Journal of Intelligent Learning Systems and Applications, 2023, 15, 144-164
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Deep Neural Network Based Spam Email Classification Using Attention Mechanisms

Md. Tofael Ahmed[1], Mariam Akter[1], Md. Saifur Rahman[1], Maqsudur Rahman[1], Pintu Chandra Paul[1], Miss. Nargis Parvin[2], Almas Hossain Antar[1]

[1]Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh
[2]Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh
Email: tofael@cou.ac.bd, mariamakter340@gmail.com

How to cite this paper: Ahmed, Md.T., Akter, M., Rahman, Md.S., Rahman, M., Paul, P.C., Parvin, Miss.N. and Antar, A.H. (2023) Deep Neural Network Based Spam Email Classification Using Attention Mechanisms. *Journal of Intelligent Learning Systems and Applications*, **15**, 144-164.
https://doi.org/10.4236/jilsa.2023.154010

Received: September 1, 2023
Accepted: November 3, 2023
Published: November 30, 2023

Abstract

Spam emails pose a threat to individuals. The proliferation of spam emails daily has rendered traditional machine learning and deep learning methods for screening them ineffective and inefficient. In our research, we employ deep neural networks like RNN, LSTM, and GRU, incorporating attention mechanisms such as Bahdanua, scaled dot product (SDP), and Luong scaled dot product self-attention for spam email filtering. We evaluate our approach on various datasets, including Trec spam, Enron spam emails, SMS spam collections, and the Ling spam dataset, which constitutes a substantial custom dataset. All these datasets are publicly available. For the Enron dataset, we attain an accuracy of 99.97% using LSTM with SDP self-attention. Our custom dataset exhibits the highest accuracy of 99.01% when employing GRU with SDP self-attention. The SMS spam collection dataset yields a peak accuracy of 99.61% with LSTM and SDP attention. Using the GRU (Gated Recurrent Unit) alongside Luong and SDP (Structured Self-Attention) attention mechanisms, the peak accuracy of 99.89% in the Ling spam dataset. For the Trec spam dataset, the most accurate results are achieved using Luong attention LSTM, with an accuracy rate of 99.01%. Our performance analyses consistently indicate that employing the scaled dot product attention mechanism in conjunction with gated recurrent neural networks (GRU) delivers the most effective results. In summary, our research underscores the efficacy of employing advanced deep learning techniques and attention mechanisms for spam email filtering, with remarkable accuracy across multiple datasets. This approach presents a promising solution to the ever-growing problem of spam emails.

Keywords

Spam Email, Attention Mechanism, Deep Neural Network, Bahdanua

Attention, Scale Dot Product

1. Introduction

Digital networking techniques play a crucial role in email communication. Ray Tomlinson, the credited inventor of email [1], introduced this revolutionary form of communication in 1971. The widespread use of email for sharing information, ideas, and written correspondence is a global phenomenon. Email serves as a means for technologically enabled interpersonal communication. It facilitates the transmission of various types of content, including text, files, images, web links, and multimedia assets, to individuals or networked groups of recipients. It's worth noting that sending emails to specific individuals or groups typically comes at no additional cost.

Recent research indicates a substantial growth in global email subscriptions, reaching 3.9 billion in 2019. Forecasts suggest continued expansion, surpassing 4.48 billion by 2024. This anticipated increase underscores the growing importance and widespread use of email as a preferred method for information transmission and participation. Notably, an estimated 281 billion emails were exchanged daily in 2018.

The evolution of technology has significantly impacted communication, including the use of email. However, email spam has posed challenges to the effectiveness of this medium. Spam emails are often used to send unwanted and, at times, malicious messages. Many internet users register their email addresses on websites and receive notifications to stay informed and safeguard against online threats. While most unsolicited emails are harmless, users must exercise caution when receiving emails that pose risks to their online identity and data.

The problem of email spam has persisted since the early 1990s and is projected to account for 90% of global email traffic by 2014. Spam emails have been found to waste recipients' time, storage, and network resources. Research from March 2020 indicates that 53.95% of emails sent were spam. Globalization has played a significant role in the rapid growth of internet usage and spam emails. What was once merely annoying unsolicited advertising has now evolved to include fraudulent schemes, malicious software, and phishing attempts aimed at stealing personal information [2] [3].

Statistics from Kaspersky Lab indicate that half of all emails are spam, while Cisco Talos reports that spam emails account for 85% of all emails, surpassing 200 billion daily. It is anticipated that by 2023, 50% of global email traffic will be spam, as indicated by Statista [4]. The Message Labs Intelligence Report reveals that 88% of email traffic is spam.

Spam has detrimental effects on businesses, causing annoyance to users, compromising communication accuracy and effectiveness, reducing work productivity, consuming network bandwidth, depleting server storage and processing ca-

pacity, facilitating the spread of malicious software, and causing financial losses due to phishing, DoS (Denial of Service), and directory harvesting attacks.

To mitigate the impact of spam, email management often involves the use of spam filter software, which helps users manage their email inboxes by marking spam or determining the relevance of messages for further reading. This software is instrumental in reducing the influence of spam on users' productivity and overall email experience by enabling informed email selection [5].

As the issue of spam email continues to grow daily, it becomes evident that existing research in this area is insufficient. This study presents an attention-based approach to spam email classification. The proposed method, termed "spam email classification using attention mechanisms", enhances the accuracy and efficiency of spam email detection. Our approach improves performance by identifying the most relevant elements and patterns in email content through attention mechanisms. This research implements a novel neural networks model that incorporates attention processes, a technique commonly used in machine translation, but here applied to spam email classification. Our work successfully classifies spam emails using attention techniques.

The primary contributions of this study include:

• The utilization of deep neural networks (RNN, LSTM, GRU) for spam email filtering, incorporating Bahdanua, Luong, scaled dot product, and scaled dot product self-attention mechanisms.

• The use of different ROC curves to compare Attention Mechanisms in datasets such as Trec spam, Enron spam email, SMS spam collection, Ling spam, and a substantial custom dataset.

2. Related Work

In a study by Islam *et al.* [1], the focus was on spam detection methods. This research centered on utilizing four machine learning models and two deep learning models to identify spam terms and assess their effectiveness in detecting and categorizing spam communications. Multiple datasets were examined, including the Trec spam dataset, Enron dataset, PU dataset, and Ling spam dataset, each serving a specific purpose in exploring the subject matter.

Machine learning models such as logistic regression, XgBoost, support vector machine, and random forest were comprehensively studied and widely used across various fields. Logistic regression, a popular linear model, is frequently employed for binary classification tasks. XgBoost, an ensemble learning technique, combines weak learners to build robust prediction models. Support vector machines (SVMs) are effective models capable of handling both linear and non-linear classification problems by identifying optimal hyperplanes for separating different classes.

Deep learning methods, such as word embedding and LSTM, were also employed in this study, and the performance of all models was deemed satisfactory.

Another study conducted by Martino *et al.* [2] delved into the identification of

spam emails. The research aimed to explore the impact of adversarial entities in dynamic environments known for dataset shifts. The findings revealed that dataset changes can significantly affect a system's generalization capabilities, with error rates potentially reaching as high as 48.81%.

Kuchipudi *et al.* [3] provided an in-depth analysis of spam filters, with a particular focus on three intrusive methodologies: synonym substitution, ham word insertion, and spam word spacing. Various strategies were implemented to optimize spam filters' effectiveness in detecting and eliminating unwanted email communications.

In the realm of natural language processing, various methodologies and computational algorithms have been developed and employed to analyze textual data. One commonly used technique for tasks like text categorization and sentiment analysis is the Naive Bayes algorithm, based on Bayes' theorem and the assumption of conditional independence between features in a document given the class label.

