Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2024.050452

ARTICLE

A Hierarchical Two-Level Feature Fusion Approach for SMS Spam Filtering

Hussein Alaa Al-Kabbi[1,2], Mohammad-Reza Feizi-Derakhshi[1,*] and Saeed Pashazadeh[3]

[1]Computerized Intelligence Systems Laboratory, Department of Computer Engineering, University of Tabriz, Tabriz, 51368, Iran

[2]Ministry of Education Iraq, General Direction of Vocational Education, Al-Najaf, 54001, Iraq

[3]Department of Computer Engineering, University of Tabriz, Tabriz, 51368, Iran

*Corresponding Author: Mohammad-Reza Feizi-Derakhshi. Email: mfeizi@tabrizu.ac.ir

Received: 07 February 2024 Accepted: 27 May 2024 Published: 06 September 2024

ABSTRACT

SMS spam poses a significant challenge to maintaining user privacy and security. Recently, spammers have employed fraudulent writing styles to bypass spam detection systems. This paper introduces a novel two-level detection system that utilizes deep learning techniques for effective spam identification to address the challenge of sophisticated SMS spam. The system comprises five steps, beginning with the preprocessing of SMS data. RoBERTa word embedding is then applied to convert text into a numerical format for deep learning analysis. Feature extraction is performed using a Convolutional Neural Network (CNN) for word-level analysis and a Bidirectional Long Short-Term Memory (BiLSTM) for sentence-level analysis. The two-level feature extraction enables a complete understanding of individual words and sentence structure. The novel part of the proposed approach is the Hierarchical Attention Network (HAN), which fuses and selects features at two levels through an attention mechanism. The HAN can deal with words and sentences to focus on the most pertinent aspects of messages for spam detection. This network is productive in capturing meaningful features, considering both word-level and sentence-level semantics. In the classification step, the model classifies the messages into spam and ham. This hybrid deep learning method improve the feature representation, and enhancing the model's spam detection capabilities. By significantly reducing the incidence of SMS spam, our model contributes to a safer mobile communication environment, protecting users against potential phishing attacks and scams, and aiding in compliance with privacy and security regulations. This model's performance was evaluated using the SMS Spam Collection Dataset from the UCI Machine Learning Repository. Cross-validation is employed to consider the dataset's imbalanced nature, ensuring a reliable evaluation. The proposed model achieved a good accuracy of 99.48%, underscoring its efficiency in identifying SMS spam.

KEYWORDS

SMS spam detection; hierarchical attention network; text classification; natural language processing

1 Introduction

Short Message Service (SMS) spam has become prevalent in mobile communication systems. Based on a recent publication by Slicktext [1], the usage of SMS is widespread, with approximately five billion people utilizing this communication channel. The number of SMS users is projected to

reach 5.9 billion by 2025. Unfortunately, the increased prevalence of SMS usage has also resulted in a surge in malicious activities such as spam and smishing. These activities inconvenience users and pose significant financial risks to individuals and businesses [2]. The primary objective of the senders behind these spam messages is to illicitly acquire personal or financial information via SMS, often through the inclusion of fraudulent content, malicious links, or malware.

The paper highlights three main categories of SMS-based spam: (i) General SMS spam, which includes unwanted messages used for bulk marketing and spreading false information; (ii) premium rate scams that deceive individuals into dialing high-cost numbers or registering for expensive services under pretenses, and (iii) phishing or smishing tactics, in which recipients are sent texts prompting them to contact specific numbers as a ploy to obtain sensitive data for nefarious objectives [3]. The detection of SMS spam becomes essential to maintain user privacy and security. Traditional rule-based and keyword-based methods for SMS spam detection have yet to prove sufficient to handle the evolving nature of spam messages, which often employ low computational techniques to evade detection [4].

Artificial intelligence tools have been developed to assist in various fields, including healthcare [5], social networks [6], network security [7], and other real-life applications. Natural Language Processing (NLP) and deep learning advancements have opened up new opportunities for improving SMS spam detection in recent years. Deep learning models, such as recurrent neural networks (RNN) and transformer-based architectures [8], have performed remarkably in various text classification tasks. These models can effectively capture text messages' semantic and contextual information, enabling more accurate spam detection [9]. Spammers have begun adopting novel writing styles to evade SMS spam detection approaches. By modifying linguistic patterns, grammar usage, and content structure, spammers aim to create messages that bypass filters [10]. This dynamic shift in writing techniques presents a significant challenge for existing spam detection methods, as they often rely on historical data and recognizable patterns. In response to these evolving tactics, this paper presents a novel approach to addressing this challenge through a two-level SMS spam detection system based on words and sentences. Recognizing that words make sentences and sentences make documents [11], our approach allows us to detect all attempts by fraudsters to bypass spam detection tools, such as using different words or changing the location of words within sentences.

The proposed two-level SMS spam detection method consists of five steps: preprocessing, RoBERTa word embedding, two-level feature extraction using CNN for word level and BiLSTM for sentence level, feature fusion and selection through Hierarchical Attention Network (HAN), and classification. The innovative part of our approach is using a Hierarchical Attention Network (HAN) [12]. Which innovatively integrate the characteristics extracted from (word-level) and (sentence-level) dimensions. This synthesis is not merely combinative but is competitively evaluative, with the HAN's attention mechanisms intricately assessing the prominence of each linguistic element. We evaluate the performance of the proposed model on the UCI SMS dataset [13], which contains more than 5000 labeled SMS messages. The results demonstrate the effectiveness of the proposed method in accurately detecting SMS spam. To summarize the research insights, this study focuses on the following Research Questions (RQs):

- **RQ1:** How does the integration of a two-level feature fusion approach improve the accuracy and robustness of SMS spam detection, especially for short texts lacking contextual information?
- **RQ2:** How can we confirm that the two-level method excels in SMS spam detection?

The main contributions of this paper and possible Answers to the Research Questions (ARQs) are summarized as follows:

- **ARQ1:** Hierarchical Feature Integration—The proposed two-level model innovatively combines feature extraction at both the word and sentence levels through a competitive fusion mechanism within the Hierarchical Attention Network. This allows for a more nuanced representation of textual features, which is critical for accurate SMS spam detection.
- **ARQ2:** Superior Performance—Empirical evaluations on the UCI SMS dataset demonstrate that the model achieves state-of-the-art performance, with an accuracy of 99.48%, indicating its capability to manage various types of SMS spam.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in SMS spam detection. Section 3 briefly explains the techniques used in the proposed method. Section 4 describes the methodology in detail. Section 5 presents the experimental setup, and Section 6 presents the evaluation results. Section 7 for the desiccation, and Finally, Section 8 concludes the paper, highlighting the contributions and future work directions.

