra, and S. K. Meher, "An attention-based CNN-LSTM model for subjectivity detection in opinion-mining," Neural Comput. Appl., vol. 33, no. 24, pp. 17425–17438, Dec. 2021, doi: 10.1007/s00521-021-06328-5.