Liu *et al.* [4] conducted a study where they employed the Transformer model for SMS spam detection, utilizing the SMS spam collection dataset along with UtkMI's Twitter spam dataset. Different models, including logistic regression, random forest, naive Bayes, support vector machines, LSTM, CNN-LSTM, and spam transformer, were evaluated. The spam transformer exhibited superior performance in terms of accuracy, recall, and F1-Score.

Anticipated for 2022, a research paper by Hua Shenet *et al.* [6] is set to make use of the Twitter social honeypot dataset and Kwak's dataset, exploring various deep learning designs and techniques to enhance the efficiency of neural networks.

Yong Fang *et al.* [7] conducted a study in 2019 utilizing the Enron dataset and the spam assassin tool. The WMT 2014 collection included datasets for the English-German and English French language pairs, providing around 4.5 million and 36 million sentence pairs, respectively.

Through the analysis of these metrics, researchers can assess system effectiveness and make informed decisions regarding system optimization and enhancement. Research has indicated that the THEMIS model outperforms alternative models. Machine translation has been significantly improved through the incorporation of attention mechanisms, with Vaswani *et al.*'s seminal work [8] contributing to the field.

In 2016, Zichao Yang [9] and their team conducted a study using various datasets, including Yelp reviews, IMDB reviews, Yahoo Answers, and Amazon reviews. Several models have been developed for text classification tasks, including the hierarchical attention network, GRU-based sequence encoder, and various Bag-of-Words (BOW) methods. SVM (Support Vector Machines) is frequently utilized as a classifier in text classification assignments. Additionally, CNNs (Convolutional Neural Networks) have been applied to text classification tasks, operating at both word and character levels.

Recent research has focused on understanding attention distribution within Recurrent Neural Networks (RNN). In 2023, Sravani *et al.* [10] conducted a study using a private email dataset to visualize attention weight distribution. Several models, including RCNN, attention mechanisms, and NLP models, were used to generate visual representations [11] of pair plots and feature correlations. Numerous studies have indicated the effectiveness of RCNN models in various tasks.

Global and local attention mechanisms play a crucial role in improving the performance of various models and algorithms. These mechanisms are used in natural language processing, computer vision, and machine learning. Performance evaluation in natural language processing and machine translation tasks often involves the use of metrics such as PPI, BLEU, AER, and others.

In another study by Vinitha *et al.* [12], different artificial neural network (ANN) models were evaluated, including feedforward neural networks (NN), multi-layer perceptron (MLP), recurrent neural networks (RNN), and long short-term memory (LSTM) networks. The LSTM model exhibited high accuracy, reaching a rate of 97.4% in the investigation.

Mani *et al.* [13] [14] extensively explored the use of K-Nearest Neighbors (KNN) and Gated Recurrent Units (GRU) for spam identification, with KNN achieving an accuracy of approximately 90%. Vinoth *et al.*'s research [15] focused on email spam detection and proposed the Feature Subset Selection with Deep Learning-based Email Spam Detection and Classification (FSSDL-ESDC) models. The study emphasized preprocessing techniques such as tokenization and stop word removal.

In a study by Abdullah Sheneamer [16], the application of machine learning and deep learning classifiers in email spam filtering was explored. Various classifiers, including Support Vector Machines (SVM), Naive Bayes, decision trees, random forests, XgBoost, Long Short-Term Memory (LSTM), LSTM with glove, and Convolutional Neural Networks (CNN) with glove, were evaluated for their effectiveness in email spam detection.

Sultan Zavrak *et al.* [17] presented an innovative methodology that combines hierarchical attention mechanisms with hybrid deep learning techniques to enhance spam identification in email correspondence. Their study explored various models, such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and Feedforward Neural Networks (FT).

In recent research by Nashit Ali and their team [18], a combination of Feature Transformation (FT) and Hierarchical Attention Networks (HAN) was found to offer optimal performance. The study focused on email classification and involved feature extraction and alignment, specifically by selecting important sentences using deep learning techniques.

In a study by Mohammad Zavvar *et al.* [19], a combined approach involving particle swarm optimization (PSO), artificial neural network (ANN), and sup-

port vector machine (SVM) was proposed for email spam detection using the UCI dataset. SVM achieved an AUC score of 0.9307.

3. Research Methodology

3.1. Dataset

Data collection is a crucial component of the proposed methodology in Figure 1, as the effectiveness of a model can sometimes be contingent upon the acquisition of a sufficient amount of relevant data. In order to effectively train our model, a substantial amount of data is necessary. Hence, the data emerges as the paramount element in this context. As part of our research methodology, From Table 1, we gather data from various online platforms, including Kaggle and the UCI machine repository. The datasets contain a significant volume of data, comprising two distinct columns: email and spam. In the context of spam classification, the spam column is typically classified into two categories: Ham (0) and Spam (1).

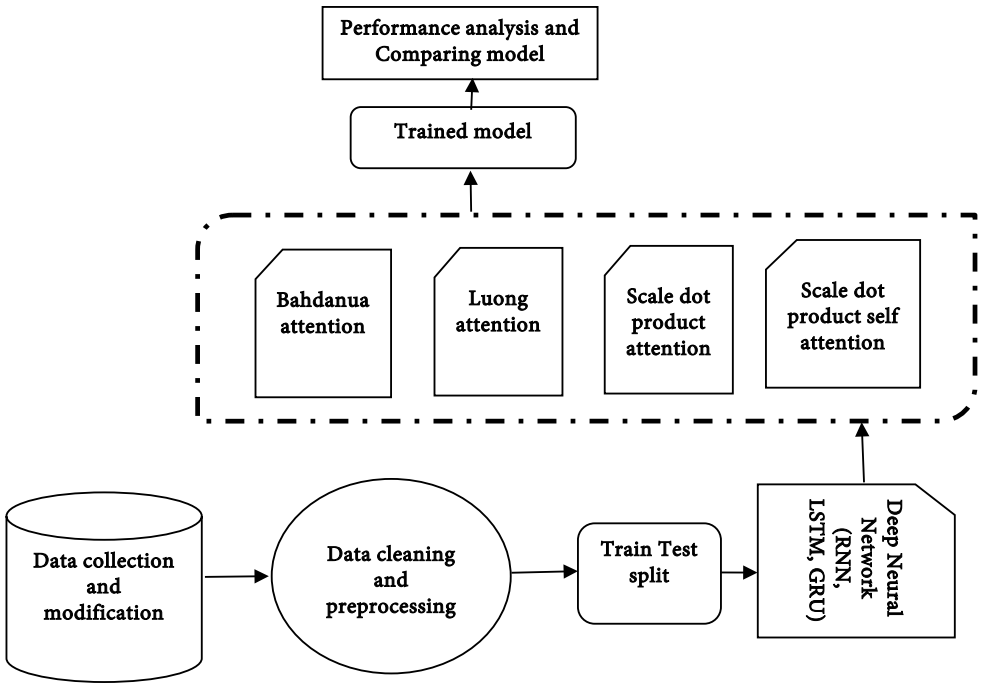

Figure 1. Flow diagram of our proposed method.

Table 1. Dataset statistics.