2 Related Work

The field of SMS spam detection has evolved significantly, witnessing a transition from rule-based methods to deep learning techniques. In the early stages, rule-based approaches utilized predefined patterns and keywords to identify spam messages. However, these methods had limitations in adapting to evolving spam tactics and handling noisy data. We divided the related work into two groups.

2.1 Traditional Methods

SMS spam filtering has been a long-standing research subject, with traditional machine learning methods like SVM [14], Naive Bayes [15], and decision trees [16] being proposed. However, these approaches need complex feature engineering and have difficulty dealing with noisy or imbalanced data [17]. Most of these studies aimed to improve the classifier's architecture rather than giving priority to feature extraction. A new method was recently presented by Ali et al. [18], who proposed a unique combination of traditional machine-learning techniques for SMS spam detection. Specifically, it uses Multiple Linear Regression (MLR) for feature weighting and an Extreme Learning Machine (ELM) for classification. This method achieved an accuracy of 98.7% on the UCI SMS dataset. Also, Hosseinpour et al. [19] proposed an ensemble learning method based on logistic regression and random forest algorithms. The ensemble learning approach achieved an accuracy of 98.06%. Pudasaini et al. [20] combined Relevance Vector Machine (RVM), SVM, Naive Bayes, and KNN, with a majority vote determining the final output. This research conducts a thorough comparative analysis of text classification algorithms for effective spam detection, emphasizing TF-IDF vectorization for preprocessing. The RVM stands out, achieving an F1-score of 97.51% in the UCI spam SMS dataset.

2.2 Deep Learning Methods

With the emergence of deep learning, researchers have explored new approaches to tackle SMS spam detection challenges. Liu et al. [21] introduced a modified version of the Transformer for SMS spam detection. Their comprehensive analysis encompassed various existing methods and evaluated them against datasets like SMS Spam Collection v.1 and UtkMl's Twitter dataset [22]. This method achieved an accuracy of 98.92% for the SMS spam dataset. Srinivasarao et al. [23] present a new model in text mining for spam and ham message differentiation. It introduces a fuzzy-based recurrent neural network with Harris Hawk optimization (FRNN-HHO) for classification. Post-classification sentiment analysis is performed to improve accuracy. In experimental evaluation using SMS, Email, and spam-assassin datasets, this method achieved an accuracy of 98.61% for SMS spam detection.

Giri et al. [24] propose four neural network models (CNN BUNOW, CNN-LSTM BUNOW, CNN GloVe, and CNN-LSTM GloVe) for distinguishing spam from non-spam messages using SMS Spam Collection v.1 dataset. The models are trained and tested on different train-test splits. CNN-LSTM BUNOW performs best among four models with an accuracy of 99.04%, 99.01%, 98.92%, and 98.44% for 85%–15%, 80%–20%, 75%–25%, and 70%–30% train-test splits, respectively.

Debnath et al. [25] aim to address the SMS spam issue and improve detection accuracy. Various machine learning and deep learning models, including LSTM and BERT, are utilized on a UCI dataset to classify SMS spam. The proposed deep learning approach achieves high accuracy rates of 99.28% with BERT and 98.84% with LSTM. Ghourabi et al. [26] propose a deep learning model, CNN-LSTM, to detect SMS spam messages effectively. It combines CNN and LSTM to handle text messages for Arabic and English datasets. Experimental results demonstrate its performance, achieving an accuracy of 98.37%. Abayomi-Alli et al. [27] propose a deep learning approach that utilizes a Bidirectional Long Short-Term Memory (BiLSTM) model for SMS spam detection. The study involves two datasets: the ExAIS_SMS [28], a unique indigenous dataset, and the well-known UCI dataset. The proposed method leverages the distinctive characteristics of BiLSTM to achieve a high classification rate in detecting spam SMS messages. The BiLSTM attained an accuracy of 98.6% on the UCI SMS dataset. Wei et al. [29] propose a lightweight deep neural model called Lightweight Gated Recurrent Unit (LGRU) for SMS spam detection. They incorporate enhancing semantics retrieved from external knowledge (WordNet) to augment the understanding of SMS text inputs for better classification. LGRU achieved an accuracy of 98.87%. Ardeshir-Larijanie et al. [30] introduce a novel integration of hybrid classical-quantum transfer learning with NLP utilizing a pre-trained BERT model and a variational quantum circuit for text classification. This approach achieved an overall AUC-ROC of 95%.

Some papers deal with SMS spam detection using private datasets such as [31] or using local language datasets [32,33].

3 Background

The "Background" section of the paper provides a concise overview of the deep learning approaches integral to the proposed model.

3.1 RoBERTa

The Robustly Optimized BERT Pretraining Approach (RoBERTa) is a Large Language Model (LLM) Chabot trained to be more robust to adversarial training. It was developed by Facebook AI and released in 2019 [34]. RoBERTa is based on the BERT architecture but with several modifications, including more training data and longer training sequences. As a result of these modifications, RoBERTa has been shown to outperform BERT on several downstream tasks, including natural language inference (NLI), question answering (QA), and sentiment analysis [35]. RoBERTa is a powerful LLM that can be used for various tasks, such as generating text, translating languages, answering questions, summarizing text, and classifying text.

3.2 Word-Level CNN

Word-level CNN is a convolutional neural network (CNN) designed explicitly for text classification tasks. CNN is well-suited for text classification tasks because it can learn to extract features from text data that are relevant to the task [36]. CNN can learn complex features from word embedding, making it a powerful tool for text classification. They have been shown to achieve state-of-the-art results on various tasks, including sentiment analysis, spam detection, and topic classification. Word-level CNNs effectively detect SMS spam because they focus on capturing features from the crucial words in the message. They excel at understanding the significance of specific words and their combinations, allowing them to differentiate between spam and legitimate messages based on these word-level features. Additionally, word embedding enhances their ability to interpret the meaning of words in the context of SMS content [37].

3.3 BiLSTM

Bidirectional LSTM (BiLSTM) is a type of recurrent neural network (RNN) that can be used for text classification tasks. RNN is well-suited for text classification tasks because they can learn long-range dependencies in text data. Unlike other RNNs, BiLSTM processes text data in both forward and backward directions, allowing for a more comprehensive extraction of information [38]. BiLSTM first converts the text data into a sequence of word embedding. Word embedding is dense vector representations of words that capture the semantic and syntactic relationships between words in the sentence. The output of the BiLSTM is a sequence of hidden states. The hidden states contain information the BiLSTM has learned about the text data.