Datasets Name	Total number of data	Ham emails	Spam emails	source	Ham (%)	Spam (%)
Trec Spam dataset 2007	75,419	25,220	50,199	Kaggle	33.4	66.6
Enron dataset	33,716	16,545	17,171	Kaggle	49	51
Large Customize dataset	46,076	36,038	10,038	custom dataset	78.2	21.8
SmsSpamCollection UCI dataset	5572	4825	747	Uci machine	86.6	13.4
Lingspam dataset	2893	2412	481	Kaggle	83.4	16.6

3.2. Recurrent Neural Network (RNN)

A type of artificial neural network called a recurrent neural network (RNN) can solve deep learning sequence prediction issues. The value of the epochs displays the total number of training examples that have gone through one forward pass and one backward pass. Using Equation (1), a group of neurons were produced and are found in the hidden layer.

$$N_h = \frac{N_s}{\alpha * \left(N_i + N_o\right)} \tag{1}$$

where N_i is the total number of input neurons, N_o is the total number of output neurons, N_s is the total number of training dataset samples, and α is an arbitrary scaling factor. Sigmoid, ReLu, and Tanh activation function types are the most prevalent. In Equation (2) is a definition of the sigmoid activation function.

$$F\left(x\right) = \frac{1}{1+\exp^{-x}} \tag{2}$$

3.3. Long Short-Term Memory (LSTM)

A type of neural network called an LSTM uses the output of the previous step as an input for the current step. Only the values closer to 1 will be forwarded to the cell state for spam categorization after preprocessing, as the values closer to zero will not be forwarded as they include irrelevant information in Equation (3).

$$f_t = \sigma\left(w_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

Input gate: The new input email that has been provided and the previous hidden state are used in the following phase to determine what additional information should be added to the cell state using the sigmoid and the tanh function in Equation (4, 5), additionally passing the candidate input c_t to the cell state multiplies the new input email data.

$$i_t = \sigma\left(w_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{4}$$

$$\hat{C}_t = \tanh\left(w_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{5}$$

Cell state: Initially multiplied the previous cell state $C_{t\text{-}1}$ by the forget gate f_t. If the multiplication results in a value of zero, the information contained in the cell state is lost. If not, point-by-point addition will be carried out using the input gate's output to update the network with a new cell state C_t that contains all necessary data in Equation (6).

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{6}$$

Output gate: The value of the following hidden state is finally controlled by the output gate. The network determines the data the hidden state should contain to anticipate spam email based on the final output produced in Equation (7).

$$O_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

Hidden state output: The newly formed cell state and hidden state are subse-

quently transferred over to the following time step as in Equation (8)

$$h_t = O_t * \tanh\left(C_t\right) \tag{8}$$

3.4. Gated Recurrent Unit

While LSTM is more effective when dealing with datasets that contain longer sequences, GRU [20] is quicker and uses less memory than it. GRUs also provides a solution to the vanishing gradient problem, which affects conventional recurrent neural networks (information used to update network weights). At time step *t*, each hidden layer is calculated using the following formulas:

Update gate:

$$z_t = \sigma\left(w_z \cdot \left[h_{t-1}, x_t\right]\right) \tag{9}$$

Reset gate:

$$r_t = \sigma\left(w_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{10}$$

New memory:

$$\tilde{h}_t = \tanh\left(w \cdot \left[r_t * h_{t-1}, x_t\right]\right) \tag{11}$$

Final memory:

$$h_t = \left(1 - z_t\right) * h_{t-1} + z_t * \tilde{h}_t \tag{12}$$

In this formula, W stands for the weight vector, $*$ for element-by-element multiplication, and σ is the sigmoid function, x_t is the input and h_{t-1} is the hidden state.

3.5. Attention Mechanism

The attention model essentially operates on the notion of attention, which involves focusing more intensely on a small number of items while ignoring others. The standard encoder-decoder model's drawback is addressed by the application of the attention mechanism. The next step is to create a context vector (c_t) that will help forecast the current target word (y_t) by capturing pertinent source-side data. While the context vector c_t is handled differently in both models, the information from the source-side context vector c_t and the target hidden state h_t are combined using a straightforward concatenation layer to create the following attentional hidden state:

$$\tilde{h}_t = \tan\left(W_c\left[c_t; h_t\right]\right) \tag{13}$$

In order to create the predictive distribution denoted by, the attentional vector $\tilde{h}_t$ is then sent through the softmax layer.

$$p\left(y_t \mid y_{<t}, x\right) = \text{softmax}\left(W_s \tilde{h}_t\right) \tag{14}$$

3.6. Bahdanua Attention Mechanism (BAMec)/Additive Attention Model/Global Attention

Bahdanau *et al.* [21] only tested one alignment function, the concate product;

nevertheless, we will demonstrate later that there are better options. The Bahdanau attention model's architecture is described below. S_{t-1} is the hidden decoder state where previous time step is $t-1$. Each decoder step generates a distinct context vector c_t at time step t to provide target word y_t. An annotation h_i that concentrates on the i-th word out of the entire number of words and catches the important details weight value assigned to h_i is $\alpha_{t,i}$ where current time step t. The attention score $e_{t,i}$. produced by the given model $a(.)$ demonstrates how well S_{t-1} and h_i match.

$$e_{t,i} = a\left(s_{t-1}, h_i\right) \tag{15}$$

$$\alpha_{t,i} = \text{softmax}\left(e_{t,i}\right) \tag{16}$$

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i \tag{17}$$

where c_t is the context vector.

3.7. Luong Attention Mechanism (LAMec)/Multiplicative Attention Model/Dot Product Attention Model

Dot-product attention, in its simplest version, computes attention weights for every query as the dot-product of the query to all keys. After that, the key dimension is treated with the SoftMax function [22]. After that, these attention weights are multiplied by the following values:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(QK^{\text{T}}\right)V \tag{18}$$

Here, $Q \in \mathbb{R}^{n \times d_{\text{model}}}, K \in \mathbb{R}^{n \times d_{\text{model}}}, V \in \mathbb{R}^{n \times d_{\text{model}}}$ are, respectively, the matrices of n queries, keys, and values. The dot-product between queries and keys may increase significantly for big values of d_{model}. As a result, the SoftMax function is pushed into the saturated region, where its gradients are incredibly small. This results from the SoftMax function's exponentiation of specific query-key dot-products. They introduce scaled dot product attention, where QK^{T} is scaled by $\frac{1}{\sqrt{d_{\text{model}}}}$, because this could be detrimental to training.

3.8. Scale Dot Product Attention Mechanism (SDP)

Scale Dot-Product [4] [7], the goal of attention is to prevent the significant expansion of dot-product when the dimensions of queries and keys d_k is high, with a scaling factor of $\frac{1}{\sqrt{d_k}}$. Transformer's Multi-Head Attention is another significant advance. In the previous exercise, the queries, keys, and values were given direct attention, and their dimension was d_{model}. The sum of all these h values is then projected back onto a dimension of the d_{model}. The full operation of the Transformer's attention mechanism is described as follows:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \tag{19}$$

$$\text{MultiHead}\left(Q, K, V\right) = \text{Concat}\left(\text{head}_1, \cdots, \text{head}_h\right)W^o \tag{20}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{21}$$

To project d_{model} dimension queries, keys, and values to d_k, d_k, and d_v dimensions, respectively, the W_i^Q, W_i^K, and W_i^V are parameters matrices found in linear projection layers. In both the original Transformer and the one we modified to detect email spam, $d_k = d_v = d_{\text{model}}/h$.

4. Performance Matrix

Various performance measures can be used to evaluate the effectiveness of spam email classification. For models, the detection of the emails is visualized using a performance matrix. Performance matrix [23] is composed of True Negative (TN), True Positive (TP), False Positive (FP), False Negative (FN).

4.1. Accuracy

The goal of the study was to determine which email spam and ham classification method had the highest accuracy. The accuracy module from the Scikit-learn library assisted in determining the precise number of emails that should have been labeled as "Spam" and "Ham". Equation (22) below can be used to quantify this.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{22}$$

where the total number of emails included in the test data serves as the equation's denominator.