3.4 Hierarchical Attention Network (HAN)

The Hierarchical Attention Network (HAN) is a pivotal component of the suggested model for Text understanding. HAN operates at both word and sentence levels, allowing it to capture hierarchical relationships within text data. The model effectively highlights essential words and sentences within a message by employing attention mechanisms, enabling the acquisition of crucial information [39]. HAN utilizes a GRU (Gated Recurrent Unit) at the word level to capture sequential dependencies among words within a sentence. Attention mechanisms are subsequently applied to assign varying weights to individual words, reflecting their significance in the overall message representation. At the sentence level, another attention mechanism is deployed to assess the importance of each sentence in the message. This dual-level attention mechanism enables the model to prioritize sentences containing substantial information while filtering out irrelevant or less informative ones.

The collaboration of word-level and sentence-level attention within HAN provides a hierarchical representation that comprehensively captures the text's context and semantics at various levels. This makes HAN exceptionally effective in tasks such as sentiment analysis, fake news detection, and text summarization, where the hierarchical structure of the text is paramount for accurate predictions and meaningful interpretations. Fig. 1 illustrates the HAN architecture.

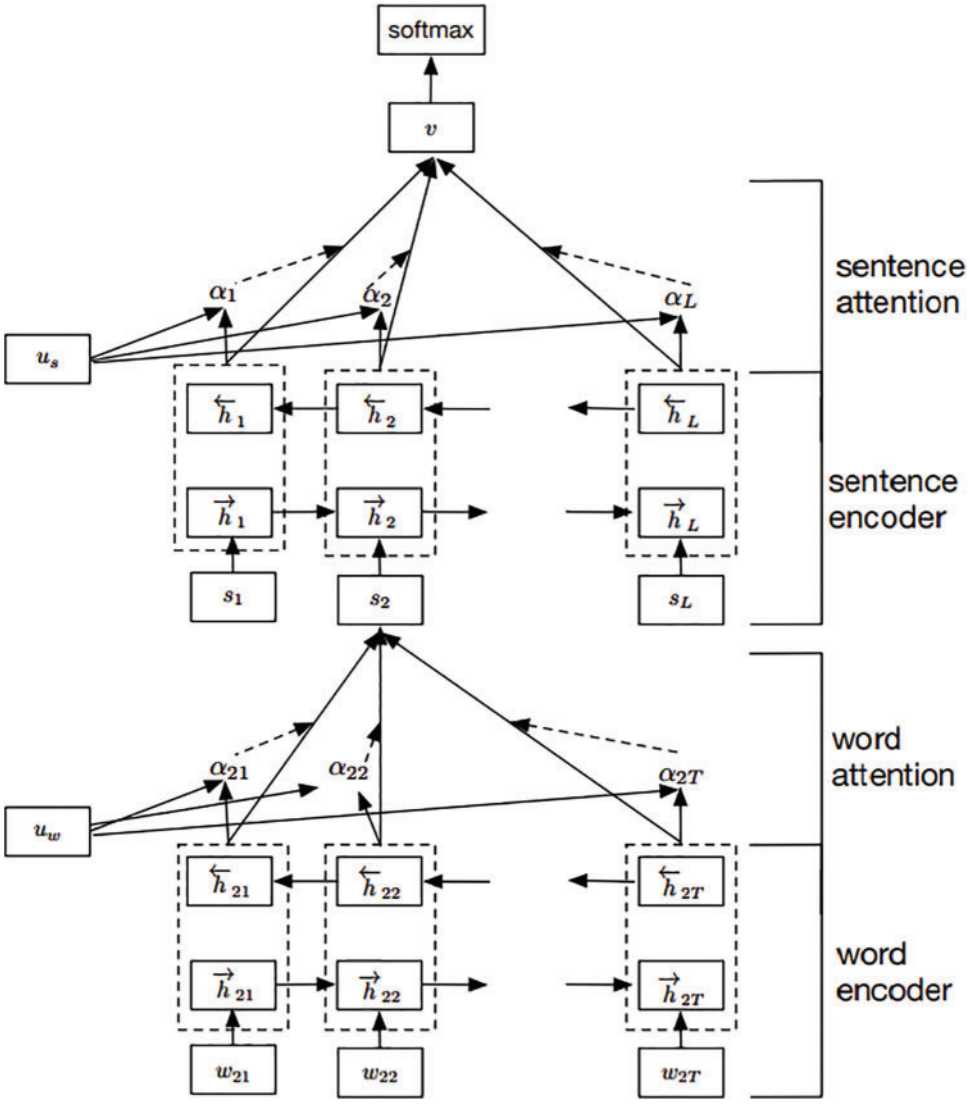

Figure 1: HAN architecture reprinted with permission from reference. [39]. Copyright 2019, ACM

4 Proposed Method

The proposed method in SMS spam detection consists of five main steps to achieve accurate classification. The method utilizes a combination of techniques to enhance performance. Fig. 2 shows the proposed model steps.

4.1 Preprocessing

The initial stage of our SMS spam detection method involves data preprocessing. This crucial process uses machine learning and deep learning models to read raw text messages for optimal analysis [9]. To enhance data quality, we undertake the subsequent preprocessing steps.

- Punctuation removal: Unnecessary punctuation marks and symbols are removed from the text messages. Removing punctuation enhances the focus on content and meaning, simplifying data representation and decreasing vocabulary size.
- Lowercasing: All text words are converted to lowercase for consistency, ensuring uniformity and standardized vocabulary usage, eliminating variations due to capitalization [40].
- Stop-word removal: Frequently occurring words such as "the," "is," and "a" (stop words) are excluded as they often lack substantial meaning. This step reduces data dimensionality and prevents interference in spam detection, emphasizing words with higher discriminatory significance [41].
- Removal: We excise symbols or characters that hold no essential role in the message's content, guaranteeing that the model remains steadfastly attuned to meaningful information. This encompasses the removal of hashtags and other extraneous symbols [42].
- Normalization: Text normalization assumes paramount significance in Short Message Services (SMS), where character limits are stringent, and senders often resort to shortcuts to economize on space and costs. This procedure involves transforming word variations, such as "u" to "you" and "2" to "to," into their standardized equivalents [43].