4.2. Precision

Calculating the correctly recognized values—that is, how many correctly identified spam emails have been separated from the provided collection of positive emails—is the precision measurement. This refers to determining the overall number of emails that were accurately identified as positive out of all emails that were positively predicted. Equation (23) defines this as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

4.3. Recall

From the total number of spam emails provided, the recall measurement calculates how many emails were accurately identified as spam. Equation (24) gives a definition for this, where "TP + FN" stands for the total number of spam emails found in the testing data.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

4.4. F1-Score

With the use of precision and recall scores, the F-measure or the value of F_β is

determined, where β is denoted by 1, and F or F1 provides the F1-score in Equation (25). The "Harmonic mean" of the precision and recall values is the F1-score.

$$F_{\beta} = \frac{\left(1+\beta^2\right)\left(\text{Precision} \times \text{Recall}\right)}{\beta^2 \times \left(\text{Precision} + \text{Recall}\right)} \tag{25}$$

4.5. AUC

Plotting the True Positive Rate (TPR) vs the False Positive Rate (FPR) at various threshold settings yields the ROC curve. TPR = TP/(TP + FN). False Positive Rate (FPR) is the proportion of positively predicted cases to all positively observed instances in the dataset. FP/(FP + TN) = FPR.

5. Results and Discussion

Table 2 provides a comparison of model performance on the TREC spam dataset, with all models surpassing previous methods. The Long Short-Term Memory (LSTM) model with the Luong attention mechanism achieves the highest accuracy at 99.61%. When it comes to precision, the research reveals that the LSTM model with SDP self-attention attains the maximum value of 99.59%. Meanwhile, studies show that the Luong attention mechanism in the GRU model results in an outstanding recall rate of 99.77%. In terms of F-measure, our research demonstrates that the LSTM model with Luong attention is the top-performer, achieving a 99.61% accuracy rate. Additionally, the LSTM model with Luong attention exhibits the largest area under the ROC curve at 99.60%. Model performance was evaluated using error analysis, including MAE, RMSE, and MSLE, with the LSTM model and Luong attention having the lowest values (0.003, 0.062, and 0.001). This research leverages the Long Short-Term Memory (LSTM) architecture with the Luong attention mechanism for the TREC Spam dataset.

In Table 3, results from a deep neural network analysis on the Enron dataset are presented. All models perform better, with the Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models with scale dot product (SDP) self-attention achieving the highest accuracy of 99.97%. Precision rates of 99.94% are observed for the GRU and LSTM models with SDP self-attention. The research indicates that recall is most effectively addressed by attention mechanisms, reaching 100%. In terms of F-measure, models using the GRU architecture with SDP self-attention mechanism perform exceptionally well, while the Long Short-Term Memory (LSTM) models with SDP self-attention also exhibit high F-measures and 99.97% accuracy. The Gated Recurrent Unit (GRU) and LSTM models with SDP self-attention have the largest area under the receiver operating characteristic (ROC) curve at 99.97%. These models also achieve the lowest error rates, as evidenced by MAE, RMSE, and MSLE values of 0.0002, 0.017, and 0.0001. The data indicate that GRU and LSTM models with SDP self-attention excel at error reduction.

Table 2. Deep neural network with attention mechanisms in trec spam dataset.

Model	Accuracy	precision	Recall	F-measure	AUC	Error		
						MAE	RMSE	MSLE
RNN + Bahdanua attention	99.38	99.38	99.38	99.38	99.37	0.006	0.078	0.002
RNN + Luong attention	99.55	99.53	99.56	99.54	99.54	0.004	0.067	0.002
RNN + SDP attention	99.28	99.04	99.53	99.28	99.27	0.007	0.084	0.003
RNN + SDP self-attention	99.22	98.97	99.49	99.23	99.22	0.007	0.088	0.003
LSTM + Bahdanua attention	99.39	99.21	99.56	99.39	99.38	0.006	0.078	0.002
LSTM + Luong attention	99.61	99.55	99.67	99.61	99.60	0.003	0.062	0.001
LSTM + SDP attention	99.54	99.48	99.61	99.54	99.54	0.004	0.067	0.002
LSTM + SDP self-attention	99.55	99.59	99.5	99.55	99.54	0.004	0.067	0.002
GRU + Bahdanua attention	99.45	99.55	99.36	99.46	99.46	0.005	0.073	0.002
GRU + Luong attention	99.42	99.09	99.77	99.43	99.42	0.005	0.075	0.002
GRU + SDP attention	99.42	99.33	99.51	99.42	99.41	0.005	0.076	0.002
GRU + SDP self-attention	99.53	99.38	99.68	99.53	99.52	0.004	0.068	0.002

Table 3. Deep neural network with attention mechanism in enron dataset.

Model	Accuracy	precision	Recall	F-measure	AUC	Error		
						MAE	RMSE	MSLE
RNN + Bahdanua attention	99.79	99.6	**100**	99.8	99.79	0.002	0.045	0.0009
RNN + Luong attention	99.78	99.56	**100**	99.78	99.77	0.002	0.046	0.001
RNN + SDP attention	99.78	99.57	**100**	99.78	99.77	0.002	0.046	0.001
RNN + SDP self-attention	99.77	99.54	**100**	99.77	99.76	0.002	0.048	0.001
LSTM + Bahdanua attention	99.74	99.48	**100**	99.74	99.73	0.002	0.051	0.001
LSTM + Luong attention	99.78	99.57	**100**	99.78	99.77	0.002	0.046	0.001
LSTM + SDP attention	99.75	99.51	**100**	99.76	99.74	0.002	0.049	0.001
LSTM + SDP self-attention	**99.97**	**99.94**	100	**99.97**	**99.97**	**0.0002**	**0.017**	**0.0001**
GRU + Bahdanua attention	99.75	99.51	**100**	99.76	99.74	0.002	0.049	0.001
GRU + Luong attention	99.77	99.54	**100**	99.77	99.76	0.002	0.048	0.001
GRU + SDP attention	99.91	99.83	**100**	99.91	99.91	0.0008	0.029	0.0004
GRU + SDP self-attention	**99.97**	**99.94**	100	**99.97**	**99.97**	**0.0002**	**0.017**	**0.0001**

Table 3 recommends the Long Short-Term Memory (LSTM) model with Self-Attention employing the scale dot product (SDP) and the Gated Recurrent Unit (GRU) model with SDP for the Enron dataset. The experimental results demonstrate a strong association between the independent and dependent variables, indicating their significant relationship.

In Table 4, model performance on an extensive custom dataset is presented. The Gated Recurrent Unit (GRU) with scale dot product (SDP) self-attention achieves the highest accuracy at 99.01%. For precision, the GRU model with SDP self-attention scores the highest at 99.01%. The research findings reveal that the

GRU model exhibits the highest recall rate, with SDP self-attention achieving a 99.51% recall. F-measure performance is highest in models using the GRU architecture, with a scale dot product (SDP) self-attention mechanism, and the GRU model with Luong attention also achieving a high F-measure and 99.38% accuracy. The analysis shows that the Gated Recurrent Unit (GRU) model with Self-Attention employing the SDP mechanism achieves the largest area under the ROC curve at an excellent 99.02%. When it comes to error rates, the GRU model with SDP self-attention excels, with the lowest MAE, RMSE, and MSLE values of 0.009, 0.099, and 0.004, respectively. The research uses the GRU model with SDP self-attention on a substantial custom dataset.

Table 5 presents results from an investigation utilizing the SMS spam collection dataset. Several models were tested on this dataset, with the Recurrent Neural Network (RNN) with SDP attention, LSTM with Luong attention, and GRU with Luong attention achieving the highest accuracy of 99.61%. These results highlight the accurate recognition of SMS spam texts by these models. The GRU model with Luong attention demonstrates the highest precision rate, reaching 99.67%. Both LSTM models with Luong and SDP attention mechanisms achieve the highest recall rates and 99.89% accuracy. In terms of F-measure, the LSTM models with Luong and SDP attention achieve the highest values, both at 99.62% accuracy. The analysis reveals that the LSTM model with Luong attention, LSTM model with SDP attention, RNN model with SDP attention, and GRU model with Luong attention exhibit the largest areas under the ROC curve and 99.61% accuracy. Error analysis, including MAE, RMSE, and MSLE, demonstrates that these models achieve the lowest error rates, with MAE values of 0.003, 0.062, and 0.001. These findings highlight the effectiveness of these models in minimizing errors.

Table 4. Deep neural network with attention mechanism in large custom dataset.

Model	Accuracy	precision	Recall	F-measure	AUC	Error		
						MAE	RMSE	MSLE
RNN + Bahdanua attention	98.27	97.94	98.56	98.25	98.28	0.017	0.131	0.008
RNN + Luong attention	97.26	96.42	98.05	97.23	97.27	0.027	0.165	0.013
RNN + SDP attention	96.27	97.07	95.28	96.17	96.25	0.037	0.193	0.017
RNN + SDP self-attention	97.67	96.81	98.5	97.65	97.68	0.023	0.152	0.011
LSTM + Bahdanua attention	98.63	98.09	99.14	98.61	98.63	0.013	0.117	0.006
LSTM + Luong attention	98.70	98.73	98.63	98.68	98.69	0.012	0.113	0.006
LSTM + SDP attention	98.63	98.85	98.36	98.6	98.62	0.013	0.117	0.006
LSTM + SDP self-attention	98.25	97.59	98.89	98.24	98.26	0.017	0.131	0.008
GRU + Bahdanua attention	98.44	97.79	99.08	98.43	98.46	0.015	0.124	0.007
GRU + Luong attention	98.43	98.43	99.38	**99.38**	98.45	0.015	0.124	0.007
GRU + SDP attention	96.20	94.05	98.5	96.22	96.24	0.037	0.194	0.018
GRU + SDP self-attention	**99.01**	**99.01**	**99.51**	99	**99.02**	**0.009**	**0.099**	**0.004**

Table 5. Deep neural network with attention mechanism in sms spam collection dataset.

Model	Accuracy	precision	Recall	F-measure	AUC	Error		
						MAE	RMSE	MSLE
RNN + Bahdanua attention	98.67	98.46	98.9	98.68	98.67	0.013	0.115	0.006
RNN + Luong attention	99.50	99.66	99.33	99.50	99.50	0.004	0.070	0.002
RNN + SDP attention	**99.61**	99.56	99.67	99.61	**99.61**	**0.003**	**0.062**	**0.001**
RNN + SDP self-attention	99.11	98.69	99.56	99.12	99.11	0.008	0.094	0.004
LSTM + Bahdanua attention	98.28	97.61	99.01	98.31	98.28	0.017	0.130	0.008
LSTM + Luong attention	**99.61**	99.61	**99.89**	**99.62**	**99.61**	**0.003**	**0.062**	**0.001**
LSTM + SDP attention	**99.61**	99.34	**99.89**	**99.62**	**99.61**	**0.003**	**0.062**	**0.001**
LSTM + SDP self-attention	99.39	99.45	99.34	99.39	99.39	0.006	0.078	0.002
GRU + Bahdanua attention	99.16	99.23	99.12	99.17	99.17	0.008	0.091	0.003
GRU + Luong attention	**99.61**	**99.67**	99.56	99.61	**99.61**	**0.003**	**0.062**	**0.001**
GRU + SDP attention	98.22	98.99	97.47	98.22	98.23	0.017	0.133	0.008
GRU + SDP self-attention	99.16	99.56	98.79	99.17	99.17	0.008	0.091	0.003

In Figure 2, it is evident that all models exhibit improvement in performance on the Ling spam dataset. However, the GRU model with SDP attention and Luong attention stands out with the highest accuracy of 99.89%. When utilizing 100% Bahdanua attention, the GRU model achieves the maximum accuracy. Notably, at this attention level, RNNs with Bahdanua attention, SDP attention, SDP self-attention, LSTM with Bahdanua, GRU with Luong, and GRU with SDP attention all achieve the highest recall rates. The GRU models with Luong and SDP attention also excel in terms of F-measure, achieving 99.89%. When considering the ROC area, the GRU model with Luong and GRU with SDP attention outperform others at 99.89%. For the Ling spam dataset, our recommendation is the utilization of the GRU model with SDP self-attention.

In Figure 3 and Figure 4, the AUC score analysis reveals that the RNN model with SDP attention attains the highest AUC score of 99.61%. This suggests that the RNN model with SDP attention is particularly effective in handling the complexity of the large custom dataset.

Researchers conducted tests using various models with SMS spam-gathering datasets. Figure 5 demonstrates that the RNN with SDP attention, LSTM with Luong attention, and GRU with Luong attention all achieve the highest AUC scores of 99.61%. These models are proficient in classifying spam in the dataset.

Figure 6 presents the results of the investigation on the Ling spam dataset. The GRU model with Luong and GRU with SDP attention once again outshine others, achieving the highest AUC score of 99.89%.

The Enron dataset was utilized in the research work, particularly in Figure 7 and Figure 8. On the Enron dataset, LSTM and GRU models with SDP self-attention achieve the highest AUC scores of 99.97%. Lastly, in Figure 9, it is apparent that the Trec spam dataset receives the highest AUC score of 99.60 when employing LSTM with Luong attention.

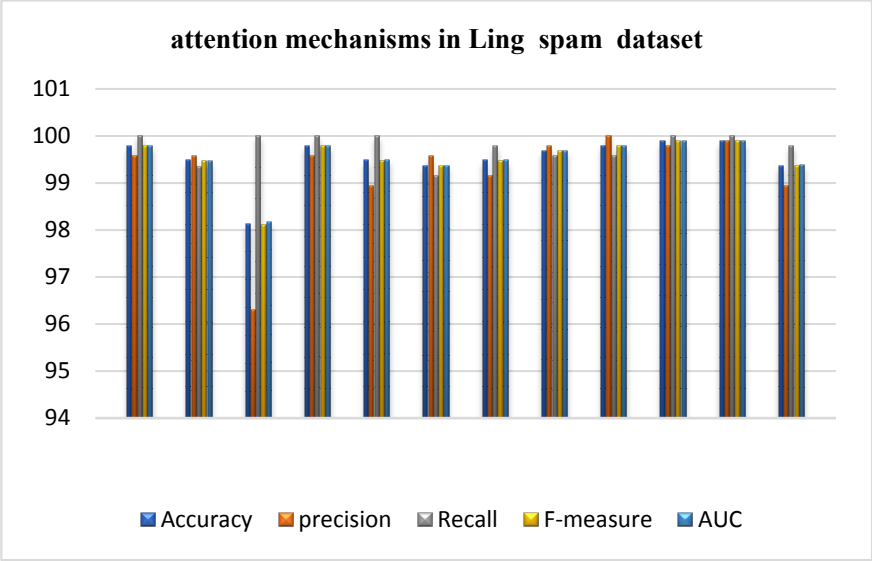

Figure 2. Bar chart of ling spam dataset using attention mechanisms.