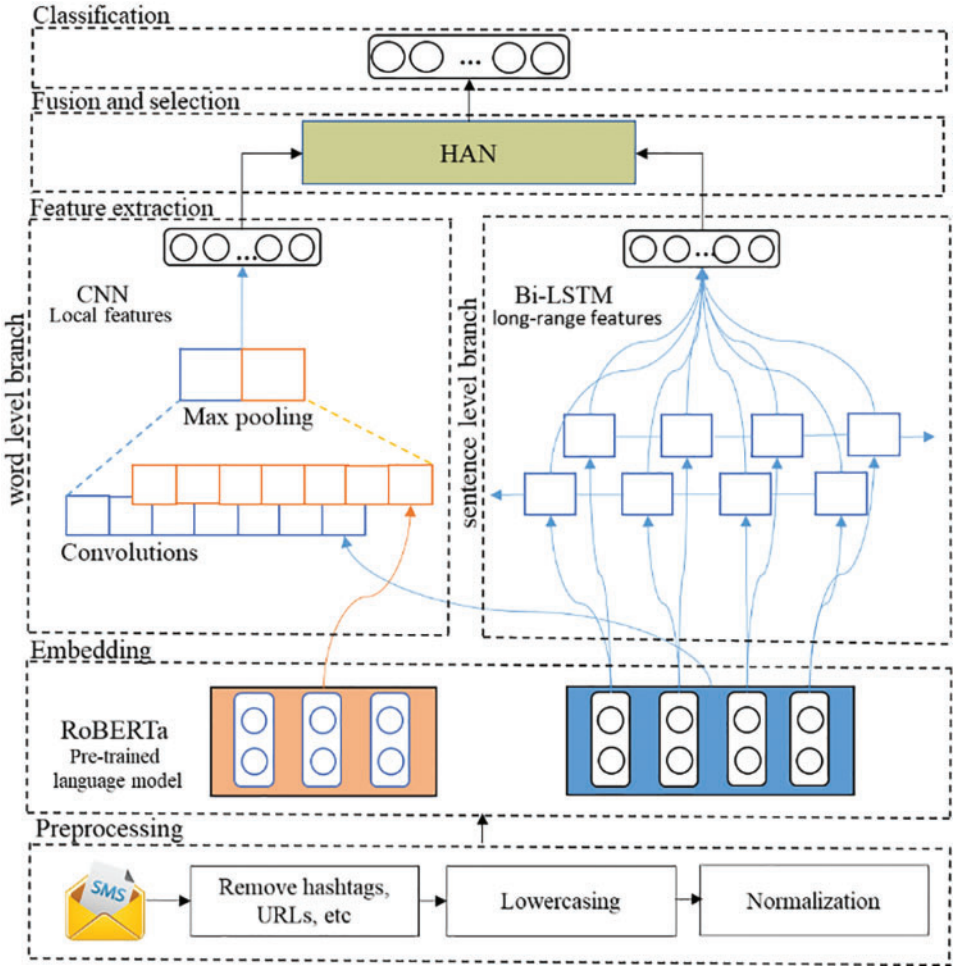

Figure 2: Proposed method framework structure

The text data is suitable for deep learning models by executing these preprocessing procedures. The numerical representations enable practical analysis and enhance our SMS spam detection approach.

4.2 Embedding

In this paper, we leverage RoBERTa, a state-of-the-art transformer-based language model, to enhance the word embedding phase of the suggested SMS spam detection methodology. RoBERTa excels in understanding the contextual relationships between words. It achieves this by representing words as dense vectors in a continuous vector space, capturing rich semantic information. Specifically, RoBERTa employs a bidirectional approach, considering each word's left and right contexts during pertaining, resulting in a highly contextualized word embedding. In our methodology, RoBERTa plays a pivotal role by transforming individual words within SMS messages into these context-aware vector representations. These embedding, which encapsulate nuanced word meanings and contextual information, serve as the foundation for subsequent stages of our model, such as feature extraction using convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks, ultimately enabling our Hierarchical Attention Network to effectively discern spam from legitimate messages by comprehensively considering the intricate relationships between words and sentences.

4.3 Parallel Feature Extraction

The proposed framework employs a dual-branch architecture for concurrent feature extraction at the word level and sentence level, thereby enabling a multifaceted analysis of textual data.

4.3.1 Word Level

We utilize Convolutional Neural Network (CNN) to refine word-level feature extraction. The efficacy of CNN lies in its ability to capture localized textual patterns by applying convolutional filters across word embedding. These filters, of varying receptive field sizes, are adept at discerning salient word combinations and syntactic structures indicative of spam content. This process is imperative for distinguishing between legitimate and spam SMS messages, as it facilitates the detection of specific linguistic markers that may be obscured in isolated word embedding. The resultant feature map from the word-level CNN encapsulates refined contextual insights, which are instrumental for the subsequent stages of our spam detection methodology.

4.3.2 Sentence Level

Building upon the local features identified by the word-level CNN, we used a sentence-level Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM augments the analytical prowess of the proposed framework by simulating extensive contextual information inherent within the SMS data. The bidirectional processing capabilities of the system enable it to understand the relationships between words in both the preceding and following contexts, thus capturing the nuanced semantic relationships in sentences. This enriched representation of sentence-level features is crucial for accurate spam discrimination. The integrated features derived from the BiLSTM constitute the foundational elements for our advanced Hierarchical Attention Network, designed to synthesize and evaluate the textual data comprehensively at both granular and holistic levels, thereby significantly elevating the precision of our SMS spam detection framework.

4.4 Fusion and Selection

HAN is a pivotal component designed to seamlessly integrate features extracted at both the word and sentence levels. The HAN leverages hierarchical attention mechanisms to weigh the importance of words within sentences and sentences within messages. By doing so, it discerns which words and sentences are most informative for spam detection, effectively filtering out irrelevant or redundant information. At the word level, attention is employed to identify significant words and their contextual importance within sentences. In contrast, at the sentence level, the network determines the relevance of each Sentence within the entire SMS message. This hierarchical attention mechanism ensures that our model concentrates on the most pertinent elements of the text, enhancing the effectiveness of spam classification. Furthermore, the HAN orchestrates feature fusion by combining the weighted word and Sentence embedding to create a comprehensive and contextually rich representation of the SMS message. This fused representation forms the basis for the final spam classification, enabling the model to make informed decisions based on local and global cues in the text data.

4.5 Classification

The final stage of the proposed SMS spam detection model is the classification, where the features are meticulously extracted and selected according to their weight by the Hierarchical Attention Network, it is time to make binary decisions regarding the nature of incoming messages. For this purpose, a fully connected layer is used as the classifier. This fully connected utilizes the fused features to perform the classification task [44]. Applying a combination of linear transformations and non-linear activation functions effectively maps the complex feature space to a decision boundary that separates spam messages from legitimate ones. The output of this layer provides a probability score, indicating the likelihood of the input message being spam or not. A suitable activation function, the sigmoid is applied to ensure that the output falls within the [0, 1] range, allowing for straightforward probability interpretation. The classification decision is made by comparing this probability score to a predefined threshold, and the message is labeled as either 'spam' or 'non-spam' based on this threshold, thus concluding the SMS spam detection process.

5 Evaluation

5.1 Dataset

The Machine Learning Repository, overseen by the University of California, Irvine (UCI), is a well-known online platform that provides many datasets for machine learning and data analysis. One of the datasets available in this repository is the UCI SMS Spam Collection [11], which consists of 5574 text messages. This dataset is classified into two categories: legitimate messages, making up 4827 instances (86.6%), and spam messages, accounting for 747 instances (13.4%). The distribution of the dataset is shown in Table 1.

Table 1: Dataset statistic

	Number of the messages	Percentage
Spam	4827	86.6%
Ham	747	13.4%
Total	5574	100%

5.2 Performance Matric

When we evaluate how well deep learning models perform, choosing the suitable measurement method is essential. There are various metrics available for this purpose, depending on the application. Sometimes, looking at just one metric may not give us a complete understanding, especially when dealing with imbalanced data. In those cases, we may need to use a combination of metrics to evaluate the models. We used well-known metrics like accuracy, precision, recall, and F1-score for our evaluation [45]. Before we delve into these metrics, it is essential to define four key terms, as shown in Table 2.

Table 2: Definitions of evaluation key terms for spam detection

Metric	Explanation
TP (True positive)	Accurately predicted spam messages.
TN (True negative)	Accurately predicted legitimate messages.
FP (False positive)	Mistakenly classified legitimate messages as spam.
FN (False negative)	Mistakenly classified spam messages as legitimate.

Here is an explanation of each metric:

1. **Accuracy:** This metric shows how accurately the model sorts out spam and legitimate messages among all its predictions. It looks at correct predictions of spam (true positives) and legitimate (true negatives) while also dealing with incorrect predictions of both. The formula computes the fraction of messages classified correctly (TN and TP) compared to all predicted messages (TN, FN, TP, and FP). It is computed as

$$\text{Accuracy} = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Number\ of\ Samples\ (TP + FP + TN + FN)} \tag{1}$$

2. **Precision:** This metric centers on the accuracy of positive predictions, particularly in pinpointing the number of predicted spam messages that are indeed spam. Reducing false positives is important. The calculation involves dividing the true positives by the total number of messages predicted as positive (including both true and false positives). It is computed as

$$\text{Precision} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{2}$$

3. **Recall:** Evaluates how well the model detects all real positive cases (spam messages), encompassing true positives and excluding false negatives. Its significance lies in minimizing the overlook of actual positive cases. The calculation divides the true positives by the total real positive cases (comprising both false and true negatives). It is computed as

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{3}$$

4. **F1-score:** Merges precision and recall, offering a well-rounded evaluation of the model's effectiveness. This becomes especially valuable when balancing precision and recall, which is essential. Using the harmonic mean, the F1-score addresses scenarios where one metric could considerably overshadow the other, achieving an equilibrium. The F1-score is computed by harmoniously factoring in precision and recall, recognizing their collective impact.

$$\text{F1-score} = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{4}$$

These metrics comprehensively assess the model's performance in SMS spam filtering, considering its accuracy and ability to Separate between spam and ham messages.

5.3 Model Configuration and Hyperparameters

Our model's architecture is specifically designed for effective SMS spam detection, leveraging advanced deep learning techniques for comprehensive textual analysis. It transforms textual data into a numerical format using RoBERTa word embedding (roberta-base), chosen for its optimal balance of computational efficiency and performance.

In the word-level feature extraction phase, a CNN employs a dual-layer setup, each layer equipped with 128 filters and initially using kernel sizes of 3. This configuration efficiently captures immediate contextual relationships within the text. The CNN's optimization process utilizes an Adam optimizer with an initial learning rate of 0.001, ensuring effective adjustment during training. For sentence-level analysis, we utilize a BiLSTM network structured with two layers and 128 units each, maintaining a consistent learning rate of 0.001. This setup effectively captures sentence dynamics by text sequences in both forward and reverse directions.

A crucial component of our approach is the Hierarchical Attention Network (HAN), which intricately combines and evaluates features at both the word and sentence levels, employing a dedicated layer of attention for each, fine-tuned at a learning rate of 0.001. This mechanism significantly enhances the model's ability to focus on the most relevant text segments for precise spam detection. The architecture includes a 256-unit fully connected layer with ReLU activation, followed by a 0.5 dropout layer to prevent overfitting [46]. Another fully connected layer with 128 units and ReLU activation precedes an additional dropout layer at the same rate. The final classification layer, equipped with 2 units and utilizing a SoftMax activation function, distinguishes between spam and non-spam messages.

6 Results and Analysis

We compared the proposed two-level SMS spam detection model with other modern approaches, including the modified Transformer, FRNN-HHO, CNN-LSTM BUNOW, BERT, and CNN-LSTM, using various performance metrics such as accuracy, precision, recall, and F-measure. The evaluation results demonstrated that our model outperformed all other models, achieving an accuracy rate of 99.48% with a high precision of 0.998, recall of 0.997, and F-measure 0.998 values. These results indicate that our model can accurately classify spam and non-spam messages with minimal false positives or false negatives, as shown in Table 3.

Table 3: Comparison performance of the SMS spam detection models

Method	Accuracy	Precision	Recall	F1-score
CNN-LSTM BUNOW	99.04%	97.3%	95.5%	96.4%
M-Transformer	98.92%	97.8%	94.5%	96.1%
FRNN-HHO	98.61%	99.7%	98.1%	98.9%
CNN-LSTM	98.37%	95.3%	87.8%	91.4%

(Continued)

Table 3 (continued)

Method	Accuracy	Precision	Recall	F1-score
BERT	99.28%	99.6%	99.2%	99.3%
LGRU	98.87%	99.7%	99%	99.4%
Proposed method	99.48%	99.8%	99.7%	99.8%

Because the dataset is imbalanced, we used a thorough 10-fold cross-validation approach to evaluate our SMS spam detection model [47]. Unlike the standard random split method, this method is well-suited for handling imbalanced datasets. It ensures that spam and legitimate messages are evenly represented in each evaluation cycle, making the results more trustworthy. The accuracy is an average of these ten cycles, offering a more robust and dependable measure of how well the model works in practical situations. Fig. 3 shows the accuracy assessment using cross-validation.

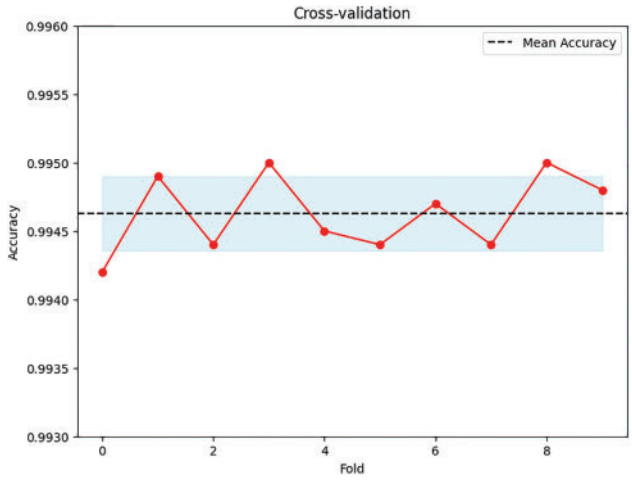

Figure 3: Accuracy assessment using cross-validation

We present a comprehensive performance evaluation of the proposed method through a series of visual figures, each specifically dedicated to a critical metric, as shown in the four parts of Fig. 4.