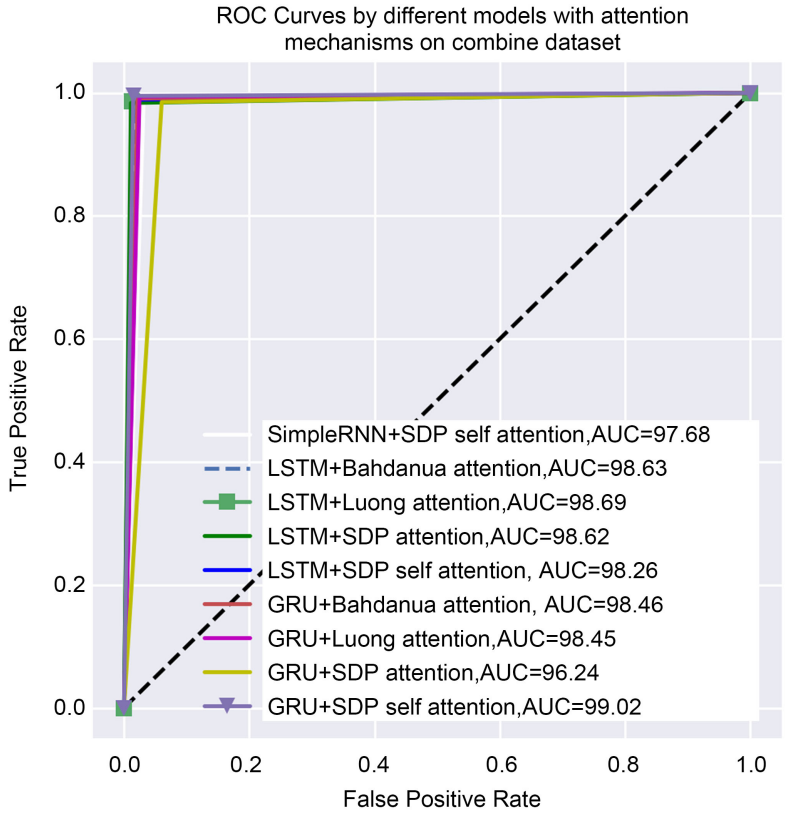

Figure 3. Roc curve of large custom dataset.

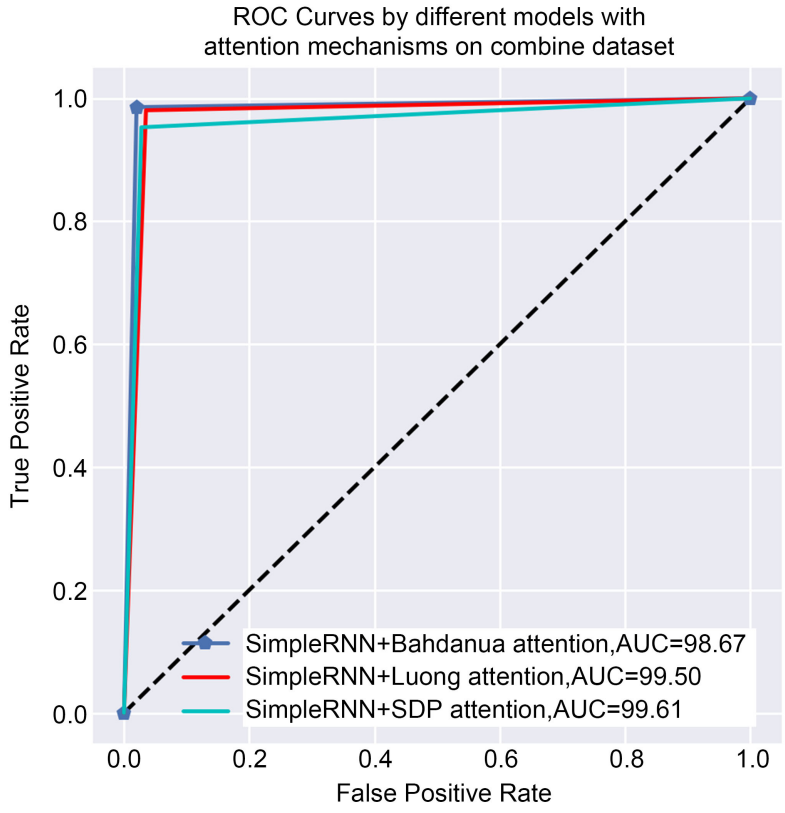

Figure 4. Roc curve of large custom dataset.

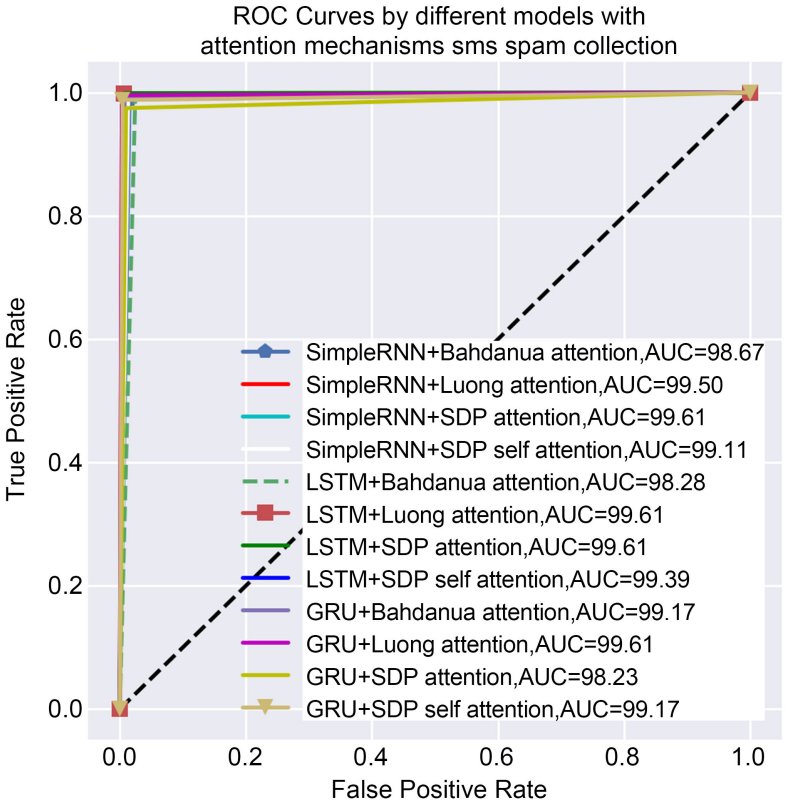

Figure 5. Roc curve of SMS spam collection dataset.

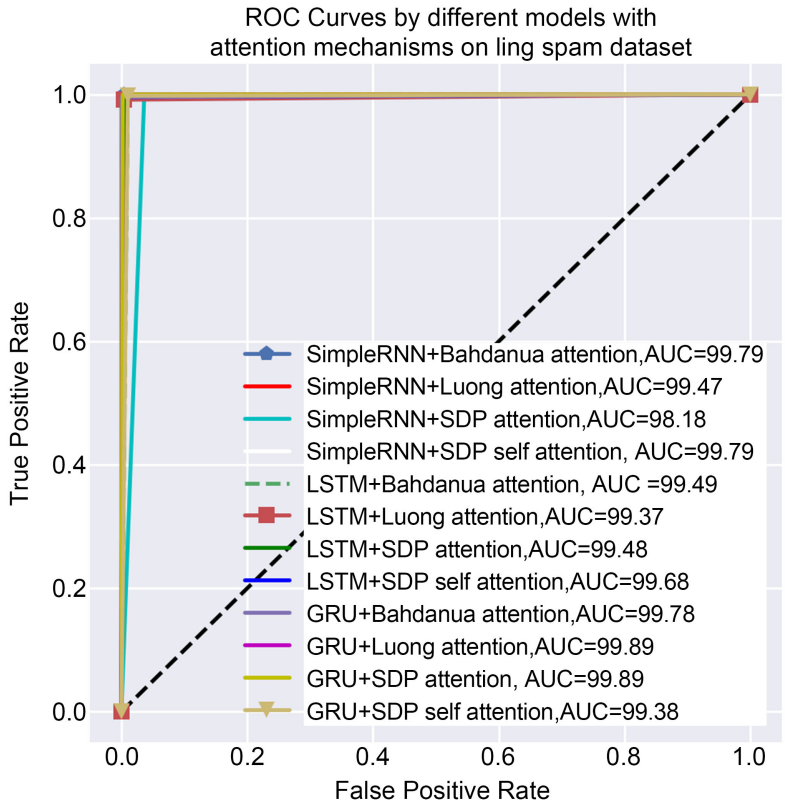

Figure 6. Roc curve of ling spam dataset.

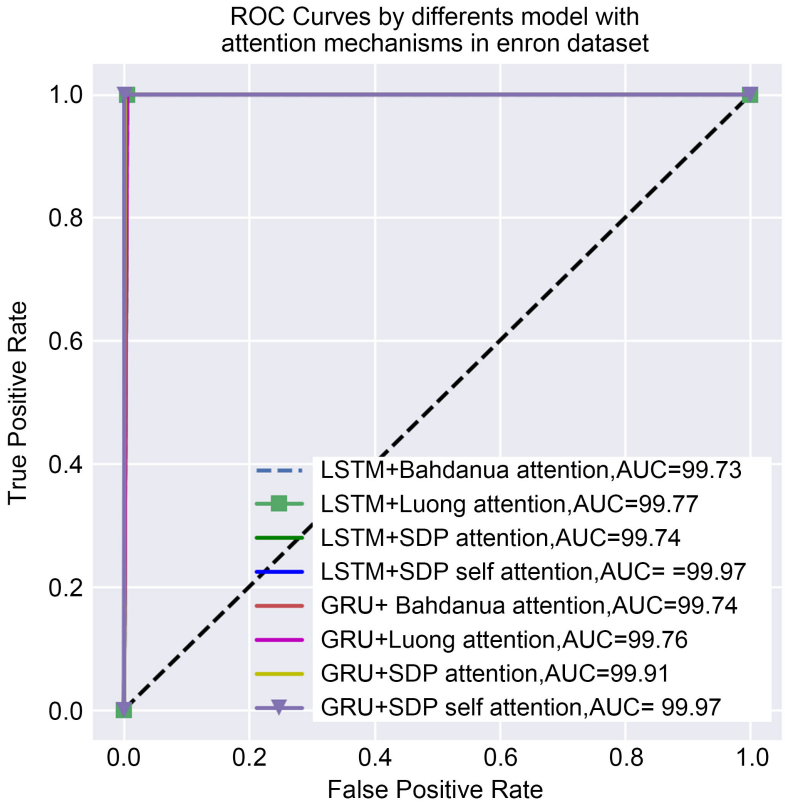

Figure 7. Roc curve of Enron dataset.

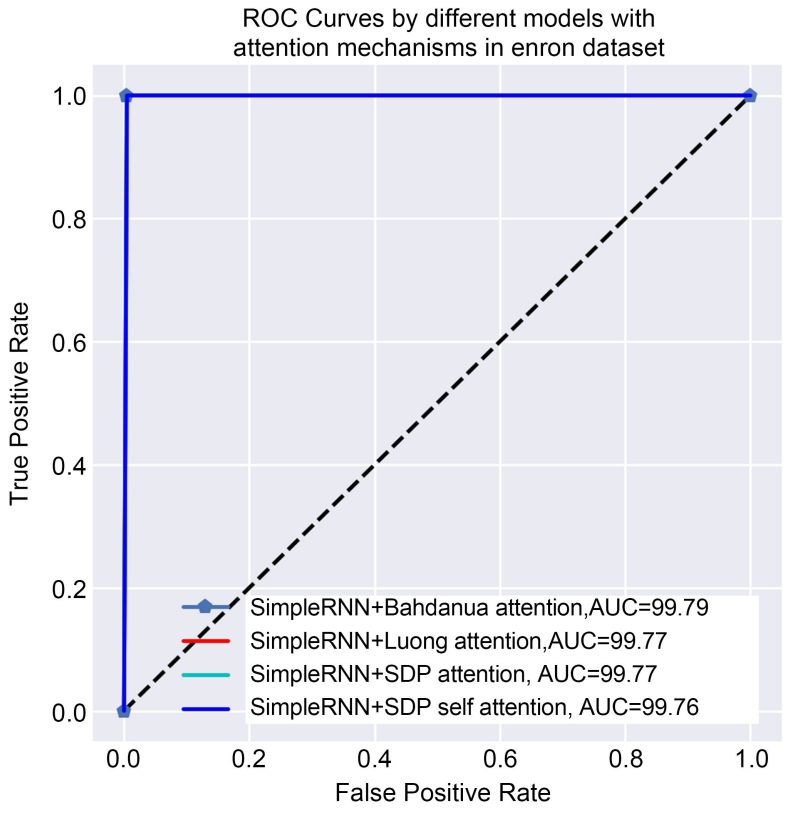

Figure 8. Roc curve of Enron dataset.

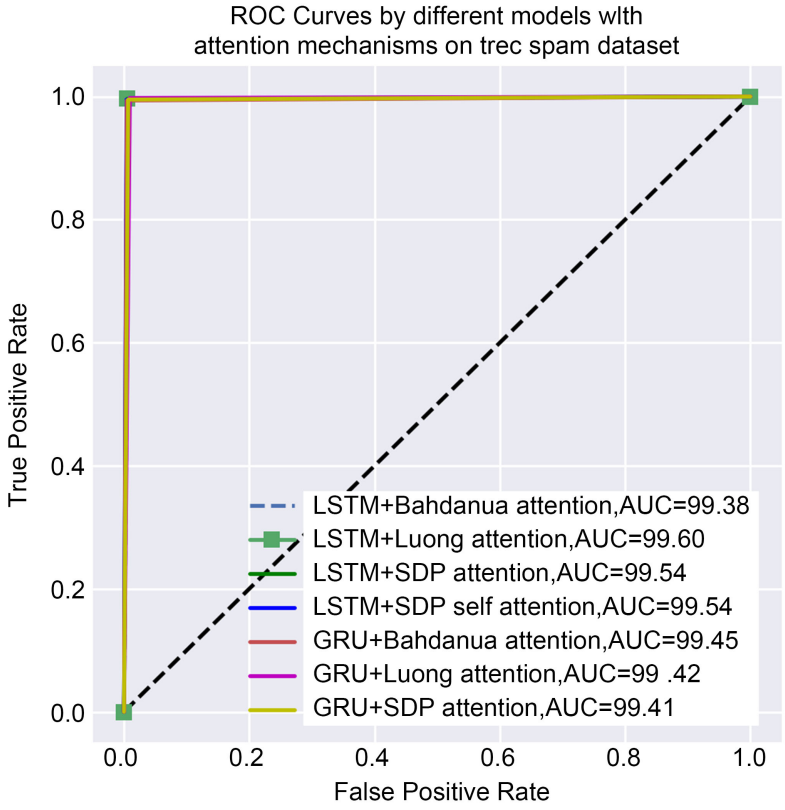

Figure 9. Roc curve of Trec spam dataset.