Part (a) of Fig. 4 illustrates accuracy results, providing a clear overview of how each approach performs in terms of accuracy. Notably, the two-level SMS spam detection method is distinctly highlighted, showcasing its exceptional accuracy compared to other methods.

Part (b) focuses on precision results, visually correlating precision performances across the evaluated approaches. The suggested approach's precision performance stands out distinctly, reaffirming its effectiveness in correctly identifying spam messages.

Part (c) visualizes the recall metric, detailing the ability of each method to identify all spam messages. Our method's superior recall performance is evident, reflecting its proficiency in comprehensively capturing spam instances.

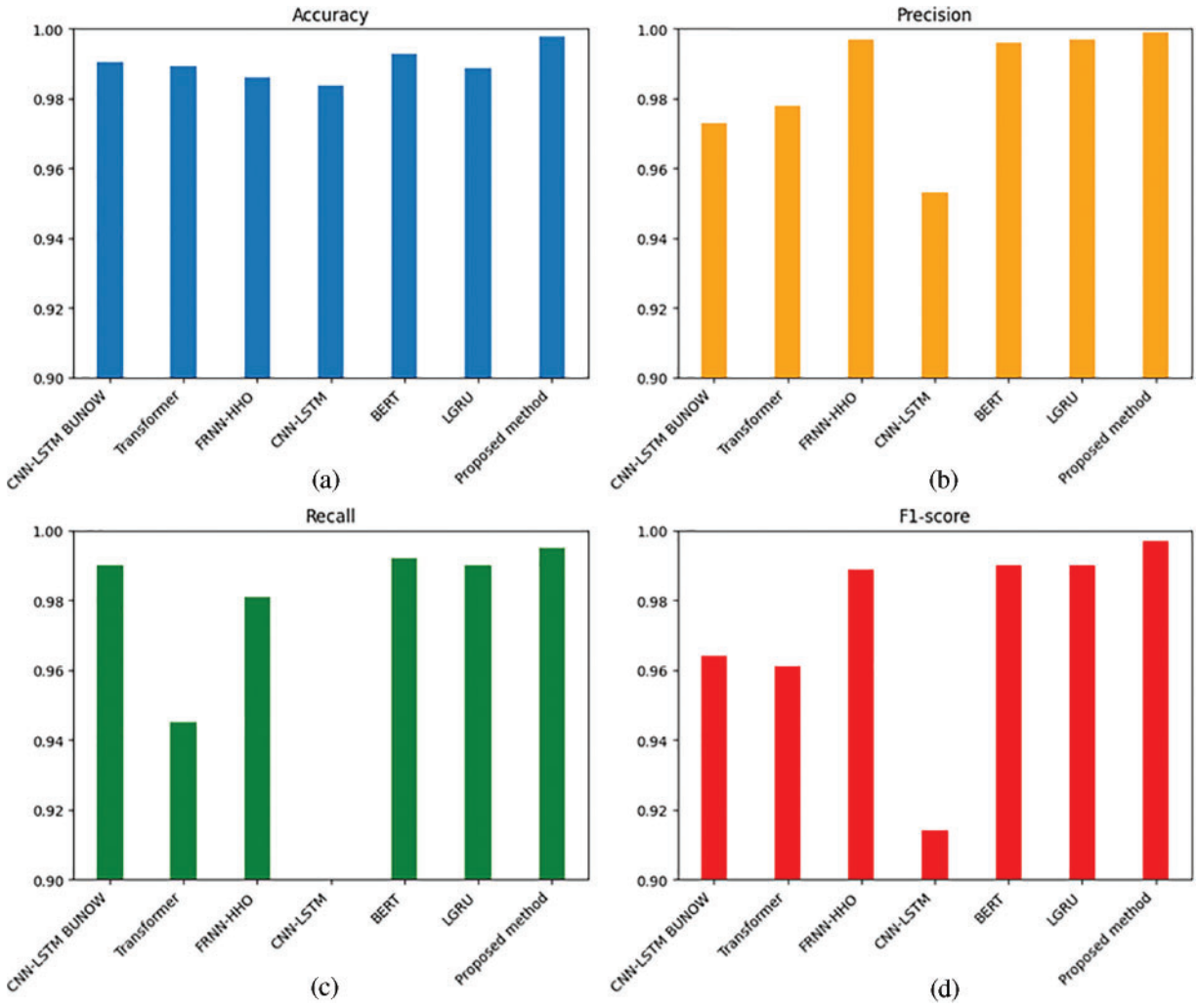

Figure 4: Performance evaluation of SMS spam detection models across various metrics

Part (d) delves into the F1-score, a balanced metric considering precision and recall. The figure illustrates how our method strikes a notable equilibrium between these aspects, further substantiating its robust performance.

The four segments of Fig. 4 offer an insightful comparison of the hierarchical two-level SMS spam detection method against other methodologies. This graphical representation effectively underscores the strengths of our approach, demonstrating its capacity to excel across multiple crucial evaluation metrics.

This study used the confusion matrix to evaluate the performance of the SMS spam detection model. It revealed the model's high precision and reliability in distinguishing between spam and non-spam messages. The matrix demonstrated the model's effectiveness in minimizing false positives and false negatives, underscoring its robustness and inaccurate message categorization, as shown in Fig. 5.

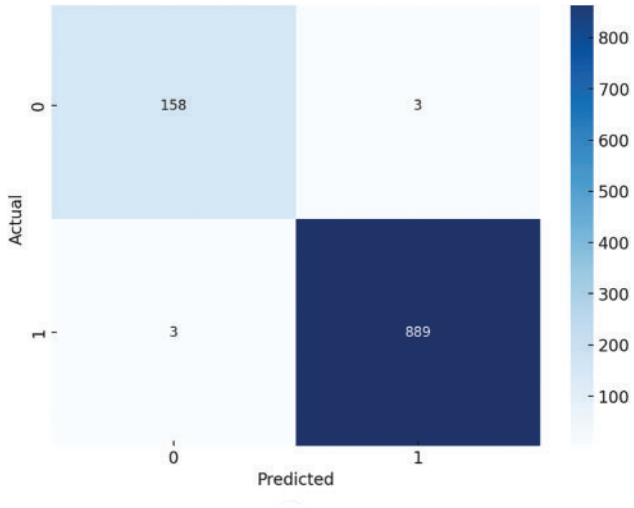

Figure 5: Confusion matrix

The performance of the proposed model for SMS spam detection was evaluated using a confusion matrix, as presented in Fig. 5. This matrix graphically displays the model's ability to differentiate between spam and legitimate (ham) messages, showcasing strong precision and minimal misclassification, affirming the effectiveness of our approach. With 158 correctly identified non-spam and 889 identified spam messages, the model demonstrates a high degree of precision in differentiating between categories. The matrix also indicates a low rate of misclassification, with only 3 instances where spam was mislabeled as non-spam and 3 instances where non-spam was incorrectly identified as spam. These minimal errors indicate the proposed model's capability for accurate message categorization. The examination of the model's errors [48] to understand where and why incorrect predictions occur is important to deeper insights into the model's performance and the nature of errors relative to the corpus.