6. Conclusions

The application employs deep neural networks in conjunction with various attention mechanisms, such as Bahdanua attention, Luong attention, Scale dot product attention, and scale dot product self-attention. Empirical evidence indicates that all these mechanisms are highly effective and successful in their design and implementation. Multiple publicly available datasets can be utilized for research purposes, including the Trec spam dataset, Enron dataset, Ling spam dataset, SMS spam collection dataset, and more. These datasets serve as valuable resources for various studies. While the focus of attention mechanisms has primarily been on machine translation, it's crucial to recognize that attention processes can be applied to problems beyond the realm of spam email classification. The accuracy and efficiency of spam email classification systems can be significantly enhanced by implementing attention mechanism-based approaches. In the case of the Enron dataset, research findings reveal that a combination of LSTM (Long Short-Term Memory) and SDP (Scaled Dot-Product) as a self-attention mechanism achieves an impressive accuracy rate of 99.97%. Furthermore, when applying a Gated Recurrent Unit (GRU) and employing Self-Attention based on the Structured Data Processing (SDP) technique, it yields the highest accuracy score of 99.01. These results stem from a comprehensive analysis of a substantial custom dataset.

In the context of the SMS spam collection dataset, the research demonstrates that utilizing LSTM with SDP attention results in optimal accuracy, reaching a remarkable 99.61%. Furthermore, the investigation of the Ling spam dataset indicates that the best performance is attained by combining the GRU (Gated Recurrent Unit) with either Luong attention or SDP (Self-Attention with Dot-Product) attention mechanisms.

Notably, when applied to the Trec spam dataset, LSTM trained with Luong attention achieves the highest accuracy score at 99.01%. Both the Trec spam dataset and the Enron dataset exhibit exceptional performance across a range of analyses, as documented by the research. Data analysis reveals that utilizing a gated recurrent neural network with a focus on scale dot product attention leads to improved performance in the evaluation process.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Islam, M.K., Al Amin, M., Islam, M.R., Ibna Mahbub, M.N., Hossain Showrov, M.I. and Kaushal, C. (2021) Spam-Detection with Comparative Analysis and Spamming Words Extractions. 2021 9*th International Conference on Reliability, Infocom Technologies and Optimization* (*Trends and Future Directions*) (*ICRITO*), Noida, 3-4 September 2021, 1-9. https://doi.org/10.1109/ICRITO51393.2021.9596218

[2] Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E. and Alegre, E. (2023) A Review of Spam Email Detection: Analysis of Spammer Strategies and the Dataset Shift Problem. *Artificial Intelligence Review*, **56**, 1145-1173. https://doi.org/10.1007/s10462-022-10195-4

[3] Farhana, K., Rahman, M. and Ahmed, M.T. (2020) An Intrusion Detection System for Packet and Flow-Based Networks Using Deep Neural Network Approach. *International Journal of Electrical & Computer Engineering*, **10**, 5514-5525. https://doi.org/10.11591/ijece.v10i5.pp5514-5525

[4] Kuchipudi, B., Nannapaneni, R.T. and Liao, Q. (2020) Adversarial Machine Learning for Spam Filters. *Proceedings of the* 15*th International Conference on Availability, Reliability and Security*, 25-28 August 2020, 1-6. https://doi.org/10.1145/3407023.3407079

[5] Liu, X.X., Lu, H.Y. and Nayak, A. (2021) A Spam Transformer Model for SMS Spam Detection. *IEEE Access*, **9**, 80253-80263. https://doi.org/10.1109/ACCESS.2021.3081479

[6] Shen, H., Liu, X.Y. and Zhang, X.C. (2022) Boosting Social Spam Detection via Attention Mechanisms on Twitter. *Electronics*, **11**, Article No. 1129. https://doi.org/10.3390/electronics11071129

[7] Fang, Y., Zhang, C., Huang, C., Liu, L. and Yang, Y. (2019) Phishing Email Detection Using Improved RCNN Model with Multilevel Vectors and Attention Mechanism. *IEEE Access*, **7**, 56329-56340. https://doi.org/10.1109/ACCESS.2019.2913705

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017.

[9] Yang, Z.C., Yang, D.Y., Dyer, C., He, X.D., Smola, A. and Hovy, E. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/N16-1174

[10] Soni, A.N. (2019). Spam E-Mail Detection Using Advanced Deep Convolution Neural Network Algorithms. *Journal for Innovative Development in Pharmaceutical and Technical Science*, **2**, 74-80.

[11] Luong, M.T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, September 2015, 1412-1421. https://doi.org/10.18653/v1/D15-1166

[12] Vinitha, V.S., Renuka, D.K. and Kumar, L.A. (2023) Long Short-Term Memory Networks for Email Spam Classification. 2023 *International Conference on Intelligent Systems for Communication, IoT and Security* (*ICISCoIS*), Coimbatore, 9-11 February 2023, 176-180. https://doi.org/10.1109/ICISCoIS56541.2023.10100445

[13] Mani, S., Gunasekaran, G. and Geetha, S. (2023) Email Spam Detection Using Gated Recurrent Neural Network. *International Journal of Prograssive Research in Engineering Management and Science*, **3**, 90-99.

[14] Urmi, A.S., Ahmed, M.T., Rahman, M. and Islam, A.T. (2022) A Proposal of Systematic SMS Spam Detection Model Using Supervised Machine Learning Classifiers. In: Bansal, J.C., Engelbrecht, A. and Shukla, P.K., Eds., *Computer Vision and Robotics*, Springer, Singapore, 459-471. https://doi.org/10.1007/978-981-16-8225-4_35

[15] Vinoth, N.A.S. and Rajesh, A. (2023) An Improvised Email Spam Detection Using FSSDL-ESDC Model. *International Journal of Intelligent Systems and Applications in Engineering*, **11**, 618-626.

[16] Sheneamer, A. (2021) Comparison of Deep and Traditional Learning Methods for Email Spam Filtering. *International Journal of Advanced Computer Science and Applications*, **12**, 560-565. https://doi.org/10.14569/IJACSA.2021.0120164

[17] Zavrak, S. and Yilmaz, S. (2023) Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method. *Expert Systems with Applications*, **233**, Article ID: 120977. https://doi.org/10.1016/j.eswa.2023.120977

[18] Ali, N., Fatima, A., Shahzadi, H., Ullah, A. and Polat, K. (2021) Feature Extraction Aligned Email Classification Based on Imperative Sentence Selection through Deep Learning. *Journal of Artificial Intelligence and Systems*, **3**, 93-114. https://doi.org/10.33969/AIS.2021.31007

[19] Zavvar, M., Rezaei, M. and Garavand, S. (2016) Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine. *International Journal of Modern Education and Computer Science*, **8**, 68-74. https://doi.org/10.5815/ijmecs.2016.07.08

[20] Dey, R. and Salem, F.M. (2017) Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. 2017 *IEEE* 60*th International Midwest Symposium on Circuits and Systems* (*MWSCAS*), Boston, 6-9 August 2017, 1597-1600. https://doi.org/10.1109/MWSCAS.2017.8053243

[21] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y. (2015) Attention-Based Models for Speech Recognition. *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Montreal, 7-12 December 2015, 577-585.

[22] Yang, H., Liu, Q.H., Zhou, S.J. and Luo, Y. (2019) A Spam Filtering Method Based on Multi-Modal Fusion. *Applied Sciences*, **9**, Article 1152. https://doi.org/10.3390/app9061152

[23] Rahman, M., Nur, S., Ahmed, M.T., Das, D. and Islam, A.T. (2022) A Feature Engineering Approach for Detecting Cyberbullying in Bangla Text Using Machine Learning. 2022 *International Conference on Recent Progresses in Science, Engineering and Technology* (*ICRPSET*), Rajshahi, 26-27 December 2022, 1-5. https://doi.org/10.1109/ICRPSET57982.2022.10188573