7 Discussion

This study developed a SMS spam filtering framework based on Hierarchical Two-Level Feature Fusion. The key characteristics and findings are highlighted below:

- This study marks the first application of a Hierarchical Two-Level Feature Fusion approach for SMS spam detection, addressing both word-level and sentence-level analysis within the constrained context of short text messages. By achieving an exceptional accuracy rate of 99.48%, our method significantly addresses the challenges posed by advanced spam techniques, thereby enhancing the security and privacy of mobile communications. This achievement underscores the method's effectiveness in parsing the limited textual content typical of SMS, a critical advantage in identifying and filtering spam.
- Our model introduces a pioneering integration of pre-trained deep learning frameworks with a Hierarchical Attention Network (HAN), setting it apart from existing methods such as BERT, CNN-LSTM, and others. It demonstrates superior performance metrics, including accuracy, precision, recall, and F1-score, illustrating the benefits of our two-level feature fusion approach.

8 Conclusions

This paper introduced the innovative Two-Level SMS Spam Detection Method, which leverages hybrid deep learning and advanced text analysis techniques to establish an accurate framework for spam detection. By integrating the power of the Hierarchical Attention Network (HAN), our method has demonstrated exceptional performance in distinguishing between spam and legitimate SMS messages. The two-level hierarchical approach adeptly captures nuanced patterns within SMS content, combining word-level features with a comprehensive understanding at the sentence level. The proposed model achieved a remarkable accuracy rate of 99.48% on the UCI SMS dataset, significantly outperforming existing methods. For future research, we envision expanding the evaluation of the Two-Level SMS Spam Detection Method to include multilingual databases, showcasing its adaptability across different languages and cultural contexts. Additionally, the integration of external contextual information, such as sender reputation or network attributes, may further enhance the model's accuracy in detecting spam. Also, we plan to enhance our spam detection model by incorporating transfer learning and external knowledge sources. This approach will utilize the rich data from pre-trained models and broaden our understanding of spam indicators, aiming to improve accuracy and adaptability to new spam trends. Exploring these techniques represents a promising direction to advance our model, ensuring it remains effective and efficient in the evolving landscape of spam detection. This study not only advances the field of cybersecurity but also lays the groundwork for broader applications in various natural language processing domains.

Acknowledgement: Thanks are due to Ali-Reza Feizi-Derakhshi for the valuable technical support.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Hussein Alaa Al-Kabbi, interpretation of results: Mohammad-Reza Feizi-Derakhshi, draft manuscript preparation: Hussein Alaa Al-Kabbi, Mohammad-Reza Feizi-Derakhshi, and Saeed Pashazadeh. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] SlickText, "44 mind-blowing SMS marketing and texting statistics," 2023. Accessed: May 15, 2023. 2023. [Online]. Available: https://www.slicktext.com/blog/2018/11/44-mind-blowing-sms-marketing-and-texting-statistics/

[2] G. Sonowal and K. S. Kuppusamy, "SmiDCA: An anti-smishing model with machine learning approach," *Comput. J.*, vol. 61, no. 8, pp. 1143–1157, Aug. 2018. doi: 10.1093/comjnl/bxy039.

[3] V. Dharani, D. Hegde, and Mohana, "Spam SMS (or) email detection and classification using machine learning," in *2023 5th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, 2023, pp. 1104–1108. doi: 10.1109/ICSSIT55814.2023.10060908.

[4] S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: Methods and data," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9899–9908, 2012.

[5] M. Woźniak, J. Siłka, and M. Wieczorek, "Deep neural network correlation learning mechanism for CT brain tumor detection," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14611–14626, 2023. doi: 10.1007/s00521-021-05841-x.

[6] A. H. J. Almarashy, M. R. Feizi-Derakhshi, and P. Salehpour, "Enhancing fake news detection by multi-feature classification," *IEEE Access*, vol. 11, pp. 139601–139613, 2023. doi: 10.1109/ACCESS.2023.3339621.

[7] H. Wu, H. Han, X. Wang, and S. Sun, "Research on artificial intelligence enhancing internet of things security: A survey," *IEEE Access*, vol. 8, pp. 153826–153848, 2020. doi: 10.1109/ACCESS.2020.3018170.

[8] M. Bani-Almarjeh and M. B. Kurdy, "Arabic abstractive text summarization using RNN-based and transformer-based architectures," *Inf. Process. Manage.*, vol. 60, no. 2, pp. 103227, 2023. doi: 10.1016/j.ipm.2022.103227.

[9] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, pp. 83, 2022. doi: 10.3390/info13020083.

[10] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: Analysis of spammer strategies and the dataset shift problem," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1145–1173, 2023. doi: 10.1007/s10462-022-10195-4.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, Jun. 2014, pp. 1188–1196.

[12] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. 2016 Conf. North Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL HLT 2016)*, San Diego, CA, USA, 2016, pp. 1480–1489.

[13] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in *Proc. 11th ACM Symp. Docu. Eng. (DocEng '11)*, New York, NY, USA, 2011, pp. 259–262.

[14] S. Y. Yerima and A. Bashar, "Semi-supervised novelty detection with one class SVM for SMS spam detection," in *Proc. 29th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Sofia, Bulgaria, 2022, pp. 1–4. doi: 10.1109/IWSSIP55020.2022.9854496.

[15] D. D. Arifin, Shaufiah, and M. A. Bijaksana, "Enhancing spam detection on mobile phone short message service (SMS) performance using FP-growth and naive bayes classifier," in *Proc. IEEE Asia Pacific Conf. Wireless Mobile (APWiMob)*, Bandung, Indonesia, 2016, pp. 80–84. doi: 10.1109/APWiMob.2016.7811442.

[16] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, "A comparative study of spam SMS detection using machine learning classifiers," in *2018 Eleventh Int. Conf. Contemp. Comput. (IC3)*, Noida, India, 2018, pp. 1–7.

[17] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019. doi: 10.1109/ACCESS.2019.2946594.

[18] Z. H. Ali, H. M. Salman, and A. H. Harif, "SMS spam detection using multiple linear regression and extreme learning machines," *Iraqi J. Sci.*, vol. 64, no. 10, pp. 6342–6351, 2023. doi: 10.24996/ijs.2023.64.10.45.

[19] S. Hosseinpour and H. Shakibian, "An ensemble learning approach for SMS spam detection," in *2023 9th Int. Conf. Web Res. (ICWR)*, Tehran, Islamic Republic of Iran, 2023, pp. 125–128. doi: 10.1109/ICWR57742.2023.10139070.

[20] S. Pudasaini, A. Shakya, S. P. Pandey, P. Paudel, S. Ghimire and P. Ale, "SMS spam detection using relevance vector machine," in *Procedia Comput. Sci.*, Halifax, NS, Canada, 2023, vol. 230, pp. 337–346.

[21] X. Liu, H. Lu, and A. Nayak, "A spam transformer model for SMS spam detection," *IEEE Access*, vol. 9, pp. 80253–80263, 2021. doi: 10.1109/ACCESS.2021.3081479.

[22] T. Xia and X. Chen, "A discrete hidden markov model for SMS spam detection," *Appl. Sci.*, vol. 10, no. 14, pp. 5011, 2020. doi: 10.3390/app10145011.

[23] U. Srinivasarao and A. Sharaff, "SMS sentiment classification using an evolutionary optimization based fuzzy recurrent neural network," *Multimed. Tools Appl.*, vol. 82, no. 27, pp. 42207–42238, 2023. doi: 10.1007/s11042-023-15206-2.
[24] S. Giri, S. Das, S. B. Das, and S. Banerjee, "SMS spam classification-simple deep learning models with higher accuracy using BUNOW and GloVe word embedding," *J. Appl. Sci. Eng.*, vol. 26, pp. 1501–1511, 2023.
[25] K. Debnath and N. Kar, "SMS spam detection using deep learning approach," in *Proc. ICHCSC 2022*, Singapore, Springer Nature Singapore, 2022, pp. 337–347.
[26] 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. 156, pp. 156, 2020. doi: 10.3390/fi12090156.
[27] O. Abayomi-Alli, S. Misra, and A. Abayomi-Alli, "A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset," *Concurr. Comput.: Pract. Exp.*, vol. 34, no. 17, pp. e6989, 2022. doi: 10.1002/cpe.6989.
[28] A. S. Onashoga, O. O. Abayomi-Alli, A. S. Sodiya, and D. A. Ojo, "An adaptive and collaborative server-side SMS spam filtering scheme using artificial immune system," *Inf Sec. J.: A Global Perspect*, vol. 24, no. 4–6, pp. 133–145, 2015. doi: 10.1080/19393555.2015.1078017.
[29] F. Wei and T. Nguyen, "A lightweight deep neural model for SMS spam detection," in *2020 Int. Symp. Netw., Comput. Commun. (ISNCC)*, Montreal, QC, Canada, 2020, pp. 1–6. doi: 10.1109/ISNCC49221.2020.9297350.
[30] E. Ardeshir-Larijani and M. M. Nasiri Fatmehsari, "Hybrid classical-quantum transfer learning for text classification," *Quantum Mach. Intell.*, vol. 6, p. 19, 2024. doi: 10.1007/s42484-024-00147-2.
[31] M. R. Julis and S. Alagesan, "Spam detection in SMS using machine learning through text mining," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 171–175, 2020.
[32] I. S. Mambina, J. D. Ndibwile, D. Uwimpuhwe, and K. F. Michael, "Uncovering SMS spam in swahili text using deep learning approaches," *IEEE Access*, vol. 12, pp. 25164–25175, 2024.
[33] E. Zafarani-Moattar, M. R. Kangavari, and A. M. Rahmani, "Neural network meaningful learning theory and its application for deep text clustering,," *IEEE Access*, vol. 12, pp. 42411–42422, 2024.
[34] P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Bert, xlnet or roberta: The best transfer learning model to detect clickbaits," *IEEE Access*, vol. 9, pp. 154704–154716, 2021. doi: 10.1109/ACCESS.2021.3128742.
[35] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis," *Appl. Sci.*, vol. 13, no. 6, pp. 3915, 2023. doi: 10.3390/app13063915.
[36] F. Tajaddodianfar, J. W. Stokes, and A. Gururajan, "Texception: A character/word-level deep learning model for phishing URL detection," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Barcelona, Spain, 2020, pp. 2857–2861.
[37] N. S. M. Nafis and S. Awang, "The evaluation of accuracy performance in an enhanced embedded feature selection for unstructured text classification," *Iraqi J. Sci.*, vol. 61, no. 12, pp. 3397–3407, 2020.
[38] G. A. Vlad, M. A. Tanase, C. Onose, and D. C. Cercel, "Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model," in *Proc. Second Workshop Nat. Lang. Process. Internet Freedom: Censorship, Disinformation, and Propaganda*, Hong Kong, China, Association for Computational Linguistics, 2019, pp. 148–154.
[39] W. Huang *et al.*, "Hierarchical multi-label text classification: An attention-based recurrent network approach," in *Proc. 28th ACM Int. Conf. Inform. Knowl. Manag. (CIKM '19)*, New York, NY, USA, 2019, pp. 1051–1060.
[40] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," *Appl. Sci.*, vol. 13, no. 7, pp. 4550, 2023. doi: 10.3390/app13074550.
[41] I. K. S. Al-Tameemi, M. R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "An efficient sentiment classification method with the help of neighbors and a hybrid of RNN models," *Complexity*, vol. 2023, pp. 1–14, 2023. doi: 10.1155/2023/1896556.
[42] C. D. P. Laureate, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *Artif. Intell. Rev.*, vol. 56, no. 12, pp. 1–33, 2023. doi: 10.1007/s10462-023-10471-x.

[43] F. Sakketou and N. Ampazis, "A constrained optimization algorithm for learning GloVe embeddings with semantic lexicons," *Knowl. Based Syst.*, vol. 195, pp. 105737, 2020.
[44] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 150, 2019. doi: 10.3390/info10040150.
[45] C. M. Fuller, D. P. Biros, and D. Delen, "An investigation of data and text mining methods for real world deception detection," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8392–8398, 2011.
[46] T. Prexawanprasut and P. Chaipornkaew, "An analytical study on email classification using 10-fold cross-validation," in *2019 5th Int. Conf. Sci. Inform. Technol. (ICSITech)*, Yogyakarta, Indonesia, 2019, pp. 38–43.
[47] I. K. S. Al-Tameemi, M. R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data," *IEEE Access*, vol. 11, pp. 91060–91081, 2023.
[48] M. R. Hossain, M. M. Hoque, and N. Siddique, "Leveraging the meta-embedding for text classification in a resource-constrained language," *Eng. Appl. Artif. Intell.*, vol. 124, no. 1, pp. 106586, 2023. doi: 10.1016/j.engappai.2023.106